DOCUMENT RESUME

ED 024 578

SE 004 992

By- Durst, Lincoln K., Ed.

Committee of the Undergraduate Program in Mathematics Geometry Conference, Part III: Geometric Transformation Groups and Other Topics.

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

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

Report No- 18

Pub Date Oct 67

Note- 174p.

EDRS Price MF-$0.75 HC-$8.80

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

Identifiers- California, National Science Foundation, Santa Barbara

This is Part III 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. The conference, attended by undergraduate mathematics teachers, involved lectures on various aspects of geometry, analyses o f material presented, and an examination of the relevance of the material to the undergraduate curriculum. In Part III of the proceedings are contained the following lectures: (1) "Transformation Groups from the Geometric Viewpoint," by H.S.M Coxeter, (2) "Two Applications of Geometry," by Herbert Busemann, (3) "Some Computational Illustrations of Geometrical Properties in Functional Iteration," by Glen Culler, (4) "Generalizations in Geometry," by Preston Hammend, (5) "The Nature and Importance of Elementary Geometry in a Modern Education," by Paul J. Kelly, and (6) "Joining and Extending as Geometric Operations (A Coordinate-Free Approach to N-Space)," by Walter Prenowitz. (RP)

COMMITTEE ON THE UNDERGRADUATE PROGRAM IN MATHEMATICS

REPORT

OCTOBER 1967

NUMBER 18

CUPM GEOMETRY CONFERENCE

PROCEEDINGS

PART III: GEOMETRIC TRANSFORMATION GROUPS AND OTHER TOPICS

Lectures by H. S. M. Coxeter and Others

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 III: GEOMETRIC TRANSFORMATION GROUPS AND OTHER TOPICS

Lectures by H. S. M. Coxeter and Others

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

This is the final volume of the Proceedings of the CUPM Geometry Conference, held at Santa Barbara in June, 1967. Part I of the Proceedings contains an Introduction by Walter Prenowitz and the lectures of Branko Grünbaum and Victor Klee on Convexity. Part II contains the lectures of Andrew Gleason and Norman Steenrod on Geometry in Other Subjects.

The texts are based on recordings made of the lectures and discussions, and were prepared for publication by the assistants, Melvin Hausner, John Reay and Paul Yale. The lecturers 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

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

Geo e 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

CONTENTS

TRANSFORMATION GROUPS FROM THE GEOMETRIC VIEWPOINT

Lectures by H. S. M. Coxeter

(Lecture notes by Paul Yale)

Lecture I. Euclidean Geometry: The group of similarities.

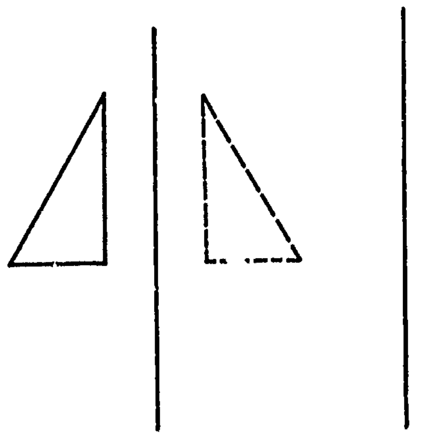
When a geometry is characterized in accordance with Klein's Erlangen program, two groups arise: a group G under which all the propositions remain valid, and a normal subgroup H under which the concepts and their properties are maintained. For instance, in the case of Euclidean geometry, G is the group of similarities (because every true theorem remains true when any similarity is applied to the figure involved) and H is the group of displacements (or direct isometries) because displacements preserve distance, area, and so on. It is accordingly desirable to classify similarities and to discuss translations, rotations, twists, and the way they can be built up from reflections.

I believe that some version of this material should be in the curriculum of our schools. I have a feeling that little bits of it should come in from the first grade on, because everything is so simple, not requiring coordinates, still less any calculus. We live in a space, so, not only for mathematicians, but also for physicists, engineers, and architects, it is desirable to have some feeling for space. Of course the first kind of space we should think about is Euclidean space. Some educators have suggested that space should bc studied in kindergarten, and the plane in first grade, since solid objects are more natural to think of than flat figures. Euclidean geometry is essentially a study of congruence and similarity, so at an early age one should see the relation between a figure and a congruent or similar figure.

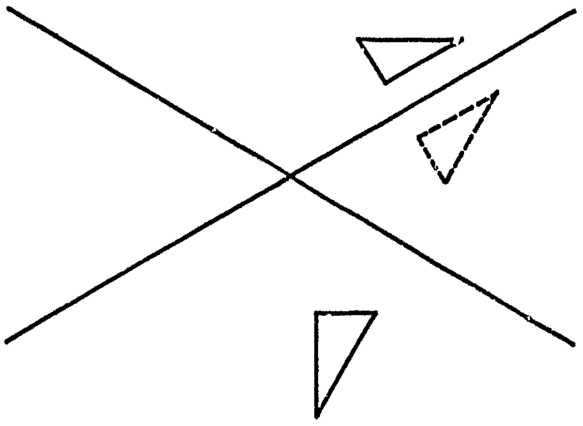
Although one would present it differently to young people, perhaps the easiest way to define a similarity (or "similarity transformation") of the

Euclidean plane is in terms of barycentric coordinates. Suppose we are given two similar scalene triangles. The transformation carrying a point with barycentric coordinates (x,y,z) for the first triangle to the point with barycentric coordinates (x,y,z) for the second is called a *similarity*. It is a one to one transformation of the plane onto itself. It has a certain ratio of magnification, μ, which means that segments are increased by the ratio μ to 1 and areas by the ratio μ^2 to 1. The similarity is *direct* if the two triangles have the same sense and is *opposite* if they have opposite senses. If $\mu = 1$, the transformation is called a congruent transformation or *isometry*. It is desirable first to classify isometries and then to classify similarities.

The first theorem is that every isometry of the plane can be expressed as the product of three or fewer reflections, a reflection being a transformation that leaves fixed every point on a line, the *mirror*, and takes points on one side of this line to points at the same distance on the other side. Reflection is an *involutory* transformation; i.e., its period is two: the image of the image is the object itself. To see that three reflections suffice, consider two congruent triangles that are related by the given isometry. You may need one reflection for the first pair of corresponding vertices, another for the second pair, and a third for the third. Since each reflection reverses sense, a product of an even number of reflections preserves sense; hence for displacements (direct isometries) the number three may be reduced to two; i.e., each displacement is the product of at most two reflections. The two mirrors may coincide, in which case the product is the identity. They may be parallel and then the product is a *translation* through twice the *distance* between the mirrors. Or they may intersect, in which case the product is a *rotation* through twice the *angle* between the mirrors. A film on this subject, called *Dihedral Kaleidoscopes*, has been made by the College Geometry Project of the



Translation



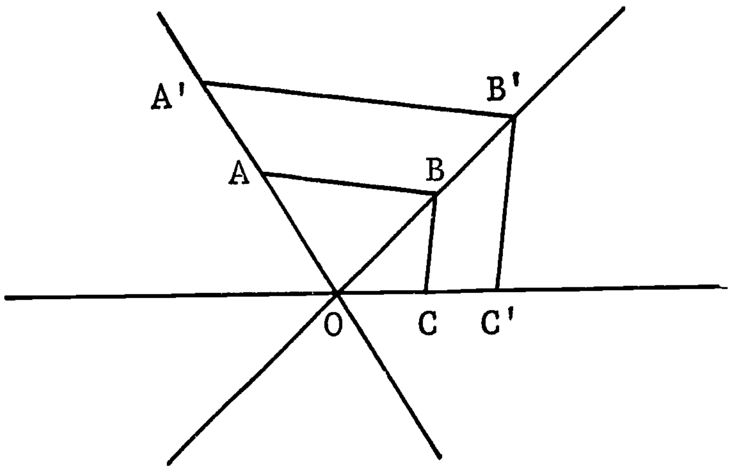
Rotation

University of Minnesota.

We now turn to products of three reflections. If the three mirrors are all parallel then the three reflections are really equivalent to one, for the translation that is the product of the first two could equally well be expressed as the product of reflections in any two lines with the same spacing, so we can take the first two mirrors and push them along until the second coincides with the third and the product then reduces to one reflection. The same sort of thing happens if we have the product of three reflections in concurrent lines. We can rotate the first two mirrors bodily around the point of intersection until the second coincides with the third and then the product of the three reflections is a single reflection.

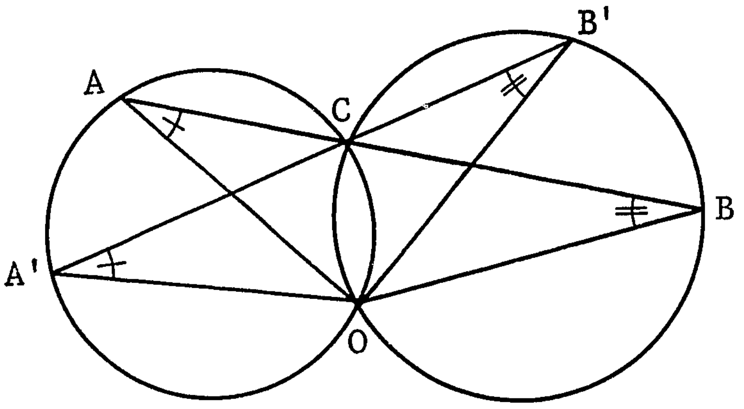
If one has the first two mirrors parallel and the third perpendicular to both, then the product of the first two is a translation and the product of all three is called a <u>glide reflection</u>. It is the operation represented by footprints in the snow: when you walk straight along a path, the relation between the left foot and the right foot is exactly a glide reflection. That is the typical opposite isometry. To see this, take three mirrors forming the three sides of a triangle. As before, the rotation which is the product of the first two reflections is the product of reflections in any two mirrors forming the same angle at the point of intersection, so we can adjust the first two mirrors so the second is perpendicular to the third. Then the product of the second and third reflections is a half-turn about the point of intersection. Thus the product of all three reflections is the product of a reflection and a half-turn, which is a glide reflection. This completes the classification of isometries in the plane.

When we pass from isometries to similarities there is the added possibility of a change of scale. The simplest change of scale is a dilatation. Using the terminology of Artin's Geometric Algebra, a dilatation is a transformation which takes every line into a parallel line; i.e., a direction-preserving transformation. The translations are the simplest dilatations and are just those dilatations with no fixed points. If a dilatation has an invariant point, O , and carries A to A', then it is easy to find how any point is transformed. This is the central dilatation with center O and ratio $\mu = OA' : OA$. Let us



denote it by $O(\mu)$ and allow μ to be negative, in which case $O(\mu)$ transforms each point into a point on the opposite side of O . In particular, $O(-1)$ is a half-turn. Any similarity can be thought of as a central dilatation followed by an isometry.

Apart from translations, every direct similarity leaves one point invariant. For example, if two maps of California are drawn to different scales on two sheets of tracing paper and superposed, then it is possible to pierce the two maps at just one point representing the same geographic position on both maps. The following simple construction determines this invariant point. If the direct similarity is not a dilatation, then we can choose two points, A and B, and their image points, A' and B', such that the two lines AB and A'B' are not parallel. Let C be their point of intersection. Consider the circle through AA'C and the circle through BB'C. Aside from a trivial case in which the circles are tangent, these two circles intersect again in another point O. The agreement of angles at A and A', and similarly at B and B', shows that the triangles AOB and A'OB' are directly similar. Therefore O is invariant for the given direct similarity. This invariant point is unique, since the only direct similarity that has more than one invariant point is the identity.



If a direct similarity S with invariant point O has ratio of magnification μ, its product with the central dilatation $O(1/\mu)$ is a direct isometry leaving O invariant, that is, a rotation about O. Therefore, S itself is a dilative rotation: the product of $O(\mu)$ and a rotation about O.

The product of the dilatation

$$x' = \mu x, \qquad y' = \mu y$$

and the rotation

$$x' = x \cos \alpha - y \sin \alpha, \qquad y' = x \sin \alpha + y \cos \alpha$$

is the dilative rotation

$$x' = \mu(x \cos \alpha - y \sin \alpha), \qquad y' = \mu(x \sin \alpha + y \cos \alpha),$$

which takes $(a,0)$ to $(\mu a \cos \alpha, \mu a \sin \alpha)$. Its n^{th} power (or n^{th} iterate) takes $(a,0)$ to

$$(\mu^n a \cos n\alpha, \mu^n a \sin n\alpha).$$

By allowing n to have all real values instead of only integer values, we can regard the dilative rotation as a continuous transformation for which the orbit of $(a,0)$ is the curve

$$x = \mu^n a \cos n\alpha, \qquad y = \mu^n a \sin n\alpha$$

or, in terms of $\theta = n\alpha$ and $c = (\log \mu)/\alpha$,

$$(1) \qquad x = ae^{c\theta} \cos \theta, \qquad y = ae^{c\theta} \sin \theta.$$

As this curve is the equiangular spiral $r = ae^{c\theta}$, the dilative rotation is sometimes called a <u>spiral similarity</u>. Allowing a to take various values between $e^{\pm c\pi}$, we obtain a "pencil" of congruent spirals, one through every point (x,y) except the origin. Thus the spirals are the orbits of all points, and not merely of a point $(a,0)$ on the x-axis.

Differentiating (1), we obtain

$$\frac{dx}{d\theta} = cx - y, \qquad \frac{dy}{d\theta} = x + cy,$$

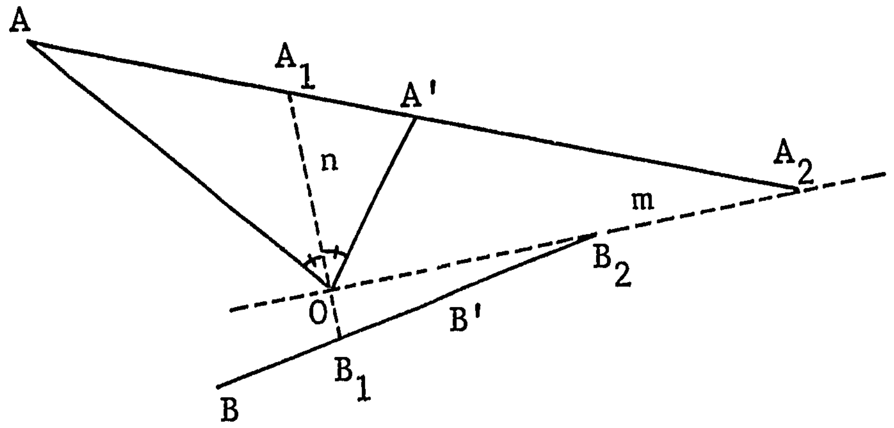
whence

$$\frac{dy}{dx} = \frac{x + cy}{cx - y}.$$

Thus the continuous spiral similarity leaving the origin invariant has for its orbits (or stream lines) the solutions of this differential equation. When $c = 0$ the similarity is simply a rotation and the orbits are concentric circles. When c is infinite the similarity is a dilatation and the orbits are straight lines through the origin.

Another approach to the construction of invariant points applies simultaneously to the two similarities, one direct and the other opposite, that transform A into A', and B into B', where $A'B' = \mu AB$ and $\mu \neq 1$. Let points A_1 and A_2 divide the segment AA' internally and externally in the ratio $1:\mu$, so that $A_1A' = \mu AA_1$ with A_1 between A and A', and $A_2A' = \mu A_2A$ with A_2 outside the segment AA'. The circle with diameter A_1A_2, called a circle of Apollonius, is the locus of all points X such that $XA' = \mu XA$. If a similarity sending A to A' and B to B' has an invariant point O, then, since $OX' = \mu OX$, O must lie on this circle. The two points of intersection of this circle and the analogous circle of Apollonius for BB' are therefore the only possibilities for the invariant points of the two similarities. Since the direct similarity has a unique invariant point, which is one of the two, the other must be invariant for the opposite similarity.

Having established the existence of an invariant point, we can easily see how it must behave and so deduce a construction for it. If an opposite similarity S with invariant point O has ratio of magnification μ, its product with the central dilatation $O(1/\mu)$ is an opposite isometry leaving O invariant, that is, the reflection in a line n through O. Therefore S



itself is a dilative reflection: the product of $O(\mu)$ and the reflection in n. In the rotation of the figure above, the mirror n, reflecting the ray OA into OA', is the internal bisector of $\angle AOA'$ and thus passes through the point A_1 that divides AA' in the ratio $OA:OA'$, which is $1:\mu$. Similarly, the line m through O, perpendicular to n, is the external bisector of $\angle AOA'$ and passes through A_2. (Since the product of reflections in n and m is the half-turn $O(-1)$, the dilative reflection is not only the product of $O(\mu)$ and the reflection in n, but equally well the product of $O(-\mu)$ and the reflection in m.) There is nothing special about A: for any point on neither n nor m, the segment joining it to its image is bisected internally by n, and externally by m, in the ratio $1:\mu$. Hence, the "axes" n and m of the dilative reflection that takes the point pair AB to $A'B'$ are simply the lines A_1B_1 and A_2B_2, and the invariant point O is their point of intersection. Incidentally, these lines, being the internal and external bisectors of the same angle, are perpendicular. Using them as coordinate axes, we deduce that any opposite similarity can be expressed as

$$x' = \mu x, \qquad y' = -\mu y.$$

This completes our classification of the similarities of the plane and we have as our theorem: Any direct similarity is either a translation or a dilative rotation, and any opposite similarity is either a glide reflection or a dilative reflection.

When we go into three dimensions the situation is closely analogous. Instead of comparing two congruent or similar triangles we naturally compare congruent or similar tetrahedra. Taking one vertex at a time, we can see that any two congruent tetrahedra are related by an isometry which is the product of at most four reflections. Of course, if they are directly congruent we need two or four reflections and if they are oppositely congruent we need one or three reflections. If one point is invariant we need one fewer reflections and, in particular, a direct isometry with an invariant point is the product of two reflections whose mirrors contain that point. Two planes through a point meet in a line, so the only direct isometries with an invariant point are the rotations about axes through that point. This is a famous theorem of Euler which can be rephrased as: any orthogonal transformation with positive determinant is a rotation.

[By analyzing the products of reflections in two, three, or four mirrors the lecturer classified all isometries of Euclidean space. In particular, he showed that every direct isometry is the product of two half-turns and that if the axes of these half-turns are skew lines then the isometry is a twist (screw displacement). See Chapter 7 of Introduction to Geometry for the details.]

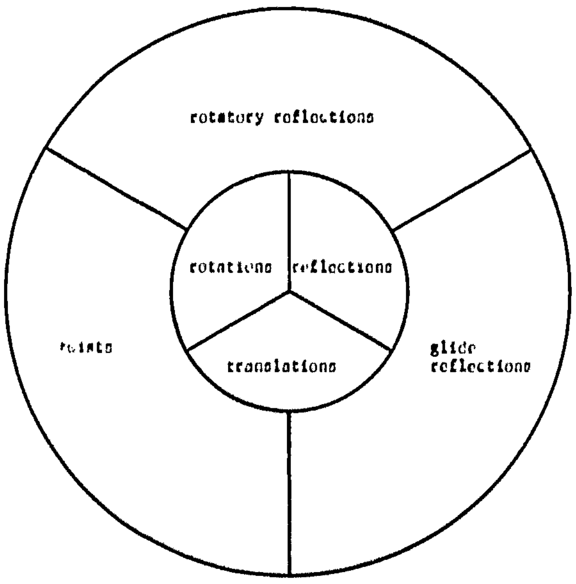
So much for isometries, now what about similarities? The following construction for an invariant point has been proposed by my former colleague at Toronto, Dr. Maria J. Wonenburger. If S is a similarity sending A to A', with ratio of magnification $\mu \neq 1$, then choose Q such that $QA' = \mu QA$ if S is direct or such that $QA' = -\mu QA$ if S is opposite. Let D denote the central dilatation with center Q and ratio $\pm 1/\mu$, plus or minus according as S is direct or opposite. The product SD is a direct isometry leaving A invariant and is therefoıe a rotation about a line, a, through A. Let α be the plane through Q and perpendicular to this line a. Both SD and D leave the plane α invariant; hence $S = (SD)(D^{-1})$ does too. The restriction of S to α is a two-dimensional similarity which by our previous discussion has a fixed point O (since we are assuming $\mu \neq 1$). But then $S \cdot O(\pm 1/\mu)$ is a direct isometry leaving O and α invariant, that is, a rotation R about the line through O perpendicular to α. We conclude that

S is a dilative rotation: the product of the rotation R and the central dilatation $O(\pm\mu)$. Thus any similarity of space which is not an isometry is a dilative rotation. The classification of similarities in space is therefore slightly simpler than in the plane, for there are only three types: translations, glide reflections, and dilative rotations. You can split this list up in various ways according to the value of μ and the amount of rotation. (See the table on page 102 of Introduction to Geometry.)

The direct dilative rotation, $R \cdot O(\mu)$ with $\mu > 0$, like its two-dimensional counterpart, can be regarded as operating continuously. The orbit of a point of general position is now a concho-spiral: a curve on a cone which cuts all of the generators at the same angle. This is the curve one reads about in Sir D'Arcy Thompson's book, On Growth and Form, in connection with the growth of shells, for each little section of the shell is similar to the section before it.

Discussion.

Johnson explained another way of classifying isometries in terms of the three fundamental types, rotations, reflections, and translations. In the schematic diagram, each of the three isometries in the outer ring is a commutative product of the isometries in the two adjacent inner regions. He also pointed out that there are obvious generalizations to n-dimension-



al Euclidean space, e.g., every isometry of E^n is the product of at most $n+1$ reflections and every direct isometry is the product of at most $[n/2]$ rotations and perhaps one translation.

Chakerian asked about the representation of direct isometries (displacements) as products of twists in higher dimensions. Coxeter replied that the analog of a twist in higher dimensions is the product of a translation and a displacement leaving a point fixed, which in higher dimensions is not as simple as a rotation. For example, in four or five dimensions the displacements leaving a point fixed are "double rotations," so even here a twist is considerably more complicated than in three dimensions. Coxeter and Gleason commented on the fact that any displacement with a fixed point can be represented as a product of rotations whose planes of rotation are mutually perpendicular.

The relevance of the principle of contraction mappings (any mapping of a complete metric space into itself which contracts distances has a unique fixed point) to the problem of finding the fixed point of a similarity was brought out by Yale. A proof of the existence and uniqueness of the invariant point of a similarity using this type of argument appears on page 103 of Introduction to Geometry. Gleason asked if the proof presented by Coxeter in this lecture applied to spaces over subfields of the reals, especially to the rational field, or if some appeal to solving quadratic equations or to completeness was necessary. After reexamining the constructions, Coxeter and Gleason agreed that they were "linear" constructions valid for any base field of odd characteristic and that the induction step for progressing from two to three dimensions is valid in general.

A discussion of various pedagogical principles involving transformations was triggered by Prenowitz's observation that Coxeter began his lecture with a comment that spatial geometry was easier than planar for kindergarten children and ended with a demonstration that similarities were easier to classify in space than in the plane. Gleason remarked that experimental evidence shows that similarity by dilatation is recognized much more readily by younger children than is similarity involving a rotation. Benson and Kelly advocated defining transformations for small subsets of space before defining them for the entire space.

Steenrod outlined a method to study the geometric structure (as a cellular complex) of the orthogonal group, $O(n)$. Define a map from $O(n)$ to the sphere S_{n-1} of dimension $n-1$ by choosing a base point x_0 in S_{n-1} and then sending each element r of $O(n)$ to the image of x_0 under r. This is a fibration of $O(n)$ whose basic fiber is $O(n-1)$ (the subgroup of $O(n)$ leaving x_0 fixed). The inverse image of any point in S_{n-1} is the coset of $O(n-1)$ sending x_0 to that point. Suppose now that we have worked out the cellular structure of $O(n-1)$, i.e., we have the cellular structure of the basic fiber, and we want to know how to extend this to the cellular structure of the entire space $O(n)$. The standard procedure is to consider for any point of the sphere the reflection in its orthogonal plane. This defines a mapping of S_{n-1} into $O(n)$ whose range is the set of reflections. Sinc antipodal points yield the same reflection, the range is like P_{n-1}. This projective (n-1)-space intersects $O(n-1)$ inductively in P_{n-2}. The difference, P_{n-1}/P_{n-2}, between P_{n-1} and P_{n-2}, is just an (n-1)-cell, so we have an (n-1)-cell, σ, of reflections determined in this fashion, hanging onto $O(n-1)$. We form the product of σ with $O(n-1)$. This maps onto $O(n)$ and, except for certain well-recognized singularities, is one to one on the open cells of highest dimension. This yields the cellular structure of $O(n)$ directly and from this one can work out such things as its homology groups.

Lecture II. The Real Affine Plane: The group of affinities.

Today I will consider the real affine plane which perhaps most simply can be thought of as the kind of plane that was described yesterday by Walter Prenowitz in terms of join and extension, together with the three additional axioms: 1. (Intermediacy) *If three points lie on a line then one lies between the other two*. 2. (Dedekind's axiom of continuity, in terms of order) *For every partition of all points on a line into two non-empty sets such that no point of either set lies between two points of the other, there is a point in one set which lies between all of the remaining points of that set and all points of the other set*. We define parallelism in a way that makes the statement of the third additional axiom much simpler. Two lines may have no common point, or one common point, or more than one common point. In the last case they coincide entirely. In the first and last cases they are said to be parallel. 3. (Playfair's axiom) *For any point A and any line m, there is a unique line through A and parallel to m*. (If the point A is on m then this parallel is just m itself.)

Since we are assuming that for any three points on a line one point is between the other two, everything could be expressed in terms of intermediacy, e.g., we could define a line segment to be the set of points between two given points. Thus we can define an affine transformation or *affinity* as a one to one transformation of the plane onto itself which preserves intermediacy. Since we are assuming continuity, it is possible to develop barycentric coordinates and show that *any* two triangles ABC and A'B'C' are related by a unique affinity sending each point P to that point P' whose barycentric coordinates with respect to triangle A'B'C' are the same as that of P with respect to triangle ABC. In particular, the only affinity leaving all three vertices of a triangle invariant is the identity.

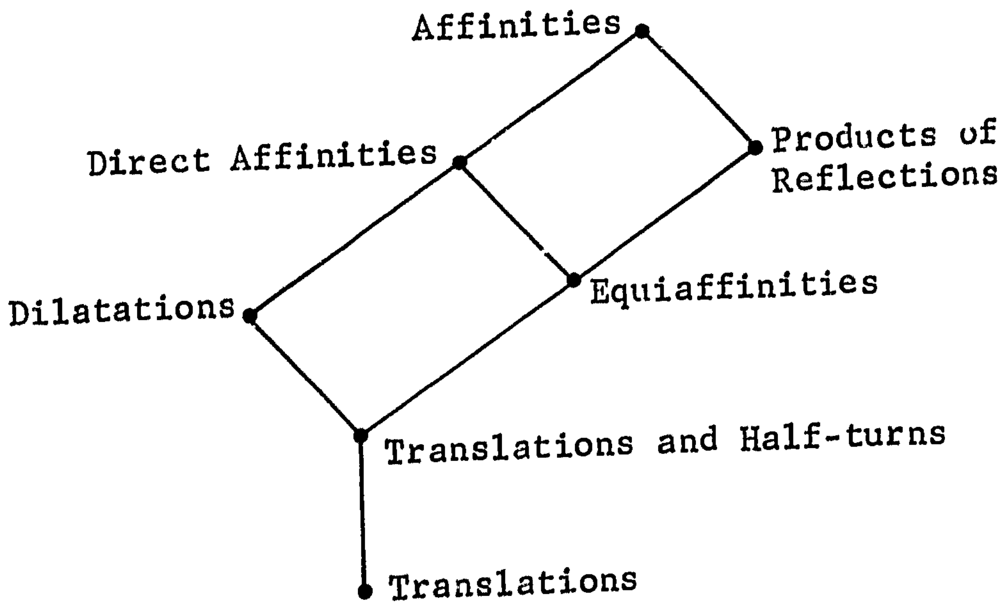
The unique affinity carrying triangle ABC to triangle ACB is called an affine reflection. A convenient symbol for this transformation is A(BC), reminding us that A is invariant while B and C are interchanged. The reflection A(BC) has a mirror, the median from A, all of whose points are invariant. A(BC) can be described as the reflection through that mirror in the direction of line BC. If A(BC) sends P to P', then the lines PP' and BC are parallel, and the lines PB and P'C' meet on the mirror. If M is on the mirror but not on PP', the symbol M(PP') has the same meaning as A(BC). In this sense, our symbol for an affine reflection is not unique.

Affinities preserve ratios of areas, and by the usual conventions the area of triangle ACB is the negative of the area of triangle ABC. Thus an affine reflection reverses the sign of every area and the product of an even number of affine reflections preserves area. A transformation preserving area (in magnitude and sign) is called an equiaffinity or (Veblen and Young) an equiaffine collineation.

Another important case of an affinity is a dilatation. Among the affinities, which map parallel lines to parallel lines, the dilatations are those that transform each line m into a line m' parallel to m. Among the dilatations we distinguish those with no invariant point, translations, from those with (at least) one invariant point, the central dilatations, exactly as in Euclidean geometry. We shall again use the symbol $O(\mu)$ for the central dilatation with center O and ratio of magnification μ, μ being a non-zero real number. Then O(1) is the identity and O(-1) is the half-turn about the pole O. We can describe O(-1) in terms of a parallelogram ABCD whose diagonals meet at O as that affinity interchanging A and C as well as

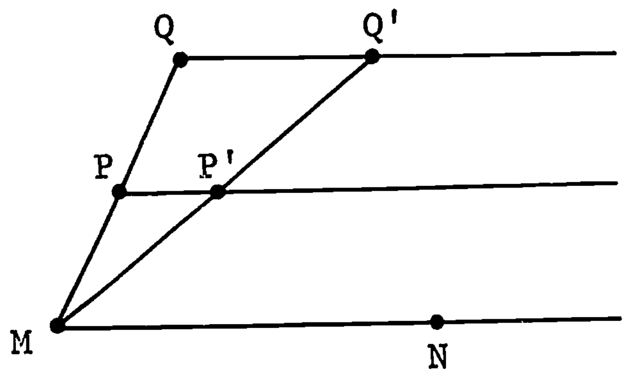
B and D. Note that the affine reflection A(BC) and the half-turn A(-1) agree on the line through A parallel to BC.

Veblen and Young [Vol. 2, p. 118] gave a table of the important normal subgroups of the group of affinities. Part of their table, with an error corrected, is as follows:

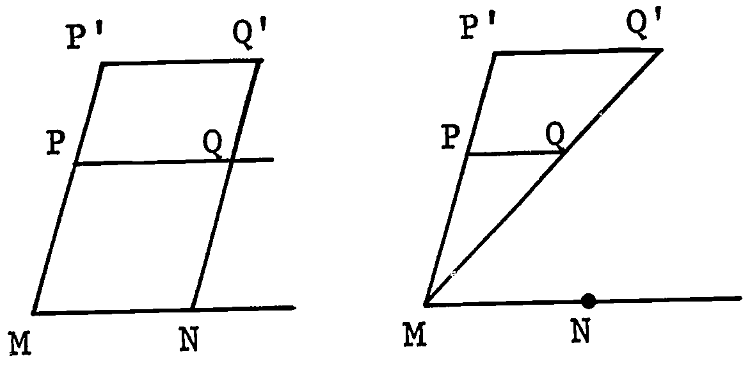


Before proving that every equiaffinity is the product of two reflections, let us investigate the possible numbers of invariant points for an affinity. Since an affinity is determined by its effect on a triangle, the only affinity leaving three non-collinear points invariant is the identity. Since ratios of lengths along a line are preserved, any affinity that leaves invariant two distinct points, M and N, leaves invariant every point on the line MN. If it transforms a point P not on this line into the point P', two cases arise according as PP' is or is not parallel to MN.

If PP' is parallel to MN, every point (not on MN) is displaced along a line parallel to MN by an amount proportional to its distance from MN, and the affinity is called a <u>shear</u>. For example, if the point Q is



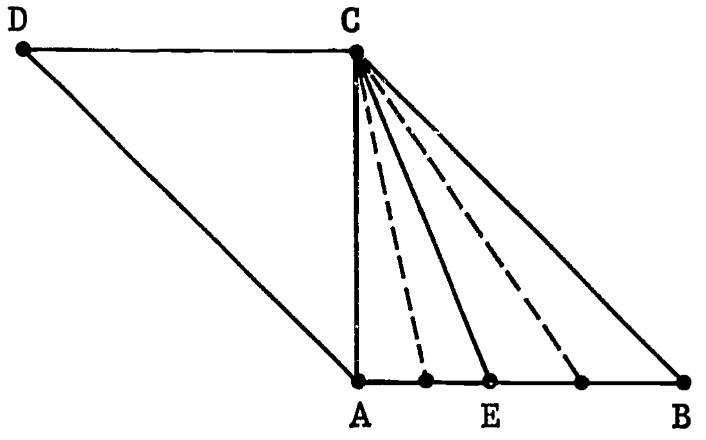
on the line MP (as in the adjacent figure), then Q' is the intersection of the line MP' and the line through Q parallel to MN. It is as if a rubber sheet fixed along MN were pulled one way on one side of MN and the opposite way on the other side. If the line MN is the x-axis of an affine coordinate system, then the coordinate form of this shear is $x' = x + \mu y$, $y' = y$. If, on the other hand, PP' is not parallel to the line MN of invariant points, then the line PQ parallel to MN is transformed into P'Q' also parallel to MN. Since the lines PP' and QQ' are left invariant, they are either parallel or meet in a point on MN, say M. The latter possibility is ruled out since the only affinity relating the homothetic triangles MPQ and MP'Q' is a dilatation for which M is the only invariant point. Hence PP' QQ' must be parallel. Thus any line parallel to PP' is invariant and the affinity is a strain in the direction of PP' leaving invariant all points on MN. If the line MN is the x-axis and the line PP' is the y-axis of an affine coordinate system, then the equations for this strain are $x' = x$, $y' = \mu y$.



To sum up, every affinity with more than one invariant point is either a shear (which reduces to the identity if $\mu = 0$) or a strain (which reduces to a reflection if $\mu = -1$). Shears and strains, like dilatations, admit a pencil of invariant lines, one through every non-invariant point. Conversely, although we shall not go into the details of the proof here, any affinity with a pencil of parallel or concurrent invariant lines must be a shear, strain, or dilatation. As a corollary of this, any equiaffinity admitting a pencil of parallel or concurrent invariant lines must be a shear, translation, or half-turn.

With these results we are now ready to present a slightly streamlined version of Veblen's proof that every equiaffinity is the product of two affine reflections. We first dispose of the easy cases, in which there is a pencil of invariant lines.

The translation sending A to B is the product of the affine reflections A(CD) and C(AB), where C and D are so chosen that ABCD is a parallelogram. Using the same parallelogram we can see that a half-turn, interchanging A and C say, is A(BD) followed by B(AC). The shear sending A to B and leaving invariant all points on the line through C parallel to AB is the product of C(AE) and C(EB), where E is the midpoint of AB (or any other point, except A and B, on the line AB). As we observed earlier, both C(AE) and C(EB) agree with C(-1) on the line through C parallel to AB; hence their product leaves invariant every point on this line.



In the case of an equiaffinity with no pencil of invariant lines, there is always a triangle ABC such that A is transformed into B, and B into C. Now suppose the given equiaffinity takes triangle ABC to BCD. We may have D = A, but even if not we must have area(ABC) = area(BCD) = area(DBC), since area is preserved both by the given equiaffinity and by the equiaffinity that cyclically permutes the vertices of a triangle. Since triangles ABC and DBC have the same area and a common "base" BC, it follows that D lies on the line through A parallel to BC. Let M be the midpoint of AD (or if A and D coincide, let M coincide with them). Then the given affinity is the product

$$M(BC) \cdot C(BD) ,$$

which clearly takes A to B, B to C, and C to D. This completes the proof that any equiaffinity is the product of two affine reflections. As an immediate corollary we see that any product of reflections can be reduced to the product of two or three reflections, and any affinity to the product of two or three reflections and a dilatation.

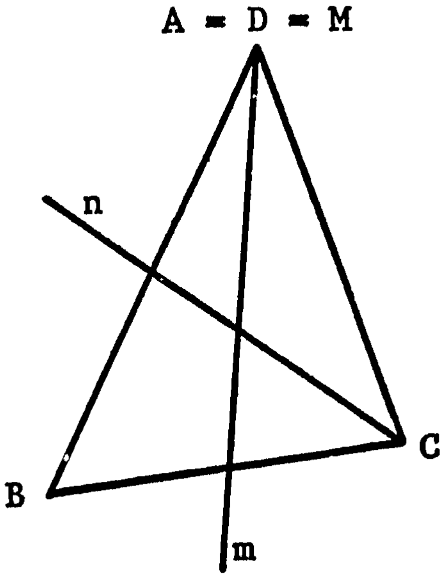
Until two weeks ago, I thought the complete classification of equiaffinities might be a rather awkward problem, but then (while sitting on a bench in Leicester Square, London) I saw something very simple which I should have thought of long ago, namely: given any triangle ABC, the various kinds of equiaffinity ABC → BCD correspond to the various positions for D on the line through A parallel to BC, and thus correspond to the various values of the ratio

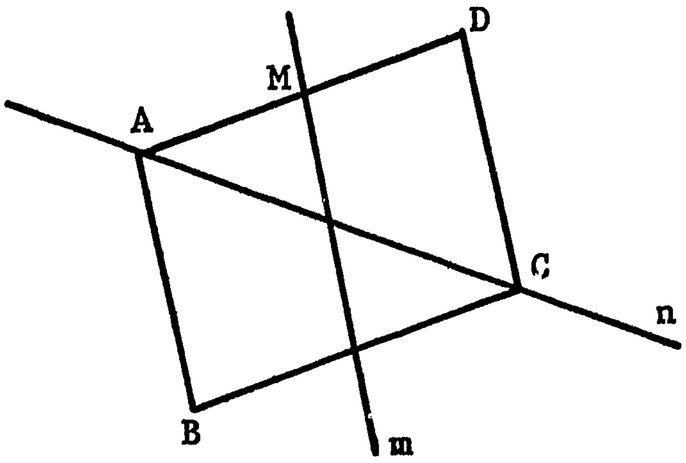
$$\lambda = \frac{AD}{BC} ,$$

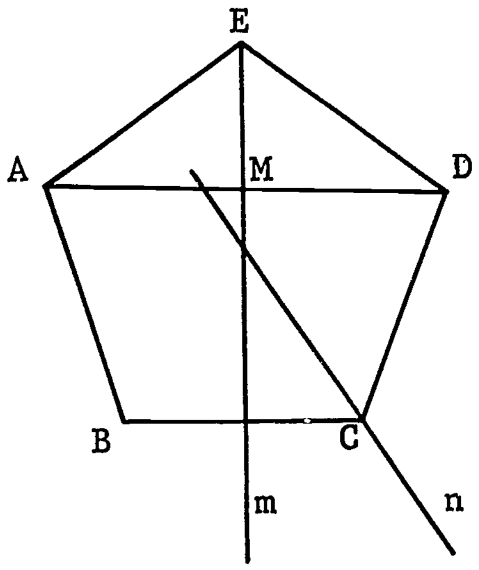
an affine invariant which may be positive, zero (when D coincides with A), or negative.

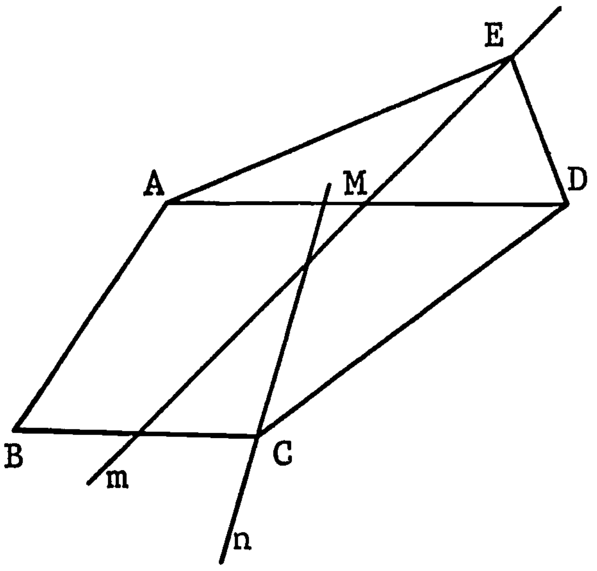
For any equiaffinity S, the successive iterates of any non-invariant point A may be regarded as the vertices of a (finite or infinite) polygon ABCD..., including, as trivial cases, the apeirogon $\{\infty\}$, which arises when S is a translation or a shear, and digon $\{2\}$, which arises when S is a half-turn so that C coincides with A. Such a polygon ABCD... is said to be affinely regular because it is transformed into itself not only by S, which generates a cyclic group, but by the two affine reflections M(BC) and C(BD), which generate a dihedral group, the complete symmetry group of the polygon. These two generating reflections, whose product is S, permute the vertices according to the permutations (BC)(AD)... and (BD)(AE)..., whose product is (ABCDE...). Thus m, the mirror for M(BC), bisects the side BC and also bisects a family of parallel diagonals beginning with AD; and n, the mirror for C(BD), bisects a family of parallel diagonals beginning with BD and AE.

By considering the special case when $\lambda = 0$, so that D coincides with C, we see that any triangle ABC is an affinely regular polygon with reflections A(BC) and C(BA). Any parallelogram ABCD with center O is affinely regular with reflections O(BC) and C(BD). A pentagon is affinely regular if and only if each diagonal is parallel to a corresponding side.



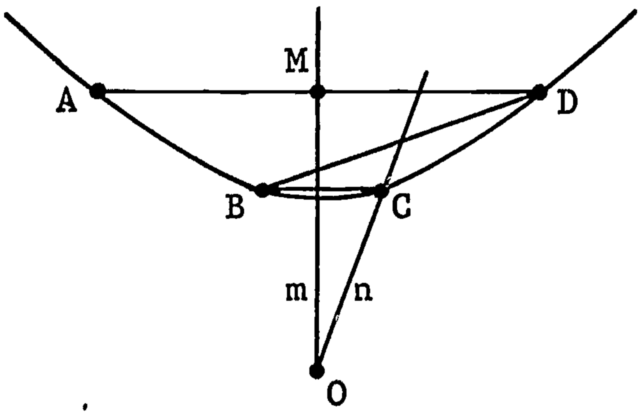


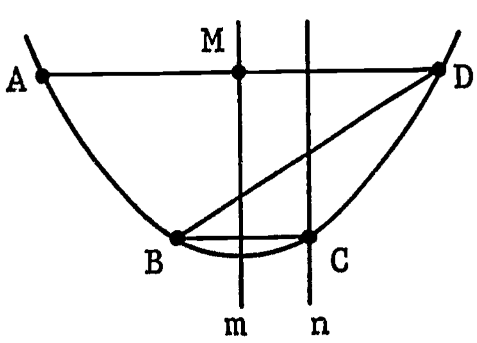




Remember that in affine geometry, although we cannot distinguish between circles and ellipses, we can distinguish between ellipses (which are closed), parabolas (whose diameters are parallel), and hyperbolas (which have asymptotes). Since lines joining midpoints of parallel chords of any conic are diameters, the conic through A, B, C with diameters m and n must pass through D (which is the other end of the chord through A parallel to BC), through E (on the chord through A parallel to BD), and so on. We see in this manner that every affinely regular polygon is inscribed in a conic, possibly degenerating into a pair of parallel lines.

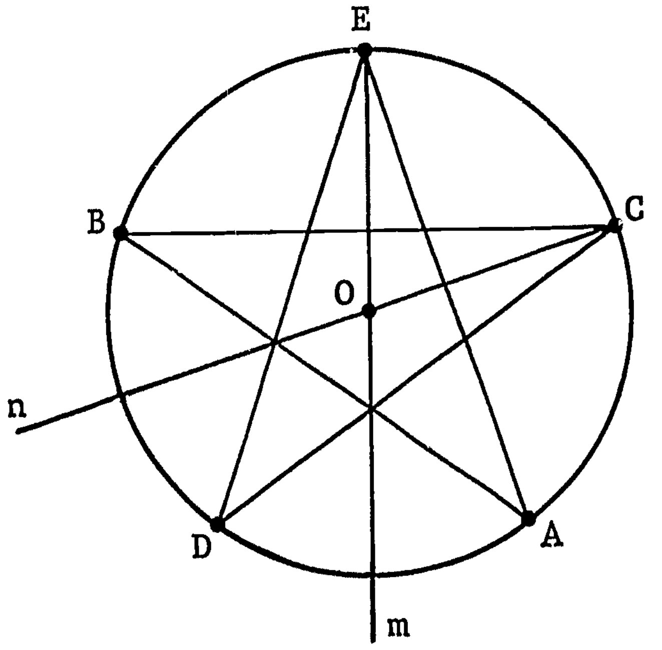
We can now distinguish five cases, depending on various possibilities for the values of the real number $\lambda = AD/BC$. If $\lambda > 3$ the diameters m and n meet in a point O on the far side of the line BC from A, and the polygon ABCD..., being inscribed in one branch of a hyperbola (with center O)



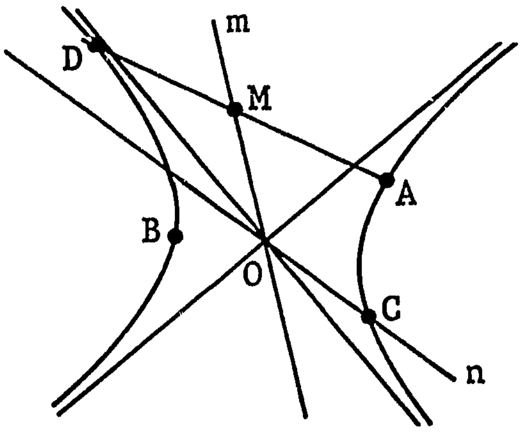


is naturally called a *hyperbolic polygon*. It is, of course, infinite. Accordingly, we call the equiaffinities with $\lambda > 3$ *hyperbolic rotations*. If $\lambda = 3$ the diameters m and n are parallel, and $ABCD\ldots$, being inscribed in a parabola, is again infinite. We call it a *parabolic polygon* and call the equiaffinity a *parabolic rotation*.

In the case $-1 < \lambda < 3$ the two mirrors meet in a point O on the same side of BC as A, and the conic with center O through A, B, and C is an ellipse. The affinely regular polygon $ABCD\ldots$ is inscribed in this ellipse, so we call it an *elliptic polygon* and call the equiaffinity an *elliptic rotation*. Since the only affinely regular polygons that close are elliptic polygons, the elliptic rotations include all possible non-involutory equiaffinities of finite period. In particular, an equiaffinity is of period three if and only if $\lambda = 0$.

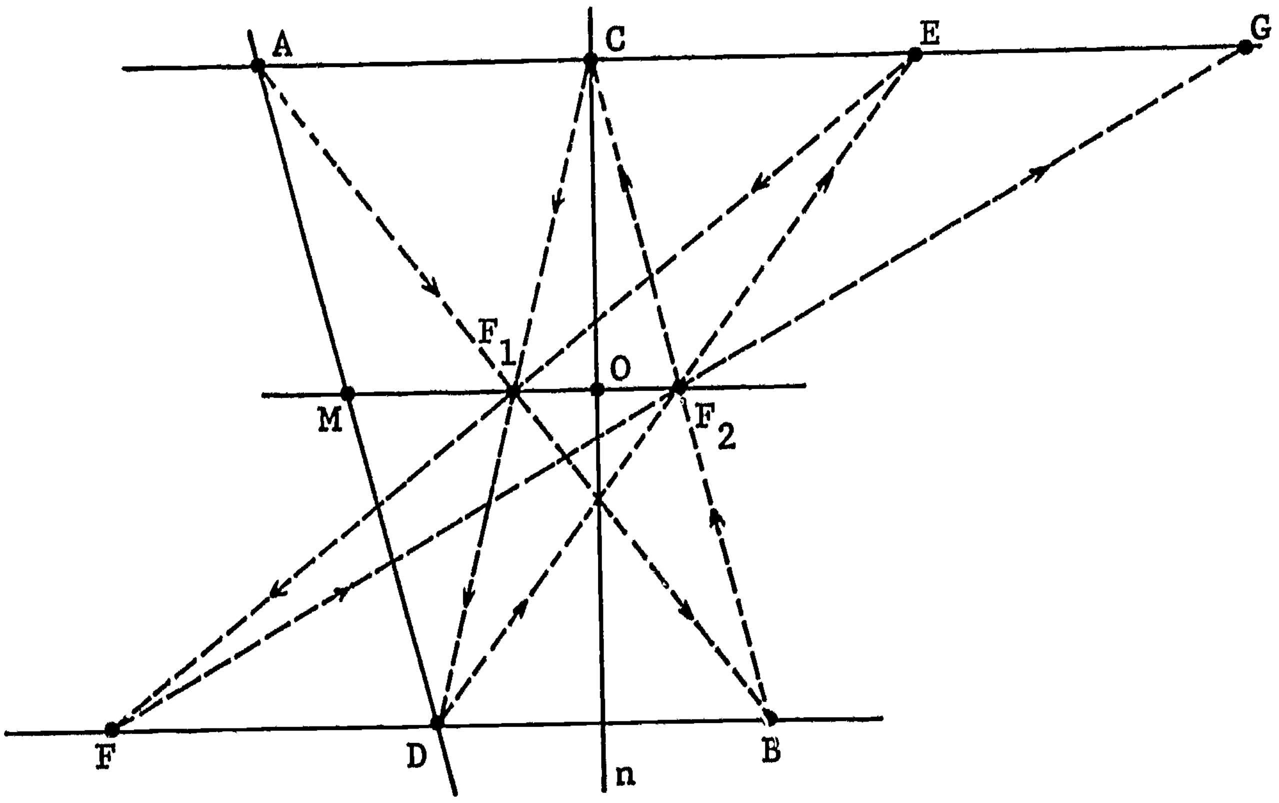


Let me defer for the moment the case $\lambda = -1$, since it is the most



amusing one, and consider the one remaining case, $\lambda < -1$. As we see in the adjacent figure, the conic is again a hyperbola, but now A and C are on one branch and B and D on the other. Thus the successive transforms of A lie alternately on the two branches of this hyperbola with ACE... progressing in one direction and BDF... in the other. For lack of a better name I call the equiaffinity a crossed hyperbolic rotation and the affinely regular polygon ABCD... a crossed hyperbolic polygon.

The equiaffinites in the interesting critical case, $\lambda = -1$, between elliptic and crossed hyperbolic rotations, I call focal rotations (following the suggestion of someone in the audience when I spoke last week to the London Mathematical Society). In this case ADBC is a parallelogram, M(BC) interchanges the two sides BD and AC, and C(BD) leaves the lines BD



and AC invariant. Thus the figure formed by the two lines AC and BD is invariant and the successive transforms of A lie alternately on these two lines. The midpoint, F_1, of AB and the midpoint, F_2, of BC are interchanged by the focal rotation and serve as the two "foci" for the focal polygon ABCD... as indicated in the diagram.

To sum up, the equiaffinities, other than translations, half-turns, and shears, may be classified into the following types:

$\lambda > 3$	$\lambda = 3$	$-1 < \lambda < 3$	$\lambda = -1$	$\lambda < -1$
Hyperbolic	Parabolic	Elliptic	Focal	Crossed Hyperbolic

For completeness we should now go on to the classification of products of three affine reflections, or, more generally, to general affinities. The conclusion seems to be that, apart from the affine counterparts of the dilative rotation and dilative reflection, the only affinities are such as can be expressed in suitable coordinates in the form $x' = \mu x$, $y' = \nu y$ for all real, non-zero, values of μ and ν. Such an affinity is of course direct or opposite according as the product $\mu\nu$ is positive or negative. I would like a synthetic proof that the general affinity is this kind of <u>double strain</u>.

Discussion.

In reply to a question by Klamkin, Coxeter elaborated as follows on the condition for an elliptic rotation to have a finite period. In terms of affine coordinates, we can take the vertices of an elliptic polygon ABCD... to be

$$(1,0), \quad (\cos\theta, \sin\theta), \quad (\cos 2\theta, \sin 2\theta), \quad (\cos 3\theta, \sin 3\theta), \ldots$$

on the ellipse $x^2 + y^2 = 1$. Then the elliptic rotation is

$$(2) \qquad x' = x\cos\theta - y\sin\theta, \qquad y' = x\sin\theta + y\cos\theta,$$

and

$$\lambda = \frac{\sin 3\theta}{\sin 2\theta - \sin \theta} = \frac{3 - 4 \sin^2\theta}{2 \cos \theta - 1} = 2 \cos \theta + 1,$$

in agreement with our inequality $-1 < \lambda < 3$. The polygon closes if and only if θ is commensurable with π, say $\theta = 2\pi/p$ where p is a rational number greater than 2. If $p = n/d$, where n and d are coprime integers, we have an affinely regular n-gon of density d, that is, a polygon $\{p\}$. In this notation, $\{3\}$ is a triangle (of any shape), $\{4\}$ is a parallelogram (given by $\lambda = 1$), $\{5\}$ is an affinely regular pentagon $(\lambda = \tau^{-1} + 1 = \tau$, where τ is the "golden section" number $\frac{1}{2}(\sqrt{5} + 1))$, and $\{5/2\}$ is an affinely regular pentagram $(\lambda = \tau^{-1} + 1 = -\tau^{-1})$ which forms, with its five lines of symmetry, the arrangement of points and lines described by Grünbaum as existing in the real plane but not in the rational plane (because τ is irrational).

Similarly, the vertices of a parabolic polygon ABCD... may be expressed as

$$(0,0), \qquad (1,1), \qquad (2,4), \qquad (3,9),\ldots$$

on the parabola $y = x^2$. The parabolic rotation is

$$x' = x + 1, \qquad y' = 2x + y + 1$$

and, of course,

$$\lambda = \frac{9 - 0}{4 - 1} = \frac{3 - 0}{2 - 1} = 3.$$

The hyperbolic polygon $(\lambda > 3)$ may be either treated like the elliptic polygon, using hyperbolic functions instead of circular functions and taking the hyperbola to be $x^2 - y^2 = 1$, or else referred to the asymptotes as coordinate axes so that the vertices are

$$(1,1), \qquad (e^t,e^{-t}), \qquad (e^{2t},e^{-2t}), \qquad (e^{3t},e^{-3t}),\ldots$$

on the hyperbola $xy = 1$. Then the hyperbolic rotation is

(3) $$x' = e^{t}x, \qquad y' = e^{-t}y$$

and

$$\lambda = \frac{e^{3t} - 1}{e^{2t} - e^{t}} = \frac{e^{2t} + e^{t} + 1}{e^{t}} = 2 \cosh t + 1 > 3.$$

Being the product of the two simple strains

$$x' = e^{t}x, \qquad y' = y$$

and

$$x' = x, \qquad y' = e^{-t}y,$$

this equiaffinity could be called a "double strain" with compensating ratios. By regarding the indefinite quadratic form xy (or $x^2 - y^2$) as the metric form for a two-dimensional Minkowskian space, we could also call this hyperbolic rotation a Lorentz transformation.

Someone asked whether the "even spacing" of the vertices of an elliptic polygon on its ellipse is determined by equal areas swept out, as in Kepler's rule for a planetary orbit. Coxeter replied that -ocal properties of conics belong to Euclidean geometry, not to affine. However, if we use the center instead of a focus, it is true. Since the vertices are permuted by an equiaffinity, the triangles AOB, BOC, COD,... all have the same area, and so do the corresponding sectors of the conic. This holds also for a hyperbolic polygon on its hyperbola. Subtracting each triangle from the corresponding sector of the ellipse, or vice versa for the hyperbola, we deduce that the "segments" bounded by the arcs and their chords all have the same area. This holds for a parabolic polygon on its parabola.

Gleason observed that the square of a focal rotation is a shear whose invariant points include the two foci (which are interchanged by the focal rotation itself). This follows from the expression of the focal rotation

(4) $$x' = -x - y, \qquad y' = -y$$

as the commutative product of the shear

$$x' = x + y, \qquad y' = y,$$

and the half-turn

(5) $$x' = -x, \qquad y' = -y.$$

In this system of affine coordinates, the focal polygon has vertices

$$(0,1), \quad (-1,-1), \quad (2,1), \quad (-3,-1), \quad (4,1),\ldots$$

and the two foci are $(\pm\frac{1}{2},0)$.

Similarly, the crossed hyperbolic rotation

$$x' = -e^{t}x, \qquad y' = -e^{-t}y$$

is the commutative product of the ordinary hyperbolic rotation (3) and the half-turn (5). Its square is the ordinary hyperbolic rotation

$$x' = e^{2t}x, \qquad y' = e^{-2t}y.$$

Killgrove asked whether there are any intrinsic reasons for the peculiar nature of a focal rotation: the only equiaffinity (without a pencil of invariant lines) in which the successive transforms of a general point do not lie on a proper conic. Coxeter replied that perhaps in part the reason for this was that we had arranged matters so that ABC was a triangle, thus excluding the digon.

There is an affinely regular star polygon $\{2 + 1/d\}$ for $d = 2,3,4,\ldots,$ which may be regarded as approximating the digon $\{2\}$ when d tends to infinity. The number

$$\lambda = 2 \cos \theta + 1, \quad \text{where} \quad \theta = \frac{2\pi}{2 + 1/d} = \frac{2d\pi}{2d + 1},$$

approaches $2 \cos \pi + 1 = -1$, which is the value for the focal polygon. Writing ϵx for x (and $\epsilon x'$ for x') in the expression (2) for an elliptic rotation, we obtain

$$x' = x \cos \theta - y \frac{\sin \theta}{\epsilon}, \qquad y' = \epsilon x \sin \theta + y \cos \theta.$$

Choosing $\epsilon = \pi/(2d + 1)$, so that $\theta = \pi - \epsilon$, and then making ϵ tend to zero, we see that the focal rotation (4) can indeed be regarded as a limiting case of the elliptic rotation.

Coxeter and Gleason discussed the similarity of the orbits of different points for a given equiaffinity. They finally agreed that the orbits of two different points on the same conic, being affinely regular polygons of the same type inscribed in this conic, are related by an elliptic, parabolic, or hyperbolic rotation according as the conic is an ellipse, parabola, or hyperbola, and that two different conics for the same equiaffinity are related by a dilatation.

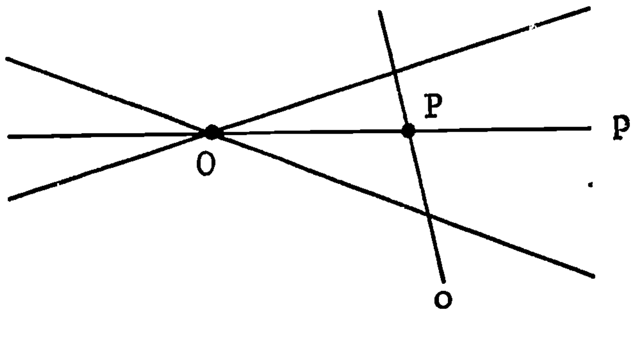
Yale considered the mapping that sends each affinity (leaving the origin invariant) to its determinant. He observed that, since this mapping is a homomorphism into the multiplicative group of non-zero reals, there are infinitely many unpleasant normal subgroups of the group of affinities, so Veblen's table is by no means a table of all normal subgroups.

Prenowitz focused attention on the possibility of developing a geometric definition of the determinant using the concept of equal area or equal volume in terms of products of an even number of reflections. Coxeter recommended Veblen's treatment of signed area as a good starting point for this. Gleason pointed out that the question (for arbitrary fields) amounts to whether or not the orbits of triangles (with respect to the group of products of even numbers of reflections) are in a natural one to one correspondence with the non-zero scalars. Everyone agreed that this is probably the situation, and the discussion closed with someone calling attention to the treatment of determinants in Artin's Geometric Algebra, a treatment which uses in place of affine reflections the fact that the equiaffine group is generated by shears (transvections).

Lecture III. Projective Geometry: The group of collineations and correlations.

I view the general projective plane as the best example of a simple and yet significant axiom system which can be developed without getting too involved in complications. I would not recommend using a complete set of axioms for Euclidean geometry in high school, for instance, but I would rather develop it informally. I think as an exercise in the use of axioms and rigid deductions there is nothing better than the general projective plane, because the axioms are so simple and the deductions fairly straightforward. The axioms I would use are that each two points determine a line, each two lines determine a point, and that there exist four points no three of which are collinear (a complete quadrangle). For most purposes I would like to include that the three diagonal points of a complete quadrangle are not collinear. Finally, I would assume what Veblen calls "Proposition P," which says that if a one-dimensional projectivity leaves three points invariant it is the identity. Of course this would require a preliminary discussion of what is meant by a projectivity. That can be done quite simply in terms of the following obvious ideas.

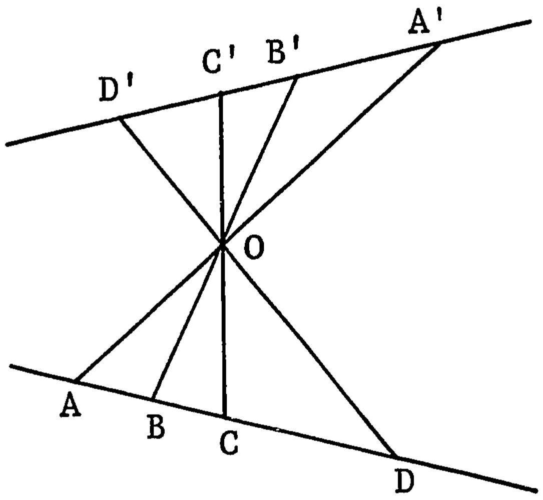
I would begin by defining an elementary correspondence as the natural relation between a pencil of lines, lines through one point O, and a range of points, which is the section of that pencil by a line o not through O. Thus I would say that each line p through O corresponds to a point P on o (and vice versa) if they are related as in the adjacent figure.



Following von Staudt, I would write $p \barwedge P$ or $P \barwedge p$. Following Poncelet, I would call any product of elementary correspondences a projectivity. This could relate a range to a range, a range to a pencil, a pencil to a pencil, or a pencil to a range. A particularly important case is the product of exactly two elementary correspondences. This special kind of projectivity, relating a range to a range or a pencil to a pencil, is called a perspectivity. Following Veblen, I would combine the symbols $p \barwedge P \barwedge p'$ or $P \barwedge p \barwedge P'$ to $p \doublebarwedge p'$ or $P \doublebarwedge P'$, respectively. In the case of a perspectivity relating the ranges ABCD... and A'B'C'D'... via the pencil of lines through O, it is convenient to write

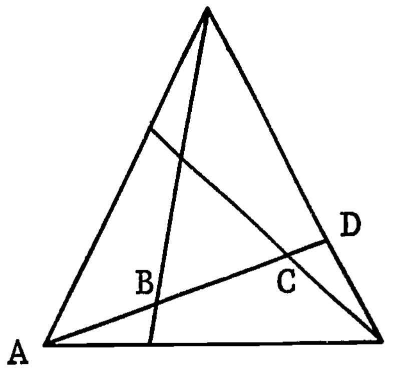
$$ABCD\ldots \overset{O}{\doublebarwedge} A'B'C'D'\ldots .$$

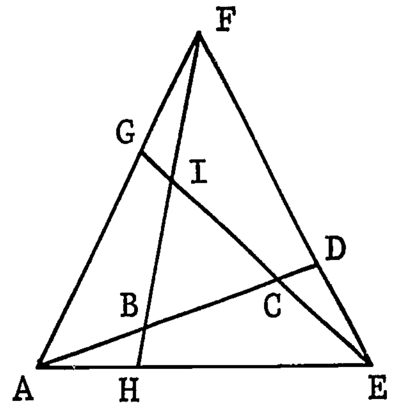
A similar convention is used for a perspectivity between two pencils. In the case of a product of more than two elementary correspondences, we revert to the original symbol $\barwedge$, since it is inconvenient to pile up the bars. But the double bar is often useful in distinguishing the special case of a perspectivity.



With several perspectivities we can of course go from one range to another, again and again, and eventually come back to a "new" range on the same line as the original range. This correspondence between "superimposed" ranges is the one-dimensional projectivity which appears in the last of the axioms, Proposition P.

As an illustration of the use of the symbols $\barwedge$ and $\doublebarwedge$, I would like to present von Staudt's important theorem that, given four points on a line,

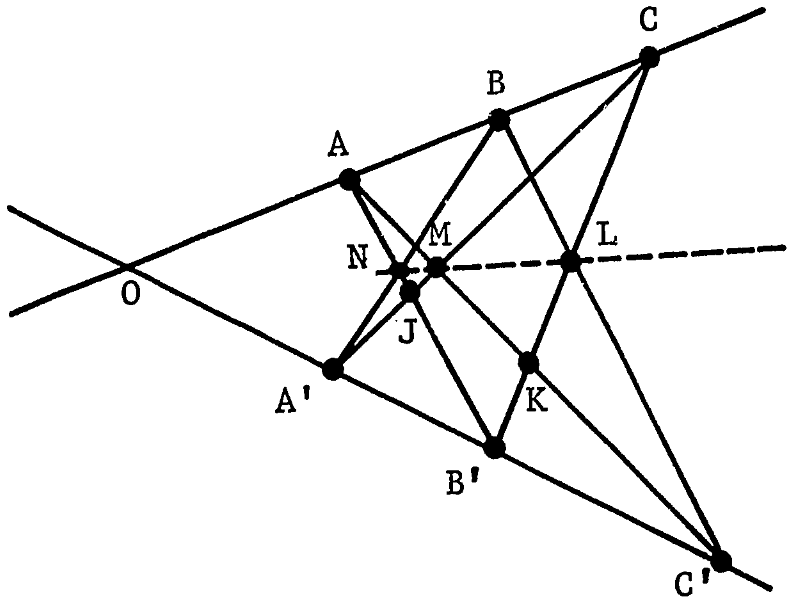




there is a projectivity that interchanges them in pairs; that is, if A, B, C, D are any four collinear points, ABCD $\barwedge$ DCBA. (In this notation, the transpositions are (AD) and (BC).) An easily remembered proof is obtained by having the adjacent symmetric figure and then using perspectivities from the three outer vertices to "rotate" the ranges on the "inner" lines. Thus if we label the extra points as indicated, ABCD $\overset{E}{\doublebarwedge}$ HBIF $\overset{A}{\doublebarwedge}$ ECIG $\overset{F}{\doublebarwedge}$ DCBA .

The final axiom, "Proposition P," supplies the only difficult step in proving the *fundamental theorem*, which tells us that a projectivity is completely determined by its effect on three points of a range (or on three lines of a pencil). From this it follows that *a projectivity relating ranges on two distinct lines must be a perspectivity if it leaves the point of intersection invariant*. Since we shall make use of this result later, let us refer to it as *the lemma*.

Two of the most famous theorems of projective geometry are those of Pappus and Desargues. Both are sometimes taken as axioms, but not in the same treatment, because the former implies the latter, though the complete deduction is difficult. The deduction of the fundamental theorem from them is difficult too. Accordingly I recommend (in a first course) taking "Proposition P" as an axiom, deducing the lemma, and continuing as follows.



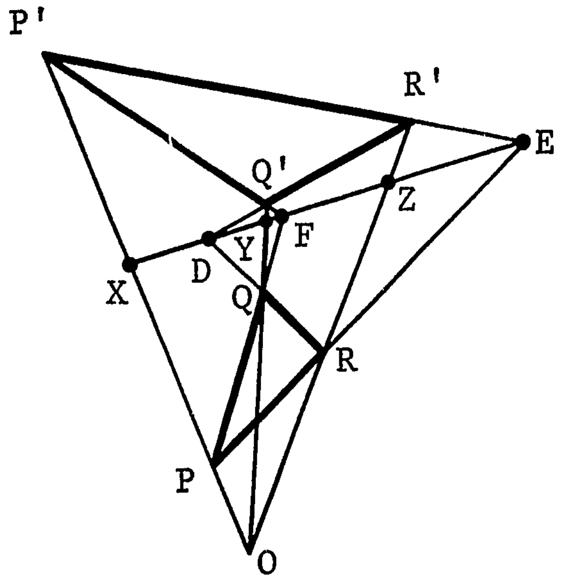
Let A, B, C be three points on one line and A', B', C' three points on another. "Cross join" them, B to C' and C to B', A to C' and C to A', A to B' and B to A', and let L, M, N be the three points of intersection of these three pairs of cross joins. Pappus' theorem asserts that L, M, and N are collinear. The proof of this is easy as soon as we have defined three new points: O, where AB meets A'B', J, where AB' meets A'C, and K, where AC' meets B'C (as in the figure above). Using the lemma, we see that the product of perspectivities

$$ANJB' \overset{A'}{\underset{\wedge}{=}} ABCO \overset{C'}{\underset{\wedge}{=}} KLCB',$$

being a projectivity $ANJB' \underset{\wedge}{-} KLCB'$ that leaves B' invariant, must be a perspectivity. The center of this perspectivity must be at M, the point in which AK and CJ meet; hence L, M, N are collinear.

Until a very few years ago I was not aware that there is a closely analogous proof of the companion theorem of Desargues. Someone pointed out to me that this proof appears in very small type at the end of one of the chapters of van der Waerden's Einführung in die algebraische Geometrie.

The theorem asserts that if PQR and P'Q'R' are two triangles perspective from a point O then the intersections, D,E,F, of corresponding sides of the two triangles are collinear. If (as in the figure below) the point O is not on the line DF, we introduce three new points, X,Y,Z, the inter-.



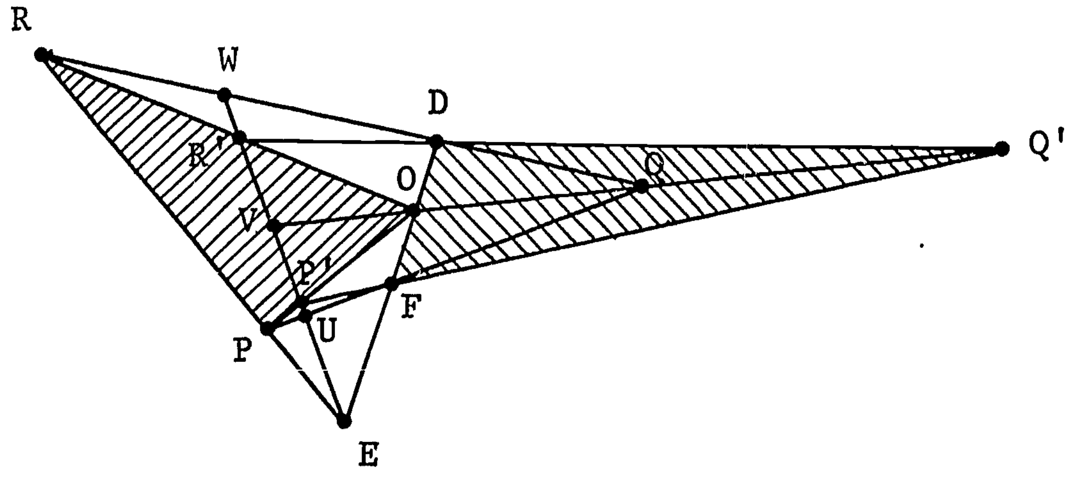
sections of the line DF with the three lines PP', QQ', RR' (which all pass through O).

The product of perspectivities

$$OPXP' \overset{F}{\doublebarwedge} OQYQ' \overset{D}{\doublebarwedge} ORZR'$$

is a projectivity $OPXP' \barwedge ORZR'$ leaving O invariant; hence, by the lemma, it must be a perspectivity. Since PR and P'R' meet in E, E must be the center of the perspectivity, and therefore E is on the line XZ which is DF.

This proof evidently breaks down if O lies on DF (as in the figure below). Let us modify it so as to avoid such complications. We notice that



Desargues' theorem is not essentially altered if we make any permutation of the letters P, Q, R, along with the corresponding permutations of P', Q', R' and of D, E, F; nor is it altered if we simultaneously interchange P and P', Q and Q', R and R'. Since the triangles PQR and P'Q'R' cannot be so mixed up that each is inscribed in the other (P' on QR, Q' on RP, R' on PQ, P on Q'R', Q on R'P', and R on P'Q'), we may assume, without loss of generality, these two perspective triangles to be so named that Q does not lie on R'P'. We can now adapt van der Waerden's proof by applying it to the two triangles POR and FQ'D, which are perspective from Q. The details are as follows. We introduce three new points U, V, W, the intersections of R'P' with the three lines PF, OQ', RD (which all pass through Q).

The product of perspectivities

$$QPUF \overset{P'}{\doublebarwedge} QOVQ' \overset{R'}{\doublebarwedge} QRWD$$

is a projectivity QPUF $\barwedge$ QRWD leaving Q invariant; hence it must be a perspectivity. Since PR and UW meet in E, E must be the center of the perspectivity, and therefore E is on the line FD. This completes the proof.

Now we go on to discuss collineations and the two-dimensional analog of the fundamental theorem which asserts (as we remarked before) that a projectivity is determined by its effect on three collinear points. A collineation is a one-to-one transformation of points to points and lines to lines that preserves incidence. We single out the particular collineations in which the correspondence induced on one range is a projectivity. Of course, any collineation relates a range of points to a range of points, but in the general projective plane you cannot be sure that this correspondence is a projectivity. If we assume the correspondence is projective for <u>one</u> range then we can show it is projective for all ranges. The two-dimensional analog of the fundamental theorem is that a projective collineation is completely determined by its effect on a complete quadrangle. (For details of the proofs of the existence and uniqueness of a projective collineation transforming one complete quadrangle to another, see pp. 50-52 of my book *Projective Geometry*.)

Projective geometry, in contrast to all the other geometries that we are discussing, admits a *principle of duality*: the possibility of interchanging points and lines consistently in any valid proposition to yield another valid proposition, the *dual*. This is justified by the fact that the axioms imply their duals.

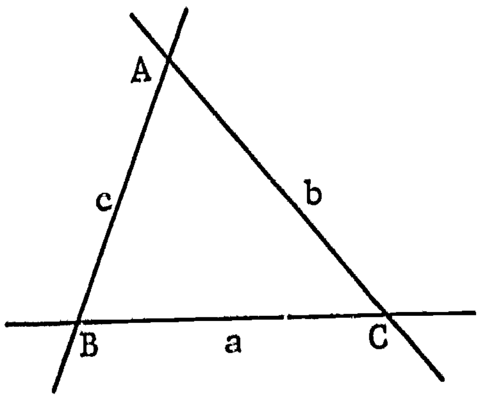
The dual figure to a complete quadrangle (a set of four points, no three collinear) is a complete quadrilateral (a set of four lines, no three concurrent). Just as Proposition P implies that there is exactly one projective collineation transforming a given complete quadrangle into another, so also does it imply that there is exactly one projective correlation transforming a given quadrangle into a given quadrilateral (see pp. 57-59 of *Projective Geometry*). A correlation is a one-to-one transformation of points to lines and lines to points that preserves incidence, and a projective correlation is a correlation in which the induced correspondence between one range and one pencil is projective.

Before leaving the subject of collineations I should perhaps have mentioned the important particular case of perspective collineations. A *perspective collineation* is one in which there is a line (the *axis*) all of whose points are left invariant. Since the identity is a trivial projectivity, a perspective collineation is certainly a special case of a projective collineation. A perspective collineation also has the dual property that all the lines through one point are left invariant, that is, a perspective collineation always has a *center* as well as an axis. If the center is on the axis the perspective collineation is called an *elation*; if the center is off the axis then it is called a *homology*. (See the discussion for the history of these two terms.)

A homology can have any period (for suitable fields). There is a nice theorem which says that a projective collineation of period two must be a *harmonic homology*, taking each point to its harmonic conjugate with respect to the center and axis. (Since we've assumed the diagonal points of a complete quadrangle are not collinear, harmonic conjugates are well defined.)

After considering projective *collineations* of period two it's natural to ask about projective *correlations* of period two. A projective correlation of period two is called a *polarity*. This is a transformation of points to lines and lines to points with the special feature that, when the capital of any letter is used for a point, the corresponding lower case letter can be used for the corresponding line without any risk of confusion, as the correspondence goes both ways round: P is transformed into p, and p is transformed into P. We speak of p as the *polar* of P, and P as the *pole* of p. Any point on p is said to be *conjugate* to P; any line through P is said to be *conjugate* to p.

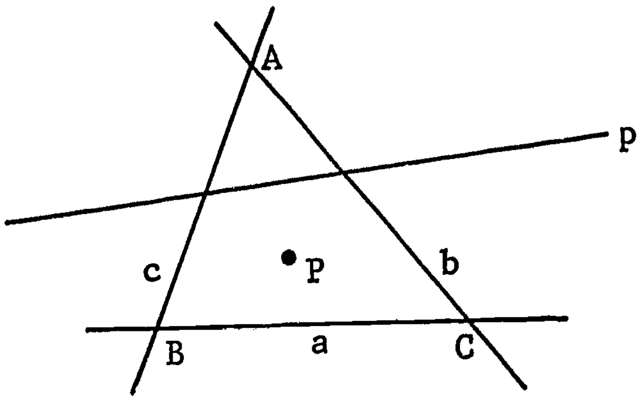
For every polarity there is at least one point A that is not *self-conjugate*, i.e., that is not on its own polar a. Given such a point, we can choose a point B on a that is not self-conjugate and let C be the point of intersection of the polars of A and B. We have now constructed a *self-polar triangle* ABC, in which each vertex is the pole of the opposite side, and therefore any two vertices



(or sides) are conjugate.

There is a partial converse to the statement that every polarity admits a self-polar triangle: if a projective correlation transforms the vertices of one triangle into the respectively opposite sides, then it must be a polarity.

Since there is a unique projective correlation relating any given quadrangle to any given quadrilateral, it follows that a polarity is determined by one self-polar triangle ABC, an extra point P, not on any side of ABC, and its polar p. In other words, given any triangle ABC, a point P not on any side, and a line p not through any vertex, the unique correlation transforming the four points ABCP into the four lines abcp is a polarity.

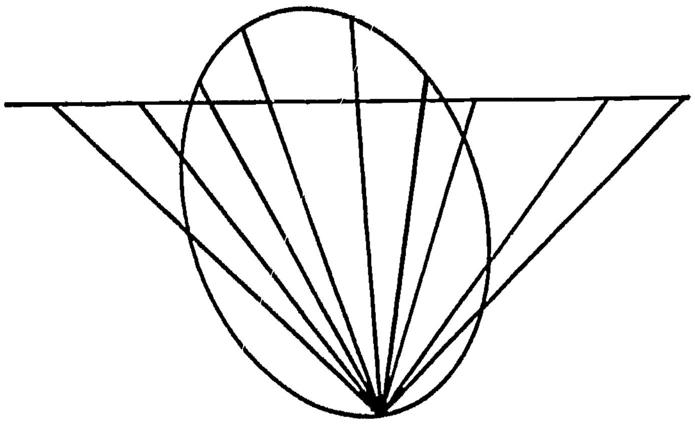


We shall find it convenient to use the symbol (ABC)(Pp) for this polarity.

There is a nice theorem, due to Veblen, which says that every projective collineation can be expressed as the product of two polarities. It is clear that the group of projective collineations is generated by pairs of polarities, but it is very interesting that each projective collineation is the product of exactly two polarities. This theorem is in the spirit of the present course, since the group associated with the projective plane (in the sense of Klein's Erlangen program) is the group of collineations and correlations: every theorem of the projective plane remains true if you apply a collineation or correlation to all the points and lines involved in it. It is natural to consider the normal subgroup of collineations and then the normal subgroup of that consisting of the projective collineations. (The lecturer presented a proof that every projective collineation is the product of two polarities. We omit it here since it is on pp. 68-70 of Projective Geometry.)

A conic can be defined as the set of self-conjugate points and self-conjugate lines in a polarity (provided this set is not empty). In other words, a conic is the set of points that lie on their polars and of lines that pass through their poles. This self-dual definition, due to von Staudt, seems to me much more convenient than Poncelet's. If at least one self-conjugate point exists then there are many, but each line contains at most two of them (pp. 61 and 72 in Projective Geometry).

There is a kind of stereographic projection from the points on a conic to the points on a line. Take one point on the conic, the pencil of lines through it, and a section of that pencil by an arbitrary line. This establishes a correspondence between the points on the conic and the points on a line, done by means of a generalized "perspectivity," as it were.



Using this correspondence you can easily see that the idea of projectivities on a line can be transferred onto the conic so that you can speak of a projectivity relating the points on a line to the points on a conic, the points on a conic to the points on another conic, or the points on a conic to the points on the same conic. The discussion of projectivites is in some ways made simpler if you use a conic instead of a line to form the range of points. For instance, projectivities of period two, called involutions, are particularly important but rather troublesome to deal with on a line. They are much easier on a conic because you can show that the lines joining corresponding points of an involution on a conic are concurrent. Thus an involution on a conic is cut out by the pencil of lines through a fixed point (not on the conic). In fact, the involution is induced on the conic by the harmonic homology whose center and axis are the fixed point and its polar.

Somewhat analogously, pairs of numbers may be said to belong to an "involution" if they have a constant sum or a constant product. After assigning the symbols $0, 1, \infty$ to three points on a given conic, we can define a formal sum and product for any two points on the conic as follows. Two pairs of points on the conic are said to have the same sum if their joining lines are concurrent with the tangent at the point ∞. Two pairs of points on the conic are said to have the same product if their joining lines are concurrent with the line joining ∞ and 0. Simple geometric considerations enable us to verify that the points on the conic (other than ∞) satisfy all the axioms for a (commutative) field. For instance, the associative laws come from Pascal's theorem concerning a hexagon inscribed in the conic. For the details, see Vol. 1 of Projective Geometry by Veblen and Young, or Chapter 11 of my Cambridge University Press book The Real Projective Plane. (There is no need for the field to be real. It may even be finite, as von Staudt noticed in 1857.)

After you have the field of points on the conic you can transfer them back to a line. I believe that this method is much easier than introducing the field of points directly on the line. Of course this approach uses Proposition P and therefore cannot be applied to the much more difficult problem of introducing coordinates into a projective plane in which Desargues' theorem is valid but Pappus' theorem is not. I do not believe that this more difficult task should be attempted in an elementary course.

Discussion.

Killgrove, Coxeter, and Busemann discussed the special case of collineations of the *real* projective plane. Since they preserve order on a line, such collineations are necessarily projective. Busemann pointed out that Klein never forgave himself for an early erroneous view in which he thought an assumption of continuity was essential for projective collineations.

Prenowitz brought up von Staudt's definition of a projectivity as a transformation preserving harmonic sets, and Coxeter remarked that although this definition and Poncelet's (given in the lecture) are equivalent in the real projective plane, they are not equivalent in the general projective plane.

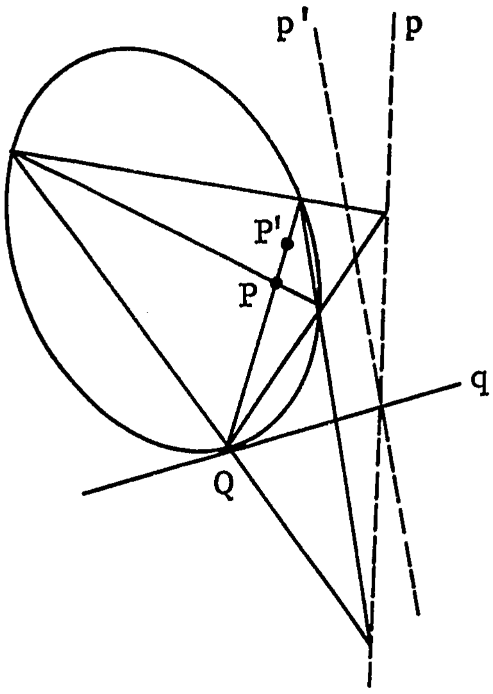
In reply to a question by Gleason, Coxeter said he believed that Poncelet used the word *homologies* for all perspective collineations, including elations. The Norwegian geometer Sophus Lie seems to have been the first to make this use of the word *elation*. Coxeter also pointed out that we use *dilatation* and not *dilation* for the same reason that we use *rotation* or *notation* and not *rotion* or *notion*. In all three cases the shorter version is poor Latin.

Blattner pointed out that one needs to be careful in projective planes with very few points on a line, for difficulties may arise in theorems or proofs when incidences are forced because of an insufficient number of points. Coxeter said that the projective plane over the field with five elements, i.e., with six points on each line, is particularly nice for illustrative purposes since it is just large enough to avoid any such unpleasant collapse. For instance, Pascal's theorem (about a hexagon whose vertices lie on a conic) is still valid in this plane; the hexagon is the conic, but otherwise everything is as usual.

Gleason discussed an interesting theorem related to the technique Coxeter advocated for introducing the field of scalars. In a general incidence plane, i.e., a plane satisfying the first three axioms of this lecture but perhaps not the axioms about the diagonal points of a quadrangle nor Proposition P, an oval is defined as a set of points such that, for each point on the oval, every line through that point, except one called the tangent, meets the oval once more. If one oval satisfies Pascal's theorem then every oval in the plane has this property.

Yale asked about the subgroup of the group of projective collineations generated by harmonic homologies. Coxeter replied that he believed it to be the full group of projective collineations and that probably only three or four harmonic homologies were needed for each projective collineation. He also noted that since an affine reflection is a harmonic homology whose center is on the line at infinity, the harmonic homologies with centers on a given line generate the projective version of the group of equiaffinities.

Klamkin stated that Desargues' theorem yields a nice construction for the straight line through two points when we are forced to use a "too short" straightedge and also a construction for a line through the inaccessible point of intersection of two "nearly parallel" lines. He remarked further that the theory of poles and polars yields nice constructions for tangents to conics using the straightedge alone. Coxeter elaborated on the last point by first giving the construction for the polar p of an exterior or



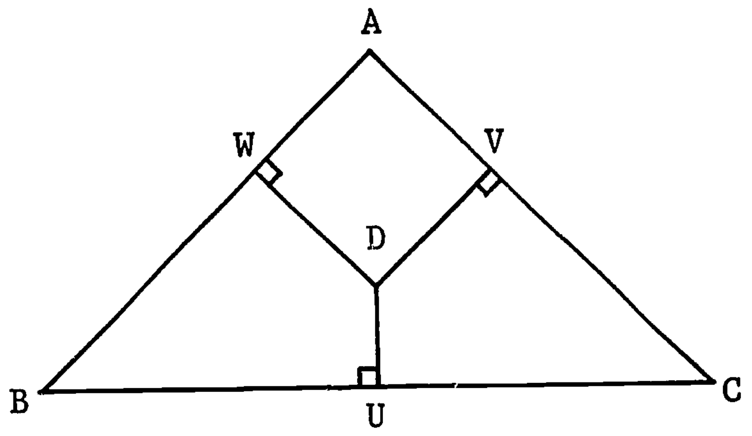
interior point P, using two secants through P. Then he remarked that the tangent at a given point Q on the conic joins Q to the point of intersection of the polars of any two points on a secant through Q. Klamkin pointed out that the tangents through an exterior point P join P to the points of intersection of the conic with the polar, p.

In reply to a question by Stratopoulos, Coxeter explained that this treatment of conics in terms of polarities could be extended to a treatment of quadrics by means of polarities in a projective space. A polarity in space is an involutory projective correlation transforming points into planes (polar planes), lines into lines (polar lines), and planes into points (poles). For the set of points that lie on their polar planes, there are now not only two but three possibilities. The set may be empty, or it may consist of the points on a quadric surface, or it may consist of all the points in the space. In the second case the quadric may be ruled (possessing generators, which are self-polar lines) or non-ruled. In the third case (a null polarity), the self-polar lines form a linear complex.

Finally, Stratopoulos commented on the spatial analogs of involutions on a conic.

Lecture IV. Inversive Geometry: The group of homographies and antihomographies.

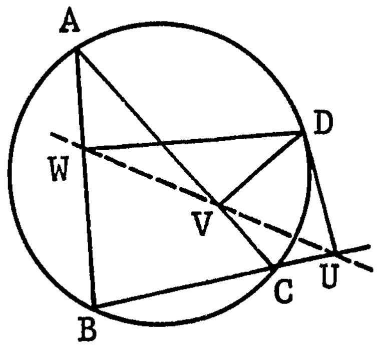
I shall begin this lecture on inversive geometry by using some results from Euclidean geometry. We recall the extended sine rule of trigonometry which tells us that BC/sin A is twice the circumradius, R, of any triangle ABC. From an arbitrary point D in the plane, drop perpendiculars to the three sides of ABC, and let U, V, W be the feet of these perpendiculars. Since AD is the circumdiameter of the triangle AWV, VW/sin A = AD.



Combining this with BC/sin A = 2R, we find BC × AD = 2R × VW. Continuing around the triangle we obtain in a similar fashion CA ·× BD = 2R × WU and AB × CD = 2R × UV. Using the triangle inequality UV + VW ≧ UW, we combine these to get Ptolemy's inequality

(1) $AB \times CD + BC \times AD \geqq CA \times BD.$

Equality is attained when the "pedal triangle" UVW collapses to form the Simson line (which I understand is not due to Simson at all but to Wallace; I think it was Cauchy who gave it this name, saying he thought it was the sort of thing that Simson was quite likely to have done). Assume that D is on the circumcircle of triangle ABC, say on the arc AC, and drop perpendiculars as before; then the points UVW are collinear, i.e., they lie on the Simson line of D. Combining these ideas we see that Ptolemy's inequality becomes an equality if and only if D is on the circle through ABC and on the arc away from B, i.e., if and only if A and C separate B and D. We use



Vailati's notation $AC \parallel BD$. There is no difficulty in proving that Ptolemy's inequality still holds in the degenerate case when ABC are collinear (so that the "circle" ABC has infinite radius).

If we divide both sides of equation (1) by $AC \times BD$, we can express it as

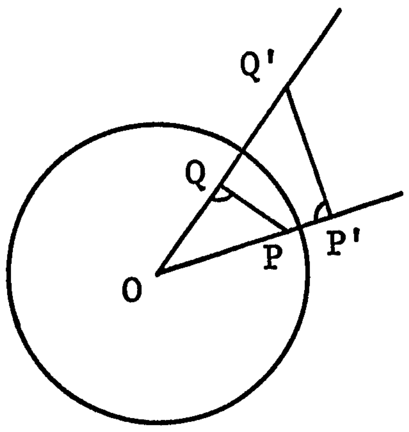
$$\frac{AB \times DC}{AC \times DB} + \frac{AD \times BC}{AC \times BD} \geqq 1,$$

or, in the notation of cross ratios (the four points still being quite arbitrary, and the distances positive),

$$\{AD, BC\} + \{AB, DC\} \geqq 1.$$

Since equality holds if and only if $AC \parallel BD$, we now have a definition of separation in terms of cross ratio.

This is important for our purposes because cross ratio is an inversive invariant. To see this, let us define inversion in a circle of radius k, centered at O, in the usual way, i.e., P and P' are inverses (of each other) if and only if $OP \times OP' = k^2$. Given two such points, we have



$$OP \times OP' = k^2 = OQ \times OQ',$$

so triangles OPQ and $OQ'P'$ are similar, and

$$\frac{P'Q'}{PQ} = \frac{OP'}{OQ} = \frac{OP \times OP'}{OP \times OQ} = \frac{k^2}{OP \times OQ}.$$

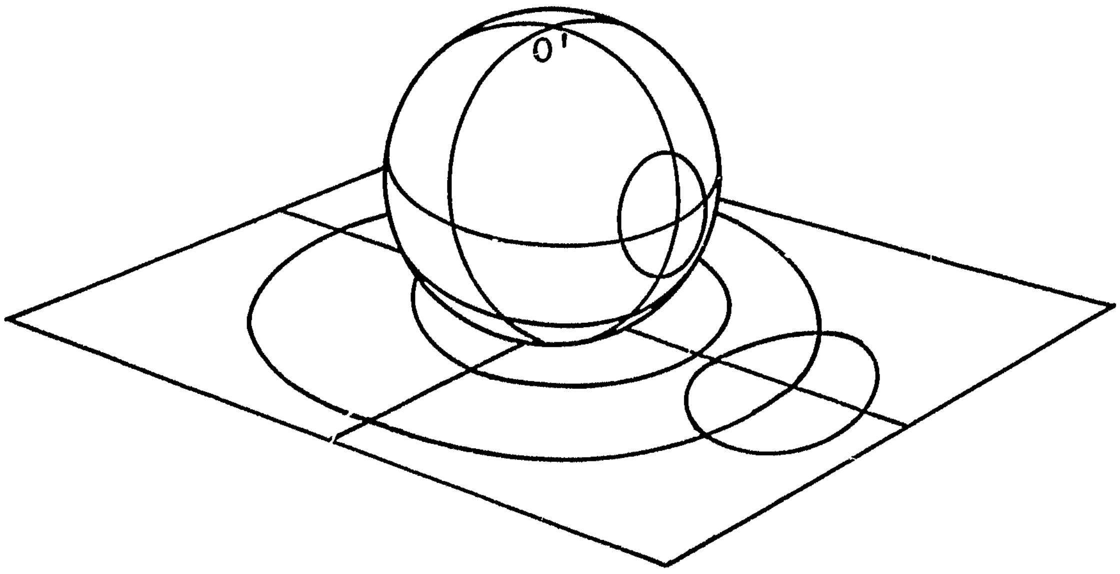
Using this you can show that if the four points PQRS are transformed by inversion into P'Q'R'S' then the cross ratios {PQ, RS} and {P'Q', R'S'} are equal. Since cross ratio is an inversive invariant and separation (on a line or circle) can be characterized by cross ratios, separation is an inversive invariant.

If we consider three distinct points A, B, C, the circle or line determined by them consists of A, B, C, all points X such that $BC \parallel AX$, all points Y such that $CA \parallel BY$, and all points Z such that $AB \parallel CZ$. But we have just seen that separation is an inversive invariant; hence circles (in the extended sense of circles or lines) are inversive invariants. In other words, circles invert into circles (or possibly lines). This seems to me to be an efficient method of getting this basic result.

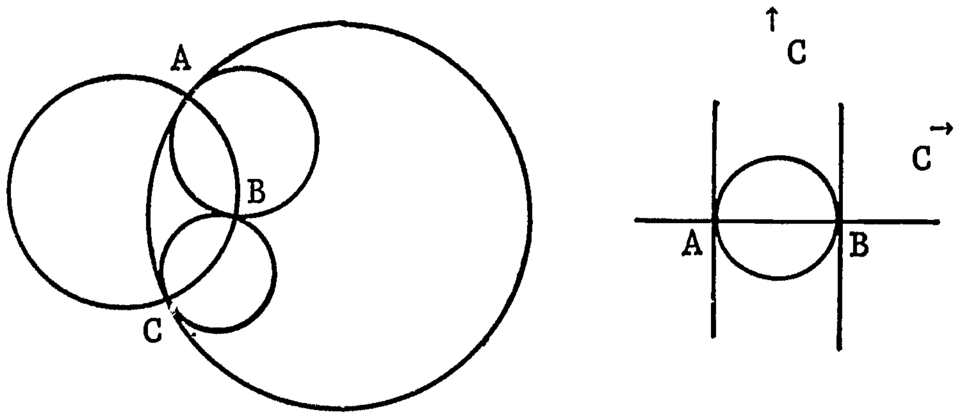
Inversion is almost a schlicht (one to one and onto) transformation of the Euclidean plane. In other words, since points on the circle of inversion are their own inverses, almost every point has an inverse: the only exception is the center of the circle, which has no inverse. In order to make inversion a schlicht transformation we invent a new kind of plane, called the inversive plane, by postulating an ideal point, just one ideal point, which is called the point at infinity. This extra point is the inverse of any point O under inversion in any circle centered at O. Now in talking about three points determining a circle we need not make an exception when the three points are collinear; for we can view a line as a special case of a circle, namely a circle through the point at infinity. This extension of the Euclidean plane can be further justified by observing that the inverse (with respect to a circle centered at O) of a circle through O is a straight line, and the inverses of points close to O on the circle are far away on the line, tending towards infinity. Two parallel lines should be thought of as two circles tangent at the point at infinity, for they are inverses of circles tangent at the center of the circle of inversion.

Inversive geometry, then, is the geometry characterized by the group generated by inversions. Just as, in Euclidean geometry, the group of similarities has a normal subgroup of displacements leaving distance invariant, and in affine geometry the group of affinities has a normal subgroup of equiaffinites leaving area invariant, so in inversive geometry the group generated by inversions has a normal subgroup consisting of homographies (or "Möbius transformations") leaving (directed) angles invariant. These are products of even numbers of inversions. The products of odd numbers of inversions are called antihomographies.

Another way of viewing inversive geometry is as the geometry of circles on a sphere in which great circles play no special role. If O and O' are antipodal points on the sphere and you project the sphere stereographically from O' onto the tangent plane at O, it becomes very clear that lines should be viewed as circles through the point at infinity, O'.



Orthogonality of circles is another property that is an inversive invariant. This invariance can be deduced as a special case of the fact that angles are preserved (or, more precisely, reversed in sign) by an inversion; but perhaps a simpler way is the following. If you have two distinct circles they may intersect (have two common points), be tangent (have one common point), or be non-intersecting (have no common points). This classification of pairs of circles is clearly an inversive invariant. Now if you are given three circles

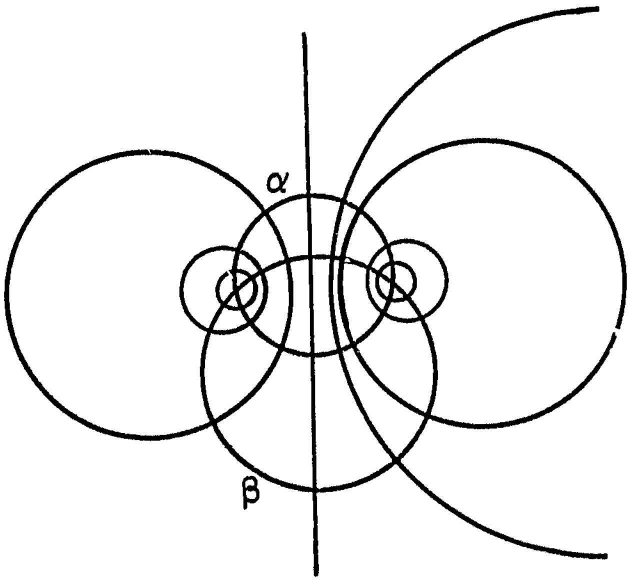


tangent in pairs at three distinct points A, B, C, then the circle determined by ABC is orthogonal to each of the others. Conversely, any two orthogonal circles can be exhibited as belonging to such an arrangement of four, e.g., by inversion in a circle with center C.

Thus orthogonality can be defined in terms of incidence and is therefore an inversive invariant, i.e., a property of inversive geometry.

Having slightly generalized the idea of a circle so that a straight line is a special case of a circle, we can give "inversive" definitions of separation and inversion. Four points A, B, C, D satisfy the relation $AC \| BD$ if and only if every circle through A and C intersects every circle through B and D. The inversive definition of inversion itself, with no appeal to distance, is somewhat similar. Points P and P' are inverses with respect to the circle ω if they are the two points of intersection of two circles orthogonal to ω. Thus to find the inverse of a point P, draw two circles through P orthogonal to ω, and where they meet again is the inverse point.

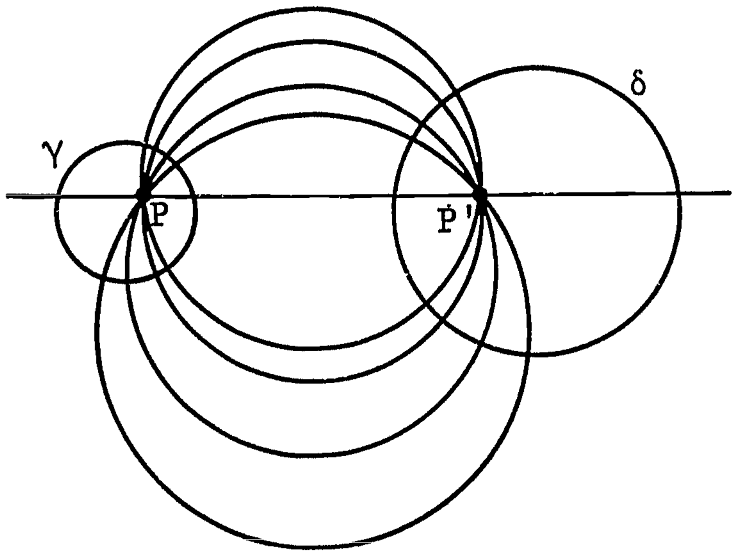
Orthogonality leads to the subject of coaxal circles. Given two circles, say α and β, if they intersect at P and P' then the circles orthogonal to both α and β form a non-intersecting pencil of coaxal circles. This



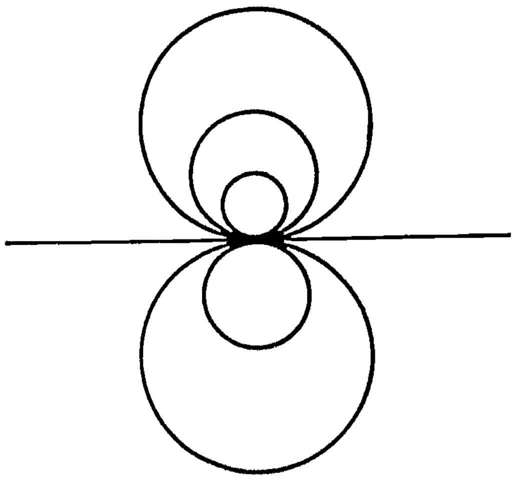
family of "circles" contains one straight line: the perpendicular bisector of PP'. It also contains two "circles of zero radius": the limiting points P and P'.

(The term pencil is consistently used for a family that contains just one member through each point of "general" position.)

If you then choose two circles, γ and δ, in the pencil and go on to consider the family of circles orthogonal to both γ and δ, this new family consists of all circles through the points P and P' (which are the limiting points of the first pencil). This intersecting pencil of coaxal circles



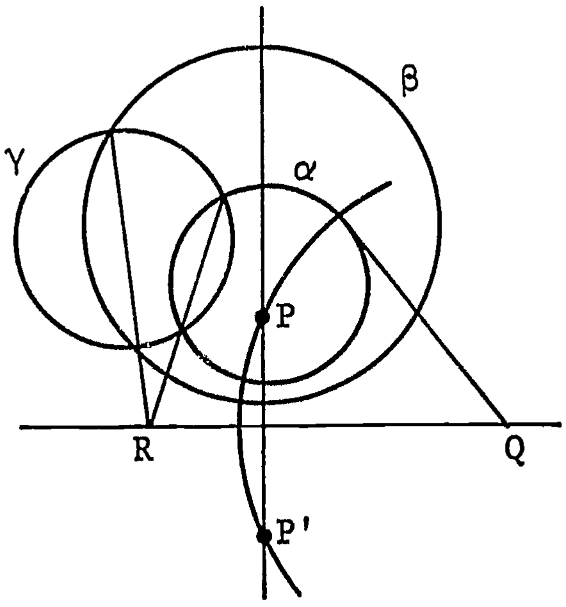
contains the line PP' and the two original circles α and β. Thus any two intersecting circles or any two non-intersecting circles determine a pencil of circles, to which they belong, and a second pencil consisting of all circles orthogonal to both. If one pencil is intersecting, the other is non-intersecting, and vice versa. In the intermediate case, where α and β are tangent, both pencils consist of tangent circles like the two families of circles touching the coordinate axes at the origin. Each family is called a <u>tangent pencil of coaxal circles</u>. If we invert such a tangent pencil in a circle centered at the point of tangency, the circles transform into lines. parallel to the one line in the pencil. Similarly, if we invert an intersecting pencil of coaxal circles in a circle centered at P then the circles transform into lines through the image of P', and if we invert a non-intersecting pencil in a circle centered at one of the limiting points the circles transform into concentric circles whose common center is the image of the other limiting point. With this we see that a pencil of parallel lines is a special case of a tangent pencil of coaxal



circles, a pencil of intersecting lines is a special case of an intersecting pencil, and a pencil of concentric circles is a special case of a non-intersecting pencil.

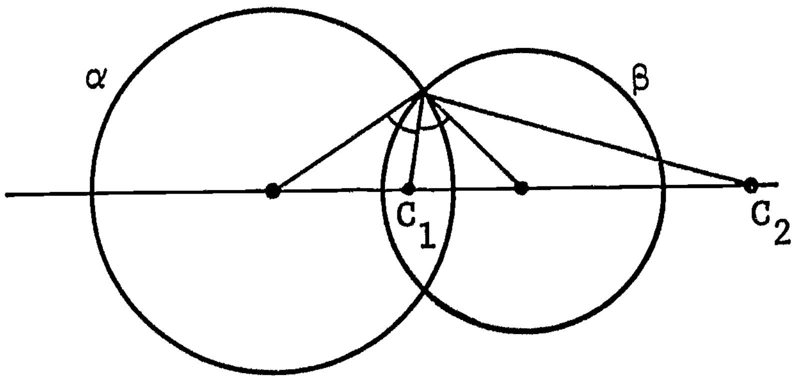
With the exception of these three special families, each pencil of coaxal circles contains one straight line called the <u>radical axis</u> of that pencil. This line, shown in each of the three diagrams above, is the locus of centers of the circles in the orthogonal pencil.

There is a simple construction of the radical axis of the pencil $\alpha\beta$ (that is, of the pencil containing the circles α and β). Draw any circle γ intersecting both α and β and let R be the point of intersection of the common chords (which are the radical axes of $\alpha\gamma$ and $\beta\gamma$). The line through R perpendicular to the line of centers of α and β is the radical axis of $\alpha\beta$. If α and β are non-intersecting let Q be any point on their radical axis and draw the tangent from Q to either α or β. The circle with center at Q and this tangent as radius is a circle belonging to the orthogonal intersecting pencil, so it cuts the line of centers of α and β in the limiting points of the pencil $\alpha\beta$. In this way we have a simple construction for the limiting points of a non-intersecting pencil of coaxal circles.

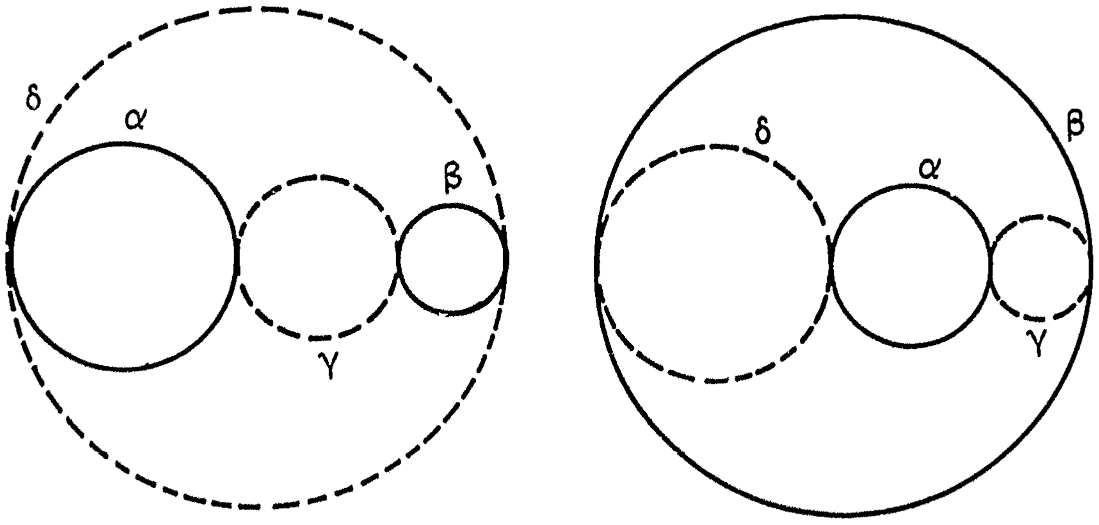


Another important idea is that of a mid-circle between two circles. I must confess that I invented this name because I am not very fond of the classical name, "circle of antisimilitude," which I believe scares people off from this simple idea. A mid-circle of two circles is simply a circle that has the effect of inverting one into the other.

Since reflection in a line is a special case of inversion, the mid-circles of a pair of intersecting lines are their angle bisectors. Any two intersecting circles may be inverted into a pair of intersecting lines, therefore any two intersecting circles have two mid-circles, orthogonal to each other. Given two intersecting circles α and β, their mid-circles, which of course are in the pencil $\alpha\beta$, may be constructed in the following manner. Draw radii for α and β to one of the points of intersection and let C_1 and C_2 be the points in which the angle bisectors, internal and external, of these radii meet the line of centers. The two circles centered at C_1 and C_2 and through the points of intersection of α and β are the two mid-circles.

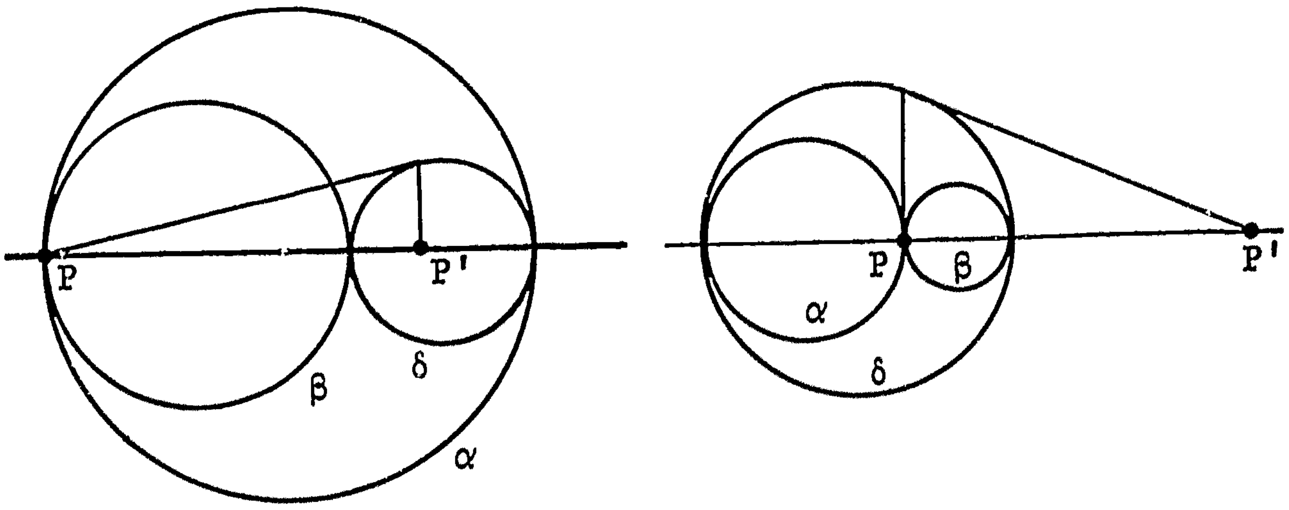


The case of intersecting circles is the only case in which there is more than one mid-circle. Two non-intersecting circles can be inverted into



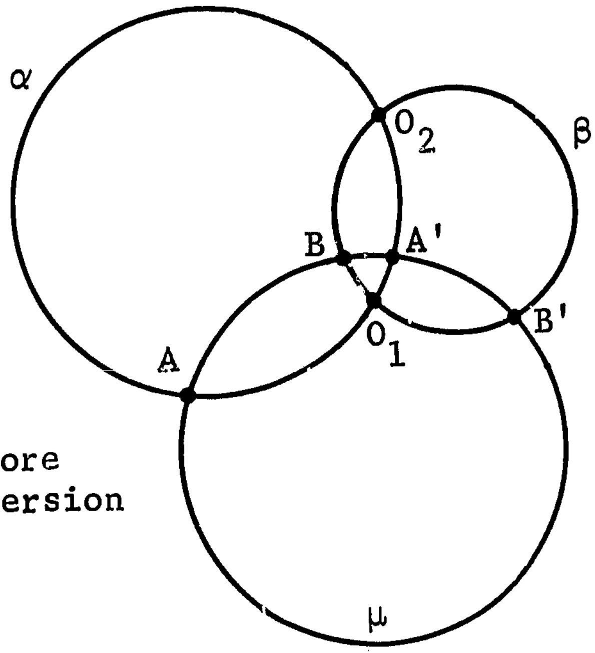
concentric circles, and there is obviously a unique mid-circle for two concentic circles. It is the circle concentric with them whose radius is the geometric mean of the radii of the two given circles. The construction of the mid-circle for two non-intersecting circles α and β is rather amusing. Let γ and δ be the two circles tangent to both α and β at points on the line of centers. Find the limiting points, P and P', for the pencil $\gamma\delta$. The circle with diameter PP' is the desired mid-circle of α and β.

Finally, suppose you have two tangent circles. You can invert them into parallel lines, and between two parallel lines there is a unique mid-line which reflects one into the other. Inverting back again transforms this mid-line to the mid-circle. The construction of the mid-circle for a pair of non-intersecting circles simplifies in the case of tangent circles. One of the two circles, say γ, degenerates to the first limiting point P, at the point of tangency of α and β. The other limiting point, P', is simply the inverse of P in δ. The mid-circle is the circle with diameter PP'. Incidentally, if α and β have their centers on the same side of their common tangent, the radius of their mid-circle is the harmonic mean of the radii of α and β.

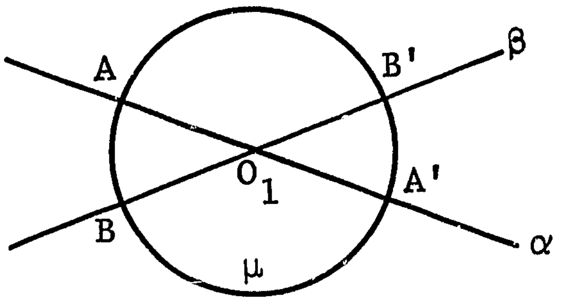


Now let us consider the rather nice problem to invert any four distinct points A, B, A', B' into the vertices of a parallelogram in such a way that A and A' are opposite vertices. We shall first consider this problem by itself and then see why it is significant. We distinguish three cases, beginning with two in which the points are in special position.

If AA' || BB', every circle through A and A' intersects every circle through B and B', and consequently the four points all lie on one circle, say μ. Consider two circles: α through A and A', and β through B and B', both orthogonal to μ. Let O_1 and O_2 be the two points of intersection of α and β. Inverting in a circle centered either at O_1 or O_2, say O_2, sends α and β to straight lines intersecting in the inverse of O_1.



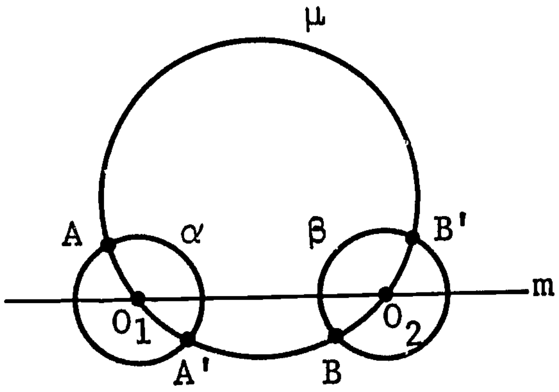
Before inversion



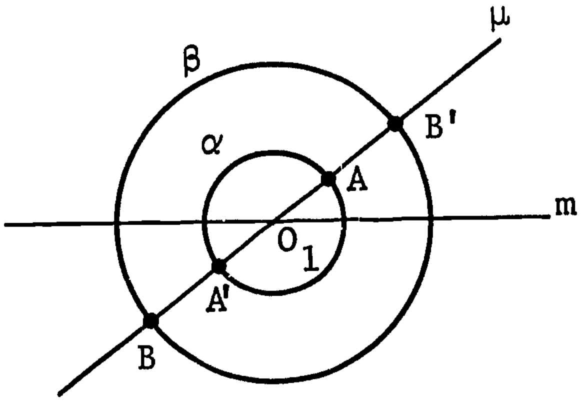
After inversion in O_2

Since μ is orthogonal to α and β, its inverse is orthogonal to the inverses of α and β and is therefore centered at their intersection. Thus the inverses of ABA'B' form a rectangle, which is, of course, a special kind of parallelogram.

In the other case involving special position, the points AA'BB' still lie on a circle μ but A and A' do not separate B and B'. Construct α and β as before, and let m be their line of centers. Invert in a circle centered at either of the two points, O_1 and O_2, in which m cuts μ, say O_2. Since m and μ are inverted into lines through the inverse of O_1,



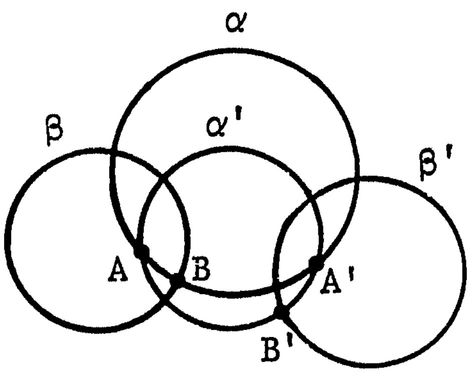
Before inversion



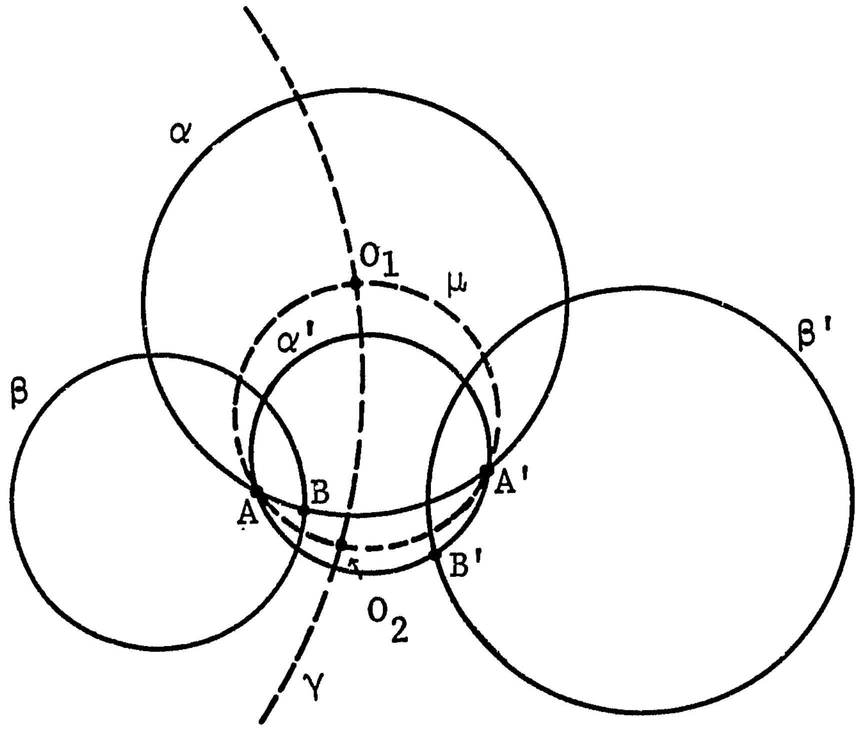
After inversion in O_2

the circles α and β, orthogonal to them, are inverted into concentric circles having this point as their common center. The four inverses of BA'AB' are therefore collinear with AB' and BA' congruent, so we have a degenerate parallelogram.

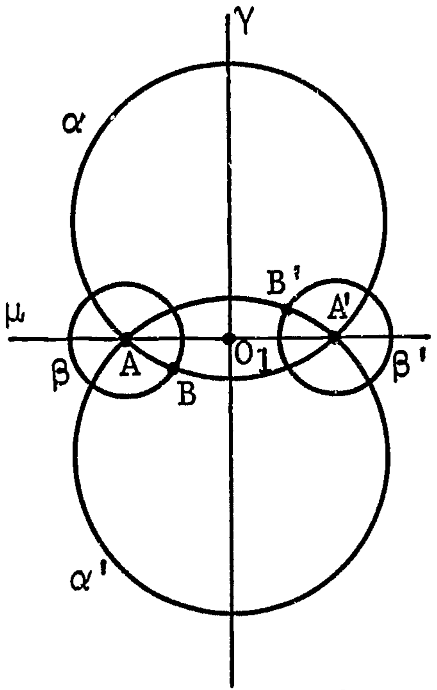
Finally we come to the most interesting case, the case in which A, B, A' B' are not concylic. Let α be the circle determined by ABA', α' be the circle determined by AA'B', β be the circle through B orthogonal to $\alpha\alpha'$, and β' be the circle through B' orthogonal to $\alpha\alpha'$.



The mid-circles of α and α' are in the pencil $\alpha\alpha'$; let μ be the mid-circle that separates B from B' (so that B is inside and B' outside, or vice versa). Let ν be the mid-circle of β and β'; then ν, being in the pencil $\beta\beta'$, is orthogonal to α, α', and μ. Let O_1 and O_2 be the points of intersection of μ and ν. When we invert in a circle centered at O_2, the inverses of μ and ν are perpendicular lines and the product of



Before inversion



After inversion in O_2

of the reflections in these two lines is a half-turn interchanging the inverses of A and A' as well as the inverses of B and B'. Thus the four inverse points are the vertices of a parallelogram.

The product of inversions in two orthogonal circles is called a <u>Möbius involution</u>. A half-turn is a special case, and by inverting in one of the two points of intersection of the two orthogonal circles we can transform any Möbius involution into a half-turn. The significance of the result we have just proved is that, given any four distinct points AA'BB', there is a Möbius involution that interchanges A and A' and also interchanges B and B'. By analogy with an involution in projective geometry, we call this Möbius involution (AA')(BB'). If we allow B and B' to coincide with O_1 (one of the points of intersection of the two orthogonal circles), then another name for this involution is $(AA')(O_1O_1)$ and, going one step further, we can use $(O_1O_1)(O_2O_2)$ as a name for the Möbius involution leaving invariant both O_1 and O_2, that is, for the product of inversions in any two orthogonal circles through these two points.

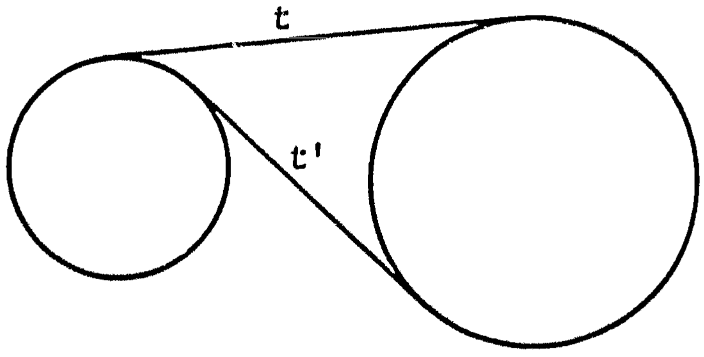
This idea will enable us to show that it is possible to transform any three distinct points ABC into any three distinct points A'B'C' by a homography, i.e., by the product of an even number of inversions. Since a Möbius involution is a product of two inversions, it is enough to transform ABC to A'B'C' by a product of Möbius involutions. Actually *two* Möbius involutions suffice, and only two cases have to be distinguished. If two of A, B, C are invariant, we can assume that A = A' and B = B'; then (AB)(CC)·(AB)(CC') is the desired product of two Möbius involutions. (It reduces to the identity if C = C'.) If, on the other hand, at least two of the three points ABC are non-invariant, then we can assume A ≠ A' and B ≠ B', and the desired product is (AB')(A'B)·(A'B')(C'D) with D the image of C' under (AB')(A'B). I found this simple proof in an old book by J. L. Coolidge, *A Treatise on the Circle and the Sphere*, published in 1916.

Noting that mid-circles bisect the angles between intersecting circles, we may ask if there is something that plays the same role for non-intersecting circles. In other words, is there some inversive invariant for two non-intersecting circles that is "bisected" by their mid-circle? One way of developing such an idea is to imbed the real plane in a complex plane so as to be able to say that the two "non-intersecting" circles intersect in two conjugate complex points. The angle between them at either point of intersection turns out to be purely imaginary, say δi; so we could ignore the i and use this δ as our inversive invariant for the circles. I think, however, that it's nicer to give a definition in terms of real numbers alone. This can be done by inverting the circles into concentric circles. In a pencil of concentric circles, the logarithms of ratios of radii are natural things to consider as "distances" between pairs of circles, since they are additive for the circles in this pencil. Accordingly, we define the <u>inversive distance</u> between any two non-intersecting circles to be the logarithm of the ratio of the radii (larger to smaller) of two concentric circles into which the given circles can be inverted. As this is independent of the particular inversion used, inversive distance is an inversive invariant. (Notice that 1 is the inversive distance betweun concentric circles of radii 1 and e.)

Inversive distance has many nice properties. For example if <u>any</u> three circles are tangent to one another at distinct points, then the two circles tangent to all three are non-intersecting circles whose inversive distance is $\delta = 2 \log (2 + \sqrt{3})$. Similarly there is a nice formula,

$$\sinh \frac{\delta}{2} = \frac{1}{2} \sqrt{\frac{r}{R}}$$

for the inversive distance δ between the incircle and circumcircle of a triangle in terms of the inradius r and circumradius R. Of course, the circumcircle encloses the incircle. If, on the other hand, two non-intersecting circles are so situated that neither of them encloses the other, then the inversive distance δ is given



$$\tanh \frac{\delta}{2} = \frac{t'}{t}$$

where t' is the length of either of the two shorter common tangents and t the length of either of the two longer common tangents. Here is an amusing paradox: although the mid-circle does bisect the inversive distance between any two non-intersecting circles, a circle may bisect the inversive distance and not be the mid-circle! These results are more fully discussed in a recent paper of mine on Inversive distance in the Annali di Matematica (4), 71(1966), and in Chapter 5 of Coxeter and Greitzer Geometry Revisited (New Mathematical Library, No. 19).

As a final remark, three nested non-intersecting circles satisfy a non-triangle inequality: the sum of the distances between the innermost pair and between the outermost pair is never more than the distance between the inner and outer circles. (This will be proved in my Presidental Address to the Canadian Mathematical Congress, August 1967.)

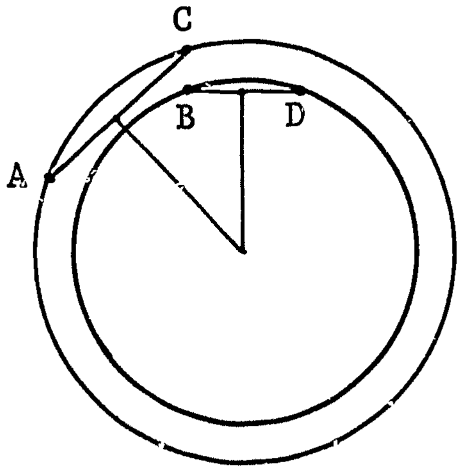
Discussion.

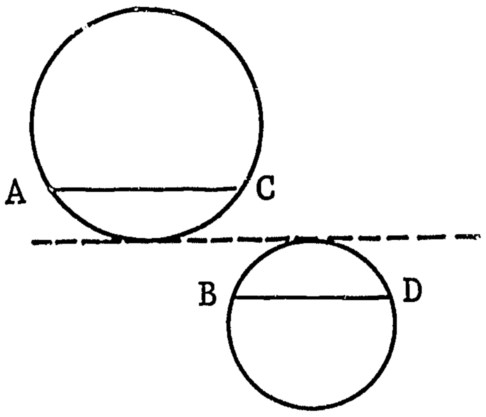
Several ways of viewing the inversive plane were brought out in the discussion. In addition to the viewpoint taken in the lecture (in which the inversive plane was approached as the Euclidean plane extended by the postulation of a single point at infinity) there are the following alternative approaches: (1) axiomatically, (2) as the geometry of circles on a sphere, without assigning any special role to great circles, (3) as the geometry of complex numbers including the "number" ∞, and (4) as the geometry of the complex projective line. Coxeter remarked that the axiomatic approach (1) was first tried by M. Pieri in 1911. The spherical approach (2) was used by H. Liebmann in the

1905 edition of his Nichteuklidische Geometrie, where he remarked that the inversions in the ∞^3 circles of the inversive plane behave like the reflections in the ∞^3 planes of a hyperbolic 3-space. The algebraic approach (3) enables us to regard a pair of points as a pair of complex numbers, and thus as a binary quadratic form. The pair of invariant points O_1O_2 of the Möbius involution (AA')(BB') thus appears as the Jacobian of two such quadratic forms. Finally, the projective approach (4) enables us to identify the theory of homographies in the inversive plane with the theory of one-dimensional projectivities.

Gleason asked if inversive geometry is essentially characterized by some of the basic properties of the inversive group in the sense that given a set and a group acting on the set, such that the group is exactly three times transitive, it should, using relatively weak incidence axioms, be possible to prove that the resulting geometry is the inversive plane. Coxeter replied that some such ideas have been published, e.g., Kerékjártó proved that the sphere is the only compact two-manifold that admits a triply transitive Lie group.

Gleason gave a solution to a problem, used on the Putnam exam a few years ago, that Coxeter had posed earlier in the week. If four points ABCD are not concyclic show that there are two circles, one passing through A and C and the other through B and D, that do not intersect. The idea of Gleason's solution was to use stereographic projection to get the points on a sphere and then look for two parallel planes, one through A and C the other through B and D. Coxeter's own solution was to consider the perpendicular bisectors of AC and BD. If they intersect use two concentric circles, and if they are parallel use two circles tangent to the mid-line of the parallel lines AC and BD. Since the points are not concyclic the





perpendicular bisectors cannot coincide.

Klamkin said that there is a classical construction, based on inversive geometry, for the center of a circle, using compass alone.

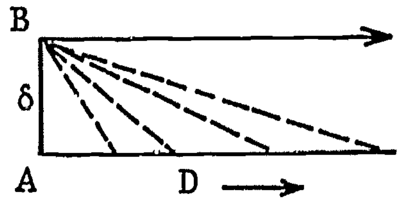
Reay showed how the "ancestor argument" used to prove the Schroeder-Bernstein theorem can be used to prove that if A and B are bounded subsets of the Euclidean plane, each with non-empty interiors, then A and B can be partitioned into non-overlapping subsets A_1, A_2 and B_1, B_2 such that A_1 is homothetic to B_1 and B_2 homothetic to A_2. The basic idea is to choose two dilatations, the first mapping A into the interior of B and the second mapping B into the interior of A and then apply the ancestor argument to obtain a one to one mapping, h, from A onto B with h made up of f and g^{-1}. The points corresponding in h via f determine the sets A_1 and B_1, those corresponding via g^{-1} determine A_2 and B_2.

Lecture V. Hyperbolic Geometry: The group of hyperbolic isometries.

Today I'd like to talk about hyperbolic geometry, the fifth type of geometry in this series of five lectures. Although I'll say a little about the hyperbolic plane, I'll focus mainly on hyperbolic 3-space, since it was in hyperbolic space that both Bolyai and Lobatchevsky worked. What I have to say can easily be extended to n dimensions.

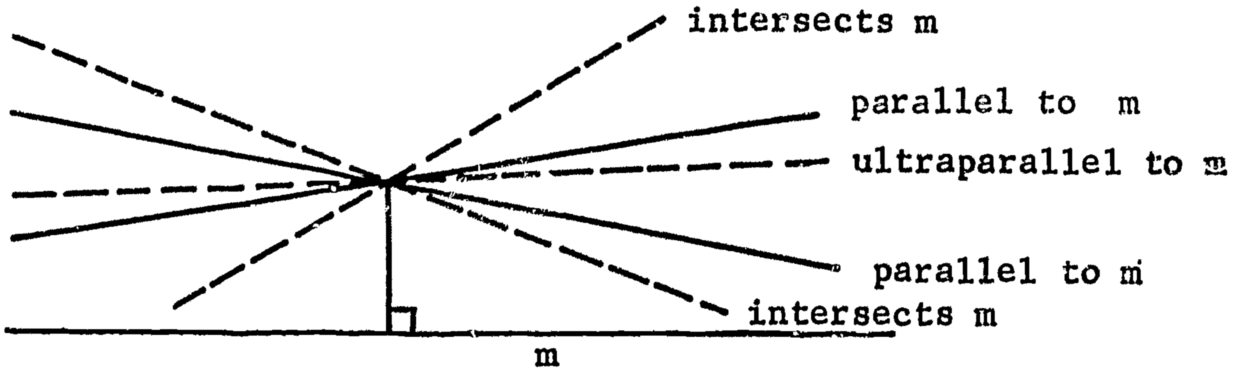
Hyperbolic and Euclidean spaces have many common properties, in fact one can do a lot of work in absolute space, which combines them. The order and continuity properties are all the same, and right angles are defined in both. The first major difference comes in the matter of parallelism.

Consider a line segment AB of length δ and choose a ray from A at right angles to AB. As a point D moves out along this ray from A the ray from B through D approaches a limit ray which in the Euclidean case is also perpendicular to AB. In the hyperbolic case, however, the angle between this limit ray and BA is less than a right angle and depends on δ. Since it is a definite function of δ we call it $\Pi(\delta)$, the angle of parallelism. We reserve the term parallel for this critical limit position. Any line BC such that



$$\Pi(\delta) < \text{angle ABC} < \pi - \Pi(\delta)$$

is said to be ultraparallel to AD. After proving that parallelism is symmetric and transitive, we can distinguish three kinds of pencils of lines in the hyperbolic plane. The first kind is the familiar pencil of all lines through a given point, called an intersecting pencil. The second, called a



pencil of parallels, is a pencil of all lines parallel in the same direction or, as Hilbert picturesquely described it, a pencil of lines with a common end. Any line has two ends in hyperbolic geometry rather than only one as in Euclidean geometry. If you move out along two rays that are ultraparallel, the distance between them may decrease for awhile but eventually it will increase. At the critical point where the distance is minimum the rays have a common perpendicular. Hilbert gave a nice geometric proof that ultraparallels always have a common perpendicular. Thus the third kind of pencil, a pencil of ultraparallels, is the pencil of all lines perpendicular to a given line.

There are three interesting types of curves in the hyperbolic plane corresponding to the orthogonal trajectories for these three types of pencils. A circle, like the familiar circle, is an orthogonal trajectory of an intersecting pencil and is the locus of all points at a constant distance from the point of intersection. The locus of points at a constant distance (on either side) from a fixed line is called an equidistant curve or sometimes a hypercycle. It is orthogonal to all lines in the pencil of ultraparallels perpendicular to this line, and each of its two branches is an orthogonal trajectory. In between circles and equidistant curves are the orthogonal trajectories of pencils of parallels, called horocycles.

The classification of isometries in the hyperbolic plane is almost the same as in the Euclidean plane. As before, each isometry is the product of at most three reflections (with lines for mirrors). A product of two reflections is a rotation, parallel displacement, or translation according as the lines intersect, are parallel, or are ultraparallel. Since ultraparallels have a common perpendicular, a translation is a product of half-turns (as in Euclidean geometry). In the hyperbolic case a translation has a unique axis and not a pencil of parallel axes. It is determined by its action on a point on this axis but not by its action on a point off the axis. The orbit of a point off the axis of a translation is on one arc of an equidistant curve. The product of three reflections is either a single reflection or a glide reflection, and the theory of glide reflections is *exactly* the same as in Euclidean geometry.

We do not have to bother classifying other kinds of similarities in hyperbolic geometry since the only similarities are the isometries. Therefore the classification of the various kinds of elements in the group of hyperbolic geometry is much simpler than in the Euclidean case. The group of isometries is generated by the reflections and contains the subgroup of direct isometries: rotations, parallel displacements, and translations.

Before leaving the hyperbolic plane I should mention a typical theorem of absolute geometry (the common ground of Euclidean and hyperbolic geometry). I think it is of great pedagogical interest to see how far one can go without committing oneself concerning parallelism. Consider the theorem about a quadrangle inscribed in a circle, which is usually stated as "the sum of two opposite angles is π." That of course is not true in hyperbolic geometry, but what is true is that *the sum of two opposite angles is the same as the sum of the other two opposite angles*. This is a nice substitute.

Now let us go on to three-dimensional hyperbolic space and consider the classification of similarities there. As in the hyperbolic plane there are no similarities except isometries and therefore it is a rather easy classification. As in Euclidean space, every isometry of hyperbolic space is the product of at most four reflections (in plane mirrors), in fact at most three if there is an invariant point. Thus every direct isometry with an invariant point is a rotation. Just as two lines in the hyperbolic plane are either intersecting, parallel, or ultraparallel, so two planes in hyperbolic space are either intersecting, parallel, or ultraparallel. *Ultraparallel* planes have a unique common perpendicular line, but *parallel* planes have neither a common point nor a common perpendicular. Thus the product of two reflections in space is a *rotation* if the mirrors intersect, a *parallel displacement* if they are parallel, or a *translation* if they are ultraparallel. The only other type of direct isometry, a product of four reflections, is a *twist* or *screw displacement*, which is the commutative product of a translation and rotation with the same axis. An opposite isometry which is the product of three reflections is either a *rotatory reflection*, a *parallel reflection* (two parallel mirrors perpendicular to the third mirror), or a *glide reflection* (two ultraparallel mirrors perpendicular to the third mirror).

With this one gets some feeling for the fact that hyperbolic geometry is homogeneous (all points are alike) and isotropic (all directions are alike). The ends of all the rays through a point O behave like points on a very large sphere enclosing everything, and the two ends of a line through O are like antipodal points on the sphere. You can think of the planes of hyperbolic space as cutting this big sphere in circles; moreover if you take a line through O, the pencil of ultraparallel planes perpendicular to this line cuts the sphere in a family of circles, among which we are tempted to regard the one with center O as a "great" circle until we recall that, the space being homogeneous, O is just like any other point. We are thus led to a geometry on the sphere in which circles play a crucial role and in which there is no distinction between great circles and small circles. In fact one can formalize this rather easily and show that the geometry of planes in a hyperbolic space is isomorphic to the geometry of circles on a sphere, i.e., to the geometry of the inversive plane. This is the isomorphism noted by Liebmann in 1905. [Editor's note. This isomorphism is discussed by Coxeter in his paper "The inversive plane and hyperbolic space," *Abh. Math. Sem. Hamburg*, vol. 29(1966), pp. 217-242.] It is a very interesting and useful isomorphism which can be used both ways--sometimes to derive results in hyperbolic space by considering their analogs in the inversive plane and other times to do the reverse.

The actual dictionary for the transition back and forth between the isometries of hyperbolic space and the homographies and antihomographies of the inversive plane appears on page 266 of the fifth edition of my *Non-Euclidean Geometry*. Let me quote a little of it. The reflections in planes of hyperbolic space correspond quite naturally to inversions in circles in the inversive plane. Products of two reflections, namely rotations, parallel displacements, and translations, correspond to *rotatory*, *parabolic*, and *dilative homographies*, since the planes are intersecting, parallel or ultraparallel according as the corresponding circles are intersecting, tangent, or non-intersecting. A twist corresponds to a *loxodromic* homography, the commutative product of a dilative homography and a rotatory homography. In this dictionary I am using the terminology given in Du Val's book *Homographies, Quaternions and Rotations*. This terminology is almost the same as in Ford's *Automorphic Functions* or in Schwerdtfeger's *Geometry of Complex Numbers*.

It is natural to go one step further and consider the inversive plane as the Euclidean plane (or the plane of complex numbers) with an extra point at infinity. You can choose any point you like in the inversive plane and call it the point at infinity and then get a corresponding Euclidean plane and represent each one of these transformations in a standard form. The basic transformation, inversion, corresponds in this way to the ordinary reflection in a line, since a line is the same as a circle from the point of view of inversive geometry. If we take this line to be the real axis in the plane of complex numbers then the standard analytic form of an inversion is $z' = \bar{z}$.

In this fashion we get the following table for homographies and direct isometries.

Direct isometries of hyperbolic space	Homographies	Direct similarities in the Euclidean plane	Analytic forms in the complex plane
rotation	rotatory	rotation	$z' = kz,\ k = e^{i\theta}$
parallel displacement	parabolic	translation	$z' = z + 1$
translation	dilative	dilatation	$z' = kz,\ k = e^{\delta}$
twist	loxodromic	spiral similarity	$z' = kz,\ k = e^{\delta+i\theta}$

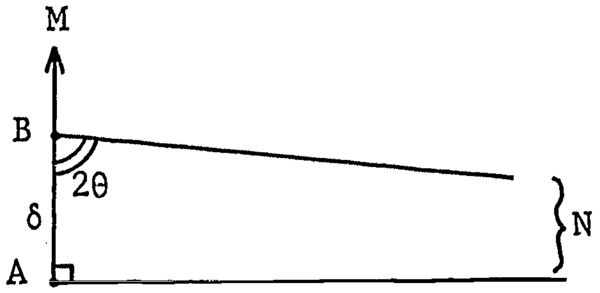
A special case of a rotation, the half-turn, corresponds to the Möbius involution, or in standard form $z' = - z$. The opposite isometries correspond to antihomographies in a similar fashion as follows.

Hyperbolic	Inversive	Euclidean	Complex
rotatory reflection	elliptic antihomography	rotatory inversion	$z' = k/\bar{z},\ k = e^{i\theta}$
parallel reflection	parabolic antihomography	glide reflection	$z' = \bar{z} + 1$
glide reflection	hyperbolic antihomography	dilative reflection	$z' = k\bar{z},\ k = e^{\delta}$

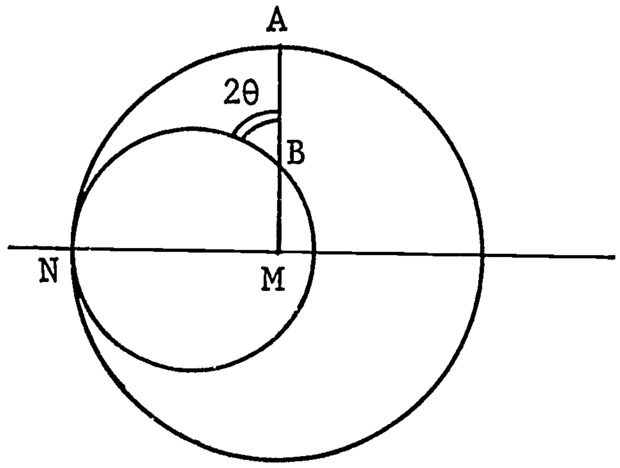
I should say a little about the metric features of this isomorphism between the two geometries of hyperbolic space and of the inversive plane. It is a conformal transformation, i.e., the angle between two intersecting planes is exactly the same as the angle between the corresponding intersecting circles. The distance between two points in hyperbolic space can be thought of as the distance between the two ultraparallel planes through these points and perpendicular to the line segment joining them. This distance is the minimal distance between the two planes. If we define this to be the "distance" between the corresponding non-intersecting circles in the inversive plane then, since it is an inversive invariant and is additive for coaxal circles, it must be the same as the inversive distance as defined in yesterday's lecture.

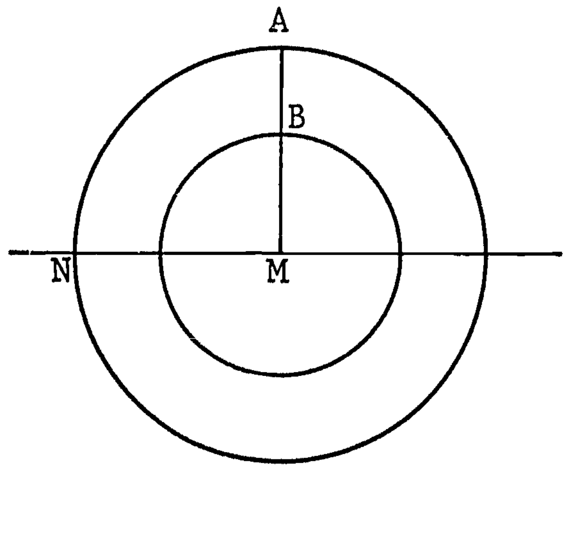
If we fix our attention on one particular hyperbolic plane, α, we can identify the lines in α with the planes perpendicualr to α, e.g., the geometry of lines in a horizontal plane is the same as the geometry of vertical planes. Since we have established an isomorphism between planes and circles, we see that the geometry of lines in a hyperbolic plane is isomorphic to the geometry of circles orthogonal to the circle representing that plane. Thus Poincaré's model is an immediate consequence of Liebmann's.

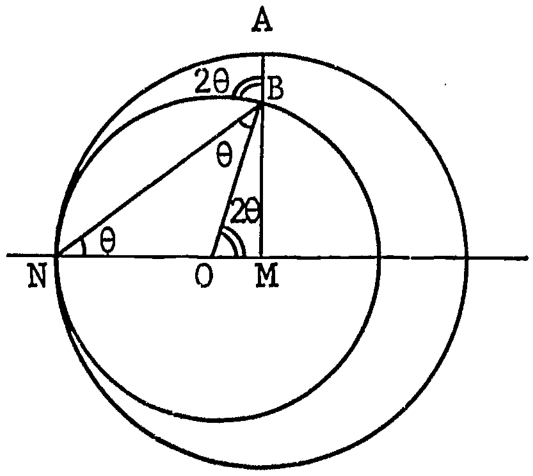
Now we have everything needed to prove the famous formula for the angle of parallelism: $\Pi(\delta) = 2 \arctan\ e^{-\delta}$. Recall the original figure in the hyperbolic plane, let N be the common end of the parallel rays from A and B and let $\Pi(\delta) = 2\theta$. Since we're working in the inversive plane we can adopt the Euclidean point of view and take the circle for our fixed plane to be a line. The lines in our plane will then



be represented conformally by circles and lines orthogonal to this line, so for example the representation of the parallel lines AN and BN is as in the adjacent diagram in which we have denoted one end of the line AB by M. Note that the line AN, orthogonal to the line AB, is represented by a circle centered at M. To find the Euclidean analog of the distance δ we recall the definition of inversive distance. The hyperbolic distance between A and B is the same as the distance between the ultraparallel planes perpendicular to AB. These planes are represented by two circles concentric







at M, whose inversive distance is $\log \frac{AM}{BM}$, thus $e^{\delta} = \frac{AM}{BM}$. If we construct lines joining B to N and the center O of the circle representing AN, then since angle BNO = $\frac{1}{2}$ angle BOM and angle BOM = 2θ, angle BNO = θ. Therefore $\cot \theta = \frac{NM}{BM}$, and since NM = AM we have $\cot \theta = \frac{AM}{BM} = e^{\delta}$. This completes the proof that $\Pi(\delta) = 2\theta = \arctan e^{-\delta}$. This proof, due to P. Szász, is much simpler than any other published proof.

To complete this lecture let me briefly mention two topics, either of which could keep us here for three or four hours.

In the Klein model of hyperbolic space one thinks of a sphere in Euclidean 3-space in which the chords of that sphere are the lines of hyperbolic space. Each plane α of Euclidean space which cuts the sphere represents a plane of hyperbolic space and can be replaced by its pole A with respect to the sphere. This is the vertex of an enveloping cone which touches the sphere along the circle cut out by the plane α. Or, if you prefer, you may think of the Klein model in terms of a non-ruled quadric in real projective space and replace each plane that cuts the quadric by its pole, a point outside the quadric. In either way the circles of inversive space and the planes of hyperbolic three space are represented by points <u>outside</u> a non-ruled quadric, so you get an <u>exterior-hyperbolic space</u> whose points represent circles in the inversive plane. Using this you can represent pencils of coaxal circles by ranges of points in the exterior-hyperbolic space. In other words, circles of the inversive plane correspond to events in a three-dimensional de Sitter's world; for in de Sitter's four-dimensional world of space-time you have events which can be mapped onto the points of an exterior hyperbolic four-space (the points outside a non-ruled quadric in real projective four-space). If you concentrate on a subspace you can cut out one dimension of the space and thus think of points outside an ordinary quadric. In this space a <u>timelike line</u> is a line which, when extended, will cut the quadric, a <u>spacelike line</u> is one which is a non-secant, and a <u>null line</u> is a tangent line. Successive events on the world-line of any observer are represented by nested circles in the inversive plane. The remark that I made yesterday, that inversive distance satisfies a non-triangle inequality for nested circles, is equivalent to the fact that a non-triangle inequality holds for timelike separations of events in de Sitter's world. The proper time of an observer going "out and back" will be less than that of one staying at home. This is of course connected with the twin paradox and related features of relativity theory.

My final remark is that the geometry of the inversive plane, i.e., the geometry determined by the linear fractional transformations, can be interpreted as the projective geometry of one dimension over the field of complex numbers. As I indicated in the lecture on projective geometry, you can think of a one-dimensional complex projective geometry as the geometry of points on a conic in the complex projective plane. Point pairs, which in this geometry determine secants of the conic, correspond to point pairs in the inversive plane which in turn correspond to lines of hyperbolic space. (Think of a line in hyperbolic space as determined by its two ends.) So you have a natural way for translating theorems about the complex projective plane into theorems about hyperbolic space. One nice example of this is the correspondence between Desargues' theorem in the complex plane and the Morley-Petersen theorem in hyperbolic space. Given a skew hexagon (in hyperbolic space) formed by six lines such that every consecutive pair are perpendicular, then each pair of opposite sides of this skew hexagon have a unique common perpendicular line. The Morley-Petersen theorem asserts that the three lines, determined in this way by the three pairs of opposite sides, have a common perpendicular. Thus we have a configuration of ten lines, each meeting three others at right angles, which is quite symmetrical and corresponds to the Desargues configuration 10_3. Of course, the Morley-Petersen theorem belongs to absolute geometry and thus holds in Euclidean space too.

Discussion.

Gleason asked how one could prove easily that two ultraparallel planes have a unique common perpendicular. Coxeter replied that here was a place where the isomorphism between hyperbolic space and the inversive plane could be used to advantage, for presumably one could set up the isomorphism without having to prove anything as complicated as this and then the following proof could be given. Pairs of planes correspond to pairs of circles which are either intersecting, tangent, or non-intersecting, so planes are either intersecting, parallel, or ultraparallel. In the third case, the case of non-intersecting circles α and β, the limiting points of the pencil $\alpha\beta$ are the unique point pair such that every circle through these points is perpendicular to α and β. Translating back to hyperbolic space, the line determined by the two ends corresponding to the two limiting points is the unique line such that every plane through this line is perpendicular to the two given ultraparallel planes.

Busemann, Coxeter, Gleason, and Prenowitz engaged in a spirited discussion of the relative merits of the Klein, Poincaré, and Liebmann models of hyperbolic space, especially in reference to the simplicity of the proof of the basic formula $\Pi(\delta) = 2 \arctan e^{-\delta}$. Busemann presented an alternate proof of the formula using the Klein model. Prenowitz and Coxeter agreed that the Liebmann model was in a slightly different class since it is an isomorphism between two geometries which can be used to provide either a model of hyperbolic space in terms of the inversive plane or to provide a model of the inversive plane in hyperbolic space. Gleason commented that the motivation given in the lecture for the Liebmann isomorphism provides a basis for proving that the Klein model is a model. Coxeter responded that a nice feature of Liebmann's presentation was that it unifies the Klein and Poincaré models, i.e., Liebmann observed that if you concentrate on the sphere of ends it doesn't matter which model you use; for the ends behave in the same way in both cases. Johnson observed that in a sense one could say that the Poincaré model is essentially an inversive model and that the Klein model is ess^ntially a projective model.

Gleason commented that he felt that in teaching this material, which he found fascinating, it would be a significant improvement to add one aspect of the subject that would relate it to other subjects. This one aspect is the way some of these transformations belong to the same one-parameter subgroups of the governing Lie group. In other words, whenever possible, attention should be focused not on one transformation alone but on the continuous family or families containing that transformation. This would lead naturally into the study of continuous groups. Coxeter replied that it was only lack of time that forced him to leave this out, but, for example, one of the things he had hoped to say was that a twist (in hyperbolic space) can be viewed as operating continuously and then the orbit of a point is a curve which one should call a helix since it has all the properties of an ordinary helix. The corresponding orbit of a loxodromic homography in the inversive plane is what one would call a loxodrome, the inverse of an equiangular spiral in the Euclidean plane. An interesting feature here is that there is only a singly infinite family of essentially different loxodromes--they depend on the angle alone. Johnson pointed out a nice way to visualize loxodromes on a sphere: follow a fixed (true not magnetic) compass bearing, not due north or south, on the sphere. Coxeter commented that a loxodrome has two poles and that the only reason an equiangular spiral seems to have only one is that the other is the point at infinity.

TWO APPLICATIONS OF GEOMETRY

Lectures by Herbert Busemann

(Lecture notes by Melvin Hausner)

Lecture I: The Simultaneous Approximation of n Real Numbers by Rationals.

Of all the problems in mathematics in which geometry can be used, I prefer those in which geometric insight makes the situation absolutely clear. The most famous example is Riemann's application of topology to the study of algebraic functions. Another example is Poincaré's application of hyperbolic geometry to the theory of automorphic functions. In this lecture and the next I shall consider two topics which need very little background but in which the application of geometry is most forceful. The first topic is a problem in approximation by rationals. The other is a problem in the calculus of variations, which will be discussed in the next lecture.

If $q > 0$ is an integer, and if $\rho_1, \ldots, \rho_{n-1}$ are $n-1$ real numbers, then there exist integers $p_1, \ldots, p_{n-1}$ such that

$$\left| \frac{p_i}{q} - \rho_i \right| \leqq \frac{1}{2q} , \qquad i = 1, \ldots, n-1.$$

To see this, we mark off on the real axis the points $0, \pm\frac{1}{q}, \pm\frac{2}{q}, \ldots$. Each point ρ_i falls in one of the intervals determined by the points $\frac{n}{q}$ and it is sufficient to take the nearest endpoint to obtain $\frac{p_i}{q}$ Indeed the distance to the nearest endpoint of an interval is never more than half the length of the interval.

Hermite proved the following stronger result. Suppose $\rho_1, \ldots, \rho_{n-1}$ are given real numbers. Then there exist integers $p_1, p_2, \ldots, p_n$ $(p_n > 0)$ such that

$$\left|\frac{p_i}{p_n} - \rho_i\right| \leqq \frac{1}{p_n^{n/(n-1)}}, \qquad i = 1,\ldots,n-1.$$

In this case, the error for large p_n is of course smaller than the $\frac{1}{2p_n}$ guaranteed by the previous argument. Actually, Hermite's result had another factor, larger than 1, on the right-hand side. We shall give Minkowski's geometric approach in what follows. It is worthy of note that the result certainly does not appear to have anything to do with geometry.

We consider a second problem also due to Hermite. Let

$$E^2(x) = \Sigma_{i,k=1}^{n} g_{ik}x_ix_k$$

be a positive quadratic form. Thus $E(x) > 0$ provided $x \neq 0$. Hermite considered the question of how small can $E(p)$ be, if p is a non-zero vector with integral coordinates. Hermite found a lower bound for such $E(p)$ which depends only on $D = \det(g_{ik})$ and the dimension n. This result also appears to have no geometric significance. However, we can see geometry entering by recalling that the volume of the ellipsoid given by $E(x) \leqq 1$ can be expressed in terms of D. If we let V_E be the volume of the ellipsoid $E(x) \leqq 1$, we have $V_E = \pi_n/\sqrt{D}$, where π_n is the volume of the unit ball ($\pi_n = \pi^{\frac{1}{2}n}/\Gamma(1+\frac{1}{2}n)$).

We can verify this in the simple case of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. Here

$$D = \begin{vmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{vmatrix} = \frac{1}{a^2b^2} \quad \text{and} \quad \pi_2 = \pi.$$

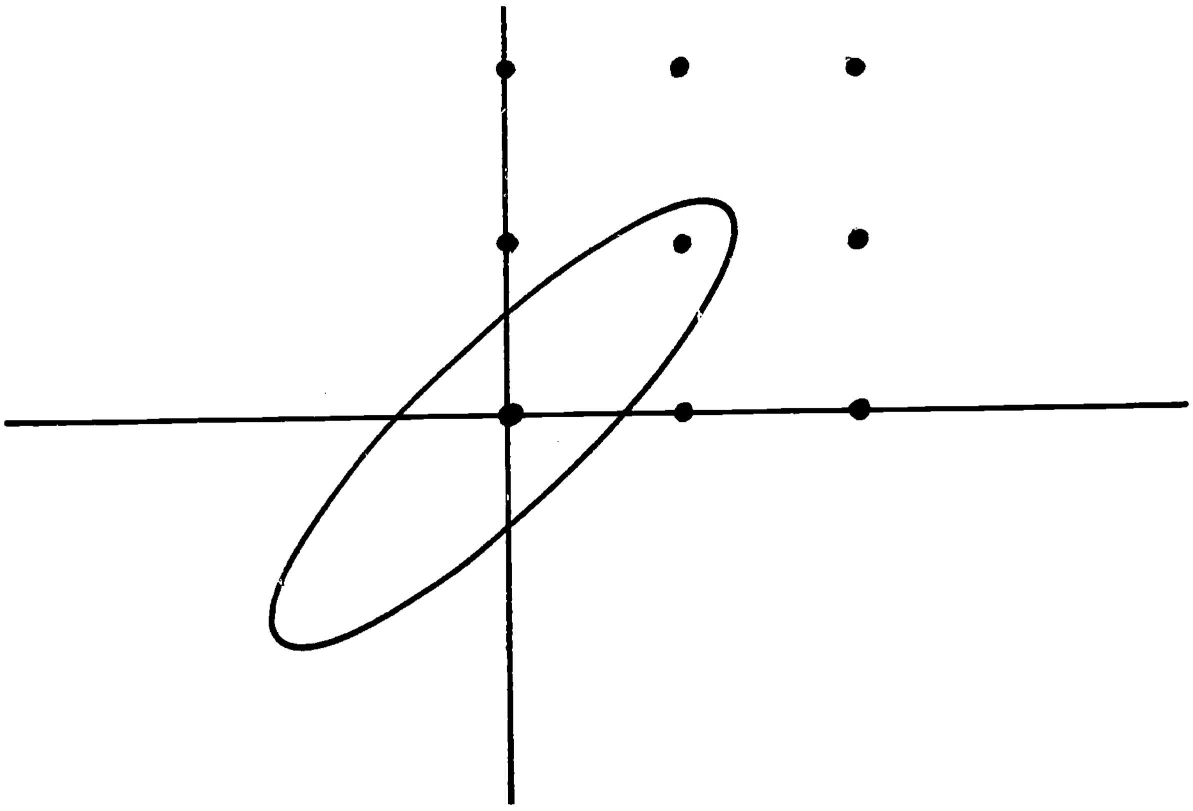
Thus the formula above gives $V_E = \pi_2/\sqrt{D} = \pi ab$.

In the general theory it is the volume V_E which enters, rather than the specific formula for it. Therefore we shall not concern ourselves with the proof of this formula. Using V_E, rather than the equivalent expression in terms of D, Hermite showed that it was possible to find a lattice point p such that

$$0 < E(p) \leqq 2/\sqrt[n]{V_E}.$$

(A lattice point is a point with integral coordinates.)

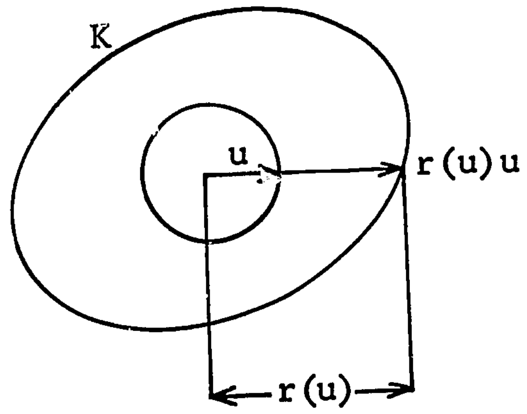
We can rephrase the theorem in the following way. If $V_E \geqq 2^n$, then the ellipse $E(x) \leqq 1$ must contain a lattice point other than the origin 0. In the diagram, any ellipse with center 0 and area $4 = 2^2$ will contain



a nontrivial lattice point, even if the ellipse is long and thin.

Minkowski observed that the same theorem is true if the ellipsoid is replaced by any centrally symmetric convex body with the origin as its center. In this case we also have the result that if $V_K \geqq 2^n$, then there exists a lattice point p in K with $p \neq 0$. Here V_K is the volume of K. In this general formulation the number 2^n is best possible. To see this we need only consider a cube centered at 0 with sides slightly less than 2. Minkowski also noted that this general theorem on convex bodies implies Hermite's result on rational approximations. We shall prove this, but we first consider a result of Minkowski on convex bodies.

We shall work in n-space. Suppose that K is a given (closed and bounded) convex body with center at the origin. We introduce a function $r(u)$, defined for all unit vectors in R^n: $r(u)$ is the largest number λ such tnat λu is in K. It is not hard to verify that $r(u)$ is continuous and that $r(-u) = r(u)$. Since $r(u)$ is continuous on the unit sphere it has a minimum value $r_1 > 0$ and a maximum value $r_2 < \infty$.



Corresponding to the function $E(x)$ for the ellipsoid, we now define the real-valued function $K(x)$ for all points of space:

$$K(0) = 0,$$

$$K(x) = \frac{|x|}{r(x/|x|)}, \quad \text{if} \quad x \neq 0.$$

Theorem 1. The function $K(x)$ has the following properties:

1. $K(x) > 0$ for $x \neq 0$.
2. $K(tx) = |t|K(x)$.
3. $K(x)$ is a convex function.

The proof of statement 1 is immediate. To prove statement 2, we note that it is trivial if $t = 0$ or if $x = 0$. Otherwise,

$$K(tx) = \frac{|tx|}{r(tx/|tx|)} = \frac{|t||x|}{r(\pm x/|x|)} = |t|K(x),$$

since r satisfies the equation $r(u) = r(-u)$. We shall find it more convenient to prove the following alternate formulation of statement 3:

3'. $K(x + y) \leqq K(x) + K(y)$.

We first show the equivalence of statements 3 and 3' (assuming statement 2). First, assuming 3', we have

$$K((1-\theta)x + \theta y) \leqq K((1-\theta)x) + K(\theta y) = (1-\theta)K(x) + \theta K(y)$$

if $0 \leqq \theta \leqq 1$. This proves that K is convex, and we have statement 3. To go the other way, assuming that K is convex, we have

$$\tfrac{1}{2}K(x + y) = K(\tfrac{1}{2}x + \tfrac{1}{2}y) \leqq \tfrac{1}{2}K(x) + \tfrac{1}{2}K(y).$$

Thus statement 3' follows from statement 3.

Before proceeding to the proof of 3', we mention two important facts about $K(x)$. These are:

4. $K(x) \leqq 1$ if and only if x is in the convex set K.

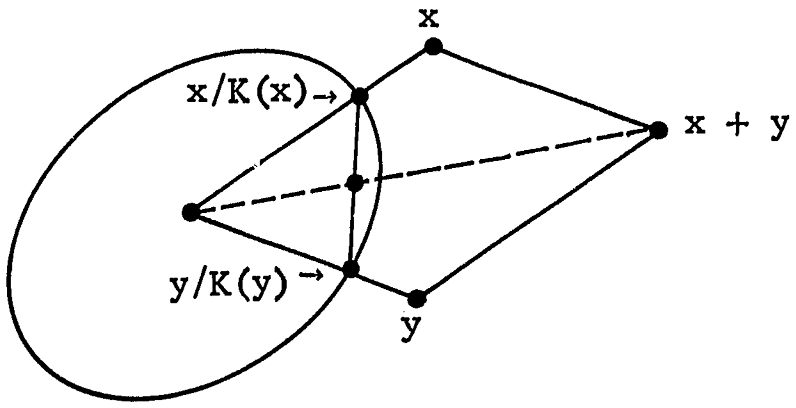
5. $\dfrac{|x|}{r_2} \leqq K(x) \leqq \dfrac{|x|}{r_1}$.

Both of these statements follow immediately from the definitions of $r(u)$ and of $K(x)$.

We now prove 3'. The statement is clearly true if $x = 0$ or if $y = 0$, so we may assume $x \neq 0$ and $y \neq 0$ with no loss in generality. The points $x/K(x)$ and $y/K(y)$ are in K, since

$$K(x/K(x)) = 1 = K(y/K(y)),$$

by property 2. Thus, using property 4, we see that they are points in K. We now express the point $\dfrac{x + y}{K(x) + K(y)}$ as a convex combination of these



two points. We have

$$\frac{x + y}{K(x) + K(y)} = \frac{K(x)}{K(x) + K(y)} \frac{x}{K(x)} + \frac{K(y)}{K(x) + K(y)} \frac{y}{K(y)} .$$

Thus $\frac{x + y}{K(x) + K(y)}$ is an element of K. Again by 4, this implies

$$K\left(\frac{x + y}{K(x) + K(y)}\right) \leqq 1.$$

Finally, using 2, we obtain the required result after multiplying the last inequality by $K(x) + K(y)$. This completes the proof of the theorem.

We remark that, conversely, if $K(x)$ is a real-valued function defined in n-space satisfying conditions 1, 2, and 3 (or 3'), it is an easy matter to show that the inequality $K(x) \leqq 1$ defines a centrally symmetric convex body. Thus there is a one-to-one correspondence between such functions and convex bodies with the origin as center. To see that $K(x) \leqq 1$ defines a convex body, we note that if $K(x) \leqq 1$ and $K(y) \leqq 1$, then

$$K((1-\theta)x + \theta y) \leqq (1-\theta)K(x) + \theta K(y) \leqq (1-\theta) + \theta = 1,$$

for $0 \leqq \theta \leqq 1$. Thus the inequality is satisfied by every convex combination of points which satisfy it.

For an ellipsoid the expression

$$E(x - y) = \sqrt{\Sigma_{ik}g_{ik}(x_i - y_i)(x_k - y_k)}$$

gives the ordinary Euclidean distance in oblique coordinates. One of Minkowski's two ideas in this development was the idea of generalizing this formula to introduce what we now call the Minkowski distance $m(x,y) = K(x-y)$. (The other main idea was the recognition that only the volume was involved in the formulation of the problem.) We note that while this distance function was introduced before Fréchet's general formulation, Minkowski proved the three conditions for a distance. Thus, defining $m(x,y) = K(x-y)$, we have

a) $m(x,x) = 0$,

b) $m(x,y) = m(y,x) > 0$ if $x \neq y$,

c) $m(x,y) + m(y,z) \geqq m(x,z)$ (the triangle inequality).

The verification is straight-forward:

a) $m(x,x) = K(x-x) = K(0) = 0$.

b) $m(x,y) = K(x-y) = K(-1(y-x)) = |-1|K(y-x) = m(y,x) > 0$ if $y \neq x$.

c) $m(x,z) = K(x-z) = K((x-y)+(y-z)) \leqq K(x-y) + K(y-z) = m(x,y) + m(y,z)$.

The distance $m(x,y)$ satisfies some further properties which we shall need. One of these properties is that it is translation invariant: only the difference $x-y$ enters into the definition. Thus

d) $m(x+a,y+a) = m(x,y)$.

More generally, if $x' = \beta x + a$ is a similitude, the Minkowski distance changes by the factor $|\beta|$.

e) $m(x',y') = |\beta|m(x,y)$ for dilations with a factor β.

A special case is a reflection in a point. This occurs when $\beta = -1$. Thus the metric is invariant under reflections in points.

We also remark that the closed ball of radius ρ about p given by the inequality $m(p,x) \leqq \rho$ can be transformed into the ball with center p and radius σ by a dilation with the factor σ/ρ. Thus the Minkowski metric has some of the features of the Euclidean metric. Of course some features are lost. For example, it is not invariant under rotations.

With this preparation, we are ready to solve the problem of Minkowski.

Theorem 2. If $K(x)$ is a function satisfying the hypothesis of Theorem 1, and if V_K is the Euclidean volume of the convex set K consisting of all points x with $K(x) \leqq 1$, then there exists a lattice point $p = (p_1,\ldots,p_n)$ such that

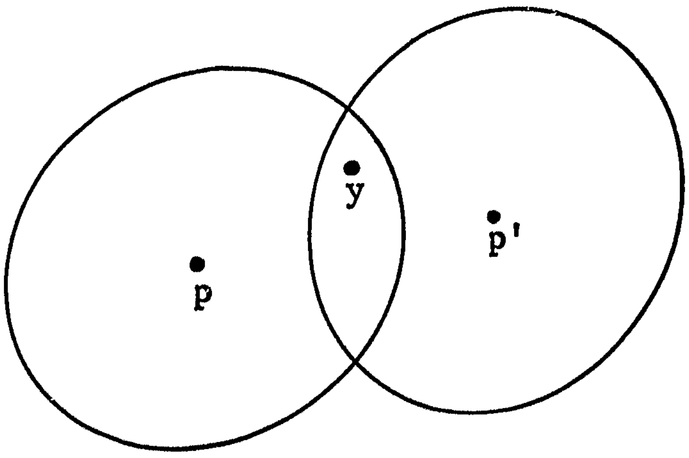
$$0 < K(p) \leqq \frac{2}{\sqrt[n]{V_K}} .$$

Another way of stating this result asserts that if the volume of the convex body K is at least 2^n, then there is a non-zero lattice point p in K. By symmetry, of course, there are at least two. In what follows p will be used to denote a lattice point. We now prove Theorem 2.

We first note that the inequality 5, $|x|/r_2 \leqq K(x)$, implies that $K(x) \to \infty$ as $|x| \to \infty$. Therefore the minimum of $K(p)$ over all non-zero lattice points exists and is a positive number. We shall let M be this minimum:

$$0 < M = \min_{p \neq 0} K(p).$$

We note further that M is the minimum Minkowski distance between two different lattice points: $m(p,p') = K(p-p')$. It follows that the closed balls of radius $\frac{1}{2}M$ about distinct lattice points p and p' do not have a common interior point. For if y were such a point, we would have $m(p,y) < \frac{1}{2}M$

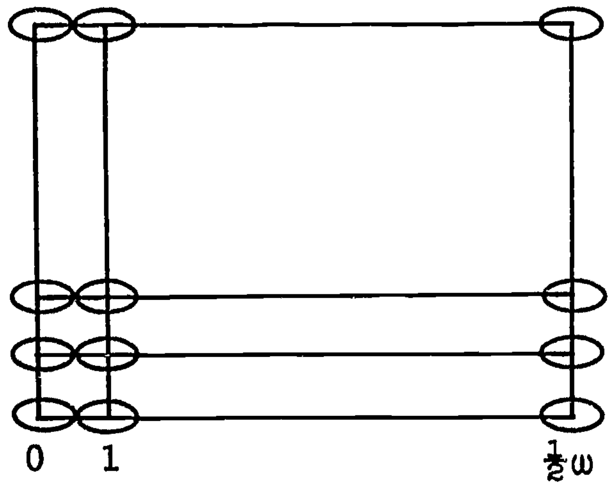


and $m(p',y) < \frac{1}{2}M$. It would follow that

$$m(p,p') \leqq m(p,y) + m(y,p') = m(p,y) + m(p',y) < \tfrac{1}{2}M + \tfrac{1}{2}M = M.$$

Thus we would have $m(p,p') < M$, which contradicts the definition of M as the least Minkowski distance between distinct lattice points.

Now let ω be any positive even integer. (In the end we shall let $\omega \to \infty$, but for the present, keep it fixed.) We now consider all of the lattice points whose coordinates are chosen from among the numbers $\{0, \pm 1, \ldots, \pm \tfrac{1}{2}\omega\}$, the lattice points inside the closed cube of length ω centered at 0. Since there are $\omega + 1 = 1 + 2(\tfrac{1}{2}\omega)$ choices for each coordinate, there are exactly $(\omega + 1)^n$ such lattice points. We call the set of these lattice points S_ω. We now cover each point p of S_ω with S_p, the closed ball (using the Minkowski metric) centered at p with radius $\tfrac{1}{2}M$, and we consider the volume of the union $\bigcup_{p \in S_\omega} S_p$. First, each ball has the same volume since the balls are translates of one another. Next, the ball centered at 0 with radius $\tfrac{1}{2}M$ is obtained from the unit ball by a dilation of factor $\tfrac{1}{2}M$. Thus the volume of S_p is equal to $(\tfrac{1}{2}M)^n V_k$, since dilations in n-space magnify volumes by the $n^{\underline{th}}$ power of their factors. Finally, since there are $(\omega + 1)^n$ lattice points in S_ω and since the spheres have no common interior points, it follows that the total volume of this set of balls is



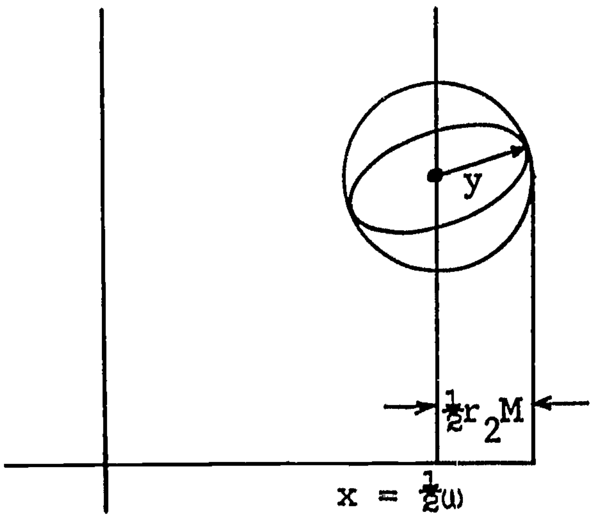
$$V = (\omega + 1)^n (\tfrac{1}{2}M)^n V_K,$$

where V_K is the volume of the unit ball.

We shall now obtain an upper bound for this volume by enclosing the set in a cube. Since the spheres at the boundary of the cube $|x| \leq \omega$ stick out for a Minkowski distance $\frac{1}{2}M$, it is necessary to estimate this as a Euclidean distance and to enlarge this cube by an appropriate amount to contain the union of our spheres. If $K(x) \leqq \frac{1}{2}M$ then, using inequality 5, we have

$$|x| \leqq r_2K(x) \leqq r_2(\tfrac{1}{2}M).$$

Therefore, if we extend the cube for a Euclidean distance of $r_2(\frac{1}{2}M)$ in each direction, the resulting cube will enclose all of our spheres. The side of this larger cube has Euclidean length $2(\frac{1}{2}\omega + r_2(\frac{1}{2}M)) = \omega + r_2M$, so the volume is $(\omega + r_2M)^n$. We therefore obtain the volume inequality



$$(\omega + 1)^n(\tfrac{1}{2}M)^nV_K \leqq (\omega + r_2M)^n.$$

Dividing by $(\omega + 1)^n$, we have

$$(\tfrac{1}{2}M)^nV_K \leqq \left(\frac{\omega + r_2M}{\omega + 1}\right)^n.$$

Finally we let $\omega \to \infty$ and obtain $(\frac{1}{2}M)^nV_K \leqq 1$ or $M \leqq 2/\sqrt[n]{V_K}$. Since M is the minimum value of $K(p)$, for $p \neq 0$, we have the result.

Again we rephrase the result: A convex body in n-space, symmetric about the origin, with volume at least 2^n, contains a non-zero lattice point.

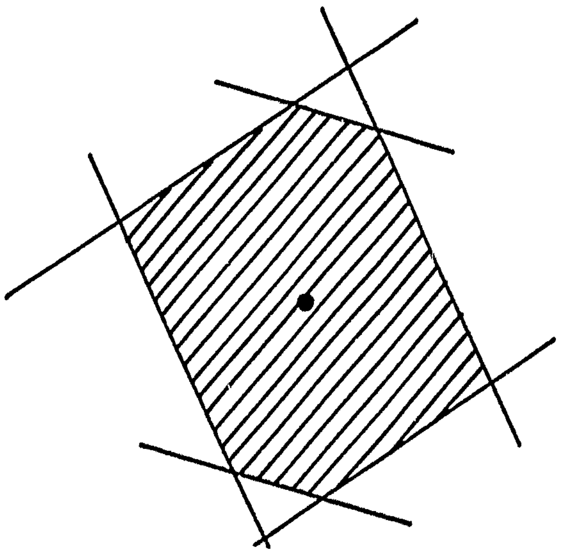
We started with the theorem of Hermite concerning the approximation of real numbers by rationals. As we shall see this is an easy consequence of Minkowski's result.

Suppose $L_i(x)$ is a non-trivial linear function of x:

$$L_i(x) = a_{i1}x_1 + \cdots + a_{in}x_n,$$
$$(a_{i1},\ldots,a_{in}) \neq (0,\ldots,0), \quad i = 1,\ldots,\nu,$$

where $\nu \geqq n$. We shall suppose that the matrix (a_{ik}) has rank n.

We now consider the system of ν linear inequalities given by $|L_i(x)| \leqq 1$. Each of these inequalities has as its solution an infinite strip, symmetric about the origin, and bounded by the hyperplanes $L_i(x) = \pm 1$, $i = 1,\ldots,\nu$. The intersection of these strips is a convex set with the origin as center. The condition that the matrix of coefficients have rank n assures us that this set is a bounded convex set. (Geometrically, the rank condition states that the normals to the bounding hyperplanes span n-space.) If we let K denote this intersection, the point x is in K if and only if x satisfies each of the ν inequalities $|L_i(x)| \leqq 1$. Therefore, applying Minkowski's theorem, we can make a statement about non-trivial integral solutions of this system of inequalities.



Theorem 3. Let $L_i(x) = \Sigma_k a_{ik}x_k$, $i = 1,\ldots,\nu$, and suppose that the matrix (a_{ik}) has rank n. Let V_K be the volume of K, the set of all points satisfying the inequalities $|L_i(x)| \leqq 1$. Then there is a lattice point

p satisfying

$$|L_i(p)| \leqq 2/\sqrt[n]{V_K}, \quad p \neq 0, \quad \text{for} \quad i = 1,\ldots,\nu.$$

With these preliminaries we turn to the theorem on approximation by rationals.

Theorem 4. Let $\rho_1,\ldots,\rho_{n-1}$ be given real numbers. Then there exist integers $p_1,\ldots,p_n$ with $p_n > 0$ such that

$$\left|\frac{p_i}{p_n} - \rho_i\right| \leqq \frac{1}{p_n^{n/(n-1)}} .$$

To prove this theorem, we specialize the discussion above to the case $\nu = n$. In this case we have n strips, meeting in a parallelepiped whose volume can be found easily. We suppose then that $L_i(x) = \Sigma_k a_{ik}x_k$, $i = 1,\ldots,n$, $\Delta = \det(a_{ik})$, and K is the convex set consisting of all points satisfying the inequalities $|L_i(x)| \leqq 1$. To find the volume of K we change coordinates, taking $x_i' = \Sigma_{k=1}^n a_{ik}x_k$. The convex set K, given by the inequalities $|\Sigma_{k=1}^n a_{ik}x_k| \leqq 1$, goes into the convex set K', given by $|x_i'| \leqq 1$. But K' is simply a cube of side 2, whose volume is 2^n. The Jacobian of the transformation is $\det(a_{ik}) = \Delta$. Thus, the formula for changing variables gives

$$2^n = \iint_{K'} dx_1'\cdots dx_n' = \iint_K |\Delta| dx_1\cdots dx_n = |\Delta| V_K.$$

Hence $V_K = 2^n/|\Delta|$. Consequently the Minkowski theorem, applied to this special case, shows that there is a non-zero lattice point p satisfying

$$|L_i(p)| \leqq 2/\sqrt[n]{V_K} = |\Delta|^{\frac{1}{n}}, \quad \text{for} \quad i = 1,\ldots,n.$$

We now choose particular linear functions as follows:

$$L_i(x) = x_i - \rho_i x_n, \quad i = 1,\ldots,n-1,$$
$$L_n(x) = x_n t^{-n}$$

where t may be any number greater than 1. We can compute Δ for this system easily:

$$\Delta = \begin{vmatrix} 1 & & & & -\rho_1 \\ & 1 & & & \cdot \\ & & \cdot & & \cdot \\ & 0 & & \cdot & \\ & & & & t^{-n} \end{vmatrix} = \frac{1}{t^n}.$$

Thus $\Delta = 1/t^n$ and $|\Delta|^{1/n} = 1/t$. Applying our general results, we obtain a system of integers p_i, $i = 1,\dots,n$, not all zero, for which

$$|p_i - \rho_i p_n| \leqq \frac{1}{t}, \quad i = 1,\dots,n-1,$$

$$\left|\frac{p_n}{t^n}\right| \leqq \frac{1}{t}.$$

Hence p_n cannot vanish, since the first $n-1$ relations would then imply that each p_i vanished. By considering $-p_i$ instead of p_i, we may assume $p_n > 0$. We therefore conclude that for each $t > 1$, there are integers $p_1,\dots,p_n$, $p_n > 0$, satisfying

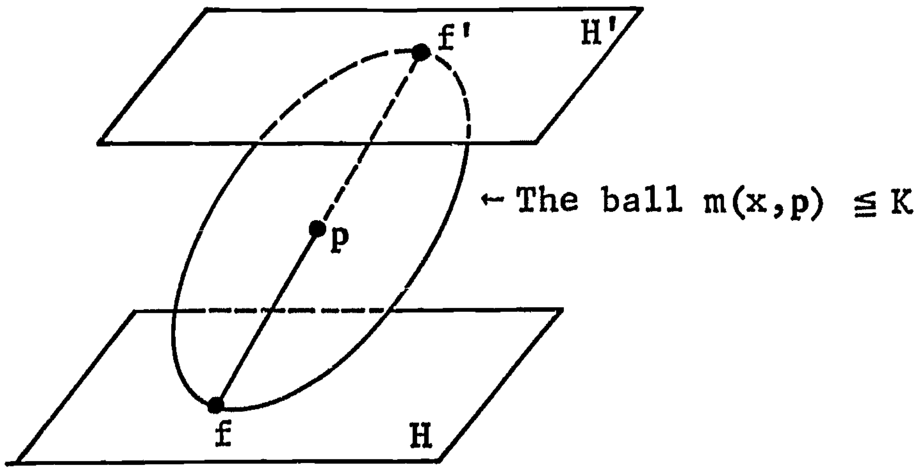
$$\left|\frac{p_i}{p_n} - \rho_i\right| \leqq \frac{1}{tp_n}, \quad 0 < p_n \leqq t^{n-1}.$$

This result is somewhat stronger than the result we set out to prove. We obtain Hermite's result by noting that $p_n^{1/(n-1)} \leqq t$:

$$\left|\frac{p_i}{p_n} - \rho_i\right| \leqq \frac{1}{p_n^{1/(n-1)} \cdot p_n} = \frac{1}{p_n^{n/(n-1)}}.$$

The result can be improved somewhat. For example, the factor $(n-1)/n$ can be introduced into the upper bound. This was done by Minkowski, also geometrically.

Before leaving the field of Minkowskian geometry, I should like to give another striking application. We defined a function $K(x)$ associated with any compact centrally symmetric convex body K; then we defined the Minkowski distance $m(x,y) = K(x-y)$. Suppose H is a hyperplane and p is a point not on H. Then it is an easy matter to show that there is a shortest Minkowski



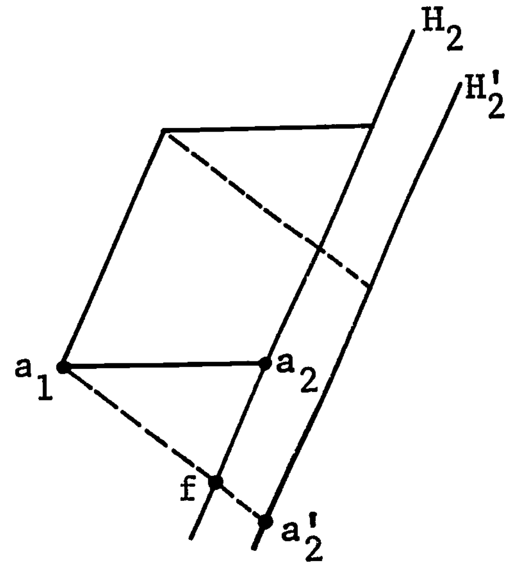
distance from p to H. (Here we use p to denote any point, not just a lattice point.) To see this, we note that the Euclidean distance from p to a point h on H approaches infinity as h approaches infinity on H. But the Minkowski distance is greater than some positive multiple of the Euclidean distance. Hence, in order to find the shortest distance, we may assume that h is in some large ball with center p. Then we can apply the usual compactness argument to show that the shortest distance is achieved. (If K has a (k-1)-dimensional face, the nearest point may not be unique.) Let us call a nearest point f on H the foot, as in Euclidean geometry, and let us call the line through p and f a perpendicular from p to H. Then, because of the properties of the Minkowski distance under dilations, it is easy to prove that this line is a perpendicular to H from any point on it. If a Minkowski ball centered at p with radius m(p,f) is constructed, then H will be a hyperplane supporting this ball at f. Because of central symmetry, the hyperplane H' obtained by reflecting H through p will support the ball at f', the reflection of f through p. The line through p and f is perpendicular to H' too. Any line parallel to pf is also perpendicular to H.

We now discuss a problem which was considered by many people and solved only in a few special cases until A. E. Taylor* observed that its solution is a trivial consequence of the ideas we have been discussing.

Problem. Let K be a compact convex centrally symmetric body. Is it possible to circumscribe a box (i.e., a parallelepiped) about K so that the center of each (n-1)-face is in K?

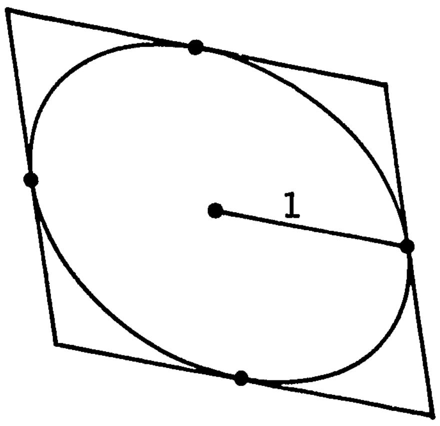
To solve this problem, we consider first the following problem, which will turn out to be useful for solving the original problem. *Given* $\rho > 0$, *construct the box of maximum volume, each of whose sides has Minkowski length* ρ.

This problem clearly has a solution, since, if one corner is kept fixed at the origin, the other vertices range over a compact set. (Note that we use the term box in an extended sense, including possibly degenerate parallelepipeds.) We claim that an edge a_1a_2 is perpendicular to the hyperplanes H_1 and H_2 which carry the (n-1)-faces that intersect the edge in its endpoints a_1, a_2. For if a_2 is not a foot of a_1 in H_2, let f be such a foot. Then $m(a,f) < \rho$. Prolong a_1f beyond f to the point a_2' for which $m(a_1,a_2') = \rho$. Then translate the (n-1)-face in H_2 so that a_2 falls on a_2'. The box spanned by the translated face and a_1 then has greater volume than the original box, but all its edges have Minkowski length ρ.



Bulletin of the American Mathematical Society, 53(1947), 614-616.

Finally, to solve the original problem, we take $\rho = 2$. The box of maximum volume will be the required box. We merely translate it so that its center is at the origin and take K as the unit ball in Minkowski geometry. Since each edge of the box has Minkowski length 2, it follows that each face is a supporting hyperplane of K. For the line from the origin, parallel to the edges and meeting a face in one point, is perpendicular to that face, since it is a translate of an edge. But this line meets the hyperplane in its center, and it meets at a Minkowski distance 1 from the origin. Hence the center of any face is in K, and the proof is finished.



Lecture II: An Application of Integral Geometry to the Calculus of Variations.

My second example of an application of geometry to another field of mathematics uses an idea from integral geometry but does not actually use integral geometry itself. This idea provides the basis for a simple proof of a theorem which had previously been considered difficult; but nobody would have thought of this proof unless he had seen integral geometry. We therefore give a brief description of integral geometry in a very simple setting.

In the plane it is reasonable to regard the area of a domain D as "the number of points in the domain." If we write $p = (x_1, x_2)$ and $dp = dx_1 dx_2$, we simply call

$$\int_D dx_1 dx_2 = \int_D dp$$

the number of points in D. In integral geometry we count lines and other objects in similar ways. That is, we introduce a suitable measure or density dL in the space of lines and evaluate $\int_X dL$ over some set X of lines. The answer is called "the number of lines in X." However, to give geometrically significant results this measure must be invariant under rotations and translations. Abstractly we have a set of objects and a group acting on these objects. We are required to introduce a density which is invariant under this group. It might be thought that one can always be found if the group is nice enough. But it is not so. For example, the linear subspaces of a fixed dimension in a projective space do not have a density invariant under the projective group. (A theorem of Chern gives general conditions for such a density to exist.)

However, we can easily find a suitable density for the lines in the plane. We first find suitable coordinates to express these lines and then introduce a density in the space of these coordinates. Natural coordinates are introduced if we use the normal form for the line:

$$x \cos \varphi + y \sin \varphi - p = 0, \quad 0 \leqq \varphi < 2\pi, \quad p \geqq 0.$$

We shall use φ and p as coordinates for the line and we shall show how to find a density which is invariant under the group of Euclidean motions.

We suppose that $f(p,\varphi)\, d\varphi\, dp$ is such a density. That is, if X is a set of lines in (φ,p) space, the number of lines in X will be given by $\int_X f(p,\varphi)\, d\varphi\, dp$. If X is a set of lines which is transformed by a motion into a set X^*, invariance requires that

$$\int_X f(p,\varphi)\, d\varphi\, dp = \int_{X^*} f(p^*,\varphi^*)\, d\varphi^*\, dp^*.$$

(It will turn out here that we can choose $f(p,\varphi) \equiv 1$ identically. However, this is an accident which depends on our choice of coordinates to describe the lines.)

We now compute what a motion does to lines and see how the motion transforms a set X. Any motion (on points) is given by the equations

$$x = x^* \cos \alpha - y^* \sin \alpha + a,$$

$$y = x^* \sin \alpha + y^* \cos \alpha + b.$$

The equation $x \cos \varphi + y \sin \varphi - p = 0$ is transformed into

$$(x^* \cos \alpha - y^* \sin \alpha + a) \cos \varphi + (x^* \sin \alpha + y^* \cos \alpha + b) \sin \varphi - p = 0,$$

which after simplification becomes

$$x^* \cos (\varphi-\alpha) + y^* \sin (\varphi-\alpha) - (p - a \cos \varphi - b \sin \varphi) = 0.$$

The coordinates φ^*, p^* of the transformed line are therefore given by the equations

$$\varphi^* = \varphi - \alpha,$$

$$p^* = p - a \cos \varphi - b \sin \varphi,$$

if $p - a \cos \varphi - b \sin \varphi \geqq 0$. If not, we do not have normal form, so we would have

$$\varphi^* = \varphi - \alpha + \pi,$$

$$p^* = -(p - a\cos\varphi - b\sin\varphi).$$

Thus

$$d\varphi^* = d\varphi$$

$$dp^* = \pm(dp + (a\sin\varphi - b\cos\varphi)\,d\varphi).$$

The Jacobian of this transformation is ±1. But since it is the absolute value of the Jacobian which appears in the formula for a change of variables in a double integral, we have

$$d\varphi^*\,dp^* = d\varphi\,dp.$$

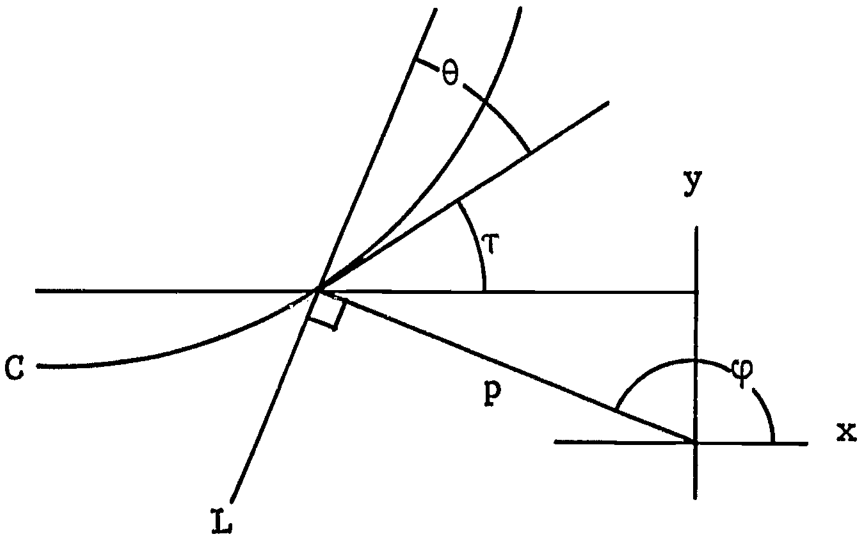
Therefore group invariance, together with the formula for changing variables in a double integral, gives

$$\int_X f(p,\varphi)\,dp\,d\varphi = \int_{X^*} f(p^*,\varphi^*)\,dp^*\,d\varphi^* = \int_X f(p^*,\varphi^*)\,dp\,d\varphi.$$

Then since X is arbitrary and any line may be moved into any other line, it follows that $f(p,\varphi) = f(p^*,\varphi^*)$ or $f(p,\varphi) = \text{constant}$. Since we can always multiply by a constant, we shall normalize so that $f(p.\varphi) = 1$. Thus $\int_X dp\,d\varphi$ is the number of lines in a set X of lines, coordinatized by p and φ. This number is left invariant if a motion is applied to X.

We now consider a typical problem in integral geometry. Given a curve C (assumed sufficiently differentiable), how many lines intersect the curve C? In the counting process we count each line as many times as it intersects the curve C. More formally, let $n(L \cap C)$ be the number of points in $L \cap C$. We wish to compute $\int n(L \cap C)\,dL$.

In order to make this computation, we choose as coordinates for a line which intersects the curve, the arc-length s along the curve and θ, the angle which the line makes with the tangent vector, measured positively. Here $0 \leqq s \leqq \lambda$, where λ is the length of C, and $0 \leqq \theta \leqq \pi$. It is necessary



to relate these coordinates with p and φ before we can make our computation. We introduce the angle τ which the curve makes with the x-axis. Then the coordinate φ of the line is clearly $\varphi = \theta + \tau \pm \frac{1}{2}\pi$. Therefore,

$$d\varphi = d\theta + \frac{d\tau}{ds}\,ds.$$

If the point $(x(s), y(s))$ lies on the line $x \cos\varphi + y \sin\varphi - p = 0$,

$$x(s) \cos\varphi + y(s) \sin\varphi - p = 0.$$

Taking differentials and recalling that $\frac{dx}{ds} = \cos\tau$, $\frac{dy}{ds} = \sin\tau$, we obtain after some simplification

$$dp = (-x \sin\varphi + y \cos\varphi)\,d\varphi + \cos(\varphi - \tau)\,ds.$$

We now compute $\frac{\partial(s,\theta)}{\partial(p,\varphi)}$ in two stages:

$$\frac{\partial(s,\theta)}{\partial(s,\varphi)} = \begin{vmatrix} 1 & 0 \\ \frac{d\tau}{ds} & 1 \end{vmatrix} = 1,$$

while

$$\frac{\partial(s,\varphi)}{\partial(p,\varphi)} = \begin{vmatrix} \cos(\varphi-\tau) & - \\ 0 & 1 \end{vmatrix} = \cos(\varphi-\tau).$$

But $\varphi = \theta + \tau \pm \frac{1}{2}\pi$. So $\cos(\varphi-\tau) = \cos(\theta\pm\frac{1}{2}\pi) = \pm \sin\theta$. Since absolute values of Jacobians are desired the sign is of no importance and

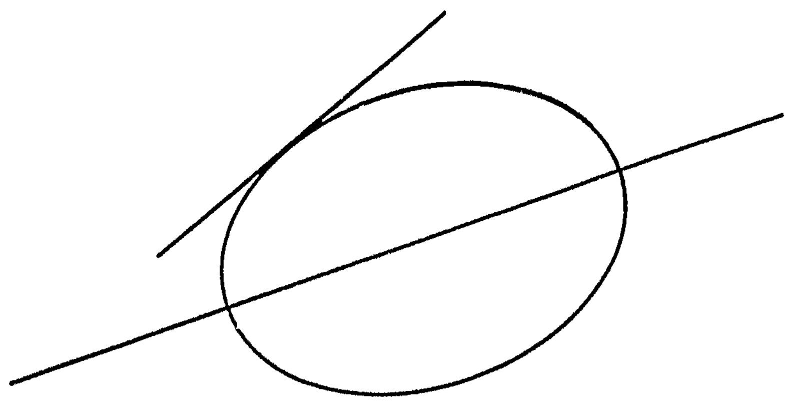
$$\left|\frac{\partial(s,\theta)}{\partial(p,\varphi)}\right| = \left|\frac{\partial(s,\theta)}{\partial(s,\varphi)}\right| \left|\frac{\partial(s,\varphi)}{\partial(p,\varphi)}\right| = \left|\cos\ (\varphi-\tau)\right| = \left|\sin\theta\right|.$$

Now we use the change of variables formula to transform $\int n(L \cap C)\, dL$ and obtain

$$\int n(L \cap C)\, dL = \int_0^\lambda \int_0^\pi |\sin\theta|\, ds\, d\theta = 2\lambda.$$

This result is very surprising. Except for a constant factor, the number of lines which intersect the curve is the length of the curve. For example, if a curve of a given length is crushed, to make it wiggle, there will be fewer lines meeting the curve, but these lines will meet it more often to compensate.

We make a few observations concerning this result. For a segment of length λ, each line which intersects the segment intersects it only once. The one line which is determined by the segment meets it infinitely often, of course, but one line by itself is of measure zero and can be ignored. For the boundary of a convex curve of length λ, each intersecting line will meet it twice, with the exception of the supporting lines. But the supporting lines form a set of measure zero, so they may be ignored. Thus for a convex set in the plane, the number of



lines which meet its boundary K is λ, the length of K: $\int_{L \cap K \neq \emptyset} dL = \lambda$.

Before leaving the subject of integral geometry, I would like to mention that this topic would make an excellent course. The problems are always well liked by all the students who can be interested at all. There are also generalizations to space. For example, the density of planes in space. The number of planes which intersect a space curve, counted as many times as they intersect the curve, is again an absolute constant times the length of the curve. Similarly the number of lines which intersect a surface, each counted as many times as it intersects the surface, is an absolute constant times the area of the surface. These results generalize to r-dimensional surfaces in n-space, where we count the number of (n-r)-dimensional planes which intersect the surface, with the correct multiplicity.

Another example is the following one. Suppose C_1 and C_2 are two smooth curves in the plane. Suppose C_2 is kept fixed, but C_1 is rigidly moved to all possible positions. We obtain a density for the various images of C_1, which we may denote dC_1. Then we find

$$\int n(C_1 \cap C_2)\, dC_1 = 4\lambda(C_1)\lambda(C_2).$$

As far as the possibilities for teaching this subject are concerned, the only difficulty is the use of the Jacobian in the change of variables formula for a multiple integral. But if we stay in the plane, even this amounts at most to a 3×3 determinant. If the elements of multilinear algebra are known, then the derivation of the formula for a change of variables is very easy and all of these results can be proved very simply.

We shall now see how these ideas can be used in the calculus of variations. Recall that if $f(x,y,y')$ is a function of three variables, one of the problems in the calculus of variations is to find a function $y = y(x)$ which is an extremal for $\int f(x,y,y')\,dx$ where the endpoints of the curve $y(x)$ are given. The classical Euler equation gives a necessary condition that y be an extremal:

$$\frac{\partial^2 f}{(\partial y')^2}\, y'' + \frac{\partial^2 f}{\partial y \partial y'}\, y' + \frac{\partial^2 f}{\partial y \partial x} - \frac{\partial f}{\partial y} = 0.$$

where x, y, and y' are taken to be independent variables for the computation of the partial derivatives. The solutions $y(x)$ of this equation may maximize the given functional, minimize it, or do neither. They are called extremals in all cases.

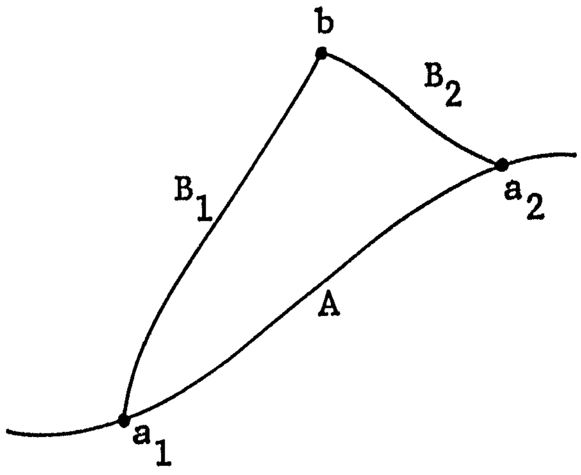
Darboux posed the following problem. Suppose that in the plane we are given a two parameter system of curves. Then can this system be regarded as the extremals of a variational problem? His method was roughly as follows. We can consider this system as the solutions of a differential equation $y'' = \varphi(x,y,y')$. If $\varphi(x,y,y')$ is substituted for y'' in Euler's equation, we obtain a partial differential equation for f, as a function of the three independent variables x,y,y'. Finally, using the appropriate existence theorem, we can solve the equation, at least locally for f. Some of the details are omitted, of course. Geometrically we would want to consider vertical lines. Also, the exact meaning of a two parameter system of curves has been left open. But it is clear that the solution by this method is local. Furthermore, all that is known is that the given curves, even locally, are extremals of the problem. Nothing is known about the question of whether they are minimal curves, etc. (Darboux was actually interested in minima.)

Before going into the precise statement of the result, I would like to sketch the idea. Our method will give <u>global</u> results, and furthermore, the given curves will turn out to be unique <u>minimal</u> curves for a variational problem.

Assume that the curves are parametrized by points in the (u,v)-plane. Thus each (u,v) is mapped into a curve $L(u,v)$ of the given system, in a one-to-one way. Since the curves are not necessarily graphs of functions $y = y(x)$, we should consider curves in parametric form, and similarly express the variational problem as a minimization of $\int F(x,y,\dot{x},\dot{y})\,dt$. What we shall find is not F itself, but the value of F integrated along a curve of the given system. Then, by differentiating, we shall be able to recover the function F. Therefore, we shall find

$$\lambda(A) = \int_A F(x,y,\dot{x},\dot{y})\,dt$$

where A is any curve, or portion of a curve, of the given system. We shall arrange to have $\lambda(A)$ equal to the length of A in a suitable geometry, and the curves of the system the "lines". In this case, the triangle inequality "the line is the shortest path between two points" will show that the curve A minimizes the length between two points. Thus we shall need to have the triangle inequality. Now suppose that a_1 and a_2 are joined by a curve of the given system. Take an arbitrary point b and join b to a_i by the curve B_i of the system. What is required is the inequality:



$$\text{Length } B_1 + \text{Length } B_2 \geqq \text{Length } A$$

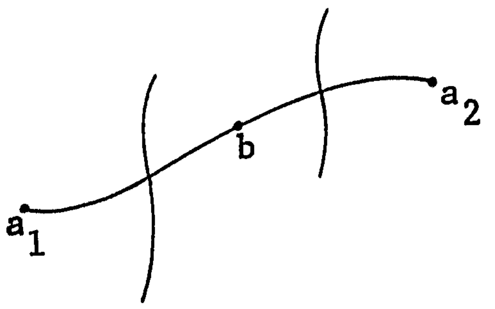
with equality if and only if b is a point of A.

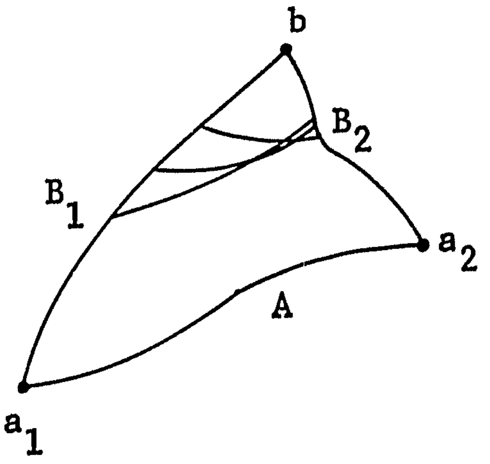
It is here that integral geometry gives us the leading idea. Although integral geometry is not used, no one could get the idea of this approach unless he has seen some of the ideas of the subject. We remember that the length of a segment is, up to a constant factor, the number of lines which intersect that segment. Therefore, all we need is a density for all the curves of the system, and we could compute this number, and hence the length of any of the curves of the system. Since we are not concerned with rotations and translations, we may choose *an arbitrary positive continuous function* $\varphi(u,v)$ as a density, and use $dL(u,v) = \varphi(u,v)\ du\ dv$. We now *define* the length $\lambda(A)$ by the formula

$$\lambda(A) = \int_{L \cap A \neq \emptyset} dL(u,v)$$

and we maintain that this gives a length function in the sense described above.

To see this, note that if b is on the curve A, then any curve which meets A meets either arc a_1b or ba_2. However, the curves which meet both





of these arcs form a set of measure zero in this density, and we have $\lambda(A) = \lambda(B_1) + \lambda(B_2)$. On the other hand, if b is not in A, every line which intersects A will intersect B_1 or B_2. This proves the triangle inequality. To obtain the strict inequality, we observe that there is an open set of lines which intersect both B_1 and B_2 and not A, and this set has positive measure. This proves the strict inequality in this case.

Thus the idea is quite a simple one, but it probably would not have been conceived unless one had already seen integral geometry.

To make the argument more rigorous, we must describe the curve system Σ more accurately. We first suppose that the curves are given by coordinates in the plane. We suppose that each curve of a system is an open Jordan curve, and may be parametrized by a function

$$P = P(t), \quad -\infty < t < \infty,$$

where

$$P(t_1) \neq P(t_2) \quad \text{if} \quad t_1 \neq t_2.$$

Furthermore, we require that $P(t_\nu)$ diverge as $|t_\nu| \to \infty$. We also require that any two distinct points belong to exactly one curve in Σ.

A flaw in the reasoning is the assumption that the curves are parametrized by the points in the plane. If $a \neq b$ and if $a_\nu \to a$ and $b_\nu \to b$ it is easy to prove that, for the curves, $L_\nu \to L$. Indeed the curves form a topological space which can be shown to be a two-dimensional manifold. But what sort of manifold is Σ? The curves of the system through a point cannot be contracted to a point. Thus Σ cannot be a plane. However, it can be proved that this space is topologically the projective plane with one point removed, just as it is in the simple case when all the curves are lines. We can then use the measure dL_p in the projective plane, and define the length

$$\lambda(A) = \int_{L \cap A \neq \emptyset} g(L)\, dL_p,$$

where $g(L)$ is an arbitrary continuous positive function, and the argument goes on as before.

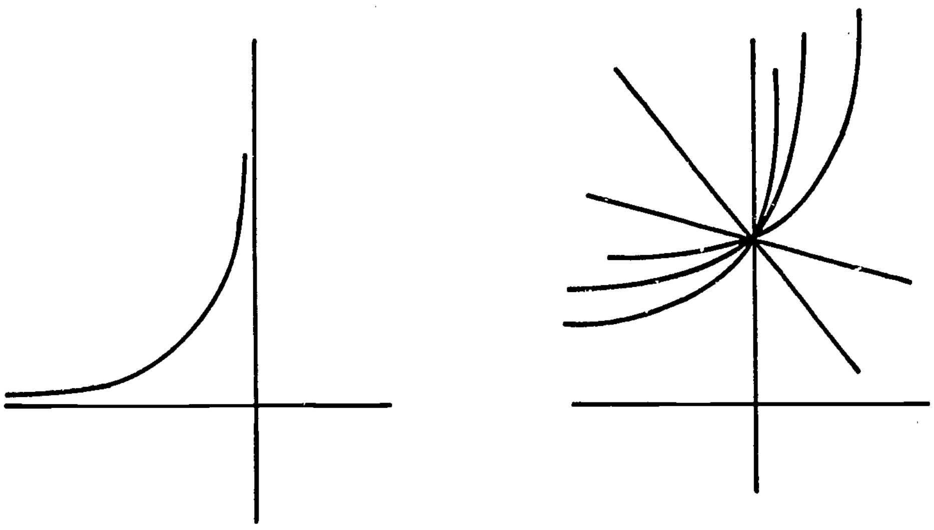
Having gone this far, we can answer some other questions. First, we can make every line of finite length if $g(L)$ is bounded. Can we make every line have infinite length in both directions? This can be done with some effort by adjusting $g(L)$.

We can consider a similar problem for curves in the projective plane. Here, the projective lines are homeomorphic to a circle, and any two distinct points determine a line. Suppose we consider in the projective plane P^2 a system Σ of curves in P^2 in which each curve is homeomorphic to a circle and for which any two distinct points lie on exactly one curve in the system. Is it then possible to find a metric in P^2 in which these curves are geodesics? (Since the projective line is not the shortest distance in the large, we can only expect a local result.) The method used to prove the result in the plane broke down in this case, and was saved with great effort by Skornyakov. However, this result is entirely trivial using the method suggested by integral geometry. Here we prove that the curve system is homeomorphic to the projective plane. We can thus use the projective density and weight it with any positive continuous function. What is the length of an entire curve? Because every curve must intersect it, the integral determining the length will be taken over the entire space of curves, and therefore the length has a constant value $2k$. This illustrates the general theorem that for any metrization of the projective plane the length of a line is a constant. An arc of a curve in Σ of length less than k strictly minimizes length.

Now assume that the system Σ has some mobility. That is, suppose there are some collineations. Of course, by a collineation in this case we mean a topological map of the underlying space of points (the plane or projective space) which sends the system Σ into itself. Is it then possible to find a metric which is invariant under all of these conditions?

In general the answer is no. For example, in the projective plane, it cannot be true of the projectivities, since a projectivity can send a segment into a proper subset of itself. However, it is possible to give a simple answer if the group is a compact group. In this case, choose one metric and then average it over the group. These considerations solve the problem for the projective plane. For if a compact space has a metric, then the group of all its metric preserving collineations must be compact. Therefore, for a system of curves Σ in the projective plane, it is necessary and sufficient that a group of collineations be contained in a compact collineation group in order for there to exist a metric invariant under these collineations.

We now give an example of a system of curves in the plane, which shows that there is no analogue for this result in the plane. The system will be invariant under all translations, but there will be no metric (for which the curves are geodesics) which is also invariant under translations. The system includes



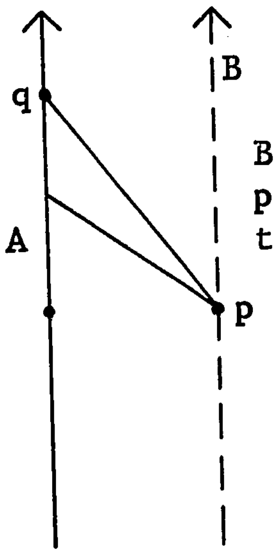
1) All lines with slope $\leqq 0$.

2) All lines whose equations are of the form $x = \text{constant}$.

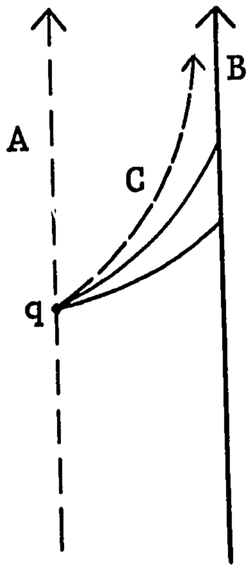
3) All translates of the branch of the hyperbola $xy = -1$, $-\infty < x < 0$.

It may be verified that this system has the property that any two distinct points of the plane lie on exactly one of these curves. But it is known that the only geometry invariant under all translations is a Minkowskian geometry, which has the straight lines as geodesics. Thus there is no translation invariant metric on the plane with this system of curves as its geodesics.

The example also illustrates a point in the theory of parallel lines which was considered here by Coxeter and which we shall need later. (We can also use the term asymptotes instead of parallels for the same idea.) It is a usual result, if $A \parallel B$, then $B \parallel A$. I should like to point out that this result depends on the mobility of the plane, that there is a sufficiently large group of motions on it. To find a parallel to a given line A in a given direction from a point p, we consider the lines pq, where q traverses A in one direction. Then the limiting line exists and is defined to be the



B is the parallel to A from p.



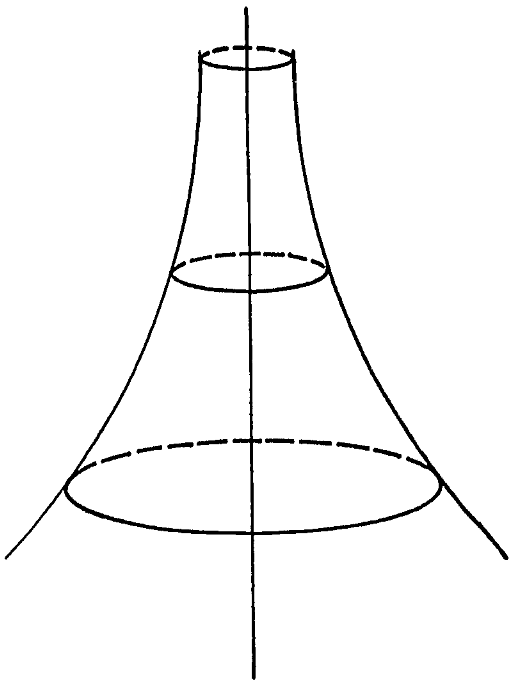
C is the parallel to B from q. A is not parallel to B.

parallel B to A through p. The diagram shows that this is not a symmetric relation in this geometry. The usual proofs from the axioms of hyperbolic geometry use mobility in some way.

If we consider the analogue of this problem on more general surfaces, we encounter a discrete group of motions. For example, consider a system of curves on a cylinder which, in the language of the calculus of variations, "has no conjugate points." This means that if we take the plane to be the universal covering space of the cylinder, then the system of curves in the plane over the given system on the cylinder is a curve system in the plane of the type we have been considering. After the cylinder is "unrolled" we have a curve system Σ in the plane. But we must obtain a metric in the plane which is invariant under the translations which correspond to the identity transformation of the cylinder. This is a discrete group, generated by one translation T. What is the general situation? Can one always find such a metric?

The answer is negative. The condition for the existence of a metric invariant under a discrete group of translations can be nicely stated, although there is not enough time to give the proof. We first define an *axis* of a translation T as a line mapped on itself by T.

Then, *if* T *does not have an axis*, the problem has a solution. For example, if we revolve a branch of a hyperbola about one of its asymptotes, we obtain a surface of negative curvature, which topologically is a cylinder. The rotation of this surface through 2π can be represented in the plane by a translation in the direction of the x-axis of length 2π. But an axis corresponding to this translation



is any of the orbits of a point on the hyperbola. These circles are not lines on the surface of negative curvature. Thus, for example, T has no axis.

But _if_ T _has an axis_, a necessary and sufficient condition that there is a metric invariant under T is that for any point P the lines L_n through P and T^nP should converge towards a line which is parallel (or asymptotic) to the axis of T.

This property may be illustrated nicely by referring to the example given above of lines and hyperbolas in the plane. We may easily verify that if a translation has a positive x-component and a positive y-component, there is no axis. For this case, therefore, there is a metric invariant under T. If the x-component is positive, but the y-component is negative, then there is an axis, but the above condition is satisfied. However, for a translation with no y-component or with no x-component, the above condition fails, and there will not exist a metric for this system of curves invariant under this translation.

Suppose, instead of a discrete group of translations generated by one translation, we have a continuous one parameter group of translations. In this case we can show that there is a metric invariant under this group which has the given family of curves in the plane as its geodesics if, for one of these translations not equal to the identity, the above criterion is satisfied. (This is clearly also a necessary condition.) The reason is that after we factor out this translation we are left with a compact group.

For my concluding example, I would like to consider a system Σ of curves on a surface of genus $p > 1$, and ask when such a system constitutes the geodesics on this surface without conjugate points. In this case, we can represent the universal covering space by the interior of the unit circle. We shall use the Poincaré model for hyperbolic geometry in order to obtain the appropriate theorem for the system Σ.

Suppose that the system Σ goes over into a system Σ' in the unit circle. Suppose that for Σ', *any two points determine a unique curve*. Just as certain translations of the plane correspond to the identity transformation of the cylinder and certain translations of the plane correspond to the identity transformation of the torus which it covers, so there are certain *covering transformations* of the hyperbolic geometry which correspond to the identity transformation of the given surface of genus p on which the curves of the system Σ lie. The system Σ' is invariant under the covering transformations. We postulate that *for each covering transformation which has an axis, the curves* L_n *determined by* p *and* $T^n p$ *must approach an asymptote to this axis*. We can prove then that each curve of the system meets the circle $|z| = 1$ in two distinct points. (This is to be interpreted in a limiting sense.) We then further assume that *a curve of* Σ' *is uniquely determined by its end-points*. With these assumptions, we can state that the given system is a system of geodesics for a suitable metric.

To see this, we note that there is a density for the lines of the hyperbolic plane which is invariant under all motions of this geometry. Since we are postulating that a curve of the system Σ' is determined by its end-points, we can transfer this density onto the curves of the system Σ'. This density can then be used to define length for curves in the system Σ'. It has the proper invariance and hence this length may be transferred onto the given surface of genus p. This solves the problem.

We conclude with the remark that the problem in the large for surfaces of genus p would be quite hopeless without this idea borrowed from integral geometry.

Discussion.

The question of generalizing this type of problem to space was considered. It was observed that in this case the situation is more complicated. It is known, for example, that it is not sufficient that two points uniquely determine a curve. For in the plane, curves separate the plane, and many arguments about crossings go through. But this is not true in space. In 3-space, Douglas gave necessary and sufficient conditions for a solution, locally, in the sense that the curves are extremals for a problem in the calculus of variations.

SOME COMPUTATIONAL ILLUSTRATIONS OF GEOMETRICAL PROPERTIES IN FUNCTIONAL ITERATION

Lecture by Glen Culler

(Lecture Notes by Melvin Hausner)

[Professor Culler's talk used, in an essential way, a keyboard connected to a digital computing machine, and a television screen which was used to display graphs, figures, and some of the instructions to the machine. There were many screens, so each member of the audience had a clear view. The resulting demonstration was most impressive and informative, but the format makes it impossible to do justice to the material by paraphrasing the talk. The following outline, then, makes no attempt to reproduce the lecture. Instead we summarize some aspects of the programming language and some of the problems covered in the lecture.

Finally, we append some photographs taken directly from the display screen. It should be pointed out that all of these examples were done on the spot. There was no preliminary programming, taping, prepared films, etc.]

1. The Machine.

The operator works on a double keyboard. The bottom half is essentially a single case typewriter keyboard, and it is used for the operands (numbers, functions, etc.). The upper keyboard is used for operators, and contains various mathematical operations such as addition, multipliication, cosine, logarithm, etc. Thus the cosine button will carry out the operation of taking the cosine of a number or list of numbers (stored in the machine) when touched. The operator keyboard has buttons corresponding to "Levels," labeled I, II,...,IX, and USER. The operators take on a different meaning according to which of these level buttons has been depressed. For example, on Level I each operand corresponds to a single number. On Level II each operand corresponds to a list of real numbers. Thus functions can be handled since a function is determined, as far as the machine is concerned, by a list of the values $f(x_i)$, $i = 1,2,\ldots,n$. On Level II the operators act on these lists. Level III corresponds to operations on complex lists. These lists may be regarded as representing polygonal paths in the plane.

For example, to operate with a single number, we use Level I. The problem of finding log (2·27), which was illustrated in the lecture, is done by first pressing the Level I button, then instructing the machine to Load 27, multiply by 2 and take the logarithm. This is accomplished by pressing:

$$\underline{\text{I}}\ \underline{\text{LOAD}}\ 27 \odot 2\ \underline{\text{LOG}}$$

(the underlined words are names of the operation buttons).

At each stage, the intermediate answer will appear on the screen if ordered to do so. The variables used are the letters A through Z which appear on the lower (operand) keyboard. Within the memory of the machine there are two data blocks designated as the X-register and Y-register. These provide storage for lists of X and Y coordinates and the contents of these registers may be selectively displayed either numerically or graphically. Figure 1 illustrates this display, where:

$$-1 \leqq X \leqq 1, \quad Y = X^2 - 1.$$

Most of the operators on Level II only transform the values in the Y-register, as appropriate for the composition of real functions. On Level III, the X and Y registers are taken together as a path in the complex plane. Both X and Y change with the operations.

The language is so set up that a function of a function is easily expressed. Thus complicated functions are easily built up out of the listed simple ones. For example, the graph of $Y = e^{-10X^2}$ can be found by first starting with the identity function $(Y = X)$, squaring, multiplying by -10, and exponentiating. All of this was done by pressing suitable buttons as follows:

II ID SQ ⊙ -10 EXP Display Return

The sequence of transformations is:

II = Prepare to do operations on lists of real numbers.

ID = Build N values from -1 to 1, equally spaced, store them in both the X and Y registers.

SQ = Square each number in the Y-register.

⊙ -10 = Multiply each number by -10.

EXP = Exponentiate each number.

Display Return = Display the present Y coordinates graphed against the present X coordinates.

The values are computed for $N \leqq 124$ numbers, which can be placed in the machine. The range of the ID operator is from -1 to 1, although this range may be easily changed by multiplying or adding constants. There is automatic scaling in the Y-axis, by successive powers of 2. At any time, the operator may find this power of 2 by "asking" the machine for it. Derivatives are found by forward differencing (there is a button, DIFF, which does this), and they may be found successively. Figure 2 shows what appears on the television screen for the successive differences of $Y = e^{-10X^2}$. In Figure 3, one can observe the effects of loss of accuracy due to the finite numerical representation of our numbers, and the use of successive differencing. Similarly, integration is obtained by summing, or by some variant, such as Simpson's rule.

2. Some Problems.

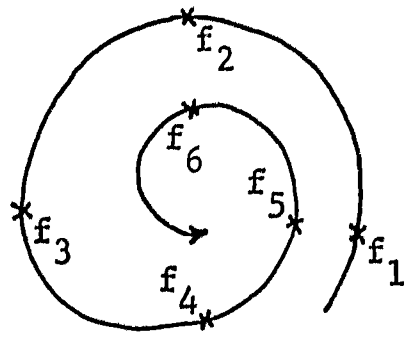
The following are some of the problems and techniques worked out during the lecture. In all cases, each step was or could have been seen on the screen. In fact, an important practical use of this procedure, other than pedagogy, is that the immediate visual solution to a problem, such as one involving successive approximations, shows the user if he is on the right track. Furthermore, this visual method often will suggest a better approach.

a. Find and plot the function and the successive derivatives of $y = e^{-10x^2}$ (Figure 2).

b. Using a smoothing procedure, reconstruct a function from one which is widely distorted. The smoothing procedure--applied repeatedly--replaced each y value y_i by the average of y_{i-2}, y_{i-1}, y_i, y_{i+1}, y_{i+2}. See Figures 4 and 5 for the results.

c. Find successive approximations to the differential equation $dy/dt = ty - 1$ using the Picard iteration scheme. Find the solution which is bounded at infinity by this method. In this example, the initial variation in the successive approximations suggested better initial guesses (see Figures 6 and 7).

d. Successive approximation is often very slow if the various approximations "spiral in" to the answer. This may be compared with a radial approach in which we may put $x_{n+1} = x_n + \lambda x_{n-1}$ for suitable λ, speeding up the process considerably. The figure illustrates the uselessness of this procedure for a spiralling approach. However, if after a fixed number of



of steps, we approximately repeat except for a constant factor, then a better approximation would seem to be the average of this number of successive approximants. In the figure, this number is four. In the problem presented, we were in function space. The integral equation

$$y(s) = g(s) + \int K(t-s)y(t)dt$$

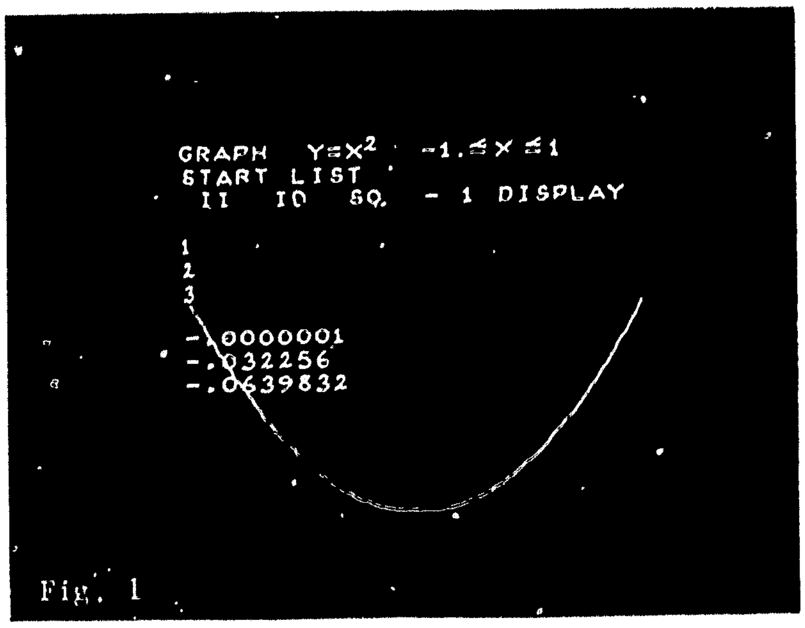
was to be solved by the iterative scheme

$$y_{n+1}(s) = g(s) + \int K(t-s)y_n(t)dt.$$

It was observed that after four steps, the new function has approximately the same shape as the original one. This averaging procedure was applied (again and again) speeding up the process. (See Figures 8, 9, 10.)

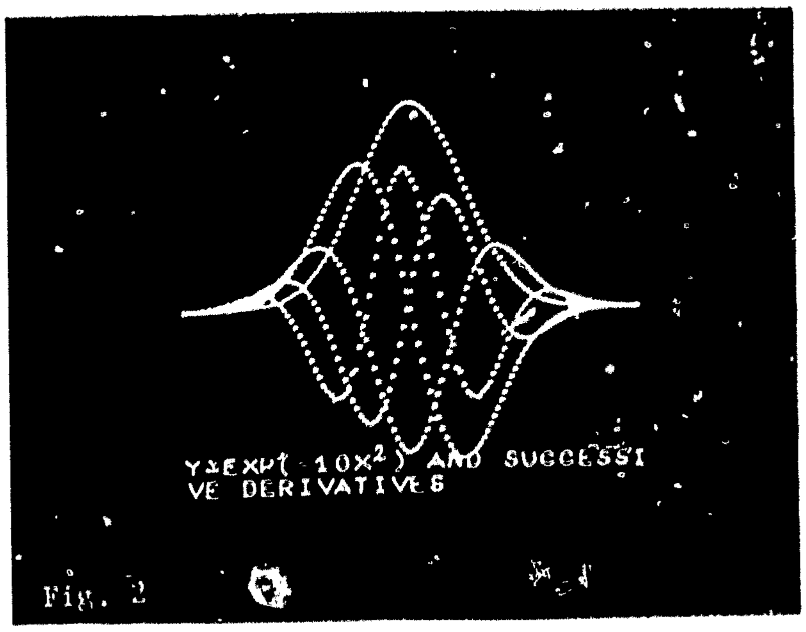
e. The effects of various conformal maps were illustrated. The maps $w = z^{1.5}, z^{1.5 \pm 0.2i}$ were illustrated by their effect on an ellipse. (See Figures 11 and 12.) Also, the effect of e^z on a rectangle in the upper half plane was illustrated, first for the simple case $0 \leqq x \leqq 1$, $0 \leqq y \leqq 2\pi$, and then after tilting this rectangle slightly by multiplying it by $0.9 + 0.1i$. (See Figures 13 and 14.)

In conclusion, the ease with which graphs and mappings could be constructed, stored, used, and visually presented was quite striking and impressive.



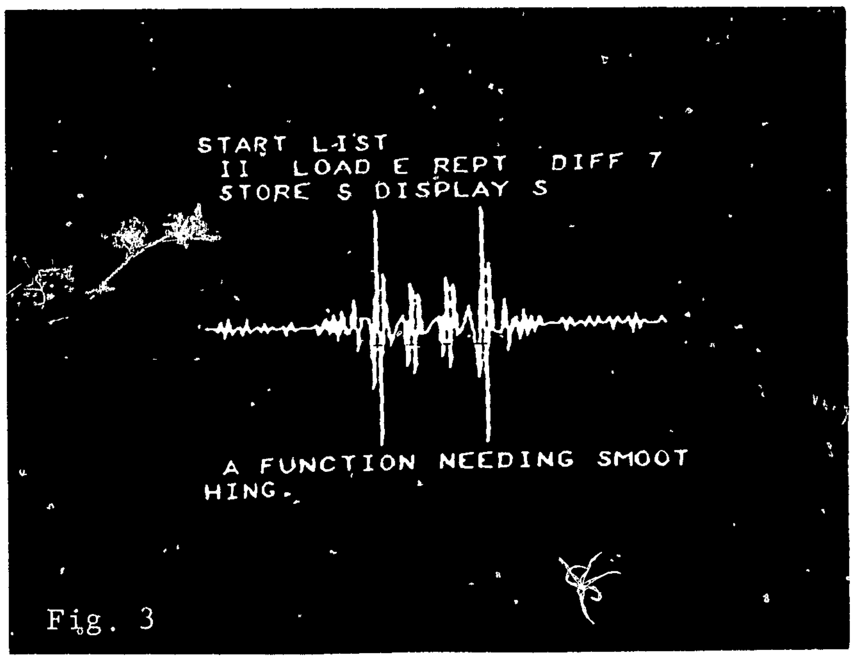
```
GRAPH  Y=X²   -1.≤X≤1
START LIST
 II   IO   SQ.   - 1 DISPLAY

1
2
3

-.0000001
-.032256
-.0639832
```

Fig. 1



```
Y=EXP(-10X²) AND SUCCESSI
VE DERIVATIVES
```

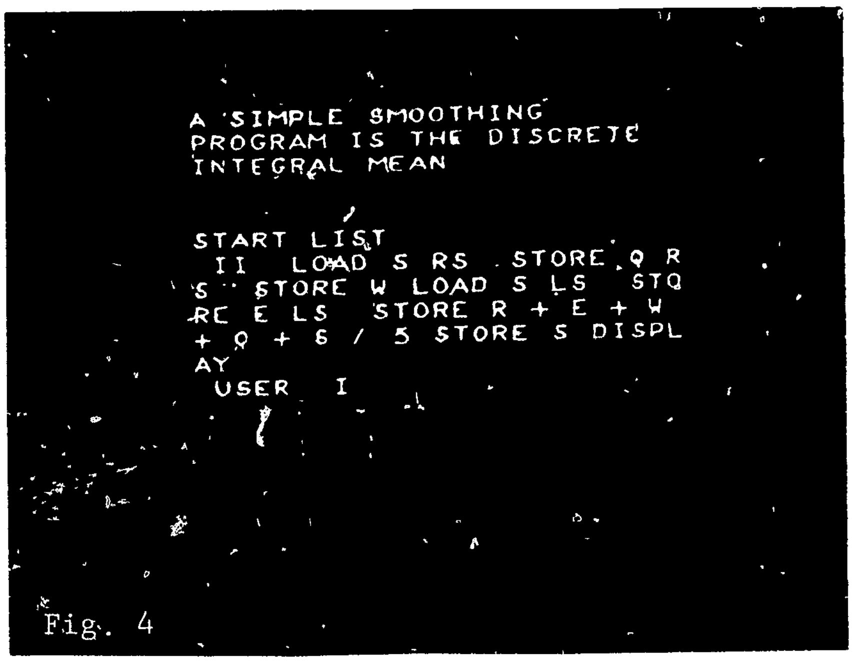
Fig. 2



```
START LIST
 II  LOAD E REPT  DIFF 7
 STORE S DISPLAY S
```

A FUNCTION NEEDING SMOOTHING.

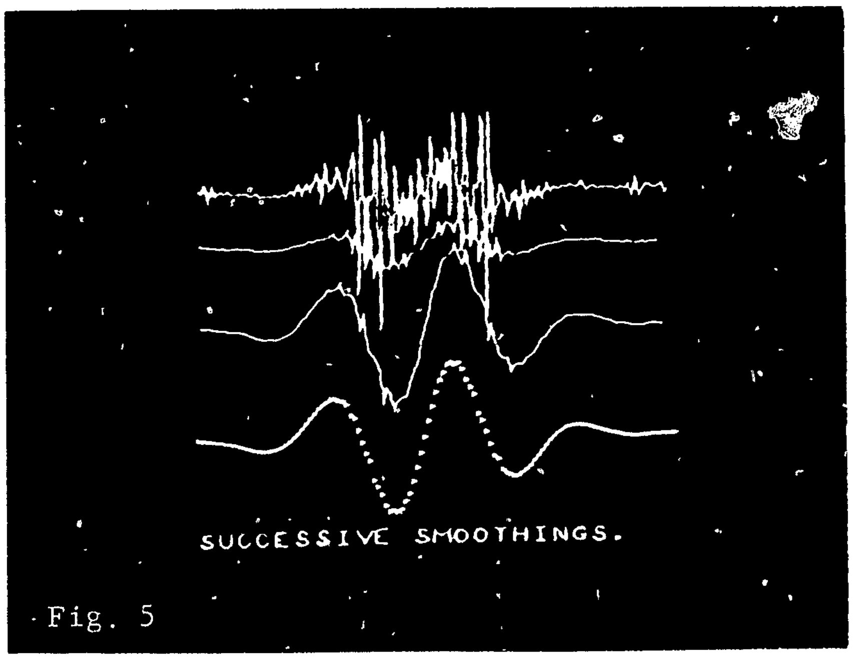
Fig. 3



A SIMPLE SMOOTHING PROGRAM IS THE DISCRETE INTEGRAL MEAN

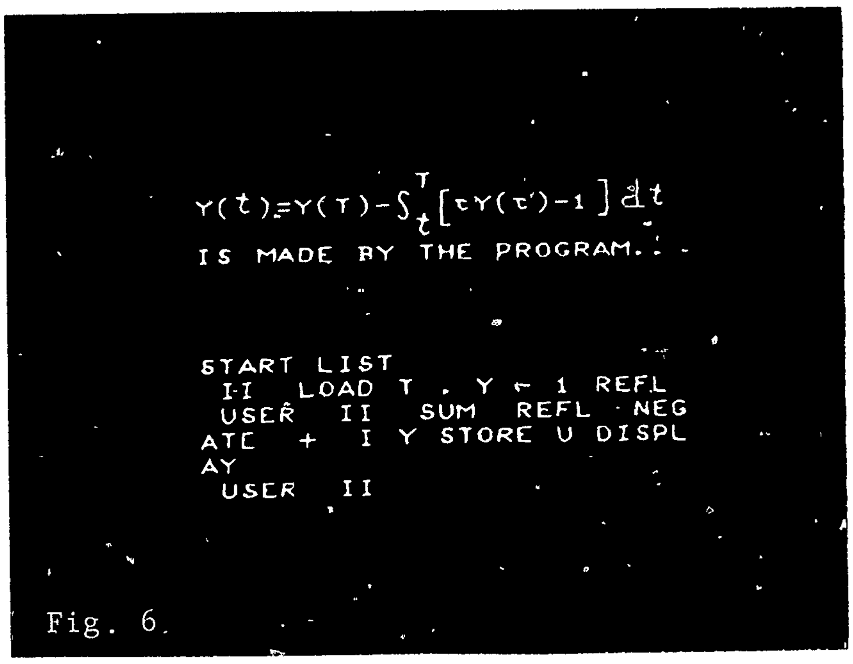
```
START LIST
 II  LOAD S RS  STORE Q R
S  STORE W LOAD S LS  STQ
RE E LS  STORE R + E + W
+ Q + S / 5 STORE S DISPL
AY
 USER  I
```

Fig. 4



SUCCESSIVE SMOOTHINGS.

Fig. 5

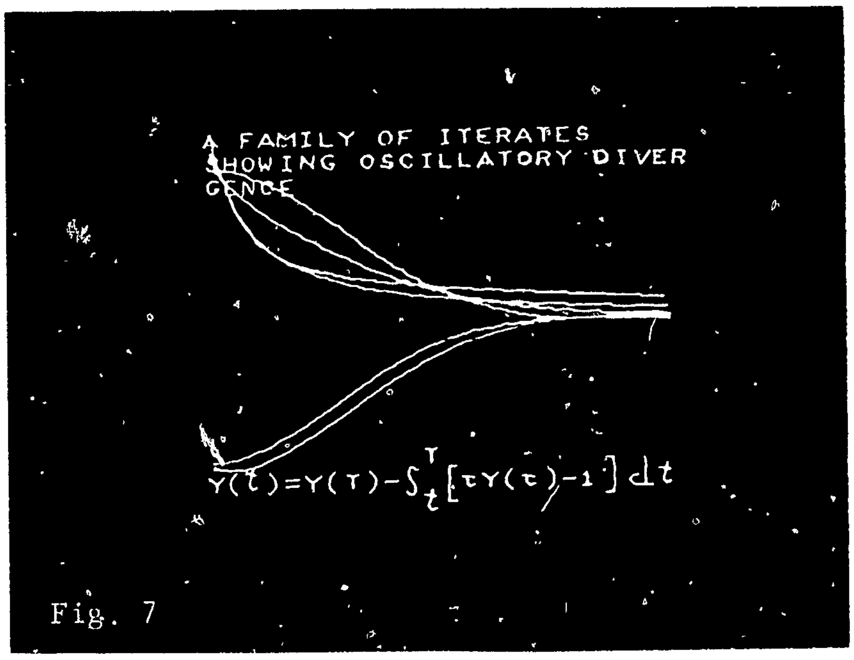


$Y(t) = Y(T) - \int_t^T [\tau Y(\tau) - 1]\, dt$

IS MADE BY THE PROGRAM...

```
START LIST
 I-I  LOAD T . Y ← 1 REFL
 USER  II  SUM  REFL  NEG
ATE  +  I Y STORE U DISPL
AY
 USER  II
```

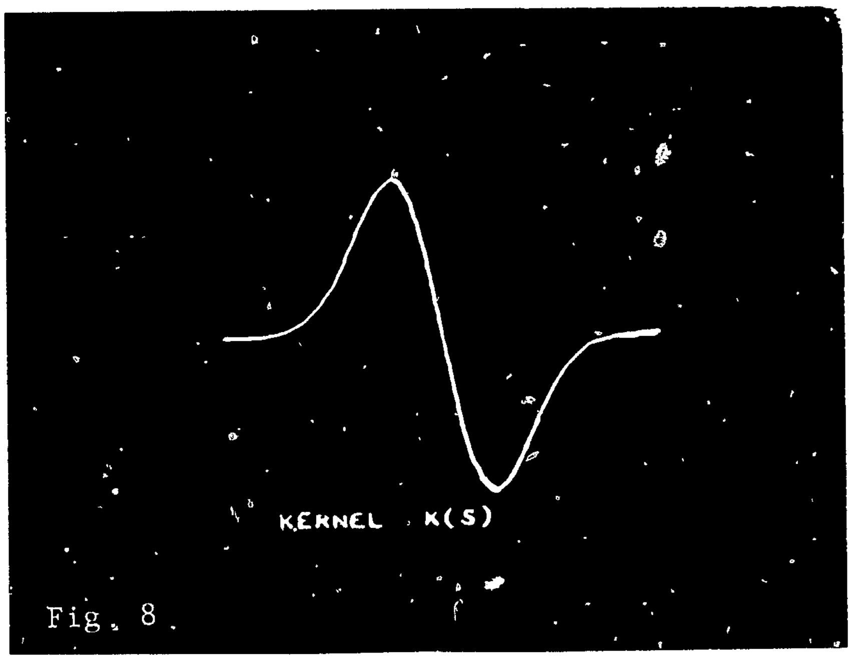
Fig. 6



A FAMILY OF ITERATES SHOWING OSCILLATORY DIVERGENCE

$Y(t) = Y(T) - \int_t^T [\tau Y(\tau) - 1]\, dt$

Fig. 7



KERNEL K(S)

Fig. 8

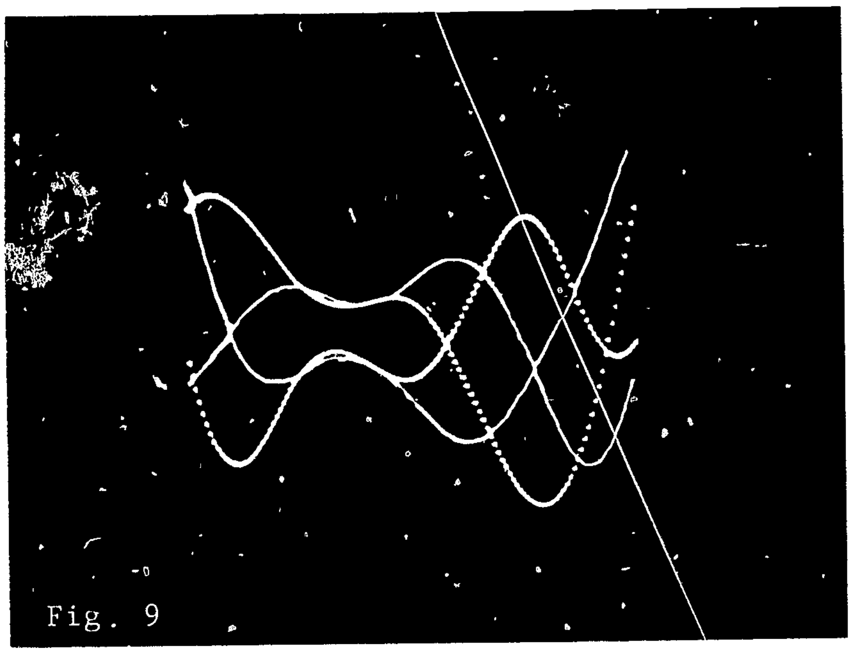


Fig. 9

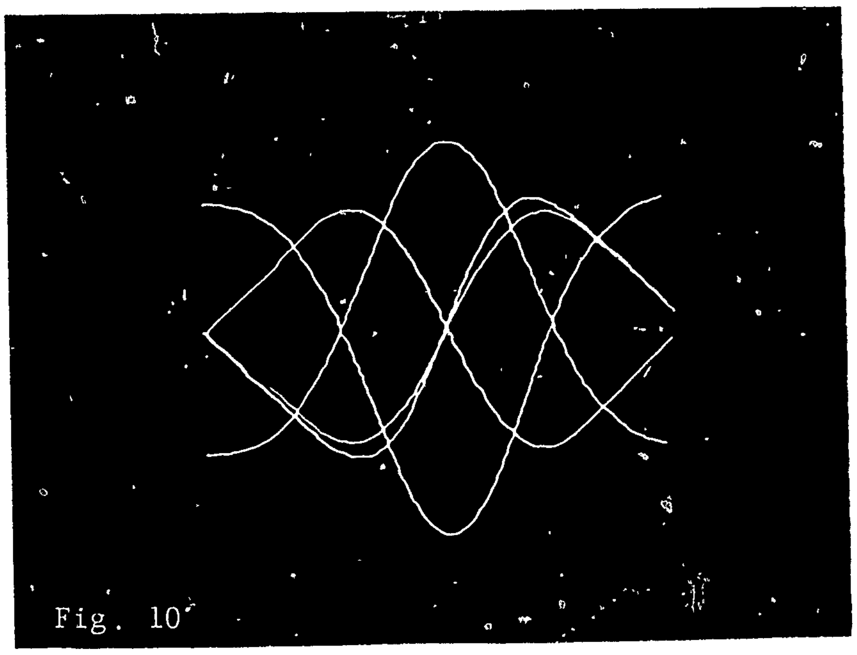


Fig. 10

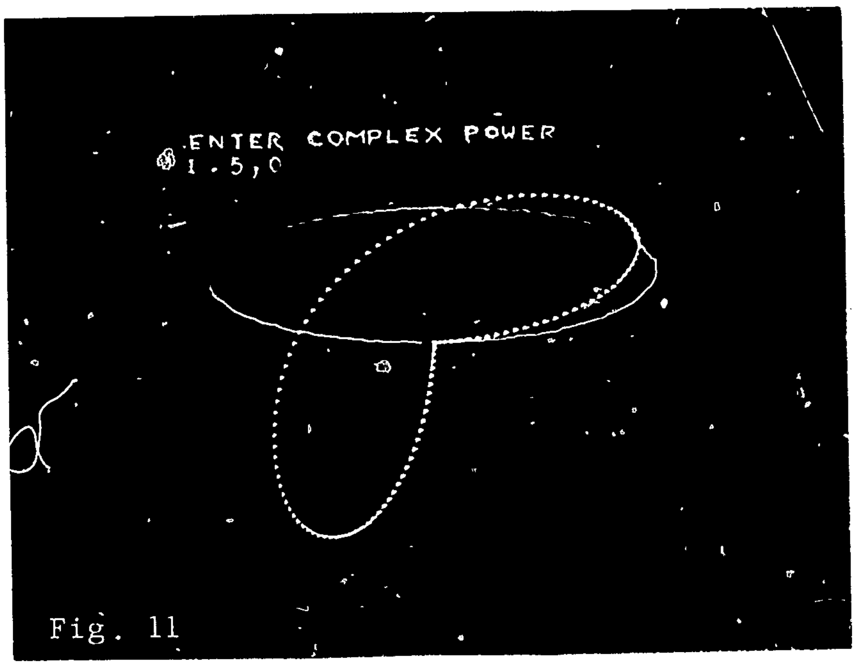


Fig. 11

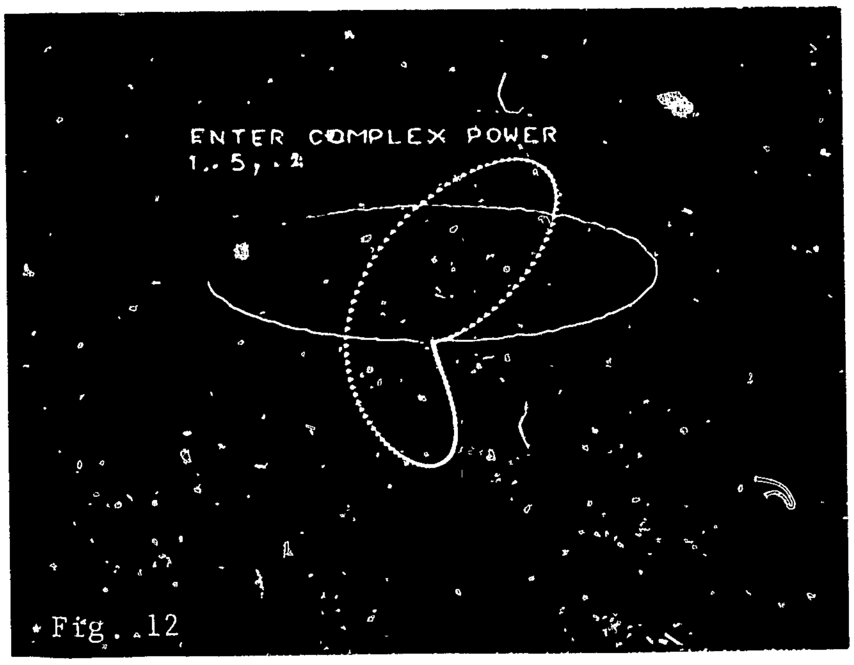


Fig. 12

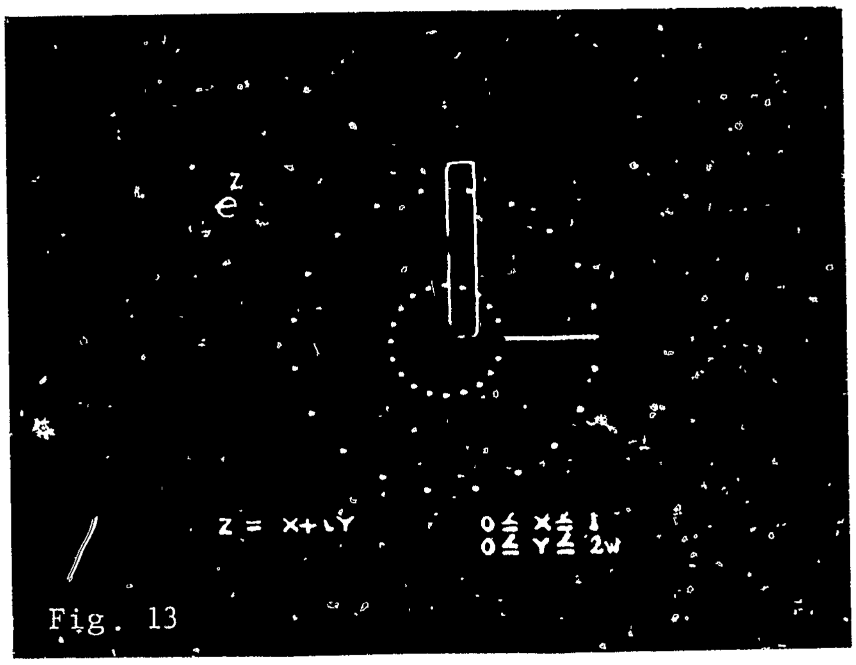


Fig. 13

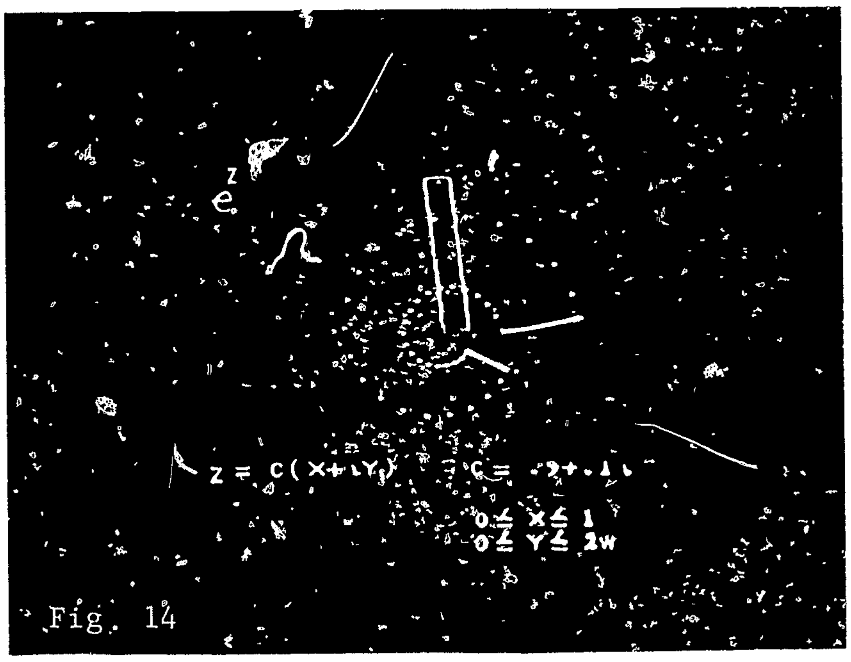


Fig. 14

GENERALIZATIONS IN GEOMETRY

Lectures by Preston C. Hammer

(Lecture notes by Melvin Hausner)

Lecture I.

My talk today is about certain general principles which I will apply to geometry. Specifically, I wish to talk about the topology of convexity, and to consider certain topological ideas which generalize to the theory of convexity.

We first note that an arbitrary intersection of convex sets is convex. Analogously, an arbitrary intersection of closed sets is closed. In topology, we have the operation of taking the closure of a set, while in the theory of convex sets, we have the operation (on sets) of taking the convex hull of a set. We are therefore led to consider mappings from subsets of a given set into the class of subsets of that given set.

Definition 1. Let E be a given set. The *power set* PE of E is defined to be the class of subsets of E. We let N denote the null set of E.

The two operations mentioned above are thus mappings f: PE → PE. We denote the closure map by u, and the convex hull map by h. Thus uX is the closure of X, while hX is the convex hull of X. We may now make the following general definition.

Definition 2. Let f: PE → PE. Then a set X is said to be *f-closed* provided $X \subseteq Y$ implies $X \subseteq fY$.

In this case, we may show that the arbitrary intersection of a class of f-closed sets is an f-closed set.

Note that in topology the u-closed sets are precisely the closed sets, while analogously, the h-closed sets of a vector space are precisely the convex sets.

We can also give properties of the convex hull operation h which correspond to similar properties of the closure operation u in topological space. These are stated below as axioms. We assume that h is a mapping of PE into PE.

Axiom 0. $hN = N$. (Recall that N is our symbol for the null set.)

Axiom 1. $hX \supseteq X$. (This is called the enlarging axiom.)

Axiom 2. $h(X \cup Y) \supseteq hX \cup hY$. (This is called the isotonic axiom.)

Axiom 3. $hhX \subseteq hX$. (This is called the subpotency axiom.)

Clearly Axioms 1 and 3 imply that h is idempotent: $h^2 = h$, or equivalently $hhX = hX$. We also note that Axiom 2 is equivalent to the property that h preserves inclusion: $X \subseteq Y$ implies $hX \subseteq hY$.

We note that if we add to Axiom 2 a sub-additivity axiom

$$h(X \cup Y) \subseteq hX \cup hY,$$

we obtain the usual Kuratowski axiom system for the closure operation. Nevertheless, the closeness of the axiom systems leads us to expect that the methods of one might be used in the other.

In general, we may define f-open sets as complements of f-closed sets. We use the notation cX for the complement of X. Thus, $c\colon PE \to PE$. We are thus led to the notion of a co-convex set. This is defined as a set whose complement is convex, or equivalently, as the complement of a convex set. Clearly the co-convex sets are the h-open sets.

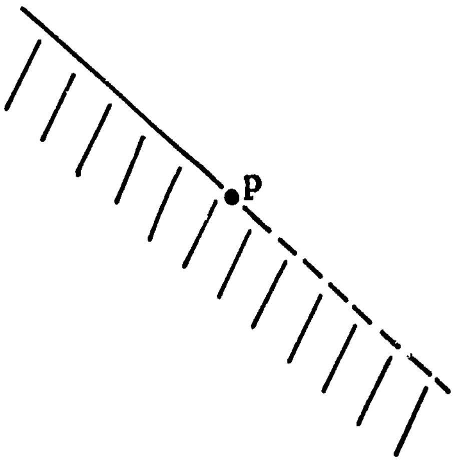
We may set up a concordance between topological ideas and ideas of convex set theory. The following is a partial listing:

Topology	Topology of Convexity
Closed set	Convex set
Open set	Co-convex set
Closure of X	Convex hull of X

Interior of X (cucX)	chcX
Isolated points	Extreme points
Dense in E	hX = E

We also note that $p \in hX$ if and only if every h-open set containing p meets X.

We now go to the Euclidean plane E^2 and consider neighborhood systems there. If $p \in E^2$, we wish to form a basis for the neighborhoods of p in the convex topology introduced above. Since we want small open sets which contain the point p, we may equivalently ask for large closed (i.e., convex) sets which exclude the point p. For the plane, we may actually find the *maximal* convex sets excluding the point p. These sets may be described as follows (see the figure): Consider any line L through p. Let S be the set consisting of one of the open half-planes determined by L, and one of the open rays in L determined by p. We call such a set a semispace S. The complement of such a space is called a co-semispace S^*. (This procedure is easily generalized to higher dimensions.)



Theorem 1. The semispaces at p are the maximal convex sets which exclude p. Thus, if X is any convex set not containing p, there exists a semispace S at p which contains X. Furthermore, no convex set can properly contain a semispace S at p and exclude p.

We shall prove the maximality of the semispace S at p. Suppose X is a convex set containing S properly. Let $q \in X$, but $q \notin S$. If we reflect the entire plane through p, it is easily seen that the plane is partitioned into S, the reflection S' of S and $\{p\}$. Since $q \notin S$, we have $q = p$ or $q \in S'$. If $q = p$, X does not exclude p and there is nothing to prove. Thus we may assume $q \in S'$. Therefore the reflection q' of q through p is in S, and hence in X. Since both q and q' are in X, it follows by convexity that their midpoint p is also in X. Thus X cannot exclude p.

To show that the semispaces are the only such maximal convex sets is not difficult. On the one hand, it may be proved as a consequence of the separation theorem. On the other hand, a direct proof is possible, and the separation theorem may be shown to be a consequence of this result.

Theorem 1 shows that a basis for the neighborhoods of p consists of all the co-semispaces S^* at p. We then have the usual topological result that $p \in hX$ if and only if $X \cap S^* \neq N$ for every co-semispace at p, a result with convexity implications. Similarly, the class of all semispaces is the minimum intersection basis for the class of all convex sets. (As usual, the empty intersection is taken to be the whole space.) Some other properties of semispaces are as follows.

1. Let S be any semispace at the origin. Then the class of all translates $\{p + S \mid p \in E^2\}$ is linearly ordered by inclusion. Furthermore, if $p \neq q$, then $p + S \neq q + S$.

2. Thus, the semispace S determines a <u>linear</u> ordering $\leqq$ of E^2, where $p \leqq q$ if and only if $q - p \in S$.

3. S is a <u>semigroup</u> with respect to vector addition. I.e., $p, q \in S$ implies $p + q \in S$. Moreover, $p \in S$ and r a positive number implies $rp \in S$. Thus S is a convex cone with the origin as a vertex. Moreover S is maximal among semigroups which exclude the origin.

4. A point p is an extreme point of a set X provided $p \in X$ and there exists a co-semispace at p which contains X.

Further Closure Functions. If we consider the plane as the Cartesian product of two lines, we may imitate topologists to obtain further results. On the line, the neighborhood base (in convexity) of a point p consists of the two closed rays emanating from p. Therefore it seems reasonable to take, as a system of neighborhoods of a point p of the plane, the four closed quadrants determined by forming the Cartesian product of each of these neighborhoods of the y-coordinate. To retain the rotational symmetry of the plane, let us stipulate that a neighborhood of a point p is any closed quadrant $\overline{Q}$ with vertex at p. Similarly, we may define neighborhoods $Q_{\frac{1}{2}}$ of p, as quadrants which are half-closed, and neighborhoods Q_0 of p as open quadrants (with the vertex p adjoined). We therefore have three different neighborhood systems for E^2. As in topology, we may define closure with the help of a neighborhood system. Thus, we make the following three definitions.

Definition 3. The f-closure of a set X is defined by the condition: $p \in fX$ if and only if $X \cap \overline{Q} \neq N$ for each closed quadrant $\overline{Q}$ with vertex at p.

Definition 4. The g-closure of a set X is defined by the condition: $p \in gX$ if and only if $X \cap Q_{\frac{1}{2}} \neq N$ for each half-closed quadrant $Q_{\frac{1}{2}}$ with vertex at p (including p).

Definition 5. The h-closure of a set X is defined by the condition: $p \in hX$ if and only if $X \cap Q_0 \neq N$ for each open quadrant with vertex at p (including p).

The above closure h is not to be confused with the previously considered h, which was the convex hull function.

All of these functions may be verified to be closure functions in the sense of our Axioms 0-3. Intuitively, the f-closure of a set X consists of all points from which it is impossible to view a full angle of $\pi/2$ of empty space (i.e., of cX).

We can now state and prove theorems analogous to the Carathéodory theorem in the plane. We first recall that theorem. If we let kX denote the usual convex closure of X, Carathéodory's theorem in the plane states that each point of kX is in some set kY where Y has three or fewer points and $Y \subseteq X$. If we let $|Y|$ denote the cardinality of Y, then the theorem of Carathéodory states that $kX = \bigcup\{kY \mid Y \subseteq X, |Y| \leqq 3\}$. Clearly the number three cannot be reduced.

<u>Theorem A</u>. If $p \in hX$, where h is the hull function based on the open quadrant neighborhoods, then there exists $Y \subseteq X, |Y| \leqq 8$ such that $p \in hY$. The number 8 is best possible.

<u>Proof</u>. If $p \in X$, the result is trivial. Now suppose $p \notin X$. Let A be a circle with center at p. We radially project X into A along the rays of p to obtain an image set Z. Clearly, $p \in hX$ if and only if $p \in hZ$. Also, if p is in the closure hY, with $Y \subseteq Z$, and $|Y| = n$, then (using the axiom of choice) we may find a set $Y_1 \subseteq X$ such that $p \in hY_1$ and $|Y_1| = n$. Therefore, with no loss in generality, we may assume that $X \subseteq A$, the circle centered at p.

For each point q on the circle A, let $A(q)$ be the open arc of A which subtends a 90° angle at p and which is "centered" at q. It is an easy matter to verify that $p \in hX$ if and only if $A = \bigcup\{A(q) \mid q \in X\}$. Thus, if $p \in hX$, the compact set A is covered by a union of open arcs $A(q)$ with $q \in X$. Hence it may be covered by a finite subset of such arcs. We let $\{A(q_1),\ldots,A(q_k)\}$ be such a set. We suppose with no loss in generality that

k is minimal, so that no set of $k-1$ arcs $A(q)$, with q in X, covers A.

For convenience, let us replace the points $q_1, \ldots, q_k$ on A by their angle coordinates $\theta_0, \ldots, \theta_{k-1}$, letting $\theta_0 = 0 = q_1$. Let $Y = \{\theta_0, \ldots, \theta_{k-1}\}$. By relabeling, if necessary, we may assume $\theta_i \leqq \theta_{i+1}$ and $\theta_{k-1} < 2\pi$. For convenience, we set $\theta_{k+i} = \theta_i + 2\pi$ $(i = 0,1,\ldots,k-1)$, and $\theta_{2k} = 4\pi$.

I claim that $\theta_{i+2} - \theta_i \geqq \frac{\pi}{2}$ for $i = 0,1,\ldots,2k-2$. To see this, note, for example, that if $\theta_2 - \theta_0 < \frac{\pi}{2}$, then θ_1 may be dropped from the set Y, retaining the covering property of Y, but reducing the number of elements in Y. This contradicts the minimality assumption on k. Therefore, we have

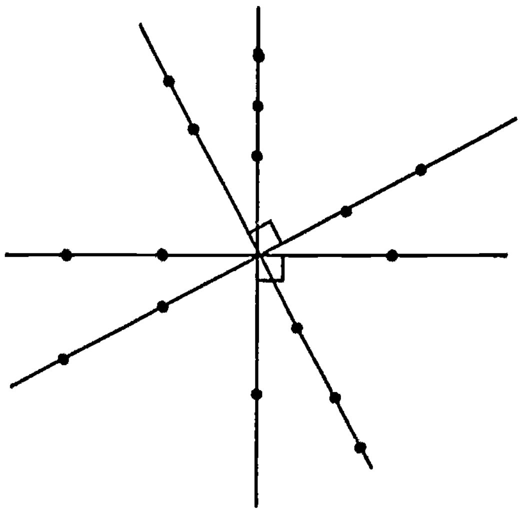
$$\sum_{i=0,2}^{2k-2} (\theta_{i+2} - \theta_i) \geqq \frac{k\pi}{2} .$$

(The notation means that i increases in steps of 2: $i = 0,2,\ldots,2k-2$.) But we also have

$$\sum_{i=0,2}^{2k-2} (\theta_{i+2} - \theta_i) = 4\pi .$$

Hence $\frac{k\pi}{2} \leqq 4\pi$. Therefore $|Y| = k \leqq 8$.

To show that 8 cannot be replaced by any smaller number, it suffices to give a set for which 8 points are necessary. The above proof shows that the only possible set of this kind is one which intersects each of eight rays through p, formed by two pairs of perpendicular lines, and which lies on these lines (as indicated in the figure). It is easily seen that for any such set, eight points is the minimum necessary to span p. In any other case, at most seven



points are required.

By refining the arguments above, we can show that the "Carathéodory number" for the hull function g based on half-open quadrants is 7. Thus, if $p \in gX$, there exists a subset Y of X with $p \in gY$ and $|Y| \leqq 7$. The number 7 is minimal. However, the Carathéodory number for the hull function f is $\aleph_0$. This may easily be seen by choosing the set X to be the half-open 270° arc with center at p. Clearly, no finite subset will contain p in its hull. The entire procedure can be generalized to sectors subtending an angle θ. The result is stated here.

Theorem B. Let $0 < \theta \leqq \pi$. Let f_θ, g_θ, h_θ be hull functions corresponding to closed, half-closed, and open sector neighborhoods. The neighborhood of a point p is taken to be a sector having angle θ and vertex at p. Then

$$h_\theta X = \bigcup\{h_\theta Y \mid Y \subseteq X, |Y| \leqq [4\pi/\theta]\},$$
$$g_\theta X = \bigcup\{g_\theta Y \mid Y \subseteq X, |Y| \leqq n = \text{maximum integer} < 4\pi/\theta\},$$
$$f_\theta X = \bigcup\{f_\theta Y \mid Y \subseteq X, |Y| \leqq \aleph_0\}.$$

In each formula, the bound on $|Y|$ cannot be decreased.

Further Generalizations. It is possible to generalize many of the above procedures. The objective is to find out more about why certain results in convexity hold and to use these insights for the development of systems applicable to a wider range of problems.

Suppose E is any set. Let $R \subseteq E \times E$ be a transitive relation on E. We define $u(p) = \{q \mid (p,q) \in R\}$. Then we may interpret $u(p)$ as a "cone" with vertex at p. We may also define an associated "closure" function $f: PE \to PE$ by the condition $p \in fX$ if and only if $X \cap u(p) \neq N$. (Again, N denotes the null set of E.)

Theorem. The function f defined above preserves inclusion and is subpotent (i.e., $ffX \subseteq fX$ for all $X \subseteq E$). Moreover, f is a closure function if and only if R is reflexive.

We omit the proof of this result.

Definition. A point $p \in X$ is a u-extreme point of X provided that either $X \cap u(p) = N$ or $q \in X \cap u(p)$ implies $(q,p) \in R$. Each transitive relation R generates a strict order relation $<$ defined to be the condition $p < q$ if and only if $(p,q) \in R$ but $(q,p) \notin R$. A set A is u-compact if each complete chain of elements of A has a maximum element.

We may generalize the results above by considering many "cones" at p. Let $T = \{R_i \mid i \in I\}$ be a collection of transitive relations in E, indexed by I. We define the sets $u_i(p)$ and the function f_i relative to the relation R_i. We then define a new function $f: PE \to PE$ by the formula $f = \cap f_i$. Equivalently, $p \in fX$ provided $X \cap u_i(p) \neq N$ for all $i \in I$. Thus $p \in fX$ if X meets all the "cones" associated with p.

Theorem. The function f associated with a collection T of transitive relations as defined above is isotonic and subpotent. f is a closure function if and only if each $R \in T$ is reflexive.

We omit the proof.

In what follows we shall formulate a few of the definitions and theorems (without proofs) concerning this general situation of a collection T of transitive relations on E indexed by I.

Definition. p is a T-extreme point of X if $p \in X$ and p is u_i-extreme point for some $i \in I$. For any $X \subseteq E$, we let vX be the set of extreme points of X.

Theorem. The function v defined above is shrinking $(vX \subseteq X)$ and idempotent $(vvX = vX)$.

Definition. A subset A of E is upper T-compact if A is u_i-compact for every $i \in I$.

Theorem. A necessary and sufficient condition that $fY = fX$ is that $Y \cap u_i(p) \neq N$ for all $i \in I$ if and only if $p \in fX$. A necessary and sufficient condition that $fY = fX$ when $Y \subseteq X$ is that $Y \cap u_i(p) \neq N$ for every $p \in fX$. In particular, $fvX = fX$ if and only if $vX \cap u_i(p) \neq N$ for all $i \in I$ and $p \in fX$.

I point out that I know of no theorem stated in convexity theory which is as powerful vis-a-vis extreme points as this theorem is when specialized to convexity theory.

Theorem. Let X be an upper T-compact subset of E. Then $fvX = fX$.

We conclude by giving a wide class of examples. Suppose $(E,\cdot)$ is a semigroup. If S is any sub-semigroup, we may define the relation $R = \{(p,q) \mid q \in pS\}$. It is then easy to verify that the associative law implies that R is transitive. However, transitive relations need not, in general, be of this type.

Discussion.

The theorem on extreme points was given explicitly for convex sets, as follows:

Theorem. Let X be a convex set, and let Z be the set of extreme points of S. Then X is the convex hull of Z if and only if $Z \cap S_p^* \neq N$ for every co-semispace about every point p of X.

It was pointed out that neither boundedness nor compactness was required. As an illustration, the interior of a parabola, together with a set of points dense on the boundary, is a set which is the convex closure of its extreme points.

REFERENCES

P. C. Hammer, Maximal convex sets. Duke Mathematical Journal 22(1955) 103-106.

____________, General topology, symmetry, and convexity. Transactions of the Wisconsin Academy of Sciences, Arts & Letters. 44(1955) 221-255.

____________, Semispaces and the topology of convexity. Proceedings of Symposia in Pure Mathematics, Volume 7 (Convexity), 1963.

____________, Isotonic spaces in convexity. Proceedings of the 1965 Copenhagen Colloquium on Convexity (W. Fenchel, Ed.), 1967.

Sister Gregory Michaud and P. C. Hammer, Extended topology: Transitive relations. Nieuw Archief voor Wiskunde (1967).

P. C. Hammer, Extended Topology: Carathéodory's theorem on convex hulls. Rendiconti del Circolo Matematico di Palermo 14(1965) 34-42.

Lecture II.

Today we shall confine ourselves to the study of connectedness and its generalizations. Historically, connectedness was used for a long time before a formal definition was given. Probably the geometers needed it first. In analysis arc-wise connectedness was a useful notion. G. Cantor once proposed a definition which was not adopted. (His definition would make a set connected if its closure is connected in the current topological sense.) If we use the notation fX for the closure of X, and N for the null set of a space M, then the current definition (due to Riesz, according to Thron's book) is as follows.

Definition. The sets X and Y are said to be separated provided

$$fX \cap Y = N = X \cap fY.$$

A set Z is connected if it is not the union of a separated pair of non-empty sets. Equivalently, Z is connected provided that if $Z \subseteq X \cup Y$, where X and Y are separated, then $Z \subseteq X$ or $Z \subseteq Y$.

If we introduce the binary relation R (on the power set of M), defined by

$$R = \{(X,Y) \mid X \text{ and } Y \text{ are separated}\},$$

we can rephrase the definition in a way which is capable of being suitably generalized. Thus Z is connected provided that if $Z \subseteq X \cup Y$ and $(X,Y) \in R$, then $Z \subseteq X$ or $Z \subseteq Y$. We note that if R is any relation on PM, the null set and the singleton sets are necessarily connected in this sense.

We now let M be a space, and let N denote its null set. The power set PM is the class of all subsets of M. Let R be a relation on PM, i.e., R is a subclass of $PM \times PM$.

Definition 1. R is a *separation* provided $(X,Y) \in R$, $X_1 \subseteq X$, $Y_1 \subseteq Y$ imply $(X_1,Y_1) \in R$. R is *exclusive* (or *disjunctive*) provided $(X,Y) \in R$ implies $X \cap Y = N$. R is *symmetric* provided $(X,Y) \in R$ implies $(Y,X) \in R$. Finally, R is a *Wallace separation* provided R is a symmetric and exclusive separation.

Clearly, the usual topological separation is an example of a Wallace separation. Cantor's idea was to use the (Wallace) separation $R_C = \{(X,Y) \mid fX \cap fY = N\}$.

Definition 2. Let R be a Wallace separation. The set Z is *R-connected* provided that if $Z \subseteq X \cup Y$, $(X,Y) \subseteq R$, then $Z \subseteq X$ or $Z \subseteq Y$.

We remark that we can replace all inclusions by equalities, since we can always consider $Z \cap (X \cup Y)$, $Z \cap X$, and $Z \cap Y$. However, the above definition is more useful.

We now note that the null set N and the singleton sets are necessarily R-connected for every Wallace separation. If f is the Kuratowski closure of a topological space, then the separation

$$R(f) = \{(X,Y) \mid fX \cap Y = N = X \cap fY\}$$

is called a topological separation. It is a Wallace separation and it yields the usual connected sets in the sense of topology. We also remark, in passing, that the "major" theorems about connectedness hold in the more general setting. At any rate, it is surprising what does hold. Let us start with an example.

Example. Let M be a metric space, for example the plane. Let $t > 0$ and let d denote the distance function on M. Define

$$R_t = \{(X,Y) \mid p \in X,\ q \in Y \text{ imply } d(p,q) \geqq t\}.$$

Then it is easily verified that R_t is a Wallace separation. Now it is a theorem, which we leave as an exercise, that a set A is R_t-connected if and only if for every pair of points p and q in A, there exists a sequence of points $a_1, \ldots, a_k$ in A such that $a_1 = p$, $a_k = q$, and $d(a_i, a_{i+1}) < t$ for $i = 1, \ldots, k-1$. This is surely a reasonable form of connectedness. We note that the unique decomposition into maximal R_t-connected components is an immediate consequence of this notion. Also, using this example, we may easily construct an irreducibly connected finite set joining two points.

A. D. Wallace, after whom these separations were named, discussed separations but not connectedness in 1941. Following Wallace, we introduce the function W_t by the formula

$$W_t(X) = \{p \mid (p,X) \notin R_t\}.$$

Intuitively, we take all points p close to (i.e., not separated from) X. (We have identified p with its singleton set.) In this example,

$$W_t(X) = \cup\{S(p,t) \mid p \in X\}$$

where $S(p,t)$ is the open disk of radius t centered at p. Thus $W_t(X)$ is always open. We have the following properties.

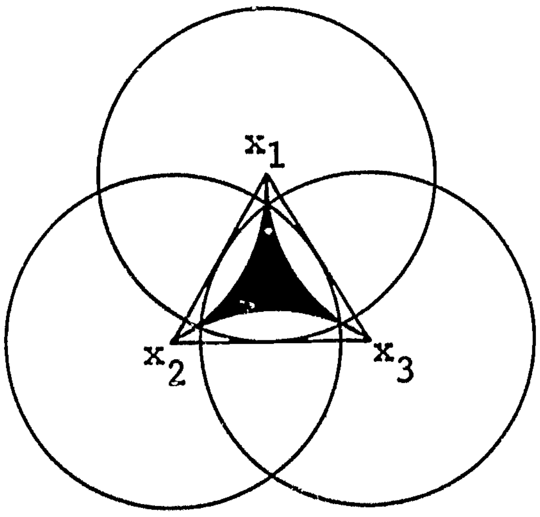
0. $W_t N = N$
1. $W_t X \supseteq X$
2. $W_t(X \cup Y) = W_t(X) \cup W_t(Y)$

This latter result may be generalized to arbitrary unions. We note that W_t is not idempotent, and it is therefore not a closure function. Wallace showed that for T_1-spaces, W_t would have to be a closure function if R_t were a topological separation. Thus, this separation is not a topological separation.

We now introduce the conjugate Wallace function W_t^*.

Definition 3. $W_t^*X = \cap\{cW_tY \mid (X,Y) \in R_t\}$.

We recall that we are using the notation cX for the complement of X. We may also describe W_t^*X as the set of all points which are separated from all sets Y which are separated from X. It can also be shown that $W_t^* = cW_tcW_t$ (i.e., $W_t^*X = cW_tcW_tX$ for every set X). It may also be shown that W_t^* is a closure function in the sense of my first talk. We illustrate W_t^* in the accompanying diagram. The set X is taken to be the vertices of an equilateral triangle of side s. Suppose $t < s$ but t is near s. Then W_tX is the union of the three open disks centered at the points of X. The set W_t^*X is the shaded region. We can make several assertions about W_t^*. First, W_t^*X is always a closed set containing the usual closure of X. In the plane, $W_t^*X = X$ if X is closed and convex. $W_0^* = \bigcap_{t>0} W_t^*$ is the usual (topological) closure function.



General Theory. We now return to a general Wallace separation R. We let $\mathcal{C}(R)$ be the class of R-connected sets.

Theorem A. If R is a Wallace separation, there exists a unique maximal Wallace separation R* such that $\mathcal{C}(R^*) = \mathcal{C}(R)$.

Thus there is an equivalence relation among Wallace separations: Namely, two Wallace separations are equivalent if they lead to the same class of connected sets. The theorem asserts that within each equivalence class, there is a unique maximal separation. In order to prove this theorem, we first prove the following lemma, of interest in itself. It states that we may make any class of sets connected, although the process will introduce other connected sets in a minimal way. Note that maximizing the separated sets will minimize the connected sets.

Lemma. Let $\mathcal{C}_0$ be a subclass of PM. Then there exists a unique Wallace separation R such that $\mathcal{C}(R) \supseteq \mathcal{C}_0$ which is maximal with respect to this property.

Proof. Let $R = \{(X,Y) \mid X \cap Y = N,$ and if $X \cup Y \supseteq A \in \mathcal{C}_0,$ then $A \subseteq X$ or $A \subseteq Y\}$. Then clearly $\mathcal{C}(R) \supseteq \mathcal{C}_0$. Furthermore, no larger separation will do. Q.E.D.

In Theorem A, replace $\mathcal{C}_0$ by $\mathcal{C}(R)$, apply the lemma, and we have the result.

We now define the notion of R-connectedness for classes of sets.

Definition 4. A subclass $\mathcal{B}$ of PM is *R-connected* if $\mathcal{B}$ is not the union of two disjoint subclasses $\mathcal{B}_0$ and $\mathcal{B}_1$ such that the union sets $X_i = \bigcup_{X \in \mathcal{B}_i} X$ are non-empty and R-separated.

We then have the following result.

Theorem B. A necessary and sufficient condition that the union of a class of connected sets be connected is that the class be connected.

Theorem B gives all of the component decomposition theorems. For example, if a class of R-connected sets has a point in common, its union is connected.

We now consider maps preserving connectedness. We point out that in the topological case, continuous maps preserve connectedness, but it is also true that wildly discontinuous maps may also preserve connectedness.

Theorem C. Let M and M_1 be spaces with Wallace separation R and S respectively. Let R* be the unique maximal separation in M which yields the same connected sets as R (Theorem A). Then a necessary and sufficient condition that $t: M \to M_1$ preserve connectedness is that, if X and Y are subsets of M such that $(X,Y) \notin R^*$, then $(tX,tY) \notin S$.

It is possible to define the functions W and W* for any Wallace separation R, by using the same formal definition as in the example. In analogy with the theorem that the closure of a connected set is connected, we have the following result.

Theorem D. If X is R-connected, then WX and W*X are R-connected.

In this case, since W and W* are not necessarily idempotent, the various iterates may yield more connected sets.

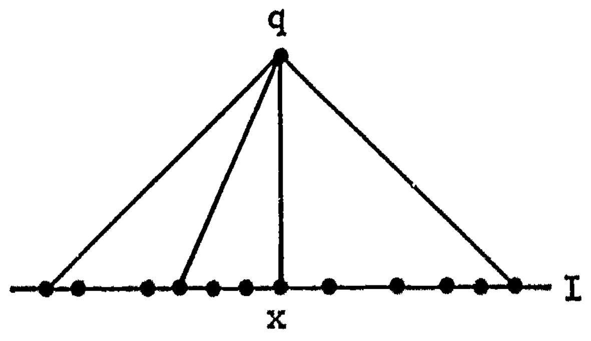
We conclude with an example due to Mrowka.

Theorem. In the plane $M = E^2$, no topology yields as its class of connected sets the class of arc-wise connected sets.

For the proof, we need only assume that we have a Fréchet space. This means that the expansive function f satisfies the two conditions $fX \supseteq X$ and $f(X \cup Y) \supseteq fX \cup fY$. The separation Rf is defined by

$$Rf = \{(X,Y) \mid fX \cap Y = N = X \cap fY\}.$$

We assume that there is such an expansive function f which yields the arc-wise connected sets as all of its Rf-connected sets. We let I be a line segment in the plane, and choose q not on the line determined by I. We split I into dense, complementary subsets of I: $I = B \cup C$, where $B \cap C = N$, and B and C are dense in I. Define



$A_1 = \bigcup_{q \in B} \overline{qx}$, where $\overline{qx}$ is the closed line segment joining q and x. Then A_1 is arc-wise connected, since any two of its points may be connected by a polygonal path through q. Thus A_1 is Rf-connected. If we choose a point $p \in C$, then $\{p\} \cup A_1$ is not arc-wise connected. Thus $\{p\} \cup A_1$ is not Rf-connected, and $\{p\}$ and A_1 are separated. Then $p \notin fA_1$. Since p is arbitrary, $C \cap fA_1 = N$. But $A_1 \supseteq B$, since each $x \in B$ is the end-point of the segment $\overline{qx}$. Hence $fA_1 \supseteq fB$. It follows that $C \cap fB = N$. But in the same way, we may show that $fC \cap B = N$. It follows that B and C are a separated pair of sets, the union of which is I, a line segment. This is the contradiction, since I is arc-wise connected, hence Rf-connected.

Remark. This example shows that arc-wise connected sets in E^2 cannot be exactly the connected sets for any neighborhood space. Yet, if in the Lemma used to prove Theorem A, $\mathcal{C}_0$ is the class of all arcs in E^2, then the minimum class $\mathcal{C}(R)$ of connected sets containing $\mathcal{C}_0$ is the class of arc-wise connected sets. This shows that connectedness should not be restricted by the topological definition even for the purposes of topology!

REFERENCES

P. C. Hammer, General topology, symmetry, and convexity. Transactions of the Wisconsin Academy of Sciences, Arts & Letters.. 44(1956) 221-255.

__________, Extended topology: The Wallace functions of a separation. Nieuw Archief voor Wiskunde (3) 9(1961) 74-86.

__________, Extended topology: Connected sets and Wallace separations. Portugaliae Mathematica. 22(1963) 167-187.

THE NATURE AND IMPORTANCE OF ELEMENTARY GEOMETRY IN A MODERN EDUCATION

Lectures by Paul J. Kelly

Lecture I (Synopsis).

It is my belief that we have so far failed, and failed badly, to exploit the educational potential of geometry in the basic high school course. We have failed our own self interest in the training of future mathematicians and we have failed to transmit to our children the general cultural values inherent in geometry. The early work of national study groups was strikingly successful in improving the vocabulary and the axiom systems used in the teaching of elementary geometry. But following this major achievement, most of the proposals and sample materials have been timid and unimaginative.

Among the many causes for our failures, two seem to me to be primary. The first is our failure to look at the subject of Euclidean geometry as a whole and to see it as a well identified, viable mathematical structure with modern developments of its own and with interconnections to other parts of modern mathematics. This failure causes all discussions of the high school course to begin from a restricted viewpoint. Instead of starting with a consideration of what is valuable and accessible in the subject matter, discussions start with praise or damnation of the traditional course. Either contention is almost invariably supported by arguments that emanate from a bias and reveal stereotyped thinking in which there is no reexamination of old attitudes.

The second major source of our failures is our inability to see the pedagogical possibilities implicit in the history of geometry in relation to the history of mathematics. The idea of a mathematical system goes back to Euclid, and man's understanding of the independent character of such systems originated with the discovery of non-Euclidean geometries. The work of Eudoxus certainly motivated the search for a satisfactory theory of real numbers. The problem of finding tangents to a curve was obviously one of the origins of calculus. The path from locus problems to analysis situs to topology is a clear one. The influence of geometric notions in the development of the Erlanger program is well known.

The lesson of history is plain. Geometry has time and again been the source of ideas from which entire mathematical disciplines have grown. This has been so because of man's visual mindedness and his propensity to first come upon important ideas in a graphic and intuitive setting. It is surely then a a rather natural idea to see in this seminal character of geometry a marvelous educational opportunity. Why do not we use geometry to introduce students at an early age to a wide range of vital ideas and important processes? This is the role that elementary Euclidean geometry can play incomparably better than any other subject. It is here that the defense of geometry becomes rational and powerful, needing no sentimental appeal to tradition, and justifying its existence by the standards of contemporary values.

Working with Norman Ladd, I have written a high school geometry textbook in the spirit just described. I am firmly convinced that the attitude toward geometry in this book, the pedagogical objectives, and the new content introduced represent natural directions for significant improvement in high school geometry and high school mathematics. I cannot judge, of course, how well the intentions of the book were actually implemented. In any case, it is what we intended to do that I believe merits serious consideration, whether or not we achieved those intentions. In sketching this program, I ask you to bear in mind that I am talking about a tenth grade course and that our concern was with the education of all young people and not just with the prospective mathematics major.

The unifying theme of the book is that Euclidean geometry is not only a mathematical system but a wonderful prototype of a scientific system. From this viewpoint, its importance to society is obvious. Whereas one may speak about the applications of algebra or functions, geometry is literally the matrix in which such subjects as classical physics, engineering, and architecture are inescapably imbedded.

The systems concept initiates the book's intention to use geometry as an introduction to ideas and processes of major importance. The construction of intellectual models to study various aspects of physical reality is fundamental in modern science and technology. In illustrating this, geometry has several advantages. One is the fund of knowledge about physical geometry that the student has already acquired. He knows a great deal that the model must succeed in representing. Moreover, in checking that the model is successful, in most cases he only needs to make reasonably accurate diagrams. He does not need a laboratory.

Stressing the natural relation of Euclidean geometry to physical geometry may be repugnant to certain purists. But the relation of the two subjects in no way contradicts the independence of the mathematical system nor denies the logical criterion it must satisfy. On the other hand, the relation of the two subjects explains the motivation for the way in which the mathematical system is constructed. Also, it keeps the mathematics in a context useful to the student who is not mathematically oriented. Finally, I venture to suggest that those relations most interesting in physical geometry usually correspond to the more interesting parts of the mathematics.

One natural consequence of the "model" point of view is that the book deals with Euclidean space from the outset for the obvious reason that physical space is not a plane. There is, of course, a natural sequence in the development of linear and planar relations, but the lines and planes are always in space and not the space itself.

In the initial stages of building the system, no attempt is made to "do" foundations. What is stressed is the building process itself, and the need for axioms, definitions, conventions, and notations. Aside from a few sample proofs, nearly all the foundation relations are stated as starred, unproved theorems. Thus they are axioms that the student knows can be proved, and this treatment allows one to proceed quickly. In particular, the system appears as a man-made structure, in which there are many arbitrary choices, and is not seen as something fixed and God-given.

With the foundations established there is a traditional development of plane and solid geometry following the themes of congruence, parallelism, and similarity. The student is involved in reading and writing proofs and solving problems. Many different points of logic are discussed as they appear in context.

At the half way point of the book, the principal structural relations of Euclidean three space have been established together with a faily extensive mathematical vocabulary. The student has taken part in the building process and has learned a good deal about proof writing in connection with the properties of elementary figures such as polygons and circles. However, all of this is clearly only a beginning.

If the system that has been built is to be taken at all seriously as an instrument for investigating physical geometry, then it obviously has to contain more mathematical objects than circles and polygons. A natural extension of planar objects is found in the class of plane convex figures. In finding these, the student sees the creative side of definition. A physical figure cut out from a flat piece of cardboard corresponds to a mathematical set that is planar, non-linear, and bounded (contained in the interior of some circle). Restricting the physical "cut-out," we can use convexity to express the "oneness" or connectedness of the corresponding mathematical set. Finally, as a physical figure the cut-out must contain its edge points or boundary points. The mathematical characteristic of such a boundary point P is its nearness to both the set S and the complement of S. Thus we can define P to be a boundary point to the planar set S if the interior of every circle at P intersects both S and the complement of S. Thus we finally arrive at the class of plane, convex figures, the sets that are planar and non-linear, bounded and convex, and that contain their boundary points. The boundaries of these figures now give us the extensive, and accurately defined, class of simple, closed convex curves.

Starting from a wholly natural objective, and using a few ideas of a topological character, a whole new vista has been opened to the student. Moreover, he has seen this result from a change of viewpoint. In the early work, polygons and circles were defined and then their interiors were discovered, so to speak. Now it is regions that are defined and their boundaries that are discovered. From this new point of view, all the former figures are simply special types of closed, convex curves. The former interiors now become the set of inner points to the plane convex figure.

In the rich variety of ideas in the theory of convex figures--diameters, lines of support, widths and curves of constant width, etc.--the student sees the spirit of mathematical generalization at work. Moreover, the natural extension of these concepts to space provides a beautiful example of analogy and the efficiency of good definitions. With the sphere replacing the circle, one finds at once how to define convex solids and obtains from their boundaries the extensive, and accurately defined, class of simple, closed, convex surfaces. In particular, the traditional surfaces of solid geometry, spheres, cones, cylinders, and polyhedra, now appear in a natural setting as particular convex surfaces.

After establishing a more realistic class of geometric objects, the next step is again a natural one. In the earlier study of elementary figures, the two principal notions employed were congruence and similarity. Our mass production society, if nothing else, shows that both the concepts of congruence and similarity are applicable to quite arbitrary objects. It is natural then to seek a mathematical generalization of these concepts that can be applied to our new class of curves and surfaces.

Starting with congruence, a clue to the desired generalization can be found in the physical motion of an object from one position to another. Each point of the object in the initial position has a natural association with its corresponding location within the object at the second position. We can express such associations by mathematical mappings. Once mappings, and the combining of mappings have been defined, one sees that the mappings that will generalize congruence are those that preserve distance, the isometries. Any set is, by definition, congruent to its image in an isometry. One easily establishes that reflections in points, lines and planes, rotations and translations are isometries.

The similarity mappings that change all distances by the same positive factor provide the desired generalization of similarity, since any set may be defined to be similar to its image in a similitude. It suffices to establish the dilations at a point as similitudes since one sees that the product of a motion and a similitude is a similitude.

In this mathematically simple theory, the student again sees the rich harvest of generalization. Not only do the mappings provide a natural way for expressing symmetries of general objects, they provide powerful tools for solving problems. Moreover, it is not difficult for him to see that all kinds of mappings are possible and that all sorts of invariants could be studied. He has, in fact, seen one of the tap roots of modern mathematics.

The final parts of the book deal with area, volume, and circular arc properties. Here there is the chance to show how the new concepts and methods may be used in the development of traditional theory.

The program just sketched is practical in being evolutionary rather than revolutionary. Though the spirit of the book, the variety of objectives, and some of the content, are not traditional, these changes are imbedded in a familiar framework. Much of the material is traditional and the book is deeply indebted to the early work of SMSG and other study groups.

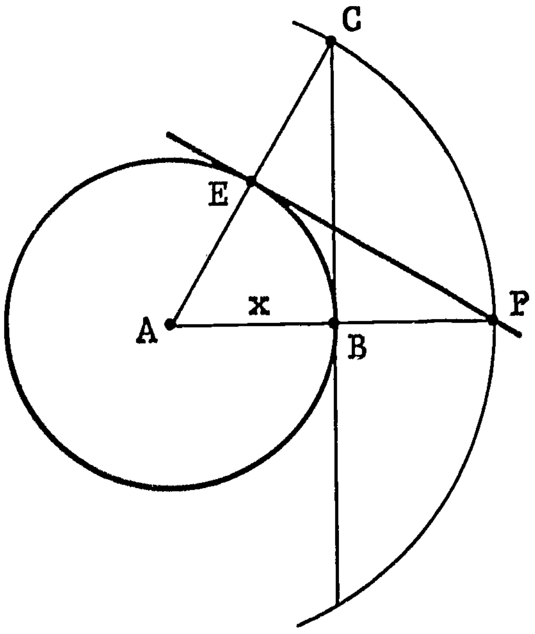
What seems to me of paramount importance is that the kind of course I have sketched introduces ideas and processes of vital importance in contemporary mathematics. It does so in a completely natural way and the course is primarily--and properly so at the tenth grade--a revelation to the student of a whole series of mathematical vistas. It is an introduction to mathematical and scientific systems, to mathematical proof and disproof, to the processes of analogy and generalization, and to the creative aspect of making definitions. It initiates topological ideas, notions related to functions and groups, and introduces the deeply fruitful concept of a transformation and the invariants of a transformation. It provides a setting within which many unsolved problems can be stated and new ones even conjectured by exceptionally good students. It pictures mathematics as a man-made subject and one which is dynamic and constantly expanding.

Not only is such a course a natural evolution at this time, it provides for future evolution in directions that are clearly desirable. When teachers have acquired familiarity with simple topological notions and transformations, a future shift of emphasis can give such concepts a more central role and a more extensive development than is practical at this time.

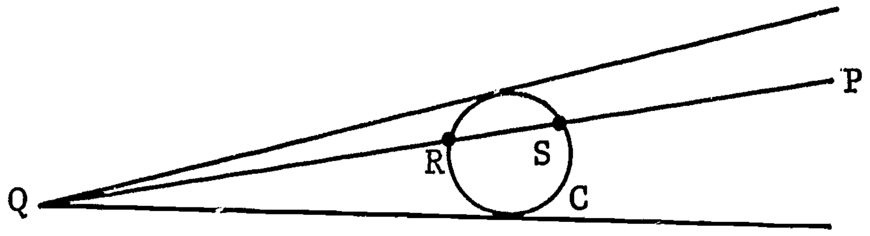
Lecture II (Excerpts) Notes by Paul Yale.

[In his second lecture Kelly presented many examples illustrating what tenth graders can do with the new material he believes should be included in a high school geometry course. The following excerpts give the flavor of the talk.]

The following example illustrates the use of reflections in a simple construction problem, the problem of constructing a tangent to the circle C(A,x) from the point P. The construction is: 1. Construct the perpendicular to the line PA through B, the point of intersection of C(A,x) and the segment PA. 2. Let C be one of the intersections of this perpendicular with the circle centered at A and through P. 3. Let E be the intersection of C(A,x) with the segment AC. The proof that the desired tangent is the line PE is obtained by reflecting the line BC in the angle bisector of angle PAC. This reflection leaves C(A,x) invariant and interchanges P and C as well as B and E. Since tangency is invariant under reflections and BC is a tangent to C(A,x) through C, the line EP is a tangent to C(A,x) through P. A nice feature of this construction is that it is also valid for constructing tangents to circles in hyperbolic geometry where you do not have the property that the angle inscribed in a semicircle is a right angle.



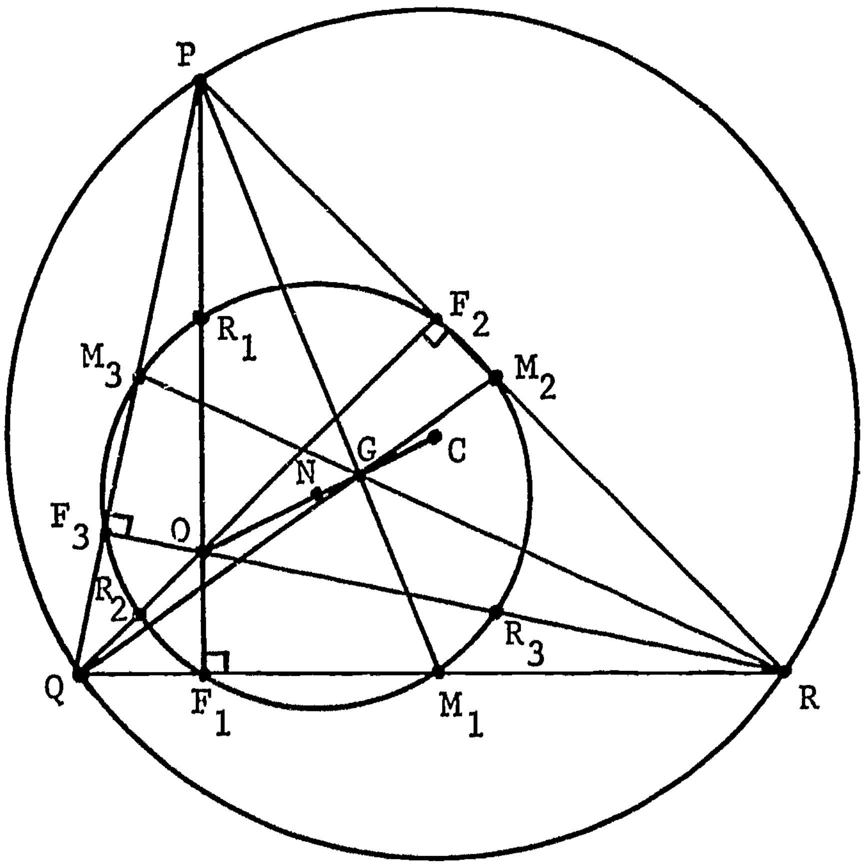
The point to the next example is the fact that the solution to an old, familiar problem often jumps out at the student when he is familiar with mapping techniques. The problem is to construct the circles through an interior point P of an angle and tangent to both sides of the angle. If C is any circle tangent to both sides of the angle, then clearly a dilation at the angle vertex Q can move C to a position where it contains P. If the two points of intersection of circle C and the line QP are R and S, then the two dilations with center Q sending R to P and S to P map the circle C to the two desired circles, since tangency is invariant under similitudes. Moreover, we see that these two circles are the only possibilities, so we have a complete solution.



The following problem illustrates nicelv the economy of mapping techniques. Consider a parallelepiped with A and B two opposite vertices and let M be the midpoint of A and B. Let C be a vertex adjacent to B and let D be the vertex opposite C. The reflection in M sends rays to oppositely directed rays and preserves distances, so it must send C to D. But it is an interchange mapping hence it also sends D back to C. The midpoint of two interchanged points must be the center of reflection, hence the diagonal DC also contains M. The same argument is applicable to the other diagonals, so all four diagonals meet at M; moreover, we have the added information that M is the center of the parallelepiped and that the solid angles at opposite vertices are congruent.

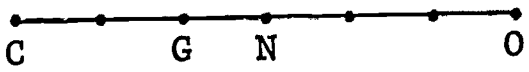
Consider now the problem of constructing an equilateral triangle with its three vertices on each of three given parallel lines, r, s, and t. Since the translates along r of any one solution are also solutions, we can pick the first vertex P on r arbitrarily. Let s' be the image of the line s under a rotation through 60° centered at P. Since s' and t are not parallel we can let Q be their intersection and let $\overline{Q}$ be the image of Q when we rotate back 60°. Since s' must come back to s, $\overline{Q}$ is on s, and the triangle $PQ\overline{Q}$ is easily shown to be the desired triangle. The simplicity of this construction is, I think, a nice example of elegance in geometry.

The classical problem of the nine point circle and the Euler line illustrates how mappings not only establish the basic result with economy, but often give extra information. Given a triangle PQR, consider the three feet, F_i,



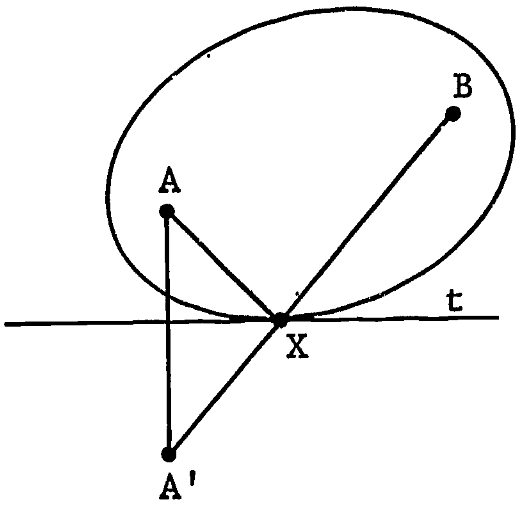
of the altitudes, the three midpoints, M_i, of the sides, and the three midpoints, R_i, of the segments joining the orthocenter O to the three vertices.

The nine points F_i, M_i, and R_i can easily be shown to be on a circle, called the nine point circle. Let N be the center of this circle. We want to explore the relations between N, O, the circumcenter C, and the centroid G. First we dilate with ratio 2 from the point O. This sends the three points R_i to the vertices and hence sends the nine point circle to the circumcircle. Thus it must send N to C and we see that O, N, and C are collinear with N the midpoint of O and C. Next we consider the reflection in G followed by a dilation with ratio 2 and center G. This product of mappings sends M_1, M_2, M_3 via M_1', M_2', M_3' to the three vertices and hence sends the nine point circle to the circumcircle. Thus the same product of mappings sends N to C and shows that G, N, and C are collinear. But with no extra effort we also see that G is one third of the way from N to C since the reflection in G sends N to the midpoint of G and C. Combining this information with that above, we see that the four points are in the order



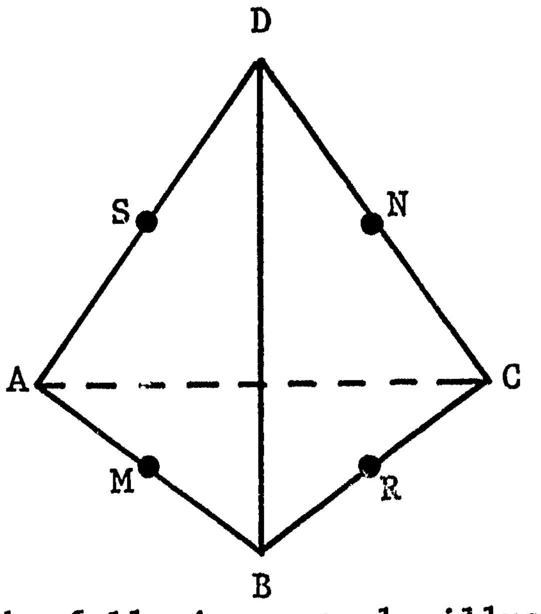
on the Euler line, and that if any two coincide they all do.

I think the following example illustrates nicely the way one can exploit a simple problem to get a more complicated result and that it illustrates techniques that we will see more of in the future. If A and B are on the same side of the line t, then we obtain the minimal path from A to B via t in the usual way by taking X to be the intersection of the line t and the line A'B, where A' is the reflection of A in the line t. This is a strict minimum, so if we let C be the ellipse through X, with foci at A and B, then X is the only point on both t and C, and therefore t is tangent to the ellipse. Because X is fixed under reflection in t, while A maps to A', the tangent t is the bisector of angle AXA'. Thus we obtain



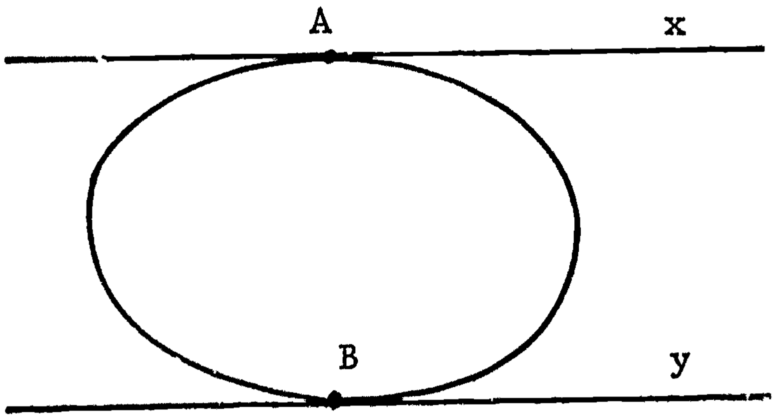
the standard focal property of an ellipse with a nice economy.

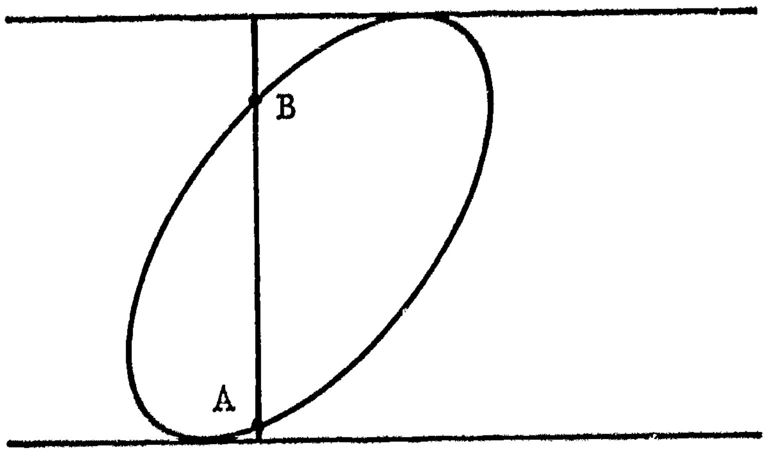
There are many examples illustrating how this new kind of material can be used as an exploratory tool. I have chosen just one, which shows how the student can discover something that is not intuitively obvious. Let A, B, C, and D be the four vertices of a regular tetrahedron. What are the axes of symmetry? Assume there is one. Since a reflection is an interchange mapping we see that a reflection cannot leave the tetrahedron invariant and have either one or three fixed points among the four vertices. It cannot leave all four vertices fixed and, since an edge is obviously not an axis of symmetry, neither can it leave two vertices fixed. Thus the four vertices are interchanged in pairs, say A ↔ B and C ↔ D. The midpoint of two points interchanged by a reflection is on the axis of reflection, so we see that the only candidates for axes of symmetry are the three lines joining midpoints of opposite edges. One easily shows these are axes of symmetry. Since it interchanges A and B and also C and D, the reflection in line MN must interchange R(the midpoint of BC) and S(the midpoint of AD). Therefore, the line RS is perpendicular to MN. Applying the same argument to



the third axis, we see that the three axes of symmetry are mutually perpendicular, a result that was not intuitively evident when we began our exploration.

The following example illustrates a typical kind of convexity problem that students can do. Given two parallel lines of support, x and y, for a convex curve, each of the lines must contain a contact point, say A and B as shown. Obviously, d(A,B) ≧ d(x,y). Since there is always a chord as long as any width, the diameter is at least as large as the maximum width. On the other hand, given A and B in the convex curve, there are parallel lines of support perpendicular to AB. These lines cannot be between A and B, hence the width is at least d(A,B). Therefore, the maximum width is at least as large as the diameter. From the two inequalities it follows that the diameter equals its maximum width. By following a similar line of reasoning, the students can show easily that the contact





points for parallel support lines of maximum width are unique and that the line joining them is perpendicular to the support lines. Two immediate consequences are that curves of constant width contain no segments and that circles are the only curves of constant width with a center. These ideas generalize of course to convex solids in space. Also, by first proving that if I is between A and B, then

$$d(X,I) \leqq \max(d(X,A), d(X,B))$$

for any point X in space, it is easily shown that the maximum width of a convex polygon or a polyhedron is the distance between some two vertices.

The new material in this course provides a rich supply of exploratory problems. The following are a few examples of problems that can be settled by a counter example. 1. A closed convex curve containing a segment has a corner. 2. Two closed convex curves cannot cross each other more than four times. 3. A linear set with infinitely many centers is a line. 4. The altitudes of a tetrahedron are concurrent. 5. An isometry without fixed points is a translation. 6. If for all $X \in R$ the distance from X to the set S is constant, then for all $Y \in S$ the distance from Y to R is constant.

Finally, let me repeat one of my major points. Although I believe that convex figures should be included in the high school curriculum and hope that future recommendations will encourage this, I am even more concerned that transformations be included. The structure of the motion group of a geometric space is essential information about the space. Transformations are easy to teach and vital to ideas that are alive today. If mathematicians said flatly that motions and similitudes are an integral part of Euclidean geometry and must be taught at the high school level, then the high school teachers would learn them. They *can* do so easily. I hope we will insist that they must do so.

JOINING AND EXTENDING AS GEOMETRIC OPERATIONS
A coordinate-free approach to n-space

Lecture by Walter Prenowitz

(Lecture notes by Melvin Hausner)

There is a widespread belief among mathematicians that the only effective way to study n-dimensional geometry is to assume the existence of a coordinate system and use coordinate methods from the start. This is very easy to understand historically. In the early development of n-dimensional geometry mathematicians studied the subject by means of analytic models and naturally used whatever procedures were at hand, algebraic and analytic, to get results. Since, in addition, the synthetic methods of classical 2- and 3-dimensional geometry do not generalize in a simple or obvious way, the belief is rarely challenged and mathematicians are not encouraged to try to discover intrinsic, coordinate-free methods for studying n-space.

This situation has two consequences that I think are harmful. First it fosters the belief that n-dimensional geometry can't be studied efficiently on an autonomous basis but can only make progress with the extrinsic support of algebra. Second it tends to inhibit studies that cover broadly many geometric systems, since the initial choice of coordinate representation restricts the geometry studied to one particular type, usually affine or Euclidean geometry.

I want to argue that the conventional view is not necessarily correct and show specifically that much of linear geometry can be treated by coordinate-free methods.

I take as my starting point the operation of joining two points to form a segment. In classical geometry we often come across the phrase "join point a to point b to form a segment." But it is the segment ab which is studied, not the joining operation. The emphasis falls on the noun "segment" rather than the verb "to join." We shall systematically study join as an operation on pairs of points which is extendible to n points.

In what follows, we shall use lower case letters $a, b, \ldots$ to denote points, and upper case letters $A, B, \ldots$ to denote sets of points. For convenience, we shall often ignore the distinction between a point a and its singleton set $\{a\}$. For example, we shall write $a \subset A$ when no confusion can occur.

Before proceeding to the formal development, we describe informally the significance of the theory in Euclidean geometry. To every ordered pair (a,b) is associated the open segment $\overline{ab}$ whose endpoints are a and b. We call $\overline{ab}$ the join of a and b and denote it $a \cdot b$ or simply ab. Implicitly, $a \neq b$ in this case. However, we certainly want the operation to be defined for all choices of a and b. Therefore, if $a = b$, we define $ab = aa = a$.

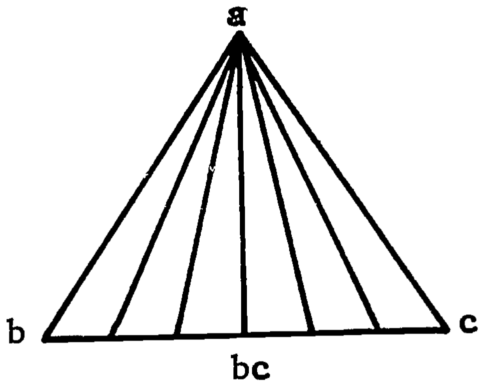


Figure 1 (a)

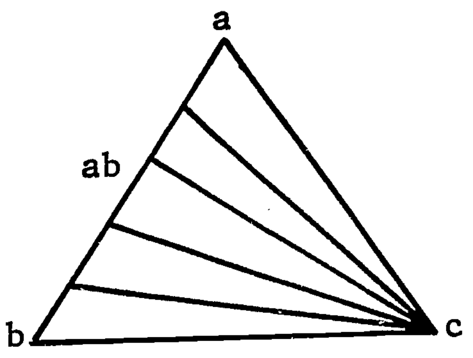


Figure 1 (b)

How would we form the join of three points a, b, c? It is natural to form bc, then join a to the individual points of bc and take the union of the joins formed to obtain the join of a, b, and c. In Figure 1(a), we can see that this operation produces the interior of the triangle abc. It is reasonable to use the notation a(bc) to denote the set of points which we obtain in this way. We can just as well form ab, then join the individual points of ab to c and take the union of the joins formed. The result, which may be denoted (ab)c, also is the interior of triangle abc (Figure 1(b)). Thus we may write a(bc) = (ab)c = abc.

We can easily extend the operation to four points. Suppose a, b, c, d are in general position. We form bcd which is the interior of triangle bcd, then join a to each point of bcd and take the union of the joins formed. The result may be denoted a(bcd) and turns out to be the interior of tetrahedron abcd (Figure 2).

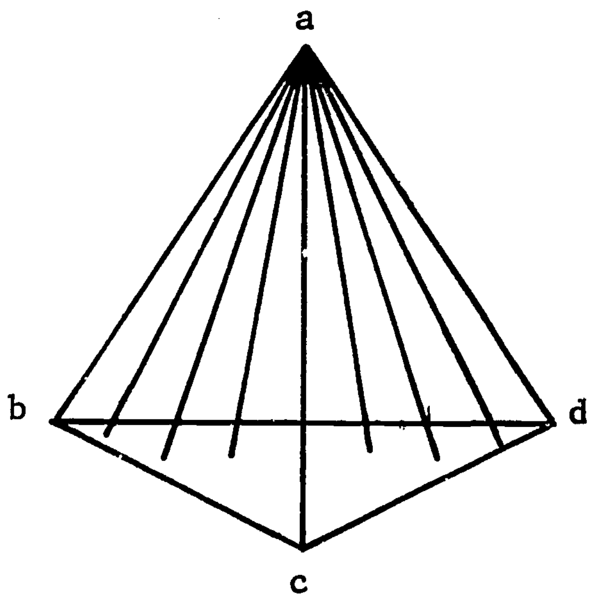


Figure 2

We might also wish to form (ab)(cd) the join of ab and cd. We therefore give the natural definition for the join of two sets:

$$AB = \bigcup_{\substack{a \subset A \\ b \subset B}} (ab).$$

Thus AB is the union of all possible joins of points in A with points in B (Figure 3). It follows that $x \subset AB$ if and only if $x \subset ab$ where $a \subset A$ and $b \subset B$. The definition covers the previously introduced idea of a join aB, such as a(bcd), since we have agreed to identify an element

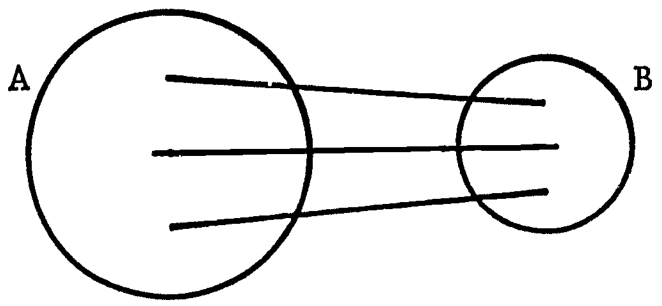


Figure 3

with its singleton set. Also the notation AB for join of sets is consistent with the notation ab for join of points when A, B are the singleton sets $\{a\}$, $\{b\}$.

Using this definition of join, we can convince ourselves that (ab)(cd) is the interior of the tetrahedron abcd, when a, b, c, d are in general position, so that (ab)(cd) = a(bcd).

While the join operation is, I think, the most important operation in elementary geometry, it is not the only important one. Next in importance is that of prolonging a segment indefinitely to form a ray. Instead of regarding this as an operation on a segment, we shall regard it as an operation on the ordered pair of endpoints of the segment. The prolongation of segment ab beyond a will be denoted a/b (see Figure 4). ab may be read "the extension of a from b" or simply "a over b." In terms of the join operation, the formal definition of a/b is simply

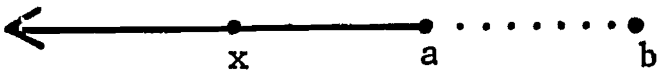


Figure 4

$$a/b = \{x \mid a \subset xb\}.$$

We observe that a/b, when $a \neq b$, is an open ray that does not contain its endpoint a. For $a \subset a/b$ implies $a \subset ab$, which is impossible since ab is an open segment. Note also that the definition is applicable when $a = b$ so that a/a has been formally defined. By the definition $x \subset a/a$ if and only if $a \subset xa$. The latter holds if $x = a$ (since $aa = a$ by definition) and only if $x = a$ (since $x \neq a$ implies xa is an open segment). Thus $a/a = a$.

It is desirable to extend the operation to sets. We define A/B (called the extension of A from B) by the formula

$$A/B = \bigcup_{\substack{a \subset A \\ b \subset B}} (a/b).$$

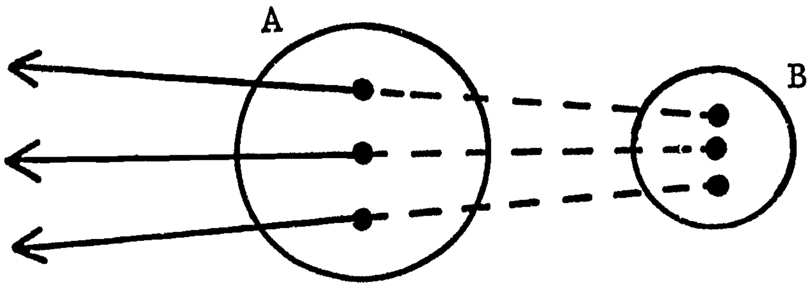


Figure 5

Note as above $x \subset A/B$ if and only if $x \subset a/b$ where $a \subset A$, $b \subset B$. Also the notation A/B is consistent with the notation a/b if $A = \{a\}$ and $B = \{b\}$. An important illustration of A/B occurs if A is simply a point a. Then A/B is a cone with vertex a whose elements are the rays a/b for all $b \subset B$ (see Figure 6). We can test the usefulness of this notation by applying it to familiar figures and basic properties in elementary geometry.

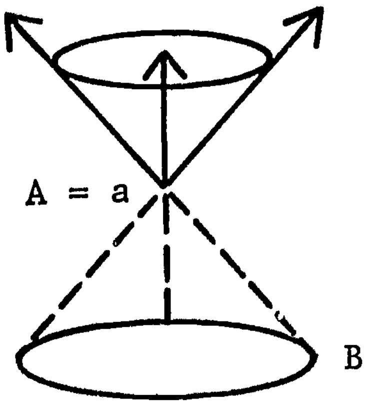


Figure 6

We note that line ab is partitioned into the sets a, b, ab, a/b, and b/a. (See Figure 7.)

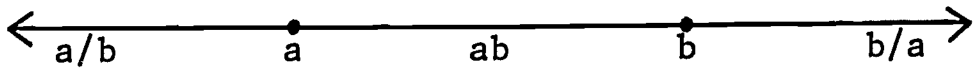


Figure 7

In order to generalize this result to plane abc we consider the expressions (ab)/c and a/(bc) written simply as ab/c and a/bc. By applying the definition of A/B above, we see that ab/c and a/bc represent regions of the plane abc as indicated in Figure 8.

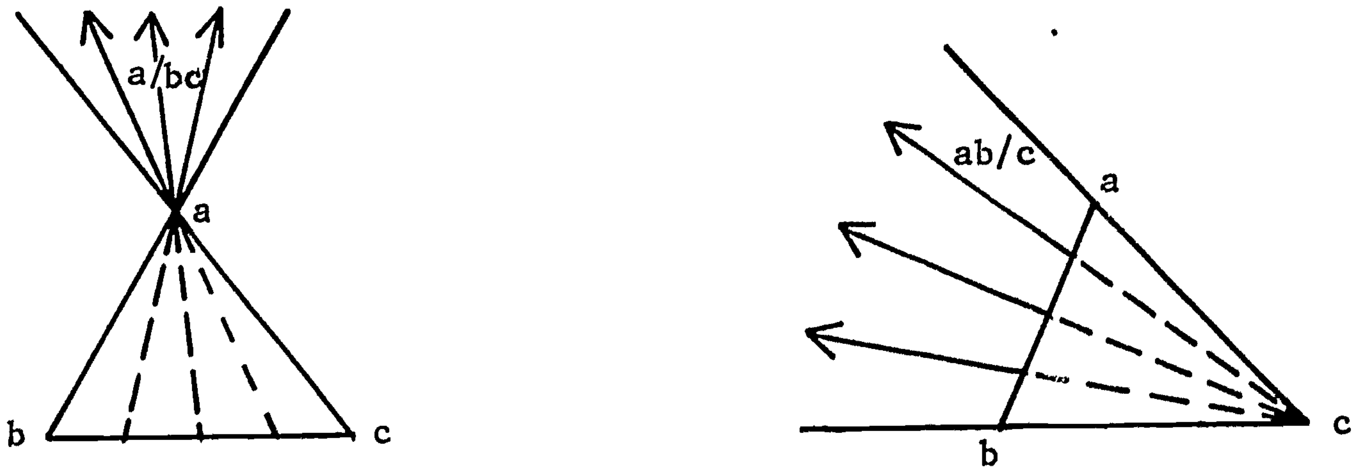


Figure 8

By using results of this type we obtain a partition of plane abc into 7 regions, 6 rays, 3 segments and 3 points which are expressed as products and certain quotients of products of a, b, and c, as indicated in Figure 9. Entirely analogous results hold in n-dimensional geometry (see [3], Section 11).

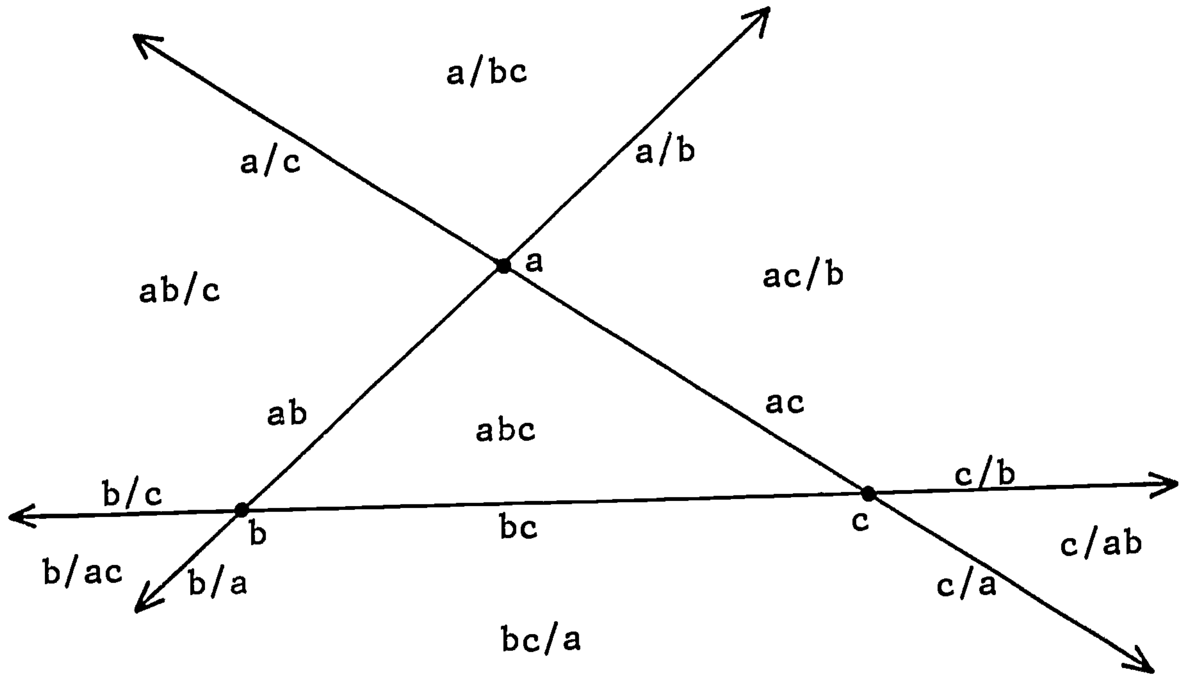


Figure 9

We can further express geometric identities with the help of this notation. For example Figure 10 illustrates $a/bc = (a/b)/c$. On the other

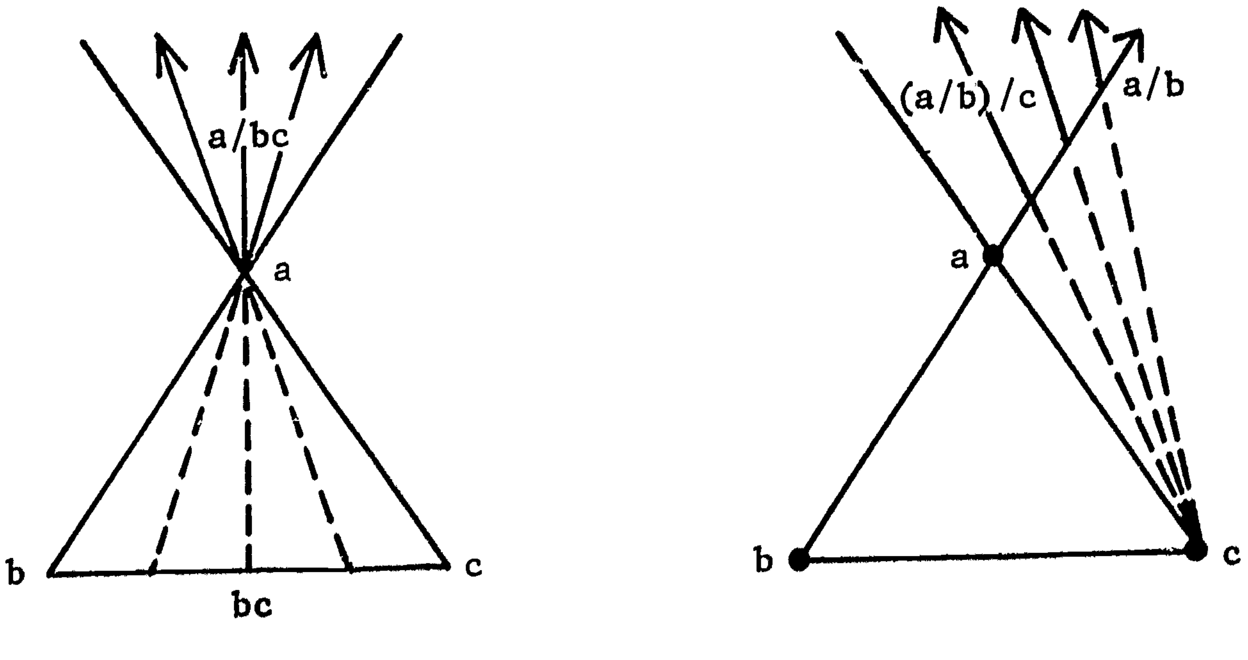


Figure 10

hand, it must not be assumed that all of the formal identities of school algebra are valid in this system. For example, we can only say $a(b/c) \subset ab/c$, as indicated in Figure 11.

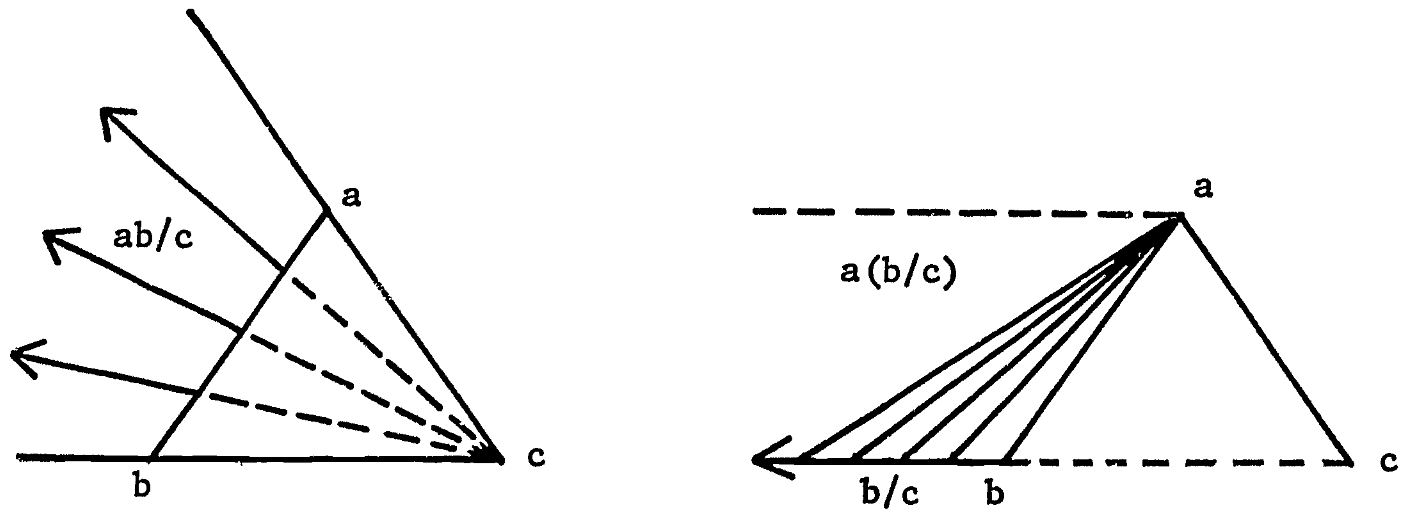


Figure 11

These considerations suggest a treatment of geometry as an abstract system (called a join system) based on a join operation. We therefore study a system $(G, \cdot)$, consisting of a set G and a binary operation which maps ordered pairs of elements of G onto subsets of G. As above, lower case letters will denote elements of G (called points) and capital letters subsets of G. We assume the following axioms.

J1: If a and b are elements of G, ab is a uniquely determined, nonempty subset of G.

J2: For any a, b in G, $ab = ba$.

In order to state the associative law, it is necessary to extend the definition of multiplication.

Definition. If $A \subset G$, and $B \subset G$, we define

$$AB = \bigcup_{\substack{a \subset A \\ b \subset B}} ab.$$

In view of our conventions concerning the identification of points with their singleton sets, this definition is easily shown to be consistent with the binary operation defined on points, that is $ab = \{a\}\{b\}$. Similarly we can regard the join aB as defined $(aB = \{a\}B)$. Note that x is in AB if and only if x is in ab for some a in A and some b in B.

J3: For any a, b, c in G, $(ab)c = a(bc)$.

It is natural to define an operation which is the inverse of join. Because the join of a and b is a set, care must be taken in framing the definition. For example, we should not expect a/b to be an element x such that $bx = a$.

Definition. For any a, b in G, $a/b = \{x \mid xb \supset a\}$.

Again, we can extend the definition to sets A and B.

Definition. If A and B are subsets of G,

$$A/B = \bigcup_{\substack{a \subset A \\ b \subset B}} (a/b).$$

It is easy to see that $x \subset A/B$ if and only if $x \subset a/b$ for some $a \subset A$ and some $b \subset B$.

J4: For any a, b in G, a/b is non-empty.

The next axiom is suggested by a triangle postulate employed by Peano. It says in effect that if, in triangle bdf, a is chosen on side bf and c is chosen on side df, then the segments ad and bc intersect. (See Figure 12.) Note the hypothesis of this triangle postulate can be

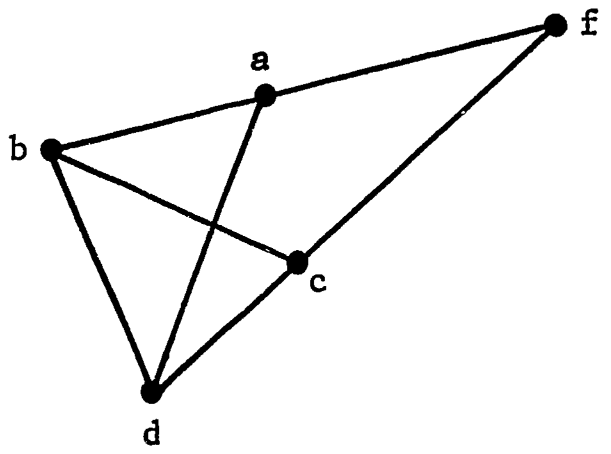


Figure 12

restated to assert that rays a/b and c/d meet.

J5: If $a/b \cap c/d \neq \emptyset$, then $ad \cap bc \neq \emptyset$.

This statement of axiom J5 is correct, but we must observe that it is very inelegant. The simple relation that two sets meet is expressed in notation which obtrudes the operation of set intersection. Therefore, I should like to introduce a new notation. I do so with some hesitation, since I don't believe that new notations should be introduced lightly. However, the notion that two sets meet or intersect seems so important in geometry and in other subjects that it seems warranted. We shall therefore write $A \approx B$ to mean $A \cap B \neq \emptyset$. As a consequence we may now restate J5 in a neater form.

J5: If $a/b \approx c/d$, then $ad \approx bc$.

We also note that the condition for $x \subset a/b$ can be simply phrased in this notation. Thus $x \approx a/b$ if and only if $xb \approx a$. Similarly $x \approx A/B$ if and only if $xB \approx A$.

The final postulate asserts our idempotent laws.

J6: $aa = a/a = a$.

As we indicated in our preliminary discussion, Euclidean geometry or affine geometry of any dimension (finite or infinite) is a model for this axiom system. We take G as the set of points of the geometry, and the join ab is defined as the open segmenc joining a to b if $a \neq b$, while the join aa is simply defined to be a. Wε can form a model from any vector space over the reals or over an ordered field. Here, we choose G as the set of elements of the vector space and we define the join ab as the set of non-trivial convex combinations of a and b. Thus

$$ab = \{x \mid x = \lambda a + \mu b, \text{ where } \lambda + \mu = 1 \text{ and } 0 < \lambda, \mu\}.$$

However, we note that the system of axioms is weak enough to cover many other systems.

Note in these examples that the axioms hold for all degenerate cases. For example, the associative law holds when a, b, and c are collinear. Indeed a, b, and c can be arbitrary. This is quite an advantage over conventional formulations of geometry in which so many statements must be qualified to exclude degenerate cases.

Now, what can be done with this system? We can, in the usual way, extend the operation to n points inductively. The generalized associative and commutative law is valid for n points and for n sets of points. In the Euclidean model, the join, $a_1 \ldots a_n$, of n points is the interior of the simplex determined by the points, provided the points are in general position (Figure 13 (a), (b), (c)). In all cases $a_1 \ldots a_n$ is the interior of the convex polyhedron "generated" by $a_1, \ldots, a_n$. For example, if a_1, a_2, a_3, a_4 are coplanar, no three of them are collinear and none of the points is in the join of the other three, then $a_1a_2a_3a_4$ is the interior of the convex quadrilateral whose vertices are a_1, a_2, a_3, a_4 (Figure 13 (d)).

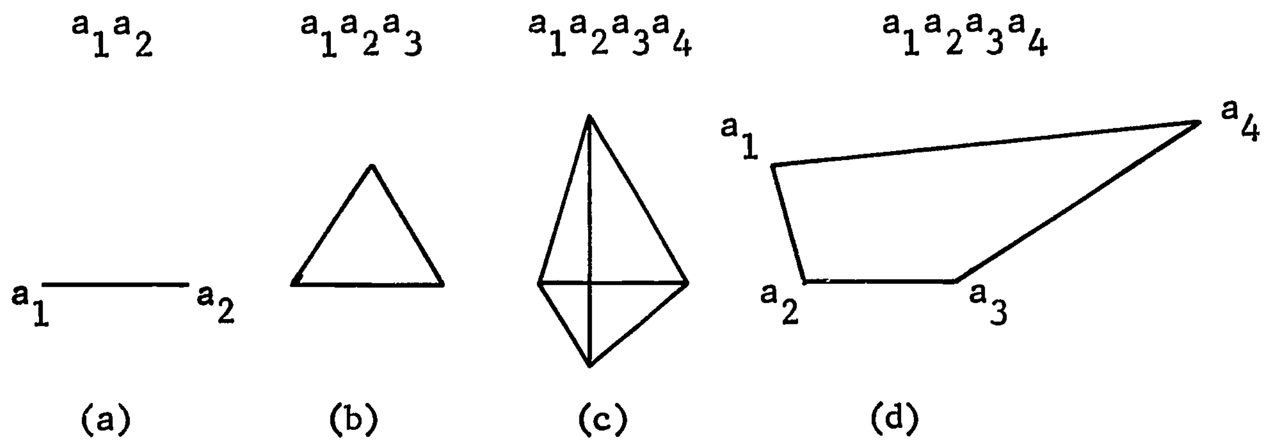


Figure 13

We now state some theorems which are derivable from the axioms.

Theorem. $a/bc = (a/b)/c$.

Theorem. $(a/b)c \subset ac/b$.

Theorem. $(a/b)(c/d) \subset ac/bd$.

Theorem. $(a/b)/(c/d) \subset ad/bc$.

All of these theorems can be extended to sets. Thus $A/BC = (A/B)/C$, $(A/B)C \subset AC/B$, etc.

In this system, what are the distinguished types of sets? Since the basic operation is join, convex sets come to the fore.

Definition. The set A is called convex if it is closed under join, that is, if $x, y \subset A$ implies $xy \subset A$.

Theorem. A is convex if and only if (1) $A \supset AA$ or (2) $A = AA$.

Theorem. If A and B are convex, then $A \cap B$, AB and A/B are convex.

Definition. The convex closure (or convex hull) of any set S is the least convex set, $[S]$, containing S. Thus $[S]$ is the intersection of all convex sets containing S.

$[S]$ may be characterized as follows: $x \subset [S]$ if and only if there are finitely many points $s_1, \ldots, s_k$ of S such that $x \subset s_1 \ldots s_k$.

It is also reasonable to consider sets which are closed under the join operation and its "inverse." This leads to the notion of linear sets (or flats).

Definition. The set A is called linear if it is closed under the operations of join and extension, that is, if $x, y \subset A$ implies $xy \subset A$ and $x/y \subset A$.

It might be considered preferable to define linear set in a more conventional way: A is linear if $x, y \subset A$ $(x \neq y)$ implies line $xy \subset A$. This is not wrong--but how is line xy to be defined? The familiar definition in Euclidean geometry (see Figure 7 above)

$$\text{line } xy = x \cup y \cup xy \cup x/y \cup y/x$$

is not suitable in our theory since the indicated set union is not necessarily closed under join and extension. (See the first model presented in the Discussion below.)

Theorem. If A and B are linear and $A \approx B$, then A/B is linear.

Theorem. If A is convex, then A/A is linear.

We may define the linear closure $\langle S \rangle$ of a set S in a way similar to the definition of the convex closure.

Definition. The linear closure of a set S is the least linear set $\langle S \rangle$ which contains S.

We can characterize the linear closure of finite sets very nicely by means of the following theorem.

Theorem. If $S = \{a_1, \ldots, a_n\}$ is a finite set, then

$$\langle S \rangle = a_1 a_2 \ldots a_n / a_1 a_2 \ldots a_n.$$

This theorem suggests a definition and formula for line in our theory. For certainly in Euclidean geometry line ab is the least linear set which contains a and b.

Definition. If $a \neq b$, ab/ab is a line.
As we suggested above, this need not equal $a \cup b \cup ab \cup a/b \cup b/a$. (See the first model in the Discussion below.)

We conclude by remarking that the notion of the interior of a convex set can be defined and studied in the theory.

Definition. Let K be convex. Then I(K), the interior of K, is defined by

$$I(K) = \{p \mid x \subset K \text{ implies } p/x \approx K\}.$$

Theorem. Let K be convex. Then $x \subset K$ implies

$$I(K) \subset xK \subset K.$$

Theorem. If K is convex, $K \cdot I(K) = I(K)$.

Corollary 1. If K is convex, I(K) is convex.

Corollary 2. Let K be convex, $x \subset I(K)$. Then $Kx = I(K)$.

Theorem. Let K be convex. Then $I(I(K)) = I(K)$.

Definition. The convex set K is open if $K = I(K)$.

Theorem. If K_1 and K_2 are open convex sets, then $K_1 \cap K_2$, K_1K_2, and K_1/K_2 are open convex sets.

It is worth noting that the proofs involve essentially straightforward applications of the axioms. They do not require study of degenerate cases, since the axioms were set up in such a way as to include all cases.

Discussion.

In answer to a question Prenowitz gave the following model in which a line is not always expressible in the form $a \cup b \cup ab \cup a/b \cup b/a$. If $(G_1, \cdot)$ and $(G_2, \cdot)$ are join systems we can form their "direct product" $(G, \cdot)$, which also is a join system, as follows. Let $G = G_1 \times G_2$, the Cartesian product of G_1 and G_2. We define join in G thus:

$$(a_1,a_2) \cdot (b_1,b_2) = (a_1b_1) \times (a_2b_2).$$

Let $(G_1, \cdot)$ and $(G_2, \cdot)$ be the real line considered as a (Euclidean) join system. Then $(G, \cdot)$ is the Cartesian plane with $a \cdot b$ interpreted as follows: (1) If a and b are on a vertical line or on a horizontal line $a \cdot b$ is the join of a and b in the usual sense in a Euclidean geometry; (2) in the contrary case, $a \cdot b$ is the interior of the rectangle with horizontal and vertical sides which has a and b as a pair of opposite vertices.

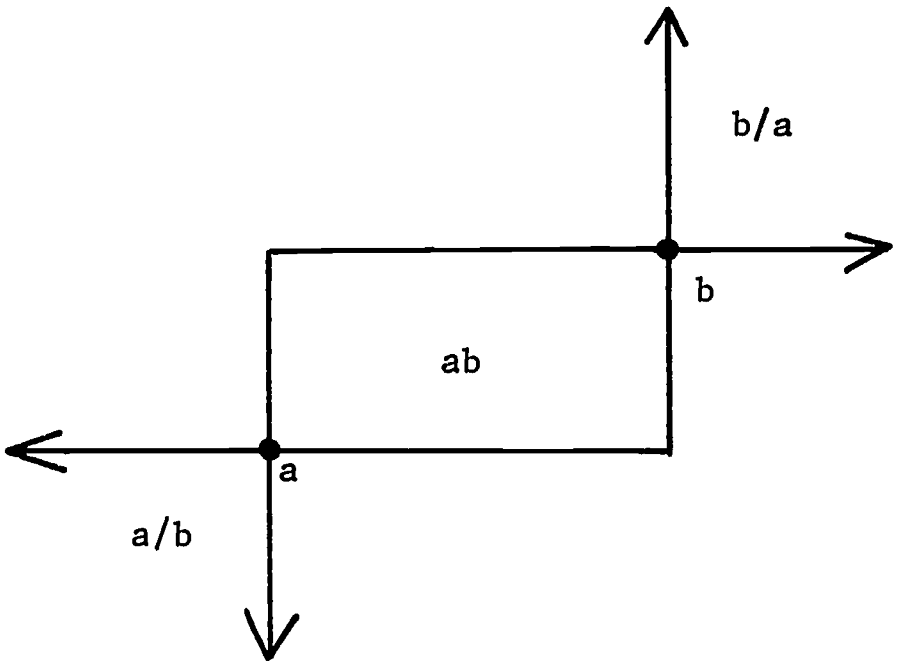


Figure 14

Figure 14 indicates that $a \cup b \cup ab \cup a/b \cup b/a$ is not a linear set, indeed it is not even convex, and certainly is not a suitable definition for line in the theory.

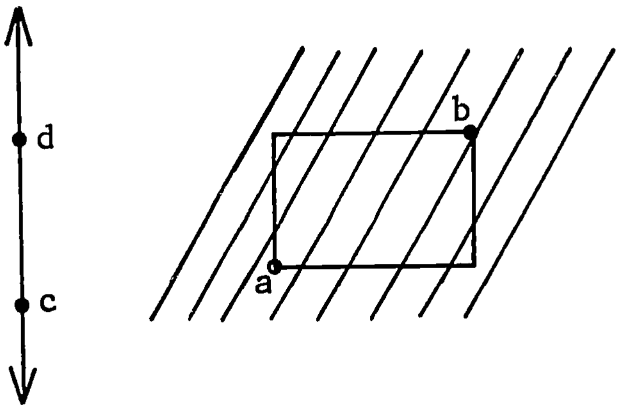


Figure 15

The model also is a counterexample for the statement that two distinct points determine a line. In Figure 15 $\langle\{a,b\}\rangle = ab/ab$, which by definition is a line, turns out to be the whole Cartesian plane. But $\langle\{c,d\}\rangle = cd/cd$ is just the vertical line cd. Thus the distinct points c and d are contained in two distinct lines, namely, $\langle\{a,b\}\rangle$ and $\langle\{c,d\}\rangle$.

We can also generalize the vector space model referred to earlier by replacing the field of reals by any partially ordered field. A partially ordered field is a field that has a distinguished subset P (whose elements are called positive) which is closed under addition, multiplication and the operation of taking reciprocals. Any vector space over a partially ordered field F becomes a join system when join is defined exactly as for a vector space over the reals. If F is not an ordered field, the resulting join system has the interesting property that its lines are not fully ordered sets of points. That is, any line contains three distinct points, none of which is in the join of the other two. A related property is that Pasch's Postulate fails in the given type of join system.

In response to a question, Prenowitz observed that Helly's Theorem fails in such "partially ordered" join systems. Helly's Theorem seems to be equivalent to the condition that the points of any line form a fully ordered set.

Another class of models noted by Prenowitz may be described as subsystem models. If G is any join system and K an open convex set in G, we obtain another model by relativizing G with respect to K (see [1] pp. 60-61).

Grünbaum asked if it was possible to add axioms systematically to get to Euclidean space. Prenowitz answered affirmatively. For example, the theory of incidence and dimension can be obtained by postulating a weak form of the principle that two points determine a line, which is related to the Steinitz exchange principle (see [1] p. 39).

If in addition we postulate that the points of a line form a fully ordered set (see [1] pp. 44-45), the familiar theory of separation of linear sets can be derived. In particular the theorem on the separation of a linear set by a linear set can be given a dimension-free treatment in the form: If A and B are linear and A covers B (that is, B is a maximal proper linear subset of A) then B separates A.

In reply to a question, Prenowitz said that the separation theorems for convex sets would probably go through in any join system for the finite-dimensional case. But the infinite-dimensional case seemed more difficult, unless some restrictions were assumed.

References

1. W. Prenowitz, A contemporary approach to classical geometry, *American Mathematical Monthly*, Vol. 68, No. 1, Part II (The Ninth Herbert Ellsworth Slaught Memorial Paper), 1961.

2. ____________, Projective geometries as multigroups, *American Journal of Mathematics* 65(1943) 235-256.

3. ____________, Descriptive geometries as multigroups, *Transactions of the American Mathematical Society* 59(1946) 333-380.

4. ____________, Partially ordered fields and geometries, *American Mathematical Monthly* 53(1946) 439-449.

5. ____________, Spherical geometries and multigroups, *Canadian Journal of Mathematics* 2(1950) 100-119.

REPORT ON CURRICULUM DISCUSSION

Walter Prenowitz

Curricular questions arose constantly throughout the period of the Conference--they were sometimes explicitly introduced by the lecturer but often appeared spontaneously in the discussion following the lecture. The questions were not always followed through, since the conference ran for only three weeks and much material was presented that was new and stimulating both in mathematical and curricular terms. Moreover, new approaches to curriculum and teaching often need to be mulled over or given a resting period before they crystallize. The Proceedings contain much interesting and stimulating material on curricular and pedagogical questions, from which I am making a selection of points which seem most important and salient.

When the conference began, no specific periods had been scheduled for the formal discussion of curriculum. Klee, who felt it was important to initiate such discussion, decided to devote the first lecture in his series on Applications of Geometry (Part I, pp. 7-12) to consideration of collegiate geometry courses, their desirable characteristics, and to the geometry of convex bodies as possible subject matter for such a course. Klee's lecture evoked much response and his suggested criteria were brought up several times during the conference as a basis for discussion of geometry courses.

Part II, Geometry in Other Subjects, contains the lectures of Steenrod on the geometric content of the freshman and sophomore mathematics courses and those of Gleason set at a somewhat higher level. There is much interesting material here, not readily accessible elsewhere, in which these eminent geometric-minded mathematicians give their views on how certain important portions of the undergraduate curriculum should be conceived, organized mathematically, and taught. It presents a perspective which ordinarily would be open to their students alone.

Several times in the course of the discussion, dissatisfaction was expressed with the treatment of linear algebra. There was a feeling that linear algebra courses too often tend to put excessive emphasis on computation within coordinate systems (matrices, determinants, etc.) at some expense to conceptual understanding. The view was expressed that conventional linear algebra courses treat geometry in an off-hand fashion and do not seem to give students an understanding of the geometric basis for the vector concept--specifically that students tend to become confused on the distinction between affine geometry (non-centered) and vector space theory (centered). It was suggested that some time be taken at the beginning of a linear algebra course for discussion of affine geometry, so that the student gets an appreciation of the affine concepts before he is introduced to the powerful vector theoretic ideas which may swallow them up. I wonder whether the ideas of affine geometry aren't sufficiently basic and important to merit inclusion in some geometry course that would be taken before the vector space concept is studied.

Several formal discussions on curriculum were held as the need for them was expressed by participants. The following recommendations were discussed and approved at these meetings.

Recommendations of the Conference

I. That a pamphlet or booklet be written in the spirit of "Geometry in Other Subjects" to be concerned with the geometric background, content and motivation in the first year calculus course.

II. That the following principles be given appropriate weight as general guidelines.

1. An upper level geometry course should, if possible, be oriented toward n-dimensional space. The framework and methods, if not themselves n-dimensional, should generalize to n-space.

2. The course should be unified by an important idea.

3. Geometry courses should make more use of intrinsic concepts and methods, such as transformations, vectorial methods, metric space methods, iterative processes such as join operations, incidence and lattice concepts.

4. Geometry courses should, so far as possible, treat topics relevant in other branches of mathematics.

5. In every course sufficient emphasis should be given to geometry as a way of viewing mathematics.

These principles are not intended to be exhaustive and may not be wholly consistent with each other but hopefully should stimulate critical discussion. They are intended to be a first word--not a last!

III. That mathematics teachers be encouraged to experiment with junior-senior level courses in the following subjects: Convex Sets, Geometry from the Transformation Group Viewpoint, Metric Linear Geometries.

Sample Course Outlines

It cannot be emphasized too strongly that the following outlines and descriptions for the three courses are suggestions for experimentation--they are not course syllabi approved either by the Conference or by CUPM for general adoption.

Convex Sets

A. Fundamental Topics.

1. Preliminaries on R^n: Basic linear algebra and affine geometry, open and closed sets, etc.
2. Basic properties of convex sets: Supporting hyperplanes, supporting functions, distance functions, convex functions, convex hulls, boundary structure, polytopes, duality.
3. Separation and support theorems, extreme and exposed points.
4. Helly's Theorem, Radon's Theorem, Carathéodory's Theorem.
5. Convergence of convex sets, approximation by polytopes.
6. Applications to other branches of mathematics including analysis.
7. Isoperimetric problems.
8. Discussion of unsolved problems.

B. Additional Topics.

9. Mixed volumes, symmetrization.
10. Polytopes.
11. Packing and covering.
12. Sets of constant width.
13. Variants and generalizations of Helly's Theorem, Radon's Theorem, Carathéodory's Theorem.

The topics listed are intended to give an indication of the potentialities of the macerial and are not necessarily definitive. Some topics might be shifted from one list to the other depending on the length of the course, the level of the students and the teacher's interests. The treatment would be framed, when the material became too difficult, in R^2 and R^3 but would still point naturally toward R^n.

Geometry from the Transformation Group Viewpoint

The study of a branch of geometry as illumined by examination of its underlying transformation group. A natural choice of geometry would be the Euclidean plane and space, the isometries and similarities forming the transformation groups. Topics studied would include transformations and their classification; groups and subgroups of transformations; geometric invariants; special attention would be given to discrete groups of transformations. Other possible choices of subject to be studied from this viewpoint might be affine geometry, projective geometry, inversive geometry or the classical non-Euclidean geometries. The course could also develop--depending on the interests of the teacher and limitations of time--some of the properties of transformations over a general field, particularly linear, affine and projective transformations.

Metric Linear Geometries

This is an attempt to present the essence of the concept of projective metrics by requiring only a minimal knowledge of projective geometry. The basic systems studied might be "subgeometries" of affine rather than projective geometry. The geometry is postulated to be a metric space which satisfies the following metric-incidence postulate: The strict triangle inequality holds for any three distinct noncollinear points.

Incidence geometries; affine and projective geometries. Metric spaces. Metric betweenness, metric segments, lines and great circles. Convexity. Perpendicularity. Congruence and isometries. Minkowski, Euclidean, Hilbert, and hyperbolic geometries.

Reference: Busemann and Kelly, Projective Geometry and Projective Metrics, Chapter IV. There is much attractive material here which was written fifteen years ago and probably now can be made accessible to a wider audience.