DOCUMENT RESUME

ED 143 551 SE 023 035

AUTHOR Schiffer, Max M.
TITLE Studies in Mathematics, Volume X. Applied Mathematics in the High School.
INSTITUTION Stanford Univ., Calif. School Mathematics Study Group.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE 63
NOTE 157p.; For related documents, see SE 023 028-041; Contains numerous light type

EDRS PRICE MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS *Algebra; Geometry; Instructional Materials; *Mathematical Applications; Mathematics; Optics; *Physical Sciences; *Secondary School Mathematics; *Teaching Guides
IDENTIFIERS *School Mathematics Study Group

ABSTRACT

This publication contains a sequence of lectures given to high school mathematics teachers by the author. Applications of mathematics emphasized are elementary algebra, geometry, and matrix algebra. Included are: (1) an introduction concerning teaching applications of mathematics; (2) Chapter 1: Mechanics for the High School Student; (3) Chapter 2: Growth Functions; (4) Chapter 3: The Role of Mathematics in Optics; (5) Chapter 4: Application of Matrix Algebra. Included in each chapter are background materials, examples, some teaching suggestions, and some exercises. (RH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED143551

SCHOOL
MATHEMATICS
STUDY GROUP

STUDIES IN MATHEMATICS

VOLUME X

Applied Mathematics in the High School

By MAX M. SCHIFFER
Edited by Leon Bowden

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SMSG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM

SE 023 035

STUDIES IN MATHEMATICS

Volume X

APPLIED MATHEMATICS IN THE HIGH SCHOOL

Application appropriate to the S.M.S.G. Curriculum with emphasis on the Elementary Algebra of the 9th Grade and the Geometry of the 10th Grade together with some applications of Matrix Algebra.

A sequence of lectures given to High School Mathematics Teachers by

MAX M. SCHIFFER
Professor Mathematics, Stanford University

Edited by
Leon Bowden
Assistant Professor of Mathematics, Victoria College, U.B.C.

ERIC Full Text Provided by ERIC

Financial support for the School Mathematics Study Group has been provided by the National Science Foundation.

© 1963 by The Board of Trustees of the Leland Stanford Junior University
All rights reserved
Printed in the United States of America

CONTENTS

INTRODUCTION

1. Why Should Applied Mathematics be Taught in the High School?

Jacobi (1804-51), speaking for the pure mathematician, claimed that the motive for mathematical research is "the honor of the human spirit." The same could be said of playing chess: there is no denying its aesthetic and intellectual appeal. So why is the youth who takes his chess as seriously as his mathematics thought to be misguided? Is there such a difference between moving pieces of wood about on a board and manipulating ink marks on paper?

Chess, for all its excellence, is merely a game; unlike mathematics, it is without applications. The miracle of mathematics is that paper work can be related to the world we live in. With pen or pencil we can hitch a pair of scales to a star and weigh the moon. Such possibilities give applied mathematics its vital fascination. Can any subject give the would-be mathematician -- initially, at least -- a stronger and more natural interest?

And what about the non-mathematician? Deny him introduction to this subject, and his appreciation of our cultural heritage must inevitably be inadequate. For mathematics in the broadest sense is instrumental not only to our understanding, but also to our changing the world we live in. And are we not a changing society in a changing world?

2. Difficulties of Teaching Applied Mathematics in the High School.

In our high school systems the teaching of science and the teaching of mathematics have become estranged. To apply mathematics there must be something to apply it to. To apply it there must be a field of application even though there is nothing which you can count as common scientific knowledge among your students. Yet you cannot squeeze many lectures on physics or chemistry or biology into your mathematics course. This is your first difficulty. There is a second. The bulk of mathematics which does apply to other fields is too advanced for your students; you would be talking above their heads.

3. What is the Way Around This Dilemma?

Go back to the beginnings of science, to Archimedes, to Euclid and Heron, to Galileo and Stevinus, when things were very obvious, very simple, and could be explained in a few words. Thus, applied mathematics in the narrow sense of the term, i.e., mechanics, is the ideal topic with which to begin, and is, accordingly, the subject of Chapter 1. For the same reasons, in Chapter 3, I take optics to illustrate the role of mathematics in formulating scientific theories and begin with Euclid's. And although in Chapter 2 (which illustrates important applications of functional and recursive equations to growth problems) things are not always so obvious and simple, and so concisely explainable, the same principle is nevertheless adhered to. Here, its application is slightly less stringent.

4. Rejection of the "Modern" Approach.

Many eyebrows will rise in horror at such a proposal. Modernists will exclaim, "Here we are living in the space age, yet you propose to teach the mathematics of antiquity." Such people never tire of pointing out that whereas today's college courses in the sciences cover twentieth-century science, today's college courses in mathematics deal with eighteenth-century mathematics. They infer that the mathematics currently taught is necessarily out of date.

The emergence of such a conclusion in a technological society is understandable; the inference is none the less fallacious.

Edison's phonograph, Ford's Model T, and the Wright brothers' aeroplane are out of date. But such machines are not out of date because they are old; they are out of date because of rapid technological progress; they have been superseded by more efficient ones of better design. The pyramids of Egypt, although old, are not out of date; progress in the pyramid building line is slow these days; the Egyptian variety, although old, has yet to be superseded. Supersession is not necessarily entailed by newness. Beset by the fad of modernity we must be ever mindful of Aladdin and the cry, "New lamps for old."

ERIC

Technology and science advance hand in hand; each helps the other over obstacles to progress. Likewise, the rudimentary chemistry and biology of the eighteenth century is now, in large measure, out of date. But not, mark you, out of date because, like the pyramids, it is old: out of date because inadequate theories have been superseded by more adequate ones. So we see the sense in present-day college teaching of the sciences giving eighteenth-century developments scant attention. The diligent reader will now exclaim, "But surely twentieth century mathematical developments supersede those of the eighteenth century, and a fortiori, supersede those of antiquity." Such an exclamation indicates grave misconception.

Mathematics is different. Old scientific theories, like old automobiles superseded by better ones, are relegated to the scrap heap. Mathematics usually conserves, seldom scraps. New mathematics is superimposed upon the old rather than the old superseded by the new. As with the successive cities of Damascus, the old is the foundation of the new. Mathematics is cumulative. Concepts thousands of years old are still in use today. Old bricks are used to make new buildings.

The mathematics most immediately applicable to the sciences is mechanics. Roughly but concisely put, mechanics is the alphabet of science. To spell out new theories we need new words, not a new alphabet.

5. Mechanics: The Alphabet of Science.

Our children are both the beneficiaries and the victims of a technological age. Pull a switch, press a button, or move a lever, and a complicated mechanism is set in motion. Turn a knob, and we see and hear the President making a speech in Washington. How does the mere turning of a knob result in the presentation of distant events? With the tremendous jump from the simplicity of primitive machines to the complexity of modern mechanisms, connections are lost sight of. The great illumination of understanding a simple machine, the insight of grasping, say, that the principle of the lever underlies prying off the lid of a can, is lost to the modern child. His docile nonunderstanding of science is limited to accepting the pronouncements of the white-coated, bespectacled man on TV, with his solemn, "Science has proved that ...."

Understanding science starts with mechanics. And mechanics, to borrow a phrase from Pólya, starts with the "congenital or inarticulate" physics we all acquire, willy-nilly, in crawling from the cradle: the experiential facts of pushing and pulling, the properties of sticks and stones; our unavoidable introduction to force, mass, weight, rigidity, flexibility, ... Here is a common background of knowledge for the teacher to exploit. His business is to make this knowledge articulate.

The brilliant, very simple, very obvious, concisely explainable mechanics of Archimedes is the natural articulation. Of course, I do not suggest that the law of the lever can lead in three easy lessons to showing how a TV set works. There is no denying the chasm between levers and electronics. Yet innovators of both had common habits of mind -- the scientific attitude. Something of this attitude can be inculcated by showing how the painstaking application and reapplication of a simple, seemingly trivial concept, the lever, can lead to something complex and deep, the theory of mechanics. In teaching mechanics we take a decisive step toward bridging the gap.

6. A Question of Rigor.

Idealization is inevitable in the applications of mathematics. By this device the complexity of a physical situation is reduced to manageable proportions. A stone becomes a point, a lever a line; knots in beams are ignored, and the wood is presumed to be precisely homogeneous. At this stage justification is practical rather than logical. And occasionally in teaching we introduce an additional assumption with an "It is obvious that ...," or a wave of the hand, and meet a too fussy objection with a shrug of the shoulders. Partial articulation. Such reasoning was good enough, *initially*, for Archimedes, for Galileo, and for Newton. Surely it is good enough for your students' high school *initiation*. Or, do you presume your students abler than Archimedes?

The mechanics of antiquity is not antiquated: its perennial youth is as young as today and as modern as tomorrow.

Liken applied mathematics to a car. Insight, intuition, imagination are its motor, its driving force; rigor, its brakes, the logical checks that control it. Of course a car without brakes is dangerous; imagination must not be allowed to run riot. We need to control imagination and to direct insight. But a car without a motor is useless. We need to drive a little before we need to brake a little.

But axiomatics are the disc brakes of mathematics -- the very latest, up-to-date-est, most rigorous of logical checks. So why content ourselves by teaching mechanics with only an axiom or two? Why not give a full-blown axiomatic treatment?

The object of axiomatics is to find explicitly the absolute minimum of assumptions necessary to a theory. With axiomatics we buy deep enlightenment; the price that must be paid is sophistication beyond the novice. An axiomatist is a man who finally ties a bow tie with the other hand behind his back. Oh yes, it can be done with one hand; oh no, it cannot be done with less. Beginners best use both hands.

Development of geometry did not patiently wait several centuries for Euclid's axiomatization, nor did it wait more than twenty succeeding centuries after Euclid for Hilbert's final dotting of the logical i's and crossing out of the illogical t's. Partial articulation of inarticulate experience necessarily preceded complete articulation.

To teach disastrously, teach with a level of rigor inappropriate to your students or your subject.

7. To Sum Up

I hope to have persuaded you that some understanding of applied mathematics (especially mechanics), liberally conceived, ought to be part of the very fabric of educated common sense, not exclusively the prerogative of the would-be math-[illegible]ics or science specialist.

Being teachers, you know full well that teaching is an art; that to teach effectively you must have the Greek sense of theater, the ability to titillate or irritate the imagination of your students, to make them articulate their experience, so that they would hitch scales to a star and weigh the moon.

Allow me to give the material I believe important to present. Only you can best know how to present it. Teachers are apt to be overawed by university people. While the latter can probably decide what material is important, it is the role of the teacher to decide what aspects can be taught in the high school, when it can be taught, and how best to teach it.

Chapter 1. Mechanics for the High School Student.

1.1 Archimedes' Law of the Lever.

We start with the simplest machine known to mankind, the lever. Supposedly, ever since man developed beyond the level of the ape he has used sticks to lever stones. The Egyptians in building their pyramids used elaborate machines consisting of a combination of levers; yet their knowledge of levers appears to have been largely inarticulate. We all know that in pushing a door shut, the nearer the point at which we push is to the line of the hinges the harder we need push. Yet how many of us realize that this common experience exemplifies the law of the lever? The hinge is the fulcrum about which the turning moment of our push counterbalances the opposing turning moment of friction at the hinge. We have experience but not the articulation.

It seems that Archimedes (287-212 B.C.) was the first in history to ask for the precise mathematical formulation of the conditions of equilibrium of the lever. To ask this question was itself a tremendous step -- to ask for mathematical laws for the behavior of a combination of sticks and stones; for here is a crucial novelty -- that number plays a role in understanding and predicting nature.

We now retrace the essential steps by which Archimedes derived his formulation. He started with the simplest case: a weightless lever with equal arms suspending equal weights. See Fig. 1.

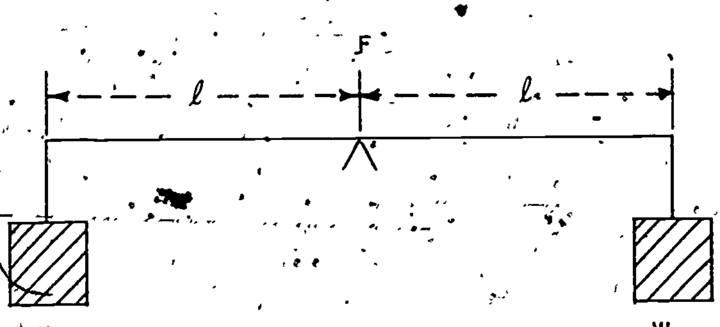

Fig. 1.

Question: Which weight sinks? By the law of insufficient reason there is no more cause for the left-hand weight to sink than the right; by the law of sufficient reason there is as much reason for the left not to sink as for the right; the figure is symmetrical. That is, the lever does not move at all.

We cannot prove this by mathematics. Resort to the law of sufficient (or insufficient) reason is really an appeal to our common experience. So, with Archimedes, we take it as axiomatic that a lever as illustrated in Fig. 1 is in equilibrium. We shall refer to this as Archimedes' Axiom.

At first sight such a beginning seems too trivial to be capable of development. And yet ....? Consider Fig. 2.

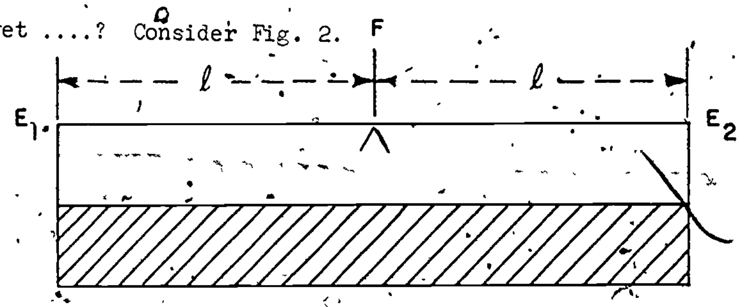

Fig. 2

Here, a homogeneous beam of constant cross section is suspended by a string at each end E_1, E_2 of the lever. If the lever tilts about its fulcrum the beam tilts with it. The same argument is again applicable: by considerations of symmetry there is no reason why the lever (and with it the beam) should tilt either way. We have equilibrium.

Next, consider carefully Fig. 3, a modification of Fig. 2.

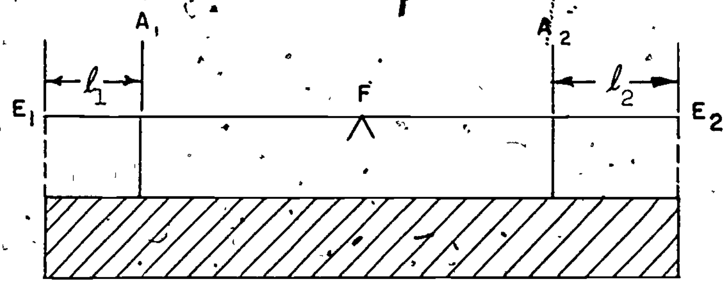

Fig. 3

What changes have been made? The beam is now suspended by strings from A_1 and A_2 (where $E_1A_1 = \ell_1$, $A_2E_2 = \ell_2$), instead of from the ends of the lever E_1, E_2. But, despite these changes, equilibrium of lever and beam remain interdependent; if the lever tilts then the beam must tilt with it; if the beam does not tilt the lever cannot tilt. The difference is that the modified figure is not symmetrical --unless A_1, A_2 happen to be symmetrically placed with respect to F (i.e., not unless $\ell_1 = \ell_2$). At this stage we need introduce further idealization; suppose the strings by which the beam is supported at A_1, A_2 to be weightless. Thus, conceptually, we regain symmetry, and consequently, equilibrium. With weightless strings, we have one body, a lever-*cum*-beam, symmetrically balanced about its fulcrum F with respect to external forces. Whether the tensions in the strings (internal forces) are equal is irrelevant.

Next, we introduce an element of specialization. Study Fig. 4 and understand it.

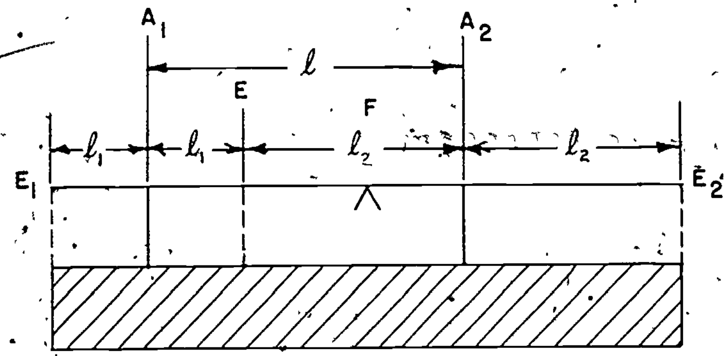

Fig. 4

We take any arbitrary point E on E_1E_2 and select A_1, A_2 to be the special points such that A_1 is the midpoint of E_1E and A_2 the midpoint of EE_2. Since $E_1A_1 = \ell_1$, $A_2E_2 = \ell_2$, it follows that $E_1E = 2\ell_1$, $EE_2 = 2\ell_2$, and since $E_1E + EE_2 = E_1E_2 = 2\ell$ it follows that

$$2\ell_1 + 2\ell_2 = 2\ell$$

i.e.,

$$\ell_1 + \ell_2 = \ell . \qquad (1)$$

But,

$$A_1F = E_1F - E_1A_1 = \ell - \ell_1 ,$$

and

$$A_2F = E_2F - E_2A_2 = \ell - \ell_2 .$$

Hence, by (1)

$$A_1F = \ell_2 , \qquad A_2F = \ell_1 . \qquad (2)$$

This result leads to consideration of Fig. 5.

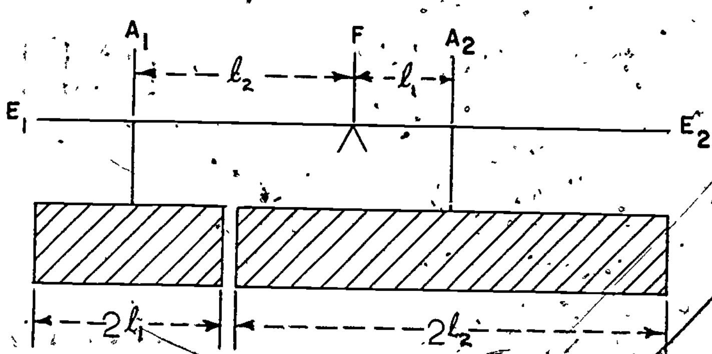

Fig. 5.

Here we conceive the beam to be dissected by the vertical plane through E. Ideally, we suppose there to be no loss of material in cutting the beam, and, consequently, no loss in weight. Thus, provided that there is no change in the distribution of material -- which would result in a change in the distribution of weight -- equilibrium will still obtain. That is, equilibrium still obtains provided that the parts of the cut beam retain the positions they had prior to the cut. Were these to rotate in vertical planes about their points of suspension, the distribution of weight would be changed. But since A_1, A_2 are the midpoints of E_1E and EE_2, we see that these are symmetrically placed with respect to their suspending strings and will remain in equilibrium. Thus the system of lever-cum-two-beams is in equilibrium.

Since the wood is supposed homogeneous, we may, without loss of generality, suppose it of unit density. Thus we have a weight $2\ell_1$ suspended at A_1 counterbalanced by a weight $2\ell_2$ suspended at A_2. That is, by (2) a weight $2\ell_1$ acting at a distance ℓ_2 from the fulcrum counterbalances a weight $2\ell_2$ acting at a distance ℓ_1 from it. This situation is illustrated by Fig. 6.

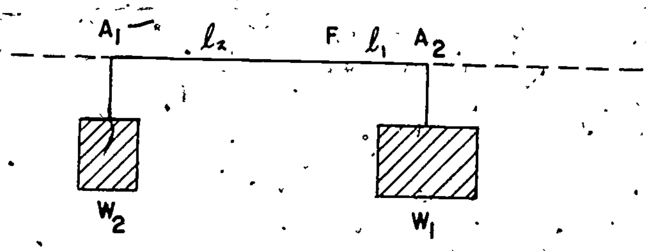

Fig. 6.

What are the conditions for equilibrium? Obviously

$$2\ell_1 \cdot \ell_2 = 2\ell_2 \cdot \ell_1 .$$

Let $W_1 = 2\ell_2$, $W_2 = 2\ell_1$ and we have

$$W_2 \cdot \ell_2 = W_1 \cdot \ell_1, \qquad (1)$$

or

right weight × length of right arm = left weight × length of left arm. The product of the weight and the length of the arm is a measure of the tendency of the weight to turn the arm about the fulcrum. $W_1\ell_1$ is said to be the moment of W_1 about F. So, alternatively put, for equilibrium

right hand moment = left hand moment.

This is Archimedes' law of the lever.

I have shown you how Archimedes' law is devised from "congenital or inarticulate" physics. Actually, this original treatment was somewhat more complicated; what I have given you is a modification due to Lagrange (1736-1813) early in the last century, when such mathematics was not below the dignity of mathematicians of the first rank.

A colleague, when teaching advanced applied mathematics at Stanford last quarter, was so intrigued by this particular proof of Archimedes' law of the lever that he spent two lecture periods discussing just the axiomatic implications of this kind of proving. This involves explication of the notion of symmetry, the distinction between forces external to and internal to a system, that nothing is changed by cutting the beam, and many other considerations. We could, for example, give an alternative proof by considering the beam to be suspended by four strings, one at E_1, two at E, and one at E_2, instead of by strings at A_1, A_2. Then a new Fig. 5 would comprise two beams each suspended by strings at its end points; one suspended from E_1 and E, the other from E and E_2. The final step would be to replace the suspension of each beam for a single string at its midpoint, i.e., strings at A_1, A_2 would replace strings at E_1, E and at E, E_2, respectively. This is an alternative way of obtaining the original Fig. 5.

I mention these matters only to show that the mathematics of Archimedes is not trivial, despite its antiquity. Here there is enough food to satisfy the hungriest thinker. (I have already pointed out the unsuitability of full-blown axiomatics for the beginner.) This is how, starting from "the obvious," the inarticulate physics about which there is common agreement, we build up our mathematics in a cumulative way. I shall further illustrate this <u>cumulative</u> process by making applications of Archimedes' law.

1.2 <u>First Application: The Centroid of a Triangle</u>.

We consider an idealized triangle, made of rigid but weightless material, lying in a horizontal plane, with a weight W suspended from each vertex. Our problem is to find the point at which the triangle can be supported without tilting from the horizontal. See Fig. 7.

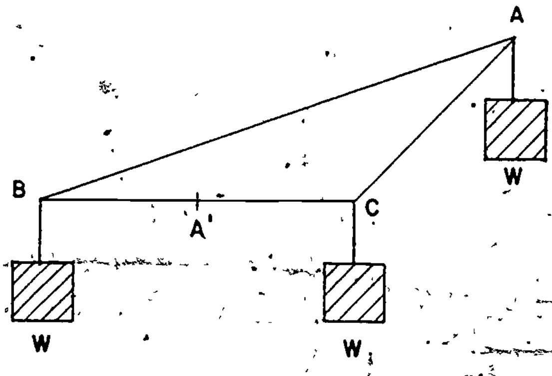

Fig. 7.

But how are we to contend with three forces all at once? We must use what we know, yet the law of the lever is applicable only to two forces. That this law may be applied, we must eliminate the effect of the third weight, say the one at A. We achieve our purpose by introducing a support at A. Now, considering A', the mid point of BC, as the fulcrum of BC, we have a lever with equal weights suspended from equal arms. Thus if the triangle is also supported at A', the points A, A', and consequently the line AA' (a median of the triangle) are fixed, so that the only motion possible is a rotation about AA'. But the forces at B, C counterbalance, so that the triangle is in equilibrium.

Obviously an upward force of W at A will counterbalance that of the weight suspended there. What upward force at A' will counterbalance the downward forces of W at A and at B? When standing on the platform of a weighing machine, your weight, as indicated by the machine, is the same no matter whether you stand on one leg or both. The total downward force is 2W, so we require 2W acting upward at A'. In short, in so far as equilibrium is concerned, the original forces are equivalent to downward forces of W at A and 2W at A'. We have reduced a problem of three forces to a problem of two forces. See Fig. 8.

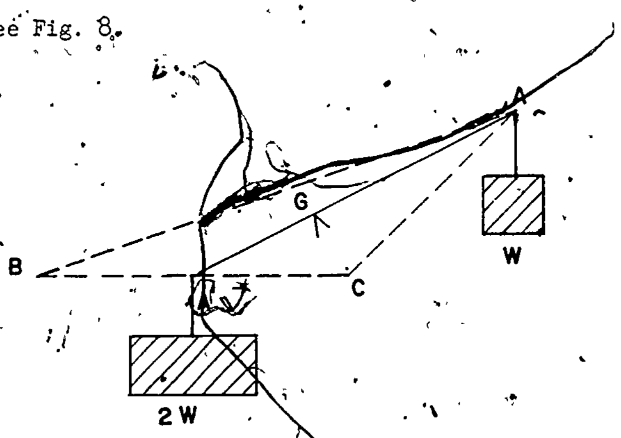

Fig. 8.

The rest is easy, for the law of the lever is immediately applicable to this pair of forces. Let G be the point on AA' such that

$$AG = 2 \cdot A'A$$

so that

$$W \cdot AG = W \cdot (2 \cdot A'A) = 2W \cdot A'A.$$

Thus the triangle is in equilibrium when suspended at G. This solves our problem. Additionally, we may remark that, since the total of the downward forces at A, A' is 3W, we have that the effect of the three equal forces of W at the vertices is equivalent to a force of 3W at G.

Articulation of our common experience and painstaking application of the law of the lever has solved our problem; yet we have by no means exhausted the results inherent in this problem. The argument by which we conclude that the point of suspension for equilibrium is G, two-thirds the way down the median, is equally applicable to the other two medians. There are no grounds for preference. Yet the forces considered can have only one resultant, consequently three medians must be concurrent at G, a point two-thirds the way down each. See Fig. 9.

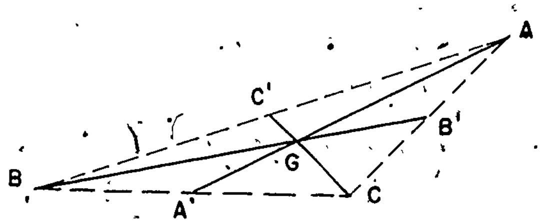

Fig. 9.

In this short deduction we see the interplay between mechanics and geometry. Not only can we use mathematics to deduce laws of nature; we can use laws of nature to deduce more mathematics. Here is an art of which Archimedes was a master.

Another result. Now suppose the horizontally placed triangle of Fig. 9 to be a lamina of homogeneous material. From a simple application of similar triangles it follows that any line segment B'C' parallel to BC is bisected by the median AA'. Consequently the thinner we make a strip with B_1C_1 as one edge, the more nearly rectangular it will become, and the more nearly its geometrical center lie on the median. And with homogenious material a rectangular lamina would, if suspended at its geometrical center, be in equilibrium. Thus we may conceive of the triangle as made up of indefinitely thin strips, with the equilibrium point of each -- and therefore the equilibrium point of all conjoined -- lying on this median. See Fig.

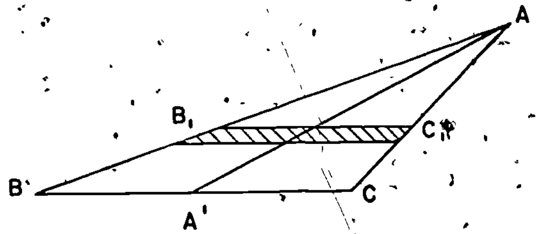

Fig. 10.

But for precisely similar reasons the equilibrium point must also lie on the other two medians, so by the foregoing result this point must be G. Thus the triangle, horizontally orientated, could be maintained in equilibrium by a force equal to its weight acting vertically upwards at G. In short, the multitude of gravitational forces acting on the various bits of the lamina act as if they were all concentrated at G. For this reason G is known as the center of gravity, or centroid, of the triangular lamina.

1.3 Second Application: The Area Under a Parabola.

Archimedes' predecessors and contemporaries had tried, unsuccessfully, to compute the area of an ellipse and the area under a hyperbola. Characteristically, Archimedes tackled the other conic section -- the parabola -- and was successful. His success caused a sensation, as well it might, for his method lies at the threshold of the integral calculus.

Unlike Archimedes, we have the notational convenience afforded by analytical geometry. The problem is to find the area under the parabola $y = ax^2$ between $x = 0$ and $x = h$, i.e., the shaded area OAB of Fig. 11. By considerations of symmetry it is visibly obvious that this is one-half the area OBB_1, one-half that between the given parabola and its mirror image in Ox, $y = -ax^2$. Carefully compare Figures 11 and 12. To any vertical strip PQ (of length ax^2) at a distance x from O in Fig. 11 there is a corresponding vertical strip P'Q' (of length 2ax) at a distance x from O', in Fig. 12. As the midpoint of PQ moves from O to A, and to use a favorite expression of

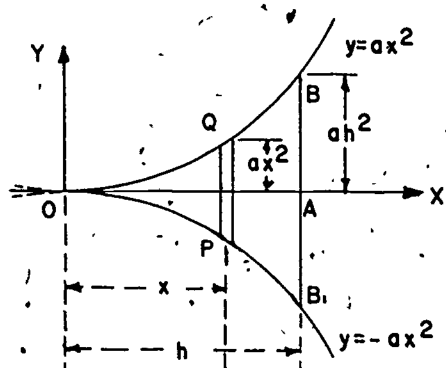

Fig. 11

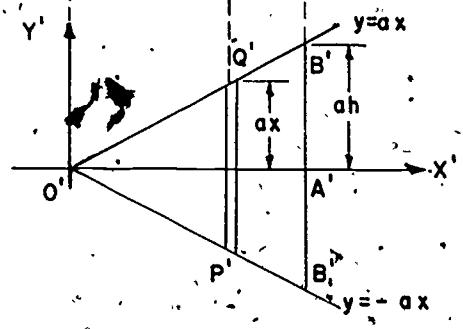

Fig. 12

Archimedes, "fills" the area OBB_1, the midpoint of $P'Q'$ moves from O' to A' and "fills" triangle $O'B'B_1'$.

Now study the conjunction of these figures in a vertical plane given by Fig. 13.

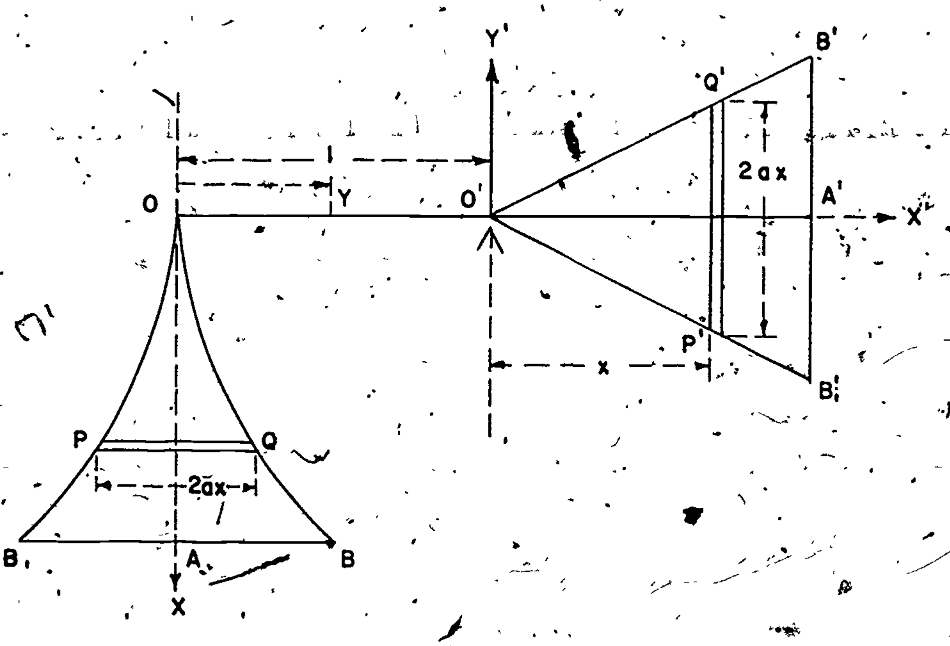

Fig. 13

The lever OA' has fulcrum at O', where OO' = 1. We suppose the corresponding typical strips PQ, P'Q' to have the same width ϵ, and the homogeneous material of both bodies to be of unit density. Thus the weight of the strip PQ is $2ax^2\cdot\epsilon\cdot 1$, and the weight of P'Q' is $2ax\cdot\epsilon\cdot 1$. But, obviously the center of gravity of PQ lies vertically below O, so that the moment of PQ about O' is

$$OO'\cdot(2ax^2\cdot\epsilon\cdot 1) = 1\cdot(2ax^2\cdot\epsilon\cdot 1) = 2ax^2\epsilon\ .$$

And since P'Q' is at a distance x from O', its moment about O' is

$$x\cdot(2ax\cdot\epsilon\cdot 1) = 2ax^2\epsilon\ .$$

Thus,

moment of PQ about O' = moment of P'Q' about O' ,

and the corresponding strips counterbalance one another. But this result holds for each and every corresponding pair! We conclude that

Moment of whole body OBB_1 about O' = Moment of $OB'B_1'$ about O' .

Let W be the weight of OBB_1. $\Delta\ OB'B_1'$ has height $OA' = h$ and base $B'B_1' = 2ah$ and therefore weight $\frac{1}{2}\cdot h\cdot 2ah\cdot 1 = ah^2$. This weight acts as if concentrated at G, the centroid of the Δ. By a previous result G is two-thirds the way along OA', and our last equation becomes

$$W\cdot 1 = \frac{2}{3}\,h\cdot ah^2\ .$$

So, remembering that our materials are of unit density, we have

$$\text{Area } OBB_1 = \frac{2}{3}\,ah^2$$

and, remembering the symmetry,

$$\text{Area } OAB \text{ under the parabola} = \frac{1}{3}\,ah^3\ .$$

We conclude with the elegant result that

$$\text{Area } OAB = \frac{1}{3}\text{ rectangle } OABC\ .$$

Fig. 14.

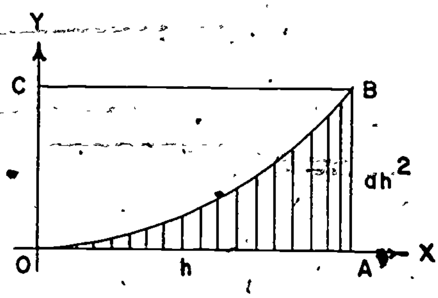

Fig. 14

Of course this proof is not completely rigorous, since the strips PQ, P'Q', of thickness ϵ, are not precisely rectangular. Yet it is intuitively evident that by making ϵ sufficiently small we make the errors of these approximations as small as we please, so that for sufficiently small ϵ the difference between the moments of PQ and P'Q' about O' may be made arbitrarily small. Further articulation would necessitate explication of the notion of limit. To say this is not to sneer at Archimedes' proof by his "mechanical method," as he called it: to the contrary, it is to suggest that intuitive proofs are often indispensable stepping stones to better ones. Archimedes was too good a mathematician to rest content with this proof; he subsequently gave a completely rigorous one by the "method of exhaustion." The discovery (by his mechanical method) of what was the right formula was necessarily prior to proving it right. To cook, first catch your hare.

Archimedes' rigorous proof for the area under the parabola, together with a dozen or so other proofs, including the volume of the sphere, were known to mathematicians of the Renaissance. That he had initially used a "mechanical method" was also known, but not the details. His cooking told nothing of his catching. Cavalieri (1598-1647) devised a way, based on the intuitive consideration that if two figures have equal corresponding strips or cross sections (e.g., PQ and P'Q' in Fig. 13), then the corresponding total areas (or volumes) are equal. It was not until 1906 that a palimpsest giving the details of Archimedes' mechanical method was discovered in Istanbul, and translated by Heiberg (1854-1928), the great Danish expert on Greek mathematical texts. Had this been available to Cavalieri, his development, and consequently that of Fermat, Newton, and Leibnitz, would have been radically different.

Let us recapitulate. We began with the question, "What is the law of the lever?" Geometry, with "inarticulate" mechanics, enabled us to find this law; successive applications of it, reducing a problem of three forces to two, to one, determined the centroid of the triangle -- and gave us, incidentally, a theorem of geometry. The notion of centroid with yet another reapplication of the law of the lever gave us the area under a parabola. This is typical of the way mathematics works: beginnings almost too trivial to take seriously, lead, with repeated applications, to new insights and new discoveries, which, with repeated application, yield yet further insight and discovery.

1.4 Third Application: The Law of the Crooked Lever.

We suppose a homogeneous beam to be freely pivoted in a vertical plane about a (horizontal) nail through its geometrical center F, with weights W_1, W_2 suspended from it as illustrated by Fig. 15 and such that

$$W_1 \ell_1 = W_2 \ell_2 . \quad (2)$$

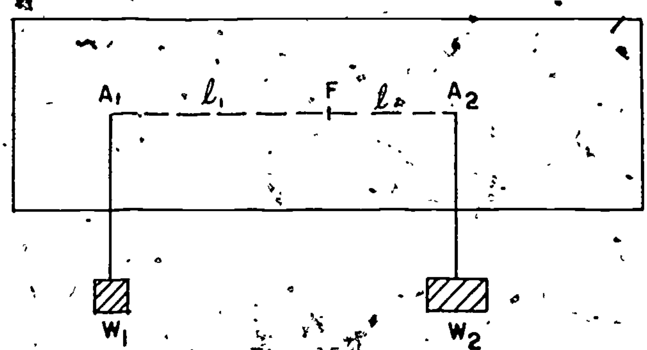

Fig. 15

The homogeneous beam being symmetrically placed about F, its weight has no effect on the equilibrium of W_1, W_2; the whole figure is in equilibrium.

What changes may we make in the suspension of W_1 without disturbing equilibrium? Supposing W_1 to be a constant weight, A_1F must remain unchanged, for otherwise the turning moment of W_1 about F would be altered. But we all know that the vertical pull of a weight on its point of suspension is unchanged by shortening or lengthening the string by which it is suspended, if the string itself is of negligible weight. Clearly, W_1' may be raised or lowered; what matters for equilibrium is that its line of action, its supporting string, passes vertically through A_1. See Fig. 16.

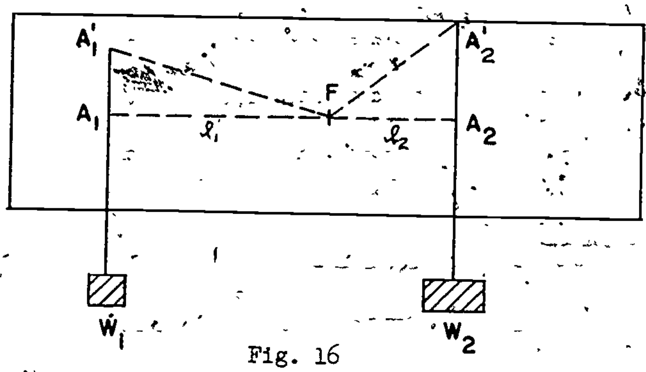

Fig. 16

Study Fig. 16. Does it matter that W_1, W_2 are now suspended from A_1', A_2', respectively, instead of from A_1, A_2? No, for the lines of actions of the two forces (and the forces themselves, of course) are unchanged.

But what is the role of the beam in this scheme of things? Being homogeneous and suspended about its geometrical center, it has no turning moment; it is, in effect, weightless. Its role is given by its rigidity, whereby the turning moments of W_1, W_2 with points of application A_1', A_2' are just the same as if these points had been A_1, A_2. It remains merely to idealize a little more to reject its substance while retaining its rigidity. In short, equilibrium is

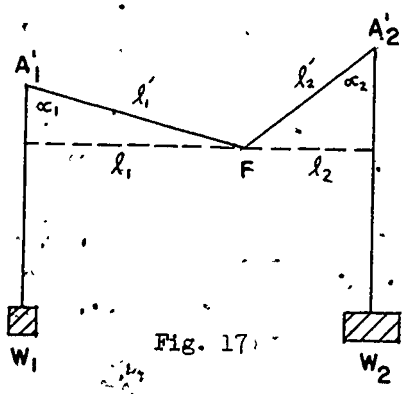

Fig. 17

where $A_1'FA_2'$ is a crooked, weightless, rigid lever.

Hence, if α_1, α_2 are the angles which $A_1'F$, $A_2'F$ make with the vertical, we have

$$\ell_1' \cdot \sin \alpha_1 = \ell_1 , \ \ell_2' \cdot \sin \alpha_2 = \ell_2$$

and, by (2)

$$W_1 \ell_1' \sin \alpha_1 = W_2 \ell_2' \sin \alpha_2 , \qquad (3)$$

the law for crooked levers. That is, the turning moment of a force is now the product of the force, the length of the arm, and the sine of the angle between them. The factor sin α is the price we pay for crookedness. Note that when α_1, α_2 are each 90°, since $\sin 90^\circ = 1$, (3) becomes

$$W_1 \cdot \ell_1' = W_2 \cdot \ell_2' .$$

Characteristically, our new result includes that from which it was deduced. Let us turn to further developments.

1.5 Galileo: The Law of the Inclined Plane.

Galileo (1564-1642) was interested in the mechanics of the inclined plane. He asked and answered the question: Given a weight W on a frictionless plane inclined at an angle α to the horizontal, what force w acting up the plane is necessary to prevent W from sliding down? See Fig. 18.

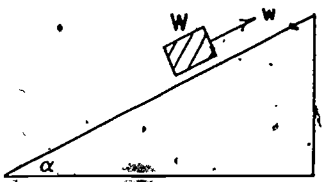

Fig. 18

Note that the precise formulation of the problem is itself a step toward solution. The inarticulate physics of bicycling makes it obvious that the steeper the incline the greater the necessary restraining force. Clearly w is a maximum when $\alpha = 90^\circ$, and must then, without any help from the incline, support the full weight of W; otherwise $w < W$. Thus it is appropriate to denote the restraining force by the smaller letter. This consideration suggests the question: Is the designation of the angle of incline by "α" entirely fortuitous? Consider the notation of (3). Can the solution of Galileo's problem be conceived of as an application of the law of the crooked lever? Yes -- given the ingenuity of Galileo.

First, since vertical forces are better understood, Galileo converts w acting up the inclined plane into a vertical force by introducing a frictionless pulley wheel and a weightless string, thus:

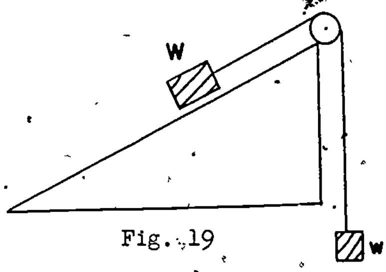

Fig. 19

This strategem may not appear at first sight to advance solution of the problem. But what is the problem? What weight w is needed to counterbalance W? If these are in equilibrium, there is a certain constraint between them. The connecting string being inextensible, if W moves up or down the incline a distance d, then w moves vertically up or down the same distance. Galileo had the great insight to see that this constraint could be realized in a different way -- by the introduction of a crooked lever. See Fig. 20.

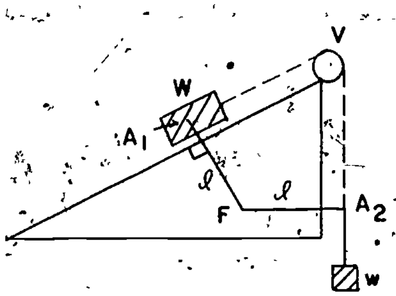

Fig. 20

A_1FA_2 is an equal-armed, crooked (and rigid but weightless) lever with fulcrum F. A_1 is the center of gravity of W, and FA_1 is perpendicular to the inclined plane; A_2 is any point on the line of action of w, and FA_2 is horizontal. To satisfy ourselves that a point F satisfying these requirements exists, it is sufficient to note that the bisector of angle A_1VA_2 is the locus of points equidistant from A_1V and A_2V.

If the lever is rotated about F in a vertical plane, since the lever is equal-armed, A_1, A_2 trace out arcs of equal circles. The smaller the rotation the more nearly these arcs approximate to straight lines, i.e., for infinitesmal rotations the ~~displacement of A_1~~ (the center of gravity of W) along the inclined plane is the same as the vertical displacement of A_2, and therefore the same as that of w. Thus the constraint realized by string and pulley may, alternatively, be realized by the crooked lever A_1FA_2. But we know the conditions for equilibrium with crooked levers, so that the problem is, in principle, solved.

Now, the details.

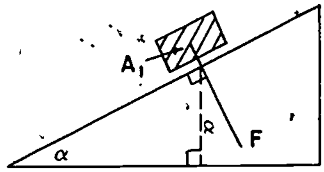

Fig. 21

From Fig. 21 it is clear that the angle between the arm A_1F and the vertical line of action of W at A_1 is α. So, by the foregoing considerations, we see that the conditions for w to maintain W in equilibrium on an inclined plane of angle α are equivalent to those for equilibrium in the following situation.

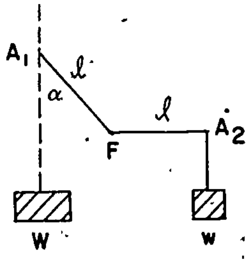

Fig. 22

ERIC

By the law of the crooked lever,

since $w\ell \cdot \sin 90° = w\ell \cdot 1$

$$w\ell = W\ell \sin \alpha$$

so that

$$w = W \sin \alpha .$$

This is the law of the inclined plane.

1.6 Stevin: The Law of the Inclined Plane.

There is another proof, a most elegant proof, due to a Dutch mathematician Simon Stevin or Stevinus (1548-1620). Although Stevin was one of the most brilliant applied mathematicians who ever lived, he is less well known than Galileo; he was not threatened with death by burning at the stake. He invented the first horseless carriage, a sailing carriage for use on the dunes of the Dutch coast; he constructed famous dikes still in use today; and feeling practical need for the facility of decimal fractions, he invented them. For him, mathematics, to be any good, had to be good for something.

Let us see how he proved the law of the inclined plane, that the force acting down it due to W, when the angle of inclination is α, is $W \sin \alpha$. His proof is based on the following figure.

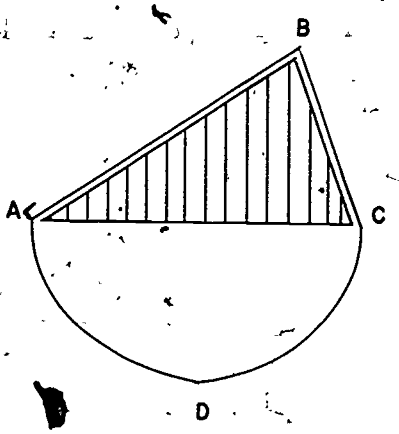

Fig. 23

Stevin was so pleased with his proof that this diagram graced as vignette, with the inscription, "It looks like a miracle, but it is not a miracle," is the title page of his treatise on mechanics. He had good cause for his pleasure; how the law of the inclined plane follows from the equilibrium of a heavy rope, with joined ends, when suspended over a triangular prism, is obvious only to a man of his genius.

Suppose the heavy rope to be in motion initially. This supposition raises the question, when will it stop rotating? Its rotation is caused by the forces acting upon it. But for every particle of rope that goes down, say, at C, an identical particle moves up at A. Thus the configuration of the rope remains unchanged, and consequently the driving forces which initially caused motion still persist. Therefore, since it is rotating initially, it must continue to rotate forever. We have a perpetual motion machine and can use its power to drive a dynamo.

We feel, as Stevin felt, that this conclusion is absurd. But either the heavy rope is in equilibrium or it is not. With him, we have no alternative but to conclude that the rope must be in equilibrium.

Undoubtedly the portion of the rope hanging below the triangle hangs symmetrically; the downward force at A is counterbalanced by an equal downward force at C. Thus, since the rope ABC is in equilibrium before the removal of the portion ADC, it must remain in equilibrium after its removal. That is, the force acting down the one incline due to the weight of the rope BA counterbalances the force acting down the other due to the weight of the rope BC. See Fig. 24.

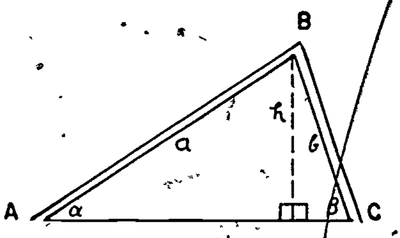

Fig. 24

The force F necessary to prevent a weight W from sliding down an inclined plane of angle θ, depends on θ. F increases as θ increases, F decreases as θ decreases; that is, F is a function of θ, say, $f(\theta)$. Also, of course, F depends on W. If for a given incline W is doubled then F is doubled. F varies both as W varies and as θ varies, that is,

$$F = W\big(f(\theta)\big) . \qquad (3)$$

The problem is to specify $f(\theta)$. See Fig. 25.

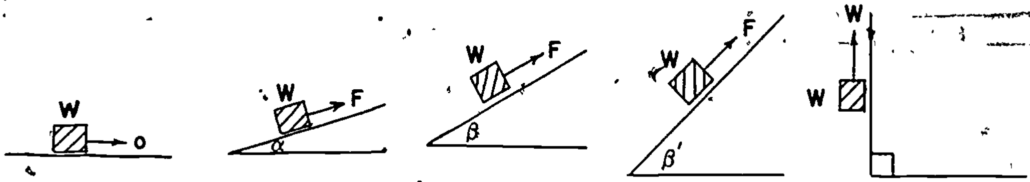

Fig. 25

Let ρ be the density of the rope, i.e., the weight of unit length, so that the weight of the rope AB is $a\rho$ and that of BC is $b\rho$. Then, since AB is inclined at angle α to the horizontal the force F_1 to prevent it slipping down is given by

$$F_1 = a\rho \cdot f(\alpha) . \qquad (4)$$

Likewise, the force F_2 to prevent BC slipping down its incline of angle β is

$$F_2 = b\rho \cdot f(\beta) . \qquad (5)$$

But since the rope AB counterbalances the rope BC

$$F_1 = F_2$$

so, by (3), (4), and (5),

$$a\rho \cdot f(\alpha) = b\rho \cdot f(\beta)$$

and from the geometry of Fig. 24,

$$a = \frac{h}{\sin \alpha} , \quad b = \frac{h}{\sin \beta}$$

giving

$$\frac{h}{\sin \alpha} \rho \cdot f(\alpha) = \frac{h}{\sin \beta} \rho \cdot f(\beta)$$

so that

$$\frac{f(\alpha)}{\sin \alpha} = \frac{f(\beta)}{\sin \beta}$$

But Stevin's argument is applicable to any arbitrary triangle ABC. No matter what non-obtuse angle α we have selected for the one incline, we are free to select β for the other incline <u>independently</u> of our first choice. If we take another case of Fig. 24 with angles α, β' we similarly deduce

$$\frac{f(\alpha)}{\sin \alpha} = \frac{f(\beta')}{\sin \beta'}$$

giving, with our first result,

$$\frac{f(\alpha)}{\sin \alpha} = \frac{f(\beta)}{\sin \beta} = \frac{f(\beta')}{\sin (\beta')}$$

i.e., $\frac{f(\theta)}{\sin \theta} = C$, where C is a constant, and any non-obtuse angle θ. Hence, in (3), we have

$$F = W \cdot C \cdot \sin \theta \qquad (6)$$

It remains to determine C. When $\theta = 90°$, W is as if suspended adjacent to a vertical wall, and clearly $F = W$, substituting in (6), we have,

$$W = W \cdot C \cdot \sin 90° = W \cdot C \cdot 1$$

therefore,

$$C = 1$$

and

$$F = W \sin \theta \; .$$

1.7 In conclusion.

Although I have made this derivation much longer than need be, I feel it well worthwhile to teach. It has the advantage of introducing the notion of a function in a natural way. I know it is "modern" to teach children that a function is an ordered pair, and that a nine-year-old sounds so sweet when he tells you so. Wouldn't it be nice if the nine-year-old knew what to do with an ordered pair? Mathematicians evolved the notion of function because they had a need; it enables them to cope with situations in which this depends on that. That this this-that dependence is up the same logical tree as the son-father relationship comes much later. In teaching, never put logical carts before heuristic horses.

Do remember the inscription Stevin gave his diagram: "It looks like a miracle, but it is not a miracle." The endless rope which does not slide upon the triangle contains, so to speak, the law of the inclined plane. Stevin's achievement was to make this unanalyzed, inarticulate knowledge, articulate. What at first sight is apparently miraculous, we see subsequently to be no more miraculous than other items we regard as self evident. His work is characteristic of the first rate in applied mathematics.

The law of the lever has many other applications, but I have no more time. I hope I have given you some insight into the driving force of mathematics and some idea of how good mathematicians go, initially, about their business.

What work have we considered? Archimedes' as simplified by Lagrange, then Galileo's, then Stevin's. The sequence is not entirely fortuitous. Mankind has found its way by groping, by trial and successive correction, by closer approximation to the truth. Oh, yes, there have been tremendous blunders in the development of mathematics and science, but broadly speaking the work of first-rate men of one era has been used as a foundation for their work by the first-rate men of the succeeding era. Mechanics, as we have said, is the alphabet of of science. Thus the sequence in which fruitful concepts evolved is a first indication of the sequence in which to teach them. The history of ideas concerns itself with all concepts, good, indifferent, and bad. To the contrary, the genetic method concerns itself only with the good ones, except insofar as their contrast with bad ideas can serve to show what makes better ideas better.

To conclude this lecture may I remind you that the initial development of mechanics was not a full-blown axiomatic treatment. Are your students abler than Archimedes?

Chapter 2. Growth Functions.

Hats, therefore hat pegs; growth, therefore growth functions. What could be a more natural introduction to the concept of function than growth problems? In the first section I show how the exponential law of growth is derived from a functional equation that arises naturally from its context. In the second section we consider an application by Maxwell of this result. Next, in considering population growth we are led from functional to recursive equations. Their use is further elegantly exemplified in the fourth section by considering the "growth" of the number of sides of a regular polygon of fixed perimeter, thereby giving Cusanus' formula for π.

2.1 The Exponential Law of Growth.

How much timber is there in a forest? Trees grow. The older the forest, the bigger the trees. The bigger the trees, the greater the amount of wood. Provided that there are no forest fires and no trees die, the volume of wood increases with time. The volume of timber depends upon when the forest was planted; it is a function of the time for which the trees have been growing.

Doesn't this situation invite introduction of a mathematical notation? We introduce one. "$N(t)$" is to be read as "the volume of timber (in cubic feet, say) in a given forest t years after it was planted." Thus $N(0)$ is the volume of timber in the given forest 0 years after it was planted, i.e., $N(0)$ is the volume of timber initially.

But the volume of timber in the given forest after t years, $N(t)$, does not depend only upon the time for which it has been growing; also, it depends upon the size of the forest originally, $N(0)$. And how does $N(t)$ depend upon $N(0)$? We take it as obvious that if a forest had originally been twice as big then it would now be twice as big as it is; if originally three times the original size, now three times its present size; if originally four times the original size, now four times the present size; and so on. More precisely, supposing that the given forest has been growing for t_1 years, its present volume of timber would be $N(t_1)$, or $2N(t_1)$, or $3N(t_1)$, or $4N(t_1)$, or whatever, according as its original volume was $N(0)$, or $2N(0)$, or $3N(0)$, or $4N(0)$, or whatever it was, respectively. Thus the relation between $N(t_1)$ and $N(0)$ is such that

$$\frac{N(t_1)}{N(0)} = \frac{2N(t_1)}{2N(0)} = \frac{3N(t_1)}{3N(0)} = \frac{4N(t_1)}{4N(0)} = \ldots = k_1, \text{ where } k_1 \text{ is a constant}.$$

so that

$$N(t_1) = k_1 N(0). \qquad (1)$$

This is the mathematical statement of the assumption that the present (at time t_1) volume of timber is directly proportional to the original amount. This assumption merely amounts to assuming non-interference of different trees in the forest, one with another; they grow independently.

If, alternatively, we had supposed the forests to grow for t_2 years, we would have concluded that

$$N(t_2) = k_2 N(0) \qquad (2)$$

where k_2 is some constant. This raises the question: Does $k_1 = k_2$? Well, suppose them equal. What follows?

$$N(t_1) = N(t_2)$$

i.e., that the volume of timber in the given forest is the same after t_2 years as it was after t_1 years. The consequence is that there would have been no growth at all for $t_2 - t_1$ years.

But with forethought we could have foreseen this consequence. (1) or (2) tells but half the story; the present size of the forest depends not only on its original size, but also on the time for which it has been growing; $N(t)$ depends on t as well as on $N(0)$. So, to tell the whole story, k_1 in (1) must depend on, or be a function of, t_1, and k_2 in (2) must depend on, or be a function of, t_2. That is, (1), (2) must be exemplifications of a law of the pattern

$$N(t) = k\ N(0)$$

where k depends on t.

Although we know <u>that</u> k depends on t, we do not know <u>how</u> k depends on t. So we must leave the nature of the relationship unspecified, and write that $k = f(t)$, giving

$$N(t) = N(0) \cdot f(t) \ . \qquad (3)$$

Note that putting $t = t_1$, $t = t_2$, successively, we get

$$N(t_1) = N(0) \cdot f(t_1) \ , \quad N(t_2) = N(0) \cdot f(t_2)$$

so that $k_1 = f(t_1)$, $k_2 = f(t_2)$

so that k_1, k_2 are constants as required by (1), (2); but that k is fixed in value for a given value of t is not to imply that k has the same fixed value for different values of t.

With hindsight, we can now see (3) as obvious. If f(t) is the volume of timber in one tree after t years of growth, then N(0) trees growing for the same period have a total value of timber of $N(0) \cdot f(t)$, the volume of the forest after t years, N(t).

We must use (3) to specify f(t). Suppose for ease of exposition that our forest was planted at the turn of the century. Then 5 years later, in 1905, its size N(5) satisfies the equation

$$N(5) = N(0) \cdot f(5) \ . \qquad (4)$$

What is its size in 1911? What, in other words, is the size (5 + 6) years after planting? Two ways of answering this question now present themselves: the one in terms of its growth since it was planted in 1900, (5 + 6) years ago; the other in terms of its <u>additional growth</u> since 1905, 6 years ago.

By (3) the first answer is

$$N(5 + 6) = N(0) \cdot f(5 + 6) \quad . \qquad (5)$$

The second answer is slightly less obvious. We now consider our forest as if it had been planted in 1905 with the initial size

$[N(0) \cdot f(5)]$ -- see (4) -- and had grown for only 6 years. By (3), we have

$$\overline{N}(6) = [N(0) \cdot f(5)] \cdot f(6) \qquad (6)$$

(the bar in "$\overline{N}(6)$" is used to remind us that "6" refers to 6 years after 1905, not 1900).

But these two answers, given by (5) and (6), must be the same, for $N(5 + 6)$, the volume of wood in our forest $(5 + 6)$ years after 1900, is the volume of wood there, $\overline{N}(6)$, 6 years after 1905. Consequently,

$$N(0) \cdot f(5 + 6) = [N(0) \cdot f(5)] \cdot f(6)$$

which gives the functional equation

$$f(5 + 6) = f(5) \cdot f(6) \quad .$$

The specific periods, 5 and 6 years, were used for ease of exposition. The argument may be repeated with the unspecified arbitrary periods t_1, t_2, giving

$$f(t_1 + t_2) = f(t_1) \cdot f(t_2) \quad , \qquad (7)$$

the functional equation that the function of the sum is equal to the product of the functions.

Note that to deduce this equation I did not need any technical knowledge of biology or forestry. That I merely made articulate what we all know even though we never stopped to think about it is evidenced by your immediate acceptance of my premises.

Let us use the functional equation, (7), to specify $f(t)$. Putting $t_1 = t_2 = t$,

$$f(2t) = f(t) \cdot f(t) = f(t)^2 ,$$

and with $t_1 = t$, $t_2 = 2t$,

$$f(3t) = f(t) \cdot f(2t) = f(t) \cdot f(t)^2 = f(t)^3 .$$

These results lead us to suppose that

$$f(\overline{n-1} \cdot t) = f(t)^{n-1}$$

the consequence of which, since

$$f(nt) = f(t+\overline{n-1}\cdot t) = f(t) \cdot f(\overline{n-1}\cdot t) \qquad \text{by (3)},$$

is that

$$f(nt) = f(t)^n . \tag{8}$$

But $f(1 \cdot t) = f(t) = f(t)^1$

so that (8) holds when $n = 1$, and consequently by the principle of mathematical induction (8) holds for every positive integer n.

Thus, we have

$$f(nt) = f(t)^n \tag{8}$$

$$f(mt) = f(t)^m \tag{9}$$

where n, m are positive integers.

Putting $t = \frac{1}{n}$ in (8)

$$f(1) = f\left(\frac{1}{n}\right)^n .$$

Taking the n^{th} root

$$f(1)^{\frac{1}{n}} = f\left(\frac{1}{n}\right)$$

Raising to the m^{th} power,

$$f(1)^{\frac{m}{n}} = f\left(\frac{1}{n}\right)^m .$$

But putting $t = \frac{1}{n}$ in (9),

ERIC Full Text Provided by ERIC

$$f\left(\frac{m}{n}\right) = f\left(\frac{1}{n}\right)^m .$$

Therefore,

$$f\left(\frac{m}{n}\right) = f(1)^{\frac{m}{n}} .$$

Putting $\frac{m}{n} = t$ and $f(1) = a$, we obtain specification of $f(t)$, namely,

$$f(t) = a^t. \qquad (10)$$

where a is a constant since $f(1)$ is a constant, and t is any positive rational since m,n are arbitrary positive integers. This is known as the exponential function.

Hence, by (3) the law of growth for our forest becomes

$$N(t) = N(0) \cdot a^t \qquad (10),$$

the value of a depending upon the kind of forest considered.

(Strictly speaking $f(t)$ has been defined only for rational values of t; but if it is conceded that a forest grows continually, then obviously (10) is to be accepted for all (real) values of t. Whether this point should be discussed or not discussed depends upon the maturity of your students.)

We have answered the question: How much timber is there in a forest? Yet it takes but slight reflection to see that the law of growth need not be applicable solely to forests. Of course, it is applicable to any phenomenon whose growth occurs as the growth of trees occur. And how do trees grow? Trees grow in such a way that the amount of growth made in any period is proportional to the amount of wood growing at the beginning of that period.

It is important to be clear on this point. Part of our inarticulate common knowledge, it is readily articulated by the law of growth. $N(t)$, $N(t + 1)$, $N(t + 2)$ being the volumes of timber in a given forest at the end of t, $t + 1$, $t + 2$ years, respectively, $N(t + 1) - N(t)$, $N(t + 2) - N(t + 1)$ are the amounts of growth in the $(t + 1)$th and $(t + 2)$th years. By (10),

$$N(t + 1) - N(t) = N(0) \cdot (a^{t+1} - a^t) = N(0) \cdot a^t(a - 1) = (a - 1) \cdot N(t) .$$

Similarly,

$$N(t + 2) - N(t + 1) = (a - 1) \cdot N(t + 1) .$$

In words: the amount of growth in the t^{th} year is $(a - 1)$ times the amount available to growth at the beginning of that year, and the amount of growth in the $(t + 1)^{th}$ year is $(a - 1)$ times the amount available to growth at the beginning of that year. But that t is measured in years is irrelevant; we could have measured t in seconds; we could have taken one-millionth of a second to be our unit of time. It follows that the amount of growth in any instant is proportional to the amount of material available to growth at the beginning of that instant; i.e., that the instantaneous rate of growth is proportional to the amount of growing material.

What grows in this way, as trees grow? If in (10) <u>a</u> were less than 1 the trees would grow smaller, i.e., decay. It is known that the rule of decay of radioactive material is proportional to the amount of material available; consequently the law of growth is applicable. When a ray of light passes through an absorbing medium, the intensity of the light is weakened by the passage; the weakening is proportional to the intensity. Thus the law of growth is also applicable here. We have

$$I(x) = I(0)\, a^x$$

where $I(0)$ is the intensity of the incident light ray at the surface of the absorbing medium, $I(x)$ the intensity at a depth x within the absorbing medium, and a (less than one) the absorption factor.

Compound, as opposed to simple, interest is another example. With simple interest the rate of growth of the investment (supposing interest to be left on deposit) is <u>constant</u> and is proportional to the capital invested <u>initially</u>. The amount of interest earned in the thirtieth year is the same as that earned in the third year. If, to the contrary, interest payable on capital is permitted to accrue as additional capital and the total capital to date (i.e., initial

ERIC

investment plus accrued interest) grows at a rate proportional to the total capital to date (not proportionally to the capital initially), then interest is said to be compounded.

If interest is permitted to accrue as capital at yearly intervals, the interest is said to be compounded annually. Of course, in this event the amount of interest earned in the thirtieth year vastly exceeds that earned in the third year, for the capital grows with the interest. The formula is

$$C_n = C_0 a^n \qquad (11)$$

where C_0 is the capital initially, C_n the total capital at the end of n years, and a a constant depending upon the annual rate of interest. If interest is compounded at more frequent (or less frequent) regular intervals, then n is to be taken as the number of times interest has been permitted to accrue as capital, C_n the total capital after the n^{th} increment, and a a constant depending upon the rate of interest for the intervals in question, semiannual, quarterly, or whatever it may be. With this application of the law of growth there is merely the difference that n is restricted to integers.

Interest could be compounded daily or at far more frequent intervals. Though your bank manager might not agree, you could argue that an instant after investing your capital you should be entitled to an instant's interest. Of course, calculated pro rata with the annual rate this would be small. Nevertheless, with your money growing continually you might be tempted to suppose that you would become infinitely rich in a year or two -- until you tempered your wishful thinking with the somber reminder that this growth would also be governed by the general law of growth. It turns out that if your capital C_0 was invested at 100% per annum compounded instantaneously, then your total capital C_n at the end of n years would be given by

$$C_n = C_0 e^n$$

where e, a number of great importance in mathematics, is the base of Naperian logarithms. This formula was first deduced by Bernoulli. Note that it exemplifies the general law, with $a = e$. The only difference is that since the increment periods are instantaneous, n is not restricted to integral values. Putting $n = 1$, we conclude that in the time C would double itself at 100% simple interest, with the interest continually compounded it becomes $C_0 e$. Approximately $e = 2.718$, so, investors please note, while the one way $100 becomes $200, the other way it grows to nearly $272.

To recapitulate: I have shown how the concept of function and the barest rudiments of functional equation theory may be used to deduce the exponential law of growth, and I have indicated fields of application.

2.2 Maxwell's Derivation of the Law of Errors.

In this section we consider Gauss' law of errors (Gauss 1777-1855). We shall find that Maxwell's (Maxwell 1831-1879) ingenious derivation of it depends upon the solution of a functional equation. This solution is an application of the exponential law functional equation considered in the last section.

When at the beginning of the last century astronomers, physicists, and surveyors started to make very precise measurements, it was realized that there is no such thing as an absolutely accurate measurement.

First consider the question of a single observation. Astronomers chart the stars as accurately as they know how, yet two astronomers seldom observe the same star as being in the same position -- though it is in the same position. The figures expressing their measurements are apt to differ in the last decimal place or two.

To come nearer home, the spring in your bathroom scale becomes fatigued and loses a little of its springiness. With changes in temperature bits of metal alter in length and so modify its mechanism. If over-conscientious about your weight, you may evade many of these contributions to inaccuracy by resorting to an equal-arm balance of appropriate dimensions. But even the arms of balances become tired and droop a little. Better designed and more carefully constructed instruments measure more accurately, yet it is always a question of more or less better; there are no absolutely accurate measuring devices. We include the human eye reading a pointer against a graduated scale.

We suppose you, afflicted by a weight-reducing fad, weighed yourself on three bathroom scales this morning, their readings being 201, 207, 204 pounds. Your problem: What was my weight this morning? Possibly you would, in the absence of a known weight with which to test the scales, take the arithmetic average, 203 pounds, as correct. You would conclude almost with certainty that you did not weigh 300 pounds, and think it very unlikely that you were as much as 250. Surely the further removed the estimated figure from 203 or thereabouts, the more unlikely its correctness.

Uninvited, the notion of probability intrudes upon the scene. Uncertain of the correct figure we cannot be certain of the error of the measured reading; the most we can ask is such questions as,"What is the likelihood that the observed reading n does not differ from the actual measure by, say, more than $\frac{1}{100}$ n?" The general answer to questions of this sort is called the law of errors. With this answer we shall be presently concerned.

Secondly, consider the question of the combination of observations. Although hundreds of physicists have made measurements from which to deduce the velocity of light, no two physcists have obtained exactly the same result. The deduced number being dependent upon several measurements, each subject to error, the final result necessarily incorporates a combination of these errors.

Consider, for simplicity, the following example. A square lamina of side 5 units is measured as having sides of 5.1 and 4.9 units. So whereas the actual area is 25 square units, the area deduced on the basis of our measurements is 24.99. Although there is a 2% error in each of our measurements there is only a $\frac{1}{25}$ of 1% error in the final result. One measurement was too big, the other too small, so that each error tends to annul the inaccuracy due to the other. But this oversimplifies; the point being that we never know with certainty the actual errors. A more realistic question is: If it is 95% certain that the error in each of our measurements does not exceed 2%, what is the probability that the error in the area calculated on the basis of these measurements does not exceed, say, 1%?

We have briefly indicated the kind of problem this line of thought leads to: now we must return to what it leads from, the probability of such and such an error in a single measurement. As we have said, the general answer to this latter question is known as the law of errors.

The law was first derived by Gauss in the masterly way characteristic of this great mathematician; but his approach to the problem was so abstract that Maxwell, among others, was only partially convinced of the correctness of his derivation. It lacked that down-to-earthness found in, for example, Stevin's deduction of the law of the inclined plane. Maxwell was led to examine Gauss' proof when he needed the law of errors to further develop statistically the kinetic theory of gases. He was concerned with the down-to-earth conception of the behavior of a gas as that of billions of molecules darting to and fro, pushing against the walls of their enclosure, so it is perhaps not too surprising that he came up with a marvelous, immediately graspable, proof. Yet on second thoughts it is most surprising; many contemporary physicists shared his dissatisfaction, but none his discovery. Such is the prerogative of genius.

From the problem of molecules impinging on the walls of their enclosure, Maxwell turned to that of bullets hitting a target. Let us consider his derivation of the law of errors.

Consider the marksman who misses the bull's-eye. Typically, the (printable) phrase he uses to describe his shot, is one of the following sort: to the right of center; on center, but too far to the left; on center, but too high; to the right of center and too high; left of center and low. He refers to his bullet's position as a combination of two errors; a horizontal and a vertical deviation from the bull's-eye. Taking our cue from him, we introduce rectangular coordinate axes with origin at the bull's-eye and x-axis horizontal. $H(x,y)$ is the position of his hit.

If a marksman is standing in a fixed position at a certain distance from the target, what is his probability of hitting the bull's-eye? First, this will depend upon the size of the bull's-eye. Surely we are agreed that, if it is no bigger than the point of a pin, then it is practically impossible to hit; and that if it is conceived of as a mathematical point, then the probability of hitting it is zero. Thus we must reformulate our questions: instead of asking, "What is the probability of hitting (0,0)?" we must ask, "What is the probability of hitting the target within the neighborhood of (0,0)?" The general question is, "What is the probability (when aiming at (0,0)) of hitting within the neighborhood of (x,y)?"

Obviously, the probability will depend upon the size of the neighborhood; take the whole world for the neighborhood of (0,0), and the marksman cannot miss. The neighborhood must be specified. It is natural to take the rectangle of sides Δx, Δy, centered on (x,y) as the neighborhood of (x,y). See Fig. 1.

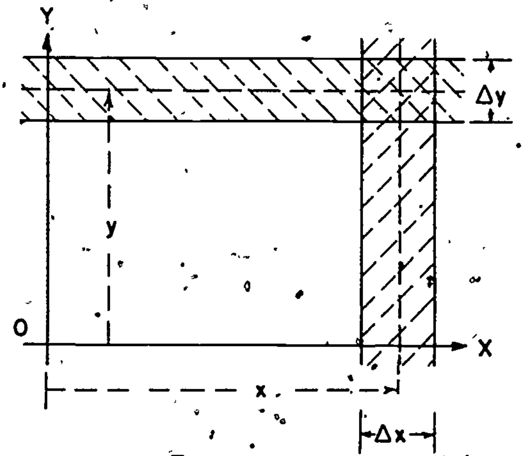

Fig. 1

Yet there remains a question. What, explicitly, do we mean by "probability"? If a marksman in firing his first 1000 rounds at the bull's-eye hits its neighborhood 60 times, does the same thing with his second 1000 shots, with his third, and fourth thousand, then we would say that his probability of a bull's-eye is $\frac{60}{1000}$. But it is a commonplace that performance varies, even for an enthusiast whose marksmanship does not improve with practice. It would be more realistic to suppose his successive scores 60, 57, 62, 59, ... To judge his expectation of a bull's-eye we would consider his performance in the long run. More generally, the probability of a shot hitting the neighborhood of (x,y) will be said to be p, if he has hit this neighborhood pn times with n shots, where n is very large.

Clearer as to what we mean by "probability," we readdress to ourselves the question: "What is the probability of a hit in the rectangular $\Delta x \times \Delta y$ neighborhood of (x,y)?" For brevity, we put this symbolically, $P(x, y, \Delta x, \Delta y)$? But aren't we really asking two questions? Or, to be more precise, are there not two (easier) questions on which the answer to our original question depends?

(1) What is the probability that a hit will lie in the rectangular strip of width Δx centered on x? Symbolically, $P(x, \Delta x)$?

(2) What is the probability that a hit will lie in the rectangular strip of width Δy centered on y? Symbolically, $P(y, \Delta y)$?

Study the conjunction of Figs. 2(1), 2(2) to give Fig. 1.

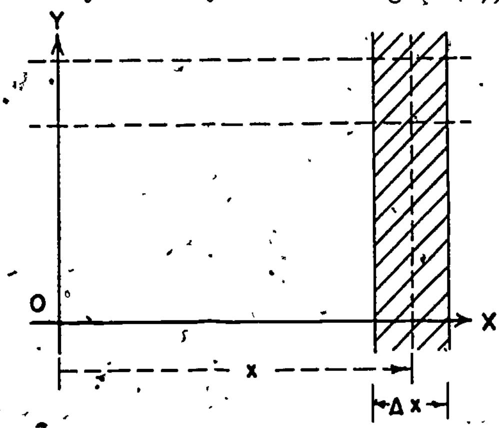

Fig. 2(1)

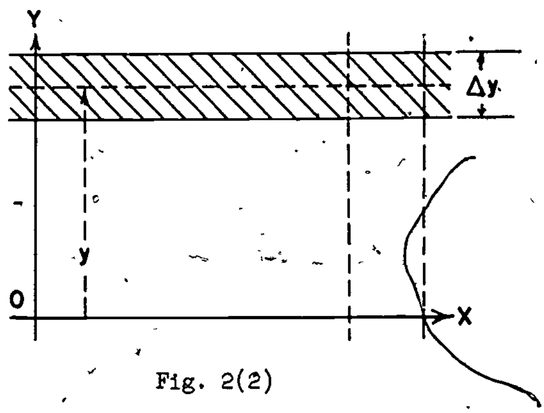

Fig. 2(2)

Does not this make it clear that our original question may be construed as: What is the probability that a hit will lie in both strips?

How, specifically, does $P(x, y, \Delta x, \Delta y)$ depend on $P(x, \Delta x)$ and $P(y, \Delta y)$? The dependence may be illustrated by a problem of throwing dice.

Suppose that the probability of throwing a 3 with a given die is $\frac{1}{6}$ and that of throwing a 4 with a second die is also $\frac{1}{6}$, what is the probability of throwing a 3 with the first and a 4 with the second? In the long run 3 turns up with a frequency of $\frac{1}{6}$, so that if we consider 36n throws, where n is large, 6n of these pairs will have a 3 uppermost on the first die (and the other 30n pairs will not). Of these 6n pairs, since the frequency of a 4 with the second die is $\frac{1}{6}$, independently of the first result, just n of them will have a 4 uppermost. Thus, just n of the 36n pairs will have a 3 uppermost on the first and a 4 uppermost on the second. In short if two independent events have probabilities of $\frac{1}{6}$ and $\frac{1}{6}$, then the conjoint event has a probability of $\frac{1}{6} \times \frac{1}{6}$. More generally, if p_1, p_2 are the probabilities of two independent events, then the probability of the combined event is $p_1 \times p_2$, the product of the individual probabilities. It follows that

$$P(x, y, \Delta x, \Delta y) = P(x, \Delta x) \times P(y, \Delta y) \quad (1)$$

It is an open mathematical secret that with two questions to answer it is best to answer them one at a time. What is $P(x, \Delta x)$? If a barn is five times as wide as its door, then surely the chance of hitting the barn is five times that of hitting the door. Or, if the door is fixed in position (i.e., the position x of its center line is fixed, say $x = x_1$) but its width Δx varies, then the probability of hitting it varies directly as its width. So, we take it that for a vertical strip whose center line is $x = x_1$,

$$P(x_1, \Delta x) = k_1 \cdot \Delta x \quad (2)$$

where k_1 is a constant with respect to Δx.

But, although the "constant of proportionality" is independent of the width Δx of the vertical strip, it is obviously not independent of the position of the strip (i.e., the x value of its center line). Consider, for example, a barn with two doors of the same size. Surely the chances of hitting the one we aim at, straight in front of us, is greater than that of the other. The farther to the side the other is, the smaller its chance of being hit. Thus, reminiscent of (1) and (2) <u>of the last section</u> we will have

$$P(x_1, \Delta x) = k_1 \cdot \Delta x$$

$$P(x_2, \Delta x) = k_2 \cdot \Delta x$$

exemplifying the pattern,

$$P(x_n, \Delta x) = k_n \cdot \Delta x$$

where k_n, although unchanged by changes in Δx, is dependent upon, i.e., is a function of, x_n. In short,

$$P(x, \Delta x) = F(x) \cdot \Delta x .$$

Suppose a barn to have three doors of the same size, the one to the left and the one to the right being equidistant from the one straight ahead of us. Surely the chance (when aiming at the middle one) of hitting the one on the left is the same as that of hitting the one on the right. The chances of a "left" error are the same as those of an equal "right" error. Mathematically,

$$P(x, \Delta x) = P(-x, \Delta x) .$$

Hence, by (3),

$$F(x) = F(-x) ;$$

that is, $F(x)$ is a symmetrical function.

Since $(x)^2 = (-x)^2$, clearly the simplest unspecified symmetrical function is $f(x^2)$. There is, for example, no gain in generality in taking $f(x^4)$, or $f(x^6)$, for these are also of the form $f(X^2)$ with $X = x^2$, $X = x^3$, respectively. Thus (3) becomes of the form

$$P(x, \Delta x) = f(x^2) \cdot \Delta x \qquad (4)$$

which indicates, for example, that the probability of a hit in the left-side strip of Fig. 3 is the same as that of a hit in the right-side strip.

The next question: $P(y, \Delta y)$? Again compare Fig. 2(2) with Fig. 2(1). What differences are there? If $x = y$, $\Delta x = \Delta y$, the strips are of the same size and at the same distance from 0. The only difference is that of direction;

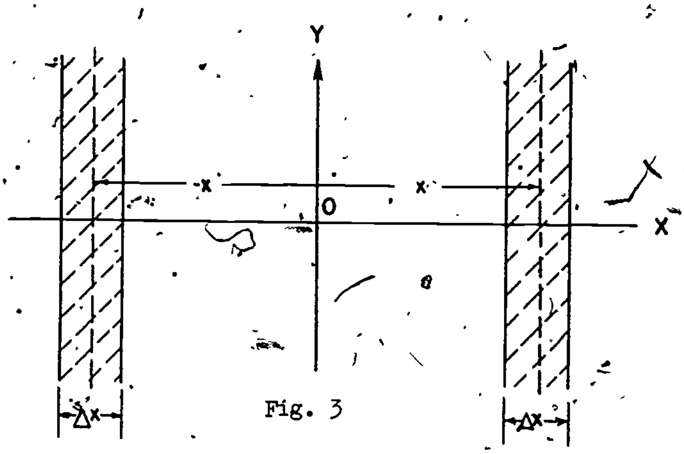

Fig. 3

the one is above, the other to the right of, O. And what role does difference play? We are agreed that a hit, (say) 3 inches left of center has the same probability as a hit 3 inches right of center; is a hit 3 inches above center more likely than 3 inches below center? Right of center was given no preference over left of center; why should above center be given preference over below center? It is natural to consider them equiprobable. This leads to another question: Is a hit 3 inches to right of center more likely than, say, 3 inches above center? Consider the circle of radius 3 inches with center O, illustrated by Fig. 4.

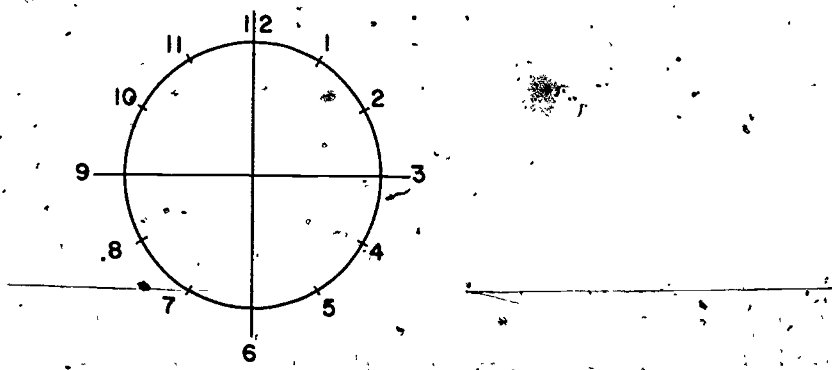

Fig. 4

In firing at O (the immediate neighborhood of), which point on this circle has the greatest probability of being hit? Has the point at "1 o'clock" more or less probability than that at "2 o'clock"? We suppose the probability of hits at any two points on a circle to be equiprobable; no direction is supposed to have preference.

Direction being considered irrelevant, it follows that the strips of Fig. 2(1), 2(2) (with $x = y$, $\Delta x = \Delta y$) are not only of the same size and at the same distance from 0, also they are similarly situated with respect to 0 in the probabilistic sense. Thus* $P(y, \Delta y)$ is determined by precisely the same function as $P(x, \Delta x)$. So, by (4)

$$P(y, \Delta y) = f(y^2) \cdot \Delta y \tag{5}$$

and hence by (1)

$$P(x,y, \Delta x, \Delta y) = f(x^2) \cdot f(y^2) \cdot \Delta x \cdot \Delta y . \tag{6}$$

At this stage Maxwell displays his ingenuity. He introduces a rotation of axes. See Fig. 5.

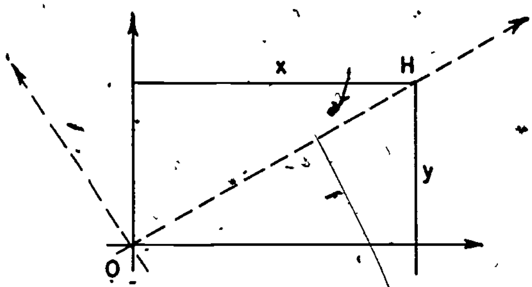

Fig. 5.

His ingenuity is that the ordinate of H relative to the new axes is zero, for H lies on the ξ-axis. $H(x,y)$ relative to the old axes is $H(\xi,0)$ relative to the new. And since the probability of a hit within the neighborhood of a point is independent of the direction of the axes to which it is referred, the probability of a hit within the immediate neighborhood of H is given by

$$P(\xi,0, \Delta\xi, \Delta\eta) = f(\xi^2) \cdot f(0) \cdot \Delta\xi \, \Delta\eta \tag{7}$$

as well as by (6). From (6), (7), we have,

$$f(x^2) \cdot f(y^2) \cdot \Delta x \cdot \Delta y = f(\xi^2) \cdot f(0) \cdot \Delta \xi \cdot \Delta \eta$$

and since the immediate neighborhood of H is described both by $\Delta x \cdot \Delta y$ and by $\Delta\xi \cdot \Delta\eta$, these terms cancel out, and imply that

$$f(x^2) \cdot f(y^2) = f(0) \cdot f(\xi^2) .$$

By Pythagoras' Theorem, $\xi^2 = x^2 + y^2$, so

$$f(x^2) \cdot f(y^2) = f(0) \cdot f(x^2 + y^2). \qquad (8)$$

(8) is a functional equation of the form

$$f(a) \cdot f(b) = K \cdot f(a + b) \qquad (9)$$

where $f(0) = K$. Here we may indulge in wishful thinking, for we recall that the functional equation for the law of growth is of the form

$$f(a) \cdot f(b) = f(a + b) . \qquad (10)$$

If K were equal to 1, then the law of errors would have the same form of functional equation as the law of growth, and consequently the solution of (8) would likewise be an exponential function.

It turns out that (8) can be reduced to the form (10). We put

$$\frac{f(x^2)}{f(0)} = g(x^2) ,$$

so that

$$f(x^2) = f(0) \cdot g(x^2)$$

$$f(y^2) = f(0) \cdot g(y^2)$$

$$f(x^2 + y^2) = f(0) \cdot g(x^2 + y^2) .$$

Substituting in (8)

$$[f(0) \cdot g(x^2)] \cdot [f(0) \cdot g(y^2)] = f(0) \cdot [f(0) \cdot g(x^2 + y^2)].$$

Dividing by $[f(0)]^2$, we obtain,

$$g(x^2) \cdot g(y^2) = g(x^2 + y^2)$$

the functional equation of the law of growth. Consequently

$$g(x^2) = a^{x^2}$$

i.e.,
$$\frac{f(x^2)}{f(0)} = a^{x^2}$$

so that
$$f(x^2) = f(0)\, a^{x^2}$$

and, by (4)
$$P(x, \Delta x) = f(0) \cdot a^{x^2} \cdot \Delta x .$$

And finally, for brevity, putting $f(0) = A$,
$$P(x, \Delta x) = A \cdot a^{x^2} \cdot \Delta x .$$

This completes Maxwell's derivation of Gauss' famous law of errors.

We discuss this law briefly. Since the chance of a large deflection is obviously smaller than the chance of a small deflection, $a < 1$. Plotting a^{x^2} as a function of x, we obtain a bell-shaped curve typical of symmetrically deviated errors. See Fig. 6.

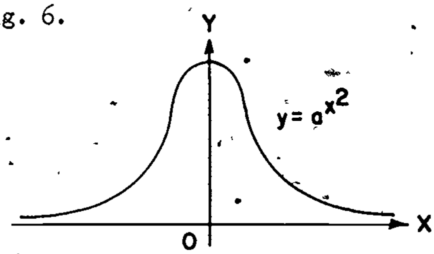

Fig. 6

This is the starting point for the development of the whole theory of the error of combinations of observations.

2.3 Differential Versus Functional Equations.

Generally speaking, scientific laws are deduced from differential, rather than functional, equations. Why? Differential equations are easy to set up; they are the mathematical answer to: What is the instantaneous change of a given state?. Functional equations are hard to come by; often, genius is required to find them. I would prefer to use differential equations; your students cannot. In consequence, my hands are tied; so, let us see what else we can do with the functional variety.

2.4. The Problem of Predicting Population Growth.

What is the law of increase of human population? The simplest, plausible assumption is that the number of people of the $(n + 1)^{th}$ generation, x_{n+1}, will be directly proportional to the n^{th}, x_n. Symbolically,

$$x_{n+1} = q \cdot x_n . \qquad (1)$$

On this basis, if x_1 is the population of the first generation considered, the population of successive generations will be

$$x_1, \; qx_1, \; q^2x_1, \; q^3x_1, \; \ldots$$

so that

$$x_n = q^{n-1} \cdot x_1 . \qquad (2)$$

If $q > 1$, the population is increasing. Again we have an exponential law.

This formula was stated in words by Malthus (1766-1834): populations of countries increase in geometric ratio. It is interesting to note that Malthus was led to his formulation by inspection of the census records of the American people, which showed a doubling of population every 50 years. His statement, simple as it is, crude as it is, had a tremendous influence on the whole of social philosophy in the 19th century.

The social philosophers of the French Revolution argued that it was mankind's duty to ease the hardship of the poor, and to abolish pestilence, plague, famine, and war, so that everyone could live happily till death of old age. Malthus thought this view greatly mistaken. What would happen with neither pestilence nor plague, with neither famine nor war? The population, increasing in geometrical ratio, would in a few years, he argued, become so vast that the earth could not feed it. The Manchester industrialists used this argument to prop up their policy of free enterprise, to increase trade while leaving the world at large to sort itself out. There could be no obligation to better the lot of the poor nor attempt to prevent famine or war; for these things, if evils, were evils necessary to prevent overpopulation. Malthus' law became the arithmetic of human misery.

Darwin also thought about the consequences of a population increasing geometrically. For him the problem had a wider context. He was as much, if not more, interested in the increase of a colony of sea birds as in the Manchester birth rate. What, he asked, controls population? The <u>dinosaurus</u> has long been extinct; the whale has survived. Ultimately he gave an answer: his theory of natural selection. There followed his theory of evolution of species. What is man's obligation to man? Is one to succor or to starve one's neighbor? The fall of the Bastille and the dark Satanic mills gave contradictory answers. By the middle of the last century even some industrialists began to question whether the evil of overworked and underpaid factory hands, living underfed in overcrowded slums, was a necessary evil. Couldn't there be a better arithmetic?

The Belgian sociologist, Verhulst, made an important observation. Catastrophies, wars, and plagues have occurred from time to time, not all the time. Between any two successive catastrophies there was a period of tranquility, say, typically, that of two or three generations. This period, had the law of increase been geometrical, would have given the population ample time to regain and surpass, before the next catastrophe, its size before the last. We illustrate with Fig. 7.

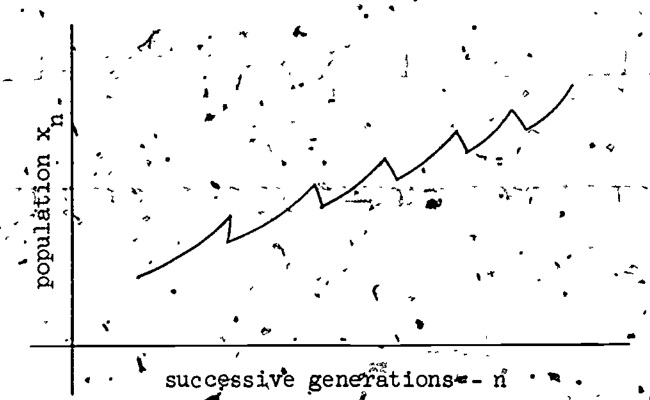

Fig. 7

But mankind has inhabited the earth for thousands of years, so that although we do not know the value of n, where x_n is the present (n^{th} generation) population, we do know that n is large. With n large, and the population before imminent catastrophe greater than that before the previous one, surely the world would now be overcrowded. Verhulst concluded that the geometric law does not give a correct account of the facts.

Discontent with the old arithmetic was the first step towards the new. Verhulst replaces

$$x_{n+1} = q \cdot x_n. \qquad (1)$$

by

$$x_{n+1} = q \cdot x_n - r \cdot x_n^2. \qquad (3)$$

What is the effect of $-r \cdot x_n^2$? The larger x_n becomes, the larger x_n^2 becomes relative to x_n, so that the larger the population the greater the braking effect of $-r \cdot x_n^2$ on its rate of growth. A vast population can only increase very slowly. To say that $-r \cdot x_n^2$ is a "slowing up" factor is to describe it in terms of its consequences; Verhulst did better. His factor is the outcome of a more painstaking analysis of population growth; he described it in terms of what causes the slowing up: competition.

Man's activities are of two kinds, cooperative and competitive. A marriage is the outcome of successful competition by a man against other men for a woman; a child is the outcome of successful cooperation by a man with a woman. Farmers and biochemists cooperate to produce greater yields of wheat; bankers and bank robbers compete for the customers' deposits. Soldiers cooperate as armies to compete against other soldiers cooperating as armies. Verhulst took the view that in the main cooperation tends to increase, and competition tends to decrease, the population.

How is the intensity of the struggle for family existence to be measured? Competition occurs when each of two or more people wants exclusively the same thing. When, for example, two married men want homes, and only one house is available. What is the probability that two men of a population x_n both want the same house? If p is the probability of either wanting it, p^2 is the probability of both wanting it. But the larger x_n, the greater the chances of a man wanting it. That to double x_n would be to double p is a plausible supposition. Yet, if p is directly proportional to x_n, then p^2 is directly proportional to x_n^2. Thus it is not unreasonable to take $r \cdot x_n^2$ as a measure of the competition.

If discerning, you may observe that we have neglected to add a competition factor $-r_1 \cdot x_n^3$, that for three competitors, and a factor $-r_2 \cdot x_n^4$, that for four competitors, and so on. True. But it is useless to set up equations which completely fit a situation if this leaves us with mathematics too difficult to handle. In applying mathematics to reality there is always a compromise: by introducing an element of idealization, or by ignoring less important factors, what is too complex is reduced to what is manageable. Often, the proper question is not, "Is a given formula dead accurate?" but rather, "Is it a sufficiently good approximation for the present investigation?" Is (3) adequate for population investigation? I am anxious to answer this question, for in so doing I shall have opportunity to exemplify that quite intricate problems can be dealt with by mere high school mathematics.

If $r = 0$, the competition factor $-r \cdot x_n^2 = 0$, and we find ourselves considering a society with the tranquility of lotus-eaters. With no competition (3) reduces to (1), so that (3) is a better formula in the sense of including (1) as a limiting case. Turn from the tranquility of everyone lotus-eating to the desperation of some not eating at all. We all know what happens if competition is so severe that there are more hands than jobs, and more mouths to feed than food to feed them: life is nasty, brutal, and for many, short. That x_{n+1} could be smaller than x_n is obvious. But, what answer does (3) give? We write it in the form

$$x_{n+1} = x_n^2 \left(\frac{q}{x_n} - r\right) .$$

This form makes it clear that when, for any given q, x_n is specified, we can select r such that $\left(\frac{q}{x_n} - r\right)$ is arbitrarily small. Consequently x_{n+1} can be made as small as we please and, a fraction, less than x_n. But there is no immediate answer to the question: Does (3) also give x_{n+2} correctly?

Unfortunately,

$$x_{n+2} = x_{n+1}^2 \left(\frac{q}{x_{n+1}} - r\right)$$

where the previous factor $\left(\frac{q}{x_n} - r\right)$ is now replaced by $\left(\frac{q}{x_{n+1}} - r\right)$. The difficulty is that the denominator of q has been changed. Since $x_{n+1} < x_n$, it follows that the factor $\left(\frac{q}{x_{n+1}} - r\right) > \left(\frac{q}{x_n} - r\right)$, but this is, of itself, insufficient to determine if $x_{n+2} < x_{n+1}$.

However, that (3) can be used to describe correctly the population at least one generation ahead in extreme states of society gives us some expectation that it will serve far ahead in intermediate states, neither completely tranquil nor thoroughly brutal. At least it does merit more systematic examination.

First we further reconcile the complex with the manageable. For intermediate states of society the change in population from one generation to the next will be so slow that $x_n \cdot x_{n+1}$ will be a good approximation to x_n^2. Consequently, we may consider

$$x_{n+1} = q \cdot x_n - r \cdot x_n x_{n+1} \qquad (4)$$

instead of (3), without introducing any really significant change in the population law. (4) is preferable as this makes for much easier mathematics.

(4) is a mixed equation; x_{n+1} occurs on both sides of the equation. Making x_{n+1} the subject of the formula, we have

$$x_{n+1} = \frac{q}{1 + rx_n} \cdot x_n \qquad (5)$$

We observe that whereas with Malthus' law x_n has a factor q, with (what is essentially) Verhulst's law the factor is $\frac{q}{1 + rx_n}$; the growth factor is no longer a constant, but depends on x_n. The longer x_n becomes, the smaller the growth factor. The population is self-regulating; overpopulation is prevented. Verhulst's law echoes his original observation.

Our problem is to find x_{n+1} in terms of x_1. With Malthus' law this as easy. Indeed, the textbooks are crammed so full with geometrical progressions that the student is apt to suppose there are no other varieties. Real problems, alas, seldom have the neat and obvious form of school exercises; to the contrary, they often come in ugly and hidden forms. How to transform the latter into the former is an essential part of the art of doing mathematics. Although the student cannot reasonably be expected to have the foresight to see that (5) is in essence geometrical, he can reasonably be required to have the hindsight.

Taking reciprocals in (5),

$$\frac{1}{x_{n+1}} = \frac{1 + rx_n}{qx_n} = \frac{1}{q} \cdot \frac{1}{x_n} + \frac{r}{q} .$$

That the reciprocals of x_{n+1}, x_n satisfy a simpler law, invites the substitutions

$$\xi_{n+1} = \frac{1}{x_{n+1}} , \qquad \xi_n = \frac{1}{x_n} ,$$

which give

$$\xi_{n+1} = \frac{1}{q} \cdot \xi_n + \frac{r}{q} .$$

But for the constant term we would have the form of Malthus' law. The thought is father of the wish. Substituting

$$\xi_{n+1} = \eta_{n+1} + \alpha , \quad \xi_n = \eta_n + \alpha ,$$

where α is an arbitrary constant, we have,

$$\eta_{n+1} + \alpha = \frac{1}{q} \cdot (\eta_n + \alpha) + \frac{r}{q} ,$$

so that,

$$\eta_{n+1} = \frac{1}{q} \cdot \eta_n + (\frac{\alpha}{q} + \frac{r}{q} - \alpha) .$$

Since α is an arbitrary constant, we are at liberty to give it whatever specification we please. But we wish the constant term of the equation to be zero; accordingly, it pleases us to define α by

$$\frac{\alpha}{q} + \frac{r}{q} - \alpha = 0 .$$

i.e., such that

$$\alpha + r = q\alpha$$

which gives

$$\alpha = \frac{r}{q - 1} , \quad \text{provided } q \neq 1 .$$

The condition that $q > 1$ merely implies that when $r = 0$, i.e., when there is no competition, the population is increasing. This supposition is acceptable and meets the proviso that $q \neq 1$. Consequently, we take α to be $\frac{r}{q-1}$ and infer that

$$\eta_{n+1} = \frac{1}{q} \cdot \eta_n . \qquad (6)$$

We have transformed the <u>form</u> of Verhulst's law to that of Malthus': the laws themselves are, of course, distinct.

Since (6) is of the same form as (1), we have, as an analogue of (2).

$$\eta_{n+1} = \left(\frac{1}{q}\right)^n \cdot \eta_1 . \qquad (7)$$

It remains merely to reverse our transformations to obtain x_{n+1} as a function of x_n. There is a gain of notational compactness by delaying the substitution for α until the end.

First, we go back from the η's to the ξ's. Since

$$\xi_{n+1} = \eta_{n+1} + \alpha$$

by (7)

$$\xi_{n+1} = \frac{1}{q^n} \cdot \eta_1 + \alpha ,$$

but,

$$\xi_1 = \eta_1 + \alpha ,$$

so that

$$\xi_{n+1} = \frac{1}{q^n}(\xi_1 - \alpha) + \alpha = \frac{1}{q^n}\xi_1 + (1 - \frac{1}{q^n})\alpha = \frac{\xi_1 + (q^n-1)\alpha}{q^n} .$$

Second, we go back from the ξ's to the x's. Since,

$$\xi_{n+1} = \frac{1}{x_{n+1}} , \quad \xi_1 = \frac{1}{x_1} ,$$

we have,

$$\frac{1}{x_{n+1}} = \frac{\frac{1}{x_1} + (q^n-1)\alpha}{q^n} = \frac{1 + (q^n-1)\alpha\, x_1}{q^n \cdot x_1}$$

and taking reciprocals,

$$x_{n+1} = \frac{q^n}{1 + (q^n - 1)\alpha\, x_1} \cdot x_1 .$$

Finally, since

$$\alpha = \frac{r}{q - 1} ,$$

we have,

$$x_{n+1} = \frac{q^n}{1 + \frac{(q^n-1)}{q - 1} r \cdot x_1} \cdot x_1 . \qquad (8)$$

Common prudence demands some check on our work. Substituting $n = 1$ in (8), we have

$$x_2 = \frac{q}{1 + rx_1} x_1 .$$

The same substitution in (5) gives the same result. It checks.

By a judicious use of the fact that (2) is a consequence of (1), we have deduced a formula for x_{n+1} when subject to Verhulst's law, in terms of x_1, q, and r. What is its significance? We suppose q greater than, but close to, 1, and r very small indeed.

First we investigate the consequences of n being small also. Since

$$\frac{q^n-1}{q-1} = 1 + q + q^2 + \cdots + q^{n-1} \approx n \qquad (q \approx 1),$$

$\frac{q^n - 1}{q - 1} \cdot rx_1$ will be small compared with unity when n is small. Thus, without charge of gross neglect we can ignore the second term of the denominator of (8) when considering the first few generations. We have

$$x_{n+1} \approx q^n \cdot x_1 ,$$

i.e., that Verhulst's law approximates to Malthus'.

Next we investigate the consequences of large n. When n is large and $q > 1$, it follows from $\frac{q^n - 1}{q - 1} \approx n$ that $\frac{q^n - 1}{q - 1} rx_1$ is large compared with unity. So without gross neglect we may ignore the first term of the denominator of (8), giving

$$x_{n+1} \approx \frac{q_n \cdot x_1}{(q^n-1) \cdot (\frac{r}{q-1})x_1} = \frac{q^n}{q^n-1} \cdot \frac{q-1}{r} .$$

But, with $q > 1$, the larger n is, the larger q^n and the nearer $\frac{q^n}{q^n-1}$ to 1. Consequently, the nearer x_{n+1} to $\frac{q-1}{r}$. Yet in neglecting the first term of the denominator, we overestimate x_{n+1}, so that no matter for how many generations the population continues it will not exceed $\frac{q-1}{r}$. Observe that this upper limit to the size of x_n is <u>independent of the size of the original population</u> x_1. Isn't this astonishing?

The graph of (8), known as the <u>logistic</u> or <u>flying S curve</u> is illustrated by Fig. 8.

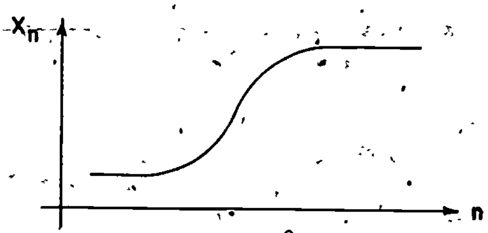

Fig. 8

How, in actual practice, do we apply (8) to predict the population of, say, the United States, a decade or a century hence? (8) gives the number of people who belong to the $(n+1)^{th}$ generation, x_{n+1}, in terms of the number of people who belong to the first generation considered, x_1. But how long an interval is there between one generation and the next? Every minute several people die and several are born. Who belongs to the present generation? To be precise "the present generation" refers to an overlapping of many generations, those generated in the years 1900, 1901, 1902, 1920, 1921, among others (if still surviving). As statistics are usually taken on a yearly basis, it is convenient to consider the population in successive years as successive generations, to take x_1 as the population for the first year considered and x_{n+1} that n years later.

Let us take 1959 as the first year and obtain the actual figures for x_1, x_2, and x_3, the population of the U.S. in 1959, 1960, and 1961, from the available table of population statistics. Substituting the figures for x_1 and x_2 in (8), we obtain a first equation relating q and r; substituting the figures for x_2 and x_3, we obtain a second. We now have two equations in the two unknowns, q and r, sufficient to determine them. (8) has been tailored to fit the facts; the growth coefficient q and the competitive coefficient r are chosen so as to describe correctly the recent population history of the U.S. If in (8) we write the figures for q, r, and x_1, our formula is ready for use. Substituting $n = 3$, we predict the population for 1962; substituting $n = 4$, we predict that for 1963.

Would it be rash to take the result of substituting $n = 100$ as more than a very tentative prediction of the population for the year 2059? Typically, growth and competition remain steady, so that a formula that has accurately described the last two years may reasonably be expected to describe the next two. But over the span of a century the growth and competitive factors have more time in which to alter, so that the long term prediction should be more cautiously regarded.

We have seen that three successive years' statistics are sufficient to determine q and r. Had we used all the statistics of the last decade, the eight periods 1952-54, 1953-55, 1954-56, ..., 1959-61 would have given us eight determinations of q and eight of r. Had these differed in the last decimal place or two we would have struck the typical figure which would give the best overall description of the decade. What fits the facts for the last ten years is surely more likely to fit the next hundred than that which fitted merely the last two.

Is Verhulst's formula reliable? Around about 1850 he made a careful population study of several European countries and of the United States. He used his law to predict their populations as far ahead as a century. Some of his predictions are famous, and justly so. For example, he calculated that France would reach a maximum population of 40 million in 1921; the event proved him correct. Despite the Civil War, his prediction for the U.S. population in 1940 was less than a million out. But ironically, his law applied to his own country, Belgium, did not work. Belgium's population curve for the century is given by Fig. 9.

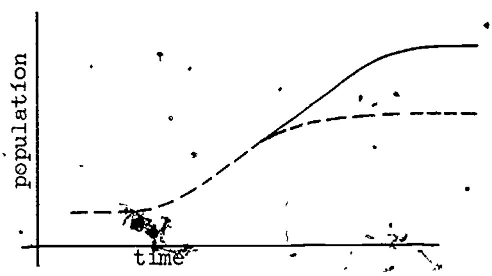

Fig. 9

How did Verhulst's prediction go wrong? Belgium switched from agriculture to industry and colonized the Congo. This distinct sociological change permanently altered the growth and competition coefficients. His application of his law continued to describe the growth of Belgium as agricultural when it was in fact industrial. Observe that the Belgian population curve is a combination of the parts of two S-curves, the earlier with agricultural, and the later with industrial conditions obtaining. How then, it may well be asked, was his prediction for the United States successful despite the Civil War? Of course, Verhulst could not know that the Civil War was going to break out a decade or so after he made his population analysis and so he could not take the changed values of the growth and competition coefficients into account. The point is that these changes of coefficients, unlike those due to a switch from agriculture to industry in Belgium, were merely <u>temporary</u>: soon after the Civil War his coefficients were again accurately descriptive. With people killed in the war his 1865 population estimate was high, but his estimate and the actual population of that time both have growths asymptotic to $\frac{q-1}{r}$. In the long run his prediction would have been correct; the run to 1940 was long enough for it to be correct within 1 million.

As promised, I have shown you that quite intricate results can be obtained without using differential equations. Actually a formula for x_{n+1} for any specified n can be obtained from (5) by using only the very simplest of algebra. Putting $n = 1, 2,$ successively, we have

$$x_2 = \frac{q}{1 + rx_1} x_1, \quad x_3 = \frac{q}{1 + rx_2} x_2$$

so that

$$x_3 = \frac{q}{1 + r\left(\frac{qx_1}{1+rx}\right)} \cdot \frac{q}{1 + rx_1} x_1$$

Multiplying numerator and denominator of right side by $1 + rx_1$,

$$x_3 = \frac{q^2}{(1+rx_1) + r(qx_1)} \cdot x_1 = \frac{q^2}{1 + (q+1)rx_1} \cdot x_1 .$$

Proceeding in this way $x_4, x_5, \ldots$ can be obtained from x_1. We go step by step along an adventurous path to find where it leads us. After patient travel the way the road runs being clearly discerned, the more ambitious student may prove the formula for x_{n+1} by mathematical induction.

2.5 Cusanus's Recursive Formula for π.

When, as in the last section, x_n, a member of a sequence, is defined in terms of earlier members of the sequence, it is said to be defined recursively. This terminology acknowledges descriptively that the sequence refers back to itself; it is, so to speak, a snake biting its own tail.

We now consider one of the most elegant recursive formulae in mathematics, namely that given by Cusanus (1401-64) in about 1450. Even though it was the first to facilitate the calculation of π better than Archimedes' formula, it is not widely known. Already more than five hundred years old, perhaps it is too modern for the "modernists." With this formula we have a hint that there was, contrary to popular historical misconception, tremendous intellectual activity before the Renaissance. Despite what the history books fail to say, without Cusanus and his ilk Galileo and Newton could not have inherited the groundwork they did in fact inherit.

Cusanus' calculation of π. It really is obvious that if a regular polygon of perimeter p is circumscribed by a circle of radius R then the more numerous the sides of the polygon the closer the approximation of p to $2\pi R$ and $\frac{p}{2R}$ to π. Surely thousands of persons before and since Archimedes must have thought of this, yet how many have found a method of effectively exploiting it to calculate π? Archimedes considered an unending sequence of regular polygons, each polygon with more sides than its predecessor, each circumscribed by the same circle; Cusanus considered an unending sequence of regular polygons, each polygon with more sides than its predecessor, but all of the same perimeter and therefore circumscribed by different circles. Whereas Archimedes found the limit of p with constant R, Cusanus found the limit of R with constant p. Both methods are elegant. Encourage the student who finds Cusanus' elegance exciting to study Archimedes' for himself.

How, specifically, did Cusanus exploit his idea? He did so in the following way. From a given circle C_1 of radius r_1 circumscribing a regular polygon of m sides and perimeter k, another circle C_2 of radius r_2 circumscribing a regular polygon of double the number of sides, but with the same perimeter, is constructed. By repetition of the procedure n times there obtains a sequence of circles $C_1, C_2, C_3, \ldots, C_{n+1}$, of radii $r_1, r_2, r_3, \ldots, r_{n+1}$, circumscribing regular polygons with constant perimeter k, of $m, 2m, 2^2m, \ldots, 2^nm$ sides, respectively. It is intuitively clear that

$$\pi = \frac{k}{2R} \quad \text{where} \quad R = \lim_{n \to \infty} r_{n+1}.$$

(Since we are now considering a sequence of circles of constant perimeter we use the letter k in preference to p.) The real problem, of course, is to determine r_{n+1}. The way in which C_2 is constructed from C_1 determines the relation between r_2 and r_1. But C_3 is constructed from C_2 as C_2 from C_1 so that r_3 has the same relation to r_2 as r_2 to r_1, and for similar reasons r_4 has the same relation to r_3 as r_3 has to r_2. Thus r_4 can be determined in terms of r_3, while r_3 can be determined in terms of r_2, and r_2 in terms of r_1, so that finally r_4 can be determined in terms of r_1. More generally, r_{n+1} is determined in terms of r_n, which in turn is determined in terms of its sequential predecessor, which in turn ..., so that finally r_{n+1} is determined in terms of r_1. The formula is recursive.

Now for the details. What, specifically, is the relation between r_2 and r_1? Fig. 10 illustrates the essentials of what we are given: the m-sided circumscribed polygon being regular, it is sufficient to consider just one of its sides.

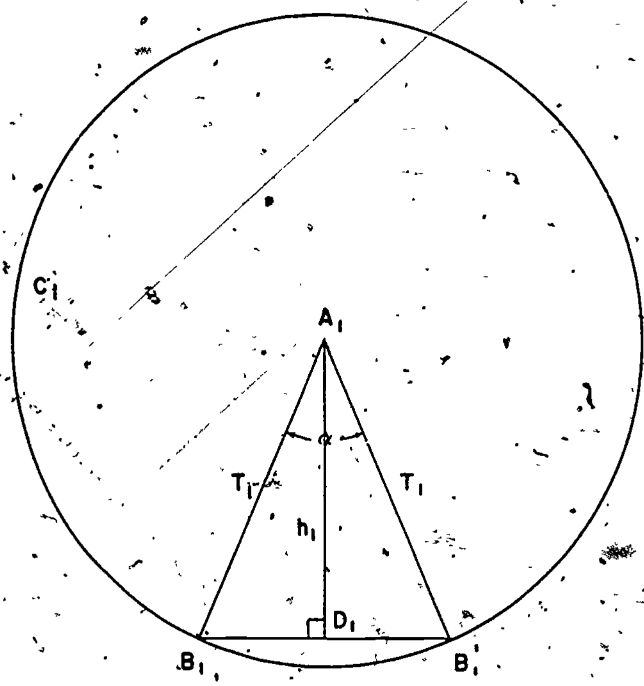

Fig. 10

We make the construction illustrated by Fig. 11.

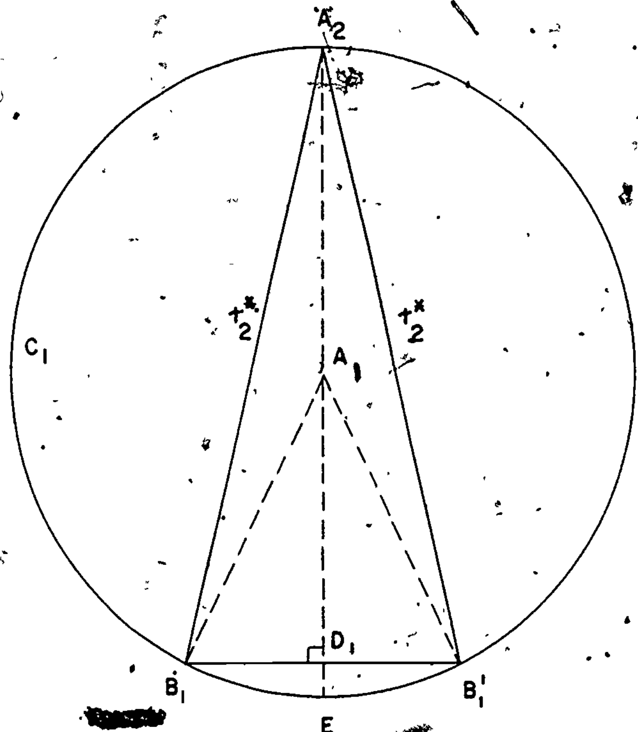

Fig. 11

Since, as Euclid tells us, angle subtended at circumference is one-half angle subtended at center,

$$\angle B_1A_2B_1' = \frac{1}{2}\angle B_1A_1B_1'$$

Consequently $2m$ such triangles as $B_1A_2B_1'$ fit together to form a regular polygon with perimeter $2m \times \overline{B_1B_1'}$, which is circumscribible by a circle C^* with center A_2, and (say) radius r_2^*. Fig. 12 illustrates the essentials. Retaining A_2 as center we now shrink Fig. 12 to half size. We now have a circle C_2 circumscribing a regular polygon with the same perimeter as, but twice the number of sides of, that circumscribed by C_1. See Fig. 13. Compare Fig. 13 with Fig. 10.

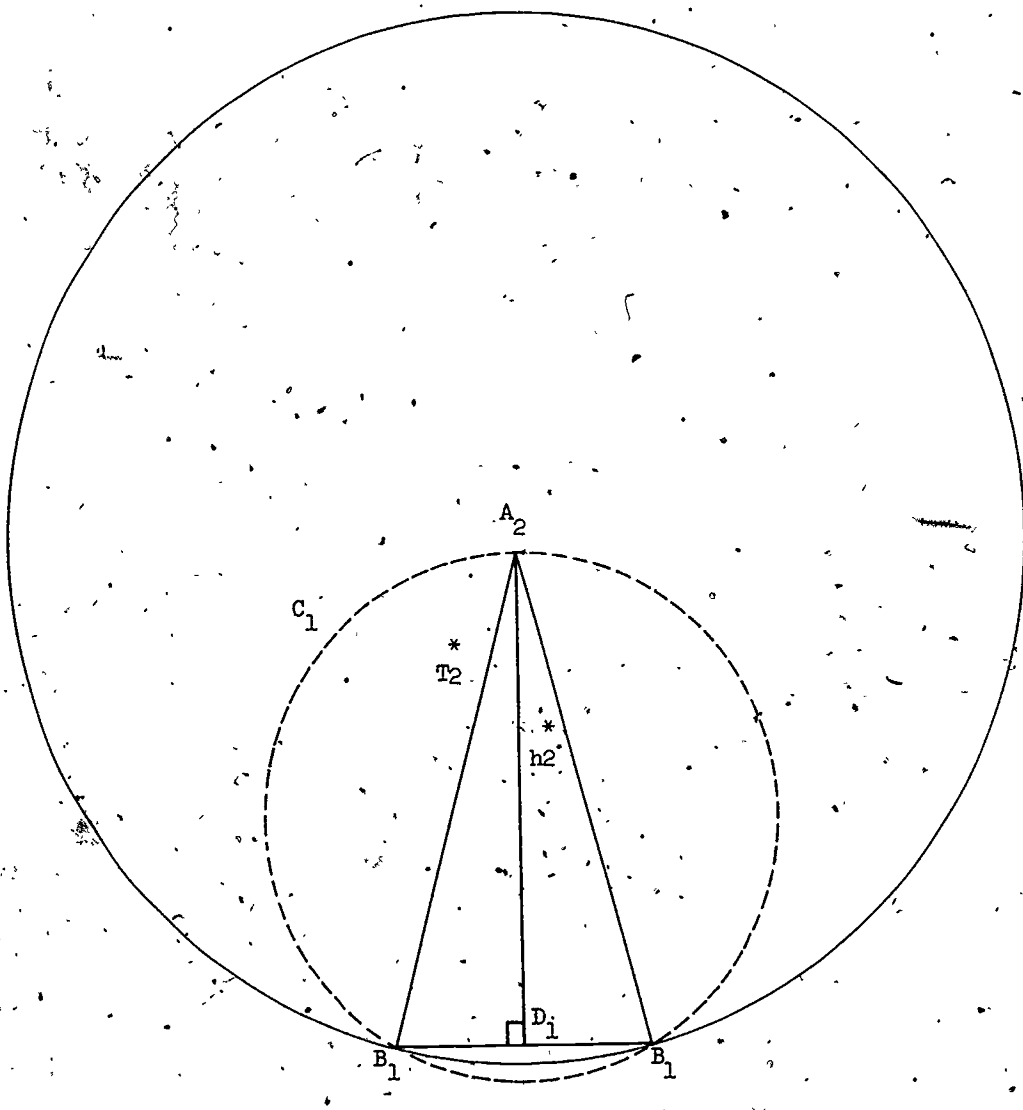

Fig. 12

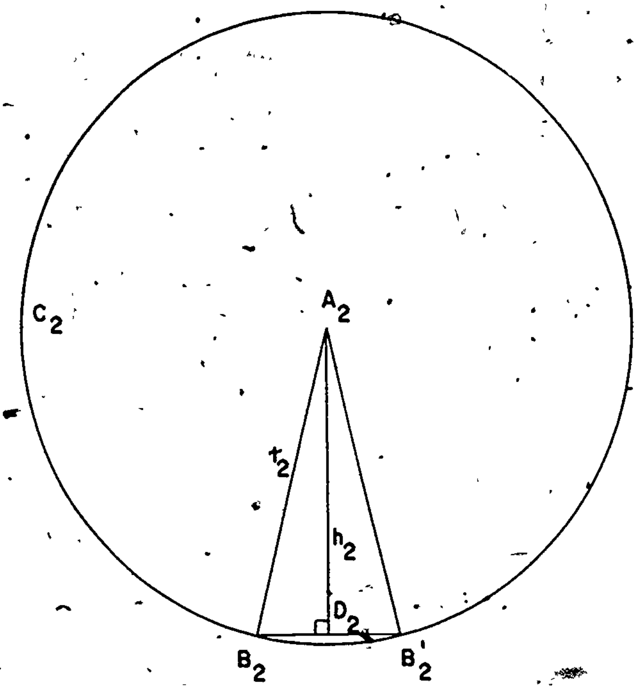

Fig. 13

The problem is to find r_2 in terms of r_1. To do this we first consider the geometry of Fig. 11. Since, as Thales tells us, the angle included in a semicircle is a right angle, A_2B_1E is a right angle (it is subtended at the circumference of C_1' by diameter A_2E). Thus triangles A_2B_1E, $A_2B_1D_1$ are both right triangles (the latter is right angled at D_1) and additionally have a common angle B_1A_2E. Therefore these triangles are similar, and consequently their corresponding sides are proportional, so that

$$\frac{A_2B_1}{A_2D_1} = \frac{A_2E}{A_2B_1}$$

i.e.,

$$\frac{r_2^*}{h_2^*} = \frac{2r_1}{r_2^*}.$$

Thus

$$(r_2^*)^2 = 2r_1 \cdot h_2^* \qquad (1)$$

But h_2^* is an uninvited bedfellow and is speedily to be replaced.

$$h_2^* = A_2D_1 = A_2A_1 + A_1D_1$$

i.e.,

$$h_2^* = r_1 + h_1 . \qquad (2)$$

We have related the measurements of Fig. 10 to those of Fig. 12; we wish to relate them to those of Fig. 13. But Fig. 13 was obtained from Fig. 12 by reducing everything to half size, so that

$$r_2 = \frac{1}{2} r_2^*$$

$$h_2 = \frac{1}{2} h_2^* .$$

Hence, from (2) we have

$$h_2 = \frac{r_1 + h_1}{2} \qquad (3)$$

and from (1),

$$r_2^2 = \frac{1}{4} (r_2^*)^2 = \frac{1}{4} \cdot 2r_1 \cdot 2h_2$$

so that

$$r_2 = \sqrt{r_1 \cdot h_2} . \qquad (4)$$

This derivation discloses our motive for using a star notation: to emphasize the transitory role of r_2^* and h_2^*.

(3) and (4) give r_2 in terms of r_1 (and h_1). The intrusion of the h's is an incidental complexity that must not be permitted to obscure the leading idea; in specifying the relation between r_2 and r_1 (and h_1) we have reached the heart of the matter. In repeating our procedure to obtain C_3 from C_2 as C_2 was obtained from C_1, r_3 will have the same relation to r_2 as r_2 has to r_1, and in obtaining C_4 from C_3, r_4 will have the same relation to r_3 as r_3 has to r_2 and r_2 has to r_1. Consequently, for C_{n+1} we have

$$h_{n+1} = \frac{r_n + h_n}{2} \qquad (5)$$

$$r_{n+1} = \sqrt{r_n \cdot h_{n+1}} \qquad (6)$$

Let us recapitulate. To avoid the verbosity of saying that h_n is the altitude of any triangle whose vertex is the center of C_n and whose base is one of the sides of the regular polygon circumscribed by C_n, let us refer to h_n as the altitude of C_n. Then, if C_1 is a circle of radius r_1 and altitude h_1 circumscribing a regular polygon of m sides, and perimeter k, by repeating n times the process considered above we form a sequence of circles $C_1, C_2, C_3, \ldots, C_{n+1}$ of radii $r_1, r_2, r_3, \ldots, r_{n+1}$, (and altitudes $h_1, h_2, h_3, \ldots, h_{n+1}$) circumscribing regular polygons of m, $2m$, $2^2m, \ldots, 2^nm$ sides, respectively, where h_{n+1}, r_{n+1} satisfy (5), (6) for $n = 0, 1, 2, \ldots, n$.

To calculate π, i.e., $\frac{k}{2R}$, it merely remains to determine R, where $R = \lim_{n \to \infty} r_{n+1}$. It is convenient to take C_1 as circumscribing a regular hexagon, i.e., to take $m = 6$, and to take $r_1 = 1$. See. Fig. 14.

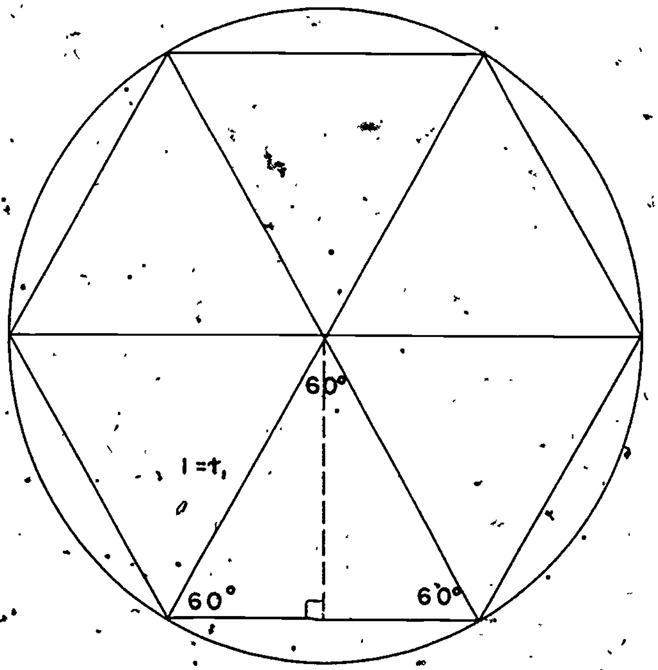

Fig. 14

Here h_1 is evidently the altitude of an equilateral triangle of unit side. By simple calculation we find $h_1 = \sqrt{\frac{3}{2}}$. The reader is now in a position to calculate a sequence of successively better approximations to R, and hence, to π.

This raises the question of the accuracy of approximations. It really is intuitively obvious that as n increases the angle at the vertex of each triangle constituting the regular polygon circumscribed by C_n will get smaller and smaller, and therefore r_n and h_n more and more nearly equal, giving

$$\lim_{n \to \infty} r_n = \lim_{n \to \infty} h_n ,$$

i.e., that r_n and h_n both converge to R. And since the hypotenuse is the greatest side of a right angle,

$$r_n > h_n .$$

See Fig. 15.

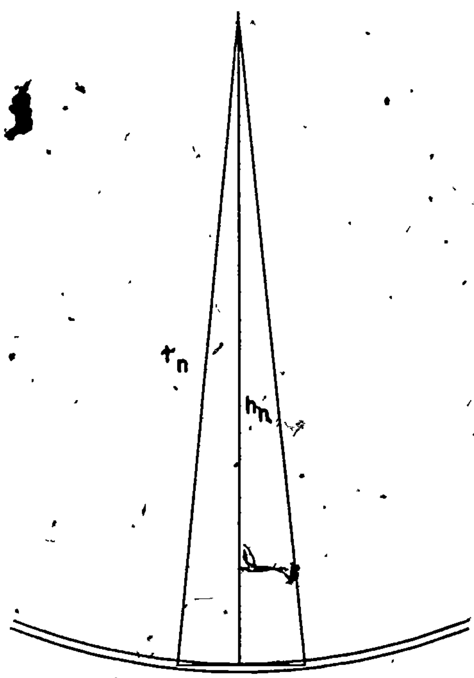

Fig. 15

Considering the polygon, h_n is the radius of the circle it inscribes and r_n is the radius of the circle circumscribing it: ultimately, in the limit surely these circles coincide. Thus it seems reasonable to suppose that r_n decreases,

h_n increases and

$$r_n > R > h_n$$

The astonishing thing is that we are able to anticipate that, for example, for the hexagon, with $r_1 = 1$, and so perimeter $k = 6$, the repeated applications of (5) and (6) will converge to $\frac{6}{2\pi}$, i.e., to $\frac{3}{\pi}$.

We may add that neither Cusanus nor Descartes (1596-1650) (who made extensive use of Cusanus' formulae) worried overmuch about convergence: they were confident of their intuition.

With no more than an elementary knowledge of inequalities we can prove convergence. The crux of the matter is that the difference between r_n and h_n gets smaller and smaller. But having the abhorrence for square roots that Pythagoras had for bean eating, we prefer to consider the difference of r_n^2 and h_n^2.

By (5), (6),

$$r_{n+1}^2 - h_{n+1}^2 = r_n h_{n+1} - \left(\frac{r_n + h_n}{2}\right)^2$$

but by (5)

$$r_n(h_{n+1}) = r_n\left(\frac{r_n + h_n}{2}\right)$$

so

$$r_{n+1}^2 - h_{n+1}^2 = r_n\left(\frac{r_n + h_n}{2}\right) - \left(\frac{r_n + h_n}{2}\right)^2$$

$$= \left(\frac{r_n + h_n}{2}\right)\left(r_n - \frac{r_n + h_n}{2}\right)$$

$$= \left(\frac{r_n + h_n}{2}\right)\left(\frac{r_n - h_n}{2}\right)$$

$$= \frac{1}{4}(r_n^2 - h_n^2).$$

Thus

$$r_3^2 - h_3^2 = \frac{1}{4}(r_2^2 - h_2^2)$$

and

$$r_2^2 - h_2^2 = \frac{1}{4}(r_1^2 - h_1^2).$$

so that

$$r_3^2 - h_3^2 = \frac{1}{4^2}(r_1^2 - h_1^2) .$$

Proceeding in this way, after n steps, we have

$$r_{n+1}^2 - h_{n+1}^2 = \frac{1}{4^n}(r_1^2 - h_1^2) . \qquad (7)$$

But r_1 the hypotenuse of the triangle is greater than h_1 the altitude (see Fig. 10), so that $r_1^2 - h_1^2$ is positive and hence

$$r_{n+1} > h_{n+1} . \qquad (8)$$

Consequently $r_n - h_n > 0$, so by (5)

$$h_{n+1} - h_n = \frac{1}{2}(r_n - h_n)$$

and therefore

$$h_{n+1} > h_n \qquad (9)$$

Squaring (6) and dividing by $r_n \cdot r_{n+1}$

$$\frac{r_{n+1}}{r_n} = \frac{h_{n+1}}{r_{n+1}}$$

but by (8)

$$1 > \frac{h_{n+1}}{r_{n+1}} .$$

so

$$1 > \frac{r_{n+1}}{r_n}$$

i.e.,

$$r_{n+1} < r_n . \qquad (10)$$

Arithmetic confirms intuition. (9) shows that successive values of h_n increase (so that if there exists an R, $h_n < R$), while (10) shows that successive values of r_n decrease (so that if there exists an R, $R < r_n$) and

(7) shows that if either r_n or h_n converges then both converge to the same limit. Conjointly these results imply that there is an R such that for all n, $h_n < R < r_n$, and that

$$\lim_{n \to \infty} h_n = R = \lim_{n \to \infty} r_n .$$

Let's be specific. Taking the hexagon as our initial polygon, with $r_1 = 1$ and (consequently) $h_1 = \frac{3}{2}$, by (7)

$$r_{n+1}^2 - h_{n+1}^2 = \frac{1}{4^n} (1^2 - \frac{3}{4}) = \frac{1}{4^{n+1}} .$$

so that

$$r_{n+1} - h_{n+1} = \frac{1}{4^{n+1}} \cdot \frac{1}{r_{n+1} + h_{n+1}} .$$

But, by (9) $h_{n+1} > h_1 = \sqrt{\frac{3}{2}}$, and by (8) $r_{n+1} > h_{n+1} > \sqrt{\frac{3}{2}}$, so that $r_{n+1} + h_{n+1} > \sqrt{\frac{3}{2}} + \sqrt{\frac{3}{2}}$ and $\frac{1}{r_{n+1} + h_{n+1}} < \sqrt{\frac{1}{3}}$.

Therefore,

$$r_{n+1} - h_{n+1} < \frac{1}{4^{n+1}} \cdot \frac{1}{\sqrt{3}} < \frac{1}{4^{n+1}} .$$

That is to say that the difference between the radii of the circumscribing and inscribed circles of the polygon obtained after n steps from the initial hexagon is $< \frac{1}{4^{n+1}}$. Convergence is rapid.

Since $2\pi R = k$ and in this example the perimeter of the hexagon is 6 (as remarked earlier), $R = \frac{6}{2\pi} = \frac{3}{\pi}$, so that the successive values of h_n increase to $\frac{3}{\pi}$ while the successive values of r_n decrease to it. That is,

$$h_n < \frac{3}{\pi} < r_n ,$$

so that

$$\pi < \frac{3}{h_n} , \quad \frac{3}{r_n} < \pi$$

i.e.,

$$\frac{3}{r_n} < \pi < \frac{3}{h_n} .$$

ERIC Full Text Provided by ERIC

Even the case $n = 1$ is interesting

$$\frac{3}{1} < \pi < \frac{3}{\sqrt{\frac{3}{2}}}$$

i.e.,

$$3 < \pi < 2\sqrt{3} \approx 3.4 .$$

Surely the reader will want to work out $n = 2, 3, 4$ (and maybe others) for himself.

Finally, with the suggestion that the reader take a second look at Fig. 15 and the reminder that

$$\lim_{\theta \to 0} \frac{\sin \theta}{\theta} = 1$$

he is urged to prove that in the general case the limit of r_n and the limit of h_n, i.e., R, is given by

$$R = \frac{\sqrt{r_1^2 - h_1^2}}{\arccos \frac{h_1}{r_1}} .$$

2.6 Arithmetic and Geometric Means.

M_A the arithmetic mean of $a_1, a_2, a_3, \ldots, a_n$ is defined by

$$M_A = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n} .$$

M_G the geometric mean of these quantities is defined by

$$M_G = \sqrt[n]{a_1 \cdot a_2 \cdot a_3 \ldots a_n}$$

Thus, for example, (5) of the last section states that h_{n+1} is the arithmetic mean of r_n and h_n, while (6) states that r_{n+1} is the geometric mean of r_n and h_{n+1}.

If all the quantities $a_1, a_2, a_3, \ldots,$ are equal then

$$M_A = \frac{na_1}{n} = a_1 \quad \text{and} \quad M_G = \sqrt[n]{a_1^n} = a_1$$

so that $M_A = M_G$. If not all the quantities are equal then $M_A > M_G$. This is very easily proved in the simple case $n = 2$. For the two quantities a, b we have

$$M_A = \frac{a + b}{2}, \quad M_G = \sqrt{ab}$$

therefore

$$M_A - M_G = \frac{1}{2}(a + b - 2\sqrt{ab}) = \frac{1}{2}[(\sqrt{a})^2 - 2\sqrt{a}\sqrt{b} + (\sqrt{b})^2]$$

$$= \frac{1}{2}(\sqrt{a} - \sqrt{b})^2$$

but $\frac{1}{2}(\sqrt{a} - \sqrt{b})^2 > 0$ unless $a = b$. This proves the proposition.

What are the uses of these means? If n independent measurements are made of the same quantity, if, for example, $a_1, a_2, a_3, \ldots, a_n$ are the n numbers independently obtained for the distance of the sun from the earth, then the arithmetic mean is the most reliable estimate. Gauss' argument to this effect is well known. Less well known is his application of the geometric mean. This follows.

How is a weight W to be accurately determined by using badly made scales? How, for example, with scales of which one arm is longer than the other? We suppose that W, when placed in the right and left pans, counterbalances weights W_1, W_2, respectively. What is the actual weight of W? Study Figs. 16 and 17.

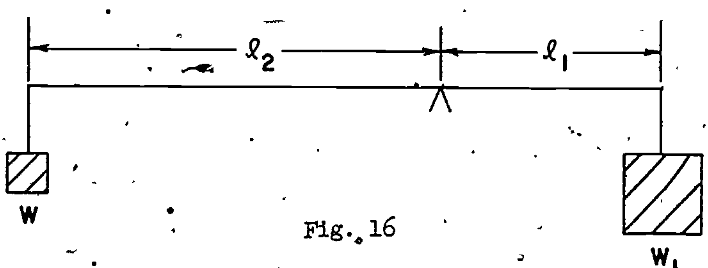

Fig. 16

For equilibrium in Fig. 16 we have

$$\ell_2 W = \ell_1 W_1 \qquad (1)$$

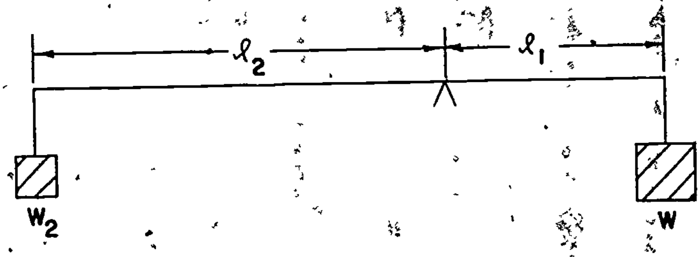

Fig. 17

and for equilibrium in Fig. 17,

$$\ell_1 W = \ell_2 W_2 \quad (2)$$

Now we come to Gauss' important observation. Multiplying (1) by (2)

$$\ell_1 \ell_2 W^2 = \ell_1 \ell_2 W_1 \cdot W_2$$

so that

$$W = \sqrt{W_1 \cdot W_2} .$$

Thus W is independent of the lengths of the arms. Use of the geometric mean rectifies this imprecision of the scales.

Chapter 3. The Role of Mathematics in Optics.

To illustrate the part played by mathematics in the construction of theories in science, I wish to consider the development of optics.

3.1 Euclid's Optics.

We begin with Euclid (c. 300 BC). Not unnaturally for a geometer, he wished, as doubtlessly had many geometers before him, to apply geometry to optics. Unlike the others he was successful. Conceiving light as propagated in straight lines enabled him to apply geometry to optics. On second thoughts this statement cannot stand. Until Euclid had applied geometry to optics there was, to use the Irish idiom, no such subject as optics. Nowadays, when using diagrams is an ingredient of educated common sense, of course it is obvious that light is propagated in straight lines. If light rays could not be represented by lines, optical phenomena could not be illustrated by diagrams. We, with the arrogance of hindsight, cannot begin to understand Euclid's foresight in making his basic assertion that light is rectilinearily propagated. When the needle in the haystack has been pointed out to us, we are prone to suppose that finding it was no problem at all.

Physical objects that more or less crudely approximate to straight lines readily come to mind, for example, a taut wire. But surely a shaft of sunlight piercing the shutters of a darkened room is singularly apt. Isn't this the perfect example? Euclid must have been well pleased with his observation. Yet note that his basic assertion embraces metaphysical speculation as well as physical observation. We see only the shafts of light at which we look; we do not see the shafts with which we look. We cannot observe the rays with which we observe, yet Euclid claims all rays to be propagated in straight lines. Such metaphysical assumptions regarding unobservables are acceptable in so far as they facilitate understanding of observables. Is his postulate obvious? Your answer depends upon how much or how little you think about it.

ERIC Full Text Provided by ERIC

Given that rays of light are straight lines, how, Euclid asked, is the direction of a ray striking the surface of a plane mirror related to that of the reflected ray? See Fig. 1.

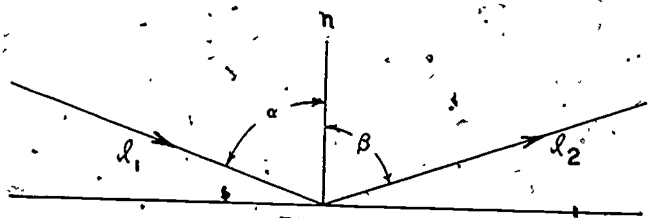

Fig. 1.

This figure reduces optics to geometry. The lines ℓ_1, ℓ_2, n, represent the incident ray, the reflected ray, and the normal to the surface at the point of incidence, respectively. The angle α between incident ray and normal is termed the angle of incidence, while the angle β between reflected ray and normal is termed the angle of reflection. What is the relation between β and α?

Euclid found by experiment that ℓ_2 lies in the plane determined by ℓ_1 and n. Thus ℓ_1, n, and ℓ_2 in Fig. 1 may be considered to lie in the plane of the paper. To determine ℓ_2 uniquely, it remains to specify β. As the result of many experiments Euclid found that $\beta = \alpha$, i.e., that angle of reflection is equal to angle of incidence. This is the famous law of reflection as formulated by him in his Optics.

Although this law was based on a large number of experiments we must remember that Greek technology was rudimentary, their measuring instruments imprecise, and their plane mirrors imperfect. What assurance had Euclid that β was precisely equal to α? He had the comforting security of experiment backed by belief. He held a possibly only half-articulate, but certainly deep-seated, belief about the nature of things; that Nature is not fortuitous; that her laws will have simplicity and elegance. With the courage of conviction he asserted his law to hold exactly for perfectly plane mirrors. But many of Euclid's contemporaries, even if equally courageous, had grave doubts whether his law is right. Some with different metaphysics doubted whether there could be laws of nature at all.

3.2 Heron: The Shortest Path Principle.

To add grounds for belief we introduce Heron of Alexandria who lived a generation or so after Euclid. (His birth and death dates are uncertain.) A man who played a far greater role in the development of science than that usually ascribed in the textbooks, he built the first automaton, made the first attempt at building a steam engine, developed trigonometry and applied it extensively. A man with both feet on the ground, he was forever stressing the possibilities of applying mathematics.

Heron gave a proof of Euclid's law of reflection. His proof consists of showing that both of Euclid's laws, that

E_1 Light is propagated rectilinearily

E_2 Angle of Reflection = Angle of Incidence

are consequences of the principle proposed by Heron himself, that

H Light takes the shortest path possible.

Here we have what is probably the first example of the unifying trend so characteristic of science. Surely either of E_1, E_2 could be true without the other. Is it not perfectly reasonable to conceive of light being propagated in straight lines without $\beta = \alpha$, and conversely? But H could not be true without both being true. Whereas belief in the truth of both E_1, E_2 merely affords grounds for believing H, believing H necessitates believing both E_1, E_2. Moreover, the complete formulation of E_2 is complicated, while H, like E_1 is simple. Is it not easier to believe one statement of a certain kind than twenty or two of the same or a more complicated kind? It is in this sense that Heron "proved" Euclid's law of reflection.

The proof that E_1 follows from H is obvious. Since the shortest distance between any two points A and B (in free space) is the straight line AB that joins them, light, in moving from A to B by the shortest path possible, is necessarily propagated rectilinearily. Fig. 2 is self-explanatory.

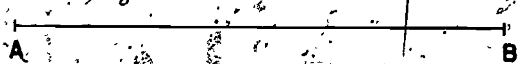

Fig. 2.

The proof that E_2 follows from H is not obvious. We suppose light to travel from A via some point P in the mirror surface to B. PN is the normal to the mirror surface at P. See Fig. 3.

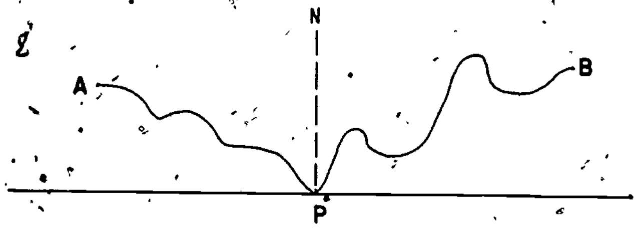

Fig. 3.

If the light did not become incident to the mirror surface, then the light could not be reflected from it. Here, in asserting that a ray takes the shortest path possible from A to B, we cannot mean the shortest of all possible paths (the straight line AB), we must mean the shortest possible path via the <u>surface of the mirror</u>. Thus to prove that E_2 is a consequence of H is to prove that, if APB is the shortest path possible (via the mirror surface), then the angles made by the <u>straight</u> lines AP, PB with PN are equal.

First we show that the lines AP, PB, cannot be wiggly. The distance from A to B via P will be a minimum when AP and PB are both minima; for if both were not minima their sum could be decreased. But the minimum distance between any two points is the straight line joining them, so that the distance from A to B via P can be a minimum only if AP and PB are both straight lines. Accordingly, we exclude wiggly lines from further consideration.

This leads us to the crux of the proof. What is the position of P such that the sum of the straight line distances AP, PB is a minimum? At this stage we avail ourselves of Heron's ingenuity by introducing an auxiliary point B', the mirror image of B. That is to say, B' is the point on the normal from B to the mirror as far below the surface as B is above it. See Fig. 4.

Since MC is perpendicular to BB' and C is the midpoint of BB', MC is the perpendicular bisector of BB'; i.e., MC is the locus of points equidistant from B and B'. Therefore, no matter what point P is on MC,

$$PB = PB'$$

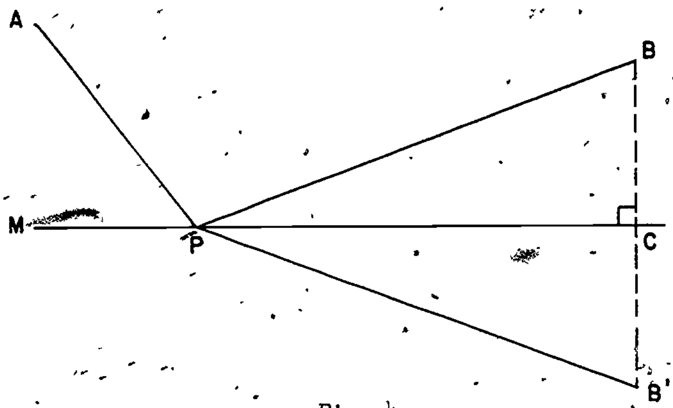

Fig. 4

and consequently

$$AP + PB = AP + PB' .$$

The former will be a minimum only when the latter is a minimum. But the shortest distance between A and B' is the straight line joining them, so that the latter, and consequently the former, will be minima when P is collinear with A and B'.

It remains merely to show that when APB' is a straight line, the angles made by AP and BP with the normal at P are equal. Study Fig. 5.

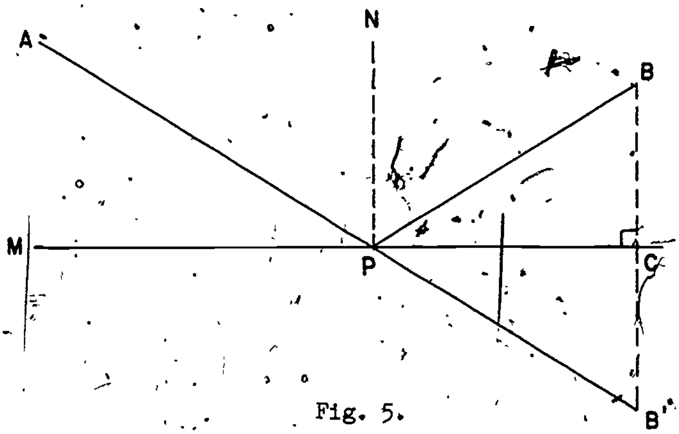

Fig. 5.

Since APB' is now a straight line, MPA and B'PC are vertically opposite angles, so that

$$\angle MPA = \angle B'PC .$$

But

$$\angle BPC = \angle B'PC ,$$

by symmetry (and also by congruence of triangles PBC and PB'C, from two sides and their included angle, PC = PC, $\angle PCB = 90^\circ = \angle PCB'$, CB = CB'), so that

$$\angle MPA = \angle BPC .$$

Therefore the complement of the former is equal to the complement of the latter, i.e.,

$$\angle APN = \angle BPN .$$

This completes the proof that E_2 is implied by H.

The critical reader may well ask, "How did Heron hit upon the idea of the auxiliary point B'?" But haven't we all seen swan and reflection floating double on a placid lake? The swan's image is the same size as the swan, but upside down. In terms of Fig. 5, if MC represents the lake surface and CB the swan, then CB' represents the swan's image; in particular B' is the image of B. To an observing eye at A looking along AP, B appears to be on AP produced at B'. To see B "in" a reflecting surface is to see it as if it were at B' and there were no reflecting surface. The concept of mirror image enables us, in effect, to throw away the mirror and reduce the problem of a reflected ray's path to that of a nonreflected ray. By E_1 the path of light from A to B' (when no mirror intervenes) is the straight line AB', the shortest path possible. We can but suppose that Heron had such considerations as these in mind when he pondered the problem; for ponder the problem he did.

3.3 Ptolemy and Refraction.

The further development of optics leads us to the work of the great Alexandrian astronomer, Ptolemy, who flourished 127 to 141 or 151 AD. Shortly after the time of Heron deep interest in astronomy raised other questions concerning the nature of light. Ptolemy found from his observations of the stars that the propagation of light near the earth's surface is not precisely rectilinear, but slightly curved. On the analogy of a straight stick partially immersed in water, appearing bent, he ascribed the curvature of light to passage through layers of air of different density.

Textbook writers would have us believe that the Greeks were interested only in the things that they could see but not touch. To the contrary, a vast amount of experimental work was done in Alexandrian times: Ptolemy, to better understand the effect of change in density on the bending of light rays by the atmosphere, conducted experiments to measure the deflection of light rays in passing from air to water. See Fig. 6.

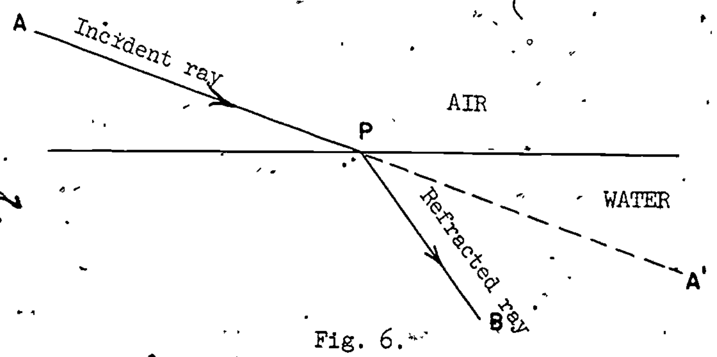

Fig. 6.

Upon penetrating the surface of the water the incident ray does not continue along AP (produced), but at an angle to it. The deflected ray, is said, to use the commonly accepted term, to be refracted. Possibly the reader is disposed to take ∠BPA' as a measure of the refraction. Ptolemy did not do this. Refraction is sufficiently similar to reflection to merit analogous terminology. With both phenomena there is a ray incident to a surface, and therefore an angle of incidence. The only difference is that whereas with reflection the ray after incidence is determined above the surface, with refraction it is deflected below it. Is it not therefore natural to measure angles for both phenomena with reference to the normal to the surface; to use the same definition of angle of incidence for both; and since a ray deflected upwards is measured against the upward normal, to measure against the downward normal a ray deflected downwards? Thus in Fig. 7, α is said to be the angle of incidence and β, the angle of refraction.

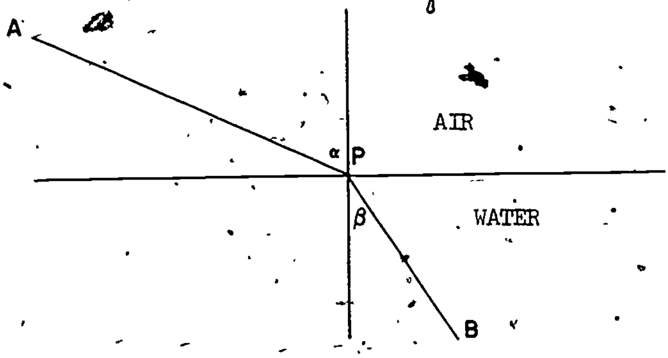

Fig. 7.

Ptolemy found that β depends upon α ; a change in the angle of incidence results in a change in the angle of refraction. Mathematically put, β is a function of α, say, $\beta = f(\alpha)$. As a first step toward specification of $f(\alpha)$ Ptolemy made extensive tabulation of the ordered pairs, α with the corresponding β. Despite more and yet more experiments with extensive and yet more extensive tabulation, the law of ordering continued to elude him. Finally he had to give up.

3.4 Kepler and Refraction.

More than a thousand years later the problem was tackled by Kepler (1571-1630), an astronomer justly famous, who had genius at finding the functional relation governing the most recalcitrant of ordered pairs. Allow me to illustrate his capacity.

Year after year he worked away, conjecturing and checking, until finally he hit upon hypotheses that fit his observational data. He showed that each planet describes an ellipse having the sun at one of its foci, and that the areas described by the radii drawn from a planet to the sun are proportional to the time taken by the planet to describe them. For each planet he knew r, the maximum distance of its elliptical orbit from the sun, and for each he calculated the planetary year T, the time it takes to complete a full orbit. He tabulated T and r. He asked himself, What is the functional relation between them? He found the answer. Incredible man.

That he may have some measure of Kepler's achievement, the reader is asked to seek the relation of T to r for the following tabulation.

T	r
287,496	484
1,601,613	1,521
2,146,689	1,849
4,251,528	2,916
4,721,632	3,136
7,414,875	4,225
9,261,000	4,900

Not obvious, eh? Alas, to find is to seek successfully. After hours or days of unsuccess we likely concede that such problems demand a Kepler. But these are neat and tidy figures, tailor-made for the occasion; devoid of messy decimals, our tabulation has none of the more-or-less-ness of the observational data of Kepler's problem. His was difficult.

A hint. Our r column contains naught but perfect squares. Is the relation of T to r now obvious? No, the "obvious" conjecture is wrong; T is not also a perfect square. No, neither is T the sum of two perfect squares. T, it so happens, is a perfect cube. Advantageously we rewrite our tabulation.

T	r
66^3	22^2
117^3	39^2
129^3	43^2
162^3	54^2
168^3	56^2
195^3	65^2
210^3	70^2

Is the relation between T and r now apparent to you? That depends upon your discernment. Perhaps you notice that neglecting exponents the first number of each ordered pair is three times the second, that

$$\frac{66}{22} = \frac{117}{39} = \frac{129}{43} = \cdots = 3 .$$

Thus, for example,

$$66 = 3 \cdot 22$$

so that the entry for T <u>with exponent</u> is

$$66^3 = 3^3 \cdot (22^3) .$$

What a pity the 22 is cubed instead of squared. Thinking wishfully, we write

$$(66^3)^2 = (3^3)^2 \cdot (22^2)^2 . (22)^2 .$$

It is left to the reader to show that

$$T^2 = 729 \cdot r^3$$

satisfies our tabulation.

Kepler's tabulation, though difficult, was governed by the same proportionality. He found that

$$T^2 = k \cdot r^3$$

where k is a constant. This is his famous third law that the square of the time of revolution of a planet about the sun is proportional to the cube of that planet's maximum distance from it. Although our tabulation with nice whole numbers devoid of observational error inadequately illustrates his achievement, it does afford some hint why Kepler's discovery cost him nearly a decade of incessant toil.

With equal enthusiasm Kepler turned to the refraction problem of specifying β in terms of α. Knowing his ability, we anticipate his success. His formula works well for small α, but the greater α becomes the greater its inaccuracy. For α greater than 15° its inaccuracy is unacceptable. It is a makeshift affair; even Kepler was unsuccessful.

3.5 <u>Fermat</u>: <u>The Quickest Path Principle</u>.

Although the reader is understandably impatient to learn the correct formula, the development of science is not to be hurried. Solution of long standing problems is attendant upon the winds of fresh discovery; the new ideas of a lively intelligence, stimulated by the intellectual ferment of its day. The lively intelligence was Fermat's (Fermat 1601-1665); the intellectual ferment of its day, the question, "Does light have a velocity or is its propagation instantaneous?"

Possibly Galileo (1564-1642) was the first to tackle this question experimentally. At night on a mountain top he signalled with a lantern to a colleague on an adjacent mountain. His colleague, on seeing the light of Galileo's lantern, uncovered his own. Galileo tried to measure the interval between dispatch of his signal and receipt of his colleague's. As near as he could tell, light is instantaneous. To us the experiment is incredibly naive, but Galileo did not know that the time for say, two 10-mile light journeys, is of the order of one ten-thousandth of a second. He experimented to find out.

This live issue captured Fermat's attention. Suppose, he pondered, light is not instantaneously propagated, but has a velocity. Further suppose this velocity to be constant. What then? Time is distance divided by velocity; the shortest path is the quickest. The supposition that light takes the shortest time has precisely the same consequences as Heron's principle that it takes the shortest path. But alternatively, suppose that the velocity of light while constant for any given medium, is different for different media. In particular, suppose that light in water has a velocity different from light in air. What then? With travel in both air and water the shortest path conceivable is not the quickest. A bent line is longer than the straight line between the same end points; simply because a refracted ray is refracted it cannot take the shortest path: does it take the quickest? If so, the consequences of Heron's shortest path principle still hold, and perhaps refraction is also explicable.

Further thought gives this conjecture further plausibility. E_1 could be expressed as a minimum principle. A straight line bends neither to the one side or the other; it has zero curvature. So, instead of saying that light is propagated rectilinearily, why not say that light takes the path of minimum curvature? Heron's minimum principle that light takes the shortest path is more embracing and covers reflection as well as E_1: Why not an even more embracing minimum principle that covers refraction as well as reflection and E_1?

Fermat's conjecture explains at least as much as Heron's. Does it explain more? What precisely are the implications for refraction of the minimum principle that light takes the quickest path possible? What is the quickest path? Consider the plight of a golfer who in driving from the fairway at A, hooks (not slices--the golfer is left-handed) his ball into the box at B. See Fig. 8.

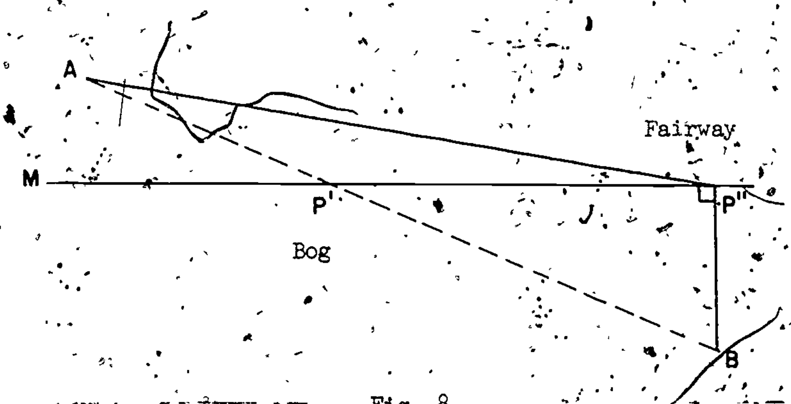

Fig. 8.

How best can he retrieve his ball? Not by taking the shortest route AP'B, but by taking the route AP"B which minimizes the amount of bog that he, frantically determined as golfers are, must flounder through waist deep. Clearly this is the quickest possible route. Compare it, for example, with AP'B. His longer walk AP" across the fairway where the going is easy and therefore rapid takes him at most a minute or two more, but his shortest bog route, P"B, saves him an hour or two of floundering. The extra time spent in walking farther across the fairway is negligible compared with the time saved from battling through bog. Similarly AP"B compares favorably against any other route. It is intuitively evident that floundering is so exasperatingly slow that as increase in the floundering distance cannot be compensated for by the corresponding decrease in fairway travel. Therefore the quickest route has the minimum of bog travel, i.e. that in which BP" is perpendicular to MP". We have solved the quickest path problem in the extreme case where the golfer's fairway velocity V_1 is very large compared with (since almost zero) V_2, his bog velocity.

What is the other extreme case? That in which bog is replaced by fairway, so that V_2 is increased to V_1. Then, of course, the quickest path is the shortest path AP'B. And what about intermediate cases? Surely the quickest route from A to B is APB where P moves rectilinearly from P' to P", as V_2 decreases from V_1 to nearly zero.

Suppose the bog replaced by bracken and gorse. Off the fairway the going is not so desperately bad as floundering through bog, but more arduous than fairway walking; we expect P to be intermediate between P' and P". Were the going rougher than it is off the fairway, our golfer would go farther out of his way (i.e., deviate farther from the shortest route AP'B) to cut down the amount of rough, time-consuming, terrain he need travel across; were it less rough he would go less far out of his way. Less time-consuming terrain would necessitate a smaller, more time-consuming terrain a greater, deviation. It is intuitively clear that as V_2 decreases from V_1 to nearly zero, the quickest route is such that P moves from P' to P".

We suppose APB to be the quickest route from A on the fairway to B in the rough. See Fig. 9.

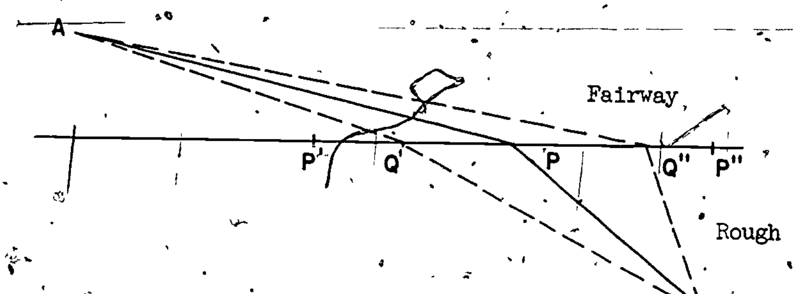

Fig. 9.

Let us compare AQ'B with the quickest route. In taking the former route our golfer has less fairway to stride across, namely AP-AQ', so that his time saving on fairway travel is $\frac{AP - AQ'}{V_1}$. But with less fairway travel he has more of the rough to cross, namely BQ'-BP, so that his extra time spent in the rough is $\frac{BQ' - BP}{V_2}$. Although he gains $\frac{AP - AQ'}{V_1}$ in striding across the fairway, he loses $\frac{BQ' - BP}{V_2}$ in struggling through the rough. Yet all told, he must lose more time than he gains, for otherwise the latter route could not be

ERIC

the quickest. That is to say

$$\frac{BQ' - BP}{V_2} - \frac{AP - AQ'}{V_1} = t' \qquad (1)$$

where t' is the time by which route $AQ'B$ exceeds the quickest. Moreover, the closer the route $AQ'B$ to the quickest, the quicker it becomes; that is, the closer Q' approximates to P, the more t' decreases toward zero. In short, (1) is such that positive t' tends to zero as Q' tends to P.

Next let us compare $AQ''B$ with the quickest route. With the former route our golfer has $AQ'' - AP$ of extra fairway travel, which loses him $\frac{AQ'' - AP}{V_1}$, but $BP-BQ''$ less struggling in the rough, which gains him $\frac{BP - BQ''}{V_2}$. But all told he must lose more time than he gains, for otherwise the latter could not be the quickest route. That is to say

$$\frac{AQ'' - AP}{V_1} - \frac{BP - BQ''}{V_2} = t'' \qquad (2)$$

where t'' is the time by which route $AQ''B$ exceeds the quickest. And as with the previous comparison, (2) is such that positive t'' tends to zero as Q'' tends to P.

Let us compare the left side comparison $AQ'B$ of the quickest route with the right side comparison $AQ''B$. The condition (2) is equivalent to

$$-\frac{AP - AQ''}{V_1} - \frac{BQ'' - BP}{V_2} = t'',$$

and consequently, equivalent to

$$\frac{BQ'' - BP}{V_2} - \frac{AP - AQ''}{V_1} = t''. \qquad (3)$$

Compare (1) with (3). We observe that the former, when $Q' = Q$ and $t' = t$, and the latter, when $Q'' = Q$ and $t'' = t$, is the condition

$$\frac{BQ - BP}{V_2} - \frac{AP - AQ}{V_1} = t, \qquad (4)$$

where positive t tends to zero as Q tends to P. That is to say, the relation between AQB and the quickest route is governed by (4) no matter whether Q is to the left or to the right of P. We have answered the question, "What is the quickest path?" The quickest path APB is such that when compared with any other path AQB, condition (4) is satisfied.

Few mathematicians, even among those of the first rank, could claim invention of the calculus. Fermat is one of the few. To solve the problem of the quickest path he invented the method of the calculus of variations. To the basic idea of this method, the reader has, in following the plight of our golfer, been afforded an intuitive introduction. With the fairway replaced by air, the rough by water, and our golfer by a ray of light, (4) is immediately applicable to the problem of a refracted ray taking the quickest path possible.

See Fig. 10.

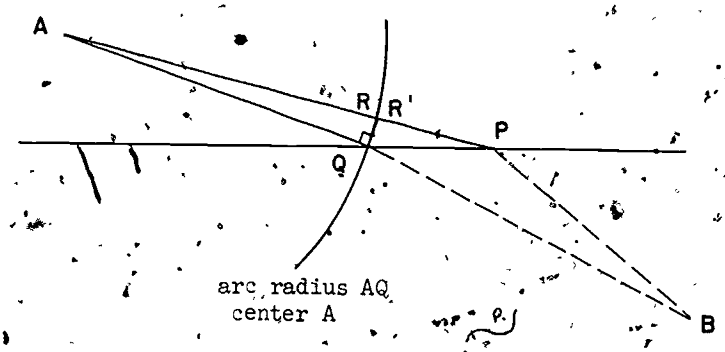

Fig. 10.

The circle with center A and radius AQ cuts AP in R, and its tangent at Q (perpendicular to AQ, of course) cuts AP in R'. $\angle PQR' = \gamma$.

What happens as Q moves closer and closer to P, i.e., as $Q \to P$? Since exterior angle of triangle is equal to the sum of the two interior opposite angles,

$$\angle QR'P = 90^\circ + \angle QAP .$$

But, as $Q \to P$, clearly $\angle QAP \to 0$, so that $\angle QR'P \to 90^\circ$, and consequently, $R'P \to QP \cdot \sin\gamma$. Moreover, as $Q \to P$, $RP \to R'P$, i.e., $AP - AQ \to R'P$; consequently, $AP - AQ \to QP \cdot \sin\gamma$. Therefore,

$$\frac{AP - AQ}{V_1} \to QP \cdot \frac{\sin\gamma}{V_1} \quad \text{as} \quad Q \to P . \qquad (5)$$

Next, study Fig. 11.

ERIC
Full Text Provided by ERIC

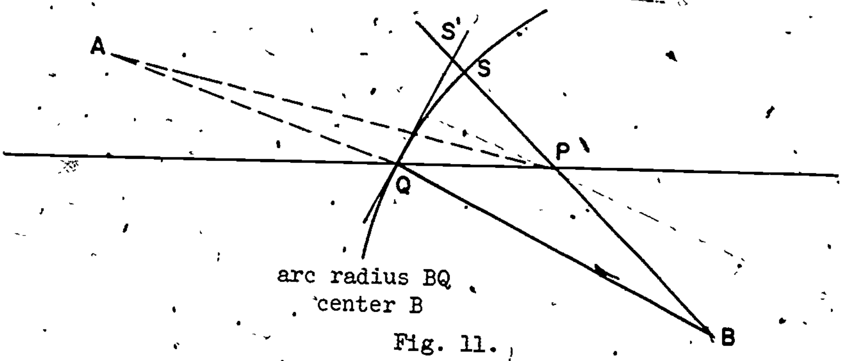

Fig. 11.

The circle with center B and radius BQ cuts BP (produced) in S, and its tangent at Q (perpendicular to BQ, of course) cuts BP (produced) in S'. $\angle PQS' = \delta$.

It is left as an exercise for the reader to show in a precisely similar way that

$$\frac{BQ - BP}{V_2} \to QP \cdot \frac{\sin \delta}{V_2} \quad \text{as} \quad Q \to P . \qquad (6)$$

By (4)

$$\frac{BQ - BP}{V_2} - \frac{AP - AQ}{V_1} \to 0 \quad \text{as} \quad Q \to P$$

so that by (5) and (6)

$$QP \cdot \frac{\sin \delta}{V_2} - QP \cdot \frac{\sin \delta}{V_1} \to 0 \quad \text{as} \quad Q \to P .$$

That is to say that, when the routes AQB, APB are arbitrarily close, the difference

$$QP \left\{\frac{\sin \delta}{V_2} - \frac{\sin \gamma}{V_1}\right\}$$

is arbitrarily small; the closer AQB is to the quickest route, the more nearly true that

$$\frac{\sin \delta}{V_2} - \frac{\sin \gamma}{V_1} = 0 .$$

In other words, for the quickest route equality holds; i.e.,

$$\frac{\sin \delta}{\sin \gamma} = \frac{V_2}{V_1} . \qquad (7)$$

Since in the limiting position the routes AQB, APB are arbitrarily close, $\angle QR'P$ differs from a right angle by an arbitrarily small amount, so that the situation of Fig. 12 obtains.

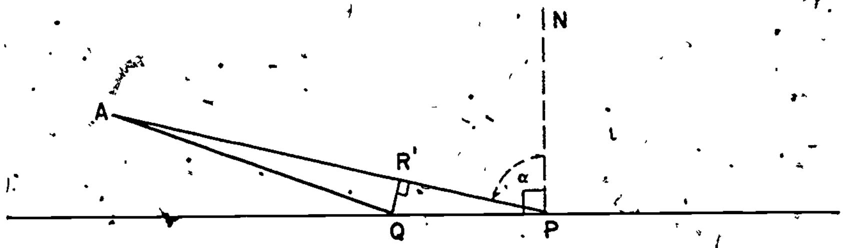

Fig. 12.

Thus α, the angle of incidence of the light ray AP with the normal PN, and γ are both complements of $\angle$QPR'. Therefore

$$\alpha = \gamma. \tag{8}$$

Similarly, in the limiting position the situation of Fig. 13 obtains.

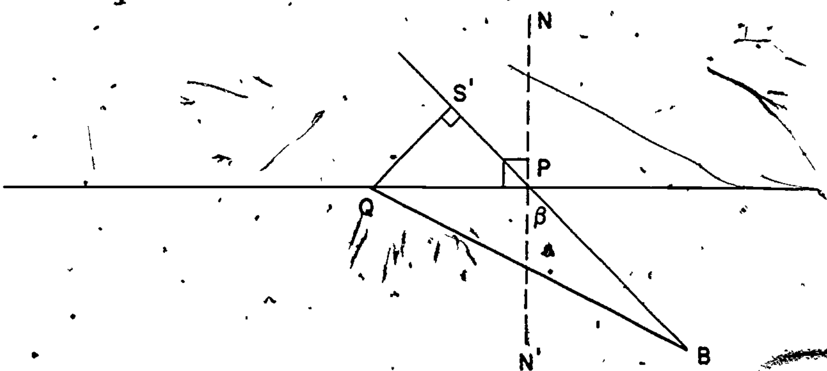

Fig. 13.

Consequently, δ and $\angle$S'PN are both complements of $\angle$QPS', and the latter is vertically opposite to $\angle$BPN', the angle of refraction β. Therefore

$$\beta = \delta. \tag{9}$$

Substituting (8), (9), in (7), we have

$$\frac{\sin \beta}{\sin \alpha} = \frac{V_2}{V_1}. \tag{10}$$

This is Fermat's law of refraction, which may be alternatively expressed

$$\sin \beta = K_1 \cdot \sin \alpha, \quad \text{where} \quad K_1 = \frac{V_2}{V_1} \tag{11}$$

i.e., that $\sin \beta$ is directly proportional to $\sin \alpha$, where the constant of proportionality is $\frac{V_2}{V_1}$.

Fermat's law was later rediscovered independently by both Snell and Descartes, and was used by the latter to explain the phenomenon of the rainbow. It won rapid acceptance with contemporary scientists. From (10) we have

$$\beta = \arcsin \left(\frac{V_2}{V_1} \sin \alpha \right).$$

That Kepler failed to conjecture such a complicated functional relationship between β and α occasions no surprise. What is surprising is that he was so successful as to find a formula of tolerable accuracy for $\alpha < 15^\circ$.

3.6 Newton's Mechanistic Theory of Light.

With Newton (1642-1727) science came of age. Understanding the starry heavens was within man's grasp. It was almost as if Newton with his three laws and few axioms could, as Jesus with his two loaves and five fishes, work miracles. He explained the ebb and flow of the waters of the deep and the passage of the fiery bodies in the firmament above. Nature lost her mystery; man his impotence. The solar system is a gigantic piece of clockwork, and Newton had discovered how it ticks. Newton's mechanics is the key to everything around the sun; must it not be the key to everything under the sun? To the enthusiasm born of his success, his laws and axioms were as clear as day. The minimal path principles of Heron and Fermat were still darkly mysterious. Surely the optics of Euclid, Heron, and Fermat could be explained mechanistically. Newton thought so.

Newton begins at the beginning. The first thing to explain is the rectilinear propagation of light. His first law states that a body moving with uniform velocity in a straight line will continue to do so unless acted upon by external forces to change that motion. Are not Euclid's and Newton's first laws remarkably similar? With characteristic ingenuity Newton makes the former as an immediate consequence of the latter by introduction of the supposition that a ray of light consists of minute bodies, particles, or corpuscules. Because of this supposition Newton's theory of light is known as the corpuscular theory.

How does Newton account for Euclid's law of reflection? According to his corpuscular theory, an incident ray of light is reflected because of the collision of its constituent particles with those constituting the surface of the mirror. It is, of course, sufficient to consider the fate of a single incident particle, for all the others will behave in the same way under similar circumstances.

First let us consider a special case, that of an incident ray normal to the mirror. See Fig. 14.

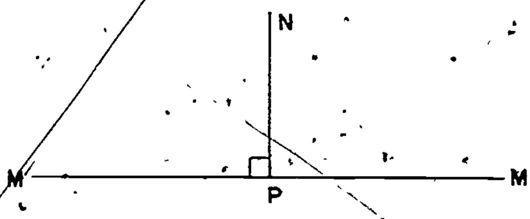

Fig. 14.

What happens when a constituent particle of a ray travelling along NP reaches P? Its attempt to penetrate the surface MM' perpendicularly downwards is resisted solely by forces acting perpendicularly upwards (due to the constituent particles of the mirror surface in the neighborhood of P). Consequently the particle returns along the normal.

We now turn to the general case. It is assumed that V_1, the velocity of a light ray in air, is constant irrespective of its direction relative to the mirror. Thus the problem is the following. A particle travelling with velocity V_1 along AP at an angle α to the normal is reflected with velocity V_1 along BP which makes some angle β with the normal. What is the relation between β and α? See Fig. 15.

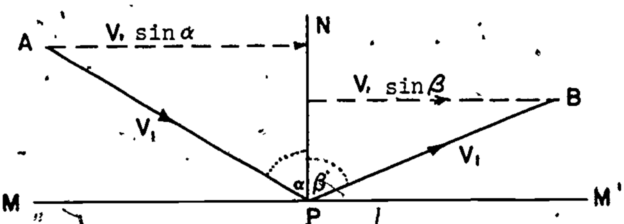

Fig. 15.

Despite the fact that the particle at P now attempts to penetrate MM' obliquely, Newton insists that the structure of this surface is such that the resistance to penetration is solely by forces acting perpendicularly upwards -- just as in the special case considered above. What are the consequences of his insistence? Forces acting in the direction PN have no components in the direction MM', so that the forces (if any) acting in this direction on the particle at P before impact are unchanged by impact. Therefore the velocity of the particle parallel to MM' when part of the reflected ray, is just the same as when part of the incident ray. Motion parallel to the surface remains unchanged. Equating component velocities parallel to MM', we have

$$V_1 \sin \beta = V_1 \sin \alpha$$

But, by hypothesis, the resultant velocity of the particle when reflected is V_1 at an angle β to PN. Hence it is clear from Fig. 16 that $\beta = \alpha$.

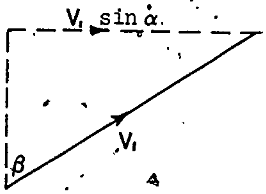

Fig. 16.

Next, refraction. What is the difference between reflection and refraction? Whereas in the latter the incident ray is successful in penetrating the surface, in the former it is not. Newton treats these phenomena similarly. No matter whether or not penetration is successful, Newton continues to insist that the only forces opposing penetration, even if oblique, act perpendicularly to the surface. Consequently, for refraction as for reflection, motion parallel to the surface remains invariant. And whereas the refracted ray differs from the reflected ray by being propagated in water instead of air, so that its component velocity parallel to MM' is $V_2 \cdot \sin \beta$ instead of $V_1 \cdot \sin \beta$, the incident ray is the same in both cases. See Fig. 17.

ERIC

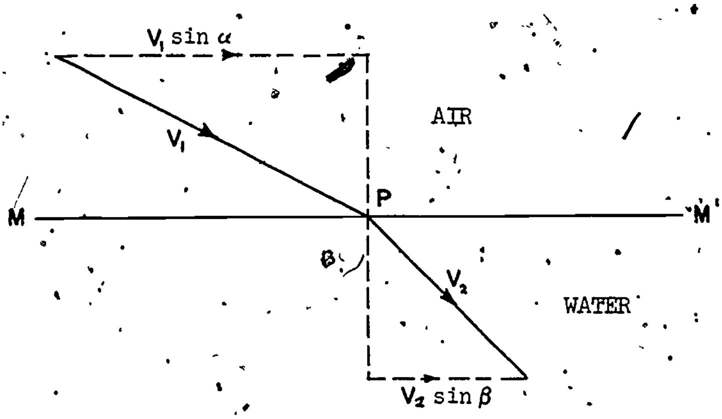

Fig. 17.

Therefore, equating component velocities parallel to MM' ,

$$V_2 \cdot \sin\beta = V_1 \cdot \sin\alpha \,,$$

giving

$$\sin\beta = K_2 \cdot \sin\alpha \,, \quad \text{where} \quad K_2 = \frac{V_1}{V_2} \,. \qquad (12)$$

3.7 Fermat Versus Newton: Experimentum Crucis.

Thus Newton, like Fermat, concludes that $\sin\beta$ is directly proportional to $\sin\alpha$. However, comparison of (11) with (12) also shows that, although their formulae have the same form, Newton's constant of proportionality is the reciprocal of Fermat's. And it is an experimental fact that the refracted ray is bent towards the normal, i.e., $\beta < \alpha$, so that $\sin\beta < \sin\alpha$. Consequently a formula of the form

$$\sin\beta = K \cdot \sin\alpha$$

cannot be correct unless the constant of proportionality K is less than unity. If (11) be correct, $\frac{V_2}{V_1} < 1$; if (12), $\frac{V_1}{V_2} < 1$. Whereas Fermat's formulae cannot be correct unless the velocity of light in water is less than the velocity in air, Newton's cannot be correct unless the precise opposite is the case.

Newton had no difficulty in finding an argument to vindicate his own theory. A particle of light when in air is traveling in a homogeneous medium, so that the forces acting upon it are constant; there being no acceleration the net force must be zero. Similarly in water. But when a particle is passing from one medium to another there is a change from one homogeneity to another, so that the forces acting upon it momentarily are not constant. Water has greater density than air, its particles are more tightly packed. When a particle reaches the neighborhood of the interface, ahead of it is an accumulation of matter, behind it, a sparsity. Consequently, since the more the mass the greater the attraction, the particle has momentarily a terrific acceleration and speeds up from V_1 to V_2. Having passed through the interface, once again there is no net force and the particle continues with constant velocity V_2.

It was a good argument as long as it lasted; it lasted rather more than a century. And then technological advances made it possible to show experimentally that light is slower in water than in air. This was the experimentum crucis. The basis of Newton's argument, that light consists of particles, is untenable. We must add that Fermat enunciated his quickest path principle a quarter of a century before it was known experimentally that the propagation of light is not instantaneous.

3.8 To Recapitulate.

I have traced the development of elementary optics over the centuries up to the formulation of Fermat's and Newton's theories. Both explain rectilinear propagation of light; both account for the law of reflection; both give the same kind of formula for refraction: yet they are rivals. Rivals, for with regard to refraction they differ in detail. Here is a situation typical of science's history; a conflict of theory only to be resolved by determination of fact. That is the role of crucial experiment.

But when a consequence of a theory is in question, the basis of the theory is also in question. In rejecting Newton's consequences for refraction as contrary to fact, we must reject the basis of these consequences--the corpuscular nature of light. That Fermat's theory could explain all the facts vindicated his quickest path principle. Later this developed into the wave theory of light--that the constitution of light is not corpuscles, but waves.

Alghouth space does not permit consideration of further developments, I hope to have shown you something well worth showing of the role of mathematics in the evolution of science. Mathematics sharpens our seeing of logical consequences and focuses our attention on appropriate experimentation; an aid to vision, it is the eyeglass of the mind.

3.9 The Role of Science in Mathematics.

I wish to end with a curious twist. From the role of mathematics in science, we turn to the role of science in mathematics; for despite an abundance of material, how science gives grounds for mathematical theorems is little known. Convenient to our purpose is the problem of how to construct a tangent to the ellipse.

If two pegs F_1, F_2 are hammered into the ground and a cord tied to both of them is kept taut by a stick P, the movement of P under this restraint marks out an elliptical flower bed. See Fig. 18.

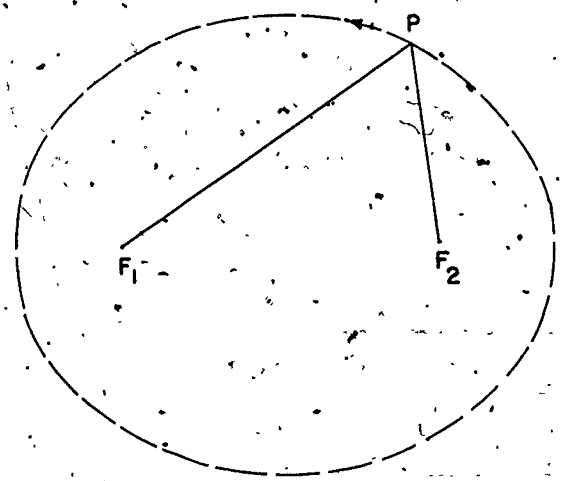

Fig. 18.

This method of construction exhibits the usual, generative, definition of the ellipse. The locus of a point P such that the sum of its distances from two fixed points F_1, F_2 (called the foci) is constant, is said to be an ellipsoid; when P is restricted to one plane through F_1, F_2, its locus is said to be an ellipse. Traditionally the constant sum is taken to be 2a, giving the equation of the ellipse as

$$F_1P = PF_2 = 2a.$$

The circle is a special case of the ellipse, the ellipse a generalization of the circle. When F_1, F_2 become coincident

$$F_1P + PF_2 = 2 \cdot F_1P = 2a$$

so that F_1 (and F_2) become the center of a circle of radius a. This suggests that properties of the circle will be limiting cases of properties of the ellipse. What light does this suggestion throw on the problem of constructing a tangent at P to the ellipse? See Fig. 19.

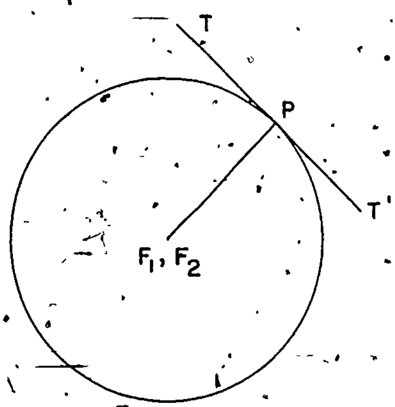

Fig. 19

The tangent at P to the circle with center F_1 is perpendicular to the radius F_1P. How do we go from this limiting case to the general? Equally well we could say that the tangent is perpendicular to F_2P, or that it is perpendicular to both F_1P and F_2P. But obviously the tangent can be perpendicular only to one of these lines when they are no longer coincident. Which one? Surely they have equal claims. What is an acceptable compromise? That F_1P, F_2P are equally inclined to the tangent, i.e., that $\gamma = \delta$. See Fig. 20.

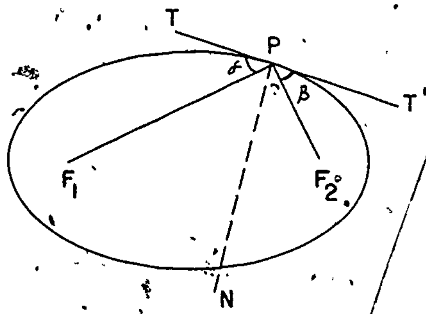

Fig. 20.

This conjecture has merit, for it is consistent with the limiting case; when F_1, F_2 become coincident $\gamma = \delta = 90^\circ$. But to suppose that $\gamma = \delta$ is equivalent to supposing their complements to be equal, i.e., that $\alpha = \beta$. See Fig. 21.

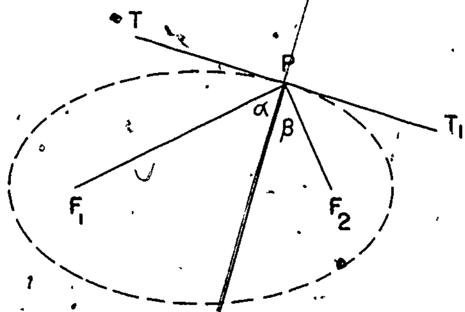

Fig. 21.

Are not the principal ingredients of this figure familiar? Could it not be interpreted as illustrating the law of reflection? We now use science to do mathematics. The phrase "to throw light," hitherto construed as a figure of speech, is now to be taken literally. Come to think of it, what more perfect exemplification of a mathematical straight line than a ray of light is there? We suppose a ray of light F_1P to be reflected at P from a mirror TT'. If the reflected ray does in fact pass through F_2, then TT' (the tangential line to the ellipse at P) is the normal at P to the bisector of $\angle F_1PF_2$, and we have solved our problem.

Does the reflected ray pass through F_2? We recall that reflection is a consequence of the shortest path principle. Thus, if Q is the point on the mirror from which an incident ray F_1Q is reflected through F_2, then F_1QF_2 must be the shortest path possible (via the mirror) from F_1 to F_2. It remains to show that the shortest path is such that Q is coincident with P, the point at which TT' is tangential to the ellipse.

Consider Fig. 22.

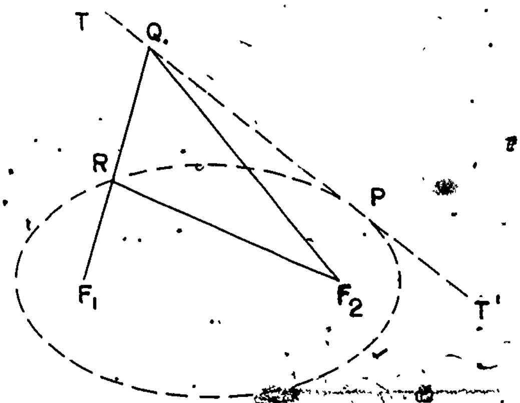

Fig. 22.

It is evident that any point Q (on TT') not coincident with P must lie outside the ellipse; therefore suppose F_1Q to cut the ellipse at R. Since RF_2 is the shortest path from R to F_2,

$$RQ + QF_2 > RF_2 .$$

Consequently, adding F_1R to both sides of the inequality,

$$(F_1R + RQ) + QF_2 > F_1R + RF_2$$

i.e.,

$$F_1Q + QF_2 > F_1R + RF_2 .$$

But R is on the ellipse, so that by definition

$$F_1R + RF_2 = 2a .$$

Therefore,

$$F_1Q + QF_2 > 2a$$

whereas, P being on the ellipse,

$$F_1P + PF_2 = 2a .$$

Since light takes the shortest path, it follows that Q must be coincident with P. That is, a ray of light from one focus, incident to a mirror

ERIC

(tangential to the ellipse) at its point of contact, is reflected through the other. This completes the proof of our conjectured construction of a tangent to the ellipse.

Fermat was the man who first raised and substantially answered the wider question of how to find tangents to plane curves in general. To solve this problem for any curve whose function is an algebraic polynomial he invented the differential calculus. Yet it is refreshing with the present density of calculus textbooks to find that a construction for the ellipse can be established without resort to differentiation. The solution by optics, given above, was the earliest.

That in Fig. 21 $\alpha = \beta$ has several practical applications. The key to these applications is that for a silvered ellipse the immediate (elliptical) neighborhood of P will reflect light as if it were the surface at P of the mirror tangential to the ellipse at that point. Consequently, no matter what its direction, a ray passing through one focus will be reflected at the ellipse through the other. The heat of a fire at F_1 although radiated in all directions will be reconcentrated at F_2. If no radiation is dissipated _en route_ and none lost in contact with the ellipses silvered surface, F_2 is as hot as F_1. A reflecting ellipse with a fire at one focal point has a fire at both; _Focus_ is the Latin for fireplace or hearth. Similarly for an auditorium with an ellipsoidal cupola, F_1, F_2 are known as the whispering points. Since sound is reflected in the same way as light, a dispersed and therefore weakened whisper from F_1 will be inaudible in all other parts of the room except at F_2 where the whisper is reconcentrated. See Fig. 23.

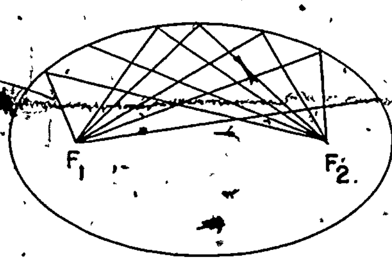

Fig. 23.

It is often instructive to go to the limit. We found it profitable to consider the limiting or degenerate case of the ellipse where F_1 and F_2 become coincident; we now go to the other extreme and suppose them to be as far apart as possible. With F_1 fixed, the farther F_2 is moved from it, the more elongated the ellipse and the more nearly parallel PF_2 to the axis VF_1. See Fig. 24.

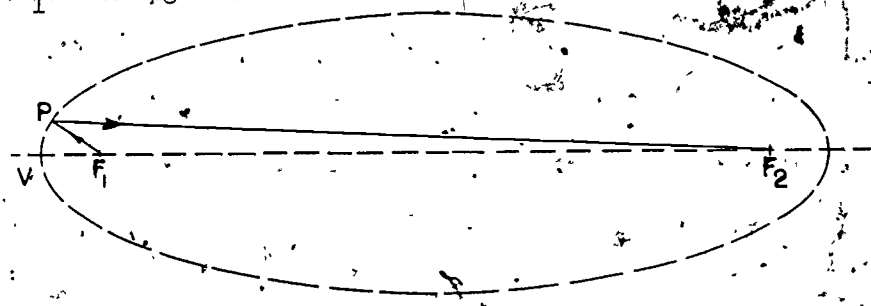

Fig. 24.

Finally, with F_2 at infinity, the ellipse has degenerated into what is known as the parabola and PF_2 has become parallel to the axis. See Fig. 25.

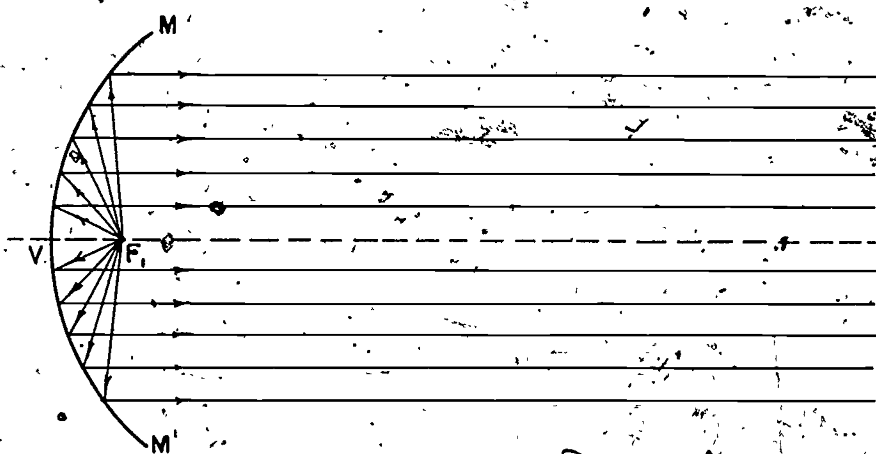

Fig. 25.

Thus, given a point source of light at F_1, the reflection from a silvered parabola MVM' is a beam parallel to the axis of the parabola VF_1. Rotation of the parabolic mirror about its axis generates what is known as a paraboloid of revolution. This, of course, reflects a solid beam of light from a point source at F_1 parallel to its axis, and is exemplified by the motorcar head-lamp. And conversely, since the rays radiated from a distance source are almost parallel, they are accumulable within the immediate neighborhood of F_1. A paraboloidal reflection could with equal justice be termed a paraboloidal accumulator. Radio rays, individually weak, can be collectively magnified into a strong signal. As well as essential to radar listening devices, the paraboloidal reflector is the basis of the radio telescope.

Chapter 4. Applications of Matrix Algebra.

Although my purpose in this lecture is to show what matrices are good for, not to teach matrix theory as such, I shall assume merely that you recognize a matrix when you see one and can readily perform matrix, row times column multiplication for very simple matrices.

My aim is two-fold, my lecture has two parts. In Part 1, Mathematics with Matrices, my principal objective is to convince you that matrices are much more than a kind of mathematical noughts-and-crosses designed to delight examiners and depress examinces, that matrix technique really does facilitate doing mathematics. In Part 2, From Matrix Theorem to Relativity Physics, I aim to show how this facility, used with bold imagination, devastates comfortable, commonplace conceptions of our physical world.

Part 1. Mathematics with Matrices

4.1 Why Use Matrices?

Have you ever tried using a lump of rock to drive a six-inch nail into a four-inch beam? It is easier with a hammer. Easier because the hammer is designed expressly for the job, designed to have good balance, to handle well, to effect a neater job with less effort. Its design, deceptively simple, is dependent upon giving much thought to questions of rigidity, distribution of weight, and center of percussion. Hard thinking goes into its design; hard work is simplified by its use.

Matrices, too, are deceptively simple. Some clever fellows gave much thought to devising a notation that handles well and a technique that does a tidier, more effortless job. Yes, matrices take the slog out of nailing equations. And, as with driving nails, there is no need to take anyone's word for it; experience is conclusive. Presently, you will have the experience. Not to start our mathematical carpentry with rusty nails, we first review.

4.2 Rotation of Rectangular Axes.

Given a plane, we introduce a rectangular coordinate system x,y with origin 0. If from any arbitrary point $P(x,y)$ we drop a perpendicular to the x-axis, then the length of this perpendicular is y (the ordinate of P) and the distance from its foot to 0 is x (the abscissa of P). A diagram makes it clear that to any given point P there corresponds just one pair of coordinates (x,y) and, conversely, that to any given (x,y) there corresponds just one point P. See Fig. 1.

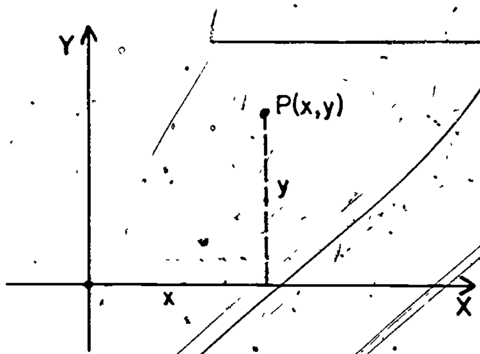

Fig. 1

Next we introduce a new rectangular coordinate system $\bar{x},\bar{y}$ with the same origin 0. Although the position of P remains unchanged, relative to the new coordinate system, it has new coordinates $(\bar{x},\bar{y})$. See Fig. 2.

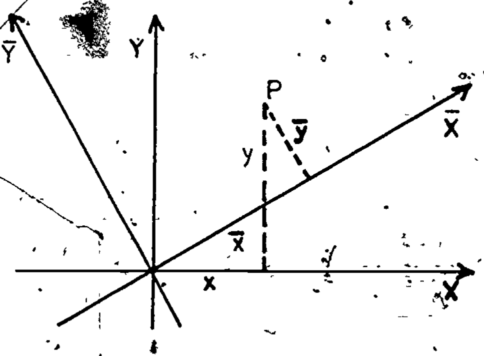

Fig. 2

This situation very naturally raises the question, what is the relation between the old and new coordinates of P? If we are given P's old coordinates (x,y) then P is fixed and so its new coordinates $(\bar{x},\bar{y})$ must, in principle at least, be determinate. Conversely, given P's new coordinates $(\bar{x},\bar{y})$, what are its old coordinates (x,y)? What is the transformation from the one coordinate system to the other?

Easy reasoning shows that the rule for going from the coordinates in the one system to those in the other, say from (x,y) to $(\bar{x},\bar{y})$, must be a linear transformation. That is to say, the rule must be a system of linear equations in $\bar{x},\bar{y}$ and x,y of the form

$$\bar{x} = Ax + By$$
$$\bar{y} = Cx + Dy$$

where A, B, C, and D are numbers independent of $\bar{x},\bar{y}$ and x,y.

Why is this? Because this is the only sort of system which can be inverted. Since P has unique coordinates in both systems the formulae must also determine a unique value for x and for y when $\bar{x}$ and $\bar{y}$ are given. If A, B, C, and D were not independent of x and y, then the formulae would contain quadratic or more complicated terms in x or y, so that given values of $\bar{x}$ and $\bar{y}$ would not necessarily determine x and y uniquely.

If the reader will use these explicit formulae for $\bar{x}$ and $\bar{y}$ to derive explicit formulae for x and y, he will get formulae of the pattern

$$x = \bar{A}\bar{x} + \bar{B}\bar{y}$$
$$y = \bar{C}\bar{x} + \bar{D}\bar{y}$$

where $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$ are expressed in terms of A, B, C, and D.

You know well how to solve a system of two simultaneous linear equations in the two unknowns x and y. You will obtain the above two equations if you express x and y in terms of $\bar{x}$ and $\bar{y}$. It is easy to express the new constants $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$ by means of A, B, C, and D. It might be a good exercise for you to carry out the calculation. We do not do it here because we do not need these formulae.

Of course, the next question is: What are A, B, C, and D? Since they are independent of the coordinates of P, they may be found by specializing. A well-chosen point will reduce the labor of calculation.

Let us take the point on the x-axis at unit distance from the origin. This point, call it P_0, therefore has coordinates (1, 0) in the old system. See Fig. 3.

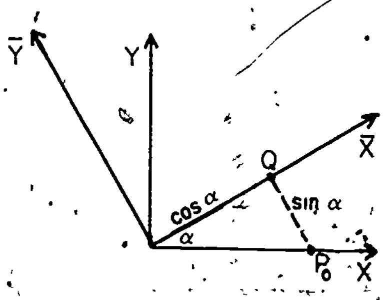

Fig. 3

What are P_0's coordinates in the new system? Its abscissa $\bar{x}$ is, of course, the distance from 0 to Q, where Q is the foot of the perpendicular dropped from P_0 to the $\bar{x}$-axis. So, taking our first equation

$$\bar{x} = Ax + By$$

with $x = 1$, $y = 0$,

$$OQ = A \cdot 1 + B \cdot 0 = A.$$

To go farther we need to know the angle made by the new axis with the old. Let this be α. From the obvious geometry of Fig. 3, since $OP_0 = 1$,

$$OQ = \cos \alpha.$$

Consequently,

$$A = \cos \alpha.$$

Next, what is the ordinate $\bar{y}$ of P_0? Since P_0 lies below the $\bar{x}$-axis, it is the negative of the perpendicular P_0Q, that is,

$$\bar{y} = -\sin \alpha.$$

So, taking our second equation

$$\bar{y} = Cx + Dy$$

with $x = 1$, $y = 0$,

$$-\sin\alpha = C \cdot 1 + D \cdot 0 = C.$$

Of the four numbers A, B, C, and D in the transformation, we have already found two, A and C. It remains to find B and D. And isn't it obvious what to do? The y-axis is just as good as the x-axis; it should be given equal consideration. So, having taken a point one unit along the x-axis, we now take a point one unit along the y-axis. Let P_1 be the point with coordinates (0,1) in the old system. See Fig. 4.

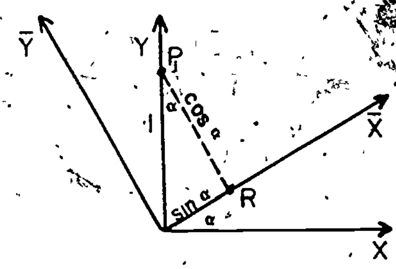

Fig. 4

What are P_1's coordinates in the new system? Its abscissa $\bar{x}$ is, of course, the distance from O to R, where R is the foot of the perpendicular dropped from P_1 to the $\bar{x}$-axis. So, substituting $\bar{x} = OR$ and $x = 0$, $y = 1$, in the first equation of the transformation

$$OR = A \cdot 0 + B \cdot 1 = B.$$

And since $\angle OP_1R$ and the angle α between OX and $O\bar{X}$ are both complements of $\angle P_1O\bar{X}$, $\angle OP_1R = \alpha$. Consequently, with $OP_1 = 1$

$$OR = \sin\alpha$$

so that

$$B = \sin\alpha.$$

Similarly, since $\bar{y} = \cos\alpha$, we obtain from the second equation of the transformation that

$$D = \cos \alpha.$$

We have found A, B, C, and D. We conclude that the required transformation is:

$$\left.\begin{aligned} \bar{x} &= (\cos \alpha)x + (\sin \alpha)y \\ \bar{y} &= (-\sin \alpha)x + (\cos \alpha)y. \end{aligned}\right\} \qquad (1)$$

Our refresher course is completed; we have scraped the rust off our nails.

4.3 Transformation with and without Matrices.

We now use matrix algebra to write (1) in a slightly simplified form. We get

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \qquad (1')$$

for multiplication of the second matrix on the right side by the first on the right side gives a column matrix whose terms are

$$(\cos \alpha)x + (\sin \alpha)y$$

and

$$(-\sin \alpha)x + (\cos \alpha)y$$

which equal $\bar{x}$ and $\bar{y}$, respectively.

The first element of this column matrix is obtained, as the reader doubtlessly recalls, by multiplying x (the first element of the column of the second matrix) by $\cos \alpha$ (the first element of the first row of the first matrix), by multiplying y (the second element of the column of the second matrix) by $\sin \alpha$ (the second element of the first row of the first matrix), and by adding these products together. Yes, it's easier to do than to state. The second element is similarly obtained by using the elements of the second row of the first matrix instead of its first row elements. In brief, multiply columns by rows.

ERIC

Do we gain anything? (1*) does look a little more compact than (1). Is a bracket less laborious to write than an equality sign? Really, this is splitting hairs. What matters is not so much how a tool looks, but how it handles. How well does it handle? Let us do some mathematical carpentry to find out; let us deduce -- I beg your pardon, let us nail home -- the basic trigonometric angle sum formulae for $\cos(\alpha + \beta)$ and $\sin(\alpha + \beta)$. For comparison let us do the job twice; once driving our nails with a hammer, i.e., using matrix transformations such as (1*), and once hitting our nails with a stone, i.e., using non-matrix transformations such as (1).

Suppose that the x,y rectangular coordinate system is first rotated through an angle α to give a second $\bar{x},\bar{y}$ system and then through an additional angle β to give a third $\bar{\bar{x}},\bar{\bar{y}}$ system. See Fig. 5.

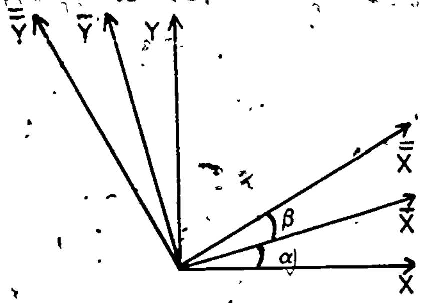

Fig. 5

Regarding the second system as the old and the third system as the new, since the latter makes an angle β with the former, by virtue of (1), i.e., substituting β for α, $\bar{x}$ for x, $\bar{y}$ for y, and $\bar{\bar{x}}$ for x, $\bar{\bar{y}}$ for y, we have

$$\left.\begin{aligned} \bar{\bar{x}} &= (\cos\beta)\bar{x} + (\sin\beta)\bar{y} \\ \bar{\bar{y}} &= (-\sin\beta)\bar{x} + (\cos\beta)\bar{y} \end{aligned}\right\} \quad (2)$$

and, by virtue of (1*)

$$\begin{pmatrix} \bar{\bar{x}} \\ \bar{\bar{y}} \end{pmatrix} = \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}. \quad (2^*)$$

Next, taking the first system to be the old and the third system to be the new, since the latter makes an angle $\alpha + \beta$ with the former, by virtue of (1), i.e., substituting $\alpha + \beta$ for α; $\bar{\bar{x}}$ for x, $\bar{\bar{y}}$ for y, x for $\bar{x}$, and $\bar{y}$ for y, we have

$$\left.\begin{aligned} \bar{\bar{x}} &= (\cos \overline{\alpha + \beta})x + (\sin \overline{\alpha + \beta})y \\ \bar{\bar{y}} &= (-\sin \overline{\alpha + \beta})x + (\cos \overline{\alpha + \beta})y. \end{aligned}\right\} \qquad (3)$$

Similarly, by virtue of (1*)

$$\begin{pmatrix} \bar{\bar{x}} \\ \bar{\bar{y}} \end{pmatrix} = \begin{pmatrix} \cos \overline{\alpha + \beta} & \sin \overline{\alpha + \beta} \\ -\sin \overline{\alpha + \beta} & \cos \overline{\alpha + \beta} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \qquad (3^*)$$

Note that this far the differences are merely notational, but this is the parting of the ways. From here on the non-matrix method is more laborious.

First, we continue with the matrix method. Using (1*) to eliminate $\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$ from (2*), we have

$$\begin{pmatrix} \bar{\bar{x}} \\ \bar{\bar{y}} \end{pmatrix} = \begin{pmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{pmatrix} \left\{ \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \sin \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right\}. \qquad (4^*)$$

Hence, from (3*) and (4*), we have, in view of the associative law of matrix multiplication,

$$\begin{pmatrix} \cos \overline{\alpha + \beta} & \sin \overline{\alpha + \beta} \\ -\sin \overline{\alpha + \beta} & \cos \overline{\alpha + \beta} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{pmatrix} \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \qquad (5^*)$$

so that

$$\begin{pmatrix} \cos \overline{\alpha + \beta} & \sin \overline{\alpha + \beta} \\ -\sin \overline{\alpha + \beta} & \cos \overline{\alpha + \beta} \end{pmatrix} = \begin{pmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{pmatrix} \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}. \qquad (5^{**})$$

To determine $\cos \overline{\alpha + \beta}$ it remains merely to multiply the first row of the first matrix (on the right side) into the first column of the second. We get

$$\cos \overline{\alpha + \beta} = \cos \beta \cos \alpha + \sin \beta (-\sin \alpha)$$

which, to show due respect to the alphabet, we write

$$\cos \overline{\alpha + \beta} = \cos \alpha \cos \beta - \sin \alpha \sin \beta.$$

Similarly, multiplying the first row into the second column, we get

$$\sin \overline{\alpha + \beta} = \cos \beta \sin \alpha + \sin \beta \cos \alpha,$$

i.e.,

$$\sin \overline{\alpha + \beta} = \sin \alpha \cos \beta + \cos \alpha \sin \beta.$$

Next, we continue with the non-matrix method. Using (1) to eliminate $\bar{x}$ and $\bar{y}$ from (2) is a much more strenuous affair than using (1*) to eliminate $\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix}$ from (2*). How much more strenuous you can find out only by doing the algebra for yourself; your mental muscles will not tire by watching me work. I'll wait.

Your labors correctly completed, we both have

$$\left.\begin{aligned}\bar{\bar{x}} &= (\cos \alpha \cos \beta - \sin \alpha \sin \beta)x + (\sin \alpha \cos \beta + \cos \alpha \sin \beta)y \\ \bar{\bar{y}} &= -(\sin \alpha \cos \beta + \cos \alpha \sin \beta)x + (\cos \alpha \cos \beta - \sin \alpha \sin \beta)y\end{aligned}\right\} \quad (4)$$

Hence, from (3) and (4), we have

$$\left.\begin{aligned}&(\cos \overline{\alpha + \beta})x + (\sin \overline{\alpha + \beta})y \\ &\qquad = (\cos \alpha \cos \beta - \sin \alpha \sin \beta)x + (\sin \alpha \cos \beta + \cos \alpha \sin \beta)y \\ &(-\sin \overline{\alpha + \beta})x + (\cos \overline{\alpha + \beta})y \\ &\qquad = -(\sin \alpha \cos \beta + \cos \alpha \sin \beta)x + (\cos \alpha \cos \beta - \sin \alpha \sin \beta)y\end{aligned}\right\} \quad (5)$$

And since these equations hold for arbitrary x and y, taking $x = 1$, $y = 0$, the first gives us immediately the formula for $\cos \overline{\alpha + \beta}$, the second the formula for $\sin \overline{\alpha + \beta}$.

How much extra work does the non-matrix method entail? Quite a lot; we have both done it, we know. But let us see precisely what this extra work is. (5) written directly in matrix notation is

$$\begin{pmatrix}\cos \overline{\alpha + \beta} & \sin \overline{\alpha + \beta} \\ -\sin \overline{\alpha + \beta} & \cos \overline{\alpha + \beta}\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}$$

$$= \begin{pmatrix}\cos \alpha \cos \beta - \sin \alpha \sin \beta & \sin \alpha \cos \beta + \cos \alpha \sin \beta \\ -(\sin \alpha \cos \beta + \cos \alpha \sin \beta) & \cos \alpha \cos \beta - \sin \alpha \sin \beta\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}. \quad (5')$$

Notice anything remarkable? Well, compare (5') with (5*). We must conclude that

$$\begin{pmatrix} \cos\alpha\cos\beta - \sin\alpha\sin\beta & \sin\alpha\cos\beta + \cos\alpha\sin\beta \\ -(\sin\alpha\cos\beta + \cos\alpha\sin\beta) & \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{pmatrix}$$

$$= \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}. \qquad (5'')$$

But, (4) written directly in matrix notation is

$$\begin{pmatrix} \bar{\bar{x}} \\ \bar{\bar{y}} \end{pmatrix} = \begin{pmatrix} \cos\alpha\cos\beta - \sin\alpha\sin\beta & \sin\alpha\cos\beta + \cos\alpha\sin\beta \\ -(\sin\alpha\cos\beta + \cos\alpha\sin\beta) & \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Thus (4) is, in effect, (4*) <u>with the first pair of matrices on its right multiplied out</u>.

What do you conclude? Think about it. Whereas by using matrices, we obtain (5*) without having to multiply out the product

$$\begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix},$$

to obtain (5) without using matrices <u>necessitates multiplying out the non-matrix equivalent of this product</u>. Put paradoxically, whereas the use of matrices avoids computation of the matrix product, the avoidance of matrices necessitates it. Isn't it easier to drive nails with a hammer than with a stone?

4.4 <u>Orthogonal Matrices</u>.

Consider (5**). This is a curious equation. The multiplication of two matrices of the form

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

gives another matrix of the same form; they constitute a group. The performance of the operation $\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$ on $\begin{pmatrix} x \\ y \end{pmatrix}$ (see (1*)), followed by the performance of a second and similar operation, namely, $\begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix}$ on the result of the first operation (see (4*)), is equivalent to the performance of the single operation $\begin{pmatrix} \cos \overline{\alpha + \beta} & \sin \overline{\alpha + \beta} \\ -\sin \overline{\alpha + \beta} & \cos \overline{\alpha + \beta} \end{pmatrix}$ on $\begin{pmatrix} x \\ y \end{pmatrix}$ (see (3*)). These operations give the transformations of the coordinates of P for rotations of axes through an angle α, an angle β following an angle α, and an angle $\alpha + \beta$. But, of course, a rotation through α followed by a rotation through β has the same outcome as a single rotation through $\alpha + \beta$. This is the reasoning underlying the deduction of (5*) from (3*) and (4*). Isn't it obvious from the geometrical point of view that rotation transforms must constitute a group?

Of course, the transform for a rotation θ will have θ for an ingredient, but why $\cos \theta$ and $\sin \theta$? Of course, α, β, and $\alpha + \beta$ will be ingredients of our transforms, but why their cosines and sines?

Reconsider the derivation of (1*). Take another look at Fig. 3. If OQ is to be the $\bar{x}$ value of P_0, by definition P_0Q must be parallel to $O\bar{Y}$. Yet if $O\bar{X}$, $O\bar{Y}$ were not perpendicular, the angle at Q would not be a right angle, so that OQ would not be equal to $\cos \alpha$ and P_0Q would not be $\sin \alpha$. Thus, we come to see that

$$\begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}$$

is necessarily the pattern of matrices with which we can handle transformations of coordinates induced by rotations of rectangular axes. And since mathematicians are disposed to use the word <u>orthogonal</u> rather than <u>rectangular</u> or <u>right-angled</u>, matrices of this pattern are said to be <u>orthogonal matrices</u>.

Right angles are very special angles; right-angled axes very special axes; we must expect orthogonal matrices to have very special properties. They do. Look at the pattern again. The first row is such that

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

the second, such that

$$(-\sin \theta)^2 + (\cos \theta)^2 = 1.$$

While adding the product of the elements in each column, we have

$$\cos\theta\cdot(-\sin\theta) + \sin\theta\cdot\cos\theta = 0.$$

These properties are characteristic; if a matrix has them it is orthogonal; if it doesn't, it isn't. Formally, a matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

is said to be orthogonal if and only if

$$A^2 + B^2 = 1$$
$$C^2 + D^2 = 1$$
$$AC + BD = 0.$$

4.5 A Most Important Theorem.

Given that

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

is subject to the very special condition that $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ is orthogonal, ought not we anticipate some very special relation between the new and the old coordinates of P? Look at the question geometrically. See Fig. 6.

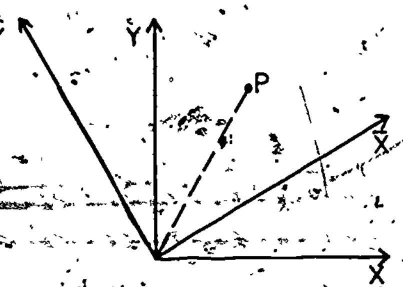

Fig. 6.

The origin O and the point P remain fixed no matter what the rotation of the orthogonal axes, so that the distance OP remains unchanged. But if

OP remains unchanged, so does its square. And we all remember one of the very first formulae we learned in coordinate geometry, namely, that the square of the distance of a point from the origin is the sum of the squares of its coordinates. Calculating in the old x,y coordinate system,

$$(OP)^2 = x^2 + y^2$$

and in the new $\bar{x},\bar{y}$ system,

$$(OP)^2 = \bar{x}^2 + \bar{y}^2$$

so that

$$\bar{x}^2 + \bar{y}^2 = x^2 + y^2.$$

We have the result:

Given that

$$\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix} = \begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix},$$

if $\begin{pmatrix}A & B\\ C & D\end{pmatrix}$ is orthogonal, then $\bar{x}^2 + \bar{y}^2 = x^2 + y^2$.

Can we say more? Well, we can at least suspect more. Suppose that the pairs of axes to be oblique instead of orthogonal. It is no longer true, in general, that

$$(OP)^2 = x^2 + y^2$$

or that

$$(OP)^2 = \bar{x}^2 + \bar{y}^2.$$

Of course, it could conceivably still be true that $\bar{x}^2 + \bar{y}^2 = x^2 + y^2$, but isn't this most unlikely? We conjecture that $\bar{x}^2 + \bar{y}^2$ cannot (for arbitrary x and y) be equal to $x^2 + y^2$ unless $\begin{pmatrix}A & B\\ C & D\end{pmatrix}$ is indeed orthogonal.

(The reader who has used oblique axes will recall that if the angle between them is ω, then in consequence of the Cosine Rule, calculating in the old x,y system

$$(OP)^2 = x^2 + y^2 + 2xy \cos \omega$$

and in the new $\bar{x},\bar{y}$ system

$$(OP)^2 = \bar{x}^2 + \bar{y}^2 + 2\bar{x}\bar{y}\cos\omega.$$

Thus $\bar{x}^2 + \bar{y}^2 = x^2 + y^2$ if and only if $xy = \bar{x}\bar{y}$, so that our suspicion is seen to be well founded.)

Combining fact with fancy we anticipate:

Given that

$$\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix} = \begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix},$$

$\bar{x}^2 + \bar{y}^2 = x^2 + y^2$ if and only if $\begin{pmatrix}A & B\\ C & D\end{pmatrix}$ is orthogonal. We have committed ourselves to an opinion. Until we know whether we are right or wrong, how can we decently rest?

We rewrite the given matrix equation thus

$$\bar{x} = Ax + By$$

$$\bar{y} = Cx + Dy.$$

Squaring both equations and adding, we get

$$\bar{x}^2 + \bar{y}^2 = (A^2x^2 + 2ABxy + B^2y^2) + (C^2x^2 + 2CDxy + D^2y^2)$$

$$= (A^2 + C^2)x^2 + (B^2 + D^2)y^2 + 2(AB + CD)xy. \qquad (6)$$

If $\begin{pmatrix}A & B\\ C & D\end{pmatrix}$ is orthogonal, by definition, $A^2 + B^2 = 1$, $C^2 + D^2 = 1$, and $AB + CD = 1$, whereupon (6) gives

$$\bar{x}^2 + \bar{y}^2 = x^2 + y^2.$$

Not surprising, but we would have been surprised at the contrary. The substitution does serve as some sort of check on our algebra, doesn't it?

If for arbitrary x,y

$$\bar{x}^2 + \bar{y}^2 = x^2 + y^2$$

(6) gives

$$1 \cdot x^2 + 1 \cdot y^2 + 0 \cdot xy = (A^2 + C^2)x + (B^2 + D^2)y + 2(AC + BD)xy.$$

Equating coefficients of this identity

for x^2,

$$1 = A^2 + C^2$$

for y^2,

$$1 = B^2 + D^2$$

and for xy,

$$0 = 2(AB + CD)$$

so that,

$$0 = AB + CD.$$

The matrix is orthogonal; our conjecture is a theorem.

4.6 A Matter of Notation.

We have seen that some orthogonal matrices are characterised by the pattern

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

To compute the elements of this matrix given $\cos\theta$, it is, of course, natural to use trigonometric tables. If, for example, $\cos\theta = 0.5172$, we use the cosine table to find θ, the angle where cosine is 0.5172, and then the sine table to find the sine of this angle. But tables are not always at hand. How can we get along without them? Yes, by using

$$\sin\theta = \sqrt{1 - \cos^2\theta}.$$

With $\cos\theta = 0.5172$, we have, without using tables,

$$\sin\theta = \sqrt{1 - 0.5172^2}.$$

To indicate our dispensation from the need to use tables, we put $z = \cos\theta$, in consequence of which $\sin\theta = \sqrt{1 - \cos^2\theta} = \sqrt{1 - z^2}$, and write the above typical orthogonal matrix with the notation

$$\begin{pmatrix} z & \sqrt{1 - z^2} \\ -\sqrt{1 - z^2} & z \end{pmatrix}$$

We have said the same thing, yet with a different emphasis.

Has, now, every orthogonal matrix such a representation by means of one variable z? Given an arbitrary orthogonal matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$, let us replace the letter A by z. Since $A^2 + B^2 = 1$, we see that $B = \sqrt{1 - z^2}$ where the square root may be taken either positive or negative. Next, the condition $AC + BD = 0$ now reads $zC + \sqrt{1 - z^2}D = 0$. Hence, we find $\frac{C}{D} = -\frac{1}{z}\sqrt{1 - z^2}$; this fact can now be stated as follows. There exists a constant α such that $C = \alpha\sqrt{1 - z^2}$, $D = -\alpha z$. Since, finally, $C^2 + D^2 = 1$, we conclude $\alpha^2(1 - z^2 + z^2) = 1$ and $\alpha^2 = 1$. This allows us two choices for α, namely, $\alpha = +1$ and $\alpha = -1$. Hence, the most general orthogonal matrix has the form

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} z & \sqrt{1 - z^2} \\ -\sqrt{1 - z^2} & z \end{pmatrix} \text{ or } = \begin{pmatrix} z & \sqrt{1 - z^2} \\ \sqrt{1 - z^2} & -z \end{pmatrix}.$$

Only the first kind of orthogonal matrices occur under rotations of the coordinate axes.

What is the meaning of the second kind of orthogonal matrices? Let us specialize and choose the convenient value $z = 1$. Thus, we are led to the matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ which is of the exceptional form. As a matter of fact, the knowledge of this one particular orthogonal matrix allows us to bridge over from all orthogonal matrices of the first kind to all orthogonal matrices of the second kind. Indeed, the rules of matrix multiplication yield the identity

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} z & \sqrt{1 - z^2} \\ -\sqrt{1 - z^2} & z \end{pmatrix} = \begin{pmatrix} z & \sqrt{1 - z^2} \\ \sqrt{1 - z^2} & -z \end{pmatrix}$$

as you may verify as an exercise in matrix multiplication. Thus, each orthogonal matrix of the first kind becomes an orthogonal matrix of the second kind by multiplication with this particular matrix.

Let us now interpret the meaning of the transformation

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

which takes, in non-matrix form, the following shape:

$$\bar{x} = x, \qquad \bar{y} = -y.$$

You can easily verify that this transformation takes place if we keep our coordinate axes but direct the positive y direction in the opposite sense. In other words, we reflect the y-axis on the x-axis as if the x-axis were a mirror. Clearly, under such a coordinate transformation the distance from the origin is also preserved. The transformation considered is called a reflection. We have thus proved that every orthogonal matrix can be expressed as a matrix belonging to a rotation or as a matrix which is the product of an arbitrary rotation and the special reflection matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

In general, the mathematician bewares of changes of coordinate systems which involve a reflection. One is accustomed to drawing the y-axis and x-axis in such a position that the first axis is obtained from the second by a rotation in counterclockwise sense. This is the so-called positive sense of rotation. If we make a transition to a new coordinate system by reflection, we change the orientation of the coordinate axes. They go now over into each other by rotating the x-axis in the clockwise (negative) sense.

We can assemble the insight obtained in this section in the following theorem:

<u>Theorem A</u>: Given that

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

is a transformation which preserves the orientation of the coordinate areas and preserves the distance from the origin $x^2 + y^2 = \bar{x}^2 + \bar{y}^2$. This holds if and only if

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} z & \sqrt{1 - z^2} \\ -\sqrt{1 - z^2} & z \end{pmatrix}.$$

In Consequence of Laziness.

We are now in a position to see how a man who knows his matrices can deduce (1).

He argues:

Because there is a one-to-one correspondence between $(\bar{x},\bar{y})$ and (x,y) the required transformation must be linear, i.e., of the form

$$\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix} = \begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}.$$

And since the orientation of the coordinate axes is not changed and the distance OP remains invariant,

$$\bar{x}^2 + \bar{y}^2 = x^2 + y^2.$$

We have in consequence of Theorem A

$$\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix} = \begin{pmatrix}z & \sqrt{1 - z^2}\\ -\sqrt{1 - z^2} & z\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}$$

i.e.,

$$\bar{x} = zx + \sqrt{1 - z^2}\cdot y$$
$$\bar{y} = -\sqrt{1 - z^2}\cdot x + z\cdot y.$$

To assign geometrical significance to z, it is convenient to take $x = 1$, $y = 0$, whereupon

$$\bar{x} = z$$
$$\bar{y} = -\sqrt{1 - z^2}.$$

From Fig. 3 it is obvious that $\bar{x} = \cos\alpha$, $\bar{y} = -\sin\alpha$ (the minus sign because QP_0 has the opposite sense to $O\bar{Y}$), so that

$$z = \cos\alpha$$
$$\sqrt{1 - z^2} = \sin\alpha.$$

Thus, (1) follows immediately.

Effortless? This is hitting a nail with a power-driven hammer. For us muscle-driven hammers are a thing of the past. And bashing away with stones?

Oh, that sort of thing belongs to cave-man mathematics.

Although (1) could have been deduced even more succintly, I have preferred the present argument because it is echoed -- albeit faintly -- in a subsequent argument.

4.8 To Sum Up.

By considering orthogonal transformations in some detail, I have tried to typify what matrices are good for in mathematics. And by emphasizing the rough analogy between nailing with and without hammers and doing algebra with and without matrices, I have tried to make it easier to appreciate the facility afforded by matrix technique. But do not mistake analogy for mathematical appreciation; in particular, there is no substitute for working out and pondering over the mathematics of Section 4.3 for yourself.

Part 2. From Matrix Theorem to Relativity Physics.

Here my main objective, you will recall, is to show how relativity theory arises out of matrix algebra used with bold imagination. This part is more difficult, although not more difficult mathematically. More difficult because, unlike Part 1, it demands that you -- how shall I put it? -- unthink firmly- even if uncritically-held notions.

The basic relativity problem arises out of trying to state with mathematical precision what we can mean when we use the phrase "at the same time" or say that two events were simultaneous. And what on earth has this to do with matrix algebra? A good question, a very good question, but let us not get ahead of ourselves. It is better first to appreciate how the problem arose; the strangest motive for the reception of new notions is the failure of old ones.

4.9 The Michelson-Morley Experiment.

A. A. Michelson (1852-1931), awarded the 1907 Nobel Prize for Physics, was one of the world's greatest experimental physicists. He is perhaps best introduced by the following anecdote: Asked by a father if his son should be encouraged to continue his studies to become a physicist, Michelson is said to have replied, "No, I advise your son not to study physics. It is a dead subject. What there is to know, we know-- except that possibly we could measure a few things to the sixth decimal place instead of the fourth."

The irony of the story is that Michelson is the man whose experiments led to such a revolution that we have learned more about physics in the last sixty years than in all the preceding centuries.

But this story is revealing as well as ironical. Michelson was a man with a passion for accuracy, a man who measured everything to the sixth decimal place. He had, in particular, in the late 1870's, by most ingenious experimentation measured the velocity of light with hitherto unheard-of accuracy. The velocity of the earth in its journey round the sun having been determined with fair accuracy from astronomical data, Michelson's next ambition was to remeasure it himself -- to the sixth decimal place. With this objective in sight, in 1881, assisted by Morley, he made the experiment that was to make them famous: the Michelson-Morley experiment.

The concept upon which this experiment was based was simple. Suppose that we put a transmitter T and a receiver R a certain distance apart on the surface of the earth and measure the time taken for a signal, a flash of light, to go from T to R. See Fig. 7.

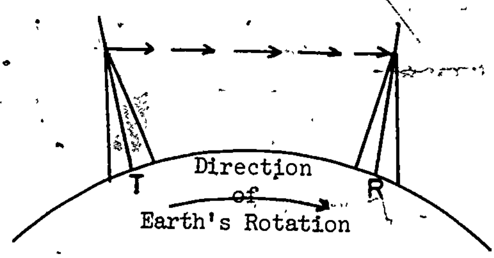

Fig. 7

The signal sent from T with the enormous velocity of light c has to overtake R which is moving ahead with velocity v, the velocity of the earth. Therefore it has velocity c - v relative to R and in consequence will take a little more time to reach R than it would if the earth were at rest. And if a return signal is sent from R back to T it is approached by T as it approaches T, so that its velocity relative to T is c + v. See Fig. 8.

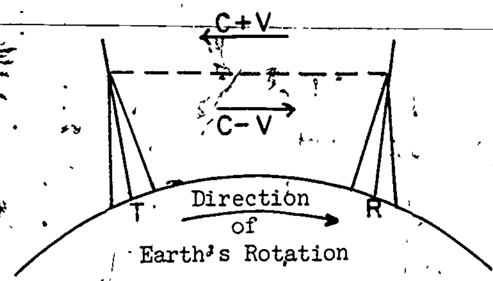

Fig. 8

Therefore the return signal will take less time than it would if the earth were at rest and still less time than the initial, outgoing signal does.

Let us describe briefly the ingenious experimental arrangement to carry out this observation. See Fig. 9.

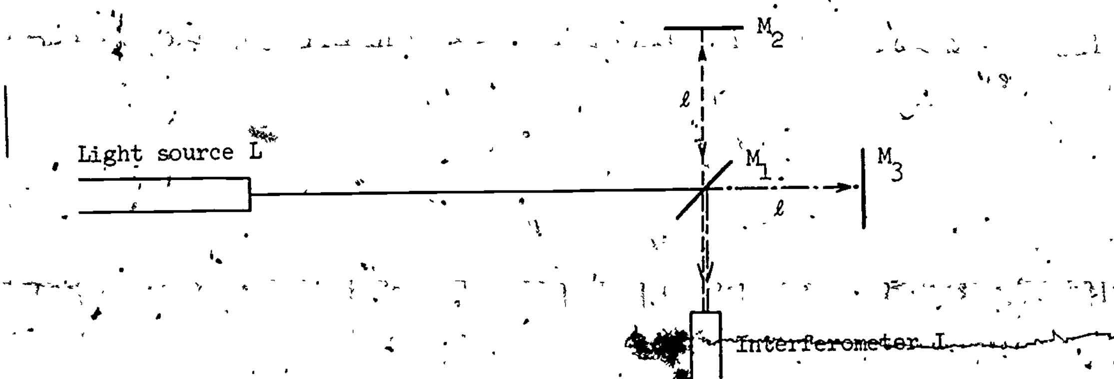

Fig. 9

We have a light source L which sends a light ray in the direction of the motion of the earth. This ray falls on a mirror M which is partially transparent and stands under an angle of 45° against the incoming ray. Thus, a part of the light is reflected under 90^o to a mirror M_2 and a part goes through to a mirror M_3. Observe that the light ray has now been split up into two rays; one moving along the line M_1M_2, that is, perpendicular to the earth's motion, and the other moving along M_1M_3 parallel to the earth's motion. Both rays are reflected again at the mirrors M_2 and M_3 and return to the transparent mirror M_1. Now, part of the vertical light ray M_2M_1 passes through M_1 and goes to the objective of an interferometer J. At the same time a part of the light ray M_3M_1 is reflected at M_1 and also enters into the same interferometer. Thus, we mix in J the light of two different travel histories. The two types of light differ in their part by the difference in time which is necessary to travel from M_1 to M_2 and back as compared to the time which it takes to travel from M_1 to M_3 and back. You do not need to know the operation of an interferometer. It is sufficient to know that such an instrument is sensitive enough to compare light rays coming from the same origin but having spent different times in travel. Being mathematicians, we shall rather calculate the expected difference in travel time which the instrument will measure.

Let ℓ be the distance between the mirrors M_1 and M_3, and M_1 and M_3 which, as you see, we assume to be equal. The travel time from M_1 to M_3 and back is evidently

$$T = \frac{\ell}{c - v} + \frac{\ell}{c + v} = \frac{2\ell c}{c^2 - v^2}$$

since in the forward motion light should travel with the lesser relative velocity $c - v$ and in return with the larger velocity $c + v$, as we discussed before.

It is more difficult to find the travel time from M_1 to M_2 and back. Let us look at the experiment from a point in outer space, so that we do not participate in the motion of the earth. At the moment when the ray leaves the mirror M_1, this mirror has the position $M_1^{(1)}$ in space, and M_2 sits at the point $M_2^{(1)}$. But, suppose it takes the time $\frac{1}{2}t$ until the ray hits the mirror M_2. During this time the mirror M_2 has shifted in the direction of the earth's

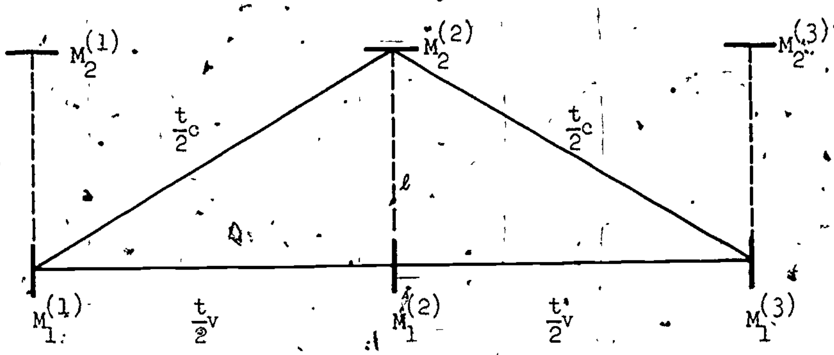

Fig. 10

motion and sits at the point $M_2^{(2)}$ in space. It reflects the light back to M_1; by reason of symmetry it will take the same time $\frac{1}{2}t$ to return from M_2 to M_1. But when the light reaches M_1, its position in space will be at $M_1^{(3)}$. We know that the distance $M_1^{(1)}$ to $M_1^{(3)}$ is given by vt since t is the total travel time and since the earth moves with the speed v. On the other hand, the light which travels with speed c had to cover in the time $\frac{1}{2}t$ the distance $M_1^{(1)}$ to $M_2^{(2)}$. Thus, in the right triangle $\Delta\left(M_1^{(1)}M_1^{(2)}M_2^{(3)}\right)$ all three sides are known as indicated in Figure 10. By the Pythagorean theorem, we have

$$\ell^2 = \frac{1}{4}t^2(c^2 - v^2), \text{ that is, } t = \frac{2\ell}{\sqrt{c^2 - v^2}}.$$

The travel time T from M_1 to M_3 is not the same as that needed to go from M_1 to M_2. The ratio of travel times is

$$\frac{T}{t} = \sqrt{1 - \frac{v^2}{c^2}}$$

This square root which occurs here in our elementary considerations is characteristic for relativity theory and will occur later in quite a different way.

The above reasoning allowed Michelson to predict a time difference in travel time and to adjust his instruments in such a way that the effect could be safely measured.

Although the idea behind this experiment is so simple, the refinements necessary to achieve the accuracy that Michelson demanded made the actual experimental set-up a hive of ingenuity. As I have said, Michelson was one of the world's greatest experimental physicists. Indeed, he achieved such accuracy that he would have been able to determine v, the earth's velocity, even if it had been moving only one tenth as fast as it does.

Michelson made the measurement and created a scandal in physics. What value for v did he get? Zero. Yes, ZERO. The flash of light takes precisely the same time to go from T to R as from R to T. But this is preposterous. Even the small boy who steals apples from an orchard appreciates the importance of relative velocity -- even if he cannot spell the words. He knows perfectly well that to escape a good hiding he must continue to run away from, not towards, the wrathful farmer hard in his pursuit. But surely there can be no difference in principle between being chased by a farmer and a flash of light? The flash is more fleet of foot, that's all.

Physicists could not believe their eyes. The Michelson-Morley experiment was repeated again and again. Again and again the answer was zero. This was against all understanding of physics. How, for goodness' sake, could the velocity of light relative to a moving object be the same when overtaking the object as when moving towards it? Despite heated discussion, the cold fact is, that the velocity of light is invariant.

4.10 What Time is it?

After a discussion of the Michelson-Morley experiment and its conceivable consequences had been prolonged in scientific journals for some twenty years, Einstein came up with a penetrating remark. "What," he asked, "do we mean by saying that two events happened at the same time. How do we know that everybody can agree what the time is at this very instant?"

In this age of jet travel it is a commonplace experience that different longitudes have different times. A telephone call from San Francisco to New York immediately confirms a different clock reading there. This communication, made by electricity at the speed of light, is so rapid that for all practical purposes the West Coast inquirer hears the East Coast answerer's reply at the same time as it is spoken in New York. At the same time the two clocks record different times, yet neither clock is wrong. Somehow or other clock readings are dependent upon an agreement about how we measure time. This remark is silly or subtle, as you please. Doesn't it sound peculiar to say that at the <u>same time</u> different clocks can <u>correctly</u> record <u>different times</u>? I am mindful of the visiting philosophy professor who, in concluding his discourse on <u>Time</u> with the remark, "So you see, gentlemen, I do not know what Time is," looked at his watch -- and dashed to catch his train.

Even if we dismiss the-different-times-at-the-same-time paradox as merely verbal, it is none the less a fact that with interplanetary travel a realistic probability the business of synchronising clocks becomes of practical importance. And cosmic voyaging introduces a complication not encountered in terrestial travel. Whereas the time lag in hearing in San Francisco what is said in New York is about $\frac{1}{50}$th of a second, the time lag in interplanetary communication (by radio waves with the velocity of light) is a matter of minutes, and that between the earth and the stars, months.

Suppose, for example, that a radio signal sent by E, an observer on earth, to A, an astronaut in outer space, 60 × 186,000 miles away, takes 1 minute. If E sends his signal when his clock records 12 o'clock, then his signal reaches A when his (E's) watch records 12:01. A, in receiving the signal sets his clock at 12:01. To confirm receipt of E's signal A immediately signals back. And since the distance between A and E remains unchanged, this return signal also takes 1 minute. Therefore E receives A's acknowledgement at 12:02 by his (E's) clock.

This, you will say, is all very simple. Surely there's no difficulty. At 12:01 E says to himself, "A is now receiving my signal sent at 12:00 by my clock and setting his clock at 12:01, the same time as mine." And from E's point of view isn't his conviction established beyond doubt by his clock reading 12:02 when he receives A's return signal?

The point is that whereas E knows that he himself sends a signal at 12:00 by his clock and knows he receives A's confirmatory signal at 12:02, E does not know that A received his (E's) signal when his (E's) clock read 12:01. Oh yes, he is convinced, but he does not know. He cannot be in both places at once to find out. He has no method of direct verification.

To synchronize clocks by means of light or radio signals, we must know the velocity with which our signals are transmitted, but to determine this velocity, we must know how long the transmission takes. To attempt to synchronize clocks without knowing the velocity of light and to determine the velocity of light without using a clock is just as futile as to try to produce hens without eggs and eggs without hens.

It is arguable that if eyewitnesses to E's signalling A are separated by great distances, then they must report vastly different, yet equally reliable, opinions of the time indicated by his (E's) clock when his signal reaches A. All very confusing. To go into great detail is to invite great confusion. Physicists went into very great detail. Many on-paper experiments were made in which people frantically set their watches as they hastily got on and off trains, trams, boats, and bicycles, scheduled for immediate departure at velocities near that of light. Scientific journals were full of these wild excursions.

Of course, it is easy to poke fun. The physicists were able, serious-minded people, trying to figure out an important problem. Their real difficulty was conceptual rather than mathematical; quite literally, they didn't know what they were talking about. Whereas we all know well enough how to use the concept of time in everyday conversation, we are at a loss when we come to map its logical geography.

4.11 The Space-Time Transformation Problem.

The matter was finally cleared up in 1905 by Einstein. He did two things of the greatest possible importance for physics: (1) He saw more clearly than any of his contemporaries what the basic problem is and gave it precise mathematical formulation; (2) He solved it. The first is by far the more difficult achievement.

This section I shall devote to (1), the next to (2).

Although Einstein (1879-1955) was an imaginative thinker with his head in the clouds, he had both feet on the ground. He did not spend several years in the Swiss Patents Office for nothing. A professor of metaphysics would have asked: "What is Time?" or, "What is the essence of Time?" or more recently, "What do we mean by Time?" Einstein, on the contrary, asked: "If a happening is observed by two persons, how are the one man's answers to the questions, 'Where?', 'When?', related is the other's? He looked for an answer in terms of measuring rods and clocks, not essences or semantics. He was a professor of physics, not metaphysics.

Although Einstein was primarily interested in observers astronomical distances apart, more homely exposition results from bringing them down to earth.

S_0 is a man who is standing at a railway track in the darkness of the night, and T is a tree by the side of the track distance x in the positive direction from S_0. See Fig. 11. The notation S_0 stands for Standing Still at the Origin of the Stationary System; "T" you can figure out for yourself.

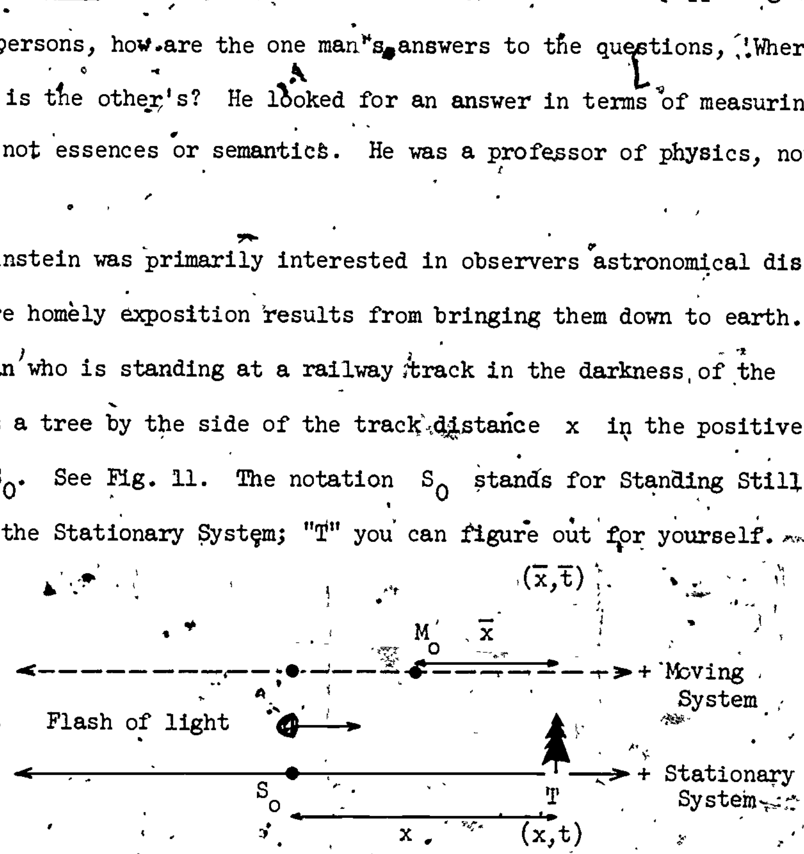

Fig. 11

ERIC Full Text Provided by ERIC

M_0 is an engine driver or motorman who drives a train along the track in the positive direction. He measures distances from where he sits in his moving cab, ahead positive, behind negative. And since he moves with his train, the distance $\bar{x}$ of the tree from him is, of course, changing as his train moves along. The notation M_0 stands for Motorman who is the Origin of the Moving System.

When M_0 in his locomotive thundering along the track passes S_0, they synchronize their watches; each sets his to zero hour. S_0 sets his $t = 0$, and M_0 sets his (a different watch, so we must use a different letter) $\bar{t} = 0$.

Also when M_0 is passing him (i.e., when $t = 0$, $\bar{t} = 0$), S_0 flashed a powerful lantern in the positive direction of the track. Almost immediately the tree is made visible to both S_0 and M_0 for an instant by the passing flash-- just as it would be by a flash of lightning. S_0 says that he caught a glimpse of T at a distance x from himself at time t; M_0 says that he glimpsed the tree at a distance $\bar{x}$ from himself at time $\bar{t}$. Whereas S_0 describes that the tree was momentarily visible as the event (x,t), M_0 describes it as the event $(\bar{x},\bar{t})$.

We put Einstein's question thus: "If a happening is observed by two men, how are the one man's answers to the questions 'Where?', 'When?' related to the other's?" It now takes on a more mathematical tone to become: What are $\bar{x}$ and $\bar{t}$ in terms of x and t? Or, mindful of matrices: What is the transformation from $\begin{pmatrix} x \\ t \end{pmatrix}$ to $\begin{pmatrix} \bar{x} \\ \bar{t} \end{pmatrix}$?

Very possibly you are tempted to say that, whereas x and $\bar{x}$ are different because M_0 is moving and S_0 is stationary, t and $\bar{t}$ must be the same. Do not be intimidated by practical concern with small scale terrestial experience.

And although not concerned with the color of the engine-driver's socks, you may be tempted at this stage to introduce v, the velocity of the train. This would be a mistake; Einstein kept the problem simple. We all know the maxim, "Put first things first"; he knew which things are the first things. His thinking was incisive.

What is relevant? Let us cast our minds back to orthogonal matrix transformations. Our problem then was to go from $\begin{pmatrix} x \\ y \end{pmatrix}$ to $\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$; our problem now is to go from $\begin{pmatrix} x \\ t \end{pmatrix}$ to $\begin{pmatrix} \bar{x} \\ \bar{t} \end{pmatrix}$. There is a similarity. The previous transformation is linear because of a one-to-one correspondence between $\bar{x}$ and x and $\bar{y}$ and y. Yet for any happening, for example, the momentary visibleness of the tree T, S_0 has just one description (x,t) and M_0 just one description $(\bar{x},\bar{t})$. To the unique description of an event by S_0 there corresponds a unique description of that event by M_0, and conversely. We must conclude that the required transformation is linear. It being understood, as previously, that the letters A, B, C, and D stand for numbers independent of x,y and $\bar{x},\bar{y}$ our present problem becomes:

Given that $\begin{pmatrix} \bar{x} \\ \bar{t} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix}$, find $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$.

Yes, find $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$. But subject to what conditions? The only thing we know from experience is the result of the Michelson-Morley experiment, that <u>the velocity of light is invariant</u>. And how are we to make use of this condition? We must inject it into the body of the problem.

Refer back to Fig. 11. First consider S_0's Stationary System. What is the relation between x and t? The flash of light is at $x = 0$ when $t = 0$. Where will it be after time t? Taking c, as is customary, to be the velocity of light,

$$x = ct$$

i.e.,

$$x - ct = 0.$$

This is supposing, of course, that the flash moves in the <u>positive</u> direction along the railway track. But this supposition is too restrictive for our purposes; it could well be that S_0 flashed his lantern in the opposite direction to a tree T'. See Fig. 12.

ERIC

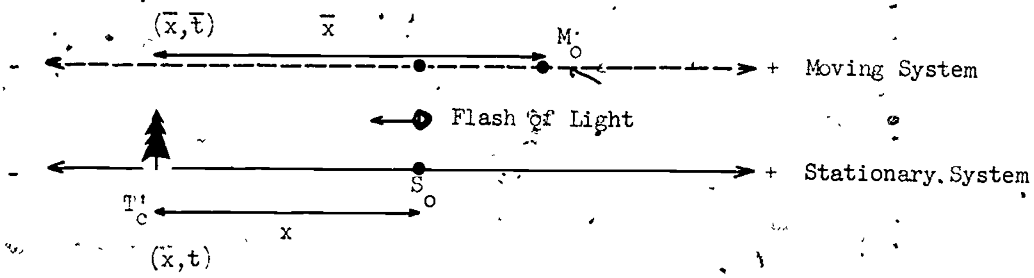

Fig. 12

If so, having due regard to signs, the velocity of the flash would be $-c$ and

$$x = -ct$$

i.e.,

$$x + ct = 0.$$

Needful of coping with either possibility, it is more convenient to handle them conjointly. By multiplying the two equations together, we have

$$(x - ct)(x + ct) = 0$$

i.e.,

$$x^2 - c^2t^2 = 0.$$

And since, if the product of two factors is 0, then at least one of them must be 0, this equation covers both the possibilities.

Next, consider M_0's Moving system. Because the velocity of light is invariant, M_0's movement makes no difference to the velocity with which a flash reaches him. In consequence, similarly, $\bar{x}$ and $\bar{t}$ are such that

$$\bar{x}^2 - c^2\bar{t}^2 = 0.$$

From the last two equations we have

$$\bar{x}^2 - c^2\bar{t}^2 = x^2 - c^2t^2.$$

This is the condition to which the required transformation is subject. Thus, the completely mathematical formulation of Einstein's space-time transformation problem is:

ERIC

Given that

$$\begin{pmatrix}\bar{x}\\ \bar{t}\end{pmatrix} = \begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}x\\ t\end{pmatrix}$$

subject to the condition that

$$\bar{x}^2 - c^2\bar{t}^2 = x^2 - c^2t^2,$$

find

$$\begin{pmatrix}A & B\\ C & D\end{pmatrix}.$$

By discovering this transformation, Einstein opened up a new world and changed our ideas of space and time.

4.12 Einstein's Solution.

Recall Theorem A:

Given that

$$\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix} = \begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}$$

preserves the orientation of the coordinate axes and the distance

$$\bar{x}^2 + \bar{y}^2 = x^2 + y^2,$$

it must necessarily have the form

$$\begin{pmatrix}A & B\\ C & D\end{pmatrix} = \begin{pmatrix}z & \sqrt{1 - z^2}\\ -\sqrt{1 - z^2} & z\end{pmatrix}.$$

There is a similarity, yet only a partial similarity, between

$$\bar{x}^2 + \bar{y}^2 = x^2 + y^2$$

and

$$\bar{x}^2 - (c\bar{t})^2 = x^2 - (ct)^2.$$

Also, just as we do not allow a transformation $\bar{x} = x$, $\bar{y} = -y$ which destroys the orientation of our coordinate system, we cannot allow a transformation $\bar{x} = x$, $c\bar{t} = -ct$ in spite of the fact that $x^2 - c^2t^2$ would be unchanged under such a substitution. Indeed, the transformation $\bar{t} = -t$ would interchange

ERIC

past and future and make clocks run backwards. It is a deep philosophical question what makes time run in a unique sense and why we cannot change its course by physical devices. We cannot discuss this problem which has been the despair of wiser men. But we can use the fact that past and future cannot be interchanged to exclude the special transformation $\bar{x} = x$, $\bar{t} = -t$ which plays the same role here as the reflection played in coordinate transformation.

Thus, had the minus signs been plus signs, with the substitutions $ct = y$, $c\bar{t} = \bar{y}$, Theorem A would have been immediately applicable and our problem solved. What a pity.

Still, wishful thinking has its uses. If we do not make a wish, we do not have a wish to come true. How can we change

$$\bar{x}^2 - c^2\bar{t}^2 = x^2 - c^2t^2$$

into

$$\bar{x}^2 + c^2\bar{t}^2 = x^2 + c^2t^2?$$

We cannot. Yet if we cannot have all our wish, can we have a part of it? We write

$$x^2 - c^2t^2 = x^2 + (-c^2t^2).$$

This is a little better; we have introduced a plus. And remembering conveniently that $i = \sqrt{-1}$, we have

$$-c^2t^2 = -1 \cdot c^2t^2 = i^2c^2t^2$$

so that

$$x^2 - c^2t^2 = x^2 + (ict)^2$$

and, similarly,

$$\bar{x}^2 - c^2\bar{t}^2 = \bar{x}^2 + (ic\bar{t})^2.$$

Therefore the condition

$$\bar{x}^2 - c^2\bar{t}^2 = x^2 - c^2t^2$$

may be replaced by the condition

$$\bar{x}^2 + (ic\bar{t})^2 = x^2 + (ict)^2.$$

Thus it would seem that the best we can do is put

$$y = ict, \quad \overline{y} = ic\overline{t}.$$

With these substitutions Theorem A becomes:

Given $\begin{pmatrix} \overline{x} \\ ic\overline{t} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ ict \end{pmatrix}$.

$\overline{x}^2 + (ic\overline{t})^2 = x^2 + (ict)^2$ [i.e., $\overline{x}^2 - c^2\overline{t}^2 = x^2 - c^2t^2$]

if and only if

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} z & \sqrt{1 - z^2} \\ -\sqrt{1 - z^2} & z \end{pmatrix}.$$

In consequence

$$\begin{pmatrix} \overline{x} \\ ic\overline{t} \end{pmatrix} = \begin{pmatrix} z & \sqrt{1 - z^2} \\ -\sqrt{1 - z^2} & z \end{pmatrix} \begin{pmatrix} x \\ ict \end{pmatrix}$$

so that

$$\overline{x} = zx + \sqrt{1 - z^2} \cdot ict$$
$$ic\overline{t} = -\sqrt{1 - z^2} \cdot x + z \cdot ict.$$

We have succeeded, at least in a formal way, in determining the necessary transformation between x,t and $\overline{x},\overline{t}$. We seem, however, to have paid a heavy price for it; there is this wretched number i. Of course, a clock can record a time t or a time ct, yet no clock can record a time ict. Must we conclude these equations to be without physical significance?

Consider $\sqrt{-6} \cdot i$. This is a sheep in wolf's clothing. Since the notation contains the letter i, it is natural to suppose $\sqrt{-6} \cdot i$ to be an imaginary number, yet $\sqrt{-6} \cdot i = \sqrt{-6} \cdot \sqrt{-1} = \sqrt{(-6)(-1)} = \sqrt{6}$, a real number. Are our transformation equations imposters, also? Let us unclothe them to find out.

Writing $\sqrt{-1}$ for i in the first equation

$$\begin{aligned} \overline{x} &= zx + \sqrt{1 - z^2} \cdot \sqrt{-1} \cdot ct \\ &= zx + \sqrt{(1 - z^2)(-1)} \cdot ct \\ &= zx + \sqrt{z^2 - 1} \cdot ct. \qquad (1) \end{aligned}$$

Making the same substitution in the second equation

$$\sqrt{-1}\cdot c\bar{t} = -\sqrt{1-z^2}\cdot x + z\sqrt{-1}\cdot ct$$

dividing both sides by $\sqrt{-1}$

$$c\bar{t} = -\frac{\sqrt{1-z^2}}{\sqrt{-1}}\cdot x + zct$$

$$= -\sqrt{\frac{1-z^2}{-1}}\cdot x + zct$$

$$= -\sqrt{z^2-1}\cdot x + zct. \qquad (2)$$

Real coefficients! Provided $z^2 \geq 1$.

Echoing the argument of Section 4.7, it remains to determine the physical significance of z. We return to the railway track. M_0 speeds through the night sitting in his engine cab. He says, "I do not budge an inch; I am $\bar{x} = 0$"; S_0 says, "Oh no, to the contrary, you are moving very fast, you have the same velocity v as your train." Putting $\bar{x} = 0$ in (1)

$$0 = zx + \sqrt{z^2-1}\cdot ct$$

i.e.,

$$x = \frac{-\sqrt{z^2-1}}{z}\cdot ct.$$

And since, when $t = 0$, $x = 0$, this equation tells us the distance that M_0 has traveled from S_0 in time t (by S_0's watch). x, the distance traveled, is proportional to t, the time taken by the factor $-\frac{\sqrt{z^2-1}}{z}c$. But, as every schoolboy knows,

distance traveled = velocity × time spent traveling.

So? The velocity v of M_0 and his train relative to S_0 and the track is given by

$$v = \frac{-\sqrt{z^2-1}}{z}c$$

where v is real when the above condition that $z^2 \geq 1$ is satisfied. We have found the physical significance of a function of z. This suffices.

With the remark that it is convenient to begin by writing (1) -- and _mutatis_ _mutandis_ (2) -- in the form

$$\bar{x} = z\left\{x - \left(-\frac{\sqrt{z^2 - 1}}{z} \cdot c\right) t\right\}.$$

I leave it to the reader to show that

$$\bar{x} = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad (3)$$

$$\bar{t} = \frac{t - \frac{vx}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}}. \qquad (4)$$

This is Einstein's solution.

In brief: Given that c, the velocity of light, is invariant, if S_0 describes an event as happening at a distance x from himself at time t by his watch and M_0, who is moving with velocity v, describes it as happening at a distance $\bar{x}$ from himself at time $\bar{t}$ by his watch, then the relations between $\bar{x},\bar{t}$ and x,t are given by Equations (3) and (4).

We recall that the completely mathematical formulation of Einstein's problem -- expressed in matrix notation -- is:

Given that

$$\begin{pmatrix}\bar{x}\\ \bar{t}\end{pmatrix} = \begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}x\\ t\end{pmatrix}$$

is subject to the condition that

$$\bar{x}^2 - c^2\bar{t}^2 = x^2 - c^2t^2,$$

find

$$\begin{pmatrix}A & B\\ C & D\end{pmatrix}.$$

It is fitting to conclude this section by giving Einstein's answer in the same notation. A moment's thought will show that (1) and (2) may be written as follows:

$$\begin{pmatrix} \bar{x} \\ \bar{t} \end{pmatrix} = \begin{pmatrix} \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}} & \dfrac{-v}{\sqrt{1-\frac{v^2}{c^2}}} \\ \dfrac{-\frac{v}{c^2}}{\sqrt{1-\frac{v^2}{c^2}}} & \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}} \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix}$$

The matrix

$$\begin{pmatrix} \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}} & \dfrac{-v}{\sqrt{1-\frac{v^2}{c^2}}} \\ \dfrac{-\frac{v}{c^2}}{\sqrt{1-\frac{v^2}{c^2}}} & \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}} \end{pmatrix},$$

first expressed in this form by the Dutch physicist Lorentz (1853-1928), is known as the Lorentz matrix. Such matrices are analogous to orthogonal matrices; they constitute a group. The given matrix, when multiplied by a similar matrix, with only v replaced, gives another matrix of the same form.

To consolidate this knowledge you are asked to work out for yourself an elaboration of Einstein's problem. We introduce a second motorman M_1, the origin of the moving system $(\bar{\bar{x}},\bar{\bar{t}})$, who drives his train in the same direction as M_0 drives his on a parallel track. See Fig. 13.

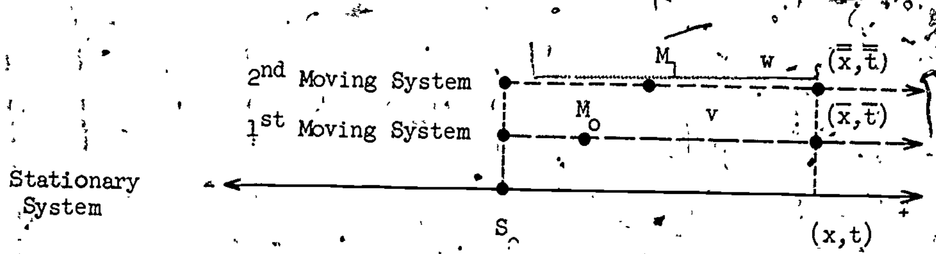

Fig. 13

When $x = 0$, $t = 0$ and $\bar{x} = 0$, $\bar{t} = 0$, also $\bar{\bar{x}} = 0$, $\bar{\bar{t}} = 0$ (i.e., M_1 passes S_0 when M_0 does, and all three synchronize their watches). Given that M_1

moves with velocity ω relative to M_0, by using Lorentz matrices in a role analogous to that of orthogonal matrices in Section 4.3, deduce formulae for $\bar{\bar{x}}$ and $\bar{\bar{t}}$ in terms of x,t. Confirm your answer by considering the motion of M_1 relative to S_0.

4.13 Einstein's Achievement.

I have taken great pains to try to make Einstein's formulation of his space-time transformation problem and his solution by matrix algebra readily intelligible to the reader who will give them his serious attention. Being wise after the event, it is difficult to appreciate the magnitude of his achievement. Now that we have the comforting assurance of a well sign-posted road that we will reach our destination, we forget that when there was no road there was no road to follow. Yet, making the road was the easy part. We forget absolutely that without a new destination, there could never have been a new road. Einstein had to see that there was a place to go to before he could figure out how to get there.

Analogy will help us to see his achievement in perspective.

Given a hammer, a bag of nails, and the instruction, "Get busy," what does a boy do? Drive a few nails in the wooden floor? Fun for a youngster yet unimaginative. Or, drive a nail in the door to improvise a hat peg? That's a more intelligent thing to do. Or, drive dozens of nails to make dozens of hat pegs? The boy who does this runs out of ideas before he runs out of nails. But what about the highly imaginative boy? He drives his hat peg nails into the door up to their heads so that their points stick out on the other side. Why, don't you see? Take the door off its hinges -- and there's a fakir's bed of nails! Not every Tom, Dick, or Harry would think of that. It takes imagination. What's that you say? A crazy idea? Come to think of it that's just what a lot of physicists at first said about Einstein's 1905 space-time transformation paper.

Einstein saw what contemporary mathematical physicists failed to see; he saw how to "get busy". He did what his contemporaries failed to do; he used matrices with bold imagination.

4.14 Important Consequences.

Equations (1) and (2), (page 144), are the basis of Einstein's Special Theory of Relativity. We may or may not be disposed to accept them. But whether or not we like it, the fact remains that these are necessarily consequences of the invariance of the velocity of light.

In this, the final section, we shall consider three major consequences of these equations, consequences that shatter our complacency. To accept the basis of the Special Theory of Relativity without accepting its consequences is illogical. If we are willing to accept the evidence of the Michelson-Morley experiment, we should likewise treat its logical consequences.

(1) Faster astronauts age more slowly.

We return to the railway track again. Suppose that S_0 sees the tree by the track momentarily made visible by the flash of light from his lantern at 12:01 by his watch, i.e., when $t = 12{:}01$. Does M_0 see it before or after S_0? Remember that they synchronized their watches when M_0 was at S_0. In other words, is $\bar{t}$ less than or greater than t?

Since S_0 remains at the origin of his system $x = 0$, so that (2) becomes

$$\bar{t} = \frac{t}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad (1)$$

And since the velocity of the train v is, of course, > 0,

$$\sqrt{1 - \frac{v^2}{c^2}} < 1, \text{ and } \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} > 1, \text{ so that}$$

$$\bar{t} > t. \qquad (2)$$

Therefore M_0 describes the event that the tree was momentarily visible as occurring later. Suppose, to be definite, M_0 says that the event occurred at

12:02. What does S_0 conclude? He says, on the basis of the event, timed by his watch as happening at 12:01 and timed by the moving watch as not taking place until 12:02, that the moving watch runs fast and that events as described by the moving man lag behind the same events as described by himself.

Although equations (1) and (2) are simple, their physical application is most difficult; it makes no concession to muddle-headedness. We must be clear that $\bar{t}$ is the time recorded by a watch that moves relative to a watch that records time t. (1) may be expressed

$$\text{moving clock time} = \frac{\text{stationary clock time}}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad (1')$$

and (2)

$$\text{moving clock time} > \text{stationary clock time.} \qquad (2')$$

Suppose the station master has a stop watch whose hand makes a full turn in one second. The motorman would think that the stop watch is slow because in his opinion the time $\bar{t}$ for such a turn is more than a second. But suppose he is given an identical stop watch. Then he will now say that his own stop watch is correct and that its hand makes one turn per second. However, the station master looking in will conclude that the engineer's stop watch is slow, since he moves relative to the train and now his time scale is increased. Thus, the total consequence of (2') is as follows. A process of physics which would take at rest an amount of time t appears to an observer moving relative to it as longer. If you ask, therefore, who of two observers is more justified in ascribing time to a given physical phenomenon, we should say that that observer will have the better judgment who rests relative to the apparatus or phenomenon which is to be judged.

Equations (1') and (2') have most important consequences for space travel at velocities near that of light.

We now suppose M_0 to be an astronaut heading straight for a distant star from the earth at S_0. When he is hurtling through outer space with velocity v,

ERIC Full Text Provided by ERIC

his clock, which measures time $\bar{t}$, is at rest relative to him, and the terrestial clock, which measures time t, is moving relative to him. This is the crucial point: In M_0's experience, $\bar{t}$ is his local or stationary clock time, t is terrestial or moving clock time. And since he moves at velocity v relative to the earth, the earth moves at $-v$ relative to him. But $(-v)^2 = v^2$, so that by (1'), for M_0

$$t = \frac{\bar{t}}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad (1^*)$$

i.e.,

$$\text{terrestial clock time} = \frac{M_0\text{'s time}}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Suppose, to take a concrete illustration with easy arithmetic, that M_0 travels with $\sqrt{\frac{99}{100}}$ the velocity of light. With $v = \sqrt{\frac{99}{100}}\,c$, $\frac{v^2}{c^2} = \frac{99}{100}$ so that

$$\sqrt{1 - \frac{v^2}{c^2}} = \sqrt{1 - \frac{99}{100}} = \sqrt{\frac{1}{100}} = \frac{1}{10}$$

and

$$t = \frac{\bar{t}}{\frac{1}{10}}$$

i.e.,

$$\bar{t} = \frac{1}{10}t.$$

Thus the duration of M_0's experience in traveling from the earth to a distant star at a velocity of $\sqrt{\frac{99}{100}}\cdot c$ is only one tenth that of the terrestial observer's experience.

Suppose that according to S_0's watch M_0 takes 200 years to reach the distant star. Had M_0 set out at the age of 25, his body would be 225 years old when he reached his destination. Surely he would have arrived a corpse. No, $t = 200$ years is the duration of the flight in the experience of S_0 and his descendants, the people who stay at home. M_0, the man who goes, lives his experience in his own time $\bar{t}$, that of the watch he takes with him. When $t = 200$, $\bar{t} = \frac{1}{10} \cdot 200 = 20$. M_0 will be 45 years old, not 225, when he reaches his destination.

That the faster an astronaut travels the more slowly he ages gives us hope of men living long enough to visit the stars, all way out beyond the solar system. Yet you may be disposed to retort: Such subtle arguments are good, clean fun, but would any hard-headed astronaut be prepared to set out on a 200 year journey because it had been argued by a few long-haired professors that he would be only 20 years older when he got there? Not very likely. If I tear a page off my desk calendar and call today the first of September instead of the first of August, it doesn't make me any older physically. Next you will be telling me that if I forget to wind my watch, then I'll stop aging when it stops -- and live forever!

No, the point is that each physical phenomenon runs its natural course in the system in which it rests, and life is a physical phenomenon. The moving astronaut lives his regular life in his space capsule. He does not have any benefit from the fact that an observer on a different system (which moves with very high speed relative to him) thinks that he lives very much longer. At this moment there are many galaxies which move relative to the earth with fantastic speed, nearly the velocity of light. If there were in such a galaxy a star with intelligent observers, they would think that we humans are practically immortal. This does not do us much good. However, for such purposes this difference in aging is a great use. While according to our system of accounting, an astronaut might need 200 years to reach a distant object, in his time scale he would need only 20 years and thus be able to survive his trip.

You may be quite bewildered and upset by our argument. But remember that your experience in life has been in systems of very slow motion, and there is nothing which could prepare your imagination to experiences of high speed travel in outer space. Wherever experience fails us, insight and intuition will fail likewise. Our only guide is our reason strengthened by mathematical argument.

ERIC

We might be wrong in our extrapolations, but until now the predictions of science have been more frequently verified than falsified.

Do you know what a radioactive substance is? It consists of a large number of atoms, of which, during a given period, a certain percentage disintegrates or dies. Uranium atoms, when placed in a cyclotron, are made to travel at nearly the velocity of light -- just as we suppose M_0 to do. It is found that uranium subjected to such cyclotron experience decays much more slowly than uranium subjected to ordinary terrestial experience. Here is evidence in favor of Einstein's time contraction formula. And isn't our aging a physiological process whose rate is that at which tissue and that sort of thing decay?

(2) No traveling faster than light.

Using (1) and (2), we divide $\bar{x}$ by $\bar{t}$. This is a nice thing to do for it eliminates the square root.

$$\frac{\bar{x}}{\bar{t}} = \frac{x - vt}{t - \frac{vx}{c^2}} = \frac{\frac{x}{t} - v}{1 - \frac{v}{c^2}\cdot\frac{x}{t}} \qquad (3)$$

The algebra is very easy; the physical interpretation -- without which the algebra is pointless -- not quite so obvious.

Once again we return to our railway track. A passenger P, traveling without a ticket, is, for reasons best known to himself, running along the train (at uniform velocity) from M_0, where he was when M_0 passed S_0. See Fig. 14.

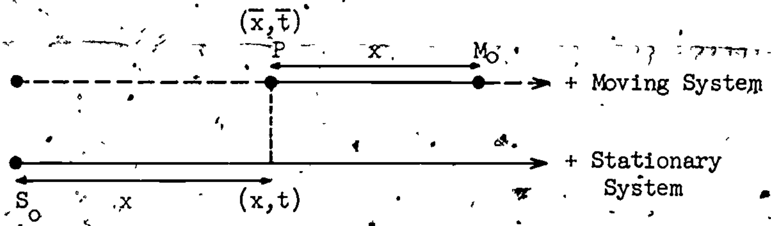

Fig. 14

Since P has coordinates $(\bar{x},\bar{t})$ on the train, he has moved distance $\bar{x}$ relative to M_0 in time $\bar{t}$. Accordingly, M_0 says that P's velocity is $\frac{\bar{x}}{\bar{t}}$. And since P has coordinates (x,t) relative to S_0, S_0 says that P has moved distance x in time t and that therefore P's velocity is $\frac{x}{t}$. For brevity, let

$$\bar{u} = \frac{\bar{x}}{\bar{t}}, \quad u = \frac{x}{t},$$

so that (3) becomes

$$\bar{u} = \frac{u - v}{1 - \frac{uv}{c^2}}.$$

We have a physical interpretation that gives P's velocity relative to the moving train in terms of his velocity relative to a fixed observer. This is the famous addition law of velocity.

What are its implications? Could we, from a rocket going at nearly the velocity of light, shoot off a rocket to go more nearly at the speed of light and from this, shoot off another to go even more nearly at the speed of light? By boosting velocity in this way, couldn't we achieve a velocity exceeding that of light? Let's use the addition law to find out.

The best that we can do for $\bar{u}$ is to take v to be a negative number. Replacing v by $-v$, the addition formula becomes

$$\bar{u} = \frac{u + v}{1 + \frac{uv}{c^2}} = \frac{u(1 + \frac{uv}{c^2}) + v - u \cdot \frac{uv}{c^2}}{1 + \frac{uv}{c^2}} = u + \frac{v(1 - \frac{u^2}{c^2})}{1 + \frac{uv}{c^2}} \qquad (4)$$

so that

$$\bar{u} > u \quad \text{provided that } 1 - \frac{u^2}{c^2} > 0$$

i.e., provided that $c > u$.

If $c = u$, then $\bar{u} = u = c$. Alternatively, putting $u = c$ in (4)

$$\bar{u} = \frac{c + v}{1 + \frac{v}{c}} = c.$$

We must conclude that it is impossible to exceed the velocity of light.

(4) is of intrinsic mathematical interest. If u and v are two given velocities, their combined velocity is given by the formula

$$f(u,v) = \frac{u + v}{1 + \frac{uv}{c^2}}.$$

This function is a kind of generalized sum of u and v. It satisfies the commutative and associative laws of addition:

$$f(u,v) = f(v,u); \quad f\big(u,f(v,w)\big) = f\big(f(u,v),w\big). \qquad (5)$$

If u and v are each less than c, the same will be true for this sum velocity.

An experimental verification for this law of addition of velocities can be found in an experiment by Fizeau which was, in fact, performed before the theory of relativity was even formulated. As you probably know, the velocity of light in water u is slightly lower than the velocity of light in empty space c. Suppose now that we send a ray of light through a body of water which moves itself in the direction of the ray with the velocity v. According to classical physics, the total velocity of the light ray should be u + v since the ray runs through the water with speed u, and the whole arrangement is carried forward with the velocity v. Fizeau carried out a very precise measurement of the velocity of such a ray in a moving medium. But to his surprise he discovered the following fact. The velocity of the light in the moving fluid was

$$\bar{u} = u + v\left(1 - \frac{u^2}{c^2}\right). \qquad (6)$$

Again, u is the velocity of light in the water and v is the stream velocity of the water.

Consider now the addition law (4). Observe that the flow velocity v is very small compared to the velocity of light c. Hence $\frac{uv}{c^2}$ is a very small number. We may use the geometric series formula to write

$$\frac{1}{1 + \frac{uv}{c^2}} = 1 - \frac{uv}{c^2} + \left(\frac{uv}{c^2}\right)^2 - \left(\frac{uv}{c^2}\right)^3 + \cdots .$$

We commit a very small error if we put

$$\frac{1}{1 + \frac{uv}{c^2}} \approx 1 - \frac{uv}{c^2}.$$

Observe that with the approximation the addition law (4) becomes the formula (6) established by Fizeau. The error due to our approximation is so small that the experimenter could not possibly observe it. Thus, Fizeau's formula is a brilliant justification for the addition law (4) of velocities.

But now we should add a remark which shows particularly well the great power of mathematical theory. Suppose a good mathematician had heard of Fizeau's experiment and of his formula (6) but had never heard of relativity theory and of the law (4) of addition of velocities. He might look at (6) and muse about its meaning. It is something like an addition law for u and v, but it does not satisfy the commutative and associative laws (5). The mathematician might suspect that (6) is only approximately true and is a good approximation to an addition law $f(u,v)$ in the case of small v. He might then ask the question: What is the function $f(u,v)$ which satisfies (5) and becomes very nearly (6) for small values of v? It can be shown that the only possible choice for $f(u,v)$ is the function (4). Thus, the mathematician could have deduced the correct law of addition of velocities from an approximate experimental formula. Think this over! You will understand why scientists call mathematics our sixth sense with which to experience reality.

(3) Energy has mass.

Fundamental to the study of even the most elementary dynamics is Newton's famous law that the force acting on a body is proportional to the mass times the acceleration of the body:

$$F = m \cdot a$$

In consequence, if a given body is acted upon by a constant force, it has a constant acceleration. But, if its acceleration is constant, then its velocity continually increases, so that finally it will go faster than light. On the other hand, if we accept the well verified Michelson-Morley result that the velocity of light is invariant, we are forced to accept its logical consequence that nothing can go faster than light. Something must happen to reduce the body's acceleration at high velocities.

To concentrate on the acceleration a, we isolate it by writing Newton's law in the form

$$\frac{F}{m} = a.$$

Since a must decrease for high velocities, so must the ratio $\frac{F}{m}$. But, by hypothesis, F is a fixed force, so what do you conclude? Yes, that for high velocities at least m must increase as the velocity of the body increases. However, it would be most odd if the increase were not continuous. In consequence, the only explanation Einstein could find is that mass must increase with velocity. Mass must become progressively harder to push with increasing velocity.

In his famous paper of 1905 Einstein argued that to the mass of a body at rest m_0 must be added the energy E of the body times $\frac{1}{c^2}$ to give it mass m_v, its mass at velocity v:

$$m_v = m_0 + \frac{E}{c^2}.$$

But the difference between masses m_0 and m_v is surely a mass, so that energy itself must be transformable into mass:

$$E = mc^2.$$

At that time this was a fantastic idea. Nobody had ever before thought that mass and energy could be eggs out of the same basket. Energy is mass in motion; mass itself is something that can be weighed, static, on a pair of scales. Surely energy is not the sort of stuff which can be weighed? Even in 1905 when Einstein was startled by his own idea he suggested that a study of radioactive substances -- where tremendous energies are hidden -- would very possibly show that energy can be transformed into mass and mass into energy. Forty years later, in 1945, this was all too dramatically verified; his thesis that $E = mc^2$ was the basis of calculation for the atomic bomb as well as for atomic power.

ERIC

4.15 In Retrospect.

Given the Michelson-Morley experimental result; what follows? We now know the more important consequences and the simple mathematics used to deduce them. In particular, we have seen what can be done with matrices when used with bold imagination. Of course, what was done elegantly with matrices could have been done inelegantly without them; but, who wants to drive nails with stones? Yet never forget that it is the man who handles the hammer that counts. Such simple mathematics enabled Einstein to change our entire conception of the physical world and to make a prediction, forty years in advance, that heralded new mastery of our world. This is an example of the power and the glory of mathematics -- and the genius of Einstein.

ERIC