MATHEMATICS MAGAZINE

CONTENTS

CODEN: MAMGA8

VOLUME 48 • JANUARY 1975 • NUMBER 1

THE MATHEMATICAL ASSOCIATION OF AMERICA
1225 Connecticut Avenue, N.W.
Washington, DC 20036

Address Change Requested

MATHEMATICS MAGAZINE

GERHARD N. WOLLAN, *EDITOR*

EDITORIAL CORRESPONDENCE should be sent to the EDITOR, G. N. WOLLAN, Department of Mathematics, Purdue University, Lafayette, Indiana 47907. Articles should be typewritten and triple-spaced on $8\frac{1}{2}$ by 11 paper. The greatest possible care should be taken in preparing the manuscript. Authors should submit two copies and keep one copy for protection against possible loss of the manuscript. Figures should be drawn on separate sheets in India ink. They should be appropriately lettered and of a suitable size for photographing.

REPRINT PERMISSION should be requested from LEONARD GILLMAN, Mathematical Association of America, University of Texas, Austin, Texas 78712. (See also the copyright notice below.)

NOTICE OF CHANGE OF ADDRESS and other subscription correspondence should be sent to the Executive Director, A. B. WILLCOX, Mathematical Association of America, Suite 310, 1225 Connecticut Avenue, N. W., Washington, D.C. 20036.

ADVERTISING CORRESPONDENCE should be addressed to RAOUL HAILPERN, Mathematical Association of America, SUNY at Buffalo (University of Buffalo), Buffalo, New York 14214.

THE MATHEMATICS MAGAZINE is published by the Mathematical Association of America at Washington, D.C., bi-monthly except July–August. Ordinary subscriptions are $10 per year. Members of the Mathematical Association of America and of Mu Alpha Theta may subscribe at the special rate of $7.00. Back issues may be purchased, when in print, for $2.00.

Second class postage paid at Washington, D.C. and additional mailing offices.

MATHEMATICS MAGAZINE

Published by

THE MATHEMATICAL ASSOCIATION OF AMERICA
(INCORPORATED)

A JOURNAL OF COLLEGIATE MATHEMATICS

VOLUME 48
1975

Published at

MENASHA, WIS. AND WASHINGTON, D. C.

ANNOUNCEMENT

Beginning with the September-October 1953 issue, Professor Robert E. Horton has served continuously as Editor of the Problem Section of the MATHEMATICS MAGAZINE. He has requested to be relieved of this responsibility effective as of January 1 of this year. Combining a keen interest in problems with a deep conviction of the importance of maintaining the MATHEMATICS MAGAZINE and improving its quality, he played a key role in guiding the MAGAZINE through a difficult period. In addition to being Editor of the Problem Section, he also served as Editor for four years from January 1960 to December 1963 and as publisher during the year 1960. In grateful recognition of his many years of devoted and effective service to the MATHEMATICS MAGAZINE and through it to mathematics and mathematical education, Robert E. Horton is hereby elected Associate Editor Emeritus of the MATHEMATICS MAGAZINE.

BOARD OF GOVERNORS OF THE MATHEMATICAL ASSOCIATION OF AMERICA

A SIMPLE APPROACH TO ISOPERIMETRIC PROBLEMS IN THE PLANE

RICHARD F. DEMAR, University of Cincinnati

Introduction. An isoperimetric problem, as the name suggests, is a problem of finding among all regions which have a given perimeter and which satisfy some given set of conditions, that region whose area is maximum. The usual way of solving such a problem is to show that if a solution exists, then it must have a certain shape and to show that a solution does exist. In early attempts at dealing with isoperimetric problems, the existence of a solution seemed so self-evident that it was not realized that this was a problem to be considered. Thus, it was tacitly assumed that a solution did exist and the only question dealt with was the shape which the solution must have. This still is the only part of the problem amenable to attack by elementary methods. The purpose of this paper is to report an elementary method which seems to have wider application than most methods used previously, so we shall always assume that a solution exists and consider only the question of what the solution must be.

Isoperimetric problems have been considered since antiquity. There is a legend that Queen Dido solved an isoperimetric problem (by intuition) and that this led to the founding of Carthage [**5**, p. 882]. According to Coolidge [**2**], the Greek mathematician, Zenodorus, studied such problems in the second century B.C. and his results were discussed and extended by Pappus five centuries later. The invention of the calculus of variations by Lagrange in the eighteenth century gave us our most powerful tool for dealing with isoperimetric problems, especially in more general settings. However, geometric methods can be used successfully in the plane and in certain other settings as was shown by the great nineteenth century geometer, Jacob Steiner [**6**].

One rather immediate simplification of isoperimetric problems is the following theorem.

(*) *If R is a region of maximum area among all regions having perimeter p, then R is convex.*

A set S is convex if for each pair of points A and B in S, the line segment AB is contained in S. A proof of (*) is given in [**4**]. The idea of the proof is simple; namely, that if R is not convex, then there exists a region of larger area and smaller perimeter than R (Figure 1). This theorem has been used consistently to simplify

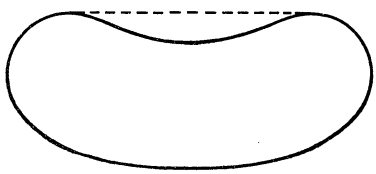


FIG. 1

the various problems which have been considered, but no one method has been used consistently to solve the various problems. Geometric methods have for the most part been specific to the problem at hand. It is our purpose to show that theorem (*) provides not just a method of simplifying many problems, but a method of solving them. It is an example of a very simple approach which has much greater power than one would at first suspect.

The manner in which (*) can be used to solve problems is this. If we want to prove that the region of maximum area satisfying given conditions has a certain property, we consider a region R satisfying the given conditions but not having the given property. If we can construct a region R' which is not convex but which has the same area and perimeter as R and satisfies the given conditions, then R cannot have maximum area since otherwise R' would be a nonconvex solution. This method was used in a limited way on a particular problem by Adler [**1**], but it is not suggested there as having wider applicability.

We illustrate the method with a trivial example by using it to show that the region of maximum area among all plane regions of perimeter p is not a square. If R is a square $ABCD$ of perimeter p, then its sides are of length $p/4$. Choose E on AB and F on BC so that the length of EF is less than $p/4$. Construct $\Delta E'F'B' \cong \Delta EFB$ with F' and E' interior points of side DC. Then the polygon $AEFCE'B'F'D$ is not convex, but it has the same area and perimeter as R. Therefore R cannot have maximum area among all regions of perimeter p (Figure 2).

We shall demonstrate the usefulness of the method by using it and slight modi-

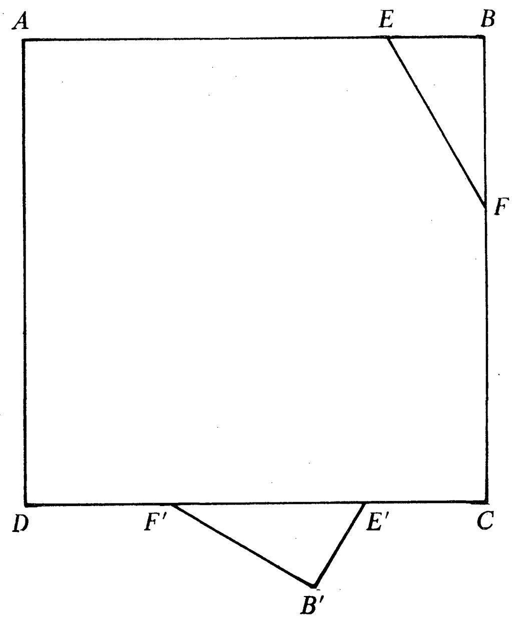


FIG. 2

fications of it to solve the following three problems: Let p be a given positive number.

1. Let n be a given positive integer. Among all n sided polygons of perimeter p, find the one having maximum area.

2. Among all plane regions of perimeter p, find the region having maximum area.

3. Given a triangle of perimeter greater than p, among all regions of perimeter p contained in the triangle, find the region of maximum area.

The first two of these problems are classical and it has long been known that the solution to the first is a regular n-gon and to the second is a circular region. The proofs given here seem to be new. The author is grateful to Professor Alfred Garvin for suggesting the third problem and for conjecturing the correct solution which he arrived at by physical experiments. Both Professor Garvin and the author believed the problem to be new. However, after the research was done and this paper was being written, the author discovered that the problem and its solution were published by Steiner in 1842 [**6**, p. 149]. Such are the hazards of working on an elementary problem. The proof given here is different from Steiner's and is believed to be new.

We have been using the term region thus far without defining it. By a *region* we shall mean the closure of a bounded open connected set whose boundary is a simple closed rectifiable curve. We shall denote the length of a line segment AB by $l(AB)$ and the measure of an angle ABC by $m(\angle ABC)$.

There are some well-known properties of the boundary of a convex set which we shall need. Let R be a convex set with boundary γ. Then γ is a simple, closed, continuous curve. Let A be a point on γ. Think of standing at A and facing the set R.

Then there is a right-hand and a left-hand direction along γ from A. A right-hand tangent at A is a ray from A tangent to γ in the right-hand direction; similarly for a left-hand tangent. One property of the boundary of a convex set is that at each of its points, it has a right-hand and a left-hand tangent [3]. The angle between these rays measured in the sector containing R will be called the interior angle at A. Since R is convex, the interior angle at every point will be at most $180°$. If we let S be any region with the property that its boundary has right-hand and left-hand tangents at every point, if there exists a point at which the interior angle is more than $180°$, then S is not convex. At every point of the boundary at which the interior angle equals $180°$, we say that γ has a tangent. Another property of the boundary of a convex set is that it contains a dense set of points at each of which there is a tangent. A subset β of γ is *dense* if between each pair of points on γ there are infinitely many points of β.

Solution to Problem 1. *Among all n-sided polygons of perimeter p, the regular n-gon has maximum area.*

Proof. We first show that the solution must be equilateral, then that it must be equiangular. Let $P_n = A_1A_2 \cdots A_n$ be an n-gon which is not equilateral. Then P_n has two consecutive sides whose lengths differ. Assume without loss of generality that $l(A_1A_2) > l(A_2A_3)$ (Figure 3). Choose C on A_1A_2 so that $l(A_2A_3) < l(A_2C)$

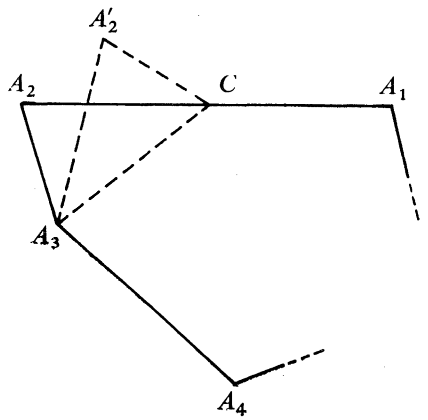


FIG. 3

$< l(A_2A_1)$. Draw A_3C. Then $m(\angle A_2A_3C) > m(\angle A_2CA_3)$. Reflect ΔA_2A_3C in the perpendicular bisector of A_3C. Let A_2' denote the image of A_2 under this reflection, so that $\Delta A_2'A_3C$ is the image of ΔA_2CA_3. Since

$$m(\angle A_2'CA_3) = m(\angle A_2A_3C) > m(\angle A_2CA_3)$$

and

$$m(\angle A_2CA_3) + m(\angle A_1CA_3) = 180°$$

$$m(\angle A_2'CA_3) + m(\angle A_1CA_3) > 180°.$$

Therefore, the polygon $P' = A_1CA_2'A_3 \cdots A_n$ is not convex. Since $\Delta A_2A_3C \cong \Delta A_2'CA_3$, polygons P_n and P' have the same perimeter and the same area. But P' has $n+1$ sides. To show P_n cannot be a solution to the problem, draw A_1A_2'. Then the polygon $A_1A_2'A_3 \cdots A_n$ is an n-gon whose perimeter is less than the perimeter of P' and therefore less than the perimeter of P_n and whose area is greater than the area of P' and hence greater than the area of P_n. Therefore P_n cannot be a solution. Thus, a solution must be equilateral.

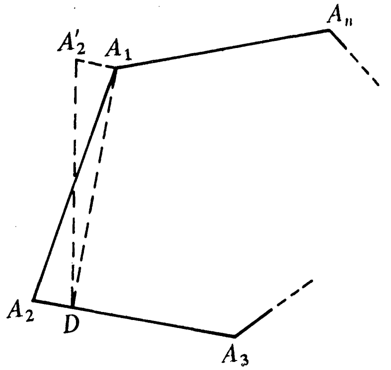


FIG. 4

That the polygon of maximum area must be equiangular can be proved in a similar way. Let $P_n = A_1A_2 \cdots A_n$ be an n-gon which is not equiangular; say $m(\angle A_nA_1A_2) > m(\angle A_2)$ (Figure 4). Then there is a number α such that $0 < \alpha < \frac{1}{2}[m(\angle A_nA_1A_2) - m(\angle A_2)]$. Choose D on A_2A_3 so that $m(\angle A_2A_1D) = \alpha$. Then

$$\begin{aligned} m(\angle A_3DA_1) &= m(\angle A_2) + m(\angle A_2A_1D) \\ &= m(\angle A_2) + \alpha \\ &< m(\angle A_2) + \tfrac{1}{2}[m(\angle A_nA_1A_2) - m(\angle A_2)] \\ &= \tfrac{1}{2}[m(\angle A_nA_1A_2) + m(\angle A_2)] \\ &= m(\angle A_nA_1A_2) - \tfrac{1}{2}[m(\angle A_nA_1A_2) - m(\angle A_2)] \\ &< m(\angle A_nA_1A_2) - m(\angle A_2A_1D) \\ &= m(\angle A_nA_1D). \end{aligned}$$

Reflect ΔA_2A_1D in the perpendicular bisector of A_1D. Let A_2' be the image of A_2. Then $m(\angle A_2'A_1D) = m(\angle A_2DA_1)$; so

$$\begin{aligned} m(\angle A_nA_1D) + m(\angle A_2'A_1D) &= m(\angle A_nA_1D) + m(\angle A_2DA_1) \\ &> m(\angle A_3DA_1) + m(\angle A_2DA_1) \\ &= 180^\circ. \end{aligned}$$

Therefore the polygon $P' = A_1A_2'DA_3 \cdots A_n$ is not convex. But since $\Delta A_2A_1D \cong \Delta A_2'A_1D$, the polygons P_n and P' have the same perimeter and area. But again P' has $n+1$ sides. If we draw A_nA_2', the polygon $P'' = A_2'DA_3 \cdots A_n$ has n sides and has greater area and smaller perimeter than P' and hence greater area and smaller perimeter than P. Therefore P is not a solution to the problem. Therefore, in order that an n-gon be a solution, it must be equiangular. This completes the proof.

Solution to Problem 2. *Among all plane regions of perimeter p, a circular region has maximum area.*

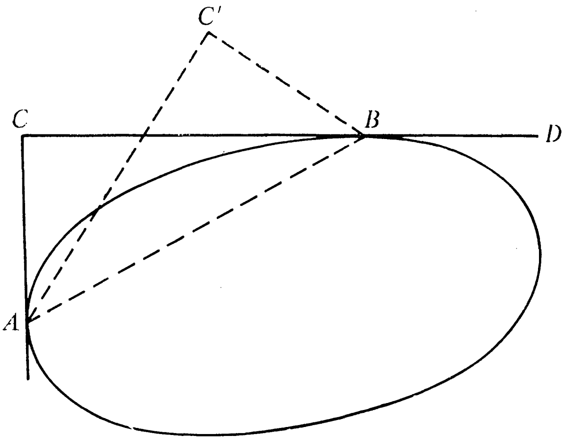


FIG. 5

Proof. Let R be a region of maximum area among all regions of perimeter p. Let γ be the boundary of R. Since R must be convex, there is a dense set of points on γ at each of which γ has a tangent. Let A and B be any two points of this dense set such that the tangents at A and B are not parallel. Let C denote their point of intersection (Figure 5). We claim that $m(\angle CAB) = m(\angle CBA)$. Suppose not; say $m(\angle CAB) > m(\angle CBA)$. Let $\overset{\frown}{AB}$ denote the part of γ inside ΔABC. Reflect ΔABC and $\overset{\frown}{AB}$ in the perpendicular bisector of AB. Let C' denote the image of C under the reflection. Then $\Delta ABC'$ is the image of ΔBAC. Let γ' be the curve obtained from γ by replacing $\overset{\frown}{AB}$ by its image under the reflection. Let R' be the region whose boundary is γ'. Since reflection preserves area and length, R and R' have the same area and the same perimeter. Since

$$m(\angle CBA) + m(\angle ABD) = 180°$$

and

$$m(\angle C'BA) = m(\angle CAB) > m(\angle CBA),$$

$$m(\angle C'BA) + m(\angle ABD) > 180°.$$

But BD and BC' are the left-hand and right-hand tangents to γ' at B, so R' is not convex. This contradicts R being the region of maximum area among all regions of perimeter p. Therefore, our assumption that $m(\angle CAB) > m(\angle CBA)$ is false. Therefore, $m(\angle CAB) = m(\angle CBA)$.

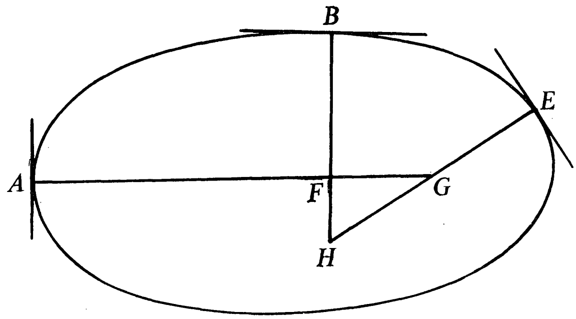


FIG. 6

If we construct perpendiculars to the tangents at A and B respectively and call the point of intersection F, then the angles FAB and FBA are complements of congruent angles and hence are congruent, so $l(FA) = l(FB)$. Now if we choose a third point E on γ at which γ has a tangent and draw a perpendicular to that tangent, letting the points of intersection with AF and BF be G and H respectively, and draw EA and EB, we obtain isosceles triangles EAG and EBH (Figure 6). Then

$$l(AF) + l(FG) = l(AG),$$

so by substitution

$$l(BF) + l(FG) = l(EG).$$

Also, $l(BF) + l(FH) = l(EG) + l(GH)$.

Subtracting the first equation from the second gives

$$l(FH) - l(FG) = l(GH)$$

or

$$l(FH) = l(FG) + l(GH).$$

Thus, ΔFGH is degenerate, and since AG, BH, and EH are distinct lines, this implies that $F = G = H$. Call the common point G. Then A, B, and E are equidistant from the point G. Thus every point on γ at which γ has a tangent is the same distance from G as is A. Since γ is a continuous curve, this implies that for every point X on γ, $l(XG) = l(AG)$. That is, γ is a circle.

Remark. This proof has an advantage over some proofs that have been given; namely, that it depends on local properties rather than global properties of γ. Essentially, it proves that γ has the same curvature at every point. Thus, it proves more than the stated result. If we consider other isoperimetric problems in which part of the boundary is restricted in some way and part of the boundary is free to assume whatever shape maximizes area, then our proof implies that each free part of the boundary is an arc of a circle. For example, the region R of maximum area among all regions having a fixed line segment AB as part of its boundary and having a given perimeter p where $p > 2l(AB)$ is a region whose boundary consists of AB together with an arc $\widehat{AB}$ of a circle with the length of $\widehat{AB}$ equal to $p - l(AB)$.

Before giving the solution to Problem 3, let us eliminate cases of this problem which have already been considered. We are given a triangular region T and a number p less than the perimeter of T and we are to find the region R of maximum area among all regions contained in T and having perimeter p. Let the inscribed circle of T have circumference c. If $p \leqq c$, then Problem 3 is the same as Problem 2 which we have already solved. Therefore in Problem 3, we assume $p > c$.

Solution to Problem 3. *Let T be a given triangular region with perimeter P and with circumference of the inscribed circle equal to c. Let p be a number such that $c < p < P$. Then among all regions contained in T and having perimeter p, the region R of maximum area has boundary γ consisting of three circular arcs all of the same radius, each tangent to two adjacent sides of the boundary of T, together with the three segments of sides of T between endpoints of these arcs* (Figure 7).

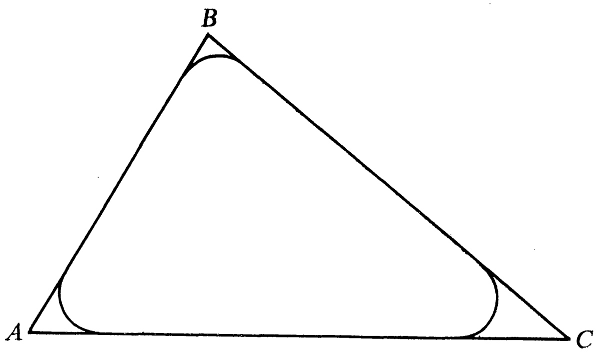


FIG. 7

Proof. Let R be a solution to the problem and let γ be its boundary. Then from the remark following the preceding solution, any part of γ which is not part of the boundary of T is a circular arc. Since $p < P$, the boundary of R cannot coincide with the boundary of T; so γ contains at least one circular arc.

We show first that γ cannot contain any vertex of T. Let T be ΔABC and suppose C lies on γ. Since R is convex, if all three vertices A, B, and C were on γ, then γ would coincide with the boundary of T which we have shown to be impossible. Therefore there is at least one vertex; say B, which is not on γ. Then there is a circular arc β with endpoints K and M on two sides of ΔABC, say BC and AB, respectively (Figure 8). Choose points E on BC and F on AC "close" to C. The meaning of "close" will be specified later. Construct a circular arc from E to F with the same radius as that of β and with the center on the opposite side of EF from C. Now

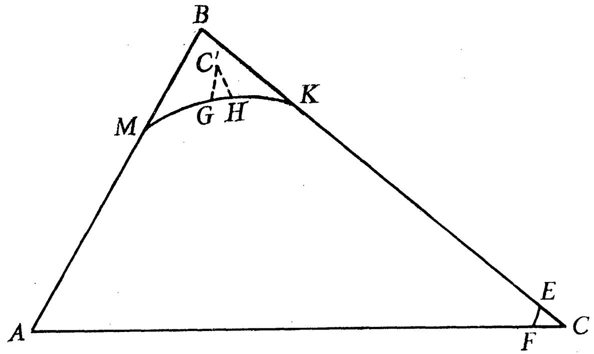


FIG. 8

choose G and H on β, not endpoints of β, so that $\widehat{GH} \cong \widehat{EF}$. Let C' be the point not in R such that $l(C'G) = l(CE)$ and $l(C'H) = l(CF)$. It is necessary that we have C' in T, so we define the term "close" used above. The points E and F can be chosen so that the following conditions hold:

(i) E is on KC,

(ii) $l(\widehat{EF}) < l(\widehat{MK})$,

(iii) $\widehat{EF}$ is not tangent to BC,

(iv) C' is in T.

We assume that these conditions are fulfilled.

Let γ' be the curve obtained from γ by replacing $FC \cup CE$ by $\widehat{EF}$ and replacing $\widehat{HG}$ by $HC' \cup C'G$. Let R' denote the region whose boundary is γ'. Since the figure whose boundary is $HC' \cup C'G \cup \widehat{GH}$ is congruent to the figure whose boundary is $FC \cup CE \cup \widehat{EF}$, R and R' have the same area and the same perimeter. But since EC is not tangent to $\widehat{EF}$, $C'G$ is not tangent to $\widehat{GH}$, so it is not tangent to β. Thus, the interior angle of γ' at G is greater than 180°; so R' is not convex. This contradicts R being a solution to the problem. Therefore no vertex of T is on γ.

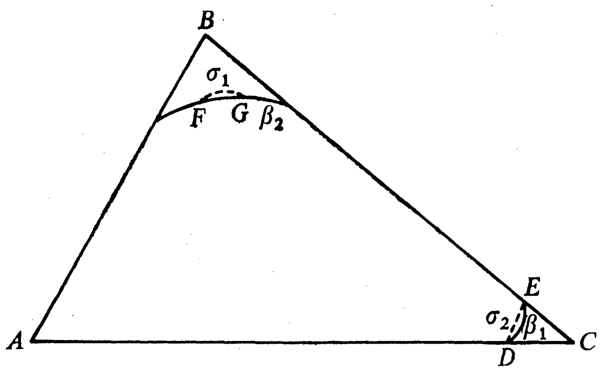


FIG. 9

We next show that all circular arcs contained in γ have equal radii. Suppose γ contains two arcs β_1 and β_2 with unequal radii, say the radius of β_1 is r_1 and the radius of β_2 is r_2 with $r_1 < r_2$ (Figure 9). Choose points D and E on β_1 and F and G on β_2 so that the chords DE and FG satisfy $l(DE) = l(FG)$. Construct arc σ_1 from F to G with radius r_1 and arc σ_2 from D to E with radius r_2. We assume that D, E, F, and G have been chosen in such a way that σ_1 and σ_2 are contained in T. Let γ' be the curve obtained from γ by replacing $\widehat{FG}$ contained in β_2 by σ_1 and replacing $\widehat{DE}$ contained in β_2 by σ_2. Let R' be the region whose boundary is γ'. Then R and R' have the same area and perimeter. But the interior angle of γ' at F (and the one at G) is greater than 180°; so R' is not convex. This again contradicts R being a solution to the problem. Therefore, every circular arc contained in γ has the same radius.

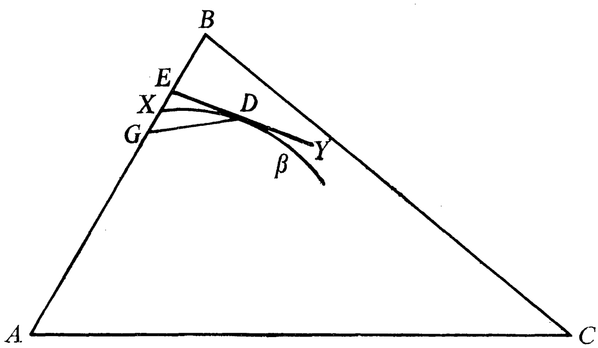


FIG. 10

Finally, we show that each circular arc contained in γ is tangent to each of the sides of T which it meets. Again let R be a solution to the problem and let γ be its boundary. Let β be a circular arc contained in γ which meets side AB of T at a point X and assume that β is not tangent to AB at X (Figure 10). Choose a point D on β (near X) and draw EY tangent to β at D meeting AB at E. Since β is not tangent to AB at X, $l(EX) < l(ED)$. Choose G on EA such that $l(EX) < l(EG) < l(ED)$. Then in ΔEDG, $m(\angle EGD) > m(\angle EDG)$. If we move along β from D to X and then continue along γ, we may move along XG since this segment may be contained in γ, or at some point of XG there may be an endpoint of another circular arc contained in γ. In either case, γ intersects DG in exactly one point other than D. Denote this point of intersection by H. (Then H may or may not coincide with G.) Let HK be the left-hand tangent to γ at H. Then $m(\angle DHK) \geqq m(\angle DGE) > m(\angle EDG)$. Reflect $(DH)_\gamma$, the part of γ inside and on ΔEDG and the left-hand tangent to γ at H, namely HK, in the perpendicular bisector of HD (Figure 11). Let K' denote the image of K. Let γ' denote the curve obtained from γ by replacing $(DH)_\gamma$ by its image under the reflection, and let R' be the region whose boundary is γ'. Again since reflection preserves distance, R and R' have the same area and perimeter. Since DK' is the right-hand tangent to γ' at D; DY is the left-hand tangent to γ' at

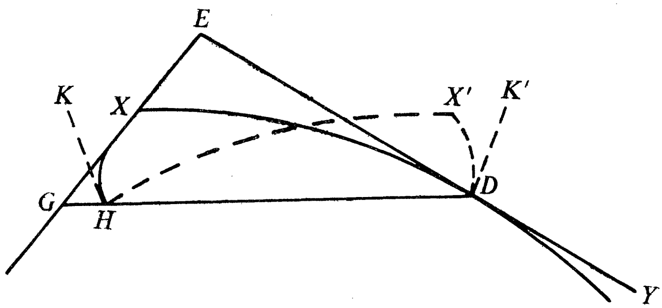


FIG. 11

D; and $m(\angle K'DG) = m(\angle KHD) > m(\angle EDH)$; and $m(\angle EDH) + m(\angle HDY) = 180°$, $m(\angle K'DH) + m(\angle HDY) > 180°$. Thus, the interior angle of γ' at D is greater than 180°, so R' is not convex. But this contradicts R being a solution to the problem. Therefore β is tangent to AB at X. Thus, each circular arc contained in γ is tangent to each of the sides of T which it meets.

We have now shown explicitly that the boundary γ of a solution R must satisfy all of the conditions stated in the solution except that the intersection of γ with each side of the boundary of T is a line segment and that there are precisely three circular arcs contained in γ. We shall show that these properties are implied from what we have already proved. Since R is convex, the intersection of γ with a side of T must be empty, a single point, or a line segment. Suppose that the intersection of γ with some side, say AB, of T is either empty or is a single point. Since A and B are not on γ and since circular arcs contained in γ are tangent to the sides of T which they meet, there exists an arc β contained in γ which is tangent to AC and to BC and whose center is on the side of β away from A and B and such that β misses side AB or is tangent to side AB. Since C is not on γ, there is another arc ω tangent to AC and to BC whose center is on the side of ω away from C. Then ω and β must have the same radius. Since both arcs are tangent to AC and to BC, this implies that they are tangent at the same points and hence $\beta \cup \omega$ is a circle. But the radius of this circle is at most c and this contradicts γ having length $p > c$. Therefore, the intersection of γ with each side of T is a line segment. For this to be the case, since no vertex is on γ, it is necessary that γ contain precisely three circular arcs. This completes the proof.

REMARK 1. It may be of interest to note that the three segments of the sides of T which are contained in γ are in the ratio $l(AB)$: $l(BC)$: $l(CA)$. If we draw a circle β tangent to AB and to AC and having the same radius as the arcs contained in γ, and if we draw DE tangent to β and parallel to BC, and draw EF parallel to AB; then the three arcs of β determined by its points of tangency with AD, DE, and EA are congruent to the three arcs contained in γ, and the line segments EF, FC, and CE are congruent to the segments of AB, BC, and CA, respectively, contained in γ (Figure 12). The area of the solution R is the sum of the area of the circle β, the parallelogram $DEFB$, and the triangle EFC.

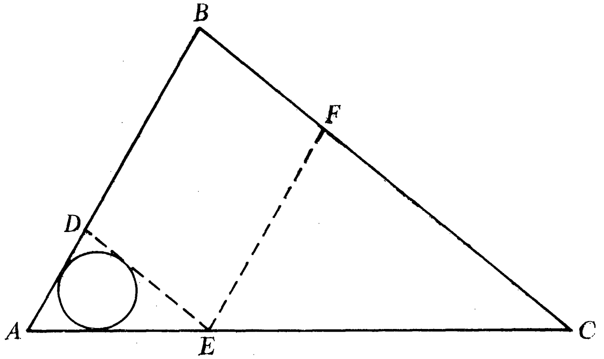


FIG. 12

REMARK 2. If the triangle T in Problem 3 is replaced by any convex polygon, it is clear that the proofs of the properties of γ still hold except that the intersection of γ with a given side of the polygon may be a single point or empty. However, it is still true that γ is composed of circular arcs of equal radii each tangent to each side that it meets together with line segments contained in sides of the given polygon.

References

1. C. F. Adler, An isoperimetric problem with an inequality, Amer. Math. Monthly, 52 (1945) 59–69.
2. J. L. Coolidge, A History of Geometrical Methods, Oxford University Press, Oxford, 1940.
3. M. A. Krasnoselskii and Ya. B. Rutickii, Convex Functions and Orlicz Spaces, Noordhoff Ltd., Groningen, Netherlands, 1961.
4. L. A. Lyusternik, Convex Figures and Polyhedra, Dover, New York, 1963.
5. J. R. Newman, ed., The World of Mathematics, vol. 2, Simon and Schuster, New York, 1956.
6. J. Steiner, Sur le maximum et le minimum des figures dans le plan, sur la sphère, et dans l'espace en général, I and II, J. Reine Angew. Math. (Crelle), 24 (1842) 93–152 and 189–250.

VENN DIAGRAMS AND INDEPENDENT FAMILIES OF SETS

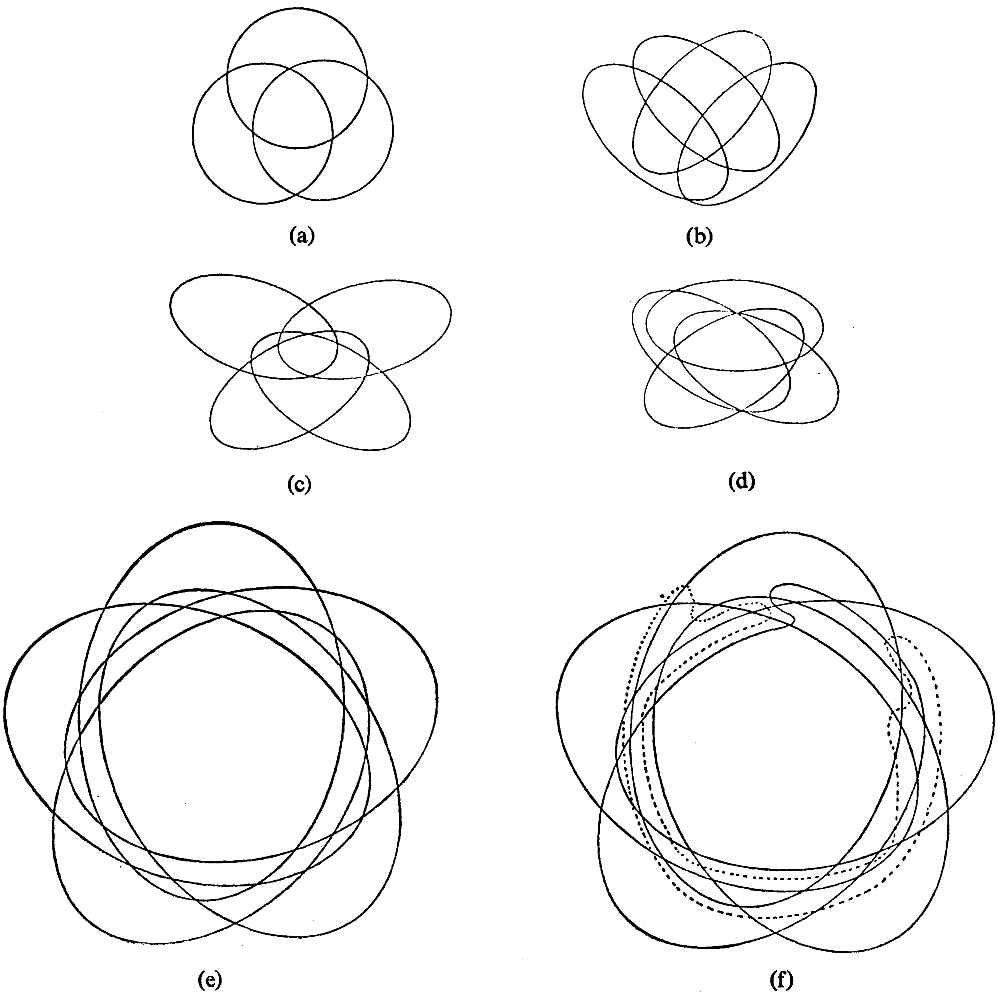
BRANKO GRÜNBAUM, University of Washington, Seattle

1. Introduction. Let $\mathscr{A} = \{A_1, \cdots, A_n\}$ be a family of n simple closed curves in the Euclidean plane. We shall say that $\mathscr{A}$ is an *independent family* provided the intersection

$$(*) \qquad X_1 \cap X_2 \cap \cdots \cap X_n \text{ is nonempty}$$

whenever each set X_j is chosen to be either the interior or the exterior of the curve A_j.

For example, the family of 3 circles and the families of 4 or 5 congruent ellipses in Figure 1 are independent.



(a) (b) (c) (d) (e) (f)

FIG. 1

Independent families and their higher dimensional analogues are interesting from several points of view; for example, Marczewski [1947] considers them in connection with certain extension problems in measure theory. We are more interested in their geometric properties, and in particular with the special types of independent families that are usually called *Venn diagrams*. Introduced by Venn [1880], the Venn diagrams are meant to facilitate and make visually accessible the relations of classes in first-order predicate calculus; various markings, hatching, etc., that are occasionally used for such purposes are of no importance for our discussions. Venn diagrams have been widely popularized by many texts on elementary logic and set theory, starting with Venn [1881]; among more modern ones we may mention Quine [1950], Birkhoff-MacLane [1953], Rosser [1953], Suppes [1957], Kenelly [1967], Gardner [1968], Roethel-Weinstein [1972]. Unfortunately and quite ironically, logicians and set theorists are mostly not interested in geometry and as a consequence their definitions are rather vague as to the precise requirements on an independent family to be considered a Venn diagram. In many of the books the discussion is limited to $n \leqq 3$, and a few even express doubts in the possibility or the feasibility of Venn diagrams with large n. Moreover, insofar as the definitions can be interpreted at all, different authors appear to have different limitations in mind when speaking of Venn diagrams, and even individual authors seem to vary their interpretations from instance to instance. For example, Venn [1880; 1881] seems to require that each A_j be a simple closed curve, but one of the examples he gives for $n = 5$ fails to have that property. In Austin [1971] each of the sets in (*) is required to be connected, but the last two illustrations fail to have that property. While Austin [1971] requires that no point should belong to 3 of the curves A_j, Henderson [1963] definitely allows that.

We find it convenient, and agreeing in spirit with Venn and many other authors, to define a *Venn diagram* of n curves in the plane as an independent family $\mathscr{A} = \{A_1, \cdots, A_n\}$ such that each of the sets $X_1 \cap X_2 \cap \cdots \cap X_n$ in (*) is an open, connected region, called a *cell* of the Venn diagram. It is easy to verify that all the bounded cells (that is, all cells except the single unbounded one) are even simply connected, as is also the complement of the unbounded cell.

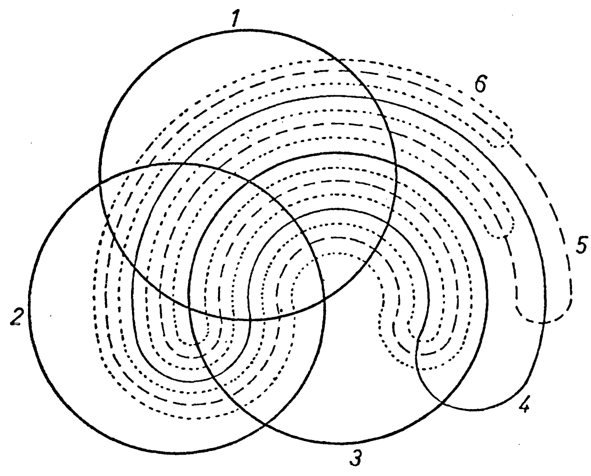


FIG. 2

The independent families in Figure 1 are easily seen to be Venn diagrams, but it is not completely obvious that Venn diagrams exist for arbitrarily large n. However, as observed already by Venn [1880, p. 8] (see also Venn [1881], Baron [1969]) a simple inductive construction may be used to establish the existence of such Venn diagrams. In Figure 2 we show Venn's inductive construction of diagrams with up to 6 curves, which should suffice to indicate the pattern Venn [1880] had in mind. Other constructions with the same purpose have been described by Berkeley [1937], More [1959], and Bowles [1971].

In the next section we shall investigate a number of questions dealing with independent families and Venn diagrams from the point of view of combinatorial geometry; two of them correct erroneous assertions found in the literature. In Section 3 we shall discuss a number of related results and open problems.

2. Venn diagrams formed by convex curves. We start with a simple observation which has several interesting consequences.

LEMMA. *If an independent family of n curves is such that each two curves meet in at most j points, then*

$$j \geqq (2^n - 2) \Big/ \binom{n}{2} = 4(2^{n-1} - 1)/n(n-1).$$

Proof. If no 3 curves have a common point, then the curves define in the plane a network with $v \leqq j\binom{n}{2}$ nodes (intersection points) and $e = 2v$ edges (arcs); therefore, by the Euler relation, the number of regions determined by this network is $e - v + 2 \leqq j\binom{n}{2} + 2$. But if the curves form an independent family the number of regions is at least 2^n, and the assertion follows for this case. The general case may be reduced to the one just considered by observing the possibility of deforming the curves so that each two still meet in at most j points, no three have a point in common, and no region is obliterated.

We may remark that the Lemma remains valid even if its assumptions are weakened to require each two curves to have an intersection consisting of at most j connected components.

If C is a convex curve (that is, the boundary of a 2-dimensional compact convex subset of the plane) we shall denote by $s(C)$ the largest number of curves in an independent family consisting of curves similar to C, and by $s^*(C)$ the analogously defined number for Venn diagrams. By $h(C)$ and $h^*(C)$ we shall denote the numbers similarly defined for families consisting of curves positively homothetic to C. Clearly $s(C) \geqq s^*(C) \geqq h^*(C)$ and $s(C) \geqq h(C) \geqq h^*(C)$ for all C.

THEOREM 1. (i) *If C is a circle then $s(C) = s^*(C) = 3$.* (ii) *For every ellipse E we have $s(E) = s^*(E) = 5$.* (iii) *For every convex curve C we have $h(C) = h^*(C) = 3$.*

Proof. Two circles intersect in at most 2 points, two ellipses in at most 4 points, and the intersection of two positively homothetic convex curves is well known to have at most 2 connected components. Therefore, we may apply the Lemma with $j = 2$, 4, and 2, respectively, and find that $s(C) < 4$, $s(E) < 6$, and $h(C) < 4$. Figure 1(a) shows that $s^*(C) \geqq 3$ for a circle C, and it is easy to construct similar examples that establish (iii). The Venn diagram in Figure 1(e) shows $s^*(E) \geqq 5$; similar diagrams may be constructed using five copies of any noncircular ellipse. This completes the proof of Theorem 1.

Remarks. The first part of Theorem 1 has been well known since Venn [1880]. However, the assertion that $s^*(E) = 4$ for ellipses E,—which is contradicted by part (ii) of Theorem 1 and by the diagram in Figure 1(d)—was also made by Venn

[1880] and repeated by many authors; among others in the article *Logic Diagrams* in Edwards [1967]. For a weak version of part (iii) see Austin [1971].

Let *k-gon* designate any convex polygon with at most k sides. We shall denote by $n(k)$ the maximal number of members in any independent family of k-gons in the plane, and by $k(n)$ the minimal k such that there exists an independent family of n k-gons. The similarly defined numbers dealing with Venn diagrams shall be denoted $n^*(k)$ and $k^*(n)$. We have

THEOREM 2. $$\lim_{k\to\infty} \frac{n(k)}{\log_2 k} = \lim_{k\to\infty} \frac{n^*(k)}{\log_2 k} = 1.$$

Proof. Since two k-gons intersect in at most $2k$ connected components, we may apply the Lemma with $j = 2k$ and deduce $\lim_{k\to\infty} n(k)/\log_2 k \leqq 1$. In order to complete the proof we shall describe the construction of a Venn diagram $\mathscr{A}(n)$ of n k-gons showing that $n^*(k) \geqq n = 2 + \log_2 k$. For $n = 4$ we take the diagram $\mathscr{A}(4)$ consisting of four quadrangles indicated by solid lines in Figure 3; we denote by $P(4)$ the quadrangle $ABCD$. For $n > 4$ we obtain $\mathscr{A}(n)$ by adding a polygon $P(n)$ to $\mathscr{A}(n-1)$. The polygon $P(n)$ is a 2^{n-2}-gon formed by crossing twice each of the 2^{n-3} sides of $P(n-1)$, with vertices situated on the polygons crossing $P(n-1)$ and sufficiently close to $P(n-1)$ to make $P(n)$ a convex 2^{n-2}-gon. (Compare Figure 3 in which the dashed polygon illustrates $P(5)$, the dotted one $P(6)$.) This completes the proof of Theorem 2.

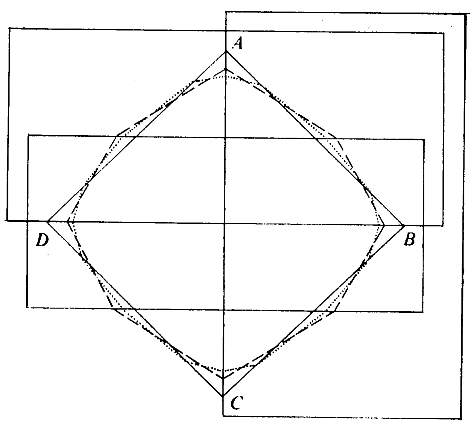


FIG. 3

Remarks. Theorem 2 appears in Rényi-Rényi-Surányi [1951]; their proof of the second part uses a different construction than the one just described, and yields a poorer bound. More importantly, however, Rényi-Rényi-Surányi [1951] base their proof of the upper bound on the statement obtained from our lemma by replacing the inequality of its conclusion by the stronger relation

(**) $$j \geqq \frac{2^{n-1}}{n-1}.$$

Unfortunately, the proof they give for (**) is fallacious, and the assertion itself is false. A counterexample to (**) with $j = n = 6$ is shown in Figure 1(f). However, the Rényi-Rényi-Surányi proof of Theorem 2 could be salvaged if relation (**) is true or independent families of n k-gons, with $j = 2k$. If this more restricted variant of (**) were established, equality could be substituted for the inequalities in the following result; but we expect that equality holds in Theorem 3 in any case.

THEOREM 3. $k(3)=k^*(3) = 3$; $k(4) = k^*(4)=3$; $k(5) = k^*(5) = 3$; $k(6) \leqq k^*(6) \leqq 4$; $k(7) \leqq 6$.

Proof. The Venn diagrams in Figures 6(a), 8(b), and 4 establish the first three parts of the theorem. A Venn diagram of six quadrangles is shown in Figure 5, while a drawing of an independent family of 7 hexagons will be supplied b the author to any interested reader.

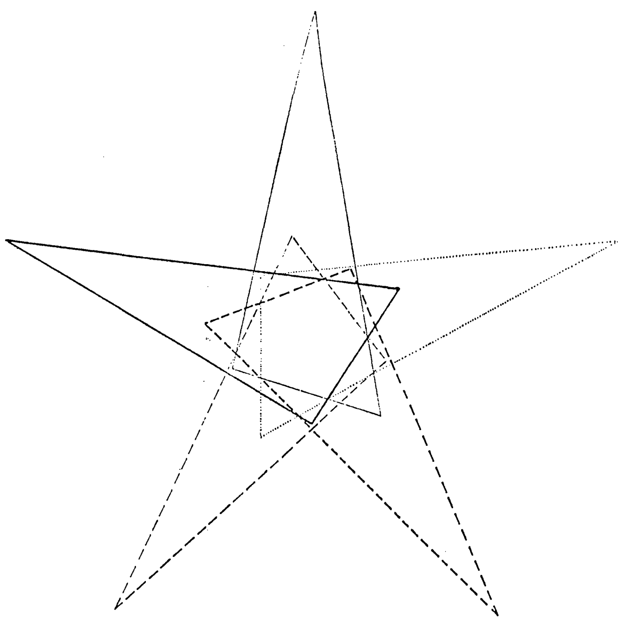


FIG. 4

3. Variations and open problems. (1) The construction we used in the proof of Theorem 2 is a modification to convex polygons of the method described in More [1959]. It may be adapted to obtain other variants of the result, such as the following:

(i) If the construction is not required to proceed beyond $\mathscr{A}(n)$, then it is easily seen that $P(n)$ may even be chosen as a 2^{n-3}-gon; therefore we actually have $n(k) \geqq n^*(k) \geqq 3 + \log_2 k$.

(ii) Let a Venn diagram $\mathscr{A} = \{A_1, \cdots, A_n\}$ be called *convex* provided all the A_j are convex curves and all the bounded cells of $\mathscr{A}$ are convex sets as is the complement of the unbounded cell. We shall denote by $n^{**}(k)$ and $k^{**}(n)$ the numbers corre-

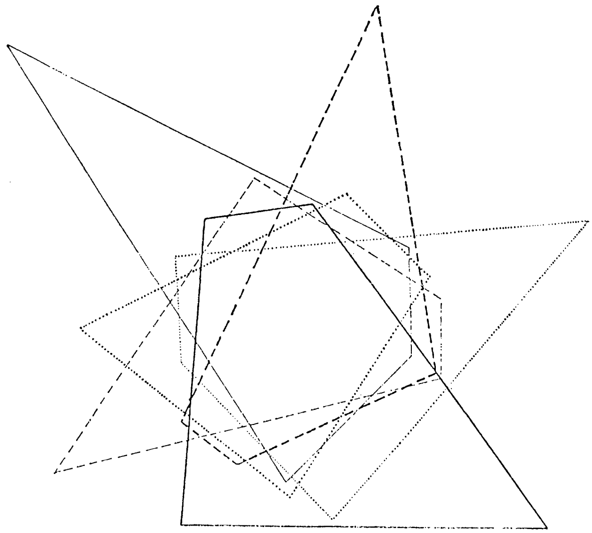
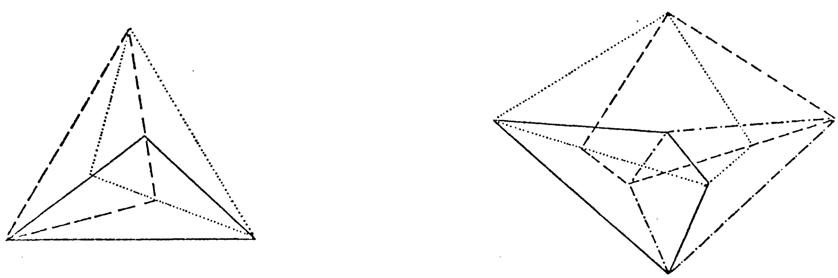


FIG. 5



(a) (b)

FIG. 6

sponding to $n^*(k)$ and $k^*(n)$ but referring to convex Venn diagrams. (It should be noted that the existence of $k^{**}(n)$ and of convex diagrams with arbitrarily large numbers of curves is not self-evident. While the existence of Venn diagrams for arbitrary n in which each cell is convex may easily be deduced from Steinitz's theorem on convex polyhedra (see, for example, Grünbaum [1967, Chapter 13]), the satisfaction of the additional requirement of convexity of the curves A_j cannot be guaranteed by that theorem.) But an easy modification of our construction may be used to establish $n^{**}(k) \geqq 2 + \log_2(k-1)$, so that Theorem 2 may be extended to $\lim_{k\to\infty} n^{**}(k)/\log_2 k = 1$. The diagrams in Figure 6 prove that $k^{**}(3) = 3$ and $k^{**}(4) \leqq 4$; actually it may be shown that $k^{**}(4) = 4$, and it is probable that $k^{**}(5) = 5$.

(2) A Venn diagram with n curves is called *symmetric* (Henderson [1963]) if all the curves are congruent and obtained from each other by rotations through multiples of $2\pi/n$ about a fixed point. The diagram of $n = 5$ triangles in Figure 4 is symmetric. Henderson [1963] gives two examples of symmetric diagrams with

$n = 5$, one formed by pentagons, the other by quadrangles; it is easy to modify them so as to obtain symmetric Venn diagrams consisting of triangles, combinatorially distinct from each other and from the one in Figure 4. (It is not known whether every symmetric Venn diagram of 5 triangles is isomorphic with one of these three.) As observed by Henderson [1963], if a symmetric Venn diagram has n curves then n is necessarily a prime number. Henderson [1963] mentions that he has found a symmetric Venn diagram of 7 hexagons. The present author's search for such a diagram has been unsuccessful, as have been attempts to clarify Henderson's claim; at present it seems likely that no such diagram exists. Concerning the analogously defined concept of a *symmetric independent family* of n curves there is no restriction that n be prime. In Figure 8(a) we have a symmetric independent family of 4 triangles (compare Austin [1971]), while Figure 7 shows a symmetric independent family of 6 quadrangles. The independent family of 7 hexagons mentioned in the proof of Theorem 3 is also symmetric. It is not known whether for every n there exists a symmetric independent family of n $k(n)$-gons; we conjecture that that is the case. But it should be stressed that the existence of symmetric Venn diagrams is still open for each prime $n \geqq 7$.

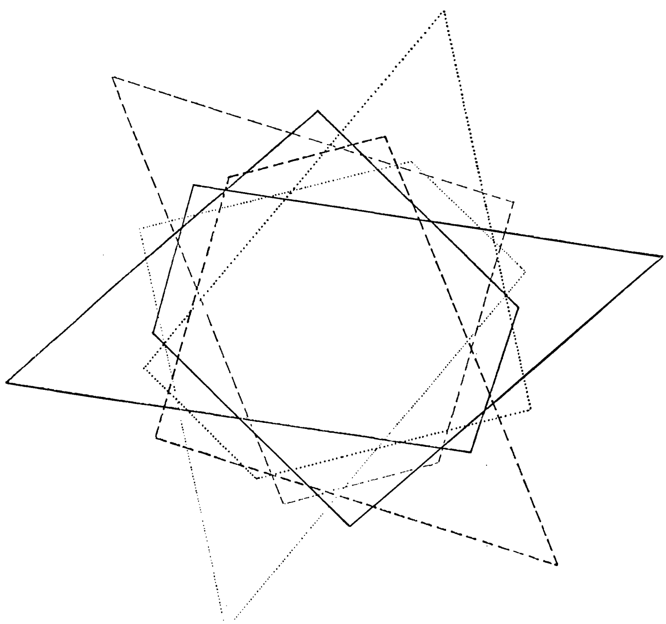
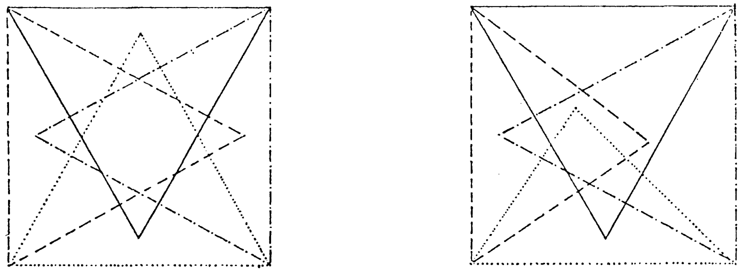


FIG. 7

(3) An independent family, or a Venn diagram, is called *simple* provided no point belongs to the boundary of three of the sets. Let $k_s(n)$ [or $k_s^*(n)$] be the smallest k for which there exists a simple independent family [or a simple Venn diagram] with n k-gons. The example in Figure 6 shows that $k_s^*(3) = 3$, and Figure 8(a) establishes $k_s(4) = 3$; the example in Figure 8(b) proves even $k_s^*(4) = 3$. No other values seem to be known.



(a) (b)

FIG. 8

(4) Concerning the functions $k(n)$ and $k^*(n)$ we venture two guesses:

(i) $k(n) = k^*(n)$ for all n.

(ii) $k(n) =]2^{n-2}/(n-1)[$ for all n, where $]x[$ denotes the smallest integer $\geqq x$.

(5) Notions of independent families and Venn diagrams may easily be extended to higher dimensions in several ways (see, for example, Weglorz [1964], Collings [1972]). We prefer the following definitions, although for $d = 2$ they are somewhat more general than the definitions we adopted in Section 1 (all the results of Section 2 remain valid with these definitions). A family $\mathscr{A} = \{A_1, \cdots, A_n\}$ of n sets in the d-dimensional Euclidean space E^d is *independent* provided (*) holds whenever X_j is chosen to be either A_j or the complement $\sim A_j$ of A_j in E^d. The independent family $\mathscr{A}$ is a *Venn diagram* provided each of the sets int $(X_1 \cap X_2 \cap \cdots \cap X_n)$ is an open d-cell (that is, is homeomorphic to the open d-ball), except that int $(\sim A_1 \cap \sim A_2 \cap \cdots \cap \sim A_n)$ is homeomorphic to the complement of a closed d-ball B^d.

The following generalization of part (i) of our Theorem 1 holds: Each independent family of d-balls has at most $d+1$ members. More precisely, using the obvious adaptation of the notation introduced in Section 2, we have $s(B^d) = s^*(B^d) = d+1$. This was established by Rényi-Rényi-Surányi [1951] (see also Anusiak [1965]), together with the theorem: *An independent family of boxes in E^d, with edges parallel to the coordinate axes, contains at most $2d$ members.*

(6) Let $k(n,d)$ denote the least k such that there exists an independent family of n sets in E^d, each set of which is a d-polytope with at most k facets ($(d-1)$-dimensional faces). Theorem 3 shows that $k(5,2) = 3$, $k(6,2) \leqq 4$ and $k(7,2) \leqq 6$, while the result on boxes in E^d mentioned above shows $k(2d,d) \leqq 2d$. Starting from the Venn diagram of 5 triangles shown in Figure 4 it is easy to deduce that $k(2d+1,d) = d+1$. If we denote by n_d the largest n such that $k(n,d) = d+1$ then by the remark just made $n_d \geqq 2d+1$. On the other hand, crude upper bounds on n_d may be obtained by using the well-known result (see Grünbaum [1971] for details and references) that j hyperplanes in E^d determine at most $\binom{j-1}{d}$ bounded cells. Therefore n_d must satisfy the inequality

$$\binom{(d+1)n_d - 1}{d} \geqq 2^n - 1.$$

Since $\binom{23}{2} = 253 < 255 = 2^8 - 1$, this implies that $n_2 \leqq 7$ (compared to the conjectured value $n_2 = 5$). Similarly, since $\binom{59}{3} = 32{,}509 < 32{,}767 = 2^{15} - 1$, we have $7 \leqq n_3 \leqq 14$, and analogously $9 \leqq n_4 \leqq 22$. It would be interesting to learn more about n_d and $k(n, d)$.

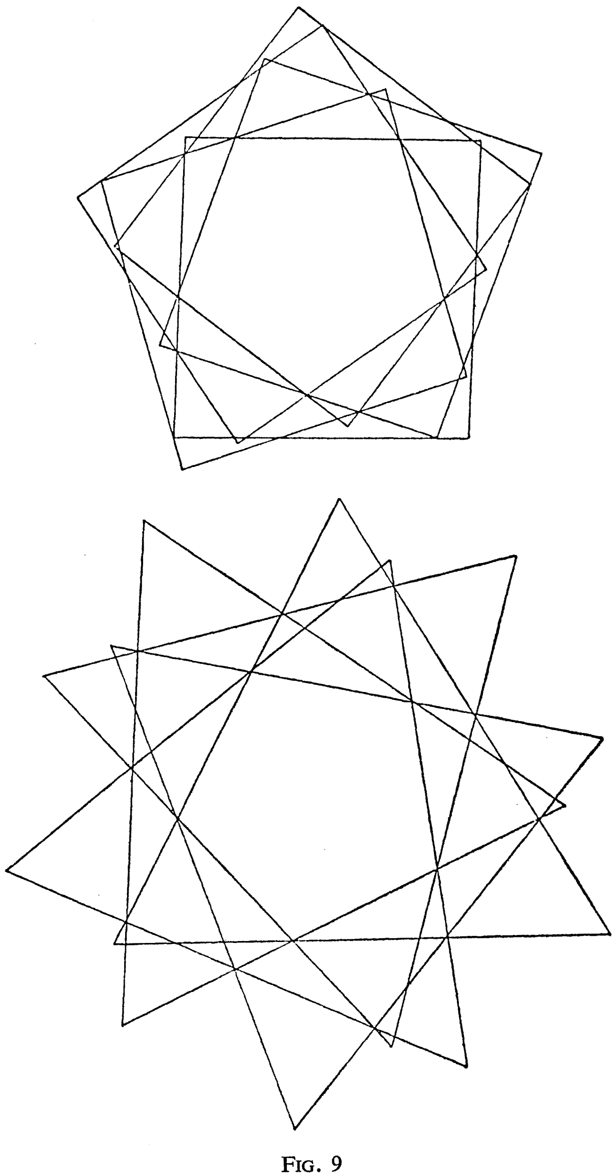


FIG. 9

(7) We denote by $h(C)$ and $h^*(C)$ the maximal number of sets in independent families or Venn diagrams consisting of sets homothetic to the convex set C. As a generalization of part (iii) of Theorem 1, and of the result of Rényi-Rényi-Surányi [1951] on balls, we conjecture that $h(C) = h^*(C) = d + 1$ for every d-dimensional

compact convex set C. (If negative ratios of homothecy are permitted, part (iii) of Theorem 1 does not hold; it would be interesting to determine the bounds in that case as well.)

(8) Another particularly attractive type of Venn diagrams or independent families comprises those consisting of congruent sets. We denote by $c(C)$ [or $c^*(C)$] the maximal number of sets in an independent family [or Venn diagram] consisting of sets congruent to C. From Theorem 1 and Figure 1(d) it follows that $c^*(E) = c(E) = 5$ for each ellipse E. It is easily seen that there is no fixed bound on $c(C)$ when C varies over planar convex sets. It would be of interest to determine whether—as seems likely—there exist planar convex sets C with $c(C) = \infty$, or even such with $c^*(C) = \infty$. For each polygon C the finiteness of $c(C)$ follows from the Lemma of Section 2. The examples in Figure 9 show that $c(T) \geqq 5$ and $c(S) \geqq 5$ if T is an equilateral triangle and S is a square. Probably equality holds in both cases. Other open problems concern the values of $c^*(T)$ and $c^*(S)$, as well as their analogues in higher dimensions.

(9) The author hopes that the facts and the spirit of the above discussion will be of some assistance to those contending with the unfortunate—though fashionable—opinions that "geometry is dead," that it is "reduced to linear algebra," or that one needs years of postgraduate study to understand nontrivial open problems in geometry. Many questions of combinatorial geometry provide not only opportunities to exercise the geometric intuition and a variety of techniques, but also afford considerable aesthetic gratification.

Research supported in part by the Office of Naval Research under Grant N00014-67-A-0103-0003.

References

J. Anusiak, 1965, On set-theoretically independent collections of balls, Colloq. Math., 13 (1965) 223–233.

A. K. Austin, 1971, Problem E 2314, Amer. Math. Monthly, 78 (1971) 904. Solution, ibid. 79 (1972) 907–908.

M. E. Baron, 1969, A note on the the historical development of logic diagrams: Leibniz, Euler and Venn, Math. Gaz., 53 (1969) 113–125.

E. C. Berkeley, 1937, Boolean algebra (the technique for manipulating "and", "or", "not" and conditions) and applications to insurance, The Record, Amer. Inst. of Actuaries, 26 (1937) 373–414.

G. Birkhoff and S. MacLane, 1953, A Survey of Modern Algebra, Macmillan, New York, 1953. 3rd ed., 1965.

L. J. Bowles, 1971, Logic diagrams for up to n classes, Math. Gaz., 55 (1971) 370–373.

S. N. Collings, 1972, Further logic diagrams in various dimensions, Math. Gaz., 56 (1972) 309–310.

P. Edwards (editor), 1967, Encyclopedia of Philosophy, Macmillan, New York, 1967, Logic diagrams by M. Gardner, Vol. 5, pp. 77–81.

M. Gardner, 1968, Logic Machines, Diagrams and Boolean Algebra, Dover, New York, 1968.

B. Grünbaum, 1967, Convex Polytopes, Interscience, New York, 1967. 1971, Arrangements of hyperplanes, Proc. Second Louisiana Conference on Combinatorics, Graph Theory and Computing, R. C. Mullin et al, ed., Louisiana State University, Baton Rouge, 1971, pp. 41–106.

D. W. Henderson, 1963, Venn diagrams for more than four classes, Amer. Math. Monthly, 70 (1963) 424–426.

J. W. Kenelly, 1967, Informal Logic, Allyn and Bacon, Boston 1967.

E. Marczewski, 1947, Indépendance d'ensembles et prolongements de mesures, Colloq. Math., 1 (1947) 122–132.

T. More, Jr., 1959, On the construction of Venn diagrams, J. Symbolic Logic, 24 (1959) 303–304.

W. V. O. Quine, 1950, Methods of Logic, Holt, Rinehart and Winston, New York, 1950. (Rev. 1959.)

A. Rényi, C. Rényi and J. Surányi, 1951, Sur l'indépendance des domaines simples dans l'espace euclidien à n dimensions, Colloq. Math., 2 (1951) 130–135.

L. F. Roethel and A. Weinstein, 1972, Logic, Sets, and Numbers, Wadsworth, Belmont, Calif., 1972.

J. B. Rosser, 1953, Logic for Mathematicians, McGraw-Hill, New York 1953.

P. Suppes, 1957, Introduction to Logic, Van Nostrand, Princeton 1957.

J. Venn, 1880, On the diagrammatic and mechanical representation of propositions and reasonings, The London, Edinburgh, and Dublin Philos. Magazine and Journal of Science, (5) 9 (1880) 1–18. 1881, Symbolic Logic, Macmillan, London, 1881. (Sec. ed. 1894.)

B. Weglorz, 1964, Nerves and set-theoretical independence, Colloq. Math., 13 (1964) 17–19.

LAGUERRE'S AXIAL TRANSFORMATION

DAN PEDOE, University of Minnesota

This transformation, as we remarked in a recent paper (Pedoe, [3]) has been mostly forgotten by geometers. It is as powerful a tranformation as inversion (transformation by reciprocal radii), being, in a sense, the dual of inversion, and was named by Laguerre *transformation par les semi-droites réciproques*. We shall call Laguerre's half-lines *rays*. They are oriented lines, and a ray touches a *cycle*, an oriented circle, at a point P if the direction of the ray at P coincides with that assigned to the cycle.

The geometry of cycles was largely developed during the 19th century, but never met with the enthusiasm accorded to the geometry of circles, and has been almost completely forgotten, although the theory contains many beautiful theorems. In this paper we have another look at Laguerre's fundamental theorem:

Given three cycles whose axis of similitude intersects none of them, a Laguerre axial transformation can be found which simultaneously maps the three cycles into three points.

In Pedoe [3] we investigated Laguerre's geometrical proof of this theorem and remarked that neither Blaschke [1], Coolidge [2] or Yaglom [5] mention it specifically in their respective discussions of axial transformations. Blaschke comes very near, as we shall see (§8), and it is his discussion of Sophus Lie's approach to the subject which has motivated this paper, in which we connect Lie's work with Laguerre's. Some unexpected views of both 2-dimensional and 3-dimensional geometry will appear during this excursion in Euclidean 3-space.

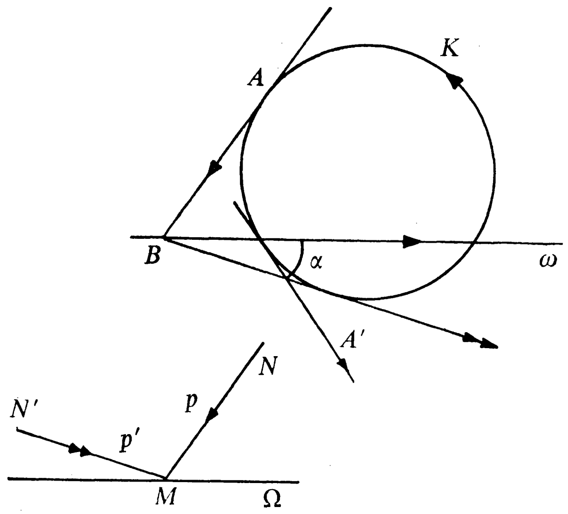


FIG. 1

1. In the Laguerre transformation a cycle K and two parallel lines, Ω, ω are given (Figure 1). The line ω is taken by Laguerre to intersect K. Given a ray NM which intersects Ω in M, there is a unique parallel ray AB which touches K at a point A and intersects ω in B. Through B there is a unique ray which touches K at a point A'. The ray $N'M$ parallel to BA' is the transform (map) of the ray NM.

The transform of $N'M$ is evidently NM, so that the transformation is involutory. Laguerre shows that the set of rays tangent to a given cycle $\mathscr{C}$ maps onto the set of rays tangent to a cycle $\mathscr{C}'$, and this at once makes the transformation an interesting one. We note that if two rays tangent to a cycle $\mathscr{C}'$ lie along the same line, but in opposite directions, the cycle $\mathscr{C}'$ must be a point, and Laguerre uses this result most effectively in proving his theorem (Pedoe, **3**, p. 262).

In contrast to circles, two given cycles $\mathscr{C}$ and $\mathscr{D}$ have at most two common tangent rays, and these naturally pass through a center of similitude on the line of centers of the cycles. If we consider three cycles $\mathscr{C}$, $\mathscr{D}$ and $\mathscr{E}$, each pair possessing common tangent rays, the three corresponding centers of similitude are collinear.

This can be seen immediately from the geometrical representation of cycles introduced by Chasles, which makes it possible to produce the Laguerre transformation as the dual of inversion, and is basic to the Lie approach.

2. In a rectangular coordinate system (X, Y, Z) we suppose our cycles to lie in $Z = 0$. We call our Euclidean space E_3, and if we adjoin the plane at infinity, which we call $T = 0$, the resulting projective space will be called S_3. The cycle center (p, q) in $Z = 0$ and radius R is mapped onto the point (p, q, R) in E_3 if the cycle we are considering is traversed positively, and onto the point $(p, q, -R)$ if it is traversed negatively. The right circular cones with vertex at either point which contain the given cycle in $Z = 0$ have semivertical angle 45°, and intersect the plane at infinity,

$T = 0$, in the fixed irreducible conic

$$\psi: X^2 + Y^2 - Z^2 = 0.$$

Tangent planes to these cones intersect $T = 0$ in tangents to ψ. This conic is fundamental to our discussion.

3. If cycles $\mathscr{C}$, $\mathscr{D}$ are represented in E_3 by the points P, Q, then there are at most two common tangent planes through the line PQ to the right circular cones with vertices P, Q which pass through the respective cycles $\mathscr{C}$, $\mathscr{D}$. These tangent planes intersect $Z = 0$ in the two tangent rays common to $\mathscr{C}$ and $\mathscr{D}$, and the line PQ intersects $Z = 0$ in the center of similitude of the two cycles. On the other hand the two tangent planes intersect $T = 0$ in the two tangent lines which can be drawn to the conic ψ from the point in which the line PQ intersects $T = 0$. For details of these observations see Pedoe (**4**, p. 428).

If P, Q, R represent the respective cycles $\mathscr{C}$, $\mathscr{D}$ and $\mathscr{E}$, the plane PQR intersects $Z = 0$ in the axis of similitude of the three cycles. If this axis does not intersect the cycle $\mathscr{C}$, say, the perpendicular distance from P', the center of the cycle $\mathscr{C}$, onto the axis of similitude, which is the intersection of the plane PQR with $Z = 0$, is greater than the distance $P'P$, which is the radius of $\mathscr{C}$. It follows that the *angle* made by the plane PQR with $Z = 0$ is then less than 45°. This remark will be of importance in establishing Laguerre's theorem (§9). Conversely, if the plane PQR makes an angle with $Z = 0$ which is less than 45°, the axis of similitude of the three cycles cannot intersect any of the given cycles, so that if it does not intersect one, it cannot intersect the others.

4. In the representation we are discussing there is a sharp distinction to be observed between planes which make an angle less than 45° with $Z = 0$ and those whose angle with $Z = 0$ exceeds 45°. Because of the connection with the special theory of relativity (Blaschke, **1**, p. 141), the former are called *space-like*, and the latter *time-like*. Planes which make an angle of 45° with $Z = 0$ are called *isotropic*. The same terms are used for lines.

There is an immediate interpretation of these terms if we consider the conic ψ in $T = 0$. The line in E_3 with direction-ratios $l : m : n$ intersects $T = 0$ in the point with homogeneous coordinates $(l, m, n, 0)$. If the line is space-like,

$$\left| n/(l^2 + m^2 + n^2)^{\frac{1}{2}} \right| < \cos(\pi/4),$$

so that

$$l^2 + m^2 - n^2 > 0.$$

We say that such points are *outside* the conic

$$\psi: X^2 + Y^2 - Z^2 = 0,$$

and observe that in fact the tangents to the conic from such an obvious outside point as $(1, 0, 0, 0)$, being given by the equation

$$(X^2+Y^2-Z^2)(1)-(X)^2=0$$

are the real pair of lines $Y^2-Z^2=0$.

Time-like lines intersect $T=0$ in points for which

$$l^2+m^2-n^2<0,$$

and these are points inside ψ, from which real tangents cannot be drawn to the conic. Isotropic lines intersect $T=0$ in points on ψ. Since two tangents can be drawn to ψ from a point outside the conic, and each tangent is the intersection with $T=0$ of an isotropic plane, it follows that *two isotropic planes can be drawn to pass through a given space-like line.* None can be drawn through a time-like line, and only one can be drawn through an isotropic line.

In a space-like plane all lines are space-like, since the line of greatest slope is space-like. Hence the line of intersection of a space-like plane with $T=0$ lies entirely outside the conic ψ. An isotropic plane contains an infinity of parallel isotropic lines, one through every point, and all other lines are space-like. A time-like plane meets $T=0$ in a line which intersects the conic ψ in two distinct points. Hence through every point of a time-like plane we can draw an infinity of space-like lines, and an infinity of time-like lines, and these sets of lines are separated by the two isotropic lines which also pass through the point.

If we consider the representation of the cycles which touch a given ray at a given point, we evidently obtain an isotropic line. The cycles which touch all points of the ray are represented by an isotropic plane which intersects $Z=0$ in the given ray, and in fact the point cycles on the ray are assumed to satisfy the tangency condition also. If the points P, Q represent cycles $\mathscr{C}$, $\mathscr{D}$, we can draw two isotropic planes through the line PQ if the line PQ itself is space-like. These planes intersect $Z=0$ in the common tangent rays of $\mathscr{C}$ and $\mathscr{D}$. Conversely, if $\mathscr{C}$ and $\mathscr{D}$ have two common tangent rays, the line PQ is space-like.

The final result needed from the theory of the representation of cycles concerns the system of cycles represented by a time-like plane. If we consider the cycles which intersect a given ray, at a given point A, at a fixed angle α, their representation is a line which makes an angle of 45° with the line in $Z=0$ along which the centers of the cycles move. If the point A varies on the ray, the representation of the cycles traces out a plane which intersects $Z=0$ along the ray, and since the lines which trace out the plane are not lines of greatest slope, this plane is time-like, making an angle $>45°$ with $Z=0$. Conversely, a time-like plane represents the system of cycles which intersect the ray of intersection of the plane with $Z=0$ at a fixed angle. The diagram which illustrates this is a favorite with writers on cycles. See Blaschke (**1**, p. 150), and Yaglom (**5**, p. 177).

5. We now discuss Lie's approach to all this. He considers the group of affine transformations in E_3 which not only map the plane $T=0$ onto itself, but also map the conic ψ onto itself. An affine map of E_3 induces a collineation in $T=0$, and since ψ is mapped onto itself, tangents to ψ are mapped onto tangents to ψ. Points in

$T = 0$ from which tangents can be drawn to ψ are mapped onto similar points. Thus the outside of ψ is mapped onto the outside, and the inside onto the inside. It follows that the type of affine map we are considering not only maps an isotropic line of E_3 onto an isotropic line, but also maps a space-like line onto a space-like line, and a time-like line onto a time-like line.

Since a space-like plane intersects $T = 0$ in a line which lies outside ψ, space-like planes are mapped onto space-like planes, and since time-like planes intersect $T = 0$ in a line which meets ψ in two distinct points, time-like planes map onto time-like planes, and, of course, isotropic planes map onto isotropic planes. We shall show later (§8) that by a suitable affine transformation of E_3 which preserves ψ, any given space-like plane can be mapped onto any other given space-like plane, and in particular, therefore, onto $Z = 0$ itself, and the Laguerre theorem will follow immediately. This affine transformation will have to correspond to a Laguerre axial transformation in $Z = 0$, so let us see how certain affine transformations of E_3 which preserve ψ induce a Laguerre axial transformation in $Z = 0$.

6. Let v be a fixed plane in E_3 which intersects $Z = 0$ in the line Ω. If q is any plane which touches ψ, let q' be the unique plane which also touches ψ, and is such that the line $q \cap q'$ lies in v. The map from q to q' defined in this way induces an affine map of E_3, an harmonic homology in fact, which maps ψ onto itself. It is the dual of the point-point map in E_3, also an harmonic homology, in which V is a fixed point, and any point Q of a given quadric surface is mapped onto the unique point Q' in which the line VQ intersects the surface again. The extension of this map to the whole of E_3 is a collineation of E_3 which maps the quadric onto itself. See Pedoe (**4**, p. 439). In our dual map the quadric surface is specialized to the conic ψ, and lies in $T = 0$, and so this plane is necessarily mapped onto itself.

We choose the fixed plane v to be time-like, so that the points of v represent cycles which intersect Ω, suitably oriented, at a fixed angle α. Let q, q' intersect $Z = 0$ in the lines p, p' respectively. Then p and p' meet on Ω, since $q \cap q'$ lies in v. The line $q \cap q'$ represents cycles which touch p, p', when suitably oriented. Since $q \cap q'$ lies in v, these cycles also intersect Ω, suitably oriented, at a fixed angle α. We only need one cycle K, intersecting a fixed ray ω parallel to Ω at the angle α to represent the whole system of cycles, and to obtain the map between the rays p and p' intersecting on Ω (Figure 1), and this is the Laguerre transformation.

7. Laguerre proves geometrically that rays tangent to a given cycle $\mathscr{C}$ map onto rays tangent to a cycle $\mathscr{C}'$. This is reproduced in Pedoe (**3**, p. 260). From the scenic viewpoint we have commandeered this is evident, since planes q tangent to ψ which pass through the point P which represents a cycle $\mathscr{C}$ are mapped by our collineation of E_3 onto planes q' tangent to ψ which pass through a point P', and such planes touch the cycle $\mathscr{C}'$ in $Z = 0$ represented by P'. But Laguerre, in the course of his proof, also shows that the cycles $\mathscr{C}$ and $\mathscr{C}'$ have Ω as their radical axis, and from this it is easy to deduce that a Laguerre axial transformation preserves the tangential distance between two given cycles. For this see Pedoe (**3**, p. 261). We know that the angle between two circles is preserved under inversion, and this is the corresponding invariant under Laguerre's axial transformation.

8. We now prove that there is an harmonic homology of E_3, of the kind we have just described in §6, which maps ψ onto itself and any given space-like plane π onto any other given space-like plane π'. Blaschke merely states, without any discussion, that there is an affine transformation of E_3 which does this (**1**, p. 150).

Let l, l' be the respective intersections of the given planes π, π' with $T=0$. We first show that there is an harmonic homology in $T=0$ which maps ψ onto itself, and the line l onto the line l'.

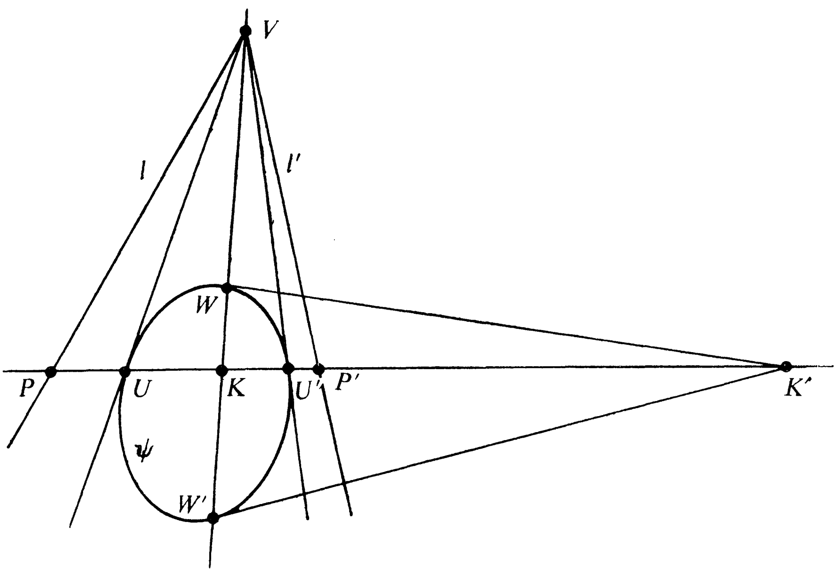


Figure 2.

Neither l nor l' intersects ψ. Let $V = l \cap l'$ and, V lying outside ψ, let VU, VU' be the tangents from V to ψ (Figure 2). Let l and l' intersect UU' at P and P' respectively. The points U, U' are inside the segment PP', and we can therefore find points K, K' which harmonically separate both U, U' and P, P'. (If circles through P, P' and U, U' respectively intersect at T, T', let the line TT' intersect PP' at O. Then

$$OP \,.\, OP' = OT \,.\, OT' = OU \,.\, OU' = c^2,$$

and the points K, K' on either side of O are determined by $OK = OK' = c$.) We know that one of K, K' is inside ψ. Let it be K, and suppose that VK intersects ψ in W and W'. The tangents to ψ at W and W' intersect on UU' at the point K'.

We can define a collineation of $T=0$ uniquely by the maps:

$$VP \to VP', \; VK \to VK, \; WK' \to WK' \text{ and } W'K' \to W'K'.$$

In this collineation $V = VP \cap VK$ is mapped onto the point $VP' \cap VK$, so that

V is mapped onto itself. The point $W = VK \cap WK'$ is mapped onto $VK \cap WK'$, that is onto itself, and the point $W' = VK \cap W'K'$ is mapped onto $VK \cap W'K'$, that is onto itself. The collineation induced on the line VK has, therefore, three fixed points, V, W, W', and is therefore a line of fixed points. This is a characteristic property of an homology.

Furthermore, $K' = K'W \cap K'W'$ is also mapped onto itself, and so this is the center of the homology, VK being the axis. All lines through K' are fixed lines, and the join of any point to its map passes through K'. Dually, any line and its map intersect on VK. Now, since our collineation maps l onto l', the point P is mapped onto the point P'. The range $\{PP', KK'\}$ being harmonic, our homology is an harmonic homology. See Pedoe (**4**, p. 317).

A conic is uniquely determined by the requirement of touching two given lines at given points on them, and of passing through a third given point. The conic ψ, which touches the fixed line WK' at the fixed point W, and the fixed line $W'K'$ at the fixed point W', and also passes through U, is mapped onto a conic which touches WK' at W, touches $W'K'$ at W', and also passes through U', the range $\{UU', KK'\}$ also being harmonic, and U' therefore being the map of U. The unique conic which satisfies these conditions is ψ, so that our collineation maps ψ onto itself. In this collineation any tangent to ψ intersects its map, which is also a tangent, on the axis VK, so that the collineation can also be defined by the requirement of pairs of corresponding tangents intersecting on VK.

Returning to E_3, a collineation is to be defined which induces the one just described in $T = 0$, but also maps the plane π onto the plane π'. We know that a collineation in $T = 0$ is uniquely defined by the assignment of four points, no three collinear, and their maps. See Pedoe (**4**, p. 295). The collineation in $T = 0$ is therefore the one we require if we map

$$V \to V,\ W \to W,\ P \to P' \text{ and } U \to U',$$

since this is what our collineation does. To obtain our map in E_3, which is uniquely determined by the assignment of five points, no four being coplanar, and their maps, we merely choose a point A in π, add it to the four points given above, and assign a point A' in π' as its map. Our collineation in E_3 induces the required one in $T = 0$, maps π onto π', and, reverting to the notation of § 6, v is a fixed plane through VK, and is therefore time-like, and all pairs of tangent planes to ψ have their lines of intersection in v, since all pairs of tangent lines to ψ in $T = 0$ intersect on VK.

9. To prove Laguerre's fundamental theorem, we suppose that three cycles $\mathscr{C}$, $\mathscr{D}$ and $\mathscr{E}$ in $Z = 0$ are such that their axis of similitude does not intersect them. The points P, Q, R representing the cycles, the plane PQR in E_3, as we saw, (§ 3), is space-like. By what we have just proved in § 8, it is therefore possible to find an affine map of E_3 which corresponds to a Laguerre axial transformation of $Z = 0$ which will map the plane PQR onto the special space-like plane given by $Z = 0$. But points in this plane represent point-cycles. We have therefore found a Laguerre axial transformation which maps the three given cycles $\mathscr{C}$, $\mathscr{D}$ and $\mathscr{E}$ simultaneously onto three points.

As an immediate consequence, one of the most elegant proofs of the Apollonius theorem follows. This was given by Laguerre, and is reproduced in Pedoe (**3**, p. 264). The field for other applications lies wide open.

References

1. W. Blaschke, Vorlesungen über Differentialgeometrie, III, Springer, Berlin, 1929.
2. J. L. Coolidge, A Treatise on the Circle and the Sphere, Chelsea, New York, 1971.
3. D. Pedoe, A forgotten geometrical transformation, L'Enseignement Math., 28 (1972) 255–267.
4. ———, A Course of Geometry for Colleges and Universities, Cambridge University Press, New York, 1970.
5. I. M. Yaglom, Complex Numbers in Geometry, Academic Press, New York, 1968.

ON (p, q)-CONTINUOUS FUNCTIONS

LOUIS H. Y. CHEN and LEONARD Y. H. YAP, University of Singapore

Every student of the calculus knows that the function f defined on $[0,1]$ by $f(x) = x$ is a one-to-one continuous function. A simple application of the Intermediate Value Theorem shows that there is no two-to-one continuous function on $[0, 1]$. Pursuing this a little further, one soon discovers that there is no p-to-one continuous function on $[0,1]$ if p is a positive integer other than one. (See [**1**].) This leads to the following problem: Given two positive integers p and q, find a necessary and sufficient condition on p and q such that there exists a continuous function on $[0,1]$ with the property that the preimage of every point in its range consists of either p or q points. We define a *(p, q)-continuous* function to be a continuous function with this property. We take $p \leqq q$ with no loss of generality. A solution to the problem is then given by the following theorem:

THEOREM. *There exists a (p,q)-continuous function on $[0,1]$ if and only if $2p \leqq q + 1$.*

First we prove that the condition is necessary.

LEMMA. *Let f be a (p,q)-continuous function on $[0,1]$, where $(p,q) \neq (1,2)$. Then the preimage of the maximum value of f consists of p points. The same is true for the minimum value of f.*

Proof. If $p = q$, then there is nothing to prove. Suppose $p < q$ and f achieves maximum value M at q points. A simple application of the Intermediate Value Theorem shows that the preimage of $M - \varepsilon$, for sufficiently small positive number ε, must contain at least $2(q-2) + 2 = 2q - 2$ points. Thus $2q - 2 \leqq q$, i.e., $q \leqq 2$. But $(p,q) \neq (1,2)$ and $p < q$ imply that $q \geqq 3$. This is a contradiction, and so f achieves the maximum value M at p points. The last assertion follows from

$$\min_{0 \leqq x \leqq 1} f(x) = - \max_{0 \leqq x \leqq 1} (-f(x))$$

and the fact that $-g$ is (p,q)-continuous if g is. This completes the proof of the lemma.

If $(p,q) = (1,2)$, then $2p \leqq q + 1$ is obviously true. Now suppose $(p,q) \neq (1,2)$. Then by the lemma, f achieves its maximum value M and minimum value m at the points $\alpha_1, \cdots, \alpha_p$ and $\beta_1, \cdots, \beta_p$, respectively. We consider two cases:

Case (1). Suppose at most one of $\alpha_1, \cdots, \alpha_p$ is an end point of $[0,1]$. Then, by the Intermediate Value Theorem, the preimage of $M - \varepsilon$, for sufficiently small positive number ε, contains at least $2(p-1)+1 = 2p-1$ points. This implies $2p - 1 \leqq q$, i.e., $2p \leqq q + 1$.

Case (2). Suppose exactly two of $\alpha_1, \cdots, \alpha_p$ are end points of $[0,1]$. Then all the points $\beta_1, \cdots, \beta_p$ are in $(0,1)$. It follows from the Intermediate Value Theorem that the preimage of $m + \varepsilon$, for sufficiently small positive number ε, contains at least $2p$ points. This implies $2p \leqq q$, and hence $2p \leqq q + 1$.

This completes the proof that $2p \leqq q + 1$ is a necessary condition.

Now to prove that the condition is sufficient we show how to construct a (p,q)-continuous function on $[0,1]$ when $2p \leqq q+1$. First for each positive integer k we define a continuous function $\psi_k\colon [0,\infty) \to [0,\infty)$ such that $\psi_k(0) = 0$, the preimage of 0 consists of k points, and the preimage of every positive number consists of $2k-1$ points.

For $k = 1$, we take ψ_1 to be the identity function $\psi_1(x) = x$ for all $x \in [0,\infty)$.

For $k \geqq 2$, we construct ψ_k in stages, as follows: Let $\{x\}$ denote the distance from x to the nearest integer and define

$$\tilde{\psi}_k(x) = \begin{cases} \{x\} & \text{for } 0 \leqq x \leqq k-1 \\ \frac{1}{2}(x-(k-1)) & \text{for } k-1 \leqq x \leqq k, \end{cases}$$

$$\psi_k(x) = \tilde{\psi}_k(x - nk) + \tfrac{1}{2}n \qquad \text{for } nk \leqq x \leqq (n+1)k,\ n = 1,2,3,\cdots.$$

The reader is advised to sketch the graph of ψ_k and check that it has the desired properties.

Define $g\colon [0,1) \to [0,\infty)$ and $h\colon [0,\infty) \to [0,1)$ by

$$g(x) = \frac{x}{1-x}, \qquad h(y) = \frac{y}{1+y}.$$

Note that g and h are one-to-one, onto, and continuous. Then $h \circ \psi_k \circ g\colon [0,1) \to [0,1)$ is continuous, $h \circ \psi_k \circ g(0) = 0$, and the preimages of 0 and y $(0 < y < 1)$ consist of k and $2k-1$ points, respectively. Since $\lim_{x \to 1^-} (h \circ \psi_k \circ g)(x) = 1$, it follows that the function ϕ_k defined by

$$\phi_k(x) = \begin{cases} h \circ \psi_k \circ g(x) & \text{if } 0 \leqq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

is continuous on $[0,1]$ and the preimages of 1, 0 and $y\,(0 < y < 1)$ consist of 1, k and $2k-1$ points, respectively.

Let the function $\tilde{f}_{1,2k-1}$ be defined on $[-\frac{1}{2},\frac{1}{2}]$ by

$$\tilde{f}_{1,2k-1}(x) = \begin{cases} -\frac{1}{2}\phi_k(-2x) & \text{if } -\frac{1}{2} \leqq x \leqq 0 \\ \frac{1}{2}\phi_k(2x) & \text{if } 0 \leqq x \leqq \frac{1}{2}. \end{cases}$$

This function is $(1,2k-1)$-continuous with $\tilde{f}_{1,2k-1}(-\frac{1}{2}) = -\frac{1}{2}$ and $\tilde{f}_{1,2k-1}(\frac{1}{2}) = \frac{1}{2}$. Therefore the function $f_{1,2k-1}$ defined by

$$f_{1,2k-1}(x) = \tfrac{1}{2} + \tilde{f}_{1,2k-1}(x - \tfrac{1}{2})$$

is $(1,2k-1)$-continuous from $[0,1]$ to $[0,1]$.

The reader may easily verify that the function $f_{1,2k+2}$ defined by

$$f_{1,2k+2}(x) = \begin{cases} \frac{1}{2} - \frac{1}{2}\phi_k(4x) & \text{if } 0 \leqq x \leqq \frac{1}{4} \\ \frac{1}{2}f_{1,3}(4x-1) & \text{if } \frac{1}{4} \leqq x \leqq \frac{1}{2} \\ \frac{1}{2} + \frac{1}{2}\phi_{k+1}(4x-2) & \text{if } \frac{1}{2} \leqq x \leqq \frac{3}{4} \\ \frac{5}{2} - 2x & \text{if } \frac{3}{4} \leqq x \leqq 1 \end{cases}$$

is a $(1,2k+2)$-continuous function from $[0,1]$ to $[0,1]$.

Let $v\colon [0,1] \to [0,1]$ be defined by

$$v(x) = \begin{cases} 1-2x & \text{if } 0 \leqq x \leqq \frac{1}{2} \\ 2x-1 & \text{if } \frac{1}{2} \leqq x \leqq 1. \end{cases}$$

Then the function $\tilde{f}_{p,2p+2k-3}$ defined by

$$\tilde{f}_{p,2p+2k-3}(x) = \begin{cases} f_{1,2k-1}(x) & \text{if } 0 \leqq x \leqq 1 \\ v(x-i) & \text{if } i \leqq x \leqq i+1, \quad i = 1,2,\cdots,p-1 \end{cases}$$

is $(p,2p+2k-3)$-continuous from $[0,p]$ to $[0,1]$. Hence the function $f_{p,2p+2k-3}\colon [0,1] \to [0,1]$ defined by

$$f_{p,2p+2k-3}(x) = \tilde{f}_{p,2p+2k-3}(px)$$

is $(p,2p+2k-3)$-continuous from $[0,1]$ to $[0,1]$.

The function v is clearly $(1,2)$-continuous so we set $f_{1,2} = v$. Also the function $f_{2,4}$ defined by

$$f_{2,4}(x) = \begin{cases} 1-2x & \text{if } 0 \leqq x \leqq \frac{1}{2} \\ \phi_2(2x-1) & \text{if } \frac{1}{2} \leqq x \leqq 1 \end{cases}$$

is a $(2,4)$-continuous function from $[0,1]$ to $[0,1]$.

For $p \geqq 3$, the function $\tilde{f}_{p,2p}$ defined by

$$\tilde{f}_{p,2p}(x) = \begin{cases} f_{2,4}(x) & \text{if } 0 \leqq x \leqq 1 \\ v(x-i) & \text{if } i \leqq x \leqq i+1, \quad i = 1,\cdots,p-2 \end{cases}$$

is $(p,2p)$-continuous from $[0,p-1]$ to $[0,1]$. Hence the function $f_{p,2p}$ defined by

$$f_{p,2p}(x) = \tilde{f}_{p,2p}((p-1)x)$$

is $(p,2p)$-continuous from $[0,1]$ to $[0,1]$.

Define $w: [0,1] \to [0,1]$ by

$$w(x) = \begin{cases} \frac{1}{2} - 2x & \text{if } 0 \leqq x \leqq \frac{1}{4} \\ 2x - \frac{1}{2} & \text{if } \frac{1}{4} \leqq x \leqq \frac{3}{4} \\ \frac{5}{2} - 2x & \text{if } \frac{3}{4} \leqq x \leqq 1. \end{cases}$$

Then the function $\tilde{f}_{p,2p+2k}$ defined by

$$\tilde{f}_{p,2p+2k}(x) = \begin{cases} f_{1,2k+2}(x) & \text{if } 0 \leqq x \leqq 1 \\ w(x-i) & \text{if } i \leqq x \leqq i+1, \qquad i = 1, 2, \cdots, p-1 \end{cases}$$

is $(p, 2p+2k)$-continuous from $[0,p]$ to $[0,1]$. Hence the function $f_{p,2p+2k}$ defined by

$$f_{p,2p+2k}(x) = \tilde{f}_{p,2p+2k}(px)$$

is $(p, 2p+2k)$-continuous from $[0,1]$ to $[0,1]$. This completes the proof.

COROLLARY. *Let f be a continuous function on* $[0,1]$ *such that the preimage of every point in the range of f contains the same number of points. Then f is one-to-one.*

Reference

1. B. R. Wenner, Continuous, exactly k-to-one functions on R, this MAGAZINE, 45 (1972) 224–225.

DENSE PACKING OF EQUAL CIRCLES WITHIN A CIRCLE

GEORGE E. REIS, Sandia Laboratories, Albuquerque, New Mexico

Introduction. In connection with the packaging of cylindrical modules in cylindrical containers, the problem of determining the diameter D of the smallest circumscribing circle that can contain N nonoverlapping equal circles, each of diameter d, for all values of N continues to recur. Several authors [**1–6**] have reported values of D/d for $N \leqq 20$. Since there is no known algorithm or formula for determining D/d for all values of N and because the reported values are conjectures, the problem was re-examined by means of an experimental procedure in which a camera iris diaphragm was used as a circumscribing circle of variable diameter and hard plastic cylinders were used to represent the enclosed circles. Patterns considered as candidates for the most compact arrays were determined by closing the diaphragm around a cluster of

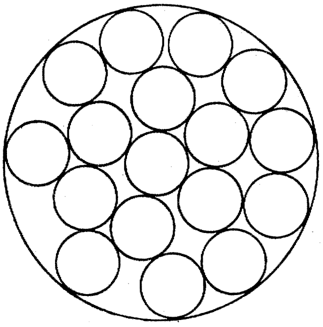


FIG. 1. $N=17$

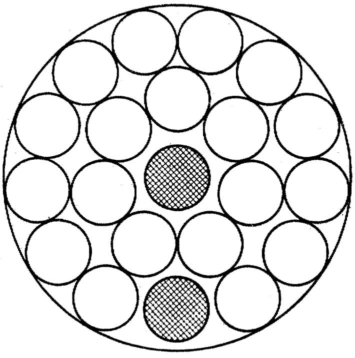


FIG. 2. $N=21$

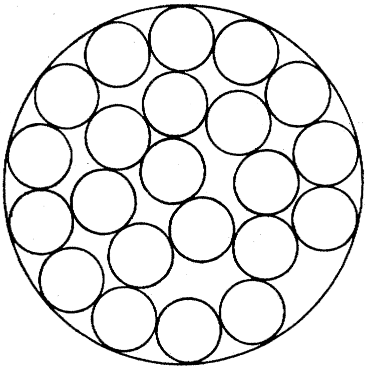


FIG. 3. $N=22$

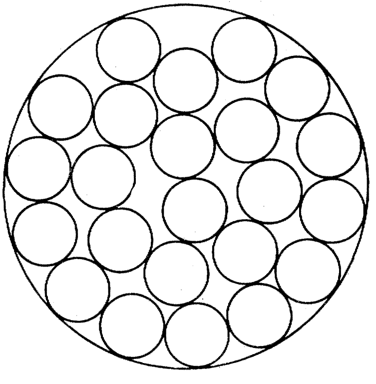


FIG. 4. $N=23$

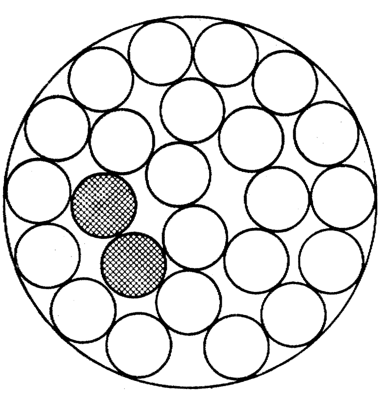


FIG. 5. $N=24$

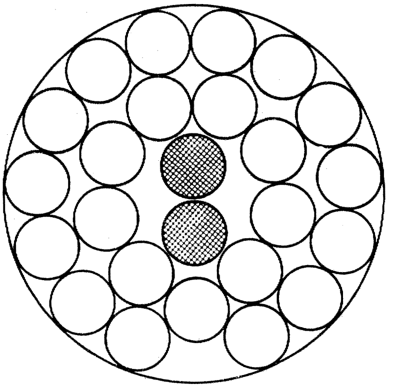


FIG. 6. $N=25$

cylinders standing on end. The coordinates of the centers of the enclosed cylinders in these patterns and the corresponding D/d's were calculated for $1 \leqq N \leqq 25$. By using this technique it was usually possible to reduce the number of patterns to be calculated, for a given N, to three or less. This investigation produced values of D/d for $N = 21, 22, 23, 24$ and 25 which have not been found in the literature and a value of D/d for $N = 17$ which is less than the smallest value found in previous reports. In the data presented below for $N = 17, 21, 22, 23, 24$ and 25, ρ is the ratio of the area of the enclosed circles to the area of the circumscribing circle, and the numerical values of D and the coordinates of the centers of the enclosed circles are given for enclosed circles of unit diameter [7].

$N = 17$. Goldberg [6] has presented the data on the densest pattern found in the literature for $N = 17$. He gives the value of D/d as 4.8085. A slightly more compact arrangement, with a D/d of 4.7920, is shown in Figure 1. The coordinates for this pattern are given in Table 1. Attempts to produce a more compact nonsymmetric pattern were unsuccessful.

$N = 21, 22, 23, 24$, and 25. Patterns for these values of N have not been found in the literature. The results of the present investigation are shown in Figures 2 through 6 and Tables 2 through 6. In the diagrams the cross-hatched circles are not rigidly constrained to the locations shown. It can be seen that the patterns for $N = 22, 23$ and 24 are not symmetric.

Comments. The techniques used by various investigators in searching for the most compact packing of equal circles within a circle involve different assumptions. In developing the data presented here an attempt has been made to avoid, in as far as possible, any assumptions in selecting the patterns considered as candidates for the most compact arrays. The use of an iris diaphragm as a variable circumscribing circle and plastic cylinders to represent the enclosed circles appears to eliminate any assumption connected with symmetry. The major objection to the use of an iris diaphragm is that its leaves do not form a perfect circle as the diameter is varied. However, the fact that a new value of D/d for $N = 17$ was found which is smaller than any published previously seems to indicate that the technique has some merit.

TABLE 1

$N = 17$

$D = 4.79203\ 37483$

$\rho = 0.74030\ 24480$

$X1$	$= 0.92425\ 56618$	$Y1$	$= 1.65548\ 52638$
$X2$	$= -0.92425\ 56618$	$Y2$	$= 1.65548\ 52638$
$X3$	$= 1.79177\ 83383$	$Y3$	$= 0.62000\ 83656$
$X4$	$= -1.79177\ 83383$	$Y4$	$= 0.62000\ 83656$
$X5$	$= 1.85799\ 61095$	$Y5$	$= -0.37779\ 68292$
$X6$	$= -1.85799\ 61095$	$Y6$	$= -0.37779\ 68292$
$X7$	$= 1.36091\ 32984$	$Y7$	$= -1.32014\ 96056$
$X8$	$= -1.36091\ 32984$	$Y8$	$= -1.32014\ 96056$

TABLE 2

$N = 21$

$D = 5.25231\ 74750$

$\rho = 0.76123\ 25612$

$X1$	$= 0.50000\ 00000$	$Y1$	$= 2.06653\ 11459$
$X2$	$= -0.50000\ 00000$	$Y2$	$= 2.06653\ 11459$
$X3$	$= 1.38939\ 40136$	$Y3$	$= 1.60938\ 97142$
$X4$	$= -1.38939\ 40136$	$Y4$	$= 1.60938\ 97142$
$X5$	$= 1.97143\ 74363$	$Y5$	$= 0.79623\ 20087$
$X6$	$= -1.97143\ 74363$	$Y6$	$= 0.79623\ 20087$
$X7$	$= 2.11737\ 52945$	$Y7$	$= -0.19306\ 17503$
$X8$	$= -2.11737\ 52945$	$Y8$	$= -0.19306\ 17503$

TABLE 1 (continued)

X9	= 0.50000 00000	Y9	= −1.82890 13060
X10	= −0.50000 00000	Y10	= −1.82890 13060
X11	= 0.00000 00000	Y11	= 1.27371 11530
X12	= 0.79275 39094	Y12	= 0.66416 93257
X13	= −0.79275 39094	Y13	= 0.66416 93257
X14	= 0.86091 32984	Y14	= −0.45412 42019
X15	= −0.86091 32984	Y15	= −0.45412 42019
X16	= 0.00000 00000	Y16	= −0.96287 59022
X17	= 0.00000 00000	Y17	= 0.05462 74985

TABLE 2 (continued)

X9	= 1.79492 43864	Y9	= −1.13964 79386
X10	= −1.79492 43864	Y10	= −1.13964 79386
X11	= 1.07541 47138	Y11	= −1.83413 03581
X12	= −1.07541 47138	Y12	= −1.83413 03581
X13	= 0.00000 00000	Y13	= 1.20050 57421
X14	= 0.88939 40136	Y14	= 0.74336 43104
X15	= −0.88939 40136	Y15	= 0.77336 43104
X16	= 1.11806 29270	Y16	= −0.23013 99413
X17	= −1.11806 29270	Y17	= −0.23013 99413
X18	= 0.50000 00000	Y18	= −1.01626 85683
X19	= −0.50000 00000	Y19	= −1.01626 85683
X20	= 0.00000 00000	Y20	= 0.00000 00000
X21	= 0.00000 00000	Y21	= −2.00000 00000

TABLE 3

$N = 22$

$D = 5.43971\ 89591$

$\rho = 0.74348\ 07966$

X1	= 0.00000 00000	Y1	= 2.21985 94795
X2	= 0.97430 34310	Y2	= 1.99461 99972
X3	= −0.97430 34310	Y3	= 1.99461 99972
X4	= 1.75089 02025	Y4	= 1.36460 96907
X5	= −1.75089 02025	Y5	= 1.36460 96907
X6	= 2.17216 65532	Y6	= 0.45767 73689
X7	= −2.17216 65532	Y7	= 0.45767 73689
X8	= 2.15264 23267	Y8	= −0.54213 20152
X9	= −2.15264 23267	Y9	= −0.54213 20152
X10	= 1.69627 95997	Y10	= −1.43192 58460
X11	= −1.69627 95997	Y11	= −1.43192 58460
X12	= 0.44241 30664	Y12	= 1.14780 67710
X13	= 1.17218 72979	Y13	= 0.46411 85591
X14	= 1.15266 30714	Y14	= −0.53569 08250
X15	= 0.69630 03444	Y15	= −1.42548 46558
X16	= 0.08666 55554	Y16	= −2.21816 70790
X17	= −0.55570 28058	Y17	= 1.08644 95745
X18	= −1.18342 93001	Y18	= 0.30801 56027
X19	= −1.03585 31613	Y19	= −0.68103 50953
X20	= −0.89568 86524	Y20	= −2.03113 71069
X21	= −0.00350 79789	Y21	= 0.25273 44715
X22	= −0.03800 94680	Y22	= −0.74667 01749

TABLE 4

$N = 23$

$D = 5.54520\ 42226$

$\rho = 0.74798\ 47534$

X1	= 0.50000 00000	Y1	= 2.21691 68582
X2	= 1.40318 93374	Y2	= 1.78767 44781
X3	= 2.03469 12956	Y3	= 1.01230 01964
X4	= 2.27223 36287	Y4	= 0.04092 30091
X5	= 2.06982 30753	Y5	= −0.93837 77454
X6	= 1.46665 06350	Y6	= −1.73598 85573
X7	= −0.50000 00000	Y7	= 2.21691 68582
X8	= −1.40318 93374	Y8	= 1.78767 44781
X9	= −2.03469 12956	Y9	= 1.01230 01964
X10	= −2.27223 36287	Y10	= 0.04092 30091
X11	= −2.06982 30753	Y11	= −0.93837 77454
X12	= −1.46665 06350	Y12	= −1.73598 85573
X13	= 0.22286 13706	Y13	= −2.26164 83294
X14	= −0.48340 51366	Y14	= −1.55370 22716
X15	= 0.55758 64321	Y15	= −1.31933 24942
X16	= 1.16075 88724	Y16	= −0.52172 16823
X17	= 1.18948 13729	Y17	= 0.47786 57416
X18	= 0.51597 19941	Y18	= 1.21704 44186
X19	= −0.14868 00752	Y19	= −0.61138 64364
X20	= −1.14698 51795	Y20	= −0.55318 91967
X21	= −1.21136 18214	Y21	= 0.44473 64759
X22	= −0.57985 98631	Y22	= 1.22011 07576
X23	= −0.03428 46956	Y23	= 0.38204 88645

TABLE 5

$N = 24$

$D = 5.65166\ 10918$

$\rho = 0.75137\ 89425$

X1	= 0.50000 00000	Y1	= 2.27145 05780
X2	= 1.40756 98060	Y2	= 1.85154 93429
X3	= 2.05493 57117	Y3	= 1.08936 99781
X4	= 2.32242 54044	Y4	= 0.12580 92564

TABLE 6

$N = 25$

$D = 5.76084\ 47666$

$\rho = 0.75329\ 94722$

X1	= 0.50000 00000	Y1	= 2.32731 83545
X2	= 1.41176 07204	Y2	= 1.91659 65646
X3	= 2.07437 59430	Y3	= 1.16763 65745
X4	= 2.37090 82880	Y4	= 0.21261 37641

TABLE 5 (continued)

X5	= 2.16059 06360	Y5	= −0.86100 86131
X6	= 1.59934 82444	Y6	= −1.68866 00965
X7	= 0.74244 97158	Y7	= −2.20414 52193
X8	= −0.50000 00000	Y8	= 2.27145 05780
X9	= −1.40756 98060	Y9	= 1.85154 93429
X10	= −2.05493 57117	Y10	= 1.08936 99781
X11	= −2.32242 54044	Y11	= 0.12580 92564
X12	= −2.16059 06360	Y12	= −0.86100 86131
X13	= −1.59934 82444	Y13	= −1.68866 00965
X14	= −0.74244 97158	Y14	= −2.20414 52193
X15	= 0.00000 00000	Y15	= 1.40542 51742
X16	= 0.90756 98060	Y16	= 0.98552 39391
X17	= −1.07118 60666	Y17	= 0.90982 43407
X18	= −1.32000 00000	Y18	= −0.08500 00000
X19	= 0.00000 00000	Y19	= −1.54324 34209
X20	= −0.19660 63070	Y20	= 0.42494 26603
X21	= 1.32365 67462	Y21	= 0.07619 91027
X22	= 1.01709 59123	Y22	= −0.87565 19650
X23	= 0.07613 17900	Y23	= −0.53714 56571
X24	= −0.85500 00000	Y24	= −0.99000 00000

TABLE 6 (continued)

X5	= 2.24902 61545	Y5	= −0.77993 08171
X6	= 1.73023 91257	Y6	= −1.63483 43313
X7	= 0.90610 19890	Y7	= −2.20122 46384
X8	= −0.50000 00000	Y8	= 2.32731 83545
X9	= −1.41176 07204	Y9	= 1.91659 65646
X10	= −2.07437 59430	Y10	= 1.16763 65745
X11	= −2.37090 82880	Y11	= 0.21261 37641
X12	= −2.24902 61545	Y12	= −0.77993 08171
X13	= −1.73023 91257	Y13	= −1.63483 43313
X14	= −0.90610 19890	Y14	= −2.20122 46384
X15	= 0.50000 00000	Y15	= 1.32731 83545
X16	= 1.22548 05506	Y16	= 0.63907 56860
X17	= 1.34434 83989	Y17	= −0.35383 43978
X18	= 0.81943 11102	Y18	= −1.20498 76391
X19	= −0.50000 00000	Y19	= 1.32731 83545
X20	= −1.22548 05506	Y20	= 0.63907 56860
X21	= −1.34434 83989	Y21	= −0.35383 43978
X22	= −0.81943 11102	Y22	= −1.20498 76391
X23	= 0.00000 00000	Y23	= −1.77816 53215
X24	= 0.00000 00000	Y24	= 0.43750 00000
X25	= 0.00000 00000	Y25	= −0.60900 00000

References

1. L. Fejes-Tóth, Lagerungen in der Ebene, auf der Kugel und im Raum, Springer, Berlin, 1953.
2. ———, Regular Figures, Macmillan, New York, 1964.
3. C. A. Rogers, Packing and Covering, Cambridge University Press, New York, 1964.
4. S. Kravitz, Packing cylinders into cylindrical containers, this MAGAZINE, 40 (1967) 65–71.
5. U. Pirl, Der Mindestabstand von n in der Einheitskreisscheibe gelegenen Punkten, Math. Nachr., 40 (1969) 111–124.
6. M. Goldberg, Packing 14, 16, 17 and 20 circles in a circle, this MAGAZINE, 44 (1971) 134–139.
7. G. E. Reis, Packing Cylindrical Modules in Circular Cylindrical Containers, SLA–73–0683, Sandia Laboratories, Albuquerque, New Mexico, September 1973.

INVARIANCE PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

PETER TAN and CONSTANTIN DROSSOS, Carleton University, Canada

1. Introduction. Maximum likelihood estimation as a method of estimating unknown parameters of a probability distribution is well known and is discussed in almost every textbook in mathematical statistics for undergraduates. It seems generally agreed that it was first introduced as a statistical method by R. A. Fisher [1], marking in many respects the beginning of modern statistical theory.

Let $f(x;\theta)$ be the probability (density or mass) function of a random variable X when $X = x$. The functional form f is usually assumed to be known except that it involves the unknown value of a parameter θ. For a given sample $\mathbf{x} = (x_1, \cdots, x_n)$ of n independent observations of X, the likelihood function of θ is the function

$L(\theta;x)$ defined by

$$L(\theta; x) = kf(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta), \tag{1}$$

where k is an indeterminate constant independent of θ which takes values in a parameter space Ω. The maximum likelihood estimate of θ, if it exists, is the value $\hat{\theta} = \hat{\theta}(x) \in \Omega$ which maximizes $L(\theta;x)$. $\hat{\theta}(X)$ considered as a function of the random variable X is called the maximum likelihood estimator of θ.

2. An invariance property of $\hat{\theta}(X)$. Fisher in his (1922) paper pointed out by an example an invariance property enjoyed by a maximum likelihood estimator but not by Bayes estimators. His example is the following. Let

$$\begin{aligned} f(x;\theta) &= \theta \qquad \text{if } x = 1, \quad 0 \leqq \theta \leqq 1, \\ &= 1-\theta \quad \text{if } x = 0. \end{aligned} \tag{2}$$

The likelihood function of θ for a given $x = (x_1,\cdots,x_n)$ is

$$L(\theta;x) = k\theta^y(1-\theta)^{n-y}, \qquad y = \sum x_i. \tag{3}$$

The maximum likelihood estimate $\hat{\theta}$ is found to be $\hat{\theta} = \bar{x}$ where $\bar{x} = \sum x_i/n$. If the parameter of interest is

$$\lambda = w(\theta) = \sin^{-1}(2\theta - 1), \quad -\frac{\pi}{2} \leqq \lambda \leqq \frac{\pi}{2},$$

the probability function (2) may be expressed in terms of λ instead of θ giving the likelihood function of λ:

$$M(\lambda;x) = k(1+\sin\lambda)^y(1-\sin\lambda)^{n-y}.$$

The maximum likelihood estimate of λ is then given by

$$\hat{\lambda} = \hat{w}(\theta) = \sin^{-1}(2\hat{\theta} - 1) = w(\hat{\theta}).$$

In general, if $\lambda = w(\theta)$ is a one-one function of θ defined on $\theta \in \Omega$, then it follows from the uniqueness of $\theta = w^{-1}(\lambda)$ that the maximum likelihood estimate of λ is given by

$$\hat{\lambda} = \hat{w}(\theta) = w(\hat{\theta}). \tag{4}$$

The invariance property (4) is given in [2] as an exercise (Problem 9.10, page 247) as follows:

"Consider a random variable $X_1,\cdots,X_n$ from a distribution which has p.d.f. $f(x:\theta)$, $\theta \in \Omega$. Let $\hat{\theta} = \mu(x_1,\cdots,X_n)$ denote a maximum likelihood statistic for θ. If $w(\theta)$ is a single-valued function of θ defined on $\theta \in \Omega$, show that $w[\mu(x_1,\cdots,x_n)]$ is that value of $w(\theta)$ which allows the likelihood function to attain its maximum. The resulting statistic $w[\mu(X_1,\cdots,X_n)]$ is called the maximum likelihood statistic for $w(\theta)$ and it is denoted by $\hat{w}(\theta)$. Thus $\hat{w}(\theta) = w(\hat{\theta})$. This fact is called the invariance property of maximum likelihood statistics, and it can be extended to single-valued functions of more than one parameter. For example $\hat{w}(\theta_1,\theta_2) = w(\hat{\theta}_1,\hat{\theta}_2)$."

The above exercise does not explicitly restrict $w(\theta)$ to be a one-one function. As an illustration, for estimating the variance $\lambda = w(\theta) = \theta(1 - \theta)$ of the Bernoulli distribution defined by (2), the invariance property (4) leads to the maximum likelihood estimate $\hat{\lambda} = \hat{\theta}(1 - \hat{\theta})$ where $\hat{\theta}$ is the sample mean $\bar{x}$. However, when $w(\theta)$ is not one-one, it is not immediately clear how we should transform the likelihood function of θ, $L(\theta; x)$, to that of $\lambda, M(\lambda; x)$, say. It is of interest to note that this exercise given in the second edition of the book by Hogg and Craig is not retained in the third edition of the book (1970).

A general proof of (4) when $w(\theta)$ is not necessarily one-one was given by Zehna in [5] who defined the likelihhood function induced by $\lambda = w(\theta)$ by

$$M(\lambda : x) = \sup \{L(\theta; x) : w(\theta) = \lambda\}. \tag{5}$$

In his review of Zehna's paper that appeared in Mathematical Review (1967), R. H. Berk commented that when $\lambda = w(\theta)$ is not one-one, "$M(\lambda)$ in general appears not to be a likelihood function associated with any random variable. That $\hat{\lambda}$ then maximizes $M(\lambda)$ is perhaps interesting but irrelevant to maximum likelihood estimation." Others, however, consider (4) a very useful property even when $w(\theta)$ is not one-one.

3. A translation property of $\hat{\theta}$ when $f(x; \theta) = f(x - \theta)$. A statistic $t(x)$ is said to have a translation property if for all $-\infty < c < \infty$, we have

$$t(g_c x) = t(x) + c, \tag{6}$$

where $g_c x = (x_1 + c, \cdots, x_n + c)$, $x = (x_1, \cdots, x_n)$. This translation property is considered by many statisticians a very desirable property for an estimator of θ when the probability function has the form $f(x; \theta) = f(x - \theta)$ for all real values of x and θ. We shall show that the maximum likelihood estimator $\hat{\theta}$ has this property.

Indeed, as pointed out to us by the referee, the property

$$\hat{\theta}(x_1 + c, \cdots, x_n + c) = \hat{\theta}(x_1, \cdots, x_n) + c \tag{7}$$

comes as a consequence of (4).

Let $\lambda = w(\theta) = \theta + c$. Then $\hat{\lambda}(x) = \hat{\theta}(x) + c$. When $f(x; \theta) = f(x - \theta)$, it is obvious that

$$\hat{\lambda}(x_1, \cdots, x_n) = \hat{\theta}(x_1 + c, \cdots, x_n + c).$$

Hence (7) obtains. Alternatively, (7) can be proved directly as follows.

Let

$$\hat{\alpha} = \hat{\theta}(x_1 + c, \cdots, x_n + c),$$

$$\hat{\theta} = \hat{\theta}(x_1, \cdots, x_n), \; \pi(x_i) = f(x_1) f(x_2) \cdots f(x_n).$$

$$\pi f(x_i + c - \hat{\alpha}) \leqq \pi f(x_i - \hat{\theta})$$

$$= \pi f(x_i + c - (\hat{\theta} + c)) \leqq \pi f(x_i + c - \hat{\alpha}).$$

Therefore

$$\hat{\alpha} = \hat{\theta} + c.$$

There are obviously other statistics which have the translation property (7). For example,

$$t(\boldsymbol{x}) = \sum x_i/n = \bar{x}\,, \text{ or } t(x) = x_1\,.$$

If X is a continuous random variable with a probability density function $f(x;\theta) = f(x-\theta)$ where x and θ both take values in the real line, then it has been shown (e.g., [4] or [3]) that under very weak assumptions on f, $\hat{\theta} = \bar{x}$ if and only if X has a normal distribution. Thus in most cases $\hat{\theta} \neq \bar{x}$. However, the translation property enjoyed by both $\hat{\theta}$ and $\bar{x}$ implies that given a coset Gx of x where

$$G\boldsymbol{x} = \{g_c\boldsymbol{x} = (x_1 + c, \cdots, x_n + c): -\infty < c < \infty\},$$

$\hat{\theta}$ and $\bar{x}$ differ only by a constant for all $\boldsymbol{x} \in G\boldsymbol{x}$.

Define $\mu(\boldsymbol{x}) = (y_1, \cdots, y_n)$ where $y_i = x_i - \bar{x}$, $i = 1, \cdots, n$. It is easily seen that

$$\mu(g_c\boldsymbol{x}) = \mu(\boldsymbol{x}) \text{ for all } g_c\boldsymbol{x} \in G\boldsymbol{x}.$$

By the translation property (7) we have

$$\hat{\theta}(x_1, \cdots, x_n) = \hat{\theta}(y_1, \cdots, y_n) + \bar{x} = K + \bar{x}, \tag{9}$$

where $K = \hat{\theta}(y_1, \cdots, y_n)$ is a constant. Theoretically, if two or more samples $\boldsymbol{x}$ belong to the same coset $G\boldsymbol{x}$, then once $\hat{\theta}(y_1, \cdots, y_n)$ is computed from one sample of data, the maximum likelihood estimates $\hat{\theta}$ for other samples can be obtained easily from (9).

4. An invariance property of $\hat{\theta}$ for a model with a group structure. Consider the probability density function

$$f(x;\theta) = (2\pi\theta^2)^{-\frac{1}{2}} \exp\{-(x - a\theta - b)^2/2\theta^2\}, \tag{10}$$

where $-\infty < x < \infty$, $\theta > 0$, a and b are given constants. It is a normal distribution whose mean is a linear function of its standard deviation θ.

Let $G = \{g_c: c > 0\}$ be a group of transformations defined by

$$g_c x_i = cx_i - cb + b, \quad i = 1, \cdots, n,$$

$$g_c\boldsymbol{x} = (g_c x_1, \cdots, g_c x_n).$$

A statistic $t(\boldsymbol{x})$ is said to be invariant under G if

$$t(g_c\boldsymbol{x}) = ct(\boldsymbol{x}) \text{ for all } c > 0. \tag{11}$$

The sample standard deviation $S(x)$ defined by

$$S^2 = \sum (x_i - \bar{x})^2/(n-1)$$

possesses the invariance property (11). We shall show that the maximum likelihood estimator $\hat{\theta}(X)$ is also invariant under G.

Let $L(\theta, x)$ be the likelihood function of θ for given x. We first observe

$$L(\theta; x) = L(c\theta; g_c x) \text{ for all } c > 0 \text{ and } x.$$

Let $\hat{\alpha} = \hat{\theta}(g_c x)$ and $\hat{\theta} = \hat{\theta}(x)$. The relation

$$\hat{\theta}(g_c x) = c\hat{\theta}(x) \tag{12}$$

comes from the following argument:

$$\begin{aligned}\sup L(\theta; g_c x) &= L(\hat{\alpha}; g_c x) = L(c^{-1}\hat{\alpha}; x) \\ &\leqq L(\hat{\theta}; x) = L(c\hat{\theta}; g_c x) \leqq \sup L(\theta; g_c x).\end{aligned}$$

Therefore

$$\hat{\alpha} = c\hat{\theta}.$$

For all x in the coset Gx of x defined by $Gx = \{g_c x : c > 0\}$, $\hat{\theta}(x)$ is related to $S(x)$ by the following relation:

$$\hat{\theta}(x) = KS(x) \tag{13}$$

where K is a constant depending only on the coset Gx. This can be shown in the following manner:

Since both $\hat{\theta}$ and S are invariant under G, we have, by (11),

$$\hat{\theta}(g_c x)/S(g_c x) = \hat{\theta}(x)/\hat{S}(x)$$

for all $c > 0$ and all x, from which (13) follows immediately.

Properties (12) and (13) enjoyed by $\hat{\theta}(X)$ are consequences of the group structure of the family $\{f(x;\theta): \theta > 0\}$ of probability distributions: if X has the density function $f(x;\theta)$, then $Y = g_c X = cX - cb + b$ has the density function $f(y;\lambda)$ where $\lambda = c\theta$ for $c > 0$.

This research was partially supported with a grant from the Department of University Affairs in Ontario and from Carleton University. The second author also acknowledges support by the Office of Scientific Research and Development of the Greek Government.

References

1. R. A. Fisher, On the mathematical foundations of theoretical statistics, Philos. Trans. Roy. Soc. London, Ser. A, 222 (1922) 309–368.

2. R. V. Hogg and A. T. Craig, Introduction to Mathematical Statistics, 2nd ed., Macmillan, New York, 1965.

3. P. Tan, Functional equations in mathematical statistics, this MAGAZINE, 45 (1972) 179–183.

4. H. Teicher, Maximum likelihood characterization of distributions, Ann. Math. Statist., 32 (1961) 1214–1222.

5. P. W. Zehna, Invariance of maximum likelihood estimators, Ann. Math. Statist., 37 (1966) 744.

A LATTICE OF CYCLOTOMIC FIELDS

ROGER CHALKLEY, University of Cincinnati

1. Introduction. The set Z^+ of positive integers with divisibility as a partial order is a distributive lattice ([3], p. 496). The set A of algebraic number fields with set inclusion as a partial order is also a lattice. We shall prove: *The set L of cyclotomic fields is a sublattice of A which is isomorphic to Z^+.*

An isomorphism of Z^+ onto L is easy to describe. For $n = 1, 2, \cdots$, let ζ_n be a primitive nth root of unity in the field C of complex numbers; and, let $Q(\zeta_n)$ be the cyclotomic subfield of C obtained by adjunction of ζ_n to the field Q of rational numbers. Then, the elements of L (with repetitions) are given by $Q(\zeta_n)$, for $n = 1, 2, \cdots$. We shall prove: *The mapping f_2 of Z^+ into L defined by*

$$f_2(n) = Q(\zeta_{2n}) \tag{1}$$

is a lattice-isomorphism of Z^+ onto L.

Throughout, let a and b denote positive integers with greatest common divisor d and least common multiple m. Then, in the lattice Z^+, the g.l.b. and the l.u.b. of a and b are

$$a \wedge b = d \quad \text{and} \quad a \vee b = m.$$

My results depend on the easily verified identity

$$\phi(a)\phi(b) = \phi(d)\phi(m), \tag{2}$$

where ϕ denotes the Euler ϕ-function.

A subfield F of C is an element of A if and only if its dimension $[F : Q]$ as a vector space over Q is finite. In particular, the cyclotomic field $Q(\zeta_n)$ satisfies

$$[Q(\zeta_n) : Q] = \phi(n), \quad \text{for} \quad n = 1, 2, \cdots; \tag{3}$$

e.g., see [2], p. 204. Given elements F_1, F_2 in A, their g.l.b. and l.u.b. are

$$F_1 \wedge F_2 = F_1 \cap F_2 \quad \text{and} \quad F_1 \vee F_2 = F_1F_2,$$

where $F_1 \cap F_2$ is their intersection and F_1F_2 is the subfield of C which they generate.

2. Verification that L is a sublattice of A.

THEOREM 1. *In the preceding context,*

$$Q(\zeta_a) \cap Q(\zeta_b) = Q(\zeta_d) \quad \textit{and} \quad Q(\zeta_a)Q(\zeta_b) = Q(\zeta_m).$$

Proof. We write $a = \alpha d$ and $b = \beta d$, where α and β are relatively prime.

(i) Let x, y be integers which satisfy $\alpha x + \beta y = 1$. With $ab = dm$ and $\alpha b = m, \zeta_m^\alpha$ is a primitive bth root of unity; thus,

$$\zeta_b = (\zeta_m^\alpha)^r \quad \text{and} \quad \zeta_m^\alpha = (\zeta_b)^u,$$

for some integers r, u. Similarly, ζ_m^β satisfies

$$\zeta_a = (\zeta_m^\beta)^s \quad \text{and} \quad \zeta_m^\beta = (\zeta_a)^v,$$

for some integers s, v. We find

$$Q(\zeta_a)Q(\zeta_b) \subset Q(\zeta_m) = Q(\zeta_m^{\alpha x+\beta y}) = Q(\zeta_b^{ux}\zeta_a^{vy}) \subset Q(\zeta_a)Q(\zeta_b)$$

and

$$Q(\zeta_a)Q(\zeta_b) = Q(\zeta_m). \tag{4}$$

(ii) Since $Q(\zeta_a)$ is a normal and separable field extension of Q, the dimension of $Q(\zeta_a)Q(\zeta_b)$ as a vector space over $Q(\zeta_b)$ equals the dimension of $Q(\zeta_a)$ as a vector space over $Q(\zeta_a) \cap Q(\zeta_b)$; e.g., see [2], pp. 196–197, Theorem 4. Thus, with (4), we have

$$[Q(\zeta_m) : Q(\zeta_b)] = [Q(\zeta_a) : Q(\zeta_a) \cap Q(\zeta_b)]. \tag{5}$$

From (3), (5), and (2), we obtain

$$\begin{aligned} &[Q(\zeta_a) \cap Q(\zeta_b) : Q] \\ &= [Q(\zeta_m) : Q(\zeta_b)][Q(\zeta_b) : Q][Q(\zeta_a) \cap Q(\zeta_b) : Q]/\phi(m) \\ &= [Q(\zeta_a) : Q(\zeta_a) \cap Q(\zeta_b)][Q(\zeta_a) \cap Q(\zeta_b) : Q]\phi(b)/\phi(m) \\ &= \phi(a)\phi(b)/\phi(m) = \phi(d). \end{aligned} \tag{6}$$

Since ζ_a^α and ζ_b^β are primitive dth roots of unity, we have

$$\zeta_d = (\zeta_a^\alpha)^j = (\zeta_b^\beta)^k,$$

for some integers j, k. This yields $Q(\zeta_d) \subset Q(\zeta_a) \cap Q(\zeta_b)$; we use (6) and (3) to complete the proof.

3. An isomorphism of Z^+ onto L. By Theorem 1, L is a sublattice of A and the mapping f_1 of Z^+ into L defined by $f_1(n) = Q(\zeta_n)$ satisfies

$$f_1(a \wedge b) = f_1(a) \wedge f_1(b) \text{ and } f_1(a \vee b) = f_1(a) \vee f_1(b).$$

Consequently, f_1 is a surjective lattice-homomorphism of Z^+ onto L.

THEOREM 2. *The mapping f_2, defined by* (1), *is a lattice-isomorphism of Z^+ onto L.*

Proof. Let E be the set of even positive integers. With

$$2a \wedge 2b = 2(a \wedge b) \text{ and } 2a \vee 2b = 2(a \vee b),$$

E is a sublattice of Z^+ and the mapping g_2 of Z^+ onto E defined by $g_2(n) = 2n$ is a lattice-isomorphism. Let g_1 be the restriction of f_1 to E. If n is odd, then $Q(\zeta_n) = Q(\zeta_{2n}) = g_1(2n)$. If n is even, then $Q(\zeta_n) = g_1(n)$. Thus, g_1 is a lattice-homomorphism of E onto L. To prove g_1 is one-to-one, suppose $g_1(2a) = g_1(2b)$. Then, we find

$$g_1(2d) = g_1(2a) \wedge g_1(2b) = g_1(2a) \vee g_1(2b) = g_1(2m)$$

and

$$\phi(2d) = [Q(\zeta_{2d}) : Q] = [Q(\zeta_{2m}) : Q] = \phi(2m).$$

We write $a = \alpha d$ and $b = \beta d$. Let P denote the product of the distinct primes which divide both $\alpha\beta$ and $2d$ (with $P = 1$ when $\alpha\beta$ and $2d$ are relatively prime). We use the formula of E. Prouhet ([1], p. 118) to obtain:

$$\phi(2d) = \phi(2m) = \phi(2\alpha\beta d) = \phi(\alpha\beta)\phi(2d)P/\phi(P),$$

$1 = \phi(\alpha\beta)P/\phi(P)$, $P = 1$, $\alpha\beta$ is odd, $\phi(\alpha\beta) = 1$, $\alpha\beta = 1$, and $2a = 2b$. Hence, g_1 is an isomorphism of E onto L. Since f_2 is the composite of the lattice-isomorphisms g_1 and g_2, this completes the proof.

References

1. L. E. Dickson, History of the Theory of Numbers, vol. 1, Chelsea, New York, 1971.
2. S. Lang, Algebra, Addison-Wesley, Reading, Mass., 1965.
3. S. MacLane and G. Birkhoff, Algebra, Macmillan, New York, 1967.

GEOMETRIC INEQUALITIES VIA THE POLAR MOMENT OF INERTIA

M. S. KLAMKIN, Ford Motor Company, Dearborn, Michigan

Introduction. Employing an inequality of Aczel, Carlitz [1] has given a rather simple and interesting proof of the Neuberg-Pedoe [2] two triangle inequality and two other such inequalities. We observe in this note that Cauchy's inequality suffices for the same purpose and in addition we give some extensions of these inequalities in two and higher dimensions with various weight factors.

Our starting point is the well-known result from mechanics that the polar moment of inertia of a system of weighted points is minimum about the centroid of that system.

Minimum polar moment of inertia. Consider a set, S, of weighted points $\{(V_i, w_i)\}$ $i = 1, 2, \cdots, n$ in Euclidean m-space, E^m, with origin O and let $\overrightarrow{OX} = \mathbf{X}$ where X is any point of E^m. If G is the centroid of S and $W = \Sigma w_i$ then

$$\mathbf{G} = \Sigma w_i \mathbf{V}_i / W.$$

The known results (these are usually attributed to Lagrange; however, they may have been known by Leibnitz) for the polar moment of inertia $I(P)$ of the set of weights about point P are given by

$$(1) \qquad I(P) = \Sigma w_i(\mathbf{V}_i - \mathbf{P})^2 = \Sigma w_i\{(\mathbf{V}_i - \mathbf{G}) - (\mathbf{P} - \mathbf{G})\}^2$$
$$= W(\mathbf{P} - \mathbf{G})^2 + \Sigma w_i(\mathbf{V}_i - \mathbf{G})^2,$$

$$I(P) \geqq \Sigma w_i(\mathbf{V}_i - \mathbf{G})^2, \tag{2}$$

with equality *iff* P coincides with G.

Letting $\mathbf{R}_i = \mathbf{V}_i - \mathbf{P}$, $A_{ij} = |\mathbf{V}_i - \mathbf{V}_j|$, the expanded forms of (1) and (2) become

$$2W\Sigma w_i\mathbf{R}_i^2 = 2W^2\overline{PG}^2 + \Sigma\Sigma w_i w_j A_{ij}^2, \tag{1'}$$

$$2W\Sigma w_i\mathbf{R}_i^2 \geqq \Sigma\Sigma w_i w_j A_{ij}^2. \tag{2'}$$

Note that (1′) and (2′) are valid for all real weights, so that (2′) is valid even when $W = 0$.

Triangle and simplex inequalities. If points V_i form a nondegenerate simplex with circumcenter P and circumradius R, then

$$R^2 = \overline{PG^2} + \frac{1}{2W^2}\Sigma\Sigma\, w_i w_j A_{ij}^2. \tag{3}$$

Now let (V_i', w_i) denote a second set of n weighted points. Then by Cauchy's inequality (assuming $w_i, w_i' \geqq 0$),

$$RR' \geqq \overline{PG}\cdot\overline{P'G'} + \frac{1}{2|WW'|}\Sigma\Sigma\sqrt{w_i w_j w_i' w_j'}\, A_{ij}A_{ij}'. \tag{4}$$

For $n = 3$, $w_i = w_i' = 1$, (4) reduces to an inequality of Carlitz referred to in the introduction. (4) can be extended to any number of sets of associated points and weights by using Hölder's inequality, e.g., for three sets, we would obtain an inequality for $(RR'R'')^{\frac{2}{3}}$.

For $n = 3$, (2′) reduces to

$$(w_1 + w_2 + w_3)(w_1R_1^2 + w_2R_2^2 + w_3R_3^2) \geqq (w_2w_3a_1^2 + w_3w_1a_2^2 + w_1w_2a_3^2) \tag{5}$$

where $R_i = |\mathbf{R}_i|$, $a_1 = A_{23}$, etc. This case corresponds to the known equivalent inequalities of Wolstenholme, Kooi, Oppenheim, Tomescu and Klamkin (see [**2**], [**3**], [**4**]). The numerous inequalities which follow from this case have been pointed out in [**3**].

If P is the orthocenter, we obtain another interesting special case of (5), i.e.,

$$\Sigma\, w_i \cos^2 A_i \geqq \{1 - \Sigma\cos^2 A_i\}\{w_2w_3 + w_3w_1 + w_1w_2\}$$

where A_1, A_2, A_3 are angles of a triangle. Then letting $w_i \cos A_i = x_i$ and noting that $1 - \Sigma\cos^2 A_i = 2\Pi\cos A_i$, we obtain

$$x_1^2 + x_2^2 + x_3^2 \geqq 2\{x_2x_3\cos A_1 + x_3x_1\cos A_2 + x_1x_2\cos A_3\} \tag{6}$$

which corresponds to an inequality of Barrow-Janić [**4**]. Another way to obtain (6) is to use (5) with P being the centroid of the triangle and note that the three medians of a triangle are congruent to sides of another triangle.

The two previous special cases also correspond to the special cases $n = 1, 2$ of the master inequality [**3**] (that these two are dual inequalities follows from the pedal triangle transformation $2A_i' = \pi - A_i$, etc.)

(7) $$x^2 + y^2 + z^2 \geqq (-1)^{n+1}\{2yz\cos nA + 2zx\cos nB + 2xy\cos nC\}$$

where n is integral. Other special cases of (5) are treated in [5].

For our last inequality, we derive one involving the volumes and circumradii of a pair of tetrahedra in a way almost identical with the Carlitz derivation of the Neuberg-Pedoe inequality. If $(a, a_1), (b, b_1), (c, c_1)$ denote the pairs of opposite edges of a tetrahedron, it is known that

$$(24VR)^2 = \{(aa_1)^2 + (bb_1)^2 + (cc_1)^2\}^2 - 2\{(aa_1)^4 + (bb_1)^4 + (cc_1)^4\}.$$

Then by Cauchy's inequality,

(8) $$24^2 VR\, V'R' \leqq \Sigma (a'a_1')^2\{-(aa_1)^2 + (bb_1)^2 + (cc_1)^2\}.$$

There is equality *iff* the triangles with side lengths (aa_1, bb_1, cc_1) and $(a'a_1', b'b_1', c'c_1')$ are directly similar.

Note added in proof. Bottema has shown that the equality condition for (8) is also equivalent to the two tetrahedra being invertible into each other.

References

1. L. Carlitz, Some inequalities for two triangles, this MAGAZINE, 45 (1972) 43–44.
2. O. Bottema, M. S. Klamkin, Joint triangle inequalities, Ford Motor Company Preprint, Sept. 1973.
3. M. S. Klamkin, Asymmetric triangle inequalities, Publ. Elektrotehn. Fak. Ser. Mat. Fiz. Univ. Beograd, No. 357–No. 380 (1971) 33–44.
4. O. Bottema, R. Ž. Djordjević, R. R. Janić, D. S. Mitrinović, P. M. Vasić, Geometric Inequalities, Wolters-Noordhoff, Groningen, 1969.
5. M. S. Klamkin, Non-negative quadratic forms and triangle inequalities, Ford Motor Company Preprint, June 1971. (Also, See Notices of Amer. Math. Soc., Oct. 1971, p. 96.)

ANOTHER GENERALIZATION OF THE BIRTHDAY PROBLEM

J. E. NYMANN, The University of Texas at El Paso

While discussing the birthday problem with a class of 100 students I decided to run an experiment with the class. Knowing that the probability that two of them would have the same birthday was very close to one, I decided on the following alternative. I asked each student, in succession, to call out his birthday and if anyone in the room had the same birthday we would stop. I offered to bet that the procedure would stop on or before the tenth student. Unfortunately (since it turns out the probability of my winning is .928) no one was willing to bet with me. The object of this note is to report on this type of generalization of the birthday problem.

For n and k positive integers with $k \leqq n$, $P(n, k)$ will denote the probability that in a group of n people, at least one pair have the same birthday with at least one such pair among the first k people. Determining $P(n, n)$ is, of course, the ordinary birthday problem and the assertion made in the preceding paragraph is that $P(100, 10) = .928$. In determining $P(n, k)$ we will make the standard assumptions that each person in the group can have his birthday on any one of the 365 days in a year (ignoring leap years) and that each day of the year is equally likely to be the person's birthday.

The sample space can be taken to be the 365^n n-tuples with integer components between 1 and 365 inclusive. The event in which we are interested is the set of n-tuples for which at least one of the first k components is repeated. By elementary counting techniques the number of elements in the complementary event (i.e., the set of n-tuples for which none of the first k components is repeated) is easily seen to be $365 \cdot 364 \cdots (365 - k + 1)(365 - k)^{n-k}$. Hence

$$P(n,k) = 1 - \frac{365 \cdot 364 \cdots (365 - k + 1)(365 - k)^{n-k}}{365^n}.$$

From knowing the solution to the ordinary birthday problem we see that if $Q_k = 1 - P(k, k)$, then

$$P(n,k) = 1 - Q_k\left(1 - \frac{k}{365}\right)^{n-k}.$$

In the table below $k(n)$ will denote the smallest value of k for which $P(n, k) \geqq 1/2$. From the solution to the ordinary birthday problem it is clear that n must be at least 23 for $k(n)$ to exist. It is also clear that $k(n + 1) \leqq k(n)$. Consequently the table begins with $n = 23$ and only those n for which $k(n)$ changes are entered. This table is constructed from a computer printout which gave $P(n, k)$ for n between 23 and 100 and $1 \leqq k \leqq n$. The entries for $n > 100$ were done separately in order to complete this table.

n	23	24	25	26	27	28	29	31	33	36	40	46	54	66	86	128	254
$k(n)$	20	17	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

The last entry in this table is at first glance rather surprising. It says that if you were to bet that in a group of n people someone else has the same birthday as you, n must be at least 254 for this to be a favorable bet. One would perhaps think, initially, that this entry should be $1 + [365/2] = 183$. However, by the solution to the ordinary birthday problem, in a group of 183 people there would be many repeated birthdays; hence, not nearly half the birthdays would be represented. Clearly, the bet described above will have favorable odds when at least half of the possible birthdays are represented among the n people. Therefore, the last entry in the table seems to show that it takes at least 254 people before the expected number of birthdays represented is 183.

NOTES AND COMMENTS

Leon Gerber writes that *Slicing boxes into cubelets*, March 1974, page 101–103, appears to duplicate the solution to Leo Moser's problem 102, this MAGAZINE 25 (1952), p. 219.

Shouro Kasahara has submitted the following even simpler proof of the main theorem (due to Edelstein) of the paper, *On a fixed point theorem for compact metric spaces* by Bennett and Fisher in the January issue. The theorem and his simpler proof are as follows:

THEOREM. *Let T be a mapping of a compact metric space (X,ρ) into itself satisfying the condition*

$$(*) \qquad \rho(Tx, Ty) < \rho(x,y),\ x \neq y;$$

then there exist a unique point z in X such that $Tz = z$.

Proof. Consider the real valued function f on X defined by $f(x) = \rho(x, Tx)$ for every x in X. If $x \neq y$ then since

$$\begin{aligned}\rho(x, Tx) &\leqq \rho(x,y) + \rho(y,Ty) + \rho(Ty,Tx)\\ &< 2\rho(x,y) + \rho(y,Ty)\end{aligned}$$

by (*), 2e have $|f(x) - f(y)| \leqq 2\rho(x,y)$ for every x, y in X which shows that f is continuous. Hence f assumes its minimum value at some z in X. Now if $Tz \neq z$ then (*) implies

$$f(Tz) = \rho(Tz, T^2z)\ \ \rho(z, Tz) = f(z),$$

which is absurd. We have thus $Tz = z$ as desired.

Roger W. Pease, Jr. writes as follows regarding the paper, *An occupancy problem involving placement of pairs of balls* by Wiggins in the March 1972 issue. A correct result in section 3 is obtained by a faulty method. The trouble begins after equation 3.1 with the phrase "average number of additional available cell pairs" which should read something like "average number of ways in which one cell pair can be added as." In the next paragraph the following two sentences cause trouble: "The cardinality of this set is $M_{n,k}^{(2)}$. Obviously each sequence of $\mathscr{S}_{n,k}^{(2)}$ will yield a sequence of $\mathscr{T}_{n\,k+1}^{(2)}$." The cardinality of $\mathscr{S}_{n,k}^{(2)}$ is not necessarily the same as that of $M_{n,k}^{(2)}$ because it is possible that one arrangement of $\mathscr{S}_{n,k}^{(2)}$ may yield several arrangements included in the sum $M_{n,k}^{(2)}$. For example, if there is a run of four empty cells, one pair of adjacent balls can be inserted in three ways. The mapping of $\mathscr{S}_{n.k}^{(2)}$ to $\mathscr{T}_{n,k+1}^{(2)}$ is not one-to-one and onto and $\mathscr{S}_{n,k}^{(2)}$ and $\mathscr{T}_{n,k+1}^{(2)}$ do not have the same cardinality. However, the final result is correct if the last sentence on page 84 is changed to read "Finally, substituting in (3.2) we have the average number of ways in which a single adjacent cell pair can be inserted (or the average number of ways in which a couple can be seated together at the concert)."

This error in Wiggins' paper was noted by the referee and was corrected by the author but the editor bungled and the incorrect version was sent to the printer and published. We apologize to the author, the referee and our readers. A paper by Roger W. Pease, Jr., dealing with a generalization of Wiggins' problem will appear in a forthcoming issue.

Simeon Reich comments as follows on his paper *On a problem in number theory* in the November 1971 issue:

For any natural number k let $g(k)$ be the smallest integer with the following property: If $n > g(k)$, then there are at least k composite natural numbers smaller than n and relatively prime to it. The existence of g is established in a simple manner in my note *On a problem in number theory*, this MAGAZINE, 44 (1971) 277–278. During a conversation Professor Paul Erdos remarked that probably $g(k) = g(k+1)$ infinitely often. In a subsequent letter I strengthened his conjecture in the following way: For any natural number m let $c(m)$ denote the cardinality of $g^{-1}(\{m\})$. Then $\sup \{c(m) : m = 1,2,\ldots\} = \infty$. In his reply (1972), Professor Erdos proved this latter conjecture using results from his joint paper with L. Alaoglu, *On highly composite and similar numbers*, Trans. Amer. Math. Soc., 56 (1944) 448–469. He also observed that an asymptotic formula for $g(k)$ can be derived from Landau's estimate of Euler's function, thus answering a query which appeared implicitly in the last paragraph of my abovementioned note.

BOOK REVIEWS

EDITED BY ADA PELUSO AND WILLIAM WOOTON

Materials intended for review should be sent to: Professor Ada Peluso, Department of Mathematics, Hunter College of CUNY, 695 Park Avenue, New York, New York 10021, or to Professor William Wooton, 1495 La Linda Drive, Lake San Marcos, California 92069. A boldface capital C in the margin indicates that a review is based in part on classroom use.

C *A Fortran IV Primer.* By Richard A. Mann. Intext Education Publishers, 1972. 207 pp.

Considering the large number of texts on the market dealing with FORTRAN IV programming, the need for another might be questioned. Yet, *A FORTRAN IV Primer* expounds the subject better than most, has a large number of desirable features, and is generally very well suited for a short course in FORTRAN programming.

Perhaps this text's strongest point is its use of examples. There are an unusually large number of these, employed whenever a new concept is introduced. They show not only valid forms, but invalid ones as well, the latter containing various and sundry common errors. In addition to the shorter examples, the text contains several sample programs, complete with flowcharts, coding, and output.

The author follows a fairly standard approach to the subject. The first two chapters are introductory in nature, dealing with the setting up of a computation, the technique of flow charting, and a brief description of the basic types of programming languages. The remaining ten chapters cover the elements of the FORTRAN IV language. Appendices dealing with material of lesser importance (for example, complex operations) are also included. In classroom use, it is possible to begin with Chapter 3 and introduce most of the ideas of Chapters 1 and 2 while covering the rest of the material. This, together with an "introduction to output" section in Chapter 4, would enable the student to get "on the machine" early in the course.

A FORTRAN IV Primer contains a number of nice touches which make for pleasant reading and efficient use. As an example, some of the limitations of the computer are illustrated by means of a not too bright instruction following "clerk", and some of the basics of programming by means of an omelet recipe. Also, a chart of "important FORTRAN forms" is given on the outside back cover for easy reference.

The book does leave room for improvement. An easily corrected oversight is the lack of answers to the exercises at the end of each chapter. These do provide good practice for the student, but he has no immediate way of determining whether or not his answers are correct. A more basic defect in the text is the poor organization of the contents. For example, material on arithmetic statements is covered first in Chapter 3 and then again in Chapter 6 for no apparent reason; and the different types of subprograms are found in Chapters 9, 10, 11, and 12, together with completely unrelated material. The instructor should be prepared to skip around.

I found that almost all of the material in Chapters 1–12 could be covered in a class meeting 100 minutes per week for 10 weeks. The author does not assume any knowledge of calculus or linear algebra, and consequently any mathematical deficiencies of the student have little effect on the matter of learning to program.

STEWART VENIT, California State University, Los Angeles

PROBLEMS AND SOLUTIONS

Readers of this department are invited to submit for solution problems believed to be new that may arise in study, in research, or in extra-academic situations. Problems may be submitted from any branch of mathematics and ranging in subject content from that accessible to the talented high school student to problems challenging to the professional mathematician. Proposals should be accompanied by solutions, when available, and by any information that will assist the editor. Ordinarily, problems in well-known textbooks should not be submitted.

The asterisk (*) will be placed by the problem number to indicate that the proposer did not supply a solution. Readers' solutions are solicited for all problems. Proposers' solutions may not be "best possible" and solutions by others will be given preference.

Solutions should be submitted on separate, signed sheets. Figures should be drawn in India ink and exactly the size desired for reproduction.

Send all communications for this department to Dan Eustice, The Ohio State University, 231 W. 18th Ave., Columbus, Ohio 43210.

To be considered for publication, solutions should be mailed before August 1, 1975.

PROPOSALS

922. *Proposed by Alan Schwartz, University of Missouri–St. Louis.*

Let $\{x_n\}$ be a sequence of nonnegative numbers satisfying

$$\sum_{n=0}^{\infty} x_n x_{n+k} \leqq C x_k$$

for some constant C and $k = 0, 1, 2, \cdots$. Prove that Σx_n converges. (Is the result still true if $k = 0, 1, 2, \cdots$ is replaced with $k = 1, 2, \cdots$?—Ed.)

923. *Proposed by Aron Pinker, Frostburg State College.*

If r and s are roots of $x^2 + px + q = 0$, where p and q are integers with $q \mid p^2$, then $(r^n + s^n)/q$ is an integer for $n = 2, 3, \cdots$.

924. *Proposed by J. Michael McVoy and Anton Glaser, Pennsylvania State University.*

How many n-tuples, $(S_1, \cdots, S_n)$, exist with $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_n \subseteq V$, where V is a set of k elements?

925. a. *Proposed by Julius G. Baron, Rye, New York.*

Prove that any non-self-intersecting cyclic octagon is such that the sum of any four nonadjacent interior angles is 3π.

b. *Proposed by Thomas E. Elsner, General Motors Institute.*

An octagon is inscribed in a circle with vertices on any four diameters. Show that each alternate pair of exterior angles is complementary.

926. *Proposed by Melvin F. Gardner, University of Toronto.*

A swimmer can swim with speed v in still water. He is required to swim for a given length of time T in a stream whose speed is $r < v$. If he is also required to start and finish at the same point, what is the longest path (total arc length) that he can complete? Assume the path is continuous with piecewise continuous first derivatives.

927. *Proposed by Roy Dubisch, University of Washington.*

Pick's formula for the area of polygonal regions whose vertices are lattice points is $\frac{1}{2}b + i - 1$ where b is the number of lattice points on the boundary and i is the number of lattice points in the interior. Show that no such formula exists for the volume of polyhedra whose vertices are lattice points even if we allow as variables, in addition to b and i, e = the number of edges, f = the number of faces, and i' = the number of lattice points in the interior of the faces.

928. *Proposed by Norman Schaumberger, Bronx Community College.*

If k is a positive integer, prove that

$$\lim_{n\to\infty} \frac{1}{n^{k+1}} \sum_{j=1}^{n} \cot^k(1/j) = \frac{1}{k+1}.$$

QUICKIES

From time to time this department will publish problems which may be solved by laborious methods, but which with the proper insight may be disposed of with dispatch. Readers are urged to submit their favorite problems of this type, together with the elegant solution and the source, if known.

Q608. If x, y, z are nonnegative and are not sides of a triangle, show that

$$1 + \frac{x}{y+z-x} + \frac{y}{z+x-y} + \frac{z}{x+y-z} \leqq 0.$$

[*Submitted by M. S. Klamkin.*]

Q609. Which real functions satisfy $f(x+y)^2 = f(x)^2 + f(y)^2$?

[*Submitted by Julian H. Blau.*]

Q610. Maximize $(7+x)(11-3x)^{1/3}$

[*Submitted by C. F. Pinzka.*]

Q611. Prove that the only solutions to the Diophantine equation, $x^3 - 2 = 6y^2$, are $x = 2$, $y = \pm 1$.

[*Submitted by Erwin Just.*]

Q612. Let $(G, \cdot)$ be a group with the following special cancellation property:

$$x \cdot a \cdot y = b \cdot a \cdot c \text{ implies } x \cdot y = b \cdot c$$

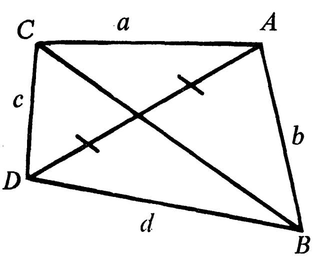
for all x, y, b, c and a in G. Prove that G is Abelian.

[*Submitted by Kenneth Taylor.*]

Q613. Prove or disprove:

If $a + b = c + d$, then $a = c$ or $a = d$.

[*Submitted by Sidney Penner.*]



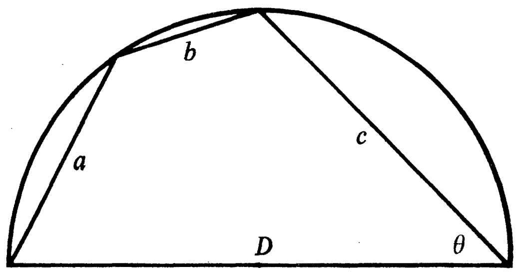
(Answers on page 58)

SOLUTIONS

Semicircular Chords

880. [November, 1973] *Proposed by Richard Corry, Clay Center, Kansas.*

Given chords a, b, c in a semicircle. Determine the diameter (D) of the circle.



I. *Solution by W. M. Sanders, Madison College, Virginia.*

Let m and n represent the diagonals of the inscribed polygon with $m^2 = D^2 - a^2$ and $n^2 = D^2 - c^2$. The relation $mn = ac + bD$ is a consequence of Ptolemy's theorem for cyclic quadrilaterals. This relation implies that D must be the positive root of the reduced cubic $D^3 - (a^2 + b^2 + c^2)D - 2abc = 0$.

II. *Solution by Thomas E. Elsner, General Motors Institute.*

Let d be the chord completing an inscribed right triangle with c and D so that $D^2 = c^2 + d^2$. Let θ be the angle shown and since the quadrilateral is cyclic the opposite angle is $\pi - \theta$. By the cosine law $d^2 = a^2 + b^2 - 2ab\cos(\pi - \theta)$, but $\cos(\pi - \theta) = -\cos\theta = -c/D$ and so D is a solution of the cubic

$$D^3 - (a^2 + b^2 + c^2)D - 2abc = 0.$$

Note. The solution of the cubic is overly tedious. Furthermore, this is an old problem appearing as an exercise in Eves' *Survey of Geometry*, p. 138 and earlier in the February 1944 AMER. MATH. MONTHLY.

Also solved by Leon Bankoff, Los Angeles, California; Martin Berman, Bronx Community College; Brother Alfred Brousseau, St. Mary's College; Donald H. Byerly, Westtown School, Westtown, Pa.; Santo M. Diano, Havertown, Pa.; Clayton W. Dodge, University of Maine at Orono; Gerson B. Robison, State University College, New Paltz, New York; Ralph Garfield, The College of Insurance, New York, New York; J. Garfunkel, Flushing, New York; Richard A. Gibbs, Fort Lewis College; George Gruber, Brooklyn, New York; J. A. H. Hunter, Toronto, Canada; Ralph Jones, University of Massachusetts; Henry S. Lieberman, John Hancock Mutual Life Insurance Company; V. Linis, University of Ottawa; Lance M. Lund, San Francisco State University; Joseph V. Michalowicz; John M. O'Malley, Jr., Sharon, Massachusetts; C. F. Pinzka, University of Cincinnati; Paul Smith, University of Victoria; Robert S. Stacy, Manzano H. S., Albuquerque, N. M.; Eric Sturley, Southern Illinois University, Edwardsville; Steven Szabo, Urbana, Illinois; Kenneth M. Wilke, Topeka, Kansas; Atila Yanik, Illinois Institute of Technology; Alexander Zujus; and the proposer.

A Root Series

881. [November, 1973] *Proposed by Raphael T. Coffman, Richland, Washington.*

A root of the equation $x^3 - 2x^2 - 5x - 4 = 0$ correct to two decimal places is 3.66. Show how to find this root using only a series, the first few terms of which are: $1 + 2 + 3 + 20 + 63 + 238 + 871 \cdots$ and show how to generate this series.

Solution by Brother Alfred Brousseau, St. Mary's College.

Form a sequence with terms $T_1 = 1$, $T_2 = 2$, $T_3 = 3$ and a recursion relation

$$T_{n+1} = 2T_n + 5T_{n-1} + 4T_{n-2}$$

following the coefficients of the transposed equation. Then the terms of this sequence give the quantities in the series. The ratio of two successive terms approaches the given root.

$$T_{20}/T_{19} = 18597639556/5077054195 = 3.66307682,$$

$$T_{21}/T_{20} = 68124582511/18597639556 = 3.66307682.$$

Also solved by the proposer.

A Prime Magic Square

882. [November, 1973] *Proposed by Charles W. Trigg, San Diego, California.*

One magic square with prime elements is

101	5	71
29	59	89
47	113	17

Form another third order magic square with positive prime elements, six of which are the same as those in the given square.

Solution by Bob Prielipp, The University of Wisconsin-Oshkosh.

The following third order magic square, which is found on p. 104 of Maxfield and Maxfield, *Discovering Number Theory*, W. B. Saunders Company, 1972, satisfies the given conditions:

83	29	101
89	71	53
41	113	59.

This magic square has "magic constant" 213. It may be of interest to note that the given magic square is the third order magic square with positive prime elements which has the smallest "magic constant" (177).

Also solved by Chuck Friesen, Lincoln, Nebraska; Meredythe Grey, Bucknell College; Bernard G. Hoerbelt; J. A. H. Hunter, Toronto, Canada; Ralph Jones; Joseph V. Michalowicz; Kenneth M. Wilke; and the proposer.

Constant Sequence

883. [November, 1973] *Proposed by Harry Pollard, Purdue University.*

Prove that if the sequence $\{N_n\}$ satisfies the recursion

$$1 + N_{n+1}^{-3} = \frac{9(N_n^2 - N_n + 1)}{(N_n - 2)^3}$$

and $N_n > 2$ then the sequence is constant.

Solution by Harald Ziehms, Naval Postgraduate School.

Let $R^* = (-\infty, +\infty) \cup \{\infty\}$, and define the function $f\colon R^* \to R^*$ by

$$F(x) = \begin{cases} (x-2)[9(x^2-x+1)-(x-2)^3]^{-1/3}, & -\infty < x < +\infty, x \neq x_p \\ \infty & ,x = x_p \\ -1 & ,x = \infty, \end{cases}$$

where x_p is the unique solution to the equation

$$9(x^2-x+1)-(x-2)^3 = 0.$$

[*Note*: $x_p \doteq 13.5420$]

Define $f^{(0)}(x) = x$, and for $j = 1, 2, \cdots$ let $f^{(j)}$ be the j-fold composition of f with itself; then $f^{(j)}(N_n) = N_{n+j}$, $j = 0, 1, \cdots$.

The function f is continuous except at x_p; it increases monotonely from -1 to $+\infty$ as x increases from $-\infty$ to x_p, and from $-\infty$ to -1 as x increases from x_p to $+\infty$; hence f is one-to-one and onto, and the inverse function $f^{(-1)}$ exists.

Let $x_a = 13$ and $x_c = f^{(-1)}(x_p) \doteq 13.5383$. Straightforward calculations show that for $x \in (-\infty, x_a] \cup [x_c, +\infty) \cup \{\infty\}$, $f^{(3)}(x) \leqq 2$; therefore, $N_n \notin (-\infty, x_a] \cup [x_c, +\infty) \cup \{\infty\}$.

The following additional facts can be established:

$$f(x_a) < x_a,\ f(x_c) > x_c,\ f'(x_a) > 1,\ f''(x) > 0 \text{ for } x \in [x_a, x_c].$$

Therefore there exists one, and only one, number x_b in the interval (x_a, x_c) such that $f(x_b) = x_b$ [$x_b \doteq 13.5382$].

From these facts it follows that for $x \in (x_a, x_b)$, $f(x) < x$, and if $x, f^{(1)}(x), f^{(2)}(x)$ are all in (x_a, x_b) then $f^{(1)}(x) - f^{(2)}(x) > x - f^{(1)}(x)$; hence for every $x \in (x_a, x_b)$ there exists a number k such that $f^{(k)}(x) \leqq x_a$ and thus $f^{(k+3)}(x) \leqq 2$. Similarly, for every $x \in (x_b, x_c)$ there exists a number M such that $f^{(M)}(x) \geqq x_c$ and thus $f^{(M+3)}(x) \leqq 2$. Therefore, $N_n \notin (x_a, x_b) \cup (x_b, x_c)$.

Thus N_n can only be x_b, and since $f^{(j)}(x_b) = x_b$ for all $j = 0, 1, \cdots, N_n = N_{n+j}$ and the sequence is constant.

[*Note*: If N_n takes on any value in $(-\infty, +\infty)$ other than x_b, then the sequence converges to $x_0 \doteq 0.7936$]

Also solved by Brother Alfred Brousseau, St. Mary's College; Robert S. Stacy, Manzano H. S., Albuquerque, N. M.; and the proposer.

Frequency of a Sine Curve

884. [November, 1973] *Proposed by K. W. Schmidt, University of Manitoba.*

A sine curve, parallel to the x-axis, has between a peak and an adjacent trough 4 points with equidistant x-components. Given the space between the x-components, and any 2 of the differences between the y-components, find the shortest possible expression for the frequency.

I. *Solution by Brother Alfred Brousseau, St. Mary's College, Moraga, California.*

Let the value of the sine function be given by

$$y = A\sin(2\pi v x + B)$$

with $x_{i+1} - x_i = d$ for $i = 1,2,3$.

Let

$$\alpha = y_4 - y_1 = 2A\cos[\pi v(x_1 + x_4 + B)]\sin 3\pi v d,$$

$$\beta = y_3 - y_2 = 2A\cos[\pi v(x_2 + x_3 + B)]\sin \pi v d.$$

Since $x_2 + x_3 = x_1 + x_4$, $\alpha/\beta = (\sin 3\pi v d)/\sin \pi v d = 3 - 4\sin^2 \pi v d$,

$$\sin^2 \pi v d = (3\beta - \alpha)/4\beta,$$

$$v = (1/\pi d)\arcsin\sqrt{(3\beta - \alpha)4\beta}\,.$$

II. *Solution by the proposer.*

Substituting the x-values

$$x_1 = a \qquad x_3 = a + 2D$$

$$x_2 = a + D \qquad x_4 = a + 3D$$

in

$$y = C_1 + C_2 e^{jwx} + C_3 e^{-jwx},$$

where C_1, C_2, C_3, w are unknowns, yields

$$Q = (y_4 - y_1)/(y_3 - y_2)$$

$$= \frac{C_2 e^{jw3D} + C_3 e^{-jw3D} - C_2 - C_3}{C_2 e^{jw2D} + C_3 e^{-jw2D} - C_2 e^{jwD} - C_3 e^{-jwD}}$$

$$= 1 + e^{jwD} + e^{-jwD}$$

$$= 1 + 2\cos wD.$$

Since $3wD > \pi$, it follows that

$$\text{frequency } w = D^{-1}\cos^{-1}((Q-1)/2).$$

(With $w = 2\pi v$ the two solutions can be shown equivalent.—Ed.)

A Sum of Permutations

885. [November, 1973] *Proposed by Stephen B. Maurer, Phillips Exeter Academy.*

Let Z_n be an additive group of integers modulo n. For what values of n do there exist permutations $f, g: Z_n \to Z_n$ such that $f + g$ is a permutation also?

Solution by J. C. Binz, Berne, Switzerland.

Such permutations exist if and only if n is odd.

Proof: If $f + g$ is a permutation, then $\{z; z = f(r) + g(r),\ r \in Z_n\} = Z_n$. Let $\sum_{i=0}^{n-1} i = s$; then $\sum_{i=0}^{n-1} [f(i) + g(i)] = s$ implies $2s = s$ and $s = 0$. But $s = 0$ if and only if n is odd.

Conversely, for an odd n, $e + e$ is the permutation

$$\begin{pmatrix} 0 & 1 \cdots 1/2(n-1) & 1/2(n+1) \cdots n-1 \\ 0 & 2 \cdots n-1 & 1 \cdots n-2 \end{pmatrix}$$

where e is the identity permutation.

Also solved by F. David Hammer, Stockton State College; Peter W. Lindstrom, St. Anselm's College; Joseph V. Michalowicz; Lloyd N. Trefethen, Harvard University; and the proposer.

Conjecture

886.* [November, 1973] *Proposed by Doug Engel, Denver, Colorado.*

In a sequence of positive integers, $N_{k+1} = N_k +$ the sum of all the distinct prime factors of N_k including 1 and N_k, $k = 0, 1, 2, \cdots$. Such a sequence is 1, 2, 3, 4, 7, 8, 11, 12, 18, 24 $\cdots$. The sequence 5, 6, 12, 18 $\cdots$ merges with the first sequence as do all sequences with $N_0 < 91$.

It is conjectured that such sequences with N_0 a positive integer merge with the basic sequence that begins with 1. Prove or disprove this conjecture. If disproven, how many independent sequences exist?

Editor: No solutions received to this problem. C. W. Trigg comments that all $N_0 < 105$ merge before or at the 24th term of the basic sequence except the sequence for 91 merges at the 38th term. We would welcome further comments.

ANSWERS

A608. We will show more generally that if x_i $(i = 1, 2, \cdots, n)$ are nonnegative and are not sides of an n-gon, then

$$1 + \sum_{i=1}^{n} \frac{x_i}{S - 2x_i} \leqq 0,$$

where $S = \Sigma x_i$. Assume that $2x_n > S$ and let $x_n = T + 2a$ where $T = x_1 + x_2 + \cdots + x_{n-1}$ and $a > 0$. Then we have to show equivalently that

$$\frac{T+2a}{2a} \geqq 1 + \sum_{i=1}^{n-1} \frac{x_i}{2(a + T - x_i)}$$

which follows from

$$\frac{x_i}{2a} \geqq \frac{x_i}{2(a + T - x_i)}.$$

A609. Only $f \equiv 0$. Evidently $f(0) = 0$. Then $f(x)^2 + f(-x)^2 = 0$, whence $f(x) = 0$.

A610. Cubing, we get $(7 + x)^3(11 - 3x)$, the product of four numbers with the constant sum 32. This will be a maximum when each number is 8, giving $x = 1$ and 16 for the maximum of the original expression.

A611. The given equation is equivalent to $(1 + y)^3 + (1 - y)^3 = x^3$. The conclusion follows after an inspection of the trivial solutions, which are the only solutions, of this equation.

A612. If e is the identity of the group then

$$(x \cdot y) \cdot x \cdot e = e \cdot x \cdot (y \cdot x).$$

The cancellation property allows us to remove the middle x and the result follows.

A613. Consider the ellipse with foci at B and C with BC as a segment of the major axis and $r = a + b = c + d$ as defining constant. In order for AD to be bisected it must either intersect BC at the center of the ellipse or be perpendicular to BC. In either case the result follows from the symmetries of the ellipse.

(Quickies on page 52)