ED 024 577

SE 004 991

By- Durst, Lincoln K., Ed.

Committee on the Undergraduate Program in Mathematics Geometry Conference, Part II: Geometry in Other Subjects.

Committee on the Undergraduate Program in Mathematics, Berkeley, Calif.

Spons Agency- National Science Foundation, Washington, D.C.

Report No- 17

Pub Date Sep 67

Note- 126p.

EDRS Price MF-$0.50 HC-$6.40

Descriptors- Calculus, *College Mathematics, *Conference Reports, Conferences, Course Content, *Curriculum, Curriculum Development, *Geometry, Instruction, *Mathematics, Undergraduate Study

Identifiers- California, National Science Foundation, Santa Barbara

This is Part II of the first volume of the proceedings of the Committee on the Undergraduate Program in Mathematics (CUPM) Geometry Conference, held at Santa Barbara in June, 1967. The purpose of the conference was to consider the status of geometry in colleges at the undergraduate level. This volume contains two lectures: "The Geometric Content o f Advanced Calculus" by Andrew Gleason and "The Geometric Content of Freshman and Sophomore Mathematics Courses" by Norman Steenrod. (RP)

COMMITTEE ON THE UNDERGRADUATE PROGRAM IN MATHEMATICS

REPORT

SEPTEMBER 1967

NUMBER 17

CUPM GEOMETRY CONFERENCE

PROCEEDINGS

PART II: GEOMETRY IN OTHER SUBJECTS

Lectures by A. M. Gleason and Norman Steenrod

MATHEMATICAL ASSOCIATION OF AMERICA

CUPM GEOMETRY CONFERENCE

Santa Barbara, California

June 12 - June 30, 1967

PROCEEDINGS OF THE CONFERENCE

Edited by Lincoln K. Durst

PART II: GEOMETRY IN OTHER SUBJECTS

Lectures by A. M. Gleason and Norman Steenrod

COMMITTEE ON THE UNDERGRADUATE PROGRAM IN MATHEMATICS

Mathematical Association of America

Financial support for the Committee on the Undergraduate Program in Mathematics has been provided by the National Science Foundation. A limited number of copies of these Proceedings is available from CUPM, Post Office Box 1024
Berkeley, California 94701.

FOREWORD

In June, 1967, CUPM sponsored a conference devoted to geometry in the undergraduate curriculum. The Proceedings of that conference are being issued in three parts, of which this is the second. An account of the background and the nature of the conference is given by Walter Prenowitz in his Introduction to the Proceedings, printed in Part I. A list of the lecture topics for the entire conference and a list of the names of the participants will be found on the following pages.

The texts printed here are based on recordings made of the lectures and the discussions, and were prepared for publication by the assistants, Melvin Hausner, John Reay, and Paul Yale. The lecturers themselves were able to make minor changes and corrections on the final sheets, but an early deadline prevented major revision or extensive polishing of the texts. The typing for offset was done by Mrs. K. Black and the figures were prepared by Mr. David M. Youngdahl.

Lincoln K. Durst
Claremont Men's College

LECTURE TOPICS

Herbert Busemann

The Simultaneous Approximation of n Real Numbers by Rationals.

An Application of Integral Geometry to the Calculus of Variations.

H. S. M. Coxeter

Transformation Groups from the Geometric Viewpoint.

Glen J. Culler

Some Computational Illustrations of Geometrical Properties in Functional Iterat'.on.

Andrew M. Gleason

The Geometric Content of Advanced Calculus

Branko Grünbaum

Convex Sets and the Combinatorial Theory of Convex Polytopes.

Preston C. Hammer

Generalizations in Geometry.

Paul J. Kelly

The Nature and Importance of Elementary Geometry in a Modern Education.

Victor Klee

Applications of Geometry

Walter Prenowitz

Joining and Extending as Geometric Operations: A Coordinate-Free Approach to n-space.

Norman E. Steenrod

The Geometric Content of Freshman and Sophomore Mathematics Courses.

MEMBERS OF THE CONFERENCE

Russell V. Benson
California State College, Fullerton

Gavin Bjork
Portland State College

John W. Blattner
San Fernando Valley State College

Herbert Busemann (Lecturer)
University of Southern California

Jack G. Ceder (Visitor)
University of California,
Santa Barbara

G. D. Chakerian
University of California, Davis

H. S. M. Coxeter (Lecturer)
University of Toronto

Glen J. Culler (Lecturer)
University of California,
Santa Barbara

Andrew M. Gleason (Lecturer)
Harvard University

Neil R. Gray
Western Washington State College

Helmut Groemer
University of Arizona

Branko Grünbaum (Lecturer)
University of Washington

Preston C. Hammer (Lecturer)
Pennsylvania State University

Melvin Hausner (Assistant)
New York University

Norman W. Johnson
Michigan State University

Mervin L. Keedy
Purdue University

Paul J. Kelly (Lecturer)
University of California,
Santa Barbara

Raymond B. Killgrove
California State College,
Los Angeles

Murray S. Klamkin
Ford Scientific Laboratory

Victor L. Klee, Jr. (Lecturer)
University of Washington

Rev. John E. Koehler
Seattle University

Sister M. Justin Markham
St. Joseph College

Michael H. Millar
Stanford University

H. Stewart Moredock
Sacramento State College

Richard B. Paine
Colorado College

Walter Prenowitz (Chairman)
Brooklyn College

John R. Reay (Assistant)
Western Washington State College

Paul T. Rygg
Western Washington State College

George T. Sallee
University of California, Davis

James M. Sloss (Visitor)
University of California,
Santa Barbara

Norman E. Steenrod (Lecturer)
Princeton University

George Stratopoulos
Weber State College

Robert M. Vogt
San Jose State College

William B. Woolf
University of Washington

Paul B. Yale (Assistant)
Pomona College

GEOMETRY IN OTHER SUBJECTS

CONTENTS

THE GEOMETRIC CONTENT OF FRESHMAN AND SOPHOMORE MATHEMATICS COURSES

Lectures by Norman Steenrod

(Lecture notes by Paul Yale)

Lecture I. Calculus.

In this series of lectures I am going to discuss the geometric content of the freshman and sophomore mathematics courses. I shall criticize what we as teachers are now doing and suggest what we might do. Let me begin with what I believe to be the chief criticisms.

1. Although geometry pervades all of mathematics and is present at every stage of a development, too often do we fail to point this out to our students. We rely on analytical formulations since we realize that they are complete and we are in a hurry to get on to other material. We do not take time to look at geometric formulations.

2. We are too greatly impressed by the rigor of analysis. We seem to feel that geometry is not rigorous, or at least that the background needed for rigor is not available. We feel that it is better not to do anything that is not rigorous. I think we are buffaloed too much by this.

3. When we do present geometry, it is too often the instructor who does the geometry while the student is merely a passive recipient. We present the geometry to him in order to explain the analysis, but then we require him to do only the analysis--no geometry.

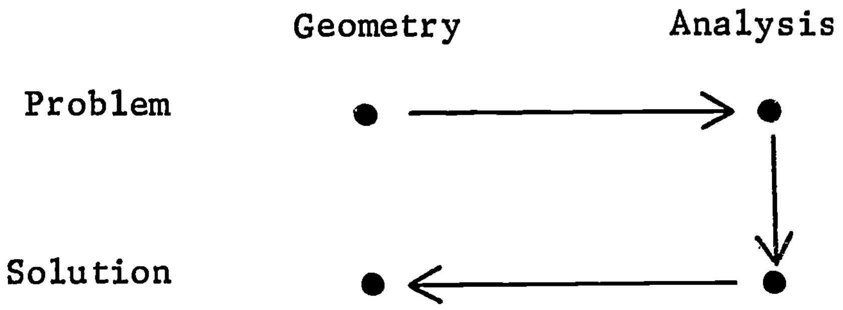
4. We tend to avoid geometric formulations of questions in examinations. Questions are difficult to formulate geometrically. Almost any time you try such a question, you find that a large group of students misinterpret it. Such questions are hard to grade because the answers are so varied. The absence of

geometric questions on final exams tends to degrade the geometric content of the course, and leads to its neglect.

Now that I have listed the main criticisms, let me take them up one at a time and fill in some details. The pervasiveness of geometry is an idea that goes back to Descartes, for a coordinate system in the plane or in space sets up an equivalence between geometry and algebra-analysis. Every geometric proposition can be translated into its algebraic-analytic analog and vice versa. I am not proposing that we lead the student through the details of the formal isomorphism between these two systems, but I am trying to remind you that the geometry is always there and to keep in mind that the geometric language for the conversion is always at hand. For example, here are a few geometric terms and their algebraic-analytic counterparts:

Geometric language	Algebraic-Analytic language
point, vector	number triple (x,y,z)
projection	coordinate, variable
surface	equation
plane	linear equation
region	system of inequalities
mapping, transformation	function
neighborhood	ϵ, δ
limit (using deleted neighborhoods)	limit (using ϵ, δ)
continuity (using neighborhoods)	continuity (using ϵ, δ)
tangent	derivative

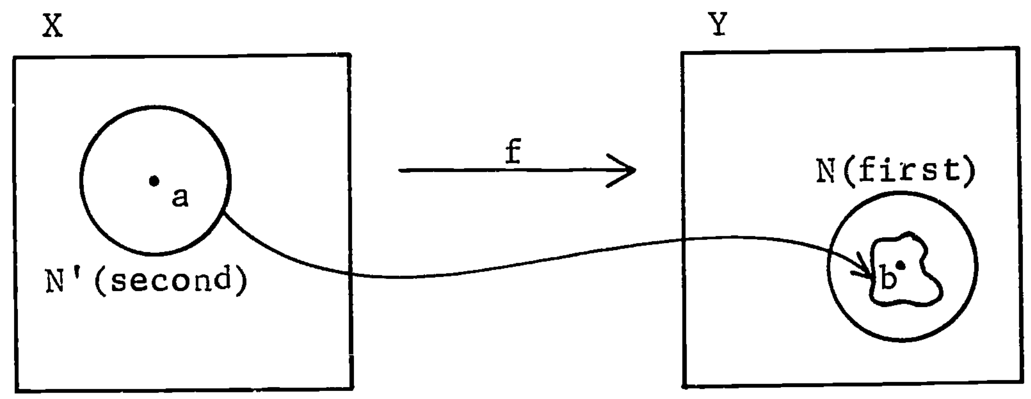
One might ask, in view of this equivalence, why bother with the geometry at all? The answer is clear to all of us. The first main reason is that many applications of calculus are to problems presented in geometric formulations; e.g., the focal properties of conics, oscillating systems, the 2-body problem. A complete solution of such a problem has three main steps: the reformulation in analytic terms, the derivation of an analytic solution, and the interpretation of the solution in geometric terms.



A second reason is psychological. Two views of the same thing reinforce one another. Most of us are able to remember the multitudinous formulas of analysis mainly because we attach to each a geometric picture that keeps us from going astray. Even better than that, the geometric view of a problem helps us to focus on the invariants and to weed out the irrelevant details. For example, a poor choice of a coordinate system may lead to a horrible mess in the analytic formulation, but with some geometric insight, we may be able to choose a much better coordinate system.

I turn now to suggestions for presenting some things from a geometric viewpoint. One of the main difficulties of teaching calculus is the problem of how to present the central notions of limit and continuity. Here I think the geometric way is clearly advantageous. Consider the geometric definition of a limit. Given a set X and a set Y, a point a in X and a point b in Y, and a mapping f from X into Y, what do we mean by "The limit of f at a is b"? The geometric answer is: For each neighborhood N of b, there is a <u>deleted</u> neighborhood N' of a which f maps into N. This is at least as simple as the standard analytic definition in terms of ϵ's and δ's; moreover, it has the following added virtues.

1. You can draw an easily-remembered picture and, if necessary, even label N as first and N' as second to emphasize that N' depends on N.



2. The geometric definition is valid for a much wider class of spaces and functions. The geometric definition reads exactly the same for all metric spaces, whereas the analyst, when he generalizes to a function mapping a k-dimensional space into an n-dimensional space, has to write out $n + k$ inequalities involving ϵ's, δ's and the coordinate versions of the function f.

To convert the definition of limit to one for continuity, we set $b = fa$ and drop the word deleted. There is no difficulty in deriving the analytic definition from the geometric one, but one can work directly with the geometric definition and easily prove the continuity of numerous functions. For example:

a. The identity function and all constant functions are continuous.

b. Composites of continuous functions are continuous.

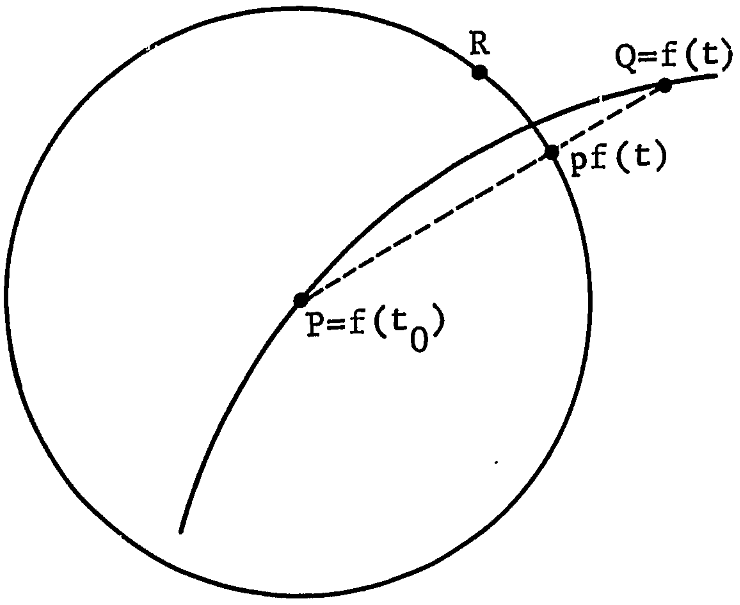
c. Isometries are continuous.

d. Any contracting map (e.g., a perpendicular projection) is continuous.

e. A mapping is continuous if it does not expand distances too much; i.e., if there is a constant k such that $d(fx,fy) \leqq kd(x,y)$ for all x,y in its domain.

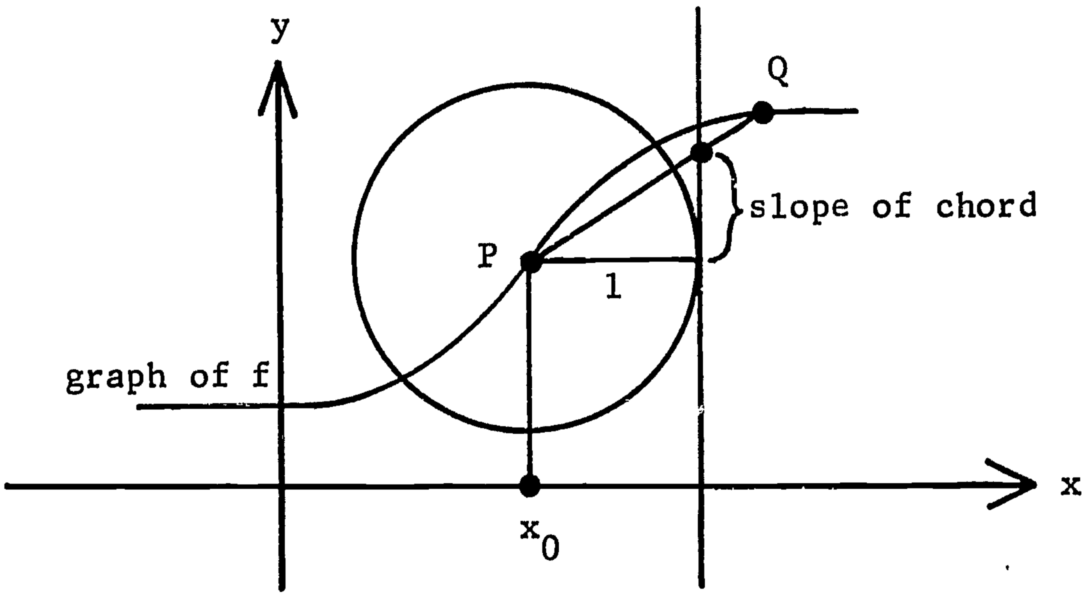
A mapping that is very useful is radial projection onto a sphere which we can prove to be continuous using the results above. Given a sphere with center c, we map each point x, $x \neq c$, to the point y where the sphere meets the ray from c through x. Outside the sphere this is a contraction, and inside the sphere distances are stretched at most by a factor of k, providing we stay outside a small sphere of radius 1/k. Therefore, radial projection onto a sphere is continuous wherever it is defined. This is a nice example of a function with an essential discontinuity which arises in a natural geometric setting.

The next critical topic in the calculus course is the tangent to a curve and its analytic companion, the derivative. I suggest the following geometric treatment of this topic. Of course, the concept of curve must come first, but let us assume we have adopted a definition in terms of continuous mappings from R into R^2 or R^3 and that we have seen enough examples to realize that this is a reasonable view. Suppose we are given a curve $f: [a,b] \to R^2$ and a point P on that curve, say $P = f(t_0)$ with $t_0 \in (a,b)$ such that $f(t) \neq P$ for t in some deleted neighborhood $(c,t_0) \cup (t_0,d)$ of t_0. We define the tangent at P as follows. For $t > t_0$, compose f with the radial projection p onto the unit circle centered at P. The composite pf of continuous mappings is continuous on the open interval (t_0,d). If pf has a limit R at t_0, then we call the ray PR the tangent from the right. We can repeat this for $t < t_0$ and define a tangent from the left. If the two one-sided tangents fit together to form a straight line, we call this line the tangent to the curve at P. Notice that when the tangent exists, then the function pf restricted to $[c,d]$ has a singularity



at t_0 with left and right limits at t_0 which are diametrically opposite points of the circle. It seems to me that this definition eliminates the vagueness usually associated with tangents in calculus courses.

It is easy to derive the analytic notion of a derivative from this geometric definition of a tangent. Given a function f, and a point P on its graph, consider the unit circle centered at P and the vertical line tangent



to that circle and one unit to the right of P. Except for the points directly above or below P, the radial projection p' from P onto this vertical line is continuous so that the composition $p'f = p'pf$ is continuous for $x \neq x_0$. It follows that pf has a limit at x_0 from the right, if and only if p'f does. This is equivalent to the statement that the slope of the tangent is the limit of the slopes of chords. If the one-sided limits exist at x_0, then we have the usual one-sided derivatives, and if these are equal, we say that f has a derivative at x_0.

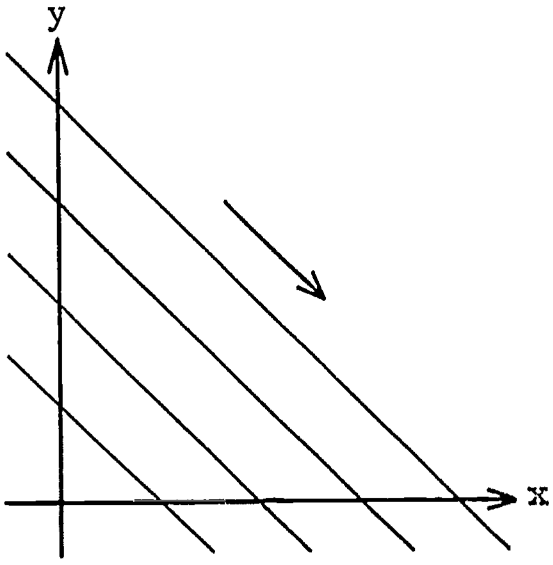
Discussion.

Hausner asserted that most of the analysts he knew would agree with this style of presentation and would have no argument with the geometric point of view. In reply, Prenowitz asked "Then where do all the calculus books come from?" Hausner answered that we have all been imitating Euler.

Woolf asked how the neighborhood definition would fit in with the algebra of the functions when sums and products are defined on the image space. Steenrod responded with the following geometric proof that if $f: R \to R$ and $g: R \to R$ are continuous, so is $f + g$.

1. It is enough to prove that addition is continuous since the function $f + g$ is then a composite of two continuous functions.

2. Addition is simply a projection of R^2 onto R^1 along the lines of constancy of $x + y$. This projection stretches distances by at most $\sqrt{2}$; hence, it is continuous.



He then asserted that he saw no difficulty in interweaving the geometric definition of continuity with the algebraic structure on the space of functions.

Stratopoulos asked for clarification of what was being proposed since initially he understood it to be that we should spend more time to present the geometric point of view in a calculus course when it is helpful, but from some of the remarks made in the discussion, it sounded as if you said we should teach calculus only this way. Steenrod replied that he proposes that we try to motivate things geometrically and carry out as much of the argument as can be done conveniently geometrically, and then apply the analytic tools. It is only for the sake of debate that we seem to take the extreme view, but we are not proposing that we teach geometry in place of analysis.

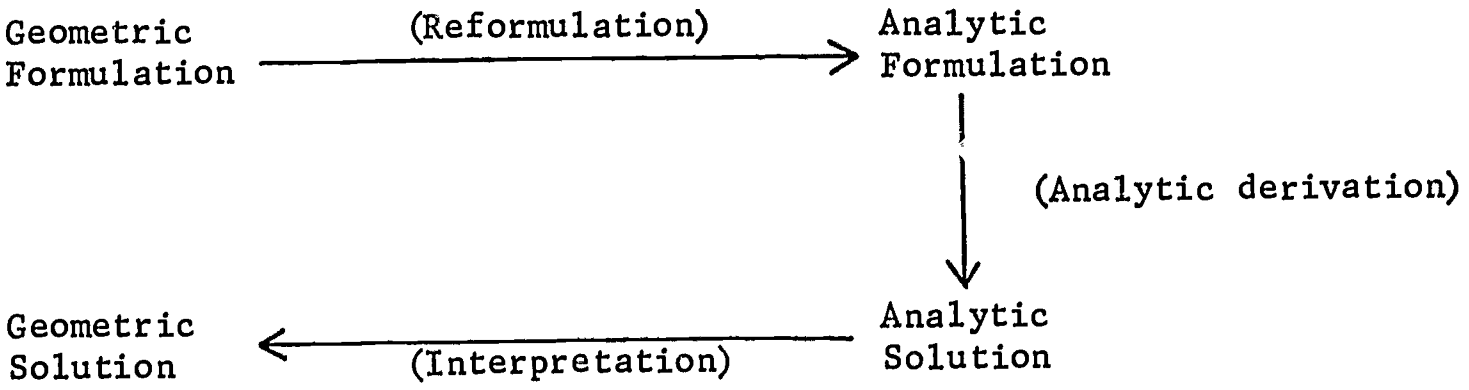
Hammer stated that he thought this presentation of tangents was very interesting from the standpoint of generality and that, among other things, it would be useful in clarifying directional derivatives.

Yale pointed out that this approach to limits, continuity, and tangents with its emphasis on mappings would be much easier to teach if the students were already familiar with and used to thinking in terms of transformations. Thus Kelly's position, that mappings should be taught in high school geometry courses, is relevant here. He then asked if Steenrod intended, as part of his proposal, to open the door to teaching calculus not only of one variable but of several variables, all intermixed together. Steenrod replied that this was right; that with this point of view, the artificial separation between calculus of one and of several variables is no longer there. Hammer pointed out the semantic difficulties we get into with our present distinction between derivatives and partial derivatives.

Prenowitz stated that he felt that this is not just a slightly different way of dealing with tangents, but, in a more general setting, a way of pinning down the elusive concept of the limit of rays. Namely, Steenrod has replaced the family of rays at P by the isomorphic representation (geometric, not analytic) using points on a sphere centered at P. The whole point is that the limit point on the sphere represents the limit ray we are interested in.

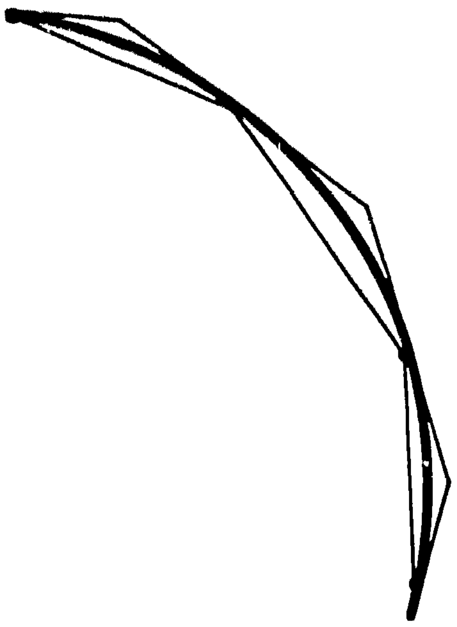
Lecture II. Calculus.

Today I shall continue my discussion of the four criticisms concerning the lack of a geometric viewpoint in current calculus courses. Yesterday I elaborated on the first (Geometry is always there but is presented too infrequently.) and pointed out the two main reasons for focusing attention on the geometric aspect as well as the analytic aspect of the subject. The first reason concerns applications. Very many problems with applications of the calculus arise geometrically. In such cases we have the diagram:



The heart of the calculus may be the middle step, but to ignore the other two is simply not to train the student in the calculus. The other main reason is the psychological one that two views of the same question reinforce each other and enable you to see and do things that one viewpoint alone would not. In this regard, recall the geometric definitions of limit, continuity, and tangent that I presented yesterday. The neighborhood definitions of limit and continuity not only have the psychological advantage of being pictorial, but also the mathematical advantage of easy generalization to more complicated spaces. Moreover, it is easy to prove, directly from the geometric definition, that many functions arising in a geometric context are continuous. From the geometric definition of the tangent one derives immediately the standard definition of the derivative. I feel that this approach, placing the emphasis on the tangent, is much better than the analytic approach in which we first define the derivative and use it to define tangents to curves.

Now let us turn to another topic that has been a sore point in calculus for many years: arc length on a circle. Thomas avoids the question in his calculus book, Apostol bases his approach on the concept of area, but I believe that we should simply meet the problem head on. Basically the problem is to show that the least upper bound of the lengths of inscribed polygons is the same as the greatest lower bound of the lengths of the circumscribed polygons. Of course it is easy enough to simply define it as one or the other of these but in order to use it you need to look at the other also, so essentially the problem is to show that they are equal.



For each partition, p, of an arc of a circle of radius r let $\ell(I_p)$ be the length of the associated inscribed polygon I_p, and let $\ell(C_p)$ be the length of the associated circumscribed polygon C_p. Using the small triangles as in the figure it is easy to show that $\ell(I_p) \leqq \ell(C_p)$. Now assume that p' refines p, i.e., that we have put some additional points on the arc. Then another triangle argument shows that $\ell(I_p) \leqq \ell(I_{p'})$ and $\ell(C_{p'}) \leqq \ell(C_p)$. Thus, to solve the problem, we only need an estimate on the size of $\ell(C_p) - \ell(I_p)$ in terms of something we can control and then show that this tends to zero.

For a partition p, label the circular arcs in order from 1 to k, let a_i denote the length of half of the $i^{\underline{th}}$ inscribed segment, and b_i the length of the corresponding half segment of the circumscribed polygon (see the

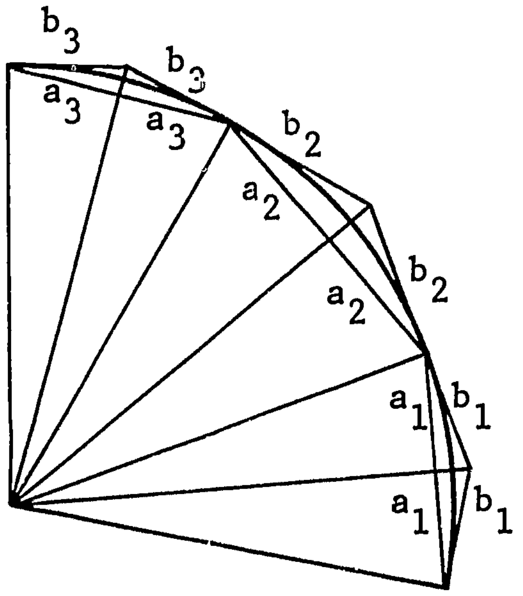


figure). By similar triangles,

$$\frac{b_i}{r} = \frac{a_i}{\sqrt{r^2-a_i^2}} ,$$

and if

$$m_p = \max (a_1,\ldots,a_k),$$

we have

$$b_i - a_i = a_i \left(\frac{r}{\sqrt{r^2-a_i^2}} - 1\right) \leqq a_i \left(\frac{r}{\sqrt{r^2-m_p^2}} - 1\right) .$$

Since $2\,\Sigma_1^k a_i = \ell(I_p)$, and $2\,\Sigma_1^k b_i = \ell(C_p)$, we obtain by summing these inequalities

$$\ell(C_p) - \ell(I_p) \leqq \ell(I_p)\,\frac{r - \sqrt{r^2-m_p^2}}{\sqrt{r^2-m_p^2}} .$$

Let c denote the length of some fixed circumscribed polygon, so that $\ell(I_p) \leqq c$ for all partitions p. The triangle inequality for the right triangle with hypothenuse r and altitude m yields $r \leqq m + \sqrt{r^2-m^2}$, hence

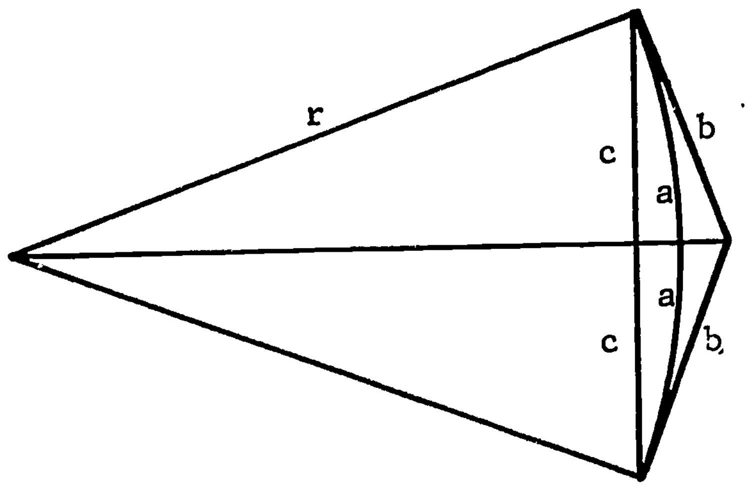
$$r - \sqrt{r^2-m^2} \leqq m .$$

Also we have $r - m \leqq \sqrt{r^2-m^2}$, so if we confine ourselves to partitions p such that $m_p \leqq r/2$, we have $\sqrt{r^2-m_p^2} \geqq r/2$. Combining these inequalities gives the required squeeze

$$\ell(C_p) - \ell(I_p) \leqq \frac{2c}{r} m_p .$$

It follows that g.l.b. $\ell(C_p)$ = l.u.b. $\ell(I_p)$.

With this definition of arc length for a circle, we have the following easy proof that the limit of the ratio of chord to arc, as the chord tends to zero, is 1. Consider for a short arc the trivial partition consisting of



its two end points. If a is half the arc length, we have by definition

$$\ell(I_p) < 2a < \ell(C_p)$$

and this implies $c < a < b$ where b, c are as shown in the figure. By similar triangles, $b = \frac{cr}{\sqrt{r^2-c^2}}$.

Now $c < a$ implies $0 < 1 - \frac{c}{a}$, and $a < \frac{cr}{\sqrt{r^2-c^2}}$ implies

$$1 - \frac{c}{a} < \frac{r - \sqrt{r^2-c^2}}{r} .$$

As in the paragraph above, $r - \sqrt{r^2-c^2} \leqq c$,

whence

$$0 < 1 - \frac{c}{a} < \frac{c}{r} .$$

Therefore $\lim_{c \to 0}\left(1 - \frac{c}{a}\right) = 0$ as required.

It is to be observed of course that we have proved the geometric form of the analytic statement

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

All calculus books endeavor to make this proposition seem reasonable; in most cases they are content with showing its equivalence to its geometric form.

The above treatment of arc length on a circle should be considered in connection with my second main criticism: we are too greatly impressed by the rigor of analytic methods. Why do we shirk doing a proper job of $\frac{\sin x}{x}$? The reason is two-fold. First we recall that the most trouble-free definition of $\sin x$ is by its Maclaurin series; and secondly, the concept of arc length for curves in general is a delicate and lengthy process involving functions of bounded variation. Since to proceed thus would be inappropriate at this early stage of the calculus, we merely state the result, wave our hands, and refer students to a more advanced course. Such an "all or nothing" attitude is not justifiable; in this case we have to do with a very special curve, the circle; and quite special methods are both adequate and appropriate.

Now let us consider the last two of my main criticisms, the problems they pose are how to get the student involved in the geometry and what to do about geometric problems on exams. These are strictly pedagogical and not mathematical problems. I suppose everyone has his own pet methods; those I offer, I recommend simply as devices that I have tried with some success.

The subject of curve sketching is treated adequately in most texts, however the material is usually confined to a portion of one chapter. If the student sketches curves for only one week, he is not going to become involved in the geometry of curve sketching. I usually adopt a program of insisting on curve sketching as a regular part of the entire course. At least one homework problem each week involves curve sketching. In addition I give two ten minute quizzes each week, and at least a third of these quizzes involve curve sketching. It is inevitable that some students will submit as a sketch a plot of three or four points connected by a sloppily drawn curve. So I tell my students that when sketching the graph of a function they must show clearly where the function is rising or falling, show where it is convex or concave, indicate the behavior of the graph as x tends to $\pm\infty$, show any horizonal or vertical asymptotes, and state clearly all symmetries.

It is easy to make curve sketching a regular part of the course because throughout the first year of the calculus we introduce functions of gradually increasing complexity: linear, polynomial, rational, algebraic, trigonometric, inverse trigonometric, exponential, and logarithmic. Also the subjects of curves in parametric form and polar coordinates offer additional and natural opportunities for more curve sketching.

One reason for heavy emphasis on curve sketching is to keep students convinced that a formula is not just a string of symbols to play games with according to certain rules. It represents something that has an existence independent of the formula, and its geometric presentation is another aspect of its existence. Also the agility they develop during the term enables one to ask geometric questions on final examinations; it does not take them forever to recognize the curve or surface specified in the problem. For these reasons I regard curve sketching as an integral part of the calculus program.

Although most calculus books treat the conic sections, many do so in one chapter somewhat late in the text. I do not understand this; the conics are prize examples for illustrating much of the material of the calculus. As soon as students have learned to differentiate polynomials, one can introduce parabolas in standard form $(4ay = x^2)$ and prove the focal property of a parabolic mirror. As soon as they can differentiate rational functions, one can introduce ellipses via the string definition $((x-c)^2 + y^2)^{\frac{1}{2}} + ((x+c)^2 + y^2)^{\frac{1}{2}} = 2a$, reduce to standard form, and prove the focal property of an elliptical reflector. A number of maximum-minimum problems involve the conics. Areas of ellipses and volumes of ellipsoids may be computed. The conics provide excellent examples for parametric equations of curves and also for polar equations. Of course a full treatment of conics (sections of cones, focus-directrix, reduction of quadratics by rot.tions and translations of coordinates) should come as a unit later on in the course; but much good material can precede this.

Another good way to get the student involved in the geometry is to play variations on formulas, so that he has to understand the geometric background of the formula if he is to solve an assigned problem. Most formulas in the text are presented in specialized form, assuming some standard organization or position with respect to the coordinate system. The case of volumes of solids of revolution is a typical example. The standard formulas are for solids obtained by rotating about one of the coordinate axes. A good problem for a ten minute quiz is to ask the student to derive the analogous formulas for rotation about a line parallel to an axis, say the line $y = -2$.

With the final exam or any other major exam we come to something which has bothered me through the years, and I am sure has bothered a lot of you. This is the control the exam seems to have on the structure of the course. Somehow the tail wags the dog. In the exam we are supposed to take a sample of what the student knows, and surely in three hours, we can do no more than take a sample. This process of sampling has a feedback effect that is very serious. The most famous example is the College Board exam and its influence on the teaching of mathematics in secondary schools. The examiners, in order to be fair to students in all parts of the country, tended to take the intersection of the topics taught in various schools and asked questions about this intersection. In the 1920's and 1930's the exam had little effect on the teaching of mathematics, but by the early fifties the feedback effect became pronounced. Sizable numbers of students were taking the exam, and schools were rated by the results. If a particular high school had a poor rating, they did something about it; they compared carefully what they were teaching with the kinds of questions asked on the exam; they altered their curriculum accordingly, and concentrated on topics of maximum frequency. The examiners, on their part, observed the shrinkage and narrowed the range of their questions accordingly. At one time it was projected that after forty years only one topic would survive this convergence process, and that would be the factoring of quadratics.

But, you say, this cannot happen in college because the instructor has charge of his course. Well, he does not, because in many schools there are freshman courses with large enrollments and many sections. To avoid troubles with young instructors giving wide varieties of grades we insist on uniform exams and uniform grading. I have seen the "feedback effect" time and again while teaching a section of the freshman course. Along comes a bright fresh Ph.D. teaching his first class. Knowing that the concept of limit is central to the calculus, he settles down and does a good job of teaching limits for two months. However on the uniform midterm exam there is only one question out of five on theoretical aspects of limits. His students do very well on that one question but not so well on the other four of a more routine nature. The average score for his students is ten points below the overall average, so he finds himself giving D's to students he thought were pretty good. Having learned his lesson, he runs a statistical analysis on the final exams for the last five years, and starts teaching his students how to turn the crank. By the end of the semester he usually brings their average up to where it should be.

I do not know how to defeat this, but I do have one suggestion to offer. Use the feedback effect to upgrade geometry by putting more of the geometric questions into the final exam and then face the problem of grading them. If, in the earlier parts of the course, on the ten minute quizzes and the homework, you have inflicted geometry on the student over and over again, then on the final exam you have some chance of getting a good reaction out of a geometric question. Let me give you two examples of the types of questions I have in mind. Instead of posing the purely analytical problem: evaluate a certain definite integral, ask instead for the area of the region between two specified curves. Similarly, instead of asking for a routine implicit differentiation, e.g., find dy/dx if $x^3 - 2xy^2 - 3y = 1$, ask the student to verify that a point, in this example (2,1), is on the curve and to find the tangent line at that point. The essential idea in this type of question is to replace the routine analytic question by a question in which part of the data appears in geometric form and in which you ask for a geometric interpretation of the answer. The second type of question I call integration in reverse: sketch and describe a region whose area is given by

$$\int_0^{\sqrt{2}} (\sqrt{4-x^2} - x)\, dx$$

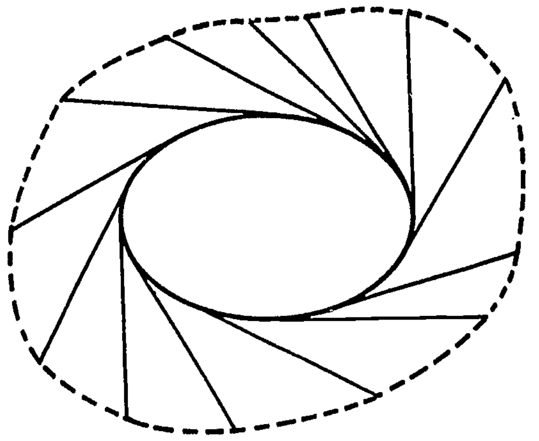
and set up an integral with respect to y for the same area. [In discussion over coffee someone pointed out that this type of question is especially effective when studying area in polar coordinates.]

This completes my comments on the first year calculus course. In the remaining three lectures I shall consider the typical second year material: linear algebra, calculus of several variables, and differential equations.

Discussion.

Gray advocated a geometric point of view when presenting a form of the completeness axiom for the real numbers, suggesting as a possibility, every bounded convex subset of the real line has endpoints.

Hammer stated the following theorem (Holditch's theorem according to Klamkin) as a nice application of polar coordinates that is seldom given in the standard calculus texts. Given a convex curve with a tangent line at every point consider the region swept out by tangent vectors

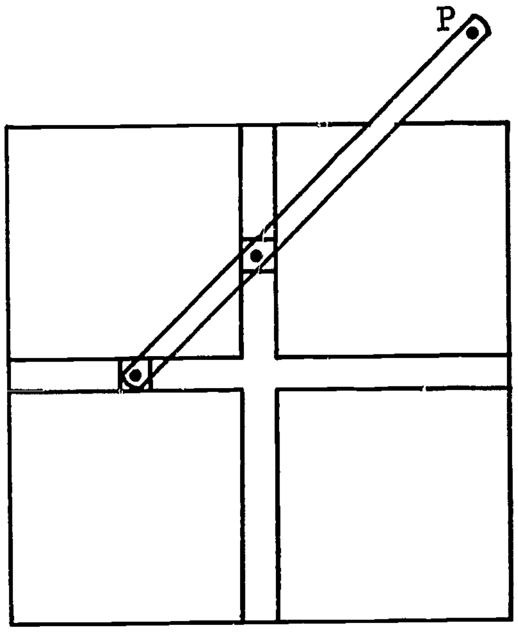


of fixed length, a. The area of this region is πa^2, independent of the convex curve we started with.

Benson, commenting that the book by Granville, Smith, and Longley was one of the few texts he has seen that had lots of problems of the type advocated, asked Steenrod if perhaps we should return to the old system in which analytic geometry was a separate course preceding the calculus. Steenrod replied that he had taught under both systems and was in favor of the combined course. He stated, that he for one tries to teach as much analytic geometry in the combined course as he used to teach in the separate course, and that it is up to us as geometers to write texts that hold up the geometric aspect of the course.

Keedy stated that at Purdue they try to encourage high schools to either teach a full year of calculus or none at all in order to avoid sending students to college who are bored for the first two weeks then lost for the rest of the semester. When asked by high schools what to do in place of a short introduction to calculus they advocate presenting some analytic geometry.

Hausner described an interesting linkage for tracing ellipses which can be used as the basis for a challenging question on parametric equations of ellipses. It consists of a block of wood with two slots at right angles and a handle attached to two sliding blocks in the slots. The problem is to show that as one turns the crank the handle (P in the diagram) traces out an ellipse. Coxeter asked if anyone present knew of an analogous device for part of a hyperbola.



Lecture III. Linear Algebra.

Until ten years ago linear algebra was a course given at the junior, perhaps senior, level for mathematics majors, but the situation is changing now and it is recognized that linear algebra is needed in the study of calculus of several variables. There are perhaps two reasons for this. First, the simplest mappings of an s-dimensional space into a k-dimensional space, other than the constant mappings which send all points in R^s to single points in R^k, are the linear and affine mappings. Almost all calculus and analytic geometry books start with the study of the straight line, $y = mx + b$. In the same way it's natural to begin the study of functions in higher dimensional spaces with the linear and affine functions so that you have a family of examples which you understand rather thoroughly. The second reason is that the general differentiable map can be analyzed locally (in the first approximation) by its differerential which is affine (or linear if related to a local coordinate system). That is, just as the local analysis of a curve in the plane replaces the curve by the tangent line, in higher dimensions the calculus analyzes a mapping locally by using the best affine approximation to the mapping--this is essentially the differential. Surely if one is to make this type of analysis one should clarify first the functions that are to be used in these approximations.

Linear algebra is needed, but it's not clear to me how much is needed or how much will be incorporated into the first two years of calculus. The procedure now at Princeton is that students planning to major in mathematics are required to take a sophomore linear algebra course. The students who are not planning to be mathematics majors, say engineers, take a regular calculus course and during the second year they have six to eight weeks of linear algebra. This meets with the hearty approval of the engineering departments. They want their students to have some familiarity with this material.

Now we come to the question of textbooks and of what material one can use in such a course, say within the calculus program. What books are available and what can be used? The situation here is definitely sad. Books on linear algebra are written by algebraists for algebra students. Of course this is not in itself a condemnation but there are two features that show up from this. One is that geometry is treated in an offhand fashion, if it is treated at all. Secondly, a semi-theoretical matter which applies to most but not all books, they confuse the presentation of theory with techniques of calculation via matrices. I will elaborate on this second point later.

First let me present an outline of the material that would be the most one would put into the sophomore course on the calculus. Perhaps not all of this material should be included but not more than this.

Of course one would start with the conceptual approach, if you will, the axiomatic approach. Discuss vector spaces, the basic operations, and the properties of the basic operations. Essentially the listing of the operations and their properties is the axiomatic approach. Examples are introducted right along with the axioms. These include R^2, R^3, R^k and function spaces. It is especially important to include function spaces since some applications of linear algebra in the calculus course are to function spaces, e.g., the space of solutions of a homogeneous linear differential equation. Next in a conceptual course one must introduce linear transformations and affine transformations (a transformation S is affine if the transformation T defined by $Tx = Sx - S0$ is linear). Of course one includes examples of these along with the definitions.

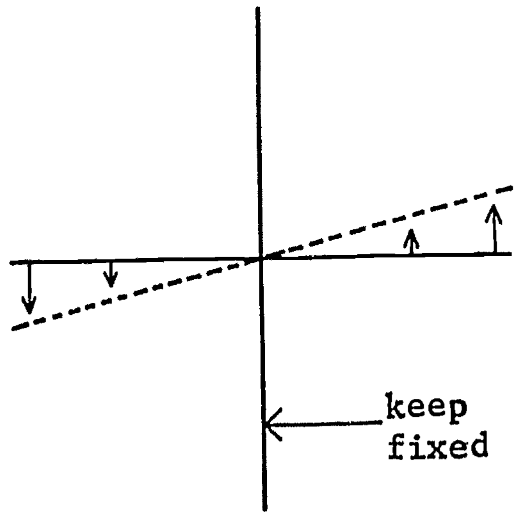
Then one turns to the analysis of the structure of a vector space by introducing linear subspaces, the concept of independence, bases for a subspace, dimension of a subspace, and the direct sum of subspaces. The work here ends with the theorem that a vector space of finite dimension, say k, is isomorphic to R^k. Thus one proves that a general (finite dimensional) vector space is really no more complicated than the examples you started with and furthermore is built up from copies of R by direct sums.

The next study is of the structure of linear transformations, and you first need a decomposition theorem. Any linear transformation $U \xrightarrow{T} V$, factors into a composite of a projection, an isomorphism, and an inclusion.

$$\begin{tikzcd}
U \arrow[r, "\text{projection}"', "\substack{U = U' + (\text{kernel } T)}"' {yshift=-12pt}] & U' \arrow[r, "\substack{\text{isomorphism} \\ V' = \text{image } T}"] & V' \arrow[r, "\substack{\text{inclusion} \\ V' \subseteq V}"'] & V
\end{tikzcd}$$

Projections and inclusions are simple geometrically, so if you feel that an isomorphism is essentially an identity then you understand <u>any</u> linear transformation and you don't need to feel baffled by the seeming complexity of linear transformations. You simply find the kernel and image, split U and V as direct sums, and the transformation is straightforward.

Of course this doesn't end the entire story, for although you can say U' is distinct from but essentially the same as V' there's a special case where it's clear they are not. Suppose you start with an automorphism of U. If you try to split this map the projection and inclusion parts are identities and the mapping you're trying to analyze <u>is</u> the isomorphism in the middle of the diagram. The spaces U' and V' are no longer distinct and identified by an isomorphism, but now the same. So now comes the analysis of automorphisms and, since it uses the same tools, the analysis of endomorphisms of a vector space. One needs examples and the first that are readily at hand are defined in terms of an inner product: orthogonal transformations (rotations and reflections), symmetric transformations (picture an orthogonal coordinate system and stretchings by perhaps different factors in the different directions), and skew-symmetric transformations (vector products with a fixed vector in R^3 is a good example). In addition to these I always enjoy presenting the shearing transformations, keep one coordinate axis in R^2 fixed and slew the other about the origin so that each vertical line slides along itself. Then one presents the structure theorems for symmetric and skew-symmetric transformations, i.e., you show that you can choose coordinates in such a way that these transformations have a simple prescribed form. Finally comes the analysis of a general automorphism, that it can be decomposed into a composite of an orthogonal transformation and a symmetric transformation. As details for this analysis one needs determinants, characteristic polynomials, characteristic values, and characteristic vectors.



This then is the outline of perhaps the most linear algebra one would put into a sophomore calculus course:

A. Vector spaces, the basic operations and their properties.

B. Examples. R^2, R^3, R^k, and function spaces.

C. Linear transformations and affine transformations including examples.

D. The structure of a vector space, subspaces, independence, bases, dimension, direct sum.

E. The analysis of linear transformations.

F. Examples of automorphisms and endomorphisms, orthogonal, symmetric, and skew-symmetric transformations, shears.

G. Structure theorems for symmetric and skew-symmetric transformations and the analysis of general automorphisms, including such details as determinants, characteristic polynomials, characteristic values, and characteristic vectors.

Now let me elaborate on my second criticism, that most linear algebra books confuse the presentation of theory and the technique of calculating via matrices. One way to see that a book makes this mistake is to see that it has a chapter on determinants and matrices before linear transformations are defined. The linear transformation is an easy conceptual thing to talk about and give examples of without matrices. The matrix is a tool for computation, i.e., it is a set of coordinates in a standard array in terms of the bases of U and V. That is, once bases have been selected in U and V so that the points in U and V have coordinates, then a linear transformation from U to V also has coordinates arranged in a rectangular array called a matrix. The matrix, important as it may be for computation, is of *no* importance in the theoretical or conceptual part of the course nor in the geometric pictures that come along. Presenting matrices before linear transformations and teaching students to work with them is comparable to trying to teach someone to play the piano on a keyboard that isn't attached to any strings. There's no feedback, the student does not see the objective and finds no pleasure in what he's doing. Now I agree that historically matrices came first, vector spaces weren't discovered in their abstract form. For many years a vector space was an R^k for some k and a linear transformation was a transformation given by a system of linear equations represented by the matrix of coefficients. Thus properties of linear transformations had to be formulated as properties of matrices. The abstract point of view, that one could proceed on a different level and work without coordinates, developed during the twenties and thirties. With this new point of view the picture became quite easy and lovely and the theory was disassociated from the mechanism of computation. Thus it is easy to see why the first books on linear algebra had to begin with determinants and matrices, but it seems to me that the conversion to the more recent and simpler view has been much too slow. I don't mind a historical presentation provided it's made clear to the student that matrices are not essential to understanding the theory and that the theory should not be confused with the computations which arise.

Another inadequacy of many texts is that the structure theorems for linear transformations are usually given only in the complex case. This case is easier and smoother because the characteristic polynomial factors into linear factors. For example, the classification theorem for general automorphisms is usually stated as: every automorphism is the product of a unitary and a Hermitian symmetric transformation. The details for the real case are omitted, but this is the case of interest at the level of second year calculus, so I think it should be included.

To establish the structure theorems for automorphisms, one must define the determinant, det T, of an endomorphism T. The question is how best to do this in a sophomore course in calculus. It may well be that one should be content with doing it only for dimensions at most 3. In such a case one would state the structure theorems in their n-dimensional form, and say that the proofs are to be given only for the first few dimensions.

There is little trouble in defining det T in dimension 3 once the scalar and cross products have been done. One defines the dot-cross (triple) product [A, B, C] of three vectors, shows that it is trilinear and alternating, that it is zero for dependent vectors, and that it gives the volume of the parallelepiped spanned by A, B, C with a sign when they are independent. Using the linearity of T and the above properties it is readily shown that the function [TA, TB, TC]/[A, B, C] of independent vectors A, B, C is a constant; this constant is called the determinant of T and we have

$$[TA, TB, TC] = (\det T)\,[A, B, C] \quad \text{for all } A, B, C.$$

An advantage of this approach is that det T takes on the geometric significance of being the ratio of the volume of the image of a parallelepiped to its original volume. Using the methods of integral calculus, this result extends to an arbitrary bounded domain D:

$$\text{volume (TD)} = (\det T)(\text{volume D}).$$

A mild disadvantage of this approach is that it makes the concept of determinant appear to depend on the choice of scalar product one is using. However if one has previously shown that any two n-dimensional spaces with scalar products are isometric, then there is an automorphism S of 3-space into itself which is an isometry of a given scalar product into a second, hence

$$\det_2 T = \frac{[TA, TB, TC]_2}{[A, B, C]_2} = \frac{[STA, STB, STC]_1}{[SA, SB, SC]_1}$$

$$= \det_1(STS^{-1}) = (\det_1 S)(\det_1 T)(\det_1 S^{-1}) = \det_1 T.$$

If one decides to do determinants in the n-dimensional case, there is much disagreement as to the best procedure. If you choose a basis and a representation of the endomorphism by a square matrix and then define determinant by a formula, you have a non-invariant definition, and you must prove properties of the determinant before you arrive at the conclusion that it doesn't depend on the chosen coordinate system. This explicit procedure for determinants, which makes you feel that you know what's going on or at least that you've got a hold on it, is not invariant in form, and rather lengthy arguments are needed to show that it is invariant.

There is an invariant approach that depends on introducing the tensor product of vector spaces, developing this algebraic operation, and extending it finally to the exterior algebra, $\Lambda^i V$, $i = 0,1,\ldots,n$ where $n = \dim V$. This is a graded algebra and $\Lambda^n V$ is one-dimensional and therefore an isomorphic copy of the ground field. Any automorphism of V induces an automorphism of the exterior algebra on V, so in particular it induces an automorphism of $\Lambda^n V$. The only automorphism of a one-dimensional space is multiplication by a scalar. The scalar that shows up here is called the determinant and gives you the invariant approach, or at least <u>an</u> invariant approach. I think that it's readily recognized that this is a bit too much for the usual sophomore course, so we need a substitute approach.

Another which is somewhat invariant in form is an extension to n dimensions of the definition in the 3-dimensional case given above. If V is an n-dimensional vector space we want a function of n vectors with scalar values such that the function is multi-linear (linear in each variable), skew-symmetric or alternating (if you switch two variables you switch the sign), and non-trivial (at least one non-zero value). The basic theorem about such functions is that they exist, moreover any two such functions differ at most by a scalar multiplier. [In response to questions the lecturer returned to this topic and outlined a proof in Lecture IV, the next lecture in these notes.] Given an endomorphism of V, let f be any one of these "volume" functions. Then the function g defined for n vectors, $x_1, x_2, \ldots, x_n$, by

$$g(x_1, \ldots, x_n) = f(Tx_1, Tx_2, \ldots, Tx_n)$$

is another volume function, so $g = \lambda f$ for some scalar λ. This scalar λ, the amount by which T alters volumes, is the determinant of T.

Discussion.

Hausner, Prenowitz, Steenrod, and Gray discussed the role that the algebra of n by n matrices should play in a linear algebra course. After discussion there seemed to be general agreement that the isomorphism between the ring of endomorphisms of V and this ring of matrices should be put on the same status as the isomorphism between V and R^n, that the students should be required in the exercises to use matrices for computations and relate their results to the underlying geometry, but that the distinction between linear transformations and matrices should be made very clear to the student. Hammer pointed out that several concepts for matrices, e.g., the strange multiplication and similar matrices, are explained by their geometric interpretation, and that the geometry of vectors and endomorphisms and the algebra of n-tuples and matrices nicely complement each other with each providing insights into the other.

Yale suggested using matrices as "shock therapy" at the beginning of a linear algebra course, and claimed that it is relatively easy to convince students that they won't get very far with matrix computations unless they understand their geometric background. As a typical device for this he suggested asking students to compute the sixteenth power of some matrix representing a simple geometric transformation in a poor coordinate system.

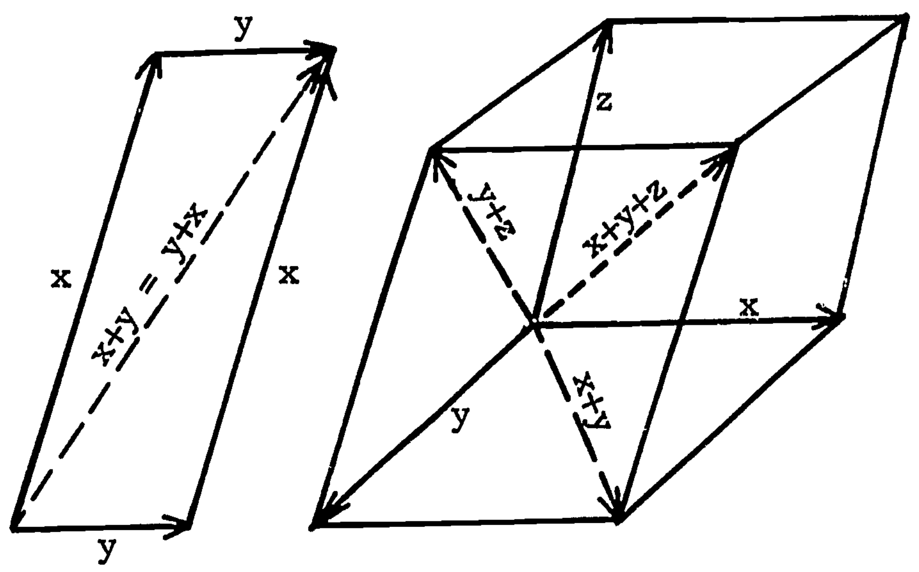
Keedy brought up the pedagogical difficulties in presenting such an abstract view of linear algebra to students who are only sophomores. Woolf stated that a course which is primarily a course in manipulating matrices may seem more efficient in the short run, the students seem to catch on to the material quicker, but that nothing sticks, a few months later the students don't remember a thing. Someone jokingly replied that nothing sticks in an abstract course either, not even the students. Gray remarked that at his school they willingly sacrificed a little on computations with matrices in order to instill a greater appreciation for and understanding of linearity. Prenowitz returned to Keedy's point and said that he felt that it would be much easier to present this type of linear algebra course if somewhere along the line, perhaps in high school, the student has a few weeks of classical vector analysis using arrows.

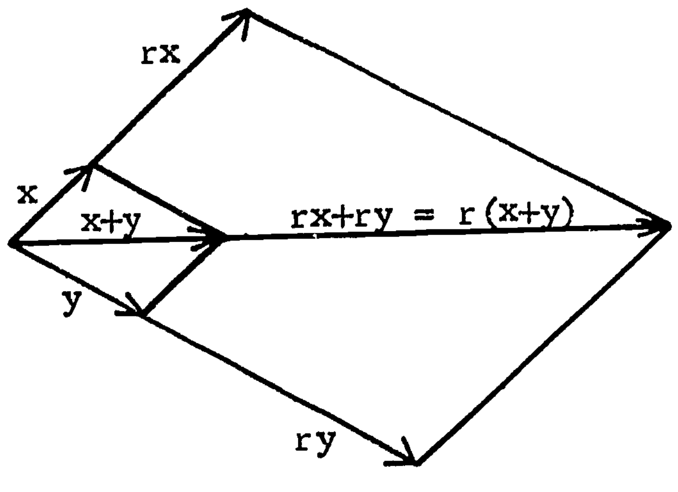
Coxeter made two suggestions concerning terminology and notation. A symmetric linear transformation could be quite naturally called a strain since, relative to a suitable coordinate system, it stretches the various coordinates by various ratios. A simple strain is a strain in which only one coordinate is stretched. He advocated using the right handed notation, $(x)T$ or x^T, instead of $T(x)$ since in the right handed notation the coordinates of a vector may be presented as a row vector rather than a slightly unnatural column vector. In this notation the coordinate symbols for the vector x and the linear transformation T go together in the same order as in xT, an order which seems more natural to him.

Prenowitz, Coxeter, and Johnson ended the discussion with a nice application of geometry in linear algebra. Every orthogonal transformation of V is the product of at most n reflections if n is the dimension of V. To see this note that an orthogonal transformation preserves distance and is determined by its action on an n-simplex with one vertex at the origin. The n vertices not at the origin can be mapped to their desired positions by a sequence of n suitably chosen reflections.

Lecture IV. Linear Algebra and Calculus of Several Variables.

I would like to look again at two points I raised in yesterday's lecture. I spoke of linear algebra books not presenting the geometric side with sufficient detail, either omitting it or treating it in an offhand fashion. Let me elaborate on this point. Most books are quite good about presenting pictures for the basic operations on vectors. They define addition of vectors in R^2 or R^3 in terms of components and then give the picture for the parallelogram rule. Similarly when they define multiplication by a scalar they present the corresponding picture of a coordinate line with the vector as the unit segment. However when they present the basic properties of the operations they usually omit the corresponding pictures. The commutative law, which one verifies algebraically in R^k by looking at components, has a very simple geometric picture associated with it. Similarly the associative law is verified algebraically and can be pictured using a parallelepiped. View the main diagonal as the sum of the diagonal of a face plus the remaining edge. The distributive law, $r(x+y) = rx + ry$, amo' its geometrically to the fact that multiplication by the scalar r is a similarity. It takes a bit of time to draw





these pictures, but by the time you've done so the algebraic formalism has taken on a geometric shape in the student's mind.

To make this clear let me take an even more trivial example, the rule of signs in arithmetic, $-(-x) = x$. I remember my son coming to me with the argument that since putting a minus sign in front of 2 makes it negative, putting another minus sign in front ought to make it still more negative. He had in mind the geometric picture of moving a point on the number line from right to left (positive to negative). His picture was inadequate because he was looking at the operation on one point at a time. Regarded as a transformation on all positive numbers, the minus sign doesn't shift all numbers uniformly to the left, instead it rotates the positive half line through 180°, pivoting at zero. Once this is seen there is only one natural extension of the operation to the negative half of the line. With this picture attached to the rule of signs, how can anyone forget the rule or be uncertain of it. In my son's class the reason given for the rule of signs was to preserve the distributive law. This is a valid mathematical reason but has little impact on a seventh grader. Distributivity means that the operation is linear; in R^1, a linear mapping is a similarity; and multiplication by -1 is in fact rigid. Surely it is better to appeal to the need to preserve rigidity than to the need to preserve distributivity.

There were several questions about the method of developing determinants that I mentioned briefly last time, so perhaps I should elaborate on it a bit. To motivate what is done in n-dimensions I suggest doing first the 3-dimensional case using the dot-cross product as described in the preceding lecture. You then ask if this can be extended to higher dimensions, that is, can the triple product be extended to an n-tuple product? First we define what is meant by an n-dimensional volume, namely a scalar valued function of n vectors in R^n that is multi-linear, skew-symmetric, and non-trivial. The objective then is to give an inductive argument showing that such a function exists and is essentially unique, i.e., unique up to a scalar multiplier.

To do this split your n-dimensional space V into the direct sum of a one-dimensional space and an (n-1)-dimensional space, V'. Let b denote a base vector of the one-dimensional space, then each vector x in V can be written uniquely as a certain scalar multiple of b plus its projection x' in V', $x = mb + x'$. Now let's assume we have a volume function f on V and are in the process of proving uniqueness inductively. Given n vectors, $x_1,x_2,\ldots,x_n$, in V we split each of them as above in the form $x_i = m_i b + x_i'$. Expanding $f(x_1,x_2,\ldots,x_n) = f(m_1 b + x_1',\ldots,m_n b + x_n')$ by multi-linearity we get 2^n terms most of which are zero, for it's easy to show that multi-linearity and skew-symmetry combined imply that $f(y_1,y_2,\ldots,y_n)$ is zero whenever a vector is repeated or, more generally, whenever the vectors are linearly dependent. Thus

$$\begin{aligned} f(x_1,\ldots,x_n) &= f(m_1 b + x_1',\ldots,m_n b + x_n') \\ &= \Sigma_{k=1}^{n}\ m_k\ f(x_1',\ldots,b,\ldots,x_n') \\ &\qquad\qquad \text{(b in the } k^{\underline{th}} \text{ place)} \\ &= \Sigma_{k=1}^{n}\ m_k(-1)^{k-1}\ f(b,x_1',\ldots,\cancel{x_k'},\ldots,x_n') \\ &\qquad\qquad\qquad (x_k' \text{ deleted}). \end{aligned}$$

Now define a function g of $n-1$ vectors in V' by

$$g(y_1,y_2,\ldots,y_{n-1}) = f(b,y_1,y_2,\ldots,y_{n-1}),$$

and then we have

$$f(x_1,\ldots,x_n) = \Sigma_{k=1}^{n}\ m_k(-1)^{k-1}\ g(x_1',\ldots,\cancel{x_k'},\ldots,x_n').$$

It is straightforward to verify that g is a "volume function" on V'. If f' is another volume function on V then expand it in the same way in terms of the analogous (n-1)-dimensional volume function g'. By the inductive hypothesis, $g' = rg$ for some scalar multiplier r, and hence $f' = rf$, so any two n-dimensional volume functions are scalar multiples of each other.

The same equation indicates how one should build an n-dimensional volume function, f, from an (n-1)-dimensional volume function g. The details of the existence proof are then straightforward.

In spite of the fact that one can start this induction from dimension one or two, I feel that in a calculus course it is valuable to develop volume in R^3 and to use the cross product in doing this.

Having completed my comments on linear algebra, I shall discuss now its applications to the calculus of several variables. Consider again the idea of the velocity vector. Let $f\colon [a,b] \to R^3$ define a curve, and let $x = f(t)$ and $x + \Delta x = f(t + \Delta t)$, i.e., x and Δx are vectors and t and Δt are scalars such that both t and $t + \Delta t$ are in (a,b). Suppose the velocity vector, $\frac{dx}{dt} = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t}$, exists and is not zero, then, since radial projection is continuous, $\lim_{\Delta t \to 0} \frac{\Delta x}{|\Delta x|}$ will also exist. Similarly, since taking lengths of vectors is continuous, if $\frac{dx}{dt}$ exists, so will $\lim_{\Delta t \to 0} |\frac{\Delta x}{\Delta t}| = |\frac{dx}{dt}|$. This leads to a geometric view of the unit tangent vector and scalar velocity (speed) without any mention of arc length, i.e., we have a geometric view of $\frac{dx}{dt} = \frac{dx}{ds}\,\frac{ds}{dt}$ without knowing anything about s.

This brings up the difficult question of how to handle arc length on curves. I looked at several calculus books and observed that most of them don't handle it. Apostol, in the second edition of his book, defines arc length as the least upper bound of lengths of inscribed polygons and proves that, if the curve is smooth, arc length is the integral of the speed. There is an alternative procedure which I've used and that I'd like to present for you to consider. Assume we're dealing with a smooth, or at least piecewise smooth, curve so $\frac{dx}{dt}$ exists at all points except perhaps a finite number of points. Also assume that $\frac{dx}{dt} \neq 0$ at any point. Then you simply define the arc length, $s(t)$, from a to t to be $\int_a^t \left|\frac{dx}{d\tau}\right| d\tau$. It's not fair to do this unless you verify that this extends the common notion of arc length, at least to the extent that the common notion has been clearly defined, say for straight lines and circles. It does agree in these two cases, it is clearly additive, so it observes the basic properties that arc length should have. A formal proof that $\frac{ds}{dt} = \left|\frac{dx}{dt}\right|$ comes immediately from this definition and the fundamental theorem of the calculus. To show that $\frac{dx}{dt} = \frac{dx}{ds}\frac{ds}{dt}$ we use the fact that $s(t)$ is strictly increasing and hence has an inverse function. It takes about one half of a page to complete the argument. The only objection I see to this approach from the geometrical point of view is that it avoids limits of inscribed polygons to which we're somehow wedded as human beings.

In this same direction I never neglect to develop the standard decomposition of the acceleration vector,

$$\frac{d^2x}{dt^2} = \frac{dx}{ds}\left(\frac{d^2s}{dt^2}\right) + (\text{normal})\left(\frac{ds}{dt}\right)^2 \varkappa ,$$

i.e., the acceleration vector is the sum of the tangential and centripetal accelerations. By this time most students have taken first year physics and recognize the normal component as the scalar velocity squared over the radius of curvature, in other words, as the "circular centripetal force."

The chief bugaboo in a course on the calculus of several variables is the chain rule. If one uses the Fréchet definition of the derivative this bugaboo tends to disappear. Consider first the case of a scalar valued function, say $f: D \to R$, D a domain in R^3. Define $f'(x,y)$, the derivative of f at x with respect to the vecto. y, as the following limit, if it exists:

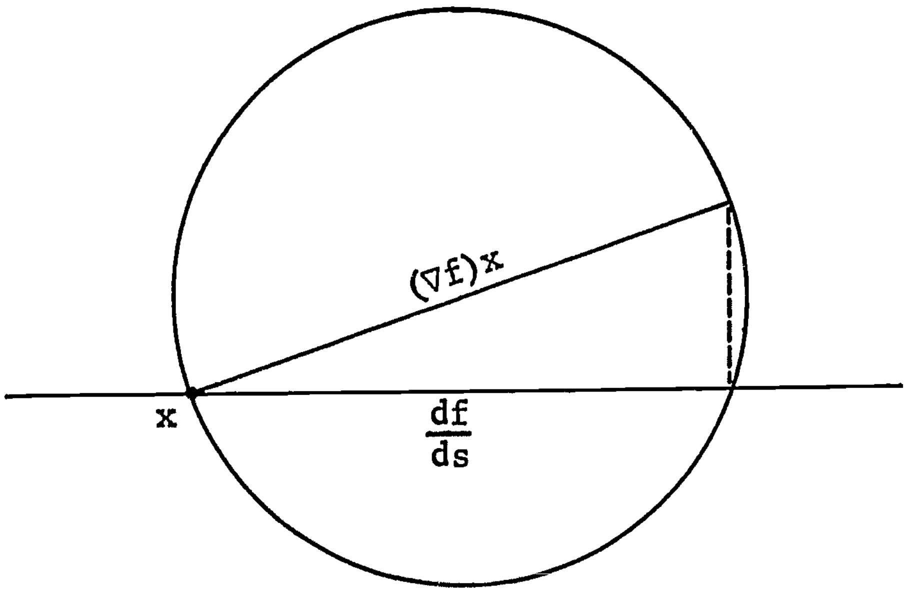
$$f'(x,y) = \lim_{h \to 0} \frac{f(x + hy) - f(x)}{h} .$$

For sufficiently small h, $x + hy$ will be in D, and essentially you are taking a derivative along a line through x using as your unit segment the vector y. This is an invariant definition which is independent of the coordinate system and uses only the linear structure of the space. The theorem one has to prove is that if $f'(x,y)$ exists for each y and is continuous in x for each y, then $f'(x,y)$ is linear in y. This linearity leads to the formula for $f'(x,y)$ in terms of a basis, <u>any</u> basis--it need not be orthonormal. If $A_1,\ldots,A_n$ is a basis and $y = \Sigma\, y_i A_i$ then

$$f'(x,y) = f'(x,A_1)\, y_1 + \cdots + f'(x,A_n)\, y_n .$$

Thus, if you know the particular derivatives, $f'(x,A_j)$, referred to as the <u>partial derivatives</u>, $\frac{\partial f}{\partial x^j}$, you can recapture all derivatives.

In this result we have the ingredients of the chain rule, for $f'(x,y)$ is the scalar product of the partial derivatives and the components of the vector y. If we now assume that the basis is orthonormal and use the fact that any linear functional (scalar valued linear function) on the space is given by the inner product of y with some fixed vector then we can write $f'(x,y)$ as the inner product of y with a fixed vector called $(\nabla f)x$, the gradient of f at x. This defines the gradient vector and one has the nice geometric theorem that the directional derivative, $\frac{df}{ds}$, in any direction is the projection of the gradient on that direction. If you draw the sphere with diameter $(\nabla f)x$, you can read off $\frac{df}{ds}$ simply by looking at the intersection of the sphere and your directional line. This picture is frequently omitted from calculus books but I find it very attractive.



When you change from scalar valued functions to functions from m-dimensions to n-dimensions say $f: D \rightarrow W$, D a domain in V, V and W vector spaces of dimension m and n respectively, then you can take the same definition of the Fréchet derivative, but now $f'(x,y)$ is in W instead of R. The happy thing about this is that practically no new analysis is required to handle this apparently much more complicated derivative than has already been done in the first case. One simply chooses a basis in W, splits W into one-dimensional pieces and takes the components of $f'(x,y)$. Each of these components is a derivative of a scalar valued function, so you've unified the concept of derivative. The same theorem (existence and continuity in x of $f'(x,y)$ for each y implies linearity of $f'(x,y)$ in y) holds in this case, just look at the components. Thus the mapping $f'(x)$ which sends y to $f'(x,y)$ is a linear transformation of V into W, or, to put it another way, the derivative f' is a mapping of D into $L(V,W)$, the space of linear transformations of V into W. In the old language, to each point of the domain of f there is assigned a Jacobian matrix at that point.

When teaching the topic of Taylor's expansion of $f: D \rightarrow R^n$ near a point

$x_0 \in D$, one can justify the statement "$f'(x_0)$ is the best linear approximation to f in a small neighborhood of x_0." Let g denote the first two terms of the expansion, i.e.,

$$g(x) = f(x_0) + f'(x_0)(x-x_0).$$

This is an affine mapping, and the remainder theorem yields

$$\lim_{x\to x_0} \frac{f(x) - g(x)}{|x - x_0|} = 0 .$$

It is readily checked that this holds for no other affine mapping.

The chain rule for the composition of $f: R^k \to R^m$ and $g: R^m \to R^n$ takes on the same form as for functions of one variable

$$(gf)'x = (g'(fx))(f'x),$$

and its geometric interpretation is most pleasing: the best linear approximation to the composition gf at a point x is the composition of the best linear approximation $f'x$ to f at x and the best linear approximation $g'(fx)$ to g at fx. Representing g' and f' by their Jacobian matrices, and recalling that multiplication of matrices corresponds to composition of linear transformations, we obtain the standard formulas

$$\frac{\partial z^i}{\partial x^k} = \Sigma_j \frac{\partial z^i}{\partial y^j} \frac{\partial y^j}{\partial x^k}$$

where $y = f(x)$ and $z = g(y) = g(f(x))$.

Turning to the special case of a vector field f in a domain $D \subset R^n$, we have $f: D \to R^n$ and $f': D \to L(R^n, R^n)$; hence $f'(x)$ for each $x \in D$ is an endomorphism of R^n. Now you can show a profit if you've covered the linear algebra in some detail, because you can now define the curl and divergence of f in an invariant fashion. The curl of f at x is simply one-half of the skew-symmetric part of $f'(x)$ and the divergence of f at x is the trace of $f'(x)$ (and also the trace of the symmetric part of $f'(x)$).

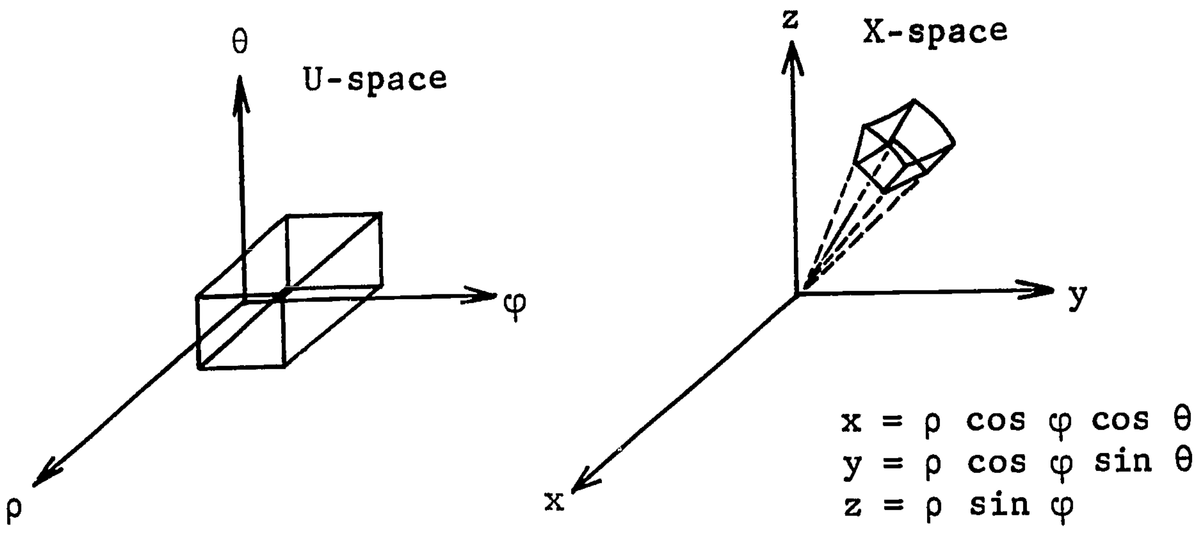
Thus the curl and divergence are simply special aspects of the general derivative.

An associated bugaboo has to do with the transformation of multiple integrals where you throw in the Jacobian determinant. For straight volume problems, for example, the standard way of writing the formula is

$$\underset{\text{U-region}}{\int\int\int} \frac{\partial(x^1x^2x^3)}{\partial(u^1u^2u^3)}\, du^1\, du^2\, du^3 \quad = \quad \underset{\substack{\text{Image in X-space}\\ \text{of the U-region}}}{\int\int\int} dx^1\, dx^2\, dx^3.$$

Now I'm not advocating that in a sophomore calculus course you prove this theorem in its full generality. The general case is much too involved, we don't have time for its proof in a calculus course, there are too many other things we must cover. But I do believe you should state the theorem, show why it is reasonable (volume $T(U) = (\det T)$ volume U), and then prove the particular cases that you want to use, especially polar, cylindrical, and spherical coordinates.

Let's review what this involves in the case of spherical coordinates. First the Jacobian of the transformation is computed and found to be $\rho^2 \cos \varphi$. Take U to be a rectangular box with sides parallel to the coordinate axes, with one corner at ρ, θ, φ and edges $\Delta\rho, \Delta\theta, \Delta\varphi$. Integrating the Jacobian over U gives $(1/3)[(\rho+\Delta\rho)^3 - \rho^3][\sin(\varphi+\Delta\varphi) - \sin \varphi]\Delta\theta$. We compute now the volume of the image V of U in (x,y,z)-space. This is an elementary solid and its volume can be worked out by the methods of the calculus. A cone with a spherical cap is a solid of revolution; if ρ is the length of the generator of the cone and $\frac{\pi}{2} - \varphi$ is the generating angle, its volume is $(2/3)\pi\rho^3(1-\sin \varphi)$. Its intersection with a wedge of angle $\Delta\theta$ has volume $(1/3)\rho^3(1-\sin \varphi)\Delta\theta$. Form now the difference of two of these taking ρ to be $\rho + \Delta\rho$ and ρ respectively, and then a second difference



$$x = \rho \cos \varphi \cos \theta$$
$$y = \rho \cos \varphi \sin \theta$$
$$z = \rho \sin \varphi$$

taking φ to be $\varphi + \Delta\varphi$ and φ respectively. The result coincides with the integral of the Jacobian over U.

This makes a pleasant exercise in the calculus, it can even be assigned as homework, and once it is done the student understands the meaning of the general case.

Discussion.

Woolf stated that he was a little uncomfortable about using a definition of arc length that is restricted to smooth curves. In reply Steenrod elaborated on one of the points he was trying to make in the lecture. Although in a sophomore calculus course we do not have time to present or prove results in their full generality we should prove them in the special cases that exhibit enough generality for the moment. For example, only a few mathematicians have worked through the details of the general theory of surface area; however, we are happy to accept the standard formula for smooth surfaces as a definition because it gives the right answers in all those special cases where the area is determined by some simpler method (flat regions, cones, cylinders, surfaces of revolution).

Klamkin and Hammer suggested that one ought to show the students that the Jacobian for rotations and translations is 1, in agreement with our intuition that rigid motions don't change volume, and verify that volumes behave the way we claim they do in the case of a linear change of coordinates. In the case of a linear change of coordinates the Jacobian is just the determinant of the transformation.

Yale pointed out that one can exploit the chain rule to simplify the computation of the Jacobian for changing from rectangular to spherical coordinates and in showing that volumes behave the way they should. The equations relating rectangular and cylindrical coordinates, $x = r \cos \theta$, $y = r \sin \theta$, $z = z$, are the same in form.as the equations relating cylindrical and spherical coordinates, $r = \rho \cos \varphi$, $z = \rho \sin \varphi$, $\theta = \theta$. Thus if one computes the Jacobian and verifies the volume relationship in changing from rectangular to cylindrical coordintes, the same computation can be used again along with the chain rule to yield the appropriate results for spherical coordinates.

In reply to Woolf's question as to whether the mapping f' from D into $L(V,W)$ should be called the derivative or differential Steenrod promised to talk further on the difference between the two in his next lecture.

Lecture V. Differential Equations.

Let me reply first to the question of last time about differentials versus derivatives. Suppose V and W are vector spaces, D is open in V, and f is a continuously differentiable function mapping D into W. We saw yesterday that the derivative, f', is a map of D into $L(V,W)$, the space of linear transformations of V into W. One could repeat this process and take the second derivative, f'', a mapping of D into $L(V, L(V,W))$, but, if you look at things this way, higher derivatives get more and more awkward to write. So one goes back and examines $L(V,W)$ in greater detail. Using the natural isomorphism between W and $R \otimes W$ and the fact that $L(V,R) = V^*$, we find that

$$L(V,W) \cong L(V, R \otimes W) \cong L(V,R) \otimes W \cong V^* \otimes W.$$

Thus f' can be regarded as a map of D into $V^* \otimes W$, f'' maps D into $V^* \otimes V^* \otimes W$, etc.

One is tempted to regard these successive derivatives as tensor functions on D, but this proves to be wrong. If we change coordinates in D, transform f' by the tensor transformation rule, and then take the derivative, we do not obtain the tensor transform of f''; second derivatives of the coordinate transformation appear in the first but not the second. The derivative somehow isn't a tensor in the proper sense. However a piece of the derivative, its skew-symmetric, or alternating, part does transform properly. These skew-symmetric tensors are called forms. A p-form on D is a function mapping D into the p^{th} component of the exterior algebra generated by V^*, $f: D \to \Lambda^p V^*$. If you take the derivative then, as we saw above with $W = \Lambda^p V^*$, we obtain a mapping f' of D into $V^* \otimes \Lambda^p V^*$. There is a natural quotient mapping (derived from the basic homomorphism of the tensor algebra of V^* onto the exterior algebra of V^*) of $V^* \otimes \Lambda^p V^*$ onto $\Lambda^{p+1}V^*$. The <u>exterior derivative</u> or differential of f is the composite of f' and this natural homomorphism, so whenever this quotient map actually collapses things the differential, df, is only an aspect of the derivative, f', and is not the same. But in the special case $p = 0$, i.e., $f: D \to R$ we have $\Lambda^1 V^* = V^*$, so no collapsing occurs and for scalar functions there is no difference between the two. There is a difference whenever $p > 0$. The gradient, curl and divergence are modifications of the exterior derivative, and the relations curl grad = 0, div curl = 0 correspond to $dd\varphi = 0$ for all φ.

A number of the things I've discussed in the last two lectures are done in some detail in the text on advanced calculus by Nickerson, Spencer, and Steenrod. This set of lithoprinted notes, was prepared in 1958 for a course at Princeton combining advanced calculus and complex variables. The first eight chapters were written and revised several times but chapters 9-13 were hastily prepared with the intent of polishing them later. In 1962 we split this course into two courses, a year's course on advanced calculus and differential geometry, and a year's course on complex variables. At the same time linear algebra was made a prerequisite, so mathematics majors are now required to take linear algebra in the sophomore year. In fact most of the material of the first eight chapters is covered now during the sophomore year. When I teach the advanced calculus and differential geometry course, we begin with chapter 9, and struggle on from there. Chapter 11 on the tensor and exterior calculus has been thoroughly revised and is available in mimeographed form. Course structures change so rapidly these days that a would-be author is easily deterred by the thought that any book he writes may soon be obsolete.

Let us turn now to a discussion of the sophomore course on differential equations. Most books I've seen on the subject begin by presenting examples of differential equations in the form of formulas. This gives the student the impression that a differential equation is just something you write on piece of paper, that it's a batch of symbols with derivatives in it, or more formally that it's a function of several variables some of which are derivatives with all this equated to zero. This is the analytic definiton. Although such a definition can be made rigorous, the approach suggests to the student that it isn't a differential equation unless and until it is written as a formula in standard notation.

It seems to me that a geometric view of differential equations can clarify the whole picture and give the student some feeling for what's going on right from the start. Suppose you have a domain D in a vector space V and a continuous mapping $f: D \to V$, i.e., a continuous vector field in D. A solution curve through a point x in D is a function, say $y = g(x,t)$, of the initial position x and time t such that $g(x,0) = x$ and $\frac{dy}{dt} = f(y)$. The equation $\frac{dy}{dt} = f(y)$ is the associated differential equation. Without proving anything at this introductory stage you can state the main facts. First of course is the existence theorem, for any x in D there is a time interval around zero and a function g such that $y = g(x,t)$ is a solution for t in this interval. The uniqueness theorem requires some condition of the Lipschitz type and says that any two solutions through x must coincide. Then you can state a group property of solutions: $g(g(x,t),s) = g(x,s + t)$. In other words, if you start at the point x, move along a solution curve through x for a time t, then move along a solution curve through the point $g(x,t)$ for a time s, you arrive at the point on the original solution curve that corresponds to the time $t + s$. Thus the two solution curves are actually the same curve and you arrive at the concept of a streamline. Another way of saying this is that a curve C followed by a point x is also the curve followed by each point of C; thus C slides or flows along itself. A point moving along C will do so with varying velocities but when it reaches x then its velocity is the prescribed velocity f(x). The picture of these streamlines filling D, one through each point, is called a steady flow.

One can give examples of steady flows on a very elementary level, examples which are very appealing. The simplest is of a constant field, $\frac{dy}{dt} = b$, b a fixed vector. In this field all vectors are parallel and have the same length, and of course the flow is simply translation with streamlines $y = x + bt$. Or consider $\frac{dy}{dt} = ry$, r a scalar, the differential equation associated with a radial field in which all vectors point away from the origin $(r > 0)$ or towards the origin $(r < 0)$. The streamlines are clearly straight lines through the origin and you can explicitly check that the solutions are $y = e^{rt}x$. Another nice example in R^3 is that of a rotating field. Choose an axis of rotation, or axis of fixed points, say the line along the vector b. Define the vector field by $f(y) = b \times y$, so that f(y) is perpendicular to both b and y. You can integrate the associated differential equation $\frac{dy}{dt} = b \times y$ explicitly in terms of sines and cosines and see that the steady flow is circular motion about the axis with angular velocity $|b|$.

It seems to me that this can be done very early in the course, say the first two or three weeks. You may feel that at the beginning you must start with mechanics, i.e., solving simple equations so as to get the students started on homework problems. But actually you can make up homework problems on this kind of material--it's not difficult.

Of course steady flow is not the whole story but it is a unifying geometric view of the subject. One needs of course the reduction theorem stating that any differential equation in the analytic sense

$$\frac{d^k y}{dx^k} = f(x,y,\frac{dy}{dy},\ldots,\frac{d^{k-1}y}{dx^{k-1}})$$

can be reduced to a steady flow by introducing new variables $y_i = \frac{d^i y}{dx^i}$ for $i = 1,\ldots,k-1$. This gives a system of first order equations which can be written as a single first order equation $\frac{dz}{dt} = F(z)$ where z is the vector whose components are $(x,y,y_1,\ldots,y_{k-1})$ and the components of $F(z)$ are $(1,y_1,\ldots,y_{k-1},f(x,y_1,\ldots,y_{k-1}))$. The only drawback with this view is that the resulting steady flow is frequently in a dimension greater than 3 where visualization is not granted to most mortals.

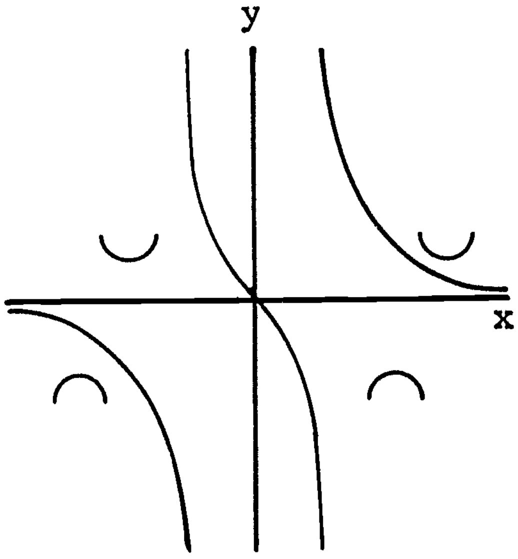
Another reason that I like the steady flow picture is that it makes it easy to explain to a class the geometric background of the game of making substitutions in a differential equation in order to reduce it to a form that is readily integrable. One is simply transforming the vector field and resulting flow to a new coordinate system, chosen with the hope that the solution curves will have a simpler and recognizable structure. To convince the student that this can be done ask him to picture a coordinate system in which the streamlines are the "lines parallel to one axis." The associated differential equation in this new coordinate system is simply $\frac{dy}{dt} = b$, i.e., in this coordinate system the flow is translation and the solution is trivial. Thus the problem of solving a differential equation and that of finding a suitable coordinate transformation are essentially equivalent.

Incidentally, while on the subject of coordinate transformation, I believe that one should always display the domain and range of the transformation as two separate rectangular coordinate systems. The image of the domain gridwork is pictured as two families of curves in the range space, and the inverse image of the range gridwork as two families of curves in the domain. The purpose of doing it this way is to emphasize the mapping aspect of the transformation. I do this even when discussing such a simple case as polar coordinates.

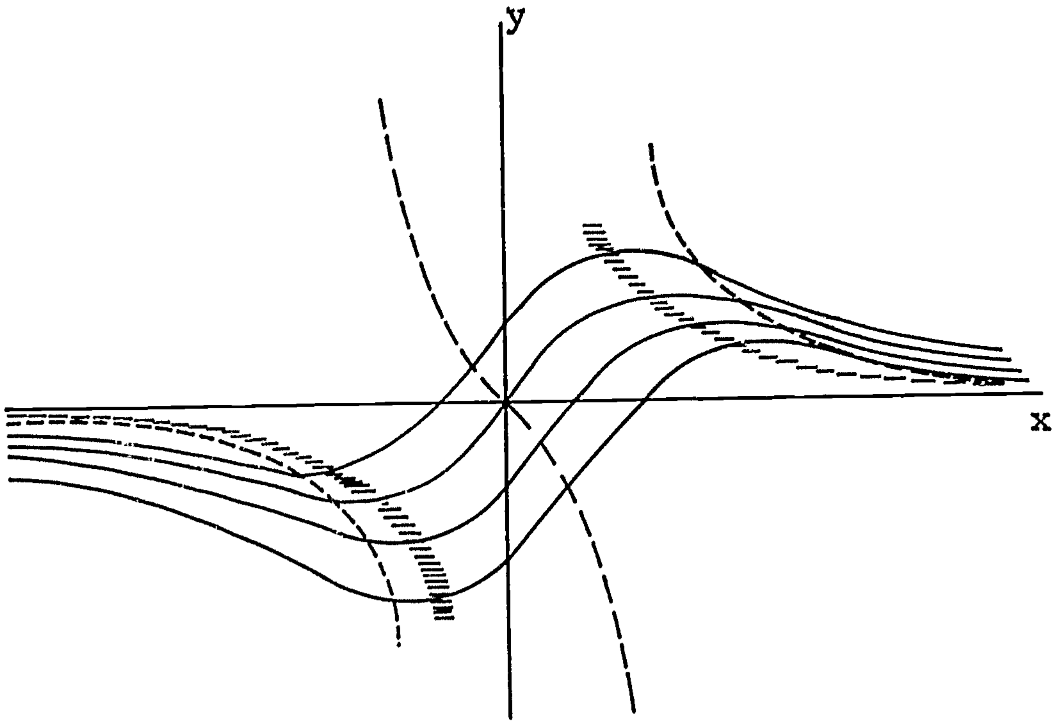
Now let me list some topics in a differential equations course which are especially geometric in nature.

1. Sketching a steady flow from its vector field.
2. Finding orthogonal trajectories.
3. The two body problem.
4. Linear systems of first order equations with constant coefficients.

The first topic is treated perfunctorily in most books, often with just a few comments about drawing the line element field and then trying to connect these line elements to form smooth curves. One exception is the old book by Piaggio which takes this up in the very first chapter. He presents details for sketching the whole family of streamlines before the student learns how to turn the crank. The main features of his procedure are: First sketch the curves of zero slope (the isoclines for zero slope), i.e., the curves of maxima and minima for the streamlines. In addition sketch the curves of inflection points, the curves $y'' = 0$. About half the people I talk to have never tried this second point, so let me give you an example. The differential equation $y' = -xy + 1$ is linear in y so you can write down the solution in terms of exponentials, however this does not give much information about the shape of the solution curves. The curve of zero slope is $xy = 1$. Differentiating we obtain $y'' = 0$ if and only if $y = \frac{x}{x^2-1}$. This curve and the convexity of the solution curves in the four regions determined by its three branches are indicated in the following sketch. With this added bit of detail the student



finds the sketching of the solution curves much more interesting.



Stream lines for $y' = -xy + 1$, showing the zero isocline and the curve of inflection points

The second topic, orthogonal trajectories, is treated adequately in most books so I won't say anything about it except that I always include it in the course. It's pretty geometry and has practical importance, e.g., lines of force versus equipotential curves and surfaces.

The two body problem is a wonderful problem from several points of view.

Consider the steps you go through to solve it. You start with any rectangular coordinate system in R^3, set up the differential equation in vector form and you have an equation with a total order of twelve. You then observe a certain kind of symmetry in the equation, do some adding and subtracting and come up with an equation for the center of mass. This is a simple equation that you can integrate and find as a result that the motion of the center of mass is a translation. Now you change to a new coordinate system whose origin is at the center of mass. You can thus assume the center of mass is fixed and ask for the motion of each particle with respect to the center of mass. You find that the differential equations for the two particles not only look alike but that the solution for one determines the solution for the other, so you've reduced the problem to the problem of one body attracted to a fixed center. The differential equation for this problem has total order six. Many books leave out these initial steps. I don't understand why they do this; it is part of the problem, and it is very nice mathematics.

You now go to work on the one body problem. The first integration is manageable in rectangular coordinates and tells you that the motion is planar. It also gives Kepler's law about equal areas. At this point you transform to polar coordinates in the plane of motion, eliminate the t parameter, and obtain a differential equation in r and θ which is linear in $1/r$. This second order linear equation can be integrated explicitly and out comes the equation of a conic section in polar form. If you've done a good job on conics, the students will be delighted to get the answer in such nice form. If you haven't, they will be mystified, and you will need to transform into rectangular coordinates to show that (if the initial velocities are not too great) the motion is along an ellipse with one focus at the center of mass. Then, using the area of the ellipse and Kepler's first law, one derives his third law.

This problem illustrates almost everything you do in a differential equations course, especially the techniques of choosing coordinates properly to simplify a problem, and making coordinate transformations to obtain explicit solutions. In addition to all this it has great cultural value.

Finally, a few words about the last topic, systems of linear differential equations of the form $\frac{dy}{dt} = Ay$, where A is a linear transformation. The studen^ts have already solved single linear equations with constant coefficients and this first order system is essentially a trivial problem compared to their previous problems. It's just in higher dimensions and now you do something about the geometry of the topic that you didn't do before. The solution can be given explicitly in the perfectly good form $y = e^{At}x$, read "y is e^{At} acting on x." Of course you have to define the exponential e^{At} in the standard way,

$$e^{At} = \Sigma_{k=0}^{\infty} \frac{1}{k!} A^k t^k .$$

Note that I view A as a linear transformation not a matrix. You can prove that this converges, without resorting to matrices, by using norms in the space of linear transformations. Since At and As commute, $e^{A(t+s)} = e^{At}e^{As}$, so the group property we spoke of earlier works out nicely.

After doing this one looks at the two dimensional case in detail. The origin is a fixed point and the question is, how do the solution curves behave in a neighborhood of the origin. It turns out that there are four types of behavior. To solve such a problem, one must determine the characteristic values and vectors of A. The latter give the straight lines that are streamlines, and the characteristic values determine the nature of the neighboring curves. This is a pretty bit of geometry which fits in neatly if you've already done the linear algebra. It gives you a fine opportunity to reestablish the idea of a steady flow. There are good treatments of this topic in the differential equations books by Kaplan and Ross. (It is sometimes listed as "phase-plane analysis.") Some books fail to include it.

Discussion.

Woolf and Hausner asked about 0-forms and whether vector-valued functions could be viewed as 0-forms. Steenrod replied that there is a generalization of the notion of differential forms called vector-valued differential forms. If V and W are vector spaces, then a vector-valued differential form of degree p defined on a domain $D \subset V$ with values in W is a function

$$\varphi : D \to (\Lambda^p V^*) \otimes W.$$

Its differential $d\varphi$ is a vector-valued $(p+1)$-form. In particular, a vector-valued 0-form is a function $\varphi : D \to W$, and $\varphi' = d\varphi : D \to V^* \otimes W$. One can multiply an ordinary p-form on D with a vector-valued q-form to obtain a vector-valued $(p+q)$-form. In this way, the vector-valued forms constitute a module over the algebra of ordinary forms.

Klamkin remarked that the radius of curvature is another nice device to help sketch a steady flow, an idea that he believes goes back to Rayleigh. The vector field for the radius of curvature at each point is readily obtained for a first order equation and can even be approximated for second order equations by approximating the derivative at each point.

The book by Horace Lamb and the American Mathematical Monthly were suggested as good sources of ideas and pictures to help visualize solutions to differential equations.

Benson initiated a long discussion of conics by observing that the two-body problem might serve as a good motivation for the study of conics and asking what other motivating ideas could be used earlier. Steenrod suggested two: analytically they're next in line after the study of straight lines, or one can look at circles and tilt them or project them to motivate ellipses. He repeated his view as stated in an earlier lecture that it is important to have the conics (at least in standard position) available early in the freshman course because they provide nice applications of the derivative. Blattner reported that his students became more excited about conics when he presented a brief treatment using Dandelion spheres as in the book by Hilbert and Cohn-Vossen. Prenowitz, Kelly, and Hausner stressed the point that the teacher's attitude is very important. Topics such as conics may seem "old hat" to us but we must be careful not to take away their natural appeal. We should be sure to give the student some idea of the tremendous amount of work behind the polished form presented. Kelly remarked that the focus-directrix definition provides a unified way of presenting all three types of conics. He showed how the variation of eccentricity, $e(x) = \left|1 - \frac{f}{x}\right|$ (assuming the y-axis is the directrix, $(f,0)$ is the focus) in terms of the x-intercept can be exploited in the early study of conics.

Kelly held up as an example of a poor attitude on the part of a teacher an anonymous algebraist who was heard to mutter, "I've got to go teach volumes by slicing. That's really a topic for butchers!" This example reminded Steenrod of one topic, how to teach students to draw good pictures, that he wanted to cover in his lectures but couldn't for lack of time. The unhappy algebraist may have been unhappy because he lacked the ability to draw pictures. If you're good at drawing pictures, then teaching such topics as volumes by slicing is enjoyable. The first place where a good picture in three dimensions is needed occurs in the topic of direction angles and cosines. Given the coordinates of two points, the problem is to draw the box with sides parallel to the coordinate planes having the two points as diagonal vertices. There is a simple mechanical technique, draw first the projection in the xy plane, erect the vertical edges and fill in the top and bottom. Techniques such as this ought to be included in calculus and analytic geometry books. Gray stated that perhaps we need a pamphlet for instructors on how to draw good pictures for calculus and linear algebra courses. Vogt pointed out that Klein, in the geometry volume of his book *Elementary Mathematics from an Advanced Standpoint*, discusses the problem of making correct drawings, cf. pages 77-80. We all could profit by studying a little descriptive geometry.

THE GEOMETRIC CONTENT OF ADVANCED CALCULUS

Lectures by A. M. Gleason

(Lecture notes by Melvin Hausner)

Lecture I.

I would like to continue Steenrod's series of lectures by considering some problems in analysis. To begin with, let me express the general opinion that the course we teach in college which is usually called "Calculus" frequently hurries into such questions as differentiation and integration, and often fails to put the proper emphasis on what the subject is all about, namely functions of a real variable, or of several real variables. The differential and integral calculus are, after all, techniques used to find out certain properties of functions, and should not be considered as ends in themselves. I believe that we are going to see, in the near future, a considerable change in the emphasis of the college calculus course which will reflect this opinion.

The easiest kind of function is the constant function, of which there is little to say. Next in the hierarchy are the linear functions, and then the quadratic functions. The linear functions (strictly, the affine functions) already exhibit their basic properties in dimension one. Or, if we consider two of these functions at a time, we get a good idea of the general situation by considering lines in the plane. Thus, to solve two linear equations in two variables, we have (geometrically) two intersecting lines, parallel lines, or identical lines. All of these situations have, as you know, straightforward generalizations in higher dimensions. For the quadratic functions, the basic situation (at least if we consider these functions one at a time) occurs in three dimensions. That is, all of the interesting phenomena concerning these functions which occur in higher dimensions already show up in three dimensions.

But this is not true of the cubic functions (of several variables) or of the polynomial functions of higher degree, where new complications occur in higher dimensions. I have a feeling that it is because we live in three dimensions, and because we can draw the pictures to see the phenomena of quadratic behavior but cannot see the basic phenomena for higher degree functions, that we know a great deal about the quadratics but very little about the higher degree functions. At any rate, this seems to be a partial explanation for our ignorance.

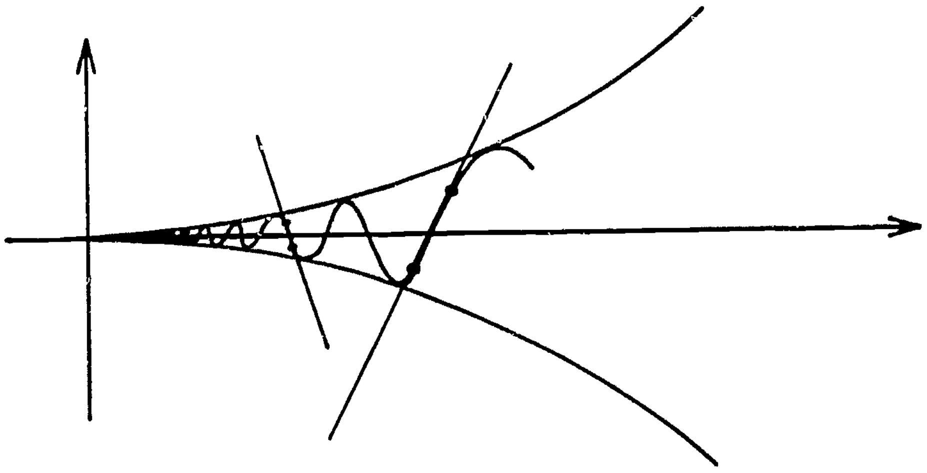
The basic thought behind calculus is to reduce questions of complicated functions to the simpler functions. We can phrase this in a sufficiently informal way, which is necessarily full of qualifications. <u>All decent functions are practically linear</u>. The theme of the calculus is to take advantage of this situation. For example, if f is a "decent" real-valued function of a real variable, and if x is in its domain of definition, and if x' is near x, then $f(x') - f(x)$ is approximately equal to $L_x(x' - x)$, where L_x is a linear function, which varies--probably non-linearly--with x:

$$f(x') - f(x) \sim L_x(x' - x).$$

Of course, this "equation" needs clarification. The usual meaning of this approximation is that the approximation "gets better" as x' gets closer to x. More formally, there is a linear function L_x (depending on f, and explicitly on x) with the property that for any x and any $\epsilon > 0$, there exists a $\delta > 0$ such that if x' is any number satisfying the condition $|x' - x| < \delta$, then

$$|f(x') - f(x) - L_x(x' - x)| \leqq \epsilon|x' - x|.$$

(Note that the weak inequality permits $x' = x$.) This is the definition of the Fréchet derivative. However, this definition is not strong enough to reject many pathological situations. For example, consider the function $f(x) = x^2 \sin \frac{1}{x}$. It is well known, and easy to prove, that $f'(0) = 0$.



Geometrically, each chord starting at the origin will have its limiting slope 0 as the other end-point approaches the origin along the curve. However, if we take two different points each approaching the origin along the curve, the limiting slope will not exist.

It is possible to change the above definition slightly to avoid this occurrence. We merely take two points x' and x'' independently approaching x. Formally, given the function f and the point x, the "strong" Fréchet derivative at x is the linear function L_x with the property that for any $\epsilon > 0$, there exists a $\delta > 0$ (depending on ϵ, f, and x) such that if $|x' - x| < \delta$ and $|x'' - x| < \delta$, then

$$|f(x'') - f(x') - L_x(x'' - x')| \leqq \epsilon|x'' - x'|.$$

Geometrically (we are thinking of f as a real-valued function of a real variable) we are permitting x' and x'' to approach x independently and it is required that the slope of the secant joining (x',y') and (x'',y'') have a limit. Here, x' and x'' may be relatively close, while each is relatively far from x. Intuitively, this is certainly a reasonable definition for smoothness of a curve. It is also true that this definition is equivalent to the assumption that f has a continuous derivative. Thus, restricting ourselves to an interval, f will have a continuous derivative in this interval if and only if f has a "strong" Fréchet derivative at every point of the interval. This definition illustrates, geometrically, why it is the C^1 functions which we wish to consider, rather than merely differentiable functions.

In the real variable case $(f: R \to R)$, the derivative L_x is a number. It is an easy matter to consider these derivatives for functions $f: R^n \to R$, and more generally if f maps R^n into R^m. All that is necessary is to replace the absolute values by norms wherever required. If $f: R^n \to R$, then the derivative L_x is a linear function from R^n into R satisfying

$$|f(x'') - f(x') - L_x(x'' - x')| \leqq \epsilon\|x'' - x'\|$$

provided $\|x' - x\| < \delta$ and $\|x'' - x\| < \delta$. Similarly, if $f: R^n \to R^m$, L_x is a linear function from R^n into R^m such that

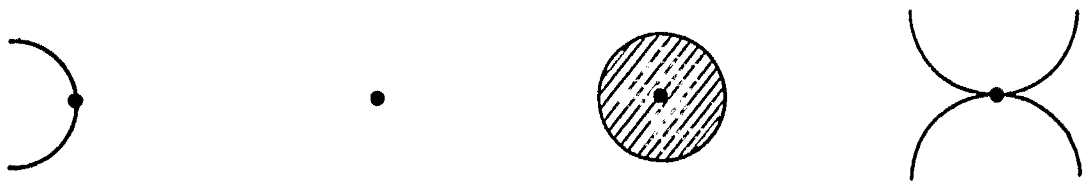
$$\|f(x'') - f(x') - L_x(x'' - x')\| \leqq \epsilon\|x'' - x'\|$$

provided $\|x' - x\| < \delta$ and $\|x'' - x\| < \delta$. Again L_x varies, in general non-linearly, with x. As in the simpler case, this leads to C^1 mappings as the primary notion rather than the secondary one.

We may rephrase our informal statement: <u>All decent functions have continuous derivatives</u>. However, we can say more, again informally. In many ways, functions of class C^1 behave just as if they were linear. We can rephrase as a metaprinciple.

<u>In cases where the affine approximations to functions are in general position, the qualitative behavior of the functions is just as if they were their affine approximations</u>.

Implicit Function Theorem. We shall illustrate the above idea by considering the implicit function theorem. Suppose we have a function $F: R \times R \rightarrow R$. The implicit function theorem gives conditions under which the equation $F(x,y) = 0$ may be solved for y in terms of x. We first note that calculus cannot answer such a question by itself. We must have an initial solution (x_0,y_0) with $F(x_0,y_0) = 0$ in order to "get a grip on the problem." The problem of discussing globally what the solutions are like is not within the scope of calculus. However, the local solutions of the equation $F(x,y) = 0$ in the neighborhood of (x_0,y_0) might not be represented as a function at all. The diagram gives some of the conceivable possibilities. The classical criterion is that if F is C^1, and if $F_2(x_0,y_0) \neq 0$, then for some interval I about x_0, there exists a C^1 function $g: I \rightarrow R$ such that $F(x,g(x)) = 0$ for all x in I and $g(x_0) = y_0$. Furthermore, g is unique among all



Possible solution sets of $F(x,y) = 0$ near (x_0,y_0).
(Given $F(x_0, y_0) = 0$.)

continuous functions $h: I \rightarrow R$ which satisfy the conditions $F(x,h(x)) = 0$ and $h(x_0) = y_0$. Also, the derivative g' is given by the formula $g'(x) = -F_1(x,g(x))/F_2(x,g(x))$ in the interval of definition. (We use the notation F_i for the "partial derivative with respect to the $i^{\underline{th}}$ variable.")

To see how this fits in with the above discussion, we write

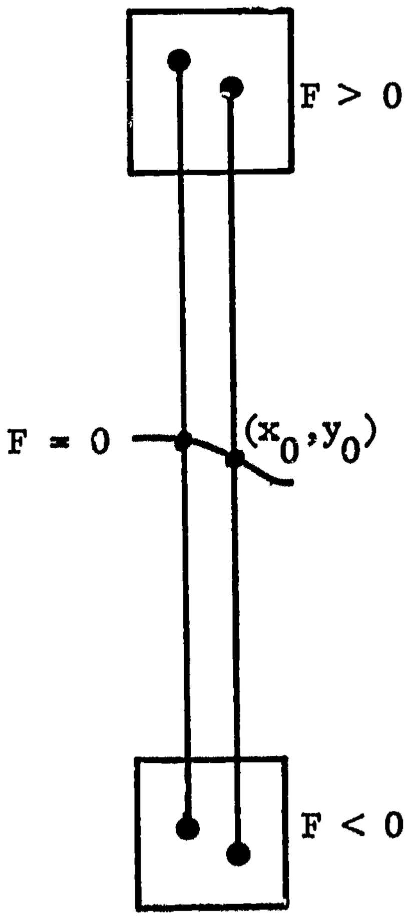
$$F(x,y) \sim F(x_0,y_0) + F_1(x_0,y_0)(x-x_0) + F_2(x_0,y_0)(y-y_0).$$

If we replace F by its affine approximation, and use $F(x_0,y_0) = 0$, we have to solve the equation

$$F_1(x_0,y_0)(x-x_0) + F_2(x_0,y_0)(y-y_0) = 0.$$

Of course, we can solve for y if $F_2 \neq 0$ at (x_0,y_0). (This is the "general position" for the affine approximations referred to in the metaprinciple.) We also note that the solution in this case is linear, and has as its derivative $-F_1/F_2$ evaluated at (x_0,y_0). Thus, when we solve the affine approximation of the given equation $F = 0$, we obtain a solution which is the affine approximation of the answer.

We recall the proof of this theorem, giving only a sketch. In the plane, we have the point (x_0,y_0) at which $F = 0$. Then, along the vertical line through (x_0,y_0), the derivative of F is not zero. Say it is positive. Therefore, the derivative is positive on the vertical line through (x_0,y_0). Hence, F is strictly monotonic on this line, positive above (x_0,y_0) and negative below it. By continuity, the signs are preserved in a (two-dimensional) neighborhood of these points. If we take any vertical line intersecting these neighborhoods, there must be a point where $F = 0$ on this line, since there is a sign



change. In fact, since $F_2 > 0$ at (x_0, y_0), we may ass me that all of this is taking place in a neighborhood in which $F_2 > 0$ (F is C^1). Thus F is strictly monotonic increasing along each of these vertical lines and $F = 0$ only once along each of these lines. This gives the uniqueness of our solution, and it also shows that we have a function $g(x)$. Then using the mean value theorem, or some equivalent result, we can show that g is differentiable and has the required derivative.

We can look at this result in another way. We imagine the smooth surface $z = F(x,y)$ which, by hypothesis, passes through the point $(x_0, y_0, 0)$. It has a tangent plane at that point. The theorem states that the surface cuts the plane $z = 0$ in a curve and that this curve has as its tangent line the intersection of the tangent plane with $z = 0$. The "general position" requirement guarantees that the tangent plane does not coincide with the plane $z = 0$, nor is the intersection with $z = 0$ a line parallel to the y-axis.

Suppose we consider the higher dimensional case where $F: R^n \to R$. Let us take $n = 4$, since the general case is no harder. It is worth noting that the difficult analysis has already been done in the case $n = 2$ discussed above. (As in many such situations, this theorem is essentially two-dimensional.) For convenience in notation we let the variables be x_0, x_1, x_2, and x_3. For partial derivatives we shall use the corresponding subscripts. Suppose $F : R^4 \to R$ and that $F(x_0^o, x_1^o, x_2^o, x_3^o) = 0$. When can we solve the equation $F = 0$ for the variable x_0? Assuming F is C^1, we replace F by its affine approximation at the given point to obtain, as the equation approximating $F = 0$, the equation

$$0 = F_0(x_0 - x_0^o) + F_1(x_1 - x_1^o) + F_2(x_2 - x_2^o) + F_3(x_3 - x_3^o) ,$$

where the partial derivatives are evaluated at the given point. To solve for

x_0, we need $F_0 \neq 0$ at the given point. In this case x_0 can be expressed linearly in terms of x_1, x_2, and x_3. The implicit function theorem, in this case, states that the non-linear equation $F = 0$ can be solved for x_0, provided the linear approximation can be solved uniquely. Again, the linear solution of the approximating equation is the linear approximation of the solution of the original equation.

We can get the result for the higher dimensional case by simply "fixing" two of the variables and concentrating only on x_0 and one of the other variables. The two-dimensional case permits us to solve in terms of this variable, and gives us a formula for its derivative. In this case we have a partial derivative, since the remaining variables are held fixed. This also shows continuity of the partial derivatives. Thus, using the theorem that continuity of the partial derivatives implies differentiability in the sense of Fréchet, we obtain the differentiability of the solution $x_0 = g(x_1,x_2,x_3)$.

I mentioned, at the beginning of this lecture, the importance of the idea of a function in the calculus. We should be careful, as far as possible to distinguish between three kinds of properties of functions. These are the global, the local, and the infinitesimal. A global property of a function is a property which concerns the function in its entire domain. A local property is a property which is stated about a function in a little patch, whose size may vary from point to point; but the property concerns the function in the whole patch. Finally, an infinitesimal property of the function is one concerning its derivatives at a point. For example, the property that a function is differentiable at a point, or that its derivative vanishes at a point, is an infinitesimal property, since no information is given about the function's behavior over any small patch. We ought to realize that the classical differential calculus concerns itself with the interplay between local and infinitesimal properties of functions. It says virtually nothing about the connection between the local and the global properties of functions. Non-calculus techniques, just as important as those of calculus, connect the local and global levels. For some reason, this is kept secret in the books. But differential calculus cannot help with this problem; it does not have the grip. (Integral calculus is a little more global in its scope.)

For example, in the implicit function theorem considered above, the hypothesis on the derivatives was an infinitesimal hypothesis, while the property of having continuous derivatives is a local property. The conclusion was a local one. Calculus does not provide us with the in-the-large (global) solution of the equation $F(x,y) = 0$. Similarly, calculus did not help us find a point (x_0,y_0) where $F = 0$.

This issue also comes up in maximum and minimum problems, and it should be mentioned when it occurs. Here, the notion of a local maximum and minimum is a local one. The theorem often used is that at a local maximum, or minimum, the derivative is 0. This is an infinitesimal condition. The sufficiency condition in terms of second derivatives is also an infinitesimal condition. However, finding the actual minimum is a global problem, and this feature should be stressed in teaching the calculus.

Discussion.

The uniqueness of the solution of $F(x,y) = 0$, and its "completeness" was brought up. Gleason gave the meaning of local uniqueness, in the possibly non-continuous sense, as follows: Let S be the set of points where $F = 0$.

Then there exists some neighborhood N of the given point (x_0,y_0), such that $N \cap S$ is the graph of some function $g(x)$ defined in an interval about x_0. Furthermore $g(x)$ is C^1 in this interval.

Concerning the continuity of the solution of $F = 0$ for more than two variables, and the relationship with the two-dimensional case, Gleason pointed out that the argument given in the two-dimensional case goes over easily in higher dimensions. The argument is essentially two-dimensional.

It was agreed that the levels global, local, and infinitesimal sometimes get mixed in hypothesis and conclusion. Despite the resulting uncertainty and even possible confusion of level, Gleason felt that the distinction was important and ought to be stressed.

Some examples of theorems with global conclusions are: The mean value theorem. The theorem that if $f' > 0$, then f is monotonic. However, it was pointed out that even these theorems use the global hypothesis that the domain is connected. Similarly, the theorem on the attainment of a maximum requires the global hypothesis of compactness of the domain.

It was thought that a careful use and phrasing of the mean value theorem would make the generalization of the implicit function theorem trivial, by replacing the real variable x by the vector variable x. However, the correct form for the mean value theorem was left open.

The generalization of the implicit function theorem when F is a vector valued function was brought up. It was agreed that the proof would be more difficult here. Gleason remarked that even here, the idea of the one-dimensional proof can be generalized. The intermediate value theorem generalizes in higher dimensions to an appropriate theorem on winding numbers or degrees of mappings.

It was pointed out that the equivalence of continuity of the partial derivatives, and the existence of a "strong" Fréchet derivative was not well known, with only a one-way implication usually stated and proved. A brief sketch of the proof was indicated (in one variable), and it was clear that in one variable, the equivalence was quite elementary.

It was mentioned that the book A First Course in Integration by E. Asplund and L. Bungart used this idea throughout. At this point Gleason remarked that he would not object to the abandonment of the traditional derivative in favor of the more restrictive strong Fréchet derivative. However, he mentioned that he had not considered the long range implications of this idea. The effect would be, at first glance, the abandonment of pathologies, which might prove worthwhile at this level. It was pointed out that the function would not even have to be defined at the given point, although it could be defined using continuity. Coxeter pointed out that the usual definition of the derivative would be discovered as the first case of Taylor's theorem. Another point mentioned in favor of the traditional definition was that it was easily computed. In general, no one was firmly in favor of a change towards the strong Fréchet derivative.

Gleason felt that there was too much emphasis in present-day calculus courses on pathological functions. The pathologies are not really in the main-stream. For example, it is fairly hard to prove that every continuous function is Riemann integrable, but it is quite easy to prove it for piecewise monotone bounded functions. The latter hypothesis covers all the functions that need to be considered in a first course.

The distinction between the general mathematical notion of a function and the scientist's notion of numbers determined by experiments should be mentioned, and in the introductory courses the problem should be what functions should be considered to represent these "physical" functions. Thus, we should raise the question "What are the possibilities for functions?"

Lecture II.

It should be noted that the geometric formulation of the alternative definition of the Frêchet derivative must be stated with care. For example, if $F: R^2 \to R$, then, in analogy with the slope of a secant through two distinct points x' and x'' approaching a given point, we might imagine that we would have a similar situation involving the direction of the plane through points on the graph of $z = F(x,y)$ corresponding to three non-collinear points (x_i, y_i) approaching the given point. However, we can easily convince ourselves that if the points are almost collinear, then we will probably obtain an almost vertical plane, which is not necessarily near the tangent plane. What is required is to keep the angles of the triangle in the (x,y) plane away from 0.

Implicit Function Theorem for Several Variables. We now recall the general theme enunciated in the last lecture, namely that C^1 functions behave like their linear approximations, provided they are in general position--that nothing vanishes "by accident." Let us now consider the implicit function theorem for mappings into a space of more than one dimension. To be specific, suppose $H: R^4 \to R^2$, and H is C^1. (Last time, we only considered real-valued functions.) We wish to solve the equation $H = 0$. Thus we have the system

$$F(x, y, u, v) = 0$$
$$G(x, y, u, v) = 0$$

where F and G are real-valued and C^1, and we wish to solve for u and v as functions of x and y. Again we suppose that we have an initial solution (x_0, y_0, u_0, v_0), since calculus cannot help us find this. If we linearize the system (i.e., replace each function by its linear approximation) we obtain the system

$$F_1(x-x_0) + F_2(y-y_0) + F_3(u-u_0) + F_4(v-v_0) = 0$$
$$G_1(x-x_0) + G_2(y-y_0) + G_3(u-u_0) + G_4(v-v_0) = 0$$

where the various partial derivatives are evaluated at the point (x_0, y_0, u_0, v_0). The <u>linear system</u> can be solved uniquely for u and v provided

$$\det \begin{bmatrix} F_3 & F_4 \\ G_3 & G_4 \end{bmatrix} \neq 0.$$

(Classically, the Jacobian of the given functions, with respect to the variables we wish to solve for, does not vanish.) The condition that the determinant does not vanish is, in this case, the condition that the functions are in general position. The determinant condition then implies the conclusion that we can solve for u and v in terms of x and y. Again we stress that we obtain a <u>local solution</u>, and that it is essential that the functions be C^1 (although some weakening of this hypothesis can be made). The conclusion of the theorem can be stated as follows. If S is the set of points in R^4 which satisfies the equations $F = 0$ and $G = 0$, then there exists a neighborhood N in R^4 of the given solution (x_0, y_0, u_0, v_0) such that $S \cap N$ is the graph of a function from some open domain of R^2 into R^2. That is there exist functions $g(x,y)$ and $h(x,y)$ such that $S \cap N$ is the set of points $(x,y,g(x,y),h(x,y))$ for (x,y) in some open domain N_1 of R^2. Furthermore, the functions g and h are C^1. As before, the solution of the approximate equation is the approximation of the solution.

The implicit function theorem very nicely illustrates the idea of a calculus course as a course in the theory of functions. This theorem is clearly a new technique for defining functions (the implicit functions). As we are aware, the student's first reaction when he is told that "we can solve for u and v" is usually confusion if not rebellion. For example, he might insist on a formula. Thus, it is helpful to take the view that we are finding new techniques for defining functions and then examining the properties of these new functions.

Inverse Function Theorem. We now consider the inverse function theorem, a theorem which may be regarded as a special case of the implicit function theorem in more than one variable, since it essentially involves solving n equations for n unknowns. However, we shall view this theorem in a different geometrical light. If F is a function from some domain of R^n, we shall use the notation $(dF)_x$ to indicate the derivative of F at the point x. $(dF)_x$ is a linear map from R^n into the appropriate vector space determined by the range of F. Then the inverse function theorem is as follows. Suppose F is a function from some domain on R^n into R^n (the same dimension). Suppose that F is C^1. This implies, for example, that for any x_0 in the domain of definition, and any $\epsilon > 0$, there exists a $\delta > 0$, such that if $\|x-x_0\| < \delta$, then

$$\|F(x) - F(x_0) - (dF)_{x_0}(x-x_0)\| \leqq \epsilon \|x-x_0\|.$$

Then if $(dF)_{x_0}$ is injective, or equivalently if it is surjective, or equivalently if $\det (dF)_{x_0} \neq 0$, then F is invertible near x_0. That is, there exists some neighborhood N of x_0 such that $F(N) = M$ is a neighborhood of $F(x_0)$, and such that if F is restricted to N and regarded as a map of N onto M, F is injective, and its inverse F^{-1} is also C^1 on M. Once again we stress that the theorem is strictly a local theorem. There are similar results for functions in C^n, where $n = 2,3,\ldots,\infty$. But these results are an immediate consequence of the C^1 theorem, since the C^1 theorem allows us to write a formula for the derivatives of F^{-1}, and this formula shows that F^{-1} has as many continuous derivatives as F. The important result is the C^1 case. The case $n > 1$ is no real extension.

There is a difference in viewpoint between the implicit function theorem and the inverse function theorem. In the former case, we tend to view the result in terms of the intersection of the contour level sets of two or more functions. In the inverse function theorem, we think of a mapping taking points of one space into the other.

The Use of Affine Spaces. We can phrase these theorems in a slightly different way which is somehow more geometric. It is unrealistic to use a vector space as the domain and range of functions, since for example the world we live in (as well as Euclidean geometry) does not have an origin. A more reasonable space to choose is an affine space. As is well known, to each affine space A, there is associated a vector space V, which I prefer to call its director space. For example, we may regard V as the set of translations in A, and it is not hard to characterize axiomatically the relationship between A and V. A familiar result is that two points P and Q of A determine a unique vector $\vec{PQ}$ of V, which may also be written $Q - P$. It is the (origin-free) affine space, or some domain of it, which should be used as the range and domain of functions. Now, if A and B are affine spaces with director spaces V and W, and if $F: A \to B$, then $(dF)_{x_0}: V \to W$. When A is a vector space to begin with, there is a natural way to identify A with V. In particular when A and B are taken to be R^n and R^m, the distinction between these spaces and their director spaces is often blurred. But the more general way of looking at F and dF does clarify the geometry.

We can illustrate with the example of the velocity vector. Suppose we regard A as Euclidean 3-space, and we have a function $R \xrightarrow{f} A$. We regard R as representing time, and we regard A as Euclidean 3-space. The velocity vector is usually computed by forming the difference quotient

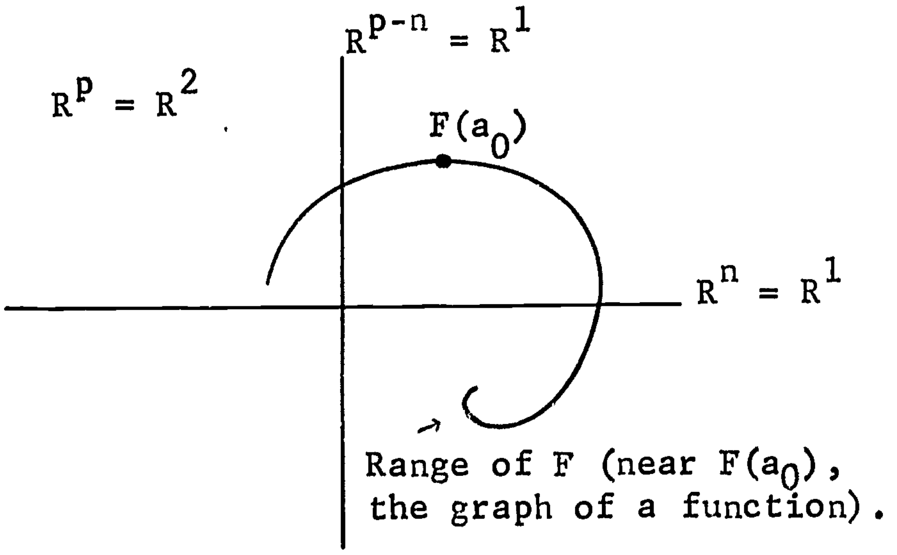
$$(f(t_1) - f(t_0))/(t_1-t_0)$$

and letting t_1 approach t_0. But the numerator is not in A. It should be regarded as the vector displacement $\overrightarrow{f(t_0)f(t_1)}$, hence an element of V, the director space of A. The limit is therefore in V, so $f'(t_0) \in V$. Thus, we have $R \xrightarrow{f'} V$. Note that $f'(t_0)$ is not in A. Every physicist knows that a velocity is not a distance--it is a different object: $f'(t)$ is in a different space.

The points of time do not constitute a v:ctor space in any natural way, but may be regarded as a one-dimensional affine space T. The director space of T will be denoted by I (for time intervals). It is also wrong to regard I as the vector space of real numbers R. The process of introducing a "unit of time" is the same as choosing a (one vector) basis in I. Once this is done, we get real numbers. These show up simply as coordinates of vectors with respect to this basis (unit of time). Thus "second" is the name of a non-zero vector of I. For the motion $T \xrightarrow{f} A$, we obtain the velocity $I \xrightarrow{df} V$. To get a numerical answer, for example in ft/sec, it is necessary to choose a basis (sec) in I, and a suitable basis in V. Then the usual way of finding the matrix of the linear transformation in terms of a given basis yields the correct numerical answer for the velocity in terms of the units chosen. Now, if we realize that for each $t \in T$, there is associated a linear map of I into V, we may regard df as a map of T into Hom(I,V). Here Hom(I,V) is the vector space of linear transformations of I into V and may be regarded as its own director space. Thus, acceleration shows up as its differential: $(ddF)_{x_0}: I \to \mathrm{Hom}(I,V)$. Since there is a natural basis for $\mathrm{Hom}(I,V)$ in terms of the basis for I and V, we obtain natural units for the acceleration vector. There is no question, for example, of combining velocity and acceleration vectors. They belong to different vector spaces. In general, then, if A and B are affine spaces with director spaces V and W respectively, and if D is some domain of A, then for any C^1 function $F: D \to B$, we have the differential map $dF: D \to \mathrm{Hom}(V,W)$. Otherwise put, $(dF)_{x_0}$ is a linear map of V into W. Parenthetically, even the notion of a norm for V and W is not required for the definition of dF, since for finite dimensional vector spaces, the topology is uniquely determined. However, carrying this out is a bit tricky.

We now consider some variations of the inverse function theorem. Suppose $F: R^n \to R^p$. As usual in our discussions, we may take the domain of definition to be any open set in R^n, and F is understood to be C^1. As discussed above, we may take $F: A \to B$ where A and B are affine spaces of dimension n and p, respectively. We take, for convenience, the director spaces of A and B to be R^n and R^p respectively. This amounts to choosing basis vectors in the director space and has the notational advantage that the dimension of these spaces is explicitly given. We thus have $(dF)_a: R^n \to R^p$. We can now state some theorems which are easily derived from the inverse function theorem by means of a few simple tricks.

First suppose $n \leqq p$ and that $(dF)_{a_0}$ (which maps R^n into R^p) has rank n, the largest possible rank. Then F maps some neighborhood of a_0 in R^n into an n-dimensional surface imbedded in R^p. Specifically, by choosing the coordinates in the right way, the image of some patch in R^n is the graph of a (smooth) function mapping a domain in R^n into R^{p-n}.

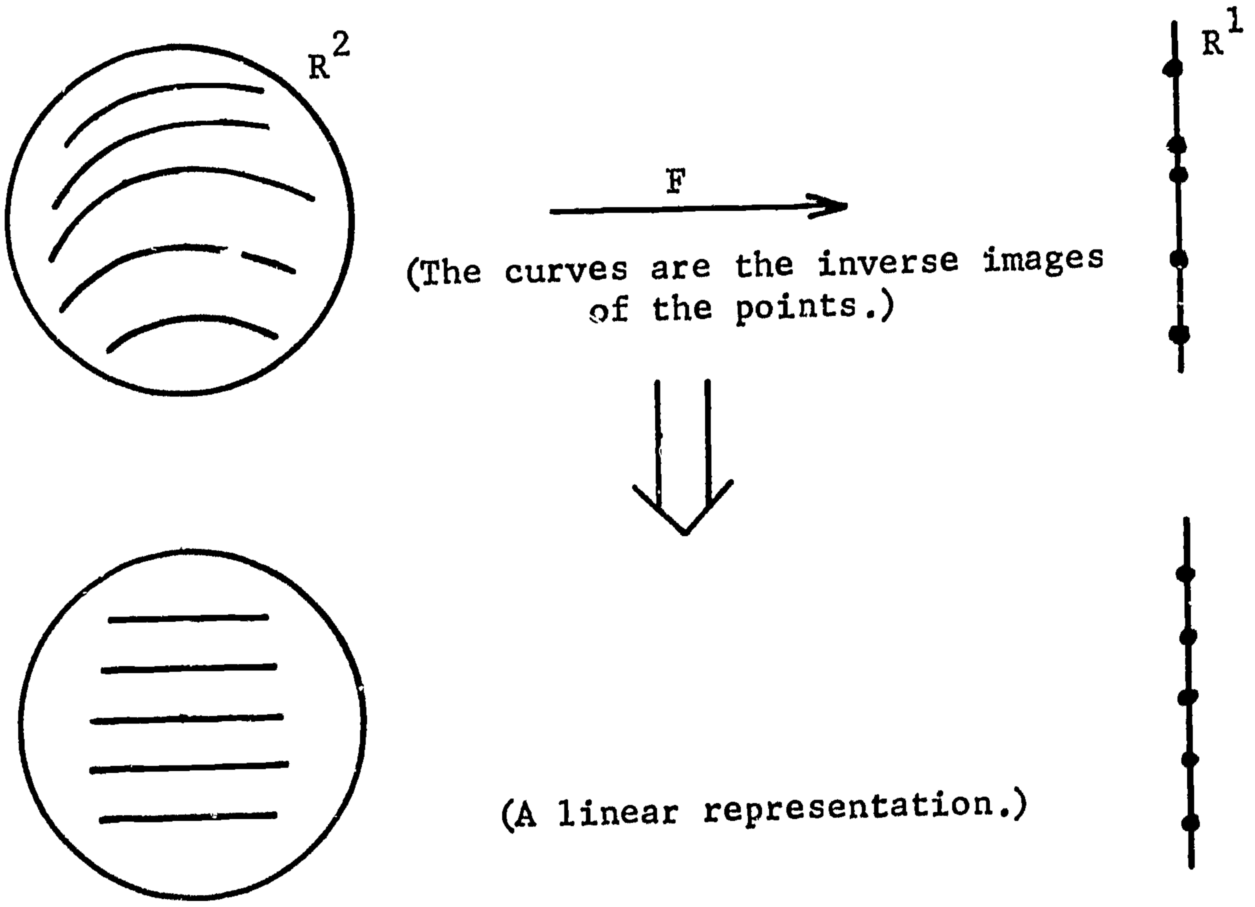


Range of F (near $F(a_0)$, the graph of a function).

Thus, we can represent the range (always locally!) by choosing a certain n "free" coordinates, and expressing the other $p - n$ in terms of them. Parenthetically, this is a very good way of defining an n-dimensional surface in p-space $(n \leqq p)$.

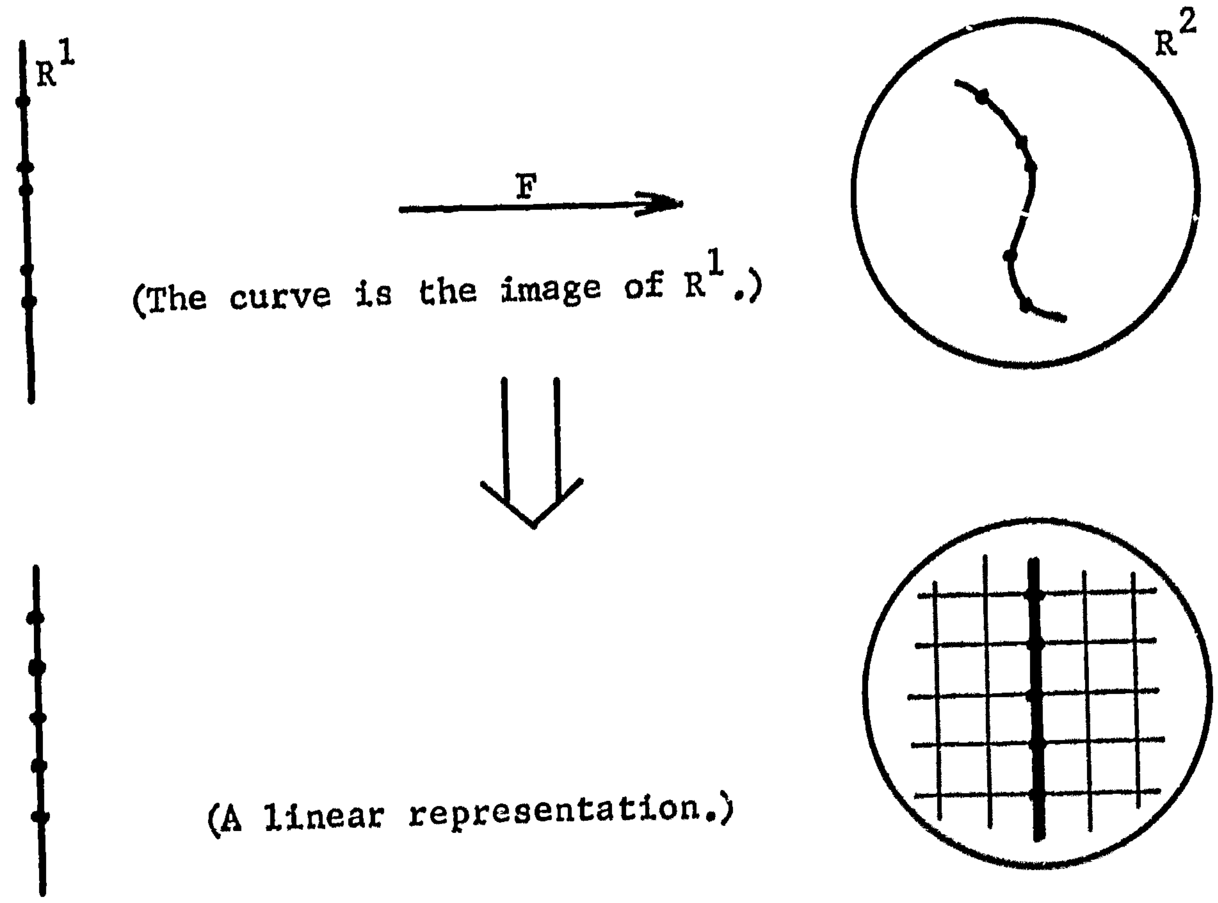
If the roles of the dimensions are reversed, i.e., if $n \geqq p$, we have the theorem that if the rank of $(dF)_{a_0}$ is p, the largest possible rank, then F is (locally) a surjection--some neighborhood of a_0 maps onto an open set. We can phrase this more precisely in terms of graphs of functions from R^p into R^{n-p}, but it is most convenient to express both of the above results in terms of curvilinear coordinates. If the maximal rank hypothesis is valid for the linear parts dF of the mapping F, then by introducing curvilinear coordinates locally in A and in B, F will actually be represented by its linear approximation. Since one can make a reasonable plea for curvilinear coordinates as a slight distortion of the usual coordinates, this fact is a reasonable statement of the theme that C^1 functions have, in general, the same behavior as their linear approximations. We illustrate some of the lower dimensional cases on the next page.

If the rank of dF is not maximal, we can still make this statement in



(The curves are the inverse images of the points.)

(A linear representation.)



(The curve is the image of R^1.)

(A linear representation.)

certain cases. The theorem is that <u>if the rank of</u> dF <u>is locally constant</u>, then, by introducing curvilinear coordinates, F can be represented by its linear part dF. If we look at the matrix definition of the rank of dF, we can see that the rank can go up locally, but never down. For if we find the largest $r \times r$ submatrix which is non-singular, this non-singularity will be maintained near the point in question. Thus, once again, this is a theorem whose hypothesis puts the function in general position. Oncе again we also remark that this is strictly a local theorem.

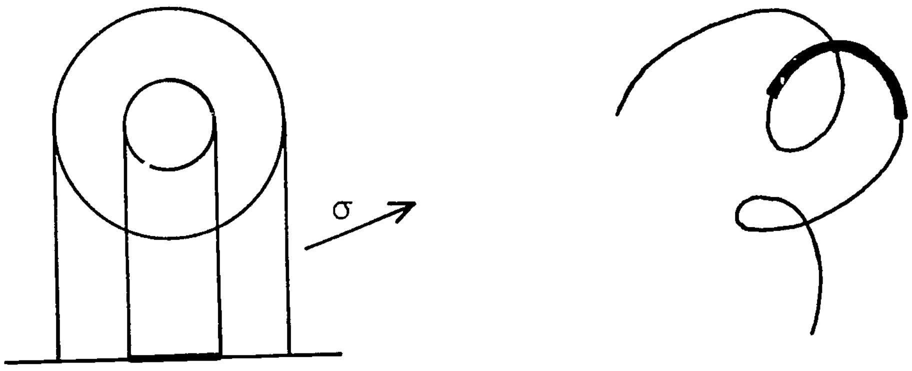
<u>Functional Dependence</u>. We now consider an example of a theorem which is widely quoted, but almost invariably incorrectly. This is the theorem on functional dependence of functions: If f, g, and h are (C^1) functions of three variables, and if the determinant of the Jacobian matrix is identically zero:

$$J\left(\frac{f,g,h}{x,y,z}\right) = \det \begin{bmatrix} \frac{\partial f}{\partial x} & \cdots \\ \cdots & \\ \cdots & \frac{\partial h}{\partial z} \end{bmatrix} = 0, \text{ identically,}$$

then there is a functional dependence among the functions, i.e., a function φ such that $\varphi(f,g,h) = 0$ identically. This theorem is false in any reasonable interpretation. Of course, we do not want φ to be a trivial function identically 0. Thus, we wish to exclude the possibility that φ vanishes in a region. We certainly want φ to be continuous. Actually there are two statements which should be added to the theorem, which are often omitted. First, the theorem is a local theorem, and second, the rank must be locally constant. Thus, if the Jacobian vanishes in a neighborhood of a point, the functional dependence may only hold in some smaller neighborhood of the point.

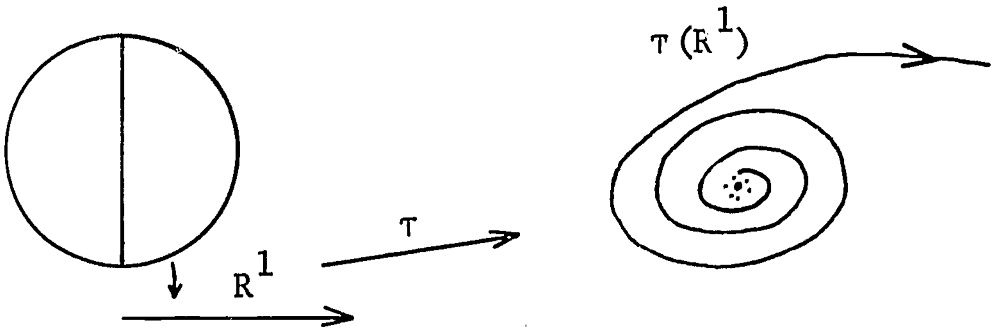
Furthermore, the hypothesis must include the condition that the rank of the Jacobian is locally constant. We now offer some counterexamples.

First, as to the local nature of the theorem, we proceed as follows. We first map the line R^1 densely and smoothly into the plane R^2 by means of some function $\sigma: R^1 \to R^2$. This may be achieved by enumerating the rational points of the plane, and mapping the integers one-to-one into these rational points. We can then extend (even in a C^∞ manner and with non-zero derivative everywhere) to a function from R^1 to R^2. We now project R^2 onto R^1 and follow this by the mapping σ, to obtain a map F of R^2 into R^2. Clearly,



dF has rank 1 everywhere, so the Jacobian vanishes. But equally clearly, if $\varphi(F) \equiv 0$, and if φ is continuous, then we must have φ identically 0, since it vanishes on the dense range of F. However, the result is locally true in this case, and the diagram illustrates the local nature of the conclusion.

We now give an example which shows that we must include the hypothesis that the rank is locally constant. Again we start on the line R^1 and construct a subsidiary map $\tau: R^1 \to R^2$. We take the ray $x > 0$, and map it in a C^∞ fashion into a spiral winding about and approaching the origin. This can be arranged so that $\tau(x)$ approaches 0 as x approaches 0. We can also arrange to have $\tau(0) = (0,0)$, and to have all of the derivatives of τ



vanish at $x = 0$. For example, we may define

$$\tau(x) = (e^{-1/x^2} \cos \frac{1}{x}, e^{-1/x^2} \sin \frac{1}{x}), \quad x \neq 0,$$
$$\tau(0) = (0,0).$$

Note that $d\tau$ has rank 0 at $x = 0$, while the rank of $d\tau$ is equal to 1 everywhere else. Again, we first project R^2 on the line R^1, and we compose with τ. For the resulting map F, dF has rank 1 everywhere except on the line $x = 0$, where its rank is 0. If we consider any small neighborhood of a point where $x = 0$, it will map onto a central portion of this spiral. Then while there is a C^∞ function which vanishes only on this spiral, all of its derivatives also vanish at the center point, and it cannot be considered a reasonable function. The functional dependence wanted requires that $d\varphi \neq 0$.

Finally, I should like to discuss the relationship between the implicit function theorem (in one dimension) and the inverse function theorem. It is an easy matter to go from the inverse function theorem to the implicit function theorem. For example, to solve the equation $F(x,y,u) = 0$ for u, we merely consider the mapping $(x,y,u) \to (x,y, F(x,y,u))$ and find its inverse. Equivalently, we invert the system $r = F(x,y,u)$, $s = x$, $t = y$. However, going back from the implicit function theorem to the inverse function theorem is tricky. For example, to solve the system $F(x,y,u,v) = 0$, $G(x,y,u,v) = 0$, the device is to solve one of these equations for one of the variables (which is possible by the Jacobian condition), substitute in the other equation, and verify that the appropriate partial derivative of the composed function does not vanish. This turns out to be the original Jacobian condition. In n variables, this becomes quite complicated. I would like to point out the underlying reason for this disparity. If we consider the infinite dimensional case, it turns out to be an easy matter to prove an appropriate version of the implicit function theorem from the inverse mapping hypothesis. All that is required is care in the formulation. However, the method of solving for one variable at a time to prove the inverse mapping theorem is hopeless in the infinite dimensional case, or at any rate would require an exceedingly great effort. Because of this, we can state that the inverse function theorem is, in some sense, the more primitive theorem.

Discussion.

(Much of the discussion occurred during the lecture proper. But we include it here since, in the spirit of Fréchet, it was tangential.)

The "affine space - vector space" formulation led to the remark that this seemed quite sophisticated to present to, say, an engineer who wanted to compute. Gleason raised the question of whether, in fact, a mountain was being made of a molehill. But he noted that one should not expect someone to figure out by himself something that is too hard to explain to him. He will have to figure this out, somehow. Physicists and engineers need this formulation more than mathematicians.

Gleason pointed out that in a Euclidean space (defined as an affine space whose director space has an inner product), it is most natural that the inner product of two vectors be regarded as a bilinear map of $V \times V$ into D, some one-dimensional vector space. In this way, the choice of unit of length is equivalent to a choice of a basis vector in D. However, there is a question of whether enough is gained to warrant doing this systematically. At any rate, it is worth pointing out to a class.

It was observed that one of the reasons origins come up is that the algebraic formulation of a vector space is very much simpler than that of an affine space. Gleason mentioned that his definition of an affihe space involved the set A, a vector space V, and a mapping of $A \times A$ into V, corresponding to the operation of forming the vector from one point to another. At one point in the development, it is shown that the vectors are the translations in a natural way.

Gleason also mentioned the example of an affine space A which is a coset of a subspace W of a vector space V. In this case, W may be regarded naturally as the director space of A.

There was a brief discussion of the Fréchet derivative in infinite dimensional space. Here the director space is taken to be a Banach space and it is not hard to verify that the Fréchet derivative of a continuous function is a bounded operator (if it exists). This follows from the definition of the Fréchet derivative. One of the complications in extending the techniques is that closed subspaces do not necessarily have closed complementary subspaces. Some details are given in Lang's book on differential geometry.

It was pointed out that the functional dependence theorem is sometimes stated for functions which are the solutions of certain differential equations.

The distinction between the derivative f' and the differential df

was brought up. Gleason argued that if $f: A \to B$, it is desired to have $df: \Lambda V \to \Lambda W$, where ΛV and ΛW are the Grassmann algebras of V and W. However, he saw no harm in using df (instead of f') also as a mapping of V into W, since the former is a natural extension of the latter.

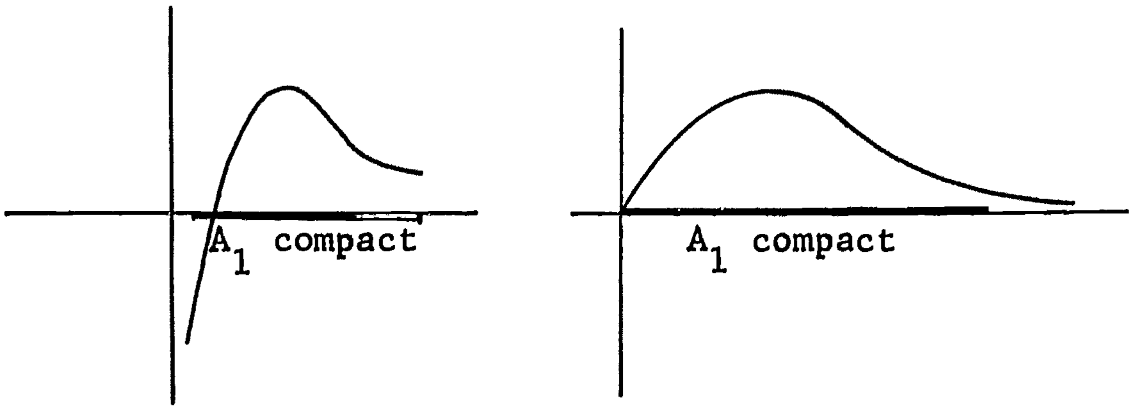
The theorem that a space curve is planar if it has torsion zero was cited as an example of another theorem which is often incorrectly stated. Gleason pointed out that it is necessary to require that the curvature is never zero.

Finally, the analogy was given between the result that the rank can never decrease locally and the result for polytopes that if vertices are moved locally the number of edges, etc., never decreases. The relationship between this semi-continuity property and the phenomena of "accidents" was pointed out.

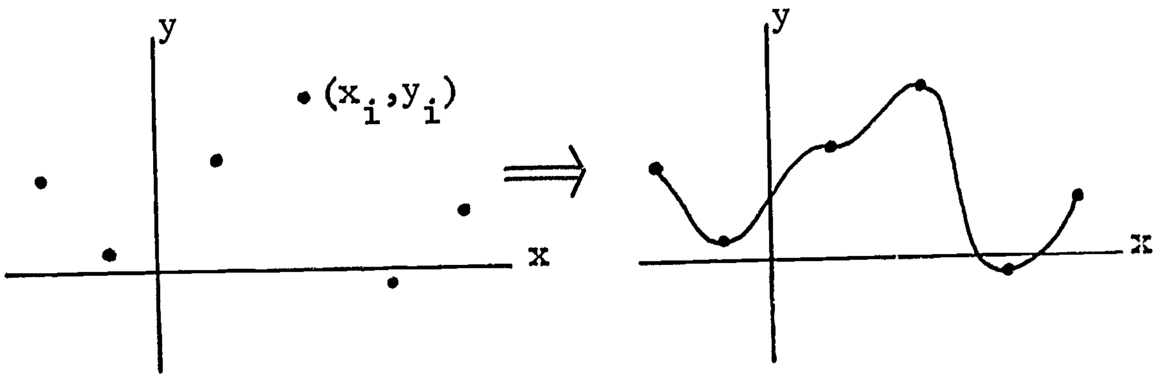
Lecture III.

Today I shall consider a topic familiar to any teacher of the calculus, namely, maximum-minimum problems. The typical setting for this problem is a function $f: A \to R$. The problem is to find the maximum value of f or the minimum value of f. A is usually some geometric space, possibly flat and possibly more complicated. The simplest case is where A is one-dimensional, and we shall consider this case first. For simplicity, we shall assume that A is an interval.

One-dimensional Maximum and Minimum Problems. An important aspect of maximum problems is the exis ence theorem: a continuous function on a compact set attains its maximum. If A is not a closed interval, then this existence theorem can often be applied by noting the behavior of f near the boundary, or infinity, and then by using the compact case on a closed sub-interval of A (see diagram). Following the existence theorem, we come to the necessary



conditions for a function to have a maximum at a point. Namely, if f attains a maximum at x_0, then either 1) x_0 is a boundary point of the domain, or 2) f does not have a derivative at x_0, or 3) $f'(x_0) = 0$. In practice we do not often consider the case 2) in a calculus course and, as we all know, the student often forgets case 1). All of these conditions are local, so that at some stage we have to sort out these "critical points" to find the (absolute) maximum. There are several sufficiency conditions for a local maximum, among which is the second derivative test. However, the most practical way of discovering the global and local maximum points is simply to plot the critical points $(x_i, f(x_i))$ and to note that between two consecutive critical points



the function is monotone. This fact is usually underemphasized, but it is enough to give a good picture of the global behavior of f, as well as to sort out the local and global maximum and minimum points (see diagram).

I should like to complain here about the way these problems are usually formulated and answered. In many important applications the problem is what is the maximum value of f? But in most texts, the questions usually ask where the maximum occurs. For example, many inequalities of the type "$f(x) \leqq M$ for all x in a certain set" are derived by using the usual techniques for finding the maximum value M of f. Students are often surprised at this turn of events because they were never asked to find $M = f(x_0)$. They only found x_0.

Maximum and Minimum Problems in Several Variables. For n variables, the situation is different in certain respects. For simplicity we shall confine ourselves to functions defined on an open domain, or the closure of such a set.

The case $n = 2$ typifies the general case. The existence theorems are essentially unchanged. The necessary condition that a function f attains its maximum at x_0 is as in the one-dimensional case, except that the derivative condition is replaced by the condition $(df)_{x_0} = 0$ (the zero linear operator). Points of non-differentiability are usually not considered. However, in this case there is no easy analogue of the monotonicity property mentioned above for the one-dimensional case. Here, for the consideration of interior points where $df = 0$, some test involving the second derivative is essential. Thus we require a consideration of $(ddf)_x$.

We now recall the theory. Starting with

$$A \xrightarrow{f} R,$$

we find the differential at a point x:

$$V \xrightarrow{(df)_x} R.$$

Regarding x as a variable point, and assuming f is C^2, we have

$$A \xrightarrow{df} \mathrm{Hom}(V,R).$$

Here $\mathrm{Hom}(V,R)$ is a vector space. Thus it may be regarded as an affine space which is its own director space. If we apply the same procedure to df we find the second derivative at a point x:

$$V \xrightarrow{(ddf)_x} \mathrm{Hom}(V,R).$$

Then, regarding x as a variable in A, we obtain

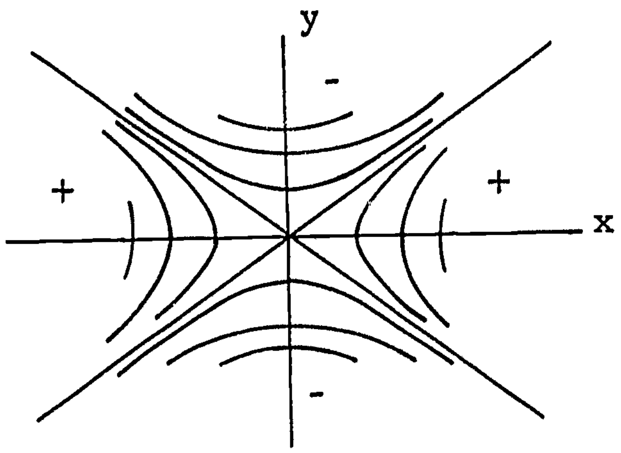
$$A \xrightarrow{ddf} H(V, \mathrm{Hom}(V,R)).$$

But this latter object is easily identified with the bilinear functions on $V \times V$, which we denote by $\mathrm{Bihom}(V, V, R)$. Furthermore, since we are dealing here with C^2 functions, $(ddf)_x$ can be shown to be a symmetric bilinear functional. This is equivalent to the result that the various partial derivative operators $\frac{\partial}{\partial x_i}$ commute on C^2 functions. If we set Symm(V, V, R) equal to the space of symmetric bilinear functions of $V \times V$, this amounts to saying that the above map factors through Symm(V, V, R). Also, the symmetric bilinear functionals are identified with the space Q(V, R) of quadratic forms on V in a natural way. Thus the diagram of the second derivative is as follows:

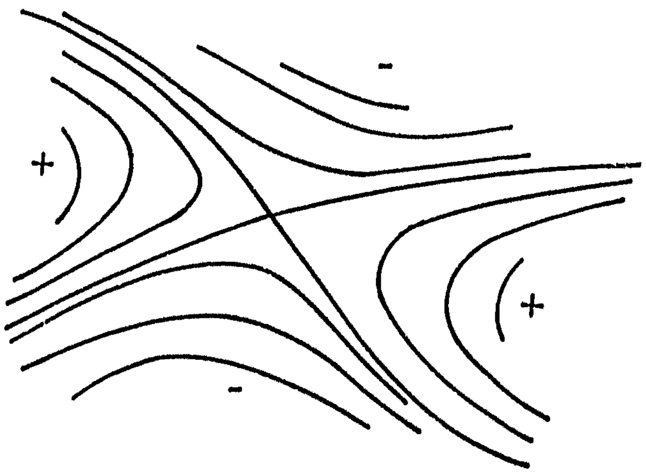
$$\begin{tikzcd}
A \arrow[r, "ddf"] & \text{Symm}(V, V, R) \arrow[r] \arrow[d, shift right, "\wr\|"'] & \text{Bihom}(V, V, R) \\
& Q(V, R) \arrow[u, shift right] &
\end{tikzcd}$$

As we know, the test for a local maximum is that the quadratic form be negative definite. If the form is positive definite, the function has a local minimum. If the form has full rank, but fails to be positive definite or negative definite, then the function has a saddle point. In an accidental situation the form is degenerate, and the classical answer is that the second derivative test fails. However, there is some sort of saddle point if the form has one negative and one positive eigenvalue.

We remarked in our first lecture that the quadratic functions are basically understood since all of the phenomena which can happen in n-space already occur for $n = 3$. Thus our ability to draw graphs permits us to look at and understand these functions. For $n = 1$ the functions are simply ax^2, and we have the natural distinctions $a > 0$, $a = 0$, and $a < 0$. In two dimensions, we are confronted with saddle points. The level lines for the functions $z = x^2 - y^2$ give an essentially new phenomenon. By the time $n = 3$, the only new thing is a degenerate quadratic with a saddle: $w = x^2 - y^2 + 0z^2$. No essentially new phenomenon occurs with more than three variables. For cubics, on the other hand, the types of singularities are more complicated. We should probably have to familiarize ourselves with cubics in quite a few variables (perhaps about six) before we could understand all the possibilities.



We often teach our students that in a neighborhood of a critical point where the quadratic form is non-degenerate, the function behaves like the corresponding quadratic form. There is a theorem due to M. Morse which states that at a critical point which is non-degenerate, the function can be made



locally equal to a quadratic form of the same type as the second differential by introducing curvilinear coordinates. Thus the contour levels for a function whose quadratic form is of the type $x^2 - y^2$ actually do look like we always say they do. Although this theorem is not very hard to prove (it is accessible to anyone who understands the inverse-function theorem), it is not widely known. Since it really clarifies the picture it ought to be part of the third year calculus course.

Convexity Approach to Maximum-Minimum Problems. We now take a slightly different approach to these problems. The Fréchet derivative at a point was a linear function. Thus we took derivatives, so to speak, in all directions and combined them in one object. But maximum-minimum problems in affine space are essentially one-dimensional problems. To see this, let us recall the following one-dimensional result:

If f is defined in an interval, and if $f''(x) \geqq 0$ at each point of this interval, then f is convex: $f(\theta x + (1-\theta)y) \leqq \theta f(x) + (1-\theta)f(y)$ for $0 < \theta < 1$. Furthermore, if $f'' > 0$, then this inequality is strict except for the trivial case $x = y$.

Geometrically, the graph lies beneath each of its chords. The proof is a rather easy consequence of the mean value theorem, and its geometric significance in terms of increasing slope is also equally clear. This result has an easy extension to many variables. Let f be a function defined on a convex set in an affine space. Suppose that on every line, the second derivative is non-negative (or strictly positive). Then in that case, f will be convex along every line. But this implies that f is a convex (or strictly convex) function. Such functions have very convenient properties from the point of view of minimum problems. For example, their contour levels enclose convex sets. For such functions we have the following result, which is very easy to prove.

If a convex function has a minimum, then the set of values where this minimum is achieved is a convex set. Furthermore, a strictly convex function *has at most one minimum point. For convex functions, a local minimum point is a minimum point in the large. Any point where* $df = 0$ *is a minimum point.*

We note that all of this information can be obtained by considering functions of one variable. We do not care about uniformity with respect to the various directions in which these derivatives are taken. The arguments used here are so general that they apply with equal ease to infinite dimensional spaces.

Summarizing, if f is defined on an affine space, or on some convex set in an affine space, then the test for a critical point is as follows: x_0 is a critical point of f if $(df)_{x_0}(g) = 0$ for every g in the director space. In more familiar language, $\frac{d}{d\lambda} f(x_0 + \lambda g)\big|_{\lambda=0} = 0$ for every direction g. This is a necessary condition for an interior point x_0 to be a minimum point. The function attains a minimum at x_0 if x_0 is a critical point and $(ddf)_x(g, g)$ is positive semidefinite for each x in the space and each direction g in the director space. In more familiar terms,

$$\frac{d^2}{d\lambda^2} f(x + \lambda g)\big|_{\lambda=0} \geqq 0$$

for every x in the space and every direction g. If this is a strict inequality for all $g \neq 0$, then x_0 is the unique minimum point. If in the positive semidefinite condition, x is restricted to some neighborhood of x_0, then we obtain the conditions for a local minimum. Note that the positive semidefinite condition is required to hold at all points of space, or in any event, for all points in a neighborhood of x_0 if only a local minimum is required.

Before leaving the subject of maxima and minima in two variables, I would like to point out that some books give the condition of positive definiteness at a critical point but do not assume that the function is C^2. Thus, positive definiteness is not implied near x_0. This is wrong, since it is possible to construct a function which has a local minimum along every line from a point, but which does not have a minimum there, since one can sneak in on the point along some parabola.

Infinite Dimensional Case: A Problem in the Calculus of Variations. We can introduce the functions df and ddf in infinite dimensional situations, but certain properties do not readily go over. For example, if Q is a strictly definite quadratic form, it is not true that a small perturbation of Q is positive definite. For example, in the space of square summable sequences $x = (x_i)$, where $\|x\| = \sqrt{\Sigma x_i^2}$, the function $Q(x) = \Sigma \frac{x_n^2}{n}$ is positive definite, but $Q(x) - \epsilon\|x\|^2$ is not positive definite for any $\epsilon > 0$. Thus even continuity of ddf is not enough to verify that x_0 is a minimum by considering ddf at x_0 alone.

However, we can illustrate the convexity approach in the following elementary theorem from the calculus of variations, which is often thought to be inaccessible to the student in some of its details.

Theorem. Of all C^1 functions f defined in the unit interval $[0,1]$ and satisfying $f(0) = 0$ and $f(1) = 1$, the unique function which has the shortest length is the straight line function $f(x) = x$.

Proof. In the class C^1 of functions defined in $[0,1]$, let A be the class of functions f such that $f(0) = 0$ and $f(1) = 1$. This is seen to be an affine subspace of C^1, and its director space is the set V consisting of functions g with $g(0) = g(1) = 0$. Geometrically, A is a coset of V, and V has co-dimension 2 since it is given by the intersection of the two hyperplanes $E_0(x) = 0$ and $E_1(x) = 0$. (A is given by $E_0(f) = 0$ and

$E_1(f) = 1.$) The linear functions E_i are evaluation functions: $E_i(f) = f(i)$, and E_0 and E_1 are evidently linearly independent.

The length function $\varphi: A \to R$ is given by the formula

$$\varphi(f) = \int_0^1 \sqrt{1 + f'(t)^2}\ dt\ .$$

We now calculate $(d\varphi)_f(g)$, and we do so along a line through f in the direction g. We have

$$\begin{aligned}(d\varphi)_f(g) &= \frac{d}{d\lambda}\varphi(f + \lambda g)\ \Big|_{\lambda=0} \\ &= \frac{d}{d\lambda}\int_0^1 \sqrt{1+(f'+\lambda g')^2}\ dt\ \Big|_{\lambda=0} \\ &= \int_0^1 \frac{f'(t)g'(t)}{\sqrt{1+f'(t)^2}}\ dt\ .\end{aligned}$$

Here, all that is required is the knowledge that we can differentiate under the integral sign and the ability to perform the required integration. The condition for a critical point is $(d\varphi)_f(g) = 0$ for all $g \in V$. Now it is an easy matter to verify that if f is a critical point, i.e., if

$$\int_0^1 \frac{f'(t)g'(t)}{\sqrt{1+f'(t)^2}}\ dt = 0$$

for all g with $g(0) = g(1) = 0$, then $f'/\sqrt{1 + (f')^2}$ is constant. The classical argument which proves this by an integration by parts is inadequate, since it assumes that this function is in C^2. But there is an easy argument which gives the result quickly. [Editorial note: The argument is given in the discussion section, following the lecture.] It follows that $(f')^2$ = constant. It follows that f' = constant. (f' is continuous; since it takes on at most two values, it is constant.) Finally $f(x) = cx$ and $c = 1$ since $f(1) = 1$. Thus $f(x) = x$ is the unique critical point for φ.

Now it is enough to verify that ddf is a positive definite quadratic form in any direction. The calculation (the so-called second variation) is simply

$$(dd\varphi)_f(g,g) = \frac{d^2}{d\lambda^2}\varphi(f + \lambda g)\Big|_{\lambda=0} = \int_0^1 \frac{g'(t)^2}{(1 + f'(t)^2)^{3/2}}\,dt.$$

This is clearly non-negative and can only be zero if $g'^2 = 0$ or $g' = 0$. Since $g(0) = 0$ we see that this quadratic form is 0 only if $g = 0$. Thus $dd\varphi$ is positive definite everywhere and we have the result: $\varphi(f)$ is simply a strictly convex function over A and has its unique minimum at its critical point.

Discussion.

Gleason remarked that one easy variation on this problem is to minimize $\psi(f) = \int\sqrt{1 + f'(t)^2}\,h(t)\,dt$ over the same class of functions. It is necessary to have $h > 0$ but, if this is so, the reasoning yields the differential equation $f'h/\sqrt{1 + f'^2} = $ constant. Otherwise the argument is the same. In the classroom it is convenient to take h as the reciprocal of a polynomial to reduce the algebra.

The lemma referred to in the lecture was stated and proved as follows.

Lemma. If ψ is a continuous function in $[0,1]$, and if $\int_0^1 \psi(t)g'(t)dt = 0$ for all $g \in V$, then $\psi = $ constant.

Proof. Define the linear functional Ψ by the equation

$$\Psi(h) = \int_0^1 \psi(t)h'(t)\,dt \qquad (h \in C^1).$$

By hypothesis Ψ vanishes on the subspace V. Therefore, since V is given by the two hyperplanes $E_0(x) = 0$ and $E_1(x) = 0$, it follows by an easy linear argument that $\Psi = \alpha E_0 + \beta E_1$. In more familiar terms,

$$\Psi(h) = \alpha h(0) + \beta h(1).$$

But clearly $\Psi(1) = 0$. Using the constant function 1 for h we obtain

$$0 = \alpha + \beta, \quad \alpha = -\beta.$$

Hence

$$\begin{aligned}\Psi(h) &= \beta[h(1) - h(0)] \\ &= \beta \int_0^1 h'(t)dt.\end{aligned}$$

But using the definition of Ψ, this yields

$$\int_0^1 (\psi(t) - \beta)h'(t)\, dt = 0 \quad \text{for all} \quad h \in C^1.$$

But now we take h such that $h' = \psi(t) - \beta$. This yields $\psi(t) - \beta = 0$, which is the result.

Johnson pointed out that the incorrect method of obtaining the critical function using integration by parts yields an answer which may then be directly verified to be a critical point. However in a non-positive definite case some critical points might conceivably be lost by this method.

Coxeter remarked, after seeing the Morse movie on pits, peaks, and passes, that a simple connection with Euler's formula was just missed. For suppose that a smooth dry planet has V pits, F peaks, and E passes. From each pass drop two heavy balls, one on each side, and consider the pattern formed by the tracks of all the balls. This is a "graph" having V vertices (each having a valency equal to the number of balls in the pit, possibly only one) and E edges (possibly one of them forming a loop or two of them joining the same pair of vertices). This graph decomposes the surface of the planet into F regions, one surrounding each peak. Therefore $V - E + F = 2$. The discussion of saddle points then led to a discussion of the Morse theory in higher dimensions and in infinite dimensional spaces.

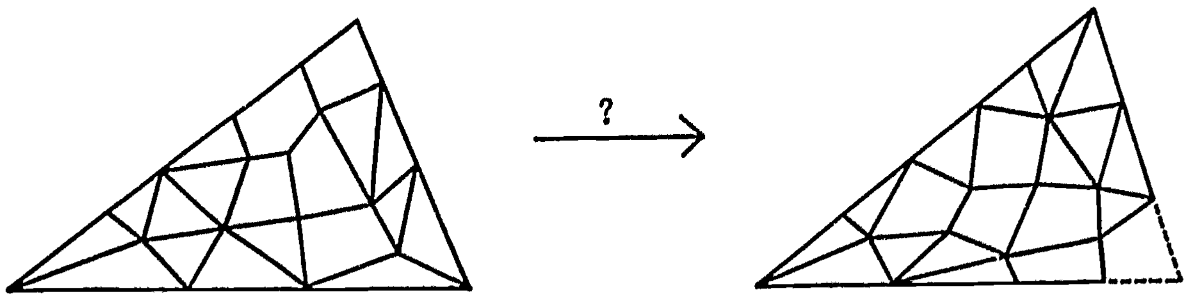
After the discussion, Gleason gave the following example (due to Whitney) of a function of three variables x, y, and t which is of order 5 at the origin but which cannot be changed into a polynomial upon introduction of curvilinear coordinates. Take $w = x(x + y)y(x + ty)(x + y(\sin t))$. Then, regardless of coordinates, the five hyperplanes $x = 0$, etc., can be recovered at the origin. Using cross-ratios on the hyperplanes $x = 0$, $x+y = 0$, $y = 0$, $x+ty = 0$, it will be possible to recover t. Similarly, it will be possible to recover $\sin t$. But clearly this is not possible for a polynomial function, since the sine is not an algebraic function.

Lecture IV.

In this lecture I should like to discuss area and volume, and some of their ramifications. I think the first thing that we should all do is to admit that the notion of area in the plane is not trivial, even if we restrict our attention to polygons. The calculus texts all take this notion for granted and then they proceed to the notion of areas of curved regions. Let us first consider some of the problems concerning area of polygons in the plane.

Area of Polygons. The area of a polygon is, of course, a certain function from the set of polygons into the set of non-negative real numbers. Congruent polygons must have the same area. Furthermore, if a polygon is cut into pieces, then the sum of the areas of these pieces is supposed to be equal to the area of the polygon. Although everybody believes that such a function exists, the belief is based, for most people, on faith and their experience with painting surfaces, transferring water from one vessel to another, etc. The question of the existence of such a measure thus arises.

The critical question is this. Suppose that a triangle is cut into many pieces. How do we know that it is impossible to rearrange the pieces in such



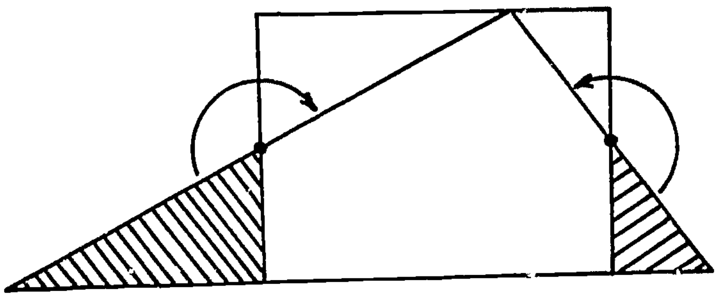
a way that we arrive at the same triangle with a chunk removed? If this could be done, then the only area function with the desired properties would be the trivial zero function. That there is a non-trivial problem here becomes quite clear when we pass to higher dimensions. The idea of physical conservation becomes meaningless for dimensions higher than three, while the Banach-Tarski example shows that even our intuition for 3-space is in error when very complicated sets are involved.

In the beginning calculus course it would take quite a long time to prove the existence of such an area function. The result is not necessary for the proof of the purely analytic theorem on the existence of the Riemann integral, but is essential for the geometric interpretation of the integral as an area. I see no reason why we shouldn't point out to our freshmen that the problem is important and non-trivial and then go on without offering a proof.

On the other hand, it is extremely easy to verify that the area function is unique, once we grant that the unit square has area one. For in this case, we can find the area of a square of dimensions 1/n by 1/n, then a rectangle with rational sides, and then a rectangle with real sides. Then the area of a triangle may be found by the usual method of Euclidean geometry. Finally, a polygon may be decomposed into triangles.

In this connection, I might mention that this approach actually needs a much smaller group than the full Euclidean group.* All that is required is the group of translations and half turns (rotations through 180°). We can illustrate it for the triangle in the diagram below. A similar technique works for any scalene triangle, although some care is needed. We can then transform any

* V. G. Boltyanskiĭ. Equivalent and Equidecomposable Figures, Boston, Heath and Company, 1963.



rectangle into a congruent one in standard position (one side "vertical" and one "horizontal"). Then the areas of these standard rectangles can be shown to have the usual area exactly as outlined above.

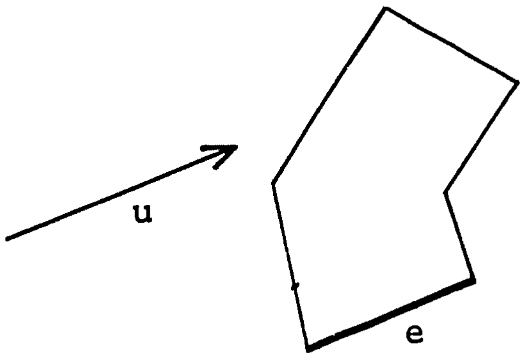
As a sidelight, we can mention that it is impossible to transform a triangle into a rectangle using such a cutting up and pasting procedure and only the group of translations. To prove this we shall first construct a function which is invariant under cutting up, pasting, and translating. Let u be a unit vector in the plane. Corresponding to u and any edge e of a polygon Π, define the number $A(u, e, \Pi)$ by the formula

$A(u, e, \Pi) = 0$ if the edge e does not have the same direction as u.

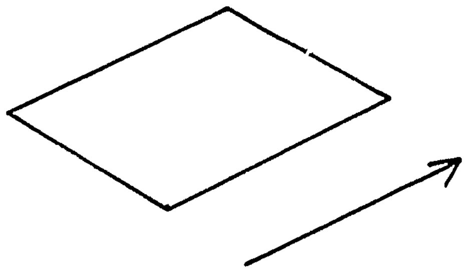
$A(u, e, \Pi) =$ length of e if the area of the polygon Π lies to the left of e when e is oriented in the same way as u.

$A(u, e, \Pi) =$ the negative of the length of e if the area lies to the right of e when e is oriented in the same way as u.

We then define $K(u, \Pi) = \Sigma_e A(u, e, \Pi)$, the sum extending over all of the edges of Π. It is then an easy matter to see that, for fixed u, $K(u,\Pi)$ is



$K(u,\Pi)$ = length of e.



$K(u,\Pi) = 0$, all u.

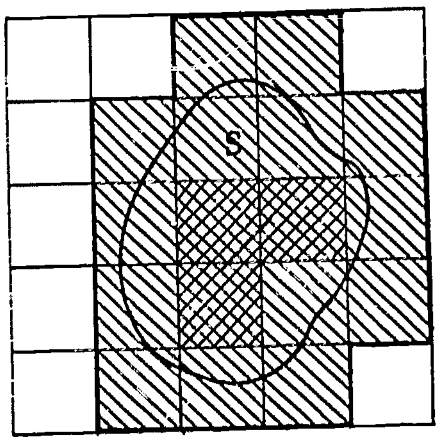
invariant under all translations. It is also seen that $K(u, \Pi)$ is invariant under the cutting and pasting operations. Thus the function $K(u,\Pi)$ is an invariant under the operations we have been considering. Now note that $K(u, \Pi) = 0$ for all u if Π is a rectangle, but $K(u, \Pi) \neq 0$ if Π is a triangle and if u is parallel to one of its sides. This completes the the proof.

Jordan Content. I would suggest that when measure is finally introduced to the calculus student the Jordan content should be developed in some detail before the Lebesgue theory. There are several reasons for this. First, it is easier. This hardly needs any elaboration. Secondly, it is more natural. The kinds of sets which arise in Lebesgue theory are unnatural when taken out of context. Of course, when a subject requires Lebesgue measure, such as the study of Fourier series, etc., Lebesgue measure should be introduced. But even in Fourier series, the student can easily be carried through a good part of the subject--for example, the convergence properties for decent functions--

before the need arises. At some point the student can honestly understand the necessity for such concepts as almost everywhere, etc., and Lebesgue measure can be introduced. Also, the theory of Jordan content is quite accessible to the student who has only the most modest idea of topology in n dimensions. Once the notion of the closure, the boundary, and the interior of a set are known, no more topology is needed.

Let us recall how we define Jordan content in n dimensions. The basic idea is that a bounded set will have a well-defined Jordan content provided that its boundary has Jordan content zero. Also, Jordan content zero may be defined in terms of finite coverings. This is a very reasonable criterion. Thus it says that a bounded set has Jordan content if its boundary isn't "thick." It includes all of the bounded sets which come up in an elementary course. It avoids all of the pathologies of a more advanced measure theory.

Briefly, here is the way I find most convenient to introduce Jordan content. This is done in n dimensions, but we can illustrate in the plane. Wc take a fixed partition of R^2, say that determined by the hyperplanes x_i = integer. We then refine this partition into squares of side ½, refine again into squares of side ¼, etc. It is enough to work with these fixed partitions of the plane, each of which is a refinement of the previous one. Now if S is any bounded set, and we work with any one of these partitions, we count how many of the squares are contained entirely in the interior of S, and also how many touch the closure of S. We multiply by



a power of $\frac{1}{2}$ to obtain an inner measure and an outer measure of the set S for the particular partition. This is a highly intuitive procedure. As we go to the next partition, it is very easy to see that the inner content increases and the outer content decreases. The content is then defined if the inner and outer sums have a common limit. A bounded set is then called a J-set if it has Jordan content. We then show that a bounded set is a J-set if and only if its boundary has Jordan content 0, and that in this case the interior of the set and the closure have the same Jordan content. This is an extremely easy theorem to prove. It is then very easy to prove that the Jordan content is additive over figures whose interiors do not overlap.

It is next shown that Jordan content is invariant under translation. (Because we took a fixed sequence of partitions there is a problem in estimation involved.)

Invariance under rotation is shown by the following clever trick. We first show that the Jordan content is characterized up to a constant factor. That is, any real valued function which is defined on all J-sets and which is additive and translation invariant is a constant multiple of the Jordan content. (This is easily shown. The volume of the unit cube will determine all volumes.) We next verify that linear transformations take J-sets into J-sets. It then follows that, for a fixed linear transformation T, the function J(TE) is a translation invariant function on J-sets E. (It is here that we depend heavily on the fact that T is linear, or that the translations are a normal subgroup in the group of affine transformations.) Thus we obtain

$$J(TE) = f(T)J(E),$$

where f(T) is a function on linear transformations and E is any J-set. At this point we can show that an orthogonal transformation preserves Jordan content by taking E = unit ball. Then $TE = E$ and $f(T) = 1$. [Editorial note: An equally elegant proof is given in the discussion.]

As a bonus we can find that $f(T)$ is the absolute value of the determinant of T. All that is required is to factor T as the product of a symmetric and an orthogonal operator and then verify that, for a symmetric operator S, $f(S)$ is the product of the absolute values of its eigenvalues. (Each eigenvalue corresponds to a stretch in one direction.) Equivalently, we may factor T in any way into simpler factors and verify the determinant formula for the factors. (The elementary operations on a matrix correspond to such a factorization.)

The Riemann Integral. The integral is defined using the usual Riemann-Darboux sums. The domain of definition is broken up into small J-sets, sums are found, and limits taken in the usual way. We also verify that, for positive functions, the integral is the Jordan content (in one higher dimension) of the set which is under the graph of the function. This is a very convenient tool, since it can reduce many problems about integration to corresponding problems about Jordan content in a higher dimension. It also shows the convenience of introducing Jordan content in all dimensions.

At this point let me emphasize that I am considering "absolute integrals." These are integrals over sets and their fundamental characterizing properties, which should be stated and proved, are:

1) $f \leqq g$ implies $\int_S f \leqq \int_S g$.

2) $\int_S k = kJ(S)$ for constants k.

3) $\int_{S \cup T} f = \int_S f + \int_T f$ where S and T are disjoint J-sets.

These properties characterize the integral and they can serve as the basis for "setting up the integral" in many physical applications. They permit us to go from the discrete to the continuous in a natural way. For example, the work done by a force f(x) acting along the x-axis is not defined, by the physicist, as $\int_a^b f(x)\,dx$. Rather, he cuts up the interval into pieces and he observes that if in any one of these pieces he replaces f(x) by the largest and smallest value of f in that interval, he will get a larger and smaller answer, respectively, for what he wants. He is led to Riemann-Darboux sums in this way. Otherwise put, he is using properties 1), 2), and 3) for the work-functional, and these properties characterize work.

Note that linearity in f does not show up in any natural way. For these applications, it seems artificial. Rather, what is used is the strong property that the integral is monotone.

The Change of Variables Formula. We now come to the formula for changing variables in a multiple integral. This is undoubtedly the hardest theorem proved in the elementary calculus course. We first reduce the problem to changing variables in the computation of J(TE), where T is a diffeomorphism. In order to avoid improper integrals, or boundary difficulties, we assume that T is defined on some open set D and that $\overline{E}$ is contained in D. We prove that TE is a J-set and we then go about proving the formula.

The best proof of the formula may not be the shortest. However, it seems to me that the right way of proving this result should be based on the reason that we believe the result in the first place. Namely, if E is cut up into small cubes, each cube is mapped onto some set. Since T is almost linear in a small cube, the volume of the image set is almost $|\det dT|$ times the volume of the cube. Therefore the formula must be

$$J(TE) = \int_E |\det dT| .$$

Again, I prefer a proof based on this idea because, as a general rule, I prefer that proof of a result which takes the natural approach to a problem and carries it through. Incidentally, it is worth noting that it is sufficient to prove the inequality $J(TE) \leqq \int_E |\det dT|$. For if this inequality is applied to T^{-1} on the set TE, the equality will be proved. This remark is a great help in the analysis.

Lower Dimensional Content in Higher Dimensions. We now come to the hard part of the subject. Namely, what should we do about content of surfaces embedded in spaces of a higher dimension? All the difficult points already come up when we consider two-dimensional surfaces in three-space. Before beginning this subject, we note that the length of a curve is rather easy. We partition the curve into sub-curves, sum the lengths of the inscribed chords, and take limits in the usual way. We can then show that the answer, for a C^1 parametrization F, is the usual formula $s = \int \|F'\|$. Of course, this is also easy to motivate, since the significance of F' as the velocity vector has already been discussed.

On the other hand, this does not generalize easily to curved surfaces. We have already mentioned at the beginning of Lecture II that three points on a surface which are almost collinear when projected onto the (x,y)-plane do not necessarily determine a plane which is near the tangent plane, hence near the surface. Thus, casual attempts to approximate the surface polygonally do not work. If we beg this question, simply assume that the surface is the graph of function $z = F(x,y)$ and define surface area by the formula $A = \int \sqrt{1 + F_1{}^2 + F_2{}^2}$, then it is not at all obvious that A is invariant under a rotation of space.

How then does one proceed? I do not have a ready answer, but I feel that we ought to focus our attention on definitions which we feel are intrinsic to the problem. Therefore I recommend the following approach. Admittedly, however, there are many problems involved. We imagine the surface suspended in a three-dimensional Euclidean space. We chop it up into decent little pieces. (This needs a definition analogous to J-sets.) Each one of these pieces is then orthogonally projected into a tangent plane at one of its points. (Intuitively, we smash the surface with a hammer to flatten it. But first we lay it on a table.) We then sum and we obtain an approximate answer. We then take a limit. This ought to be the answer. There are many problems involved and I do not know of a good, smooth, attack.

I do not like the classical answer which parametrizes the k-dimensional surface in n-space using a one-one function F, from a region in R^k into the n-dimensional Euclidean space A, which is smooth and has rank k everywhere. The area is then determined in R^k by integrating the function which corresponds to the stretching of area, namely the square root of the sum of the squares of all $k \times k$ sub-determinants of a matrix representing dF. This has to be shown invariant under change of parametrization. This is possible with some work. It is much harder to show that this is independent of change of variables in the ambient space. It involves quite a complicated identity involving the invariance of sums of squares of sub-determinants.

Before concluding, I should mention why this problem is done in Euclidean space. We know that a volume is determined up to a constant factor in an affine space. This is true for every k-plane. But there is no intrinsic normalization available suitable for all subspaces in an affine space. But in

Euclidean space we can always stipulate that the volume of the unit cube (determined by an orthonormal basis) is 1. This is a uniform prescription and normalizes volume on all subspaces of a Euclidean space.

Discussion.

Whitney's book on geometric integration theory was mentioned as a possible source for some of this material. But it was felt that this book was too hard for the audience for which the material of this lecture was intended. Spivak's book was also noted.

Coxeter observed that the definition of the integral used partitions well suited to Euclidean space and asked about the situation in hyperbolic geometry. Gleason thought that this would be well suited to Haar measure. Roughly speaking, once a fixed set is decided to have unit volume, then the volume of any set is approximated by taking a small set and comparing how many translations of it are required to cover S with how many are necessary to cover the unit volume. Gleason mentioned a friend who took this Haar measure approach with youngsters, about twelve years old, with some success. Of course, more time was available to him than is customary in a calculus course.

Prenowitz pointed out that the Haar measure approach is at least capable of wide generalization, while the usual approach is a very special procedure. Gleason observed that the Haar measure approach takes quite a bit longer and is more sophisticated. Also, the partition approach has the advantage of being very algorithmic. At each stage upper and lower bounds are obtained, so the error in the approximation by Riemann sums is known. It was also pointed out that the translations formed a simply transitive abelian normal subgroup of the affine group, and it was this property which was used in the development. Thus this offers another way of generalizing the method.

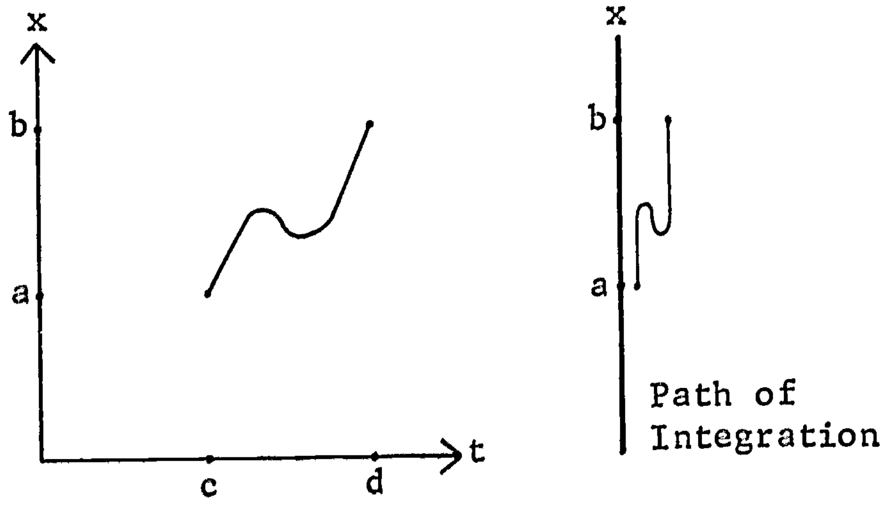
The problem of defining k-dimensional curved area in n-space from the point of view of coordinates was raised. It was thought that this approach was very natural in view of the method outlined for the change of variables formula. Gleason observed that in the change of variables situation, at least there was already a notion of content in the image space, while here it had to be defined from the beginning. The determinant identity which was needed to show that the "obvious" integral definition of area was independent of the choice of orthonormal coordinates in the large n-dimensional space was as follows. Let A be an $n \times k$ matrix, and let $\|A\|^2$ be the sum of the squares of all $\binom{n}{k}$ $k \times k$ subdeterminants of A. Let R be any $n \times n$ orthogonal matrix. Then $\|RA\| = \|A\|$. Gleason mentioned that one of the uses of exterior algebra is that such determinant identities are "swept under the rug." A natural proof of this identity would at least make the "formula definition" of area more attractive. It was observed that $\|A\|$ was equal to the k-volume of the parallelepiped formed by k vectors in n-space (modulo a simple proof), and this result easily proves the determinant identity. The question of a natural proof of this result was left open. Geometrically, it is the following generalization of the Pythagorean theorem. Let Π be a k-dimensional simplex in Euclidean n-space, and let $\|\Pi\|$ be the k-dimensional content of Π. Let orthonormal coordinates u_i be chosen. For each $j = (j_1, \ldots, j_k)$ with $1 \leqq j_1 < \cdots < j_k \leqq n$, let Π_j be the projection of Π onto the k-space spanned by $u_{j_1}, \ldots, u_{j_k}$. Then $\|\Pi\|^2 = \Sigma_j \|\Pi_j\|^2$.

Gleason noted that in the proof of the change of variables formula, when one has come down to estimating the size of TC, where C is a small cube and T is the given diffeomorphism, we can assume, with no loss in generality that dT is the identity at one point in the cube. This is so because the change of variables formula is known for linear maps, so it is always possible to apply a linear map to T and prove the result for the composition.

Yale gave the following proof that orthogonal transformations preserve volume. From the equation $J(TE) = f(T)J(E)$, we find that for reflections T, $J(E) = J(T^2E) = f(T)^2J(E)$. Therefore $f(T)^2 = 1$. Since $f(T) \geqq 0$, this gives the result for reflections. The general result follows from the simple result that any orthogonal transformation is a product of reflections.

Lecture V.

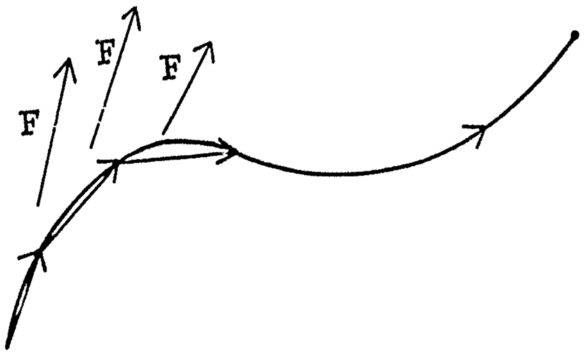
Today we shall consider the notion of a directed integral, as contrasted with an absolute integral which we considered at the last lecture. Most integrals involve chopping up the domain, weighting the value of the function we are integrating with some measure assigned to the subdomain, summing, and then taking a suitable limit as the partition grows finer. But one of the first things we do after defining $\int_{[a,b]} f = \int_a^b f$ is to permit $b < a$ by defining $\int_a^b = -\int_b^a$. This is usually done very quickly and formally. I suggest that we do a better job on this transition. It is an important step because we are passing from an absolute integral to a directed integral. The sign change is confusing and it does not clarify matters sufficiently to remark that "we count things negatively if we integrate in the wrong direction." I do not suggest that a great issue be made of this when the subject is first introduced to freshmen. However, we might remark on this transition, to the effect that for a directed integral we chop the directed interval into smaller directed intervals and use their directed lengths for the computation of the directed integral. This is a useful notion in the substitution rule, where it is usually hidden under the rug by the formalism. If we substitute



$x = x(t)$ $(c < t < d)$ to evaluate an integral $\int_a^b f(x)dx$, in some cases a di:ected integral is clearly involved (see diagram).

In higher dimensions the distinction is much clearer. For the simplest case, suppose we have a smooth curve C in the plane or in space. The absolute integral of a function f along this curve would be denoted $\int_C f\,ds$. The method of computation is to chop the curve up into little pieces, multiply a value of the function in each piece by the length of the piece, sum, and take a limit. Usually f is defined in some neighborhood of the curve and this takes place in Euclidean space so that lengths are defined. For practical purposes we can use chord length instead of arc length in the computation.

However, a line integral is based on a new idea. When we evaluate the approximating sums we must take into account the direction and sense of the chord as well as the length. The most obvious and simplest way to do this is to make this dependence linear. In Euclidean space we use a vector F and evaluate $\int_C F\cdot ds$. Thus the contribution of the vector chord is its inner product with F evaluated somewhere in the vicinity. This comes up very directly in the notion of work. F is a force field, and something is being pushed against this force from one point to another along a curve. Then the amount of work done by the field is $\int_C F\cdot ds$, using the usual approximation arguments. The Riemann sum used to evaluate such an integral uses the vector F as a weight. Its inner product is found with the small "vector chord," then sums and limits are taken. We must realize that the curve we are integrating must be taken as a <u>directed</u> curve. That is, there is a linear ordering on its points, which



may be given by a parametrization.

We first note that a vector field F was used to determine the linear operator on the directed chords, with the help of the inner product. However, in a sense this is an accident. In any vector space we always have linear operators but an inner product may not be given. Therefore we make the observation that if the curve C is in an affine space, the entire procedure can be generalized. However, instead of being given a field F of vectors in the director space V of the affine space, we must be given vectors in the dual space V^* of V.

We now review the definitions. If A is an affine space and V is its director space, a vector field is a function defined on some open set D with values in V: $D \xrightarrow{\varphi} V$. A co-vector field is a function from some open set D of A into the dual space V^* of V: $D \xrightarrow{\omega} V^*$. ω is also called a one-dimensional differential form. Then, if a co-vector field ω is given, we can integrate it along an ordered cu:ve which lies in the domain of definition of ω. The definition involves Riemann sums. The ordered curve we are thinking of is a point set with some abstractly given order on it. However, in order to allow self-intersections we can assume that we have an order parametrization for the curve. The definition of the integral involves chopping up the curve into smaller ordered curves, using the ordered end-points to obtain a vector of V, applying the value of ω in the vicinity of this smaller curve to this vector, summing, and taking limits.

We now wish to generalize this idea to higher dimensional surfaces. First, let us consider a two-dimensional surface in three-space. When the surface is chopped up into pieces, we must somehow take into consideration the orientation of each piece. All the different tangent planes for a small piece are nearly parallel, and their direction must be taken into account. For a surface in 3-space the classical device is to take a piece of the surface and to construct a vector normal to the tangent plane at one of its points. The length of this vector is taken to be the area of the little piece. However, there are two ways for this normal vector to point. Therefore some rules must be given for the orientation of the normal vector. Questions then arise concerning the possibility of doing this continuously and lead to the consideration of such objects as the Möbius band where it is impossible. However, once the normal can be consistently constructed on the surface, the method of integration is clear. $\int \varphi \cdot d\sigma$ is defined when φ is a continuous vector field and the surface is smooth. Each piece of the surface is represented by a vector orthogonal to one of its tangent planes and whose length is the area of the surface. For each piece, we evaluate φ at one point, find the inner product with the vector representing the piece, sum, and take limits.

Exterior Algebra. The techniques we have discussed do not work for k-dimensional surfaces in n-dimensional space unless $k = 1$ or $k = n - 1$. There are simply too many directions determining a k-plane unless $k = 1$ or $k = n - 1$. Therefore something new is required. I am not happy with the motivations that I know for imposing, at this point, the exterior algebras on V and on V^*. However, they can be described simply enough. To describe $\Lambda(V)$, we construct an algebra which is generated by the vectors of V, has a unit, is associative, and for which $v \wedge w = - w \wedge v$ for any vectors v and w of V. $\Lambda(V)$ is the "universal algebra" with these properties. It is the unique largest algebra with these properties. No equation is true which is not forced by the above properties. Once this is done, it is possible to decompose $\Lambda(V)$ into a direct vector space sum of homogeneous spaces $\Lambda^0(V)$, $\Lambda^1(V)$, $\Lambda^2(V)$, $\Lambda^3(V), \ldots, \Lambda^n(V)$. The elements of $\Lambda^k(V)$ are homogeneous of degree k. $\Lambda^k(V)$ is generated by the wedge products of k elements of V. In a natural way $\Lambda^0(V)$ may be regarded as the real numbers R, and $\Lambda^1(V)$ may be regarded as V itself. The dimension of $\Lambda^k(V)$ is $\binom{n}{k}$ and therefore the dimension of $\Lambda(V)$ is 2^n.

The algebra $\Lambda(V^*)$ is a little easier to define in terms of V. Namely, each space $\Lambda^k(V^*)$ is conveniently identified with the space of k-linear skew-symmetric forms on V. There are two simple ways of defining a multiplication for these forms. Both ways multiply and skew-symmetrize, but a different factor is used in the skew-symmetrizing operation. The reason is that $\Lambda^k(V^*)$ is naturally isomorphic to the space of k-linear skew-symmetric forms, but there are many natural isomorphisms. The algebraically "right" way of doing this is to consider a large space of multilinear functions of mixed degrees (the free tensor algebra) and divide by the ideal generated by all elements of the form $v \times v$. This leads to equivalence classes, and doesn't offer much motivation for the subject.

Once these constructions are made we can make the following statement, which, together with the defining relations, characterizes the exterior algebra: If W and V are vector spaces, and if φ is a linear map from W into V, then φ may be uniquely extended to a homomorphism φ of the algebra $\Lambda(W)$ into the algebra $\Lambda(V)$. In brief, the following diagram can be completed to be commutative:

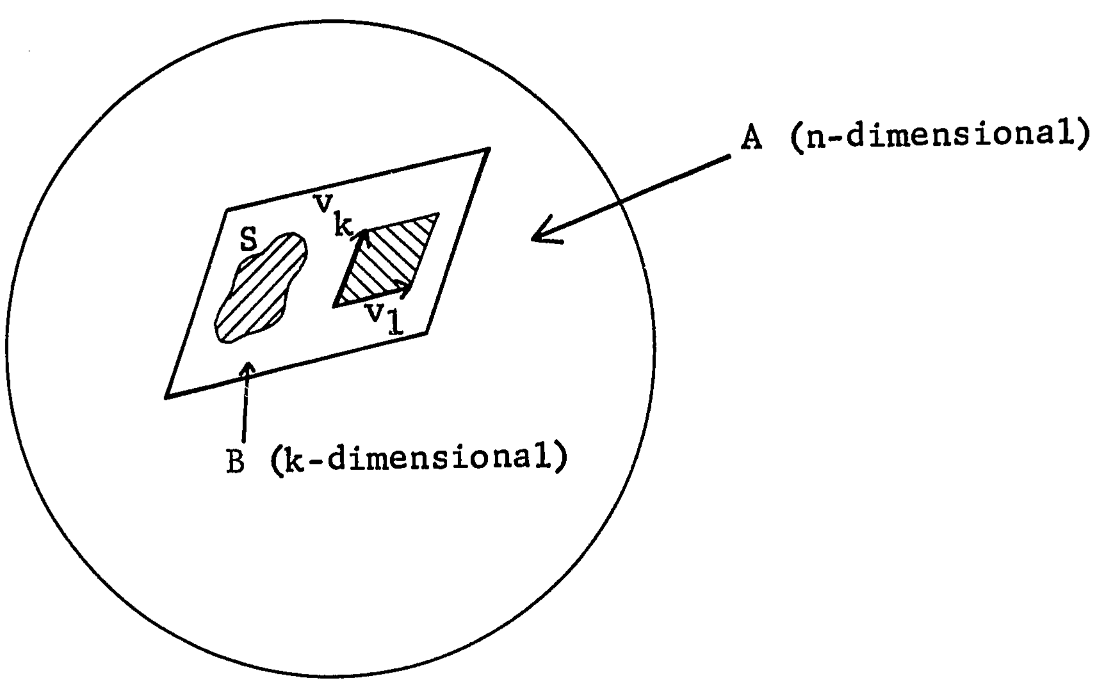
$$\begin{tikzcd}
W \arrow[r, "\varphi"] \arrow[d] & V \arrow[d] \\
\Lambda(W) \arrow[r, dashed] & \Lambda(V)
\end{tikzcd}$$

The maps from V and W into $\Lambda(V)$ and $\Lambda(W)$, respectively, are inclusion maps.

Once these spaces are constructed, there is a natural duality between $\Lambda^k(V)$ and $\Lambda^k(V^*)$. Under this duality, the inner product of $f_1 \wedge \ldots \wedge f_k$ $(f_i \in V^*)$ and $v_1 \wedge \ldots \wedge v_k$ $(v_i \in V)$ is simply $\det (f_i, v_j)$. (When the other identification of the space of skew forms with $\Lambda^k(V^*)$ is used, a factor of of $k!$ appears.)

Orientation and Vector Valued Content. We can now define what we mean by an orientation of a vector space W. Suppose W is k-dimensional. Then $\Lambda^k(W)$ has dimension one. Therefore there are two rays in $\Lambda^k(W)$ at the origin, as there are in any one-dimensional vector space over the reals. Then each of these rays is said to determine an orientation of W. Briefly, we call one of the rays the positive half. Also, if W is the director space for some affine space, this also orients that affine space. Thus if $\varphi: W \to V$ is an injection of W into V, any orientation in W induces an orientation in $\varphi(W)$, namely, the image of the ray determining the orientation in W by the natural extension of φ which was described above.

We can now define a vector valued Jordan content in an oriented k-dimensional affine space which is embedded in an affine space of dimension $n \geqq k$. (Think of the surface embedded in 3-space and go to the tangent plane.) Suppose that B is k-dimensional affine space, with director space W of dimension k, and that B is a flat in the affine space A, whose director space V has dimension n. By the above remarks, the orientation of B singles out a ray in $\Lambda^k(V)$. We shall define the vector Jordan content of a J-set in B as an element of $\Lambda^k(V)$ which is in this ray. (For planes in



3-space we went from $\Lambda^2(V)$ to a ray in V itself. This was possible because we had an inner product and an orientation of 3-space. We shall go into this duality at the end of the lecture.)

Our method is as follows. For a parallelepiped of sides $v_1,\ldots,v_k$ in B, we define the vector Jordan content to be $\pm v_1\wedge\ldots\wedge v_k$. The sign is chosen so that the answer is in the selected positive ray of $\Lambda^k(V)$. We then show that the parallelepipeds determined by $v_1,\ldots,v_k$ and by $w_1,\ldots,w_k$ have the same Jordan content if and only if $v_1\wedge\ldots\wedge v_k = \pm w_1\wedge\ldots\wedge w_k$. This statement is meaningful because Jordan content in an affine space is determined only up to a constant factor, and so a statement about equality of Jordan content can be given without reference to a specific content. The statement on wedge products can be seen to be true by using the determinant theorem on how linear transformations transform areas. Finally, if S is any Jordan set in B, we define the vector valued Jordan content of S to be the previously defined vector valued content of any parallelepiped which has the same Jordan content as S.

Surface Integrals. Once this content is defined we can define a surface integral. Suppose that A is an affine space of dimension n with director space V. In order to integrate along some k-dimensional surface in A we take D as some open set in A and we suppose that we are given a function ω from D into $\Lambda^k(V^*)$. (This is also called a k-form on D.) To integrate ω along some k-dimensional surface in D we first chop the surface up into small pieces. Each piece is not in an affine k-space, but it is almost so. In the Euclidean case we can simply project down orthogonally onto some tangent plane of the piece, but in the affine case the procedure is admittedly not clear. In any event each piece, being almost in an affine k-space almost has a k-dimensional Jordan content in $\Lambda^k(V)$. We operate on this by ω evaluated at some point of the piece to obtain a number, sum over the pieces, and take a limit. Once intrinsically defined, it is an easy matter to see how such an integral can be computed using a parametrization of the surface.

Another way of looking at this problem is as follows. We have, up to a sign, the k-dimensional Jordan content of sets in a flat space, as an element of $\Lambda^k(V)$. We wish to extend this function to curved k-dimensional surfaces in a reasonable way. We also note that in order for the above integral to be defined, it is necessary that each of the pieces have an orientation.

At this point we can clear up a matter concerning k-dimensional area. If V has an inner product, then V and V^* can be identified. Thus there is a natural way of defining an inner product in $\Lambda(V)$ and hence a norm. Now, while the vector valued Jordan content is determined only up to a sign, its norm is unique and this is the usual k-dimensional area. Thus to find the area of a curved surface, it is this norm which must be summed over the pieces. The limit is the area.

Exterior Differentiation. We now consider exterior differentiation. The following development is somewhat different from the usual one, and many objections are possible. Its advantage is that it naturally ties in with our procedure up to now. Let us denote the usual Fréchet derivative of a function f by $d_t f$ in order not to confuse it, at this point, with the exterior derivative. If ω is a k-form on A, we have

$$A \xrightarrow{\omega} \Lambda^k(V^*).$$

Taking the Fréchet derivative we have, as before,

$$A \xrightarrow{d_t\omega} \mathrm{Hom}(V, \Lambda^k(V^*)).$$

Of course, we may replace A by an open set. The elements of $\mathrm{Hom}(V, \Lambda^k(V^*))$ may be regarded as multilinear functions of $k + 1$ vectors, since we regard $\Lambda^k(V^*)$ as the space of k-linear skew-symmetric forms on V. Thus an element τ of $\mathrm{Hom}(V, \Lambda^k(V^*))$ may be regarded as the function

$$\psi(v_0, v_1, \ldots, v_k) = (\tau v_0)(v_1, \ldots, v_k).$$

This function is, however, skew-symmetric on the last k-variables but not the first. In order to find the exterior derivative of ω, we skew-symmetrize according to the formula

$$\overline{\psi}(v_0, \ldots, v_k) = \Sigma_i \ (-1)^i \psi(v_i, v_0, \ldots, \hat{v}_i, \ldots, v_k).$$

The resulting form $\overline{\psi}$ is then skew-symmetric in its $k + 1$ arguments. Hence it may be regarded as an element of $\Lambda^{k+1}(V^*)$. This is the exterior derivative $d\omega$: $d\omega = \overline{d_t\omega}$. We have

$$A \xrightarrow{d\omega} \Lambda^{k+1}(V^*).$$

The Theorems of Stokes and Gauss. We now come to the theorem of Stokes, which states that, under certain conditions,

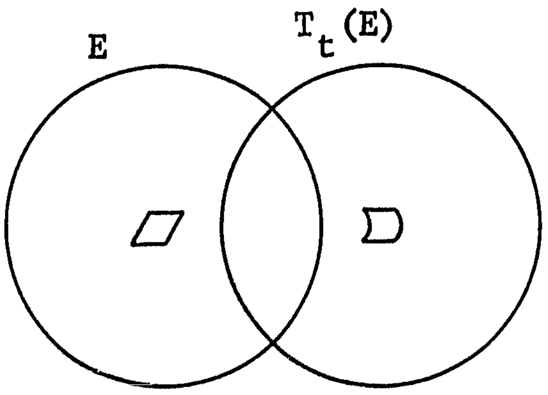
$$\int_{\partial E} \omega = \int_E d\omega.$$

I want to consider this theorem in the geometric sense, and not for the simpler case when the integration is done over singular cycles. The set E is taken to be a set whose closure is in a smooth (k + 1)-surface and whose boundary set relative to that surface is the union of finitely many sets, each of which is a J-set in a k-surface. E is assumed to be the closure of its interior (taken in the (k+1)-surface). We further assume that we can orient E. That is, we can consistently assign an orientation throughout E. (An orientation on the surface will induce one on the tangent plane.) Then there will be a natural orientation on the boundary of E. For wherever we are on the boundary, which can be expressed as a hyperplane in some curvilinear coordinate system, there are two sides of the hyperplane. These are expressible by two vectors, one pointing out of, and one pointing into, E. If the outward normal u is chosen, then we can fix the orientation in the boundary by choosing k vectors $v_1,\ldots,v_k$ in such a way that u, $v_1,\ldots,v_k$ is positively oriented. The ordering of the v's then determines the orientation on the boundary. Clearly we have a technical problem here.

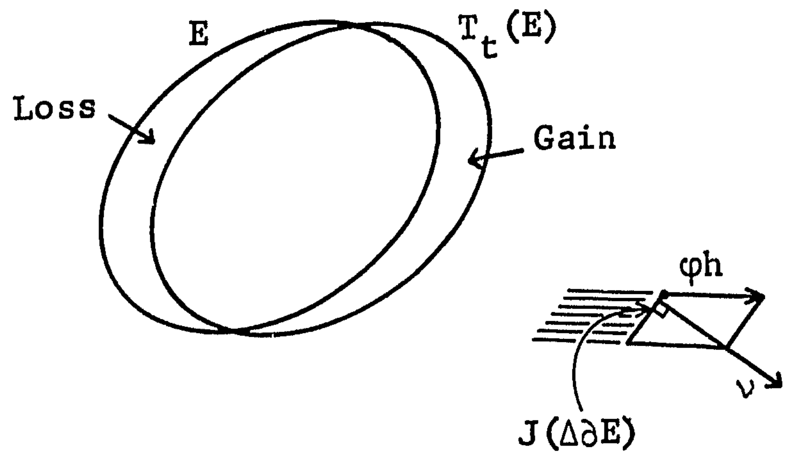
We now come to the Gauss divergence theorem, a theorem with clear geometric content, and which is almost the same theorem. In some ways, however, this theorem is a little more general. The Gauss theorem is a theorem about vector fields, not differential forms. For simplicity we shall work in Euclidean n-space. Suppose that E is a bounded region in n-space. To avoid complications, assume that E is the closure of its interior. We assume that E is bounded by a smooth variety. (This is a much harder theorem if E has corners, especially curved corners. Of course, everybody proves the theorem for a smooth boundary and then uses it for cubes.) Now suppose there is a vector field φ on E or in a neighborhood of E. Then there is a natural notion of a flow generated by this vector field. This is a function $\psi(p,t)$ whose value is a point of space. Here p is a space variable and t is a real variable representing time. The conditions on ψ are that, for fixed p, the map $t \to \psi(p,t)$ has $\varphi(\psi(p,t))$ as tangent vector. We also require that $\psi(p,0) = p$. The function $\psi(p,t)$ is called a flow. If we imagine a compressible fluid flowing in such a way that for any point p, the particle of fluid at p has velocity $\varphi(p)$, then $\psi(p,t)$ gives the position, after the elapse of t units of time, of the particle which started out at p.

We now consider the volume of E, and let time elapse. The points of E each move, and therefore we have a transformation T_t. We wish to find $\frac{d}{dt} J(T_t(E))\big|_{t=0}$.

There are two ways of calculating this derivative. They are quite different but, of course, they must lead to the same result. This gives us a nontrivial theorem. One way of calculating the derivative is to note that $J(T_t(E))$ can be calculated by the change of variables formula. As we noted, this amounts to breaking E up into little cubes and finding the volume of the image of each such cube. The global change of variable thus expresses the volume of $T_t(E)$ as an integral over E. Thus the derivative of $J(T_tE)$ is an integral over E in a natural way.



The other way of finding the derivative is to compute the change in volume directly for small t. When the set E is moved, some volume is gained and some is lost. The amount gained or lost is found at the boundary,



and can be found simply by breaking up the boundary into pieces and making the usual linear approximations. If the outward unit normal ν is constructed at each boundary point, then in a small amount h of time the net gain in volume is seen to be $J^{n-1}(\Delta\partial E)\nu \cdot \varphi h$. The sign is negative if volume is lost, so summing gives the net gain. When the details are carried out, we obtain

$$\int_E \operatorname{div} \varphi \, dJ^n = \int_{\partial E} \nu \cdot \varphi \, dJ^{n-1},$$

where ν is the outward unit normal on the boundary. This is the theorem of Gauss. It has a meaning regardless of any orientation considerations. The integrals are absolute integrals. Here div φ, the divergence of φ, is defined at any point p as the trace of the linear operator $(d_t\varphi)_p$.

This form of the Gauss divergence theorem can be extended to a non-orientable Riemannian manifold, since there is nothing in the above "proof" which involves orientation. But if orientation is assumed, then this theorem can be converted into Stokes' theorem.

Identification in Exterior Algebras. We now consider the algebraic question of what is involved when one of these theorems is converted to the other. When can we go from k-forms to k-vectors? What identifications are possible?

We have already seen that the spaces $\Lambda^k(V)$ and $\Lambda^k(V^*)$ are naturally dual spaces. It follows that *an orientation in* V *induces an orientation in* V*. For an orientation in V is simply the choice of a positive ray in

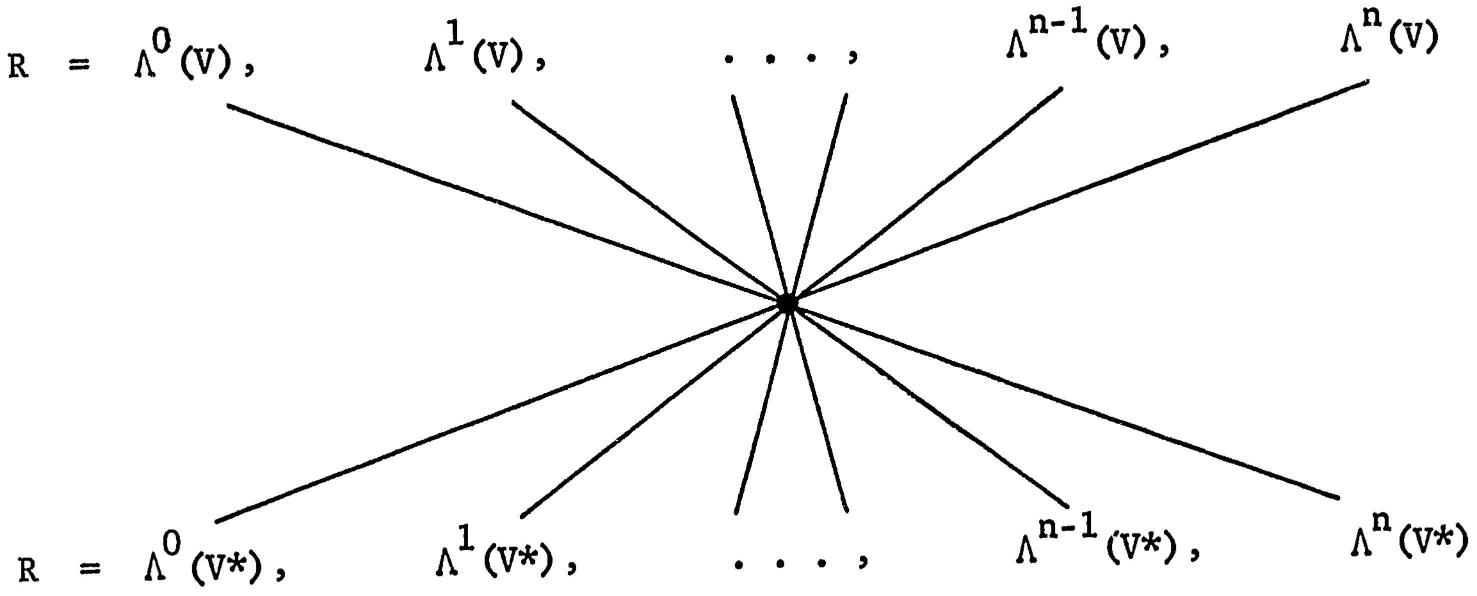
$\Lambda^n(V)$. We therefore orient V^* by choosing as the positive ray in $\Lambda^n(V^*)$ those linear functionals which are positive on the positive ray in $\Lambda^n(V)$.

If, in addition, a unit of n-volume is chosen in V, *then this determines a unit of n-volume in* V^*. For a unit of volume amounts to singling out a a certain parallelepiped and declaring it to have volume 1. As we have noted, this amounts to choosing (up to sign) a non-zero element in $\Lambda^n(V)$. But since an orientation is given, we can single out the unique positively oriented one. Thus choosing a unit of volume and an orientation in V is the same as choosing a basis vector for $\Lambda^n(V)$. Once this basis vector is chosen, we simply choose as basis vector in V^* that element whose value is 1 at the basis vector of V.

When this happens, $\Lambda^n(V)$ is identified with $\Lambda^n(V^*)$, since both are identified with the reals, as is any one-dimensional vector space once a basis is chosen. But we can say more. We have a wedge product defined between $\Lambda^k(V)$ and $\Lambda^{n-k}(V)$ with values in $\Lambda^n(V)$ which may be regarded as the real numbers. Thus these spaces may be regarded as dual spaces. However, the dual space of $\Lambda^{n-k}(V)$ is, naturally, $\Lambda^{n-k}(V^*)$. Therefore we have the identification

$$\Lambda^k(V) \cong \Lambda^{n-k}(V^*) \qquad\qquad (k = 0,1,\ldots,n),$$

provided that a unit of volume and an orientation are chosen in V. All identified spaces are indicated in the diagram on the next page; the identified spaces are connected by lines.



Identified subspaces of $\Lambda(V)$ and $\Lambda(V^*)$, given an orientation and a unit of volume in V.

If there is an inner product in V, there is no orientation implied. However, there is a unit of volume, since we may choose any orthonormal basis as determining the unit of volume. Even before the orientation is chosen, there is the obvious identification of V with V^*, hence of $\Lambda^k(V)$ with $\Lambda^k(V^*)$. Thus before an orientation of V is given we have the identification as follows.

$$\begin{tikzcd}
& \Lambda^1(V), \arrow[d, dash] & \cdots & \Lambda^{n-1}(V), \arrow[d, dash] & \Lambda^n(V) \arrow[d, dash] \\
R = \Lambda^0(V) = \Lambda^0(V^*) & \Lambda^1(V^*), & \cdots & \Lambda^{n-1}(V^*), & \Lambda^n(V^*)
\end{tikzcd}$$

Identifications if a space has an inner product. (No orientation given.)

Finally, once one of two unit vectors in $\Lambda^n(V)$ is chosen, we also obtain an orientation and we have all of the above identifications. *All homogeneous subspaces of* $\Lambda(V)$ *and* $\Lambda(V^*)$ *which have the same dimension are identified*.

In this way, the identification of the vector valued (n - 1)-Jordan content of a set located in an oriented (n - 1)-flat of an oriented Euclidean n-space is identified with a vector orthogonal to this flat whose length is the (n - 1)-content of the set. The direction of this vector depends on the orientation of the flat and of the surrounding space. Formally, we go from $\Lambda^{n-1}(V)$ to $\Lambda^{n-1}(V^*)$ to $\Lambda^1(V) = V$. The same answer is obtained if we go from $\Lambda^{n-1}(V)$ to $\Lambda^1(V^*)$ to $\Lambda^1(V) = V$. In a diagram:

$$\begin{tikzcd}
\Lambda^1(V) & \Lambda^{n-1}(V) \arrow[d] \arrow[dl] \\
\Lambda^1(V^*) \arrow[u] & \Lambda^{n-1}(V^*) \arrow[ul]
\end{tikzcd}$$

Here care must be taken in the identification! For example, V acts on $\Lambda^{n-1}(V)$ both by left multiplication and by right multiplication and, if n is even, there are two identifications possible because the wedge product is anti-commutative with respect to elements of V.

For $n = 3$ there are just two spaces left in $\Lambda(V)$ and $\Lambda(V^*)$ after all of the identifications. These are simply the scalars R and the vectors V. This explains why the classical vector field analysis did not need any forms, exterior derivatives, etc. For example, the two differential operators on the vector field φ which yield the vector field curl φ and the scalar field div φ are both found by exterior differentiation of the appropriate field of co-vectors after the appropriate identifications:

$$\begin{tikzcd}
\Lambda^1(V) \arrow[d, "i"'] & \\
\Lambda^1(V^*) \arrow[r, "d"] & \Lambda^2(V^*) \arrow[ul, "j^{-1}"']
\end{tikzcd}
\qquad\Bigg|\qquad
\begin{tikzcd}
\Lambda^1(V) \arrow[dr, "j"] & & R \\
\Lambda^1(V^*) & \Lambda^2(V^*) \arrow[r, "d"] & \Lambda^3(V^*) \arrow[u, "k^{-1}"']
\end{tikzcd}$$

If φ has values in Λ', then curl $\varphi = j^{-1}d(i\varphi)$ and div $\varphi = k^{-1}d(j\varphi)$.

We end with the remark that the people who used these operations were not in the dark about their nature. For example, they knew that one integrates the curl in connection with a surface integral and that one uses the divergence as a volume density.