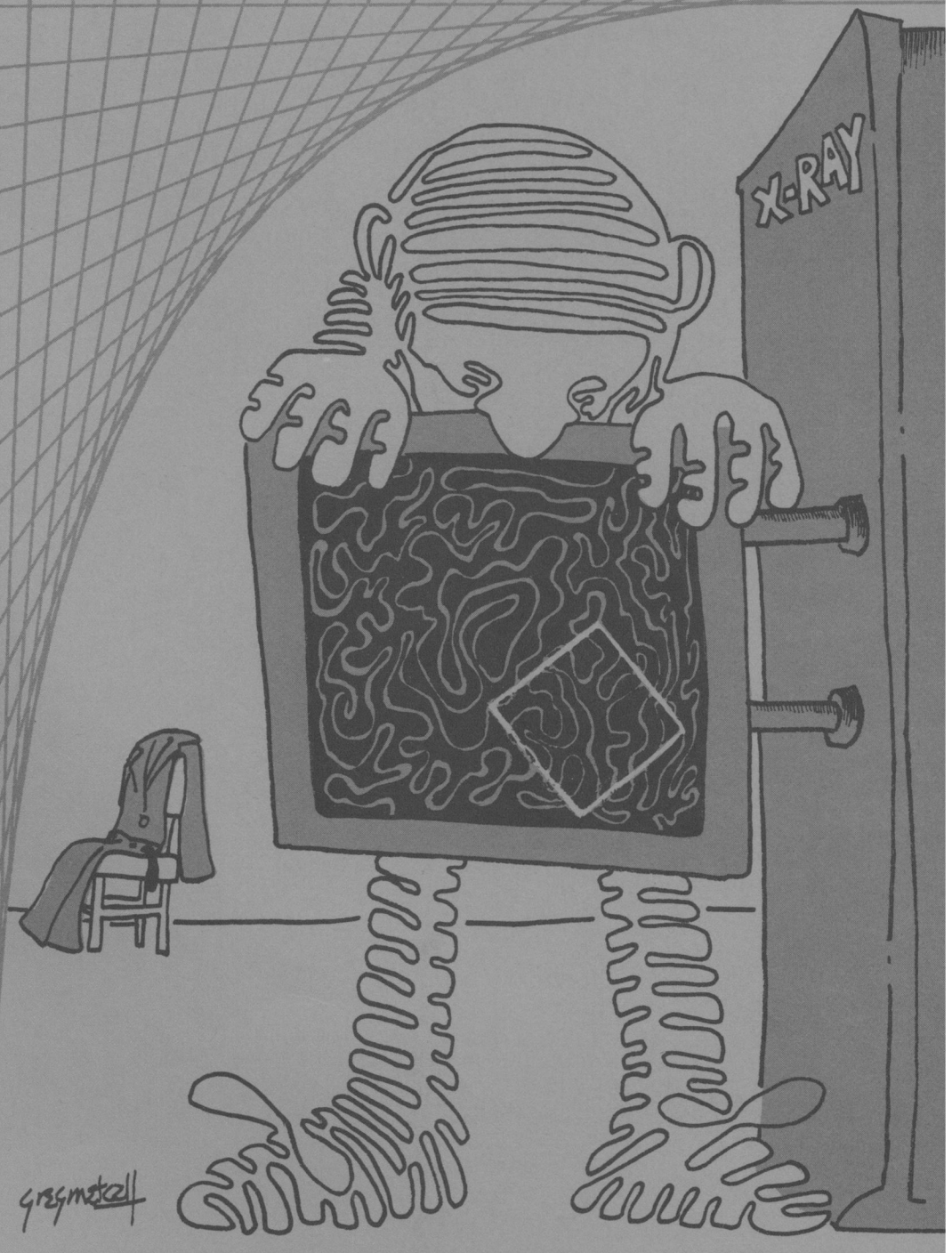
MATHEMATICS MAGAZINE



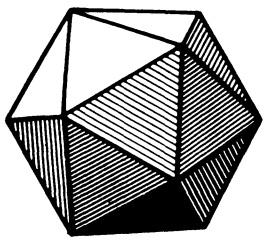
Vol. 52, No. 3
May, 1979

STOKES' THEOREM • GEOMETRY PROBLEMS

UNUSUAL SPHERES • COMPUTER DESIGNS

MATHEMATICS MAGAZINE

Volume 52, Number 3, May 1979
(ISSN—0025-570X)



COVER: Does every simple closed curve in the plane contain the vertices of some square? This superficially simple question remains unanswered, as do many other geometric questions which should have simple answers (see pp. 131–145).

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

EDITORIAL POLICY

Mathematics Magazine is a journal of collegiate mathematics designed to enrich undergraduate study of the mathematical sciences. The *Magazine* should be an inviting, informal journal emphasizing good mathematical exposition of interest to undergraduate students. Manuscripts accepted for publication in the *Magazine* should be written in a clear and lively expository style. The *Magazine* is not a research journal, so papers written in the terse "theorem-proof-corollary-remark" style will ordinarily be unsuitable for publication. Articles printed in the *Magazine* should be of a quality and level that makes it realistic for teachers to use them to supplement their regular courses. The editors especially invite manuscripts that provide insight into applications and history of mathematics. We welcome other informal contributions, for example, brief notes, mathematical games, graphics and humor.

Editorial correspondence should be sent to: Mathematics Magazine, Department of Mathematics, St. Olaf College, Northfield, Minnesota 55057. Manuscripts should be prepared in a style consistent with the format of *Mathematics Magazine*. They should be typewritten and double spaced on $8\frac{1}{2}$ by 11 paper. Authors should submit the original and one copy and keep one copy as protection against possible loss. Illustrations should be carefully prepared on separate sheets of paper in black ink, the original without lettering and two copies with lettering added; the printers will insert printed letters on the illustration in the appropriate locations.

Authors planning to submit manuscripts may find it helpful to obtain the more detailed statement of guidelines available from the editorial office.

ABOUT OUR AUTHORS

Victor Klee ("Some Unsolved Problems in Plane Geometry") is a Professor of Mathematics and Applied Mathematics, and Adjunct Professor of Computer Science at the University of Washington in Seattle. He is also a former President of the Mathematical Association of America and in 1977 received the MAA's Award for Distinguished Service to Mathematics. The article is closely related to the first of his MAA films on "Shapes of the Future." He was led to produce the films and the article, and in 1969 to start the Research Problems section of the *American Mathematical Monthly*, by his fascination with the frontiers of "elementary" mathematics and his belief that familiarity with those frontiers is an important part of mathematical culture.

Victor J. Katz ("The History of Stokes' Theorem") received his Ph.D. from Brandeis University in 1968 with a specialization in algebraic number theory. He has been at the University of the District of Columbia (formerly Federal City College) since 1968. He writes that in addition to his interest in algebraic number theory, "I have always been interested in the history of mathematics, especially since attending a summer seminar in the field in 1972. After reading the statement in E. T. Bell's *The Development of Mathematics* that a 'concise report on what has been done on [Stokes' Theorem] would occupy a chapter,' I decided to attempt to write that chapter."

BUSINESS INFORMATION. Mathematics Magazine is published by the Mathematical Association of America at Washington, D.C., five times a year in January, March, May, September, and November. Ordinary subscriptions are $12 per year. Members of the Mathematical Association of America or of Mu Alpha Theta may subscribe at special reduced rates. Colleges and university mathematics departments may purchase bulk subscriptions (5 or more copies to a single address) for distribution to undergraduate students.

Subscription correspondence and notice of change of address should be sent to A. B. Willcox, Executive Director, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Back issues may be purchased, when in print, from P. and H. Bliss Co., Middletown, Connecticut 06457.

Advertising correspondence should be addressed to Raoul Hailpern, Mathematical Association of America, SUNY at Buffalo, Buffalo, New York 14214.

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

Some Unsolved Problems in Plane Geometry

A collection of simply stated problems that deserve equally simple solutions.

VICTOR KLEE
University of Washington
Seattle, WA 98195

If $S(t)$ is the number of mathematical problems that have been solved up to time t, and $U(t)$ is the number that have been explicitly considered but still remain unsolved, then probably $U(t)/S(t)\to\infty$ as $t\to\infty$. Yet most mathematical expositions concentrate on the denominator and ignore the much larger numerator. I like to redress the balance by talking or writing about unsolved problems, especially ones that have immediate intuitive appeal and can be understood with little background. The problems presented here are all concerned with the geometry of the Euclidean plane, which is almost as fertile a source of such problems as number theory and combinatorics. The problems are not new (one dates from 1916) but probably will be new to some readers. It will be interesting to see how many years pass before the problems are solved.

This paper is divided into two parts. In the first part, the problems are presented and some background information is provided. I hope this part will be accessible to all readers of the *Magazine*. The second part contains references and additional comments which are more advanced than the material of the first part. Nevertheless, it should be accessible to most readers.

I should emphasize that there is a tremendous number and variety of "elementary" unsolved problems in plane geometry. The small sample presented here consists of the ones that I find most appealing. This paper is an adaptation of [KH]. It is dedicated to Hugo Hadwiger for his seventieth birthday and to Paul Erdős for his sixty-fifth. The writings of Hadwiger and of Erdős are ideal sources of additional problems.

Problems

1. A colorful problem

With only two colors, you can paint an entire line so that no two points at unit distance receive the same color. Simply use red for each half-open interval $[n,n+1[$ when n is even, and white when n is odd. That's easy to do mathematically. However, with real paint it would be difficult to paint half-open rather than closed unit intervals, and to do the latter would produce two points at unit distance with the same color. Even with real paint, three colors suffice. For each integer n, use red for the half-open interval $[6n/3,(6n+2)/3[$, white for $[(6n+2)/3, (6n+4)/3[$, and blue for $[(6n+4)/3, (6n+6)/3[$. A little sloppiness at ends of the interval doesn't matter, because no number in the interval $]2/3,4/3[$ is realized as the distance between two points that receive the same color.

What happens in the Euclidean plane E^2? That is:

(A) *If all the points of the plane are to be painted so that no two points at unit distance receive the same color, what is the minimum number c of colors that can be used?*

Though the exact value of c is unknown, it's easy to see $4 \leqslant c \leqslant 7$. To prove $c \geqslant 4$, we show three colors are not sufficient, and to prove $c \leqslant 7$, we tell how to paint the plane in seven colors so that no two points at unit distance receive the same color.

Suppose the plane is painted in three colors—say red, white and blue—so that no two points at unit distance receive the same color. For an arbitrary red point r, consider the configuration shown in FIGURE 1. The triangles are equilateral with side length 1, so the other two vertices of r's triangle are colored differently from r and from each other. Thus one is white and one blue, whence r' is red. By rotating the rhombus about r, we obtain an entire circle of red points r'. The circle includes two points at unit distance, and the resulting contradiction shows $c \geqslant 4$.

For an upper bound on c, consider the tessellation of the plane by regular hexagons of side length $2/5$. Paint one hexagon with color 7, and its six neighbors with colors 1 through 6. As the entire plane is tessellated by translates of this configuration, we can simply extend the coloring "by translation" (see FIGURE 2) and it turns out that no two points at unit distance receive the same color. Thus $c \leqslant 7$. We know, then, that c has one of the four values 4, 5, 6 and 7. But which one?

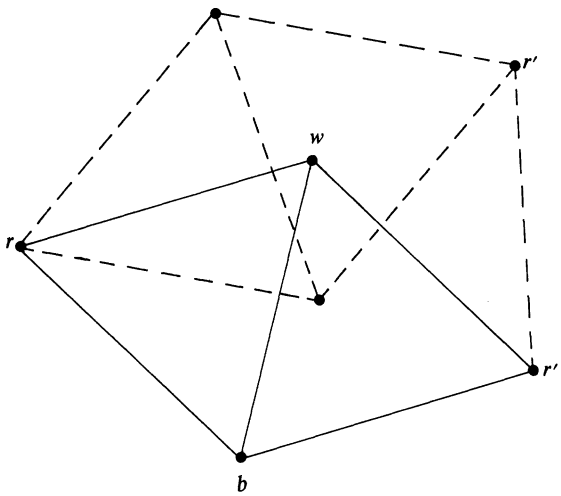


FIGURE 1.

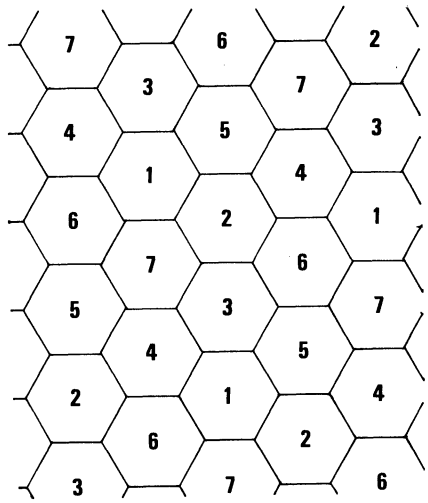


FIGURE 2.

2. A modern way of squaring the circle?

The ancient Greek problem required the construction, for a given circle C, of a square of the same area as C. "Construction" meant that, starting from a line segment whose length is equal to the radius of C, one should use ruler and compass to produce a side of a square whose area is that of C. Starting from a segment of unit length, one of length $\pi^{1/2}$ would be produced. As is now well known, that can't be done with ruler and compass. However, the problem is unsolved (rather than provably unsolvable) for some other interpretations of "construction".

Two subsets X and Y of a Euclidean space are said to be **equivalent by finite decomposition** if X can be partitioned into a finite number of sets $X_1, \ldots, X_n$ and Y into sets $Y_1, \ldots, Y_n$ such that X_i is congruent to Y_i for $1 \leqslant i \leqslant n$. Though now more than a half-century old, the following is sometimes called the "modern form of the problem of squaring the circle":

(B) *Can a circular region and a square region be equivalent by finite decomposition?*

It is known that if a circular and a square region are equivalent then they have the same area, even though the individual sets in the partitions are not required to be "nice" or even to have area in the usual sense. Thus, the situation in E^2 contrasts sharply with that in higher dimensions. In E^d for $d \geqslant 3$, any two sets that are bounded and have interior points are equivalent by finite decomposition. This is the Banach-Tarski "paradox", which is discussed briefly in the second half of the paper. It provides a striking illustration of the fact that when we make a mathematical definition based on the apparent behavior of familiar objects, we may inadvertently create mathematical "monsters" whose behavior defies intuition.

3. Equichordal points

A **chord** of a plane region R is a segment that joins two boundary points of R. A point p of R is an **equichordal point** if all chords through it are of the same length. For example, the center of a circle is an equichordal point. You might think that only a circular region could have an equichordal point, but that's far from the case. Consider a drawing arm as shown in FIGURE 3, with a marker at each end of the arm. The arm pivots at the point p and is free to move back and forth through p. If we start with the ends of the arm at points a and b and move one end along a "reasonable" arc from a to b (one that doesn't get too far from p and that satisfies some other mild conditions), then the two ends together will draw the boundary of a region having p as an equichordal point. In particular, as shown in FIGURE 3, a noncircular convex region may have an equichordal point. (Recall that a region is **convex** if it contains each segment joining two of its points.)

It is known that a plane region cannot have three equichordal points, but the following problem was posed in 1916 and is still unsettled:

(C) *Can a plane convex region have two equichordal points*?

The problem is also open for nonconvex regions, though in that case some fine points in the definition of "region" and "equichordal point" become relevant. Details are given in the second half of this paper.

A point p of a plane region is an equichordal point if and only if the sum of distances $\delta(x,p)+\delta(y,p)$ is constant for all chords $[x,y]$ through p. Analogously, call p an **equireciprocal point** if the sum $\delta(x,p)^{-1}+\delta(y,p)^{-1}$ is constant. The foci of an ellipse are equireciprocal points, but the following problem is open:

(D) *Can a nonelliptical plane convex region have two equireciprocal points*?

4. A tale of two problems

In considering the problems of this paper, it is natural to wonder whether anyone has a reasonable chance of solving them. I can't answer that, except to say that problems of this sort are great equalizers among mathematicians, for solutions usually depend on clever ideas rather than extensive knowledge of theory or development of complicated mathematical machinery.

If age is a reliable guide, the equichordal problem should be the most difficult one considered here, for it is the oldest. However, unsolved problems are by their nature unpredictable, and that may be especially true of ones that are easily stated. By way of illustration, let's consider two problems from the last century that long appeared to be very difficult despite their elementary nature.

The famous four-color problem, which originated in 1852, asks whether every planar map can be painted in four colors so that no two countries with a common boundary arc receive the same color. The problem was first mentioned in print in 1878, and the sufficiency of five colors was soon demonstrated. However, for decades, it was unknown whether any planar map really required five colors. This and Fermat's problem were the two most famous unsolved problems in mathematics. Finally, in 1976, it was proved that four colors suffice. A considerable amount

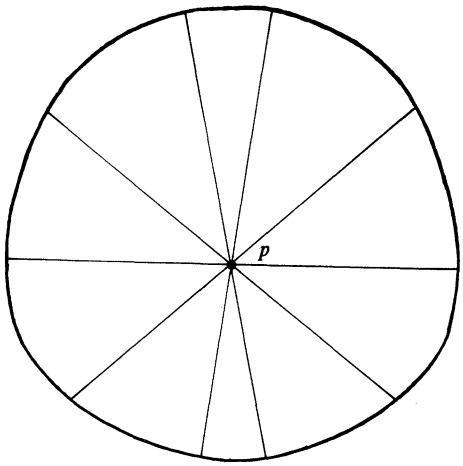


FIGURE 3.

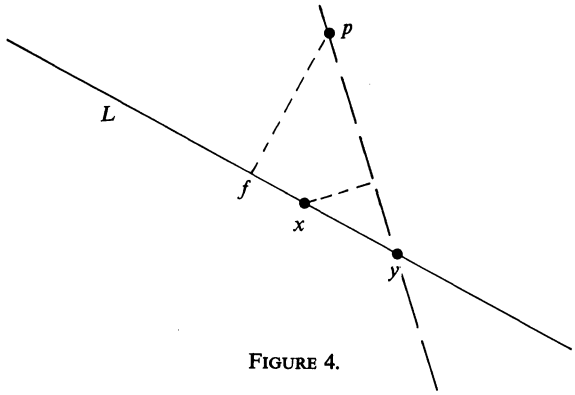


FIGURE 4.

of mathematical machinery was involved, but much of it had been available since 1880. The proof combined refinements of the machinery with many hours of electronic computation. Thus, the four-color problem has been solved, but there is still no easy solution.

The following problem was posed in 1893: If S is a finite set of points in the plane, not all collinear, must there exist an ordinary line for S? (An **ordinary line** is one that contains exactly two points of S.) The problem still appeared difficult forty years later, and was sometimes mentioned along with the four-color problem as an apparently impossible though easily stated problem. Yet now, thanks to a clever idea, it appears to be almost trivial. The solution is given in the next paragraph.

Since the points of S are not all collinear, it is possible to choose a point p of S and a line L that misses p but contains at least two other points of S. Since S is finite, there are only finitely many such choices and hence there is one for which the distance from p to L is minimum. For any such minimizing choice, the line L is ordinary. Otherwise, at least three points of S are on L and at least two of them are on the same side of the foot f of the perpendicular from p to L. If x and y are two such points and x is closer to f, the distance from x to the line py is less than the distance from p to the line $L = xy$. That contradicts the minimizing property of the pair (p, L). (See FIGURE 4)

It will be interesting to see whether any of the open problems from this paper is eventually solved in such an elegant manner. In working on the problems, one may hope to find the clever idea that leads to such a solution. However, there is also the danger that after one has spent months in a futile attack on a problem, *someone else* may find a short and elegant solution!

5. Reflections on reflections

Suppose you live in a 2-dimensional room whose walls form a simple closed polygon, and each wall is a mirror. Is there necessarily a point from which a single light source will illuminate the entire room, using reflected rays as well as direct rays? Must every point of the room have this property? More briefly:

(E) *Is every polygonal plane region illuminable from at least one of its points? From each of its points?*

These problems are unsolved, but there are nonpolygonal regions for which the answers are negative. One is shown in FIGURE 5. The upper elliptical arc has foci at p and p', the lower one at q and q'. The broken lines are the major axes of the ellipses. You'll recall that a light ray issuing from one focus of an ellipse is immediately reflected through the other focus; from this it

follows that any ray crossing the major axis between the foci is immediately reflected between the foci. Thus, in the figure, a ray that issues from any point above ⟨resp. below⟩ the lower axis cannot reach the areas marked B⟨resp. A⟩. Though this smooth, nonpolygonal region R is not illuminable from any point, the problem remains open for polygonal regions, even for those very close to the region of FIGURE 5.

A convex region is of course illuminable from each of its points, even without using reflections. But what happens when attention is restricted to a single ray rather than (as in the illumination problem) all rays issuing from a particular point? Of course, a single ray cannot cover all the points of the region, but it may come arbitrarily close to each point—in topological parlance, it may be *dense*. For example, if the region is the square shown in FIGURE 6, then every

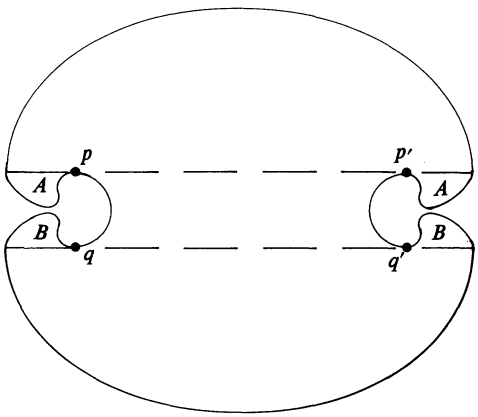


FIGURE 5.

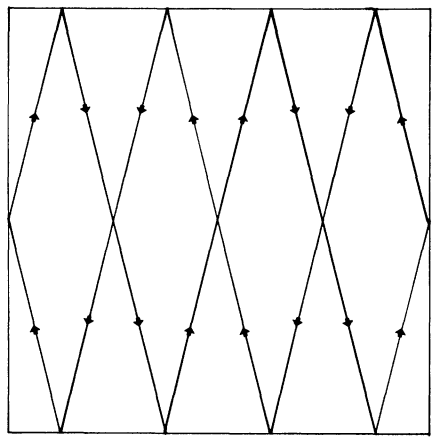


FIGURE 6.

ray with rational slope is closed, like the one shown in the figure, but every ray with irrational slope is dense. There are many unsolved problems concerning the behavior of individual light rays in convex regions, though they are usually stated in terms of billiard ball paths rather than light rays. Perhaps the simplest is the following:

(F) *Does every triangular region admit a dense light ray*?

An affirmative answer is known for triangles (in fact, for closed polygons) in which each angle measure is a rational multiple of π.

6. Forming convex polygons

A set of points in the plane is in **general position** if no three of them are collinear. A much stronger requirement is that the points are the vertices of a convex polygon. For example, the set of 10 points in FIGURE 7 is in general position, but it does not contain the vertex-set of any convex k-gon for $k \geqslant 6$.

(G) *For $n \geqslant 3$, what is the smallest number $f(n)$ such that whenever S is a set of more than $f(n)$ points in general position in the plane, S contains the vertex-set of a convex n-gon*?

It is known that $f(3)=2$, $f(4)=4$ and $f(5)=8$. For larger n, it is not intuitively obvious that any size of S is sufficient to guarantee that S contains the vertex-set of a convex n-gon. However, it has been proved that

$$2^{n-2} \leqslant f(n) \leqslant \binom{2n-4}{n-2} \tag{6.1}$$

and conjectured that $f(n)=2^{n-2}$.

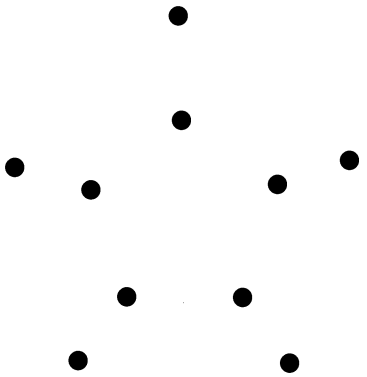


FIGURE 7.

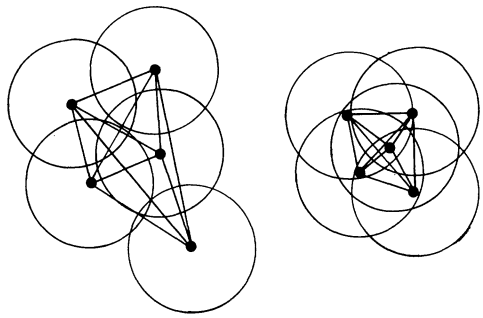


FIGURE 8.

7. Pushing disks around

The next problem deals with congruent circular disks in the plane and with the area that they cover. When two disks are far apart, the area of their union is simply the sum of the individual areas, but as they are pushed closer together, some overlapping occurs and the area of the union decreases. For two disks, the area of the union certainly cannot be increased by pushing the disks closer together. But what happens with three or more disks? Suppose their centers are tautly connected by inelastic string, as shown in FIGURE 8, and we are permitted to move the disks but not to break the string. Thus, we can decrease the distances between the centers but not increase them. Is it possible to increase the total area covered? A negative answer is expected and this is known for three disks. However, the problem is unsettled for any larger number of disks. Here is a formal statement:

(H) *If* $C_1,\dots,C_n$ *and* $D_1,\dots,D_n$ *are congruent circular disks in the plane, centered at points* $p_1,\dots,p_n$ *and* $q_1,\dots,q_n$ *respectively, and if* $\delta(q_i,q_j)\leqslant\delta(p_i,p_j)$ *for all* i *and* j, *does it follow that*

$$\text{area of } \cup_1^n D_i \leqslant \text{ area of } \cup_1^n C_i? \tag{7.1}$$

It has been proved that the area covered by the D_i's is no more than 9 times the area covered by the C_i's, but that's far from solving the problem. To understand the difficulty, note that if α denotes area then

$$\alpha(C_1\cup C_2\cup\cdots\cup C_n)=A_1-A_2+A_3-\cdots+(-1)^{n-1}A_n,$$

where A_1 is the sum of the individual areas and A_k is the sum of the areas of the k-at-a-time intersections. That is,

$$A_k=\Sigma_{1\leqslant j_1<j_2<\cdots<j_k\leqslant n}\alpha(A_{j_1}\cap A_{j_2}\cap\cdots\cap A_{j_k}).$$

In particular, A_n is the area common to all n disks. As the disks are pushed closer together, A_1 is unchanged and A_2 does not decrease, so the two-term approximation A_1-A_2 of $\alpha(C_1\cup\cdots\cup C_n)$ does not increase. However, the problem is complicated by the presence of $A_3,\dots,A_n$.

8. Pushing points around

There are many appealing problems that involve a mixture of plane geometry and number theory. My favorite is presented here.

Let us say that a set is **rational** ⟨resp. **integral**⟩ if only rational numbers ⟨resp. integers⟩ are realized as distances between points of the set. Each line in Euclidean d-space E^d contains a dense rational set, and it is known that each infinite integral set in E^d is actually contained in a

line. But what can be said about rational sets in the plane? In particular:

(I) *Can every finite subset $X=\{x_1,\ldots,x_n\}$ of the plane be closely approximated by a rational set? If so, does the plane actually contain a dense rational set?*

By **closely approximated**, we mean that for each $\varepsilon>0$ there exist points $y_1,\ldots,y_n$ such that the set $Y=\{y_1,\ldots,y_n\}$ is rational and $\delta(x_i,y_i)<\varepsilon$ for all i. In other words, we would like to push the points x_i to nearby positions y_i in order to obtain a rational set.

To gain some insight into the problem, consider the case of three noncollinear points x_1, x_2 and x_3. Of course,

$$\delta(x_i,x_k)<\delta(x_i,x_j)+\delta(x_j,x_k) \tag{8.1}$$

for each permutation (i,j,k) of $(1,2,3)$. Taking $y_1=x_1$, we may move x_2 to a nearby position y_2 so as to preserve the inequalities (8.1) and make $\delta(y_1,y_2)$ rational. Then if ρ_1 and ρ_2 are rational numbers sufficiently close to $\delta(y_1,x_3)$ and $\delta(y_2,x_3)$ respectively, and if C_i is the circle centered at y_i with radius ρ_i, the intersection $C_1\cap C_2$ will consist of two points, one of which is very close to x_3. (See FIGURE 9.) If x_3 is replaced by this nearby point y_3, a close rational approximation $\{y_1,y_2,y_3\}$ of the set $\{x_1,x_2,x_3\}$ is obtained. Thus for $n=3$ the desired rational approximation can be obtained by simply moving the points one at a time. However, this simple approach doesn't work when $n\geqslant 4$ for there are $\binom{n}{2}$ point-pairs whose distances must be controlled.

9. Inscribed squares

The next unsolved problem is my favorite among those that involve a mixture of plane geometry and topology.

We say that a square S is **inscribed** in a set X if X contains all four vertices of S. It doesn't matter how the edges of S are related to X. For example, the square indicated in FIGURE 10 is inscribed in the set C shown there. When can a plane set be expected to admit an inscribed square? Since any set with an interior point admits many such squares, we're really interested only in sets that are "thin" in some sense. The best that can be expected is an affirmative answer to the following question:

(J) *Does every plane simple closed curve C admit an inscribed square?*

Recall that a **simple closed curve** is a set that is topologically the same as a circle. It need not be very "simple" at all from the viewpoint of Euclidean geometry. For example, the set C of FIGURE 10 is a simple closed curve. One naturally tries certain continuity arguments in attempting to answer (J) affirmatively, and they are easily made rigorous when C is the boundary of a convex region. Several authors have claimed incorrectly to extend them to an arbitrary simple closed curve C, but the extension has been made correctly only when C is sufficiently smooth.

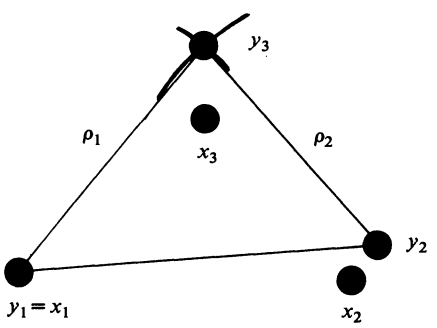


FIGURE 9.

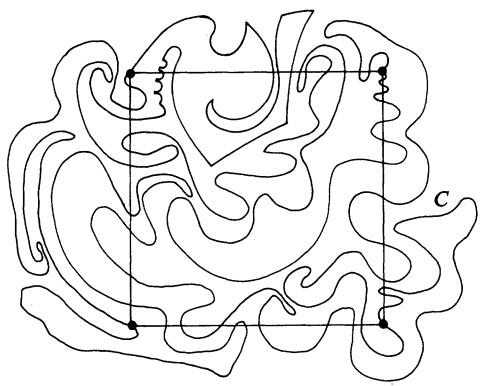


FIGURE 10.

Additional Background

1. A colorful problem

For each positive integer d, let $A_0(d)$ denote the smallest c such that all of Euclidean d-space E^d can be painted with c colors so that no two points at unit distance receive the same color. It follows from an observation of de Bruijn and Erdős [DE] that c is unchanged if "all" is replaced by "every finite subset." A short proof can be based on Tychonov's theorem asserting that the product of compact topological spaces is compact. For let P denote the set of all functions on E^d to $\{1,\ldots,c\}$, topologized as the compact product $\times_{y\in E^d} Q_y$ where each factor Q_y is equal to $\{1,\ldots,c\}$. For each finite $S\subset E^d$, let P_S denote the closed subset of P consisting of all $f\in P$ such that $f(a)\neq f(b)$ whenever a and b are points of S at unit distance. If every finite subset of E^d can be colored in c colors so that no two points at unit distance receive the same color, the family of closed sets $\{P_S$: finite $S\subset E^d\}$ has the finite intersection property and hence by compactness has nonempty intersection. Each point f of the intersection yields a painting of all of E^d in c colors so that no two points at unit distance receive the same color.

The emphasis on unit distance has been mainly for conciseness. Note that $A_0(d)$ is the smallest c such that whenever E^d is covered by fewer than c sets, then for each $\delta>0$, at least one of the sets includes two points at distance δ. A related number is $B_0(d)$, the smallest c such that whenever E^d is covered by fewer than c sets, at least one of them is a **Δ-set**, meaning that it realizes all positive numbers as distances between its point-pairs. The numbers $A_i(d)$ and $B_i(d)$ are similarly defined, except that for $i=1$ they refer to coverings by mutually congruent sets, for $i=2$ by closed sets, and for $i=3$ by sets that are both closed and congruent. Plainly $B_0(d)\leqslant A_0(d)$,

$$A_0(d)\leqslant \begin{matrix}A_1(d)\\A_2(d)\end{matrix} \leqslant A_3(d), \quad \text{and} \quad B_0(d)\leqslant \begin{matrix}B_1(d)\\B_2(d)\end{matrix} \leqslant B_3(d).$$

Problem (A), which was first discussed by Gardner [Ga] in 1960 and Hadwiger [Ha7] in 1961, asks for the determination of $A_0(2)$. It is not hard to see that $A_i(1)$ and $B_i(1)$ are equal to 2 for $i=0,1$ and to 3 for $i=2,3$. The construction based on the hexagonal tessellation of the plane shows that $A_3(2)\leqslant 7$, for it provides a covering of the plane by 7 mutually congruent closed sets such that the entire interval $]4/5,2\sqrt{7}/5[$ is omitted from the distances determined by the point-pairs of any of the sets. In [Ha1, Ha2], Hadwiger showed for all d that $B_2(d)\geqslant d+2$ and $B_3(d)\geqslant 4d$-2, and Raiskii [Ra] and Woodall [Wo] improved the first result to $B_0(d)\geqslant d+2$. For all $d\geqslant 5$, a deep study of "configurations" by Larman and Rogers [LR] and Larman [La'] led to further improvements. In particular, [LR] proved that $B_0(d)\geqslant d(d-1)/6$ and [La'] that $B_0(d)\geqslant (n-2)(n-3)(n-4)/178200$. Both of these papers contain additional results and conjectures of great interest. In particular, [La'] conjectures $B_0(d)\geqslant \frac{1}{3}(\frac{4}{3})^{3d/4}$. In a recent letter, Paul Erdős conjectures the existence of $\delta>0$ such that $A_0(d)>(1+\delta)^d$ for all d, and he reports Peter Frankl's theorem that for some $k>1$, $A_0(d)>d^k$ for all sufficiently large d.

2. A modern way of squaring the circle?

Problem (B), posed by Tarski [Ta'2] in 1925, was motivated by the theorem [Ta'1] that two polygonal plane regions are equivalent by finite decomposition if and only if they are equal in area. In this notion of equivalence, the pieces into which a set is decomposed are individually unrestricted, but they must be pairwise disjoint. For equivalence based on decompositions into polygonal regions that may intersect but not overlap, the theorem is due to J. Bolyai (1802-1860), a founder of noneuclidean geometry. For Bolyai's decomposition theory and higher-dimensional analogues, see Hadwiger [Ha5], Boltyanskii [Bo'1,2] and Meschkowski [Me']. See Dubins, Hirsch and Karush [DHK] and Sallee [Sa'] for results and unsolved problems related to Tarski's problem.

When the pieces are unrestricted, it seems conceivable that two plane regions could be equivalent by finite decomposition even though they are unequal in area. However, this is

excluded by the result of Banach [Ba1] and Morse [Mo′] that the area function of plane geometry (and even 2-dimensional Lebesgue measure) can be extended to a function μ that is:

(a) defined for all bounded subsets of E^2;
(b) finitely additive—$\mu(A \cup B) = \mu(A) + \mu(B)$ whenever A and B are disjoint bounded subsets of E^2;
(c) invariant under rigid motions—$\mu(A) = \mu(B)$ whenever A and B are congruent bounded subsets of E^2.

Though the Banach-Tarski paradox seems incredible, one might imagine that the number of sets in "paradoxical" decompositions is so large that our intuition is unreliable in dealing with such numbers. However, that is not the correct explanation. Robinson [Ro] showed that if X, Y and Z are congruent spherical balls in E^3, then X can be partitioned into five sets $X_1, \ldots, X_5$ in such a way that the pieces X_1 and X_2 can be reassembled (moved by suitable rigid motions) to form Y and the pieces X_3, X_4 and X_5, one of which consists of a single point, can be reassembled to form Z. It is hard to think of a more surprising result! The failure of our intuition does not stem from the number of pieces X_i, but from the complicated nature of some of the pieces. In particular, the volume function of solid geometry cannot be extended to a function μ that satisfies conditions (a), (b) and (c) with E^2 replaced by E^3. If the extension μ is to be defined for all bounded subsets of E^3, finite additivity or invariance under rigid motions must be sacrificed. On the other hand, finite additivity and invariance under translations can be preserved. The crux of the matter is the fact that the group of translations in E^3 (or rigid motions in E^2) is solvable, while the group of rigid motions in E^3 is not. See Sierpinski [Si], Dekker [De] and Meschkowski [Me′] for expositions of the Banach-Tarski paradox and related unsolved problems.

The Banach-Tarski paradox depends heavily on the axiom of choice, but there are other weird decompositions that are quite elementary and constructive in nature and work even in the plane. Some are described by Hadwiger, Debunner, and Klee [HDK]. Others can be derived from Banach's easily proved theorem [Ba2] that for any two sets X and Y and one-to-one mappings f: $X \to Y$ and g: $Y \to X$ there are partitions $X = X_1 \cup X_2$ and $Y = Y_1 \cup Y_2$ for which $fX_1 = Y_1$ and $gY_2 = X_2$. For example, let us say that two subsets U and V of a normed linear space E are **homothetic** if each can be obtained from the other by a dilatation—equivalently, if $U = p + \lambda V$ for some $p \in E$ and $\lambda > 0$. An immediate consequence of Banach's theorem is that if X and Y are subsets of E that are bounded and have nonempty interior, then there are partitions $X = X_1 \cup X_2$ and $Y = Y_1 \cup Y_2$ such that X_i and Y_i are homothetic for $i = 1, 2$. In view of this fact, and of Tarski's result for polygonal regions [Ta′1], it should perhaps not be surprising if there is "a modern way of squaring the circle."

3. Equichordal points

The equichordal problem was raised by Fujiwara [Fu] in 1916 and by Blaschke, Rothe and Weitzenböck [BRW] in 1917. The general strategy of attack on the problem has been to assume the existence of a plane convex region R with two equichordal points p and q, and then to derive many necessary properties of R. A sufficiently long list might show how to construct R, or might include two mutually contradictory properties and hence show R does not exist after all. For example, Wirsing [Wi] showed that R's boundary curve C is analytic, while an earlier author had claimed C could not be six times differentiable. That would have settled the problem but for errors in the earlier "proof." Note that if R does exist we could design a drawing instrument with two arms, pivoting at p and q respectively and hinged together at one end, so that as the hinge traces R's boundary, the other ends of *both* arms do the same. On the basis of experiment, this seems to be impossible. However, a proof is lacking, and Petty and Crotty [PC] have described some noneuclidean plane geometries in which a convex body *can* have two equichordal points.

Let p and q be distinct points of E^2 and let L denote the line through p and q. A construction of Hayashi [Ha^v] and Hallstrom [Ha″] apparently produces a subset R of E^2 such that $L \cap R$ is

a segment and each line through p or q intersects R in a segment of the same length as $L \cap R$. However, it seems the set R so constructed may be neither closed nor open nor convex. The ends of the mentioned segments (aside from $L \cap R$) lie on two curves, but those may oscillate wildly as they approach the line L. In particular, the boundary of R may fail to be a simple closed curve and there may be segments in L that are longer than $L \cap R$ and yet join two boundary points of R.

The deepest and most important paper on the equichordal problem is that of Wirsing [Wi], whose results do not depend on convexity; instead, they require only that R's boundary should be a simple closed curve. See Hadwiger [Ha3] and Klee [Kl1] for expositions of the problem and lists of other references. Some recent inconclusive attacks on the equichordal problem have been made by Hallstrom [Ha″] (who also considered the equireciprocal problem (D)) and McLachlan and Owens [MO].

Problem (D) was initially misstated in [Kl1], then correctly stated by Guy and Klee [GK].

4. A tale of two problems

Some accounts of the four-color conjecture attribute it to cartographers, some to A. F. Möbius. Neither attribution is supported by the careful historical research of May [Ma]. Apparently the conjecture was first formulated by Francis Guthrie, who studied mathematics at University College, London. His brother, Frederick, told the conjecture in 1852 to their teacher, A. DeMorgan, who communicated it to other mathematicians. The first published reference, in 1878, was associated with A. Cayley. The clever but erroneous solutions of Kempe [Ke] and Tait [Ta] appeared in 1879-80, and Kempe's solution was accepted until the error was pointed out by Heawood [He] in 1890. Over the years, several other research mathematicians and scores of mathematical amateurs produced erroneous "proofs" of the four-color conjecture. Errors were so common that whenever a new "solution" appeared, mathematicians would automatically assume there must be a hole in it somewhere. Thus it was especially interesting that the final solution, by Appel and Haken in 1976, was based on the original idea of Kempe (see [AH1,2,3], [AHK], [Ha′]).

Among treatments of the four-color problem that appeared before it was solved, the monographs of Ore [Or] and Heesch [He′] and the article of Coxeter [Co3] are worthy of special mention. A popular account of the solution appears in [AH2], and post-solution monographs have been written by Saaty and Kainen [SK] and Barnette [Ba′].

The sphere is easily seen to be equivalent to the plane for the purposes of map coloring, and the problem may be considered on other surfaces as well. The **chromatic number** of a surface S is the smallest number c such that any map on S can be colored with c or fewer colors. Let c_p denote the chromatic number of the surface obtained by adding p handles to a sphere or, equivalently, drilling p separate holes through a block of wood. The four-color theorem asserts $c_0 = 4$. Heawood [He] showed $c_p \leqslant H(p)$ for all $p > 0$, where $H(p)$ is the greatest integer not exceeding $(7 + \sqrt{1+48p}\,)/2$. He also showed $c_1 = H(1)$, so that the chromatic number of a torus is 7. Over the years, other mathematicians established equality for other values of p, and finally in 1968 Ringel and Youngs [RY1,2] were able to show $c_p = H(p)$ for all $p > 0$. It is interesting that this problem, dealing with all cases except $p = 0$, should have been settled before the four-color problem.

The problem on ordinary lines was posed by Sylvester [Sy] in 1893, rediscovered by Erdős in 1933, and solved by T. Gallai a few days later. The short proof given here is due to L. M. Kelly (reported by Coxeter [Col,2]). A similarly short proof was given by Lang [La] for the dual result, asserting that if a finite family of lines in the plane is such that its members are not all parallel and do not pass through a common point, then there exists a point that lies on exactly two of the lines. See Motzkin [Mo] for a fuller discussion of the problem's history.

The affirmative solution to Sylvester's question says there is at least one ordinary line for any finite plane set whose n points are not all collinear. But should there not be many such lines when n is large? Dirac [Di] conjectured there are at least $[n/2]$ ordinary lines, and Kelly and

Moser [KM] proved there are at least $3n/7$. For other results related to Sylvester's problem and Dirac's conjecture, and for additional references, see Crowe and McKee [CM], Chakerian [Ch], Grünbaum [Gr], Kelly and Rottenberg [KR], Meyer [Me″], and Erdős and Purdy [EP2]. For higher-dimensional extensions, see Motzkin [Mo], Bonnice and Kelly [BK], Edelstein [Ed] and Rottenberg [Ro]. This is only a small sample of the many papers that have been inspired by Sylvester's problem.

5. Reflections on reflections

Problem (E) was stated by Klee [Kl2] but did not originate with him. Perhaps it was first posed by E. Straus in the early 1950's. The nonilluminable region related to the ellipse is P. Ungar's modification of an idea of Penrose and Penrose [PP]. By extending the idea it is possible to construct, for each positive integer k, a plane region R_k with smooth boundary such that R_k is not illuminable from any set of k points. This was noted in [KH], and Rauch [Ra′] obtained related results.

Problems (E) and (F) involve the mathematical "light rays" of geometric optics rather than the light waves that are closer to physical reality. (See Rauch and Taylor [RT] for behavior of the latter.) It is assumed that whenever a ray meets an inner point of an edge of the polygonal boundary, the angle of reflection is equal to the angle of incidence; and when it meets a vertex, it simply ends (or it may be assumed to return along its former path).

The stated result on rays (or billiard ball paths) in squares is due to König and Szücs [KS] (see also Hardy and Wright [HW]), and was first proved by using a theorem of Kronecker on simultaneous Diophantine approximation. A more elementary proof was given by Sudan [Su]. The result on the existence of dense light rays was proved in different ways by Zemlyakov and Katok [ZK] and Boldrighini, Keane and Marchetti [BKM]. For other references, problems and results on billiard ball problems, see these two papers, Poritsky [Po], Croft and Swinnerton-Dyer [CS], Schoenberg [Sc1,2], Halpern [Ha‴] and Kreinovic [Kr].

6. Forming convex polygons

The conjecture that $f(n)=2^{n-2}$ was made by Erdős and Szekeres [ES1] in 1935. They established the right side of (6.1) in [ES1], and the left side 25 years later in [ES2]. (See also Kalbfleisch, Kalbfleisch and Stanton [KKS].) For other problems similar in spirit to (G), see [ES1], a paper by Erdős and Purdy [EP1], and other papers referred to in these. Recently Erdős has asked: For $n \geqslant 3$, what is the smallest number $g(n)$ (if it exists) such that whenever S is a set of more than $g(n)$ points in general position in the plane, S contains the vertex-set of a convex n-gon that has no points of S in its interior? Plainly $g(3)=2$ and $g(4)=4$. The existence of $g(5)$ was proved independently by A. Ehrenfeucht and by H. Harborth [Ha^{iv}2], who showed $g(5)=9$. It is unknown whether $g(6)<\infty$.

A tool of [ES1] is the fact that in any permutation of the integers $1,\ldots,n^2+1$ there is a monotone subsequence of length $n+1$. A short proof of this was given by Seidenberg [Se]. (See also Chvátal and Komlós [CV] and Mirsky [Mi′].)

7. Pushing disks around

Problem (H) is due to Thue Poulsen [TP] and Kneser [Kn], and was discussed also by Hadwiger [Ha4]. It was Kneser who established the relaxation of (H) which has an additional factor 9 on the right side, and a factor 3^d for the analogous problem in E^d. Hadwiger noted that the 1-dimensional analogue of (7.1) (dealing with congruent segments on a line) is valid, and attributed to Habicht and Kneser the fact that (7.1) holds when the centers p_i can be moved continuously to the centers q_i in such a way that, during the motion, the inter-center distances never increase. A proof of the latter result was published by Bollobás [Bo]. To understand the hypothesis of the Habicht-Kneser-Bollobás theorem, imagine that the p_i's are connected by elastic bands; then the assumption is that the p_i's can be moved to the positions q_i in such a way that no band stretches at any time.

Under the hypotheses of (H), does it follow that

$$\text{area of } \cap_1^n D_i \geqslant \text{area of } \cap_1^n C_i?$$

This problem also appears to be open, though it is known at least that if the intersection of the C_i's is nonempty, then so is the intersection of the D_i's. The analogous result in E^d was established by Kirszbraun [Ki], and in recent years it has been extended and applied in remarkable ways. See Danzer, Grübaum, and Klee [DGK] for the history up to 1963, and see Minty [Mi1,2] for more recent developments.

8. Pushing points around

Anning and Erdős [AE] proved (i) every infinite integral set in E^2 is collinear, and (ii) E^2 contains a noncollinear integral set of n points for each $n \geqslant 3$. Erdős [Er] gave a simple proof of (i) for E^d, and Steiger [St] extended (ii) to E^d by replacing 3 with $d+1$ and "noncollinear" with "not contained in any hyperplane."

Using an idea of Müller [Mü], Hadwiger and Debrunner constructed a dense rational set in the unit circle of E^2 (see [HDK]). This implies (ii) and raises the question of which plane sets can be closely approximated by rational sets. Perhaps

(a) there is a rational set that is dense in the plane,

a possibility mentioned by Hadwiger [Ha6] in attributing the problem to Erdős (who writes that he first heard it from Ulam in 1945 or 1946 (see [U1])). Perhaps (a) fails but

(b) every finite plane set can be closely approximated by a rational set,

which would settle a question attributed by Mordell [Mo] to I. J. Schoenberg. Or perhaps, as suggested by J. R. Isbell in [HDK], there is an integer $n \geqslant 5$ such that

(c) in any rational plane set of at least n points, some three points are collinear or some four are concyclic.

That would strongly negate (b) and hence (a). An example of Harborth [Ha^{iv}1] shows that (c) fails when $n=5$, but the case of $n=6$ appears to be unsettled. Problems analogous to those posed by $(a)-(c)$ are open in E^d for all $d>2$, and there are no obvious relationships among the problems in different dimensions.

As was remarked by Mordell, the problem of finding all rational sets of 4 points in the plane goes back to Brahmegupta, who was born in 598 A.D. The problem was in a sense solved by E. E. Kummer in 1848, though it is not obvious from his parametrization that every set of 4 points in E^2 is closely approximated by a rational set. The latter result was proved by Mordell [Mo] in 1959 and sharpened by Almering [Al] in 1963. Since then, several other results on rational sets have been discovered; most of them can be found through a paper by Ang, Daykin and Sheng [ADS] and its references. The difficulty of problem (I), and the fact that little progress has been made on it, is dramatically illustrated by the fact that known results provide no basis for choosing between the truth of (a) on the one hand and, on the other hand, the truth of (c) for $n=6$.

9. Inscribed squares

For all sufficiently smooth simple closed curves in the plane, the existence of an inscribed square was proved by S'nirelman [Sn] and Jerrard [Je]. For the case in which the curve is the boundary of a convex region, elementary proofs were given by Emch [Em1,2], Zindler [Zi] and Christensen [Ch']. This case had been settled earlier by O. Toeplitz, but apparently his proof was never published (see Grünbaum [Gr]). Perhaps the general case of (J) could be settled by the methods of nonstandard analysis, in the spirit of Narens's proof [Na] of the Jordan curve theorem.

It is natural to wonder about higher-dimensional analogues of (J). One aspect was settled by Bielecki [Bi], who described a 3-dimensional convex body in which no rectangular parallelepiped can be inscribed. On the other hand, Pucci [Pu] showed that every 3-dimensional convex body admits an inscribed regular octahedron. The theorems of S'nirelman and Pucci were extended to E^d by Guggenheimer [Gu].

References

[A1] J. H. Almering, Rational quadrilaterals, Indag. Math., 25 (1963) 192–199.

[AE] N. H. Anning and P. Erdős, Integral distances, Bull. Amer. Math. Soc., 51 (1945) 598–600.

[ADS] D. Ang, D. Daykin, and T. Sheng, On Schoenberg's rational polygon problem, Bull. Austral. Math. Soc., 9 (1969) 337–344.

[AH1] K. Appel and W. Haken, Every planar map is four colorable, Bull. Amer. Math. Soc., 82 (1976) 711–712.

[AH2] K. Appel and W. Haken, The solution of the four-color map problem, Sci. Amer., 237, No. 4 (Oct. 1977) 108–121.

[AH3] K. Appel and W. Haken, Every planar map is four colorable, Part I: Discharging, Illinois J. Math., 21 (1977) 429–490.

[AHK] K. Appel, W. Haken, and J. Koch, Every planar map is four colorable, Part II: Reducibility, Illinois J. Math., 21 (1977) 491–567.

[Ba1] S. Banach, Sur le problème de mesure, Fund. Math., 4 (1923) 7–33.

[Ba2] S. Banach, Un théorème sur les transformations biunivoques, Fund. Math., 6 (1924) 236–239.

[BT] S. Banach and A. Tarski, Sur la décomposition des ensembles de points en parties respectivement congruentes, Fund. Math., 6 (1924) 244–277.

[Ba'] D. Barnette, A monograph on the four-color problem, to be published by the Mathematical Association of America.

[Be] F. R. Bernhart, A digest of the four-color theorem. J. Graph Theory, 1 (1977) 207–225.

[Bi] A. Bielecki, Quelques remarques sur la note précédente, Ann. Univ. Mariae Curie-Sklodowska, Sec. A., 8 (1954) 101–103.

[BRW] W. Blaschke, H. Rothe, and R. Weitzenböck, Aufgabe 552, Arch. Math. Phys., 27 (1917) 82.

[BKM] C. Boldrighini, M. Keane, and F. Marchetti, Billiards in polygons, to appear.

[Bo] B. Bollobás, Area of the union of disks, Elem. Math., 23 (1968) 60–61.

[Bo'1] V. Boltyanskiĭ, Equivalent and Equidecomposable Figures, Heath, Boston, 1963. (Translated by A. Henn and C. Watts from the first Russian edition, Moscow, 1956.)

[Bo'2] V. Boltyanskiĭ, The Third Problem of Hilbert (Russian), Nauka Press, Moscow, 1977.

[BK] W. Bonnice and L. Kelly, On the number of ordinary planes, J. Combinatorial Theory Ser. A, 11 (1971) 45–53.

[Ch] G. Chakerian, Sylvester's problem on collinear points and a relative, Amer. Math. Monthly, 77 (1970) 164–167.

[Ch'] C. Christensen, A square inscribed in a convex figure (Danish), Mat. Tidsskr, B 1950 (1950) 22–26.

[CV] V. Chvátal and J. Komlós, Some combinatorial theorems on monotonicity. Canad. Math. Bull., 14 (1971) 151–157.

[Co1] H. Coxeter, A problem of collinear points, Amer. Math. Monthly, 55 (1948) 26–28.

[Co2] H. Coxeter, Introduction to Geometry, Wiley, New York, 1961.

[Co3] H. Coxeter, The mathematics of map coloring, J. Recreational Math., 2 (1969) 3–12.

[CS] H. T. Croft and H. P. F. Swinnerton-Dyer, On the Steinhaus billiard table problem, Proc. Cambridge Philos. Soc., 59 (1963) 37–41.

[CM] D. Crowe and T. McKee, Sylvester's problem on collinear points, this MAGAZINE, 41 (1968) 30–34.

[DGK] L. Danzer, B. Grünbaum, and V. Klee, Helly's theorem and its relatives, Convexity (V. Klee, ed.), Amer. Math. Soc. Proc. Symp. Pure Math., 7 (1963) 101–180.

[DE] N. DeBruijn and P. Erdős, On a combinatorial problem, Indag. Math., 10 (1948) 421–423.

[De] T. Dekker, Paradoxical decompositions of sets and spaces, Ph.D. thesis, University of Amsterdam, 1958.

[Di] G. Dirac, Collinearity properties of sets of points, Quart. J. Math Oxford Ser. 2, 2 (1951) 221–227.

[DHK] L. Dubins, M. Hirsch, and J. Karush, Scissor congruence, Israel J. Math., 1 (1963) 239–247.

[Ed] M. Edelstein, Generalizations of the Sylvester problem, this MAGAZINE, 43 (1970) 250–254.

[Em1] A. Emch, Some properties of closed convex curves in a plane, Amer. J. Math., 35 (1913) 407–412.

[Em2] A. Emch, On the medians of a closed convex polygon, Amer. J. Math., 37 (1915) 19–28.

[Er] P. Erdős, Integral distances, Bull. Amer. Math. Soc., 51 (1945) 996.

[EP1] P. Erdős and G. Purdy, Some extremal problems in geometry, J. Combinatorial Theory Ser. A, 10 (1971) 246–252.

[EP2] P. Erdős and G. Purdy, Some combinatorial problems in the plane, J. Combinatorial Theory Ser. A, 25 (1978) 205–210.

[ES1] P. Erdős and G. Szekeres, A combinatorial problem in geometry, Compositio Math., 2 (1935) 463–470.

[ES2] P. Erdős and G. Szekeres, On some extremum problems in elementary geometry, Ann. Univ. Sci. Budapest Eötvös Sect. Math., 3 (1960/61) 53–62.

[Fu] M. Fujiwara, Über die Mittelkurve zweier geschlossenen konvexen Kurven in bezug auf einen Punkt, Tôhoku Math. J., 10 (1916) 99–103.

[Ga] M. Gardner, A new collection of brain-teasers, Sci. Amer., 206 (October 1960) 180.

[Gr] B. Grünbaum, Arrangements, Spreads, and Hyperplanes, Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics, No. 10, Amer. Math. Soc., Providence, 1972.
[Gu] H. Guggenheimer, Finite sets on curves and surfaces, Israel J. Math., 3 (1965) 104–112.
[GK] R. Guy and V. Klee, Monthly research problems, 1969–1971, Amer. Math. Monthly, 78 (1971) 1113–1122.
[Ha1] H. Hadwiger, Ein Überdeckungssatz für den euklidischen Raum, Portugal. Math., 4 (1944) 140–144.
[Ha2] H. Hadwiger, Überdeckung des euklidischen Raumes durch kongruentre Mengen, Portugal. Math., 4 (1945) 238–242.
[Ha3] H. Hadwiger, Ungelöste Probleme Nr. 3, Elem. Math., 10 (1955) 19.
[Ha4] H. Hadwiger, Ungelöste Probleme Nr. 11, Elem. Math., 11 (1956) 60–61.
[Ha5] H. Hadwiger, Vorlesungen über Inhalt, Oberfläche und Isoperimetrie, Springer, Berlin, 1957.
[Ha6] H. Hadwiger, Ungelöste Probleme Nr. 24, Elem. Math., 23 (1958) 85.
[Ha7] H. Hadwiger, Ungelöste Probleme Nr. 40, Elem. Math., 16 (1961) 103–104.
[HDK] H. Hadwiger, H. Debrunner, and V. Klee, Combinatorial Geometry in the Plane, Holt, Rinehart, and Winston, New York, 1964.
[Ha′] W. Haken, An attempt to understand the four color problem, J. Graph Theory, 1 (1977) 193–206.
[Ha″] A Hallstrom, Equichordal and equireciprocal points, Bogazici Univ. J. Sci., 2 (1974) 83–88.
[Ha‴] B. Halpern, Strange billiard tables, Trans Amer. Math. Soc., 232 (1977) 297–305.
[Ha^{iv}1] H. Harborth, On the problem of P. Erdős concerning points with integral distances, Ann. New York Acad. Sci., 175 (1970) 206–207.
[Ha^{iv}2] H. Harborth, Konvexe Fünfecke in ebenen Punktmengen, Elem. Math., 33 (1978) 116–118.
[HW] G. Hardy and W. Wright, An Introduction to the Theory of Numbers (4th Edition), Oxford Univ. Press, 1960.
[Ha^{v}] T. Hayashi, On pseudo-central oval (Japanese), Tohoku Math. J., 27 (1926) 197–202.
[He] P. Heawood, Map-colour theorem, Quart. J. Pure Appl. Math., 24 (1890) 332–338.
[He′] H. Heesch, Untersuchungen zum Vierfarbenproblem, Bibiliographisches Institut, Mannheim, 1969.
[Je] R. Jerrard, Inscribed squares in plane curves, Trans. Amer. Math. Soc., 98 (1961) 234–241.
[KKS] J. D. Kalbfleisch, J. G. Kalbfleisch and R. G. Stanton, A combinatorial problem on convex regions, Proc. Louisiana Conf. Combinatorics, Graph Theory and Computing (R. C. Mullin, K. B. Reid and D. P. Roselle, eds.) (1970) 180–188.
[KM] L. Kelly and W. Moser, On the number of ordinary lines determined by n points, Canad. J. Math, 10 (1958) 210–219.
[KR] L. Kelly and R. Rottenberg, Simple points in pseudoline arrangements, Pacific J. Math., 40 (1972) 617–622.
[Ke] A. B. Kempe, On the geographical problem of four colors, Amer. J. Math., 2 (1879) 193–204.
[Ki] M. Kirszbraun, Über die zusammenziehenden und Lipschitzschen Transformationen, Fund. Math., 22 (1934) 77–108.
[Kl1] V. Klee, Can a plane convex body have two equichordal points? Amer. Math. Monthly, 76 (1969) 54–55.
[Kl2] V. Klee, Is every polygonal plane region illuminable from some point?, Amer. Math. Monthly, 76 (1969) 180.
[KH] V. Klee and E. Halsey, Shapes of the Future—Some Unsolved Problems in Geometry, Part I: Two Dimensions, viewer's manual for film of the same title produced by the Individual Lectures Film Project of the Mathematical Association of America, 1971.
[Kn] M. Kneser, Einige Bemerkungen über das Minkowskische Flächenmass, Arch. Math., 6 (1955) 382–390.
[KS] D. König and A. Szücs, Mouvement d'un point abandonné à l'intérieur d'un cube, Rend. Cird. Palermo, 36 (1913) 79–90.
[Kr] V. Ja. Kreinovic, Note on billiards, Notices Amer. Math. Soc., 26 (1979) A–15.
[La] D. W. Lang, No. 2577, The dual of a well-known theorem. Math. Gaz., 39 (1955) 314.
[La′] D. Larman, A note on the realization of distances within sets in Euclidean space, Comment. Math. Helv., 53 (1978) 529–535.
[LR] D. Larman and C. Rogers, The realization of distances within sets in Euclidean space, Mathematika, 19 (1972) 1–24.
[Ma] K. May, The origin of the four-color conjecture, Isis, 56 (1965) 346–348.
[Mo] E. K. McLachlan and G. K. Owens, Evidence for the existence of a convex curve with two equichordal points, Unpublished typescript, Oklahoma State Univ., 1977.
[Me′] H. Meschkowski, Unsolved and Unsolvable Problems in Geometry, Oliver and Boyd, Edinburgh and London, Frederick Ungar, New York: (Translated by J. A. C. Burlak from the first German edition, Vieweg, Braunschweig, 1960).
[Me″] W. Meyer, On ordinary points in arrangements, Israel J. Math., 17 (1974) 124–135.

[Mi1] G. J. Minty, On the extension of Lipschitz, Lipschitz-Hölder continuous, and monotone functions, Bull. Amer. Math. Soc., 76 (1970) 334–339.
[Mi2] G. J. Minty, A finite-dimensional tool-theorem in monotone operator theory, Advances in Math., 12 (1974) 1–7.
[Mi'] L. Mirsky, A dual of Dilworth's decomposition theorem, Amer. Math. Monthly, 78 (1971) 876–877.
[Mo] L. J. Mordell, Rational quadrilaterals, J. London Math. Soc., 35 (1960) 277–282.
[Mo'] A. Morse, Squares are normal, Fund. Math., 36 (1949) 35–39.
[Mo''] T. Motzkin, The lines and planes connecting the points of a finite set, Trans. Amer. Math. Soc., 70 (1951) 451–464.
[Mu] A. Müller, Auf einem Kreis liegende Punktmengen ganzzahliger Entfernungen, Elem. Math., 8 (1953) 37–38.
[Na] L. Narens, A nonstandard proof of the Jordan curve theorem, Pacific J. Math., 36 (1971) 219–229.
[Or] O. Ore, The Four Color Problem, Academic Press, New York, 1967.
[PP] L. Penrose and R. Penrose, Puzzles for Christmas, New Scientist, (25 December 1958) 1580–1581, 1597.
[PC] C. Petty and J. Crotty, Characterizations of spherical neighbourhoods, Canad. J. Math., 22 (1970) 431–435.
[Po] H. Poritsky, The billiard ball problem on a table with a convex boundary—an illustrative dynamical problem, Ann. of Math., 51 (1950) 446–470.
[Pu] C. Pucci, Sulla inscrivibilita di un ottaedro regolare in un insieme convesso limitato dello spazio ordinario, Atti. Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Natur, (8) 21 (1956) 61–65.
[Ra] D. Raiskii, The realization of all distances in a decomposition of the space R^n into $n+1$ parts, Math. Notes 7 (1970) 194–196 (Translated from Mat. Zametki, 7 (1970) 319–323.)
[Ra'] J. Rauch, Illumination of bonded domains, Amer. Math Monthly, 85 (1978) 359–361.
[RT] J. Rauch and M. Taylor, Penetration into shadow regions and unique continuation properties in hyperbolic mixed problems, Indiana Univ. Math. J., 22 (1972) 277–285.
[RY1] G. Ringel and J. Youngs, Solution of the Heawood map-coloring problem, Proc. Nat. Acad. Sci., 60 (1968) 438–445.
[RY2] G. Ringel and J. Youngs, Lösung des Problems der Nachbargebiete, Arch. Math., 20 (1969) 190–201.
[Ro] R. Robinson, On the decomposition of spheres, Fund. Math., 34 (1947) 246–260.
[Ro'] R. Rottenberg, On finite sets of points in P^3, Israel J. Math, 10 (1970) 160–171.
[SK] T. L. Saaty and P. C. Kainen, The Four-Color Problem: Assault and Conquest, McGraw-Hill, New York, 1977.
[Sa'] G. Sallee, Are equidecomposable plane convex sets convex equidecomposable? Amer. Math. Monthly, 76 (1969) 926–927.
[Sc1] I. Schoenberg, Extremum problems for the motions of a billiard ball, I, The L_p norm, $1 \leqslant p < \infty$, Indag. Math., 38 (1976) 66–75.
[Sc2] I. Schoenberg, Extremum problems for the motions of a billiard ball, II, The L_∞ norm, Indag. Math., 38 (1976) 263–279.
[Se1] A. Seidenberg, A simple proof of a theorem of Erdős and Szekeres, J. London Math. Soc., 34 (1959) 352.
[Si] W. Sierpinski, On the Congruence of Sets and Their Equivalence by Finite Decomposition, Lucknow University Studies, vol. 20, 1954.
[Sn] L. S'nirelman, On certain geometrical properties of closed curves (Russian), Uspehi Mat. Nauk., 10 (1944) 34–44.
[St] F. Steiger, Zu einer Frage über Mengen von Punkten mit ganzzahliger Entfernung, Elem. Math., 8 (1953) 66–67.
[Su] G. Sudan, Sur le problème du rayon réfléchi, Rev. Roumaine Math. Pures Appl., 10 (1965) 723–733.
[Sy] J. Sylvester, Mathematical Question 11951, Educational Times, 59 (1893) 98.
[Ta] P. G. Tait, On the colouring of maps, Proc. Roy. Soc. Edinburgh, 10 (1880) 501–503, 729.
[Ta'1] A. Tarski, On the equivalence of polygons (Polish), Przeglad Matematyczno-Fizyczny, 2 (1924).
[Ta'2] A. Tarski, Problem No. 38, Fund. Math., 7 (1925) 381.
[TP] E. Thue Poulsen, Problem 10, Math. Scand., 2 (1954) 346.
[U1] S. Ulam, A Collection of Mathematical Problems, Interscience, New York, 1960.
[Wi] E. Wirsing, Zur Analytizität von Doppelspeichenkurven, Arch. Math., 9 (1958) 300–307.
[Wo] D. R. Woodall, Distances realized by sets covering the plane, J. Combinatorial Theory Ser. A, 14 (1973) 187–200.
[ZK] A. Zemlyakov and A. Katok, Topological transitivity of billards in polygons, Math. Notes, 18 (1975) 760–764. (Translated from Mat. Zametki, 18 (1975) 291–300).
[Zi] K. Zindler, Über konvexe Gebilde, Monatsh. Math., 31 (1921) 25–57.

The History of Stokes' Theorem

Let us give credit where credit is due: Theorems of Green, Gauss and Stokes appeared unheralded in earlier work.

VICTOR J. KATZ
University of the District of Columbia
Washington, D.C. 20005

Most current American textbooks in advanced calculus devote several sections to the theorems of Green, Gauss, and Stokes. Unfortunately, the theorems referred to were not original to these men. It is the purpose of this paper to present a detailed history of these results from their origins to their generalization and unification into what is today called the generalized Stokes' theorem.

Origins of the theorems

The three theorems in question each relate a k-dimensional integral to a $k-1$-dimensional integral; since the proof of each depends on the fundamental theorem of calculus, it is clear that their origins can be traced back to the late 17th century. Toward the end of the 18th century, both Lagrange and Laplace actually used the fundamental theorem and iteration to reduce k-dimensional integrals to those of one dimension less. However, the theorems as we know them today did not appear explicitly until the 19th century.

The first of these theorems to be stated and proved in essentially its present form was the one known today as Gauss' theorem or the divergence theorem. In three special cases it occurs in an 1813 paper of Gauss **[8]**. Gauss considers a surface (superficies) in space bounding a solid body (corpus). He denotes by PQ the exterior normal vector to the surface at a point P in an infinitesimal element of surface ds and by QX, QY, QZ the angles this vector makes with the positive x-axis, y-axis, and z-axis respectively. Gauss then denotes by $d\Sigma$ an infinitesimal element of the y-z plane and erects a cylinder above it, this cylinder intersecting the surface in an even number of infinitesimal surface elements $ds_1, ds_2, \ldots, ds_{2n}$. For each j, $d\Sigma = \pm ds_j \cos QX_j$ where the positive sign is used when the angle is acute, the negative when the angle is obtuse. Since if the cylinder enters the surface where QX is obtuse, it will exit where QX is acute (see FIGURE 1), Gauss obtains $d\Sigma = -ds_1 \cos QX_1 = ds_2 \cos QX_2 = \ldots$ and concludes by summation that "The integral $\int ds \cos QX$ extended to the entire surface of the body is 0."

He notes further that if T, U, V are rational functions of only y,z, only x,z, and only x,y respectively, then "$\int (T\cos QX + U\cos QY + V\cos QZ)\,ds = 0$." Gauss then approximates the volume of the body by taking cylinders of length x and cross sectional area $d\Sigma$ and concludes in a similar way his next theorem: "The entire volume of the body is expressed by the integral $\int ds\, x(\cos QX)$ extended to the entire surface." We will see below how these results are special cases of the divergence theorem.

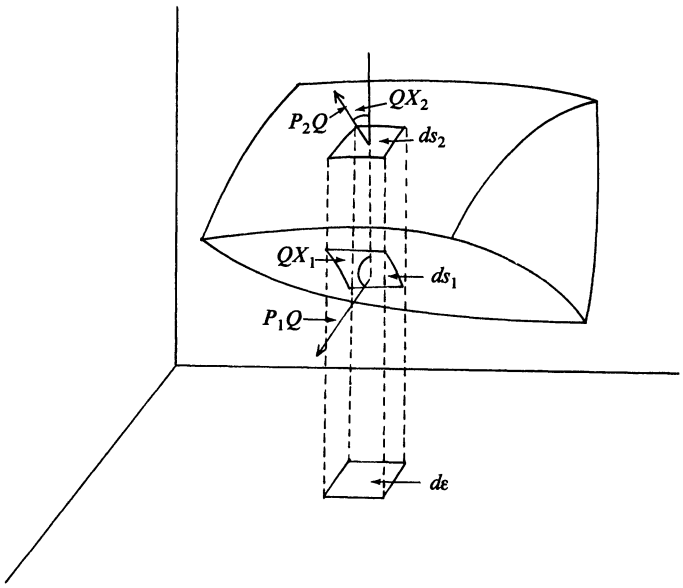


FIGURE 1

In 1833 and 1839 Gauss published other special cases of this theorem, but by that time the general theorem had already been stated and proved by Michael Ostrogradsky. This Russian mathematician, who was in Paris in the late 1820's, presented a paper **[15]** to the Paris Academy of Sciences on February 13, 1826, entitled "Proof of a theorem in Integral Calculus." In this paper Ostrogradsky introduces a surface with element of surface area ε bounding a solid with element of volume ω. He denotes by α,β,γ the same angles which Gauss called QX, QY, QZ, and by p,q,r three differentiable functions of x,y,z. He states the divergence theorem in the form:

$$\int\left(a\frac{\partial p}{\partial x}+b\frac{\partial q}{\partial y}+c\frac{\partial r}{\partial z}\right)\omega=\int(ap\cos\alpha+bq\cos\beta+cr\cos\gamma)\varepsilon$$

where a,b,c are constants and where the left hand integral is taken over a solid, the right hand integral over the boundary surface.

We note that Gauss' results are all special cases of Ostrogradsky's theorem. In each case $a=b=c=1$; Gauss' first result has $p=1,q=r=0$; his second has

$$\frac{\partial p}{\partial x}=\frac{\partial q}{\partial y}=\frac{\partial r}{\partial z}=0;$$

and his third has $p=x,q=r=0$. We also will see that Gauss' proof is a special case of that of Ostrogradsky.

Ostrogradsky proves his result by first considering $\frac{\partial p}{\partial x}\omega$. He integrates this over a "narrow cylinder" going through the solid in the x-direction with cross-sectional area $\bar{\omega}$, using the fundamental theorem of calculus to express this integral as

$$\int\frac{\partial p}{\partial x}\omega=\int(p_1-p_0)\bar{\omega},$$

where p_0 and p_1 are the values of p on the pieces of surface where the cylinder intersects the solid. Since $\bar{\omega}=\varepsilon_1\cos\alpha_1$ on one section of surface and $\bar{\omega}=-\varepsilon_0\cos\alpha_0$ on the other (α_1 and α_0 being the appropriate angles made by the normal, ε_1 and ε_0 being the respective surface elements) we get

$$\int \frac{\partial p}{\partial x}\omega = \int p_1\varepsilon_1 \cos\alpha_1 + \int p_0\varepsilon_0 \cos\alpha_0 = \int p \cos\alpha\varepsilon$$

where the left integral is over the cylinder and the right ones over the two pieces of surface (FIGURE 2). Adding up the integrals over all such cylinders gives one third of the final result, the other two thirds being done similarly. We note that this proof can easily be modified to suit modern standards, and is in fact used today, e.g., in Taylor and Mann **[24]**.

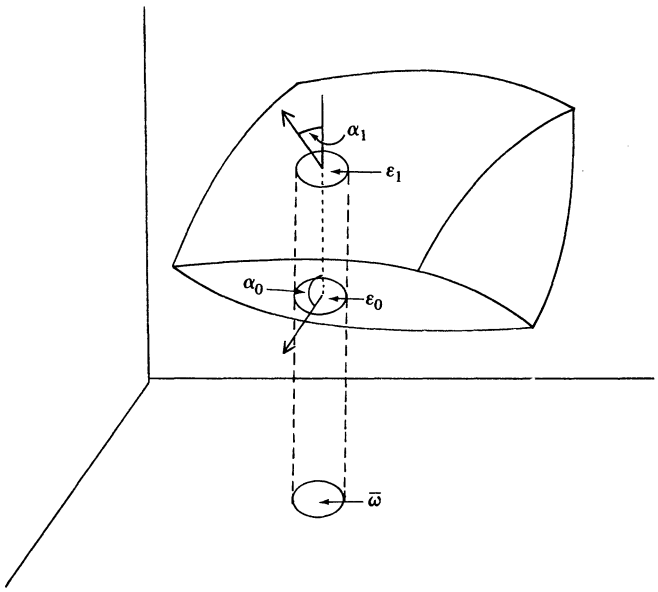


FIGURE 2

Though the above proof applies to arbitrary differentiable functions p,q,r, we will note for future reference that Ostrogradsky uses the result only in the special case where

$$p = v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x},\quad q = v\frac{\partial u}{\partial y} - u\frac{\partial v}{\partial y},\quad r = v\frac{\partial u}{\partial z} - u\frac{\partial v}{\partial z},$$

with u and v also being differentiable functions of three variables.

Ostrogradsky presented this theorem again in a paper in Paris on August 6, 1827, and finally in St. Petersburg on November 5, 1828. The latter presentation was the only one published by Ostrogradsky, appearing in 1831 in **[16]**. The two earlier presentations have survived only in manuscript form, though they have been published in Russian translation.

In the meantime, the theorem and related ones appeared in publications of three other mathematicians. Simeon Denis Poisson, in a paper presented in Paris on April 14,1828, (published in 1829) stated and proved an identical result **[19]**. According to Yushkevich in **[28]**, Poisson had refereed Ostrogradsky's 1827 paper and therefore presumably learned of the result. Poisson neither claimed it as original nor cited Ostrogradsky, but it must be realized that references were not made then with the frequency that they are today.

Another French mathematician, Frederic Sarrus, published a similar result in 1828 in **[21]**, but his notation and ideas are not nearly so clear as those of Ostrogradsky and Poisson. Finally, George Green, an English mathematician, in a private publication of the same year **[9]**, stated and proved the following:

$$\int u\Delta v\,dx\,dy\,dz + \int u\frac{dv}{dw}\,d\sigma = \int v\Delta u\,dx\,dy\,dz + \int v\frac{du}{dw}\,d\sigma$$

where u, v are functions of three variables in a solid body "of any form whatever," Δ is the symbol for the Laplacian, and d/dw means the normal derivative; the first integrals on each side are taken over the solid and the second over the boundary surface. Green proved his theorem using the same basic ideas as did Ostrogradsky. In addition, if we use again the special case where

$$p = v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x},\quad q = v\frac{\partial u}{\partial y} - u\frac{\partial v}{\partial y},\quad r = v\frac{\partial u}{\partial z} - u\frac{\partial v}{\partial z},$$

we can conclude by a short calculation that the two theorems are equivalent. Nevertheless, Green did not so conclude; he was interested in the theorem in the form in which he gave it. It would thus be difficult to attribute the divergence theorem to him.

All of the mathematicians who stated and proved versions of this theorem were interested in it for specific physical reasons. Gauss was interested in the theory of magnetic attraction, Ostrogradsky in the theory of heat, Green in electricity and magnetism, Poisson in elastic bodies, and Sarrus in floating bodies. In nearly all cases, the theorems involved occurred in the middle of long papers and were only thought of as tools toward some physical end. In fact, for both Green and Ostrogradsky the functions u and v mentioned above were often solutions of Laplace-type equations and were used in boundary value problems.

The theorem generally known as Green's theorem is a two-dimensional result which was also not considered by Green. Of course, one can derive this theorem from Green's version by reducing it to two dimensions and making a brief calculation. But there is no evidence that Green himself ever did this.

On the other hand, since Green's theorem is crucial in the elementary theory of complex variables, it is not surprising that it first occurs, without proof, in an 1846 note of Augustin Cauchy **[5]**, in which he proceeds to use it to prove "Cauchy's theorem" on the integral of a complex function around a closed curve. Cauchy presents the result in the form:

$$\int\left(p\frac{dx}{ds} + q\frac{dy}{ds}\right)ds = \pm\int\int\left(\frac{\partial p}{\partial y} - \frac{\partial q}{\partial x}\right)dx\,dy$$

where p and q are functions of x and y, and where the sign of the second integral depends on the orientation of the curve which bounds the region over which the integral is taken. Cauchy promised a proof in his private journal *Exercises d'analyse et de physique mathématique*, but he apparently never published one.

Five years later, Bernhard Riemann presented the same theorem in his inaugural dissertation **[20]**, this time with proof and in several related versions; again he uses the theorem in connection with the theory of complex variables. Riemann's proof is quite similar to the proof commonly in use today; essentially he uses the fundamental theorem to integrate $\partial q/\partial x$ along lines parallel to the x-axis, getting values of q where the lines cross the boundary of the region; then he integrates with respect to y to get

$$\int\left[\int\frac{\partial q}{\partial x}\,dx\right]dy = -\int q\,dy = -\int q\frac{dy}{ds}\,ds.$$

The other half of the formula is proved similarly.

The final theorem of our triad, Stokes' theorem, first appeared in print in 1854. George Stokes had for several years been setting the Smith's Prize Exam at Cambridge, and in the February, 1854, examination, question #8 is the following **[22]** (see FIGURE 3):

> If X, Y, Z be functions of the rectangular coordinates x, y, z, dS an element of any limited surface, l, m, n the cosines of the inclinations of the normal at dS to the axes, ds an element of the boundary line, shew that
>
> $$\int\int\left\{l\left(\frac{\partial Z}{\partial y} - \frac{\partial Y}{\partial z}\right) + m\left(\frac{\partial X}{\partial z} - \frac{\partial Z}{\partial x}\right) + n\left(\frac{\partial Y}{\partial x} - \frac{\partial X}{\partial y}\right)\right\}dS = \int\left(X\frac{dx}{ds} + Y\frac{dy}{ds} + Z\frac{dz}{ds}\right)ds$$
>
> ... the single integral being taken all around the perimeter of the surface.

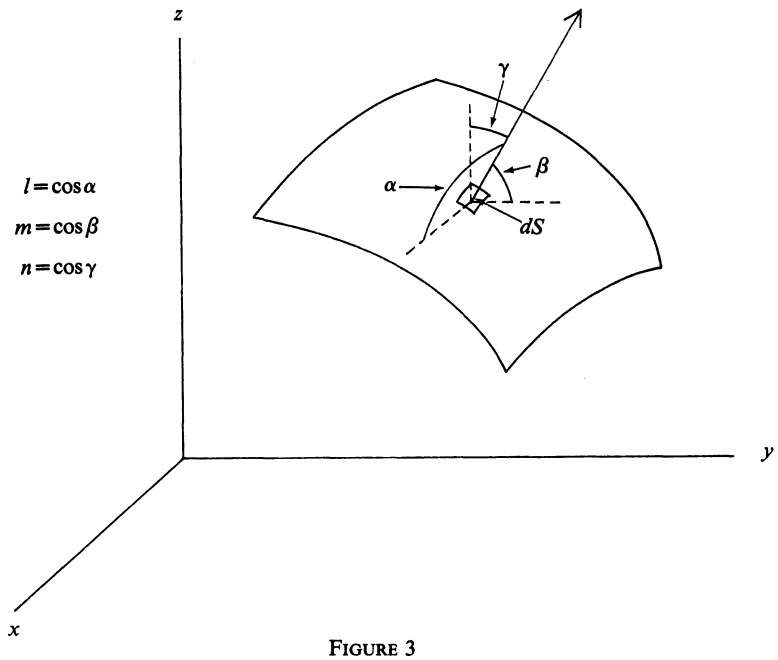


FIGURE 3

It does not seem to be known if any of the students proved the theorem. However, the theorem had already appeared in a letter of William Thomson (Lord Kelvin) to Stokes on July 2, 1850, and the left hand expression of the theorem had appeared in two earlier works of Stokes. The first published proof of the theorem seems to have been in a monograph of Hermann Hankel in 1861 **[10]**. Hankel gives no credit for the theorem, only a reference to Riemann with regard to Green's theorem, which theorem he calls well-known and makes use of in his own proof of Stokes' result.

In his proof Hankel considers the integral $\int X\,dx + Y\,dy + Z\,dz$ over a curve bounding a surface given explicitly by $z = z(x,y)$. Then

$$dz = \frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy;$$

so the given integral becomes

$$\int \left(X + \frac{\partial z}{\partial x} Z\right) dx + \left(Y + \frac{\partial z}{\partial y} Z\right) dy.$$

By Green's theorem, this integral in turn becomes

$$\int\int \left\{ \frac{\partial\left(X + \frac{\partial z}{\partial x} Z\right)}{\partial y} - \frac{\partial\left(Y + \frac{\partial z}{\partial y} Z\right)}{\partial x} \right\} dx\,dy.$$

An explicit evaluation of the derivatives then leads to the result:

$$\int (X\,dx + Y\,dy + Z\,dz) = \int\int \left\{ \left(\frac{\partial X}{\partial y} - \frac{\partial Y}{\partial x}\right) + \left(\frac{\partial Z}{\partial y} - \frac{\partial Y}{\partial z}\right)\frac{\partial z}{\partial x} + \left(\frac{\partial X}{\partial z} - \frac{\partial Z}{\partial x}\right)\frac{\partial z}{\partial y} \right\} dx\,dy.$$

Since a normal vector to the surface is given by $(-\partial z/\partial x, -\partial z/\partial y, 1)$ and since the components of the unit normal vector are the cosines of the angles which that vector makes with the coordinate axes, it follows that $\partial z/\partial x = -1/n$, $\partial z/\partial y = -m/n$, and $dS = dx\,dy/n$. Hence by substitution, Hankel obtains the desired result.

Of course, this proof requires the surface to be given explicitly as $z=z(x,y)$. A somewhat different proof, without that requirement, is sketched in Thomson and Tait's *Treatise on Natural Philosophy* (1867) without reference **[25]**. In 1871 Clerk Maxwell wrote to Stokes asking about the history of the theorem **[12]**. Evidently Stokes answered him, since in Maxwell's 1873 *Treatise on Electricity and Magnetism* there appears the theorem with the reference to the Smith's Prize Exam **[13]**. Maxwell also states and proves the divergence theorem.

Vector forms of the theorems

All three theorems first appeared, as we have seen, in their coordinate forms. But since the theory of quaternions was being developed in the mid-nineteenth century by Hamilton and later by Tait, it was to be expected that the theorems would be translated into their quaternion forms. First we must note that Hamilton's product of two quaternions

$$\mathbf{p}=x_0+x_1\mathbf{i}+x_2\mathbf{j}+x_3\mathbf{k} \text{ and } \mathbf{q}=y_0+y_1\mathbf{i}+y_2\mathbf{j}+y_3\mathbf{k}$$

may be written as

$$\mathbf{pq}=(x_0y_0-x_1y_1-x_2y_2-x_3y_3)+(x_2y_3-y_2x_3)\mathbf{i}+(x_3y_1-y_3x_1)\mathbf{j}+(x_1y_2-x_2y_1)\mathbf{k}.$$

The scalar part is denoted $S\cdot\mathbf{pq}$ and the vector part $V\cdot\mathbf{pq}$. Secondly, applying Hamilton's ∇-operator $\mathbf{i}\partial/\partial x+\mathbf{j}\partial/\partial y+\mathbf{k}\partial/\partial z$ to a vector function $\boldsymbol{\sigma}=\mathbf{i}X+\mathbf{j}Y+\mathbf{k}Z$ we get a quaternion

$$\nabla\boldsymbol{\sigma}=-\left(\frac{\partial X}{\partial x}+\frac{\partial Y}{\partial y}+\frac{\partial Z}{\partial z}\right)+\mathbf{i}\left(\frac{\partial Z}{\partial y}-\frac{\partial Y}{\partial z}\right)+\mathbf{j}\left(\frac{\partial X}{\partial z}-\frac{\partial Z}{\partial x}\right)+\mathbf{k}\left(\frac{\partial Y}{\partial x}-\frac{\partial X}{\partial y}\right).$$

Again we denote the scalar part by $S\nabla\boldsymbol{\sigma}$ and the vector part by $V\nabla\boldsymbol{\sigma}$.

Tait, then, in an 1870 paper **[23]** was able to state the divergence theorem in the form

$$\int\int\int S\cdot\nabla\boldsymbol{\sigma}\,d\zeta=\int\int S\cdot\boldsymbol{\sigma}U\nu\,ds$$

where $d\zeta$ is an element of volume, ds an element of surface, and $U\nu$ a unit normal vector to the surface. Furthermore, Stokes' theorem took the form

$$\int S\cdot\boldsymbol{\sigma}\,d\rho=\int\int S\cdot V\nabla\boldsymbol{\sigma}\,U\nu\,d\zeta$$

where $d\rho$ is an element of length of the curve bounding the surface.

Maxwell, in his treatise of three years later, repeated Tait's formulas, but also came one step closer to our current terminology. He proposed to call $S\nabla\boldsymbol{\sigma}$ the *convergence* of $\boldsymbol{\sigma}$ and $V\nabla\boldsymbol{\sigma}$ the *curl* of $\boldsymbol{\sigma}$. Of course, Maxwell's convergence is the negative of what we call the divergence. Furthermore, we note that when two quaternions $\mathbf{p}$ and $\mathbf{q}$ are pure vectors, Hamilton's $S\cdot\mathbf{pq}$ is precisely the negative of the inner product $\mathbf{p}\cdot\mathbf{q}$. (This particular idea was first developed by Gibbs and Heaviside about twenty years later.)

Putting these notions together, we get the modern vector form of the divergence theorem

$$\iiint_M(\operatorname{div}\boldsymbol{\sigma})\,dV=\iint_S\boldsymbol{\sigma}\cdot\mathbf{n}\,dA$$

where $\boldsymbol{\sigma}$ is a vector field $X\mathbf{i}+Y\mathbf{j}+Z\mathbf{k}$, dV is an element of volume, dA is an element of surface area of the surface S bounding the solid M and $\mathbf{n}$ is the unit outward normal to this surface. Stokes' theorem then takes the form

$$\int\int_S(\operatorname{curl}\boldsymbol{\sigma})\cdot\mathbf{n}\,dA=\int_\Gamma(\boldsymbol{\sigma}\cdot\mathbf{t})\,ds$$

where ds is the element of length of the boundary curve Γ of the surface S and $\mathbf{t}$ is the unit tangent vector to Γ. Obviously, a similar result can be given for Green's theorem.

Generalization and unification

The generalization and unification of our three theorems took place in several stages. First of all, Ostrogradsky himself in an 1836 paper in Crelle **[17]** generalized his own theorem to the following:

$$\int_V \left(\frac{\partial P}{\partial x}+\frac{\partial Q}{\partial y}+\frac{\partial R}{\partial z}+\cdots\right)dx\,dy\,dz\ldots=\int_S \frac{\left(P\frac{\partial L}{\partial x}+Q\frac{\partial L}{\partial y}+R\frac{\partial L}{\partial z}+\cdots\right)}{\sqrt{\left(\frac{\partial L}{\partial x}\right)^2+\left(\frac{\partial L}{\partial y}\right)^2+\left(\frac{\partial L}{\partial z}\right)^2+\cdots}}\,dS.$$

Here Ostrogradsky lets $L(x,y,z,\ldots)$ be "a function of as many quantities as one wants," V be the set of values $x,y,z,\ldots$ with $L(x,y,z\ldots)>0$ and S be the set of values with $L(x,y,z,\ldots)=0$. In modern terminology, if there are n-values, S would be an $n-1$-dimensional hypersurface bounding the n-dimensional volume V.

Ostrogradsky's proof here is similar to his first one. He integrates $\partial P/\partial x$ with respect to x and after a short manipulation gets

$$\int \frac{\partial P}{\partial x}\,dx\,dy\,dz\ldots=\int \frac{P\frac{\partial L}{\partial x}}{\sqrt{\left(\frac{\partial L}{\partial x}\right)^2}}\,dy\,dz\ldots$$

with, of course a similar expression for every other term. Then, putting $dS=\sqrt{dy^2dz^2\ldots+dx^2dz^2\ldots+dx^2dy^2\ldots+\cdots}$, he shows that

$$\frac{dy\,dz\ldots}{\sqrt{\left(\frac{\partial L}{\partial x}\right)^2}}=\frac{dx\,dz\ldots}{\sqrt{\left(\frac{\partial L}{\partial y}\right)^2}}=\cdots=\frac{dS}{\sqrt{\left(\frac{\partial L}{\partial x}\right)^2+\left(\frac{\partial L}{\partial y}\right)^2+\left(\frac{\partial L}{\partial z}\right)^2+\cdots}}$$

and concludes the result by summation.

(To understand how Ostrogradsky gets his expression for dS, we note that for a parametrized surface in three-space,

$$dS=\Delta\,du\,dv=\sqrt{\left(\frac{\partial(y,z)}{\partial(u,v)}\right)^2+\left(\frac{\partial(z,x)}{\partial(u,v)}\right)^2+\left(\frac{\partial(x,y)}{\partial(u,v)}\right)^2}\,du\,dv$$

and

$$\frac{\partial(y,z)}{\partial(u,v)}\,du\,dv=dy\,dz,\ \frac{\partial(z,x)}{\partial(u,v)}\,du\,dv=dz\,dx,\ \frac{\partial(x,y)}{\partial(u,v)}\,du\,dv=dx\,dy.$$

Hence

$$dS=\sqrt{dy^2dz^2+dz^2dx^2+dx^2dy^2}\ .$$

For more details, see **[1]**.)

If we further note that

$$\left(\sqrt{\left(\frac{\partial L}{\partial x}\right)^2+\left(\frac{\partial L}{\partial y}\right)^2+\left(\frac{\partial L}{\partial z}\right)^2+\cdots}\ \right)^{-1}\left(\frac{\partial L}{\partial x},\frac{\partial L}{\partial y},\frac{\partial L}{\partial z},\ldots\right)$$

is the unit outward normal $\mathbf{n}$ to S and if we let $\boldsymbol{\sigma}$ denote the vector function $(P,Q,R\ldots)$, then Ostrogradsky's result becomes

$$\int(\operatorname{div}\boldsymbol{\sigma})\,dV=\int\boldsymbol{\sigma}\cdot\mathbf{n}\,dS,$$

a direct generalization of the original divergence theorem.

The first mathematician to include all three theorems under one general result was Vito Volterra in 1889 in [27]. Before quoting the theorem, we need to understand his terminology. An r-dimensional hyperspace in n-dimensional space is given parametrically by the n functions $x_i = x_i(u_1, u_2, u_3, \dots u_r)$, $i = 1, \dots n$. Volterra considers the n by r matrix $J = (\partial x_i / \partial u_j)$ and denotes by $\Delta_{i_1, i_2, \dots i_r}$ the determinant of the r by r submatrix of J consisting of the rows numbered $i_1, i_2, \dots i_r$. Letting

$$\Delta = \left(\sum_{i_1 < \cdots < i_r} \Delta^2_{i_1, i_2, \dots, i_r} \right)^{\frac{1}{2}},$$

he calls $\Delta\, du_1\, du_2 \dots du_r$ an "element of hyperspace" and $\alpha_{i_1 \dots i_r} = (\Delta_{i_1 \dots i_r} / \Delta)$ a direction cosine of the hyperspace. (The α's are, of course, functions of the u's.) For the case where $r = 2$ and $n = 3$ we have already calculated Δ above in our discussion of Ostrogradsky's theorem. Since the determinants $\partial(x_i, x_j)/\partial(u_1, u_2)$ are precisely the components of the normal vector to the surface, the α_{ij} are then the components of the unit normal vector, hence are the cosines of the angles which that vector makes with the appropriate coordinate axes.

We now quote Volterra's theorem, translated from the Italian:

> Let $L_{i_1 i_2 \dots i_r}$ be functions of points in a hyperspace S_n defined and continuous in all their first derivatives and such that any transposition of indices changes only the sign. Let the forms
>
> $$M_{i_1 i_2 \dots i_{r+1}} = \sum_{s=1}^{r+1} (-1)^{s-1} \frac{\partial L_{i_1 i_2 \dots i_{s-1} i_{s+1} \dots i_{r+1}}}{\partial x_{i_s}}.$$
>
> We denote by S_r the boundary of a hyperspace S_{r+1} of $r+1$ dimensions open and immersed in S_n; by $\alpha_{i_1 i_2 \dots i_{r+1}}$ the direction cosines of S_{r+1} and by $\beta_{i_1 i_2 \dots i_r}$ those of S_r. The extension of the theorem of Stokes consists of the following formula:
>
> $$\int_{S_{r+1}} \sum_i M_{i_1 i_2 \dots i_{r+1}} \alpha_{i_1 i_2 \dots i_{r+1}} dS_{r+1} = \int_{S_r} \sum_i L_{i_1 i_2 \dots i_r} \beta_{i_1 i_2 \dots i_r} dS_r.$$

Let us check the case where $r = 1$ and $n = 3$ to see how this result generalizes Stokes' theorem. In that case we have three functions L_1, L_2, L_3 of points in three-dimensional space. The M functions are then given as follows:

$$M_{12} = \frac{\partial L_2}{\partial x_1} - \frac{\partial L_1}{\partial x_2}, \quad M_{31} = \frac{\partial L_1}{\partial x_3} - \frac{\partial L_3}{\partial x_1}, \quad M_{23} = \frac{\partial L_3}{\partial x_2} - \frac{\partial L_2}{\partial x_3}.$$

Since $r = 1$, S_r is a curve given by 3 functions $x_1(u)$, $x_2(u)$, and $x_3(u)$. So

$$\Delta = \sqrt{\left(\frac{dx_1}{du}\right)^2 + \left(\frac{dx_2}{du}\right)^2 + \left(\frac{dx_3}{du}\right)^2}$$

and $ds = \Delta\, du$. Then

$$\beta_i = \frac{\dfrac{dx_i}{du}}{\Delta} \qquad \text{for } i = 1, 2, 3 \qquad \text{and } \beta_i\, ds = \frac{dx_i}{du} ds.$$

The α_{12}, α_{31}, and α_{23} are the appropriate cosines as mentioned above. Hence the theorem will read:

$$\int_{S_2} \left(\frac{\partial L_3}{\partial x_2} - \frac{\partial L_2}{\partial x_3} \right) \alpha_{23} + \left(\frac{\partial L_1}{\partial x_3} - \frac{\partial L_3}{\partial x_1} \right) \alpha_{31} + \left(\frac{\partial L_2}{\partial x_1} - \frac{\partial L_1}{\partial x_2} \right) \alpha_{12}\, dS_2$$

$$= \int_{S_1} \left(L_1 \frac{dx_1}{du} + L_2 \frac{dx_2}{du} + L_3 \frac{dx_3}{du} \right) du,$$

the result being exactly Stokes' theorem. A similar calculation will show that the case $r=2, n=3$ will give the divergence theorem, the case $r=1, n=2$ will give Green's theorem, and the case $r=n-1$ is precisely Ostrogradsky's own generalization.

We note further that if we replace α_{ij} by $(\partial(x_i,x_j)/\partial(u,v))/\Delta$, dS_2 by $\Delta\,du\,dv$, and $(\partial(x_i,x_j)/\partial(u,v))\,du\,dv$ by $dx_i\,dx_j$, and if we set $x_1=x$, $x_2=y$, $x_3=z$, we get another familiar form of Stokes' theorem:

$$\int_{S_2} M_{23}\,dy\,dz + M_{31}\,dz\,dx + M_{12}\,dx\,dy = \int_{S_1} L_1\,dx + L_2\,dy + L_3\,dz.$$

Similarly, the divergence theorem becomes

$$\int_{S_3} L_{23}\,dy\,dz + L_{31}\,dz\,dx + L_{12}\,dx\,dy = \int_{S_2}\left(\frac{\partial L_{23}}{\partial x} + \frac{\partial L_{31}}{\partial y} + \frac{\partial L_{12}}{\partial z}\right) dx\,dy\,dz.$$

Although Volterra used his theorem in several papers in his study of differential equations, he did not give a proof of the result; he only said that it "is obtained without difficulty."

If one studies Volterra's work, it becomes clear that it would be quite useful to simplify the notation. This was done by several mathematicians around the turn of the century. Henri Poincaré, in his 1899 work *Les Méthodes Nouvelles de la Mécanique Céleste* **[18]**, states the same generalization as Volterra, but in a much briefer form:

$$\int \sum A\,d\omega = \int \sum \sum_k \pm \frac{dA}{dx_k}\,dx_k\,d\omega.$$

Here the left-hand integral is taken over the r-1-dimensional boundary of an r-dimensional variety in n-space while the right-hand integral is over the entire variety. Hence A is a function of n variables and $d\omega$ is a product of $r-1$ of the dx_i's, the sum being taken over all such distinct products. Poincaré's form of the theorem is more compact than that of Volterra in part because the direction cosines are absorbed into the expressions $d\omega$. (See **[1]** for more details.) Poincaré, like Volterra, in this and other works of the same period, was chiefly interested in integrability conditions of what we now call differential forms; i.e., in when a form ω is an exact differential.

The mathematician chiefly responsible for clarifying the idea of a differential form was Elie Cartan. In his fundamental paper of 1899 **[2]**, he first defines an "expression différentielle" as a symbolic expression given by a finite number of sums and products of the n differentials $dx_1, dx_2, \ldots, dx_n$ and certain coefficient functions of the variables $x_1, x_2, \ldots, x_n$. A differential expression of the first degree, $A_1\,dx_1 + A_2\,dx_2 + \cdots + A_n\,dx_n$, he calls an "expression de Pfaff."

Cartan states certain rules of calculation with these expressions. In particular, his rule for "evaluating" a differential expression requires that the value of $A\,dx_{m_1}\,dx_{m_2}\ldots dx_{m_n}$ be the product of A with the determinant $|\partial x_{m_j}/\partial\alpha_i|$ where the x's are functions of the parameters α_i. The standard rules for determinants then require that if any dx_i is repeated, the value is 0 and that any permutation of the dx_i's requires a sign change if the permutation is odd. For instance, Cartan concludes that $A\,dx_1\,dx_2\,dx_3 = -A\,dx_2\,dx_1\,dx_3$, or just that $dx_1\,dx_2 = -dx_2\,dx_1$.

Cartan further discusses changes of variable; if $x_1, x_2, \ldots, x_n$ are functions of $y_1, y_2, \ldots, y_n$, then

$$dx_i = \frac{\partial x_i}{\partial y_1}\,dy_1 + \frac{\partial x_i}{\partial y_2}\,dy_2 + \cdots + \frac{\partial x_i}{\partial y_n}\,dy_n, \qquad i=1,2,\ldots,n.$$

Then, for instance, in the case $n=2$, we get

$$dx_1\,dx_2 = \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\,dy_1\,dy_2.$$

One might note, on the other hand, that if one assumes a change of variable formula of this type, then one is forced to the general rule $dx_i\,dx_j = -dx_j\,dx_i$.

Finally, Cartan defines the "derived expression" of a first degree differential expression $\omega = A_1 dx_1 + A_2 dx_2 + \cdots + A_n dx_n$ to be the second degree expression $\omega' = dA_1 dx_1 + dA_2 dx_2 + \cdots + dA_n dx_n$, where, of course,

$$dA_i = \Sigma_j \frac{\partial A_i}{\partial x_j} dx_j.$$

For the case $n=3$ one can calculate by using the above rules that if $\omega = A_1 dx + A_2 dy + A_3 dz$, then

$$\omega' = \left(\frac{\partial A_3}{\partial y} - \frac{\partial A_2}{\partial z}\right) dy\, dz + \left(\frac{\partial A_1}{\partial z} - \frac{\partial A_3}{\partial x}\right) dz\, dx + \left(\frac{\partial A_2}{\partial x} - \frac{\partial A_1}{\partial y}\right) dx\, dy.$$

Comparing this with the example we gave in discussing Volterra's work, it is clear that Volterra's M_{23}, M_{31}, and M_{12} are precisely the coefficients of Cartan's ω'.

Cartan in **[2]** did not discuss the relationship of his differential expressions to Stokes' theorem; nevertheless, by the early years of the twentieth century the generalized Stokes' theorem in essentially the form given by Poincaré was known and used by many authors, although proofs seem not to have been published.

By 1922, Cartan had extended his work on differential expressions in **[3]**. It is here that he first uses the current terminology of "exterior differential form" and "exterior derivative." He works out specifically the derivative of a 1-form (as we did above) and notes that for $n=3$ Stokes' theorem states that $\int_C \omega = \int\int_S \omega'$ where C is the boundary curve of the surface S. (This is, of course, exactly Volterra's result in the same special case.) Then, defining the exterior derivative of any differential form $\omega = \Sigma A\, dx_i dx_j \ldots dx_l$ to be $\omega' = \Sigma dA\, dx_i dx_j \ldots dx_l$ (with dA as above), he works out the derivative of a 2-form Ω in the special case $n=3$ and shows that for a parallelepiped P with boundary S, $\int\int_S \Omega = \int\int\int_P \Omega'$. One can easily calculate that this is the divergence theorem, and we must assume that Cartan realized its truth in more general cases. He was, however, not yet ready to state the most general result.

The "d" notation for exterior derivative was used in 1902 by Theodore DeDonder in **[6]**, but not again until Erich Kähler reintroduced it in his 1934 book *Einführung in die Theorie der Systeme von Differentialgleichungen* **[11]**. His notation is slightly different from ours, but in a form closer to ours it was adopted by Cartan for a course he gave in Paris in 1936–37 (published as *Les Systèmes Différentiels Extérieurs et leurs Applications Géométriques* **[4]** in 1945). Here, after discussing the definitions of the differential form ω and its derivative $d\omega$, Cartan notes that all of our three theorems (which he attributes to Ostrogradsky, Cauchy-Green, and Stokes, respectively) are special cases of $\int_C \omega = \int_A d\omega$ where C is the boundary of A. To be more specific, Green's theorem is the special case where ω is a 1-form in 2-space; Stokes' theorem is the special case where ω is a 1-form in 3-space; and the divergence theorem is the special case where ω is a 2-form in 3-space. Finally, Cartan states that for any $p+1$-dimensional domain A with p-dimensional boundary C one could demonstrate the general Stokes' formula:

$$\int_C \omega = \int_A d\omega$$

(For examples of the use of these theorems, see any advanced calculus text, e.g., **[1]** or **[24]**. For more information on differential forms, one can consult **[7]**.)

Appearance in texts

A final interesting point about these theorems is their appearance in textbooks. By the 1890's all three theorems were appearing in the analysis texts of many different authors. The third of

our theorems was always attributed to Stokes. The French and Russian authors tended to attribute the first theorem to Ostrogradsky, while others generally attributed it to Green or Gauss; this is still the case today. Similarly, Riemann is generally credited with the second theorem by the French, while Green is named by most others. Before Cartan's 1945 book, about the only author to attribute that result to Cauchy was H. Vogt in **[26]**.

The generalized Stokes' theorem, first published, as we have seen, in 1945, has only been appearing in textbooks in the past twenty years, the first occurrence probably being in the 1959 volume of Nickerson, Spencer, and Steenrod **[14]**.

References

[1] R. C. Buck, Advanced Calculus, 2nd ed., McGraw-Hill, New York, 1965.

[2] E. Cartan, Sur certaines expressions différentielles et sur le problème de Pfaff, Annales École Normale, v. 16, 1899, pp. 239–332; Oeuvres, Part II, Vol. I, pp. 303–397.

[3] E. Cartan, Leçons sur les invariants intégraux, Chap. VII, Hermann, Paris, 1922.

[4] E. Cartan, Les Systèmes Différentiels Extérieurs et leurs Applications Géométriques, Hermann, Paris, 1945.

[5] A Cauchy, Sur les intégrales qui s'étendent à tous les points d'une courbe fermée, Comptes Rendus, v. 23, 1846, pp. 251–55; Oeuvres, Ire Série, Tome X, pp. 70–74.

[6] T. DeDonder, Etude sur les invariants intégraux, Rendiconti del Circolo Matematico di Palermo, v. 16, 1902, pp. 155–179.

[7] H. Flanders, Differential Forms, Academic Press, New York, 1963.

[8] C. F. Gauss, Theoria Attractionis Corporum Sphaeroidicorum Ellipticorum Homogeneorum Methodo nova tractata, Commentationes societatis regiae scientiarium Gottingensis recentiores, v. II, 1813; Werke, v. 5, pp. 1–22.

[9] G. Green, An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism, London, 1828; Green's Mathematical Papers, pp. 3–115.

[10] H. Hankel, Zur allgemeinen Theorie der Bewegung der Flüssigkeiten, Dieterische Univ. Buchdruckerei, Göttingen, 1861.

[11] E. Kähler, Einführung in die Theorie der Systeme von Differentialgleichungen, Leipzig, 1934.

[12] C. Maxwell, Letter to Stokes in G. Stokes, Memoir and Scientific Correspondence, ed. by J. Larmor, Cambridge, 1907, v. II, p. 31.

[13] C. Maxwell, A Treatise on Electricity and Magnetism, Oxford, 1873, v. I.

[14] H. Nickerson, D. Spencer, N. Steenrod, Advanced Calculus, Van Nostrand, Princeton, 1959.

[15] M. Ostrogradsky, Démonstration d'un théorème du calcul intégral, pub. in Russian in Istoriko-Matematicheskie Issledovania, v. XVI, 1965, pp. 49–96.

[16] M. Ostrogradsky, Note sur la Théorie de la Chaleur, Mémoires de L'Acad. Imp. des Sciences de St. Petersburg, ser. 6, v. 1, 1831, pp. 129–133.

[17] M. Ostrogradsky, Sur le calcul des variations des intégrales multiples, Journal fur die Reine und angewandte Mathematik, v. XV, 1836, pp. 332–354.

[18] H. Poincaré, Les Méthodes Nouvelles de la Mécaniques Céleste, v. III, p. 10, Gauthier-Villars, Paris, 1899, Dover, N.Y. 1957.

[19] S. Poisson, Mémoire sur l'Équilibre et le Mouvement des Corps Élastiques, Mémoires de l'Académie Royale des Sciences de l'Institut de France, v. VIII, 1829, pp. 357–571.

[20] B Riemann, Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse, Gottingen, 1851; Werke, pp. 3–48.

[21] F. Sarrus, Mémoire sur les oscillations des corps flottans, Annales de mathématiques pures et appliquées (Nismes), v. XIX, 1828, pp. 185-211.

[22] G. Stokes, Mathematical and Physical Papers, v. 5, p. 320, Cambridge Univ. Press, Cambridge, 1905.

[23] P. Tait, On Green's and other allied Theorems, Transactions of the Royal Society of Edinburgh, 1869–70, pp. 69–84.

[24] A. Taylor, W. Mann, Advanced Calculus, 2nd. ed., Xerox, Lexington, 1972.

[25] W. Thomson, P. Tait, Treatise on Natural Philosophy, Vol. 1, Oxford, 1867.

[26] H. Vogt, Éléments de Mathématiques Supérieures, Paris, 1925.

[27] V. Volterra, Delle Variabili Complesse Negli Iperspazi, Rendiconti della R. Accad. der Lincei, ser. IV, v. V, 1889, pp. 158–165; Opere, v. I, pp. 403–411.

[28] A. Yushkevich, Istoriya Matematiki v Rossii do 1917 goda, Moscow, 1968, p. 290.

NOTES

The Mean Value Theorem for Vector Valued Functions: A Simple Proof

WILLIAM S. HALL
University of Pittsburgh
Pittsburgh PA 15260

MARTIN L. NEWELL
University College
Galway, Ireland

It is well known that the mean value theorem in one dimension extends readily to real-valued functions of several variables, but fails for the vector-valued case. For example, let $f(t)=(\cos t,\sin t)$ and suppose there is a point ξ in $(0,2\pi)$ such that $f'(\xi)=0$. Then $-\sin\xi=\cos\xi=0$, an impossible situation. A useful and correct generalization is the inequality

$$|f(y)-f(x)| \leqslant \sup_{0<t<1} \|f'(x+t(y-x))\|\,|y-x|$$

where $f: D \subset R_n \to R_m$ is a differentiable vector-valued function on a convex open set D, f' is the matrix $\delta f_i/\delta x_j$, $i=1,2,\ldots,m$, $j=1,2,\ldots,n$, $|\ |$ is the appropriate norm (in R_n, or in R_m), $\|\ \|$ is the usual norm in the set of linear maps from R_n to R_m, and x, y are arbitrary points in the domain D.

Many undergraduate calculus and analysis texts prove the mean value theorem in the real case but omit the result above. Those that do present this more general form usually give either a "sloppy" proof, using components, or a "slick" proof with the Hahn-Banach Theorem. Here we present a direct approach, requiring only the chain rule and the mean value theorem in R. It is worth noting that f' at each point is a linear map (given by the Jacobian matrix) and that the usual norm for a linear map (matrix) is given by $\sup_{|x|=1} |Ax|$. However, other norms such as $(\Sigma a_{ij}^2)^{1/2}$ where $A=(a_{ij})$ are frequently used in advanced calculus courses. All we really use is that $|Ax| \leqslant \|A\|\ |x|$.

The result is certainly true if $f(y)=f(x)$. If not, form the function $\phi(t)$ by

$$\phi(t)=<f(y)-f(x),f(x+t(y-x))>/|f(y)-f(x)|$$

where $<,>$ denotes the inner product in R_m. Then

$$\phi(1)=<f(y)-f(x),f(y)>/|f(y)-f(x)|,$$
$$\phi(0)=<f(y)-f(x),f(x)>/|f(y)-f(x)|,$$
$$\phi'(t)=<f(y)-f(x),f'(x+t(y-x))(y-x)>/|f(y)-f(x)|.$$

In the last line we used the chain rule twice, once because f is differentiable, and once because the inner product is also differentiable. Of course, ϕ itself is well-defined because D is convex.

By the usual mean value theorem,

$$\phi(1)-\phi(0)=\phi'(t)=<f(y)-f(x),f(y)-f(x)>/|f(y)-f(x)|=|f(y)-f(x)|,$$

and by the Schwarz inequality,

$$\begin{aligned}\phi'(t) &\leqslant |f(y)-f(x)|\ |f'(x+t(y-x))(y-x)|/|f(y)-f(x)| \\ &\leqslant \|f'(x+t(y-x))\|\ |y-x| \leqslant \sup_{0<t<1} \|f'(x+t(y-x))\|\ |y-x|.\end{aligned}$$

This proves the theorem.

It is clear that if R_n is replaced by any Banach space and R_m is replaced by any real Hilbert space, then the method of the proof remains valid.

An Unusual Example of a Sphere

J. MICHAEL MCGREW
Ball State University
Muncie, IN 47306

The history of mathematics is dotted with false proofs and counterexamples which aren't really counterexamples. Occasionally, as in the early attempts to prove the four color conjecture, the effort that goes into these false starts is far from wasted, producing new tools or ideas which, if not useful for solving the intended problem, are useful in other areas of mathematics. Occasionally, too, a false counterexample can have instructional value. This paper discusses one such counterexample. For a broader analysis of the way false counterexamples can be useful in the evolution of mathematical ideas, we refer the reader to [3].

As a vital part of his characterization of planar simple closed curves, A. M. Schoenflies introduced "accessibility" [6, p. 126]. A point m in the boundary of a region G is **accessible** from G if, for every point $g \in G$, there is a path joining m and g and contained entirely within G (with the exception of the point m). The point m is said to be **accessible from all sides** with respect to G if, for each arc α in G with endpoints in the boundary of G, and for each component G' of $G-\alpha$ which contains m in its boundary, m is accessible from G'.

The following example illustrates the fact that a point may be accessible from a region without being accessible from all sides with respect to that region. Let T be the closed curve formed by the union of the following sets in cartesian coordinates (see FIGURE 1(a)):

$$\left\{(x,y): y=\sin\frac{\pi}{x}, 0<x\leqslant 1\right\},$$
$$\{(1,y): -2\leqslant y\leqslant 0\},$$
$$\{(0,y): -2\leqslant y\leqslant 1\},$$
$$\{(x,-2): 0\leqslant x\leqslant 1\}.$$

(Some irreverent authors call this a "Warsaw circle.") Let m be the point $(0,-1)$. Then m is accessible from both of the regions of which T is the boundary, but if G' is the subset of the exterior region G indicated in FIGURE 1(b), then m is contained in the boundary of G' but is not accessible from G'. If p is the point $(0,-2)$, however, then p is accessible from all sides with respect to both the unbounded region G and the bounded region I.

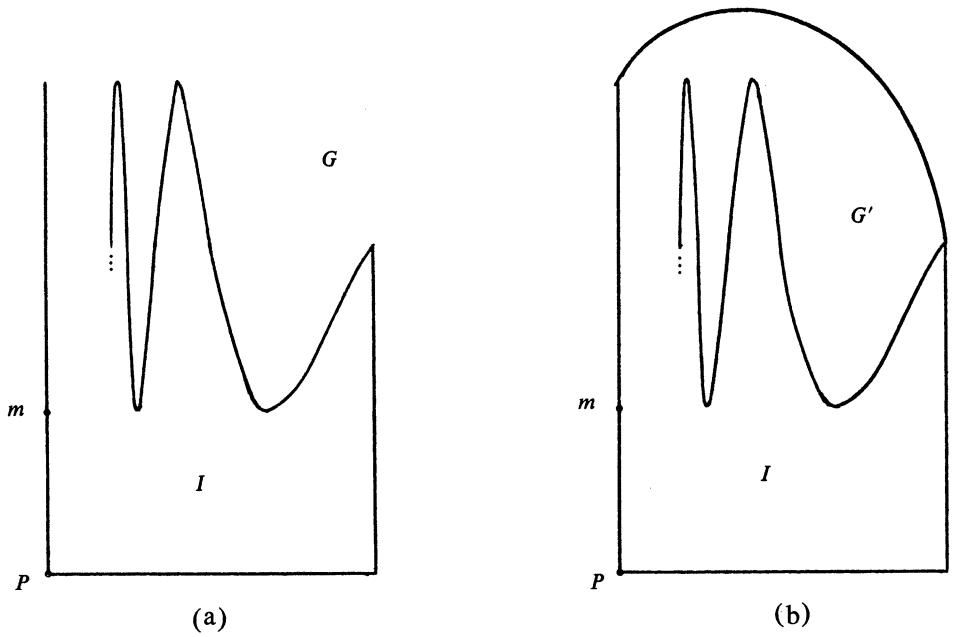


FIGURE 1.

Schoenflies gave importance to his new concept by proving that necessary and sufficient conditions for a closed and bounded subset M of the plane to be a simple closed curve are (i) that it separate the plane into exactly two regions, and (ii) that each point of M be accessible from all sides with respect to each of these regions. In studying this result we were pleasantly surprised to find an improvement by showing that condition (ii) implies M can separate the plane into no more than two regions. This permits condition (i) to be weakened considerably by replacing it with condition (i'): M separates the plane into at least two regions. (For more details, see [4, p. 60].)

One immediately asks if some modification of condition (i) to 3-space coupled with condition (ii) might also characterize 2-spheres (the analogs in 3-space of simple closed curves). The answer is no, as the example of a torus clearly demonstrates. What further conditions, then, must one impose on a set to insure that it be a sphere? With the torus in mind, a logical requirement might be that in addition to satisfying a modified condition (i) and condition (ii) the set have no "holes" in it (technically, the same homology as a sphere).

Perhaps tacitly assuming such an added hypothesis, L. E. J. Brouwer [1] thought he produced a counterexample, the example of our title, which separates 3-space into two regions such that every point of the surface is accessible from all sides with respect to each of the two regions of 3-space and which has the same homology as the 2-sphere, but yet is not a 2-sphere. A slightly modified version, which we shall call the "Brouwer sphere," B, can be expressed as the union of the following sets of points in cylindrical coordinates:

$$\{(\rho,\theta,z): 0<\rho<1, z=2+\cos\theta+(1-\sqrt{\cos^2\theta}\,)\sin(\pi/\rho)$$
$$\{(1,\theta,z): 0\leqslant z\leqslant 2+\cos\theta\},$$
$$\{(0,\theta,z): 1\leqslant z\leqslant 3\},$$
$$\{(\rho,\theta,0): 0\leqslant\rho\leqslant 1\}.$$

Various cross sections of B through the z-axis are given in FIGURE 2.

Brouwer claims without proof that at the point H with coordinates $(0,0,2)$, there is a singularity in the sense that there exists a sequence of points of B converging to H which cannot be joined by a simple arc lying entirely in B. However, the claim is not true! To see why, first

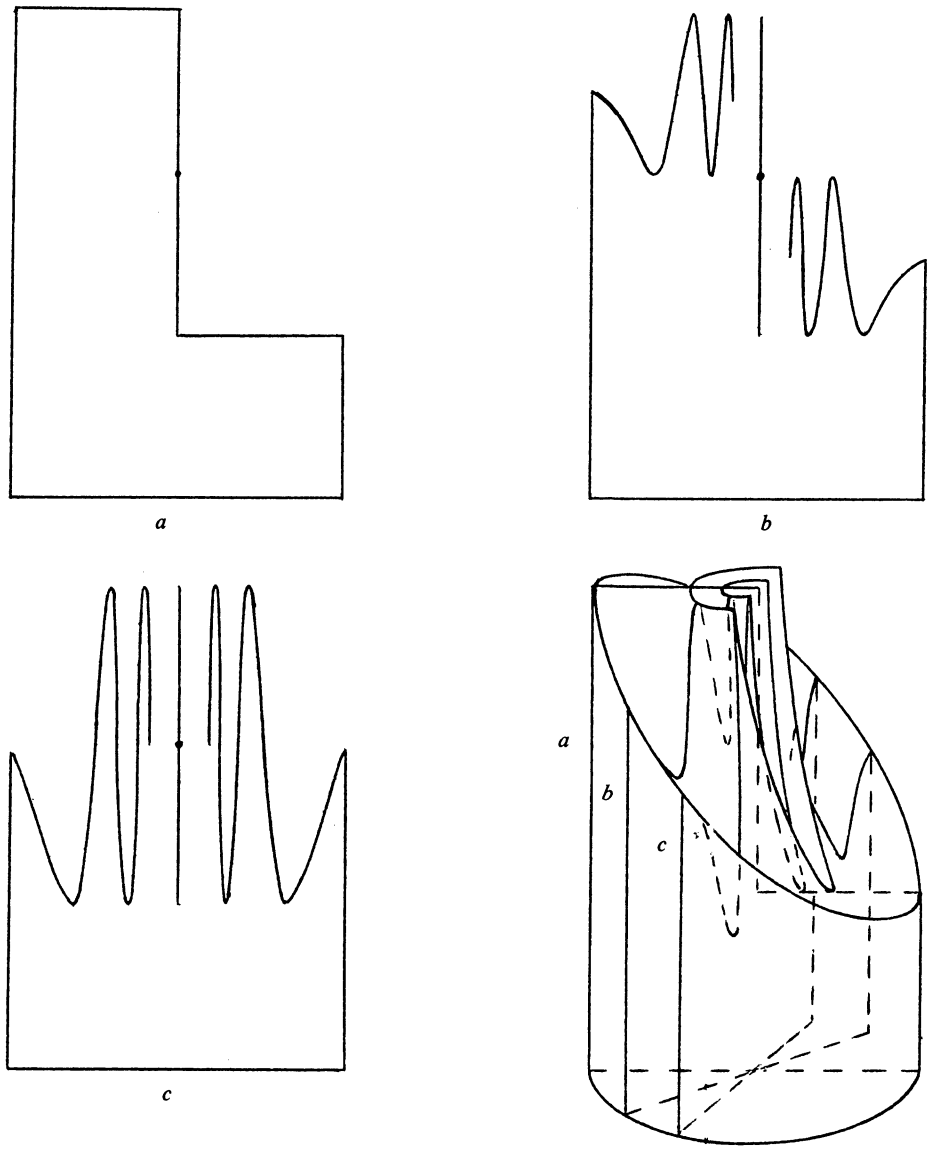


FIGURE 2.

separate B into a countable number of zones by planes parallel to the base plane $z=0$ and converging to the plane $z=2$ (for example, $z=2\pm\frac{1}{2}, 2\pm\frac{1}{3}, 2\pm\frac{1}{4}, \ldots$). Within any given zone there can be at most a finite number of points of the sequence. These points can be joined in B by a simple arc lying entirely in the given zone and not intersecting any of the previously constructed arcs. Furthermore, one can pass from one zone to any other zone along a simple arc. Now, to construct the desired simple arc containing all the points of the sequence, start with the furthest zone above H containing points of the sequence and construct a simple arc in that zone containing all the points of the sequence within that zone. Then pass along a simple arc to the furthest zone below H containing points of the sequence not already used in the construction. Alternating above and below H in this way, the arc under construction will eventually converge to the point H and will pass through all the points of the given sequence.

In fact, the Brouwer sphere is homeomorphic to the standard 2-sphere. To prove this it is sufficient to exhibit a homeomorphism from the top of the cylinder onto a planar disc. Such a

homeomorphism was first demonstrated by J. Rey Pastor [5]. The two-stage construction described below, arrived at independently by the present author, carries the same flavor as the one Pastor gave.

First map the top of the cylinder one-to-one and onto a concave hexagonal planar region, such as the one in FIGURE 3. Equations for this map (call if f) taking the point (ρ,θ,z) to the point (x,y) are as follows: for $\rho=0$, $1\leqslant z\leqslant 3$, let $x=0$, and $y=z-2$; for $0\leqslant\theta\leqslant\pi$, $0<\rho\leqslant 1$, let

$$x=-\rho, \text{ and } y=(1-\sqrt{\cos^2\theta}\,)\sin(\pi/\rho)+(1-\rho)\cos\theta;$$

and for $\pi<\theta<2\pi$, $0<\rho\leqslant 1$, let

$$x=\rho, \quad \text{and } y=(1-\sqrt{\cos^2\theta}\,)\sin(\pi/\rho)+(1+\rho)\cos\theta.$$

The only points where continuity is in question are those for which $\rho=0$. The continuity of f at a point $p=(0,\theta,z)$ can be verified by considering any sequence of points $\{p_i\}_{i=1}^{\infty}$ of B which converges to p and showing that the sequence of functional values $\{f(p_i)\}_{i=1}^{\infty}$ converges to $f(p)$.

The definition of the homeomorphism is completed by identifying the sides of the hexagon which are labelled "a" in FIGURE 3 and those labelled "b" in the obvious way to produce a disc.

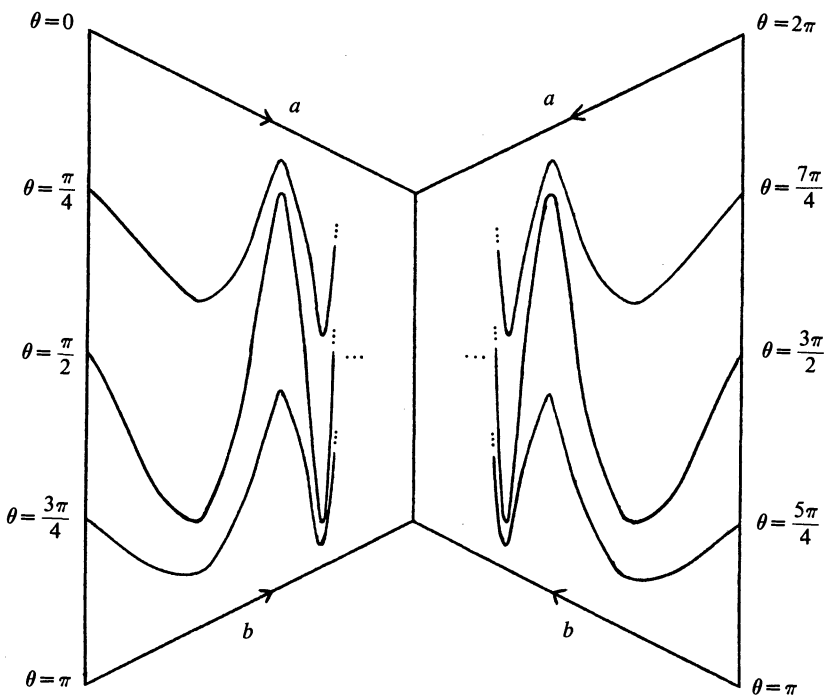


FIGURE 3.

Though the Brouwer sphere fails to be the counterexample Brouwer intended, his conjecture is still true. That is, conditions (i) and (ii), even coupled with the right homology, are not sufficient to guarantee a 2-sphere. We will present a valid counterexample below, but first we introduce a useful and necessary criterion for a surface in 3-space to be a 2-sphere, namely, that of being locally connected.

A point set X (with an associated topology) is **locally connected at a point** p in X, if for every neighborhood U of p, there is a connected neighborhood V of p with $V\subset U$. It is convenient in 3-space to think of these neighborhoods as balls with their centers at p. We say that X is **locally connected** if it is locally connected at each of its points. Clearly a sphere (or anything homeomorphic to it) is locally connected.

P. H. Doyle has provided us with the following example which has the same homology as the sphere and yet is not homeomorphic to the 2-sphere, since it is not locally connected. Consider the surface K of a cube with the curve T (defined in FIGURE 1(a)) embedded in one of its faces. All of the points of T are accessible from one of its complementary regions on K (the one corresponding to the region G in FIGURE 1(a)), but some fail to be accessible from the other (the one corresponding to region I). Delete the region with respect to which accessibility fails, and let K_0 be the resulting surface. Now take another copy of K_0 and identify the points of T on one copy of K_0 with the corresponding points of T on the other. The resulting surface is equivalent (that is, homeomorphic) to the surface of a cube with a sequence of "holes" of equal depth (say $1/2$ the height of the cube) and diameters tending toward zero "drilled" in one of its faces in such a way that the holes converge toward a line segment L on an adjoining face of the cube. (See FIGURE 4.) This surface has the same homology as the 2-sphere, but it is not locally connected at points along L. This gives us the desired counterexample.

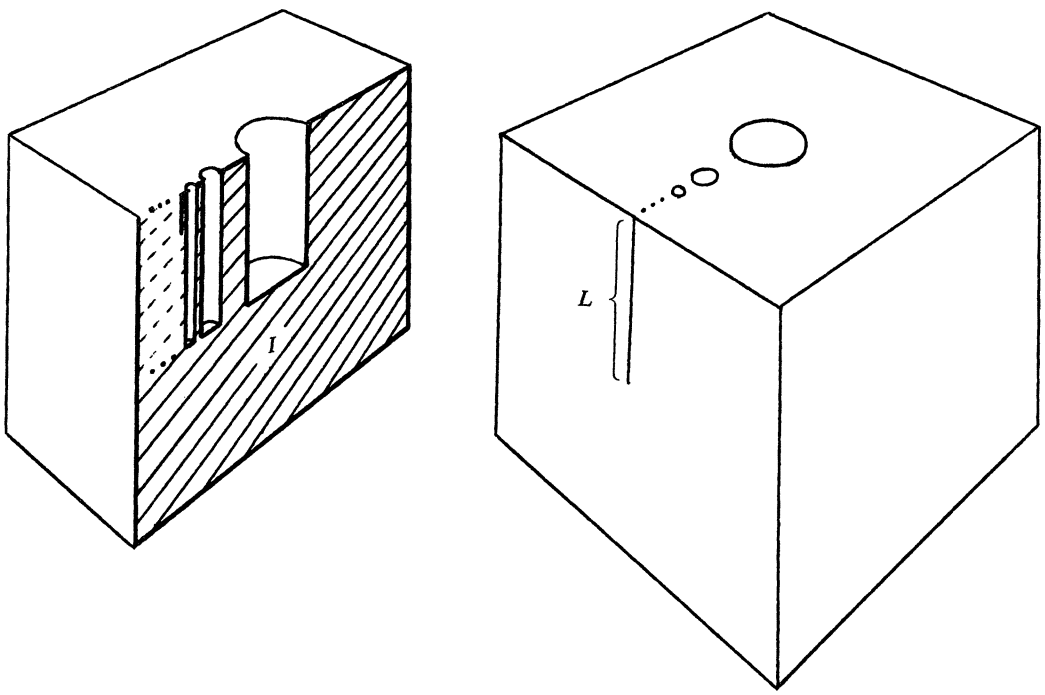


FIGURE 4.

A wise professor of mathematics is reported to have told his students, "If you can't draw a picture of what your proof is saying, then you haven't proven anything." However, as useful as a picture can be for stimulating thought and reinforcing intuition, it can also be misleading, as Brouwer's false counterexample illustrates. The most desirable mathematical demonstrations are those which use examples to motivate the ideas and then solid logic to support them.

References

[1] L. E. J. Brouwer, Über Jordansche Mannigfaltigkeiten, Math. Ann., 71 (1911) 320–327.

[2] J. G. Hocking and G. S. Young, Topology, Addison-Wesley (1961).

[3] I. Lakatos, Proofs and Refutations, Cambridge University Press (1976).

[4] J. M. McGrew, The Origins of Connectedness im Kleinen, Ph.D. Thesis, Michigan State University (1976).

[5] J. R. Pastor, Une propriété caractéristique des variétés de Jordan, Comptes Rendus de l'Acad. Sci. Paris, 192 (1931) 27–29.

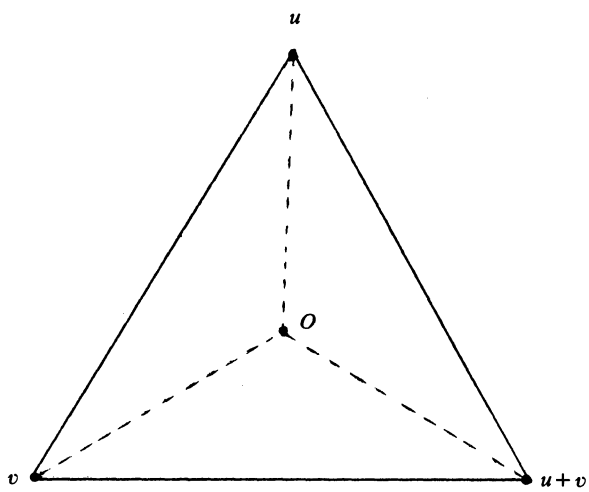
[6] A. M. Schoenflies, Die Entwicklung der Lehre von den Punktmannigfaltigkeiten, Part II, Deutsche Math. Verein.-Jahresb., Ergänzungsbände, 2 (1908) 1–331.

Finite Vector Spaces from Rotating Triangles

TONY CRILLY
Middlesex Polytechnic
Enfield, EN3 4SF England

A well-known example used to illustrate ideas in elementary group theory is the group of symmetries of an equilateral triangle. It is generated by a reflection and a rotation and is the symmetric group on three symbols. The same triangle can be regarded as representing a particularly simple vector space and we shall show that this point of view provides a geometric path to some ideas in the theory of field extensions and groups of linear transformations.

We denote by J_2 the field containing the integers 0 and 1 where addition and multiplication are carried out modulo 2. The vector space V of dimension two over this field then consists of four vectors: u, v, $u+v$, and 0. We associate these vectors with the triangle by placing the vector 0 at its centroid and the other vectors at the vertices (FIGURE 1). These vectors over J_2 should not be interpreted as directed line segments (arrows) in the usual way. Indeed, the idea of length does not work for these vectors as for the better-known vectors with scalars in the real number field. In particular, the parallelogram model in which we represent $u+v$ as the resultant of u and v in a parallelogram of vectors fails completely. However, we find it advantageous to think of vectors over J_2 as simply labelling the vertices and centroid in the triangular representation of V shown in FIGURE 1.



The vector space V.

FIGURE 1.

The reflection S of the triangle about the vertical line of symmetry gives $S(u)=u$, $S(v)=u+v$ and $S(u+v)=v$. These formulae show that S is linear! The rotation R_0 of the triangle through $2\pi/3$ in a counterclockwise direction yields $R_0(u)=v$ and $R_0(v)=u+v$ and $R_0(u+v)=u$. Again, R_0 is linear. Now each one dimensional subspace of V consists of the zero vector and a non-zero vector. In FIGURE 1 these are at opposite ends of a line segment joining the centroid to a vertex. The rotation R_0 maps these 3 subspaces cyclically and consequently it possesses no eigenvectors. Algebraically, this means that the characteristic polynomial for R_0 has no roots in J_2. The matrix R_0 (relative to the basis elements u, v taken as unit column vectors) is $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$

and its characteristic polynomial, $\det(A-\lambda I)$, is $\lambda^2+\lambda+1$. (Remember that -1 is equivalent to $+1$ in the field of J_2.) The geometric transformation R_0 has thus helped us see that the polynomial $\lambda^2+\lambda+1$ has no root in J_2.

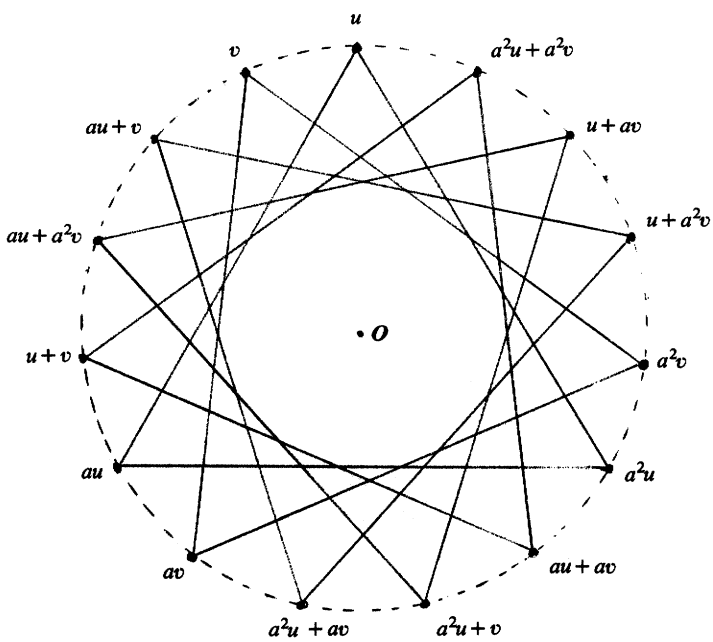
By adjoining a root, say a, of $\lambda^2+\lambda+1$ to the field J_2, we obtain the (extended) field $J_2(a)$ and we may study the new vector space $V(a)$ still generated by u, v, but with scalars in $J_2(a)$. $V(a)$ is the set $\{\alpha u+\beta v:\alpha,\beta\in J_2(a)\}$ where $J_2(a)=\{0,1,a,a^2:a^2+a+1=0\}$. An addition and multiplication table for $J_2(a)$ can be easily constructed. For example, if we add $a+1$ to both sides of $a^2+a+1=0$ (mod 2) we get $a^2=a+1$, so $a^3=a(a+1)=a^2+a=1$, again (mod 2). (An introduction to finite arithmetics and field extensions is given in [1].)

Our new space $V(a)$ has sixteen vectors, as there are four choices for each of α and β. These vectors form five one dimensional subspaces each containing three non-zero vectors with the zero vector common to each subspace. We may represent each subspace of the new vector space $V(a)$ as a centered triangle by representing $V(a)$ as the circular diagram of FIGURE 2 with the order of arrangement of its vectors chosen so that a counterclockwise rotation of $V(a)$ through $2\pi/15$ is a linear transformation. We denote this rotation by R_1 and define it by setting $R_1(u)=v$ and $R_1(v)=au+v$. The action of R_1 on all the other vectors is determined by linearity.

In what follows we shall use FIGURE 2 and diagrams like it to study the general linear group and its subgroups. These classical groups associated with finite fields have been extensively treated by Dickson [2].

There are, in fact, 180 one-to-one linear transformations of $V(a)$. We can see this by considering the possible images of u and v under a transformation T; there are 15 choices for $T(u)$, and once this has been chosen, 12 choices are available for $T(v)$. ($T(v)$ cannot map into the subspace containing $T(u)$ as T is required to be one-to-one.) These 180 transformations form a group denoted by $GL(2, 4)$, an example of the general linear group; the "2" here refers to the dimension of $V(a)$ and the "4" to the number of scalars in its underlying field $J_2(a)$.

We now investigate three special subgroups of $GL(2, 4)$ in conjunction with the geometric model for $V(a)$ shown in FIGURE 2. We first study the subgroup $HL(2, 4)$ whose elements keep



The vector space $V(a)$.

FIGURE 2.

one vector of $V(a)$ (say u) fixed. From the preceding argument, the order of $HL(2, 4)$ is 12. The generators of this group and the rotation R_1 form a set of generators for $GL(2, 4)$; to see this consider FIGURE 2 and any $T \in GL(2, 4)$. If the composition of k successive applications of R_1 to $V(a)$ is denoted by R_1^k and if $T(u)$ is the kth vector on $V(a)$ measured from u in a counterclockwise direction, we say the vector u is **rotated** to $T(u)$ by the linear transformation R_1^k. If the vector v is rotated to $T(v)$ by R_1^m and we choose to define $F \in HL(2, 4)$ by $F(u)=u$ and $F(v)=R_1^{m-k}(v)$ we obtain $T=R_1^kF$ on the linearly independent vectors u, v and hence on $V(a)$. (The convention we adopt for composing transformations is that the right hand transformation acts first.)

Our next subgroup of $GL(2, 4)$ is the subgroup consisting of those transformations which commute with every transformation, called the center of the group and denoted by $Z(GL(2, 4))$. Consider any transformation L in this subgroup and suppose $L(x)=y$ where x is any vector from $V(a)$. If the vectors x and y are linearly independent, then there is a linear transformation $T \in GL(2, 4)$ completely determined by $T(x)=ax$ and $T(y)=y$. Looking in turn at the action of TL and LT on the vector x, we find that $TL(x)=y$ and $LT(x)=ay$ whereas L is required to commute with all elements of $GL(2, 4)$ and in particular with T. We conclude that x and y are linearly dependent whereby either $y=x$, ax or a^2x. Corresponding to these values for y, there are three possibilities for L and they are represented by the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \begin{pmatrix} a^2 & 0 \\ 0 & a^2 \end{pmatrix}.$$

In the graphic framework of FIGURE 2, the first transformation is the identity and keeps all subspaces fixed; the second maps each triangular subspace into itself in a counterclockwise sense and the third acts similarly, but in a clockwise sense. It is apparent that these transformations do commute with every member of $GL(2, 4)$. Furthermore they are the only transformations that map each subspace into itself. For if $F(x)=\alpha x$, $F(y)=\beta y$ and $F(x+y)=\gamma(x+y)$ where x and y belong to different subspaces, then, by linearity, $F(x+y)=\alpha x+\beta y=\gamma(x+y)$ hence $\alpha=\gamma$ and $\beta=\gamma$.

The third type of subgroup of $GL(2, 4)$ is the one which consists of those transformations with determinant unity, called the special linear group and denoted by $SL(2, 4)$. By factoring out the determinant from any linear transformation T, we can write T as the composition of a transformation from the center with a transformation from the special linear group; thus $T=(\det T)^{1/2}IT'$ where I is the identity transformation and T' has determinant unity. The only transformation in the center which has determinant unity is the identity itself, so that $Z(GL(2,4)) \cap SL(2, 4)=\{I\}$ and also $GL(2, 4)=Z(GL(2, 4))\ SL(2, 4)$. Hence the order of $SL(2, 4)$ is $180/3=60$.

The group $GL(2, 4)$ is actually the direct product of the center and the special linear group as both these subgroups are normal in $GL(2, 4)$ and have only the identity transformation in common.

We now review our general method of construction, and procede to construct a chain of more exotic vector spaces with corresponding "pictures." The vector space $V(a)$ was obtained from the rotation R_0 of V. If we repeat the process and rotate the vector space $V(a)$ itself, we obtain a new vector space $V(b)$ from the rotation R_1 of $V(a)$ defined earlier by $R_1(u)=v$ and $R_1(v)=au+v$. The matrix of R_1 relative to u, v is $\begin{pmatrix} 0 & a \\ 1 & 1 \end{pmatrix}$ and its characteristic equation, which b must satisfy, is $\lambda^2+\lambda+a=0$. More generally, we can obtain the vector space $V(x_{n+1})$ by considering the rotation R_n of $V(x_n)$ defined by: $R_n(u)=v$ and $R_n(v)=x_nu+v$, where $x_0=1$ and x_{n+1} satisfies $\lambda^2+\lambda+x_n=0$. The vector space $V(x_n)$ itself is obtained at the nth stage of rotation; it is the two dimensional vector space over the finite field of characteristic 2 with r_n elements generated by x_n and denoted by $J_2(x_n)$.

How many elements does $J_2(x_n)$ have? In the beginning (with no rotations) we had the triangle space V with $r_0=2$ while at each rotation the number of elements in the base field was squared and so by induction $r_n=2^{2^n}$. For notational simplicity we introduce

$$p_n = 2^{2^n} - 1 \text{ and } q_n = 2^{2^n} + 1 \text{ for } n = 0, 1, \ldots .$$

In the vector space $V(x_n)$ there are q_n one dimensional subspaces each containing p_n nonzero vectors. Diagramatically, the one dimensional subspaces of $V(x_{n+1})$ each have the same form as the entire space $V(x_n)$. FIGURES 1 and 2 exhibit the special case $n=0$: $V(x_{n+1}) = V(a)$ and $V(x_n) = V$.

A one-to-one linear transformation T of the vector space $V(x_n)$ is called an involution if $T^2 = I$. In what follows we prove several results which describe the involutions of $V(x_n)$.

PROPOSITION 1. *A linear transformation T is an involution of $V(x_n)$ if and only if $T \in SL(2, r_n)$ and fixes a one dimensional subspace of $V(x_n)$.*

Proof. The statement is trivially true if $T = I$, so in what follows we assume T is not the identity. For an involution T there are linearly independent vectors x, y with $T(x) = y$ and $T(y) = x$, so, relative to the basis $\{x, y\}$ for $V(x_n)$, T is represented by the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ with characteristic equation $\lambda^2 + 1 = 0$. In the field $J_2(x_n)$ the only solution of this equation is $\lambda = 1$. Hence T fixes one subspace, interchanges the remaining r_n subspaces in pairs, and has determinant one.

Conversely, suppose $T \in SL(2, r_n)$ has the property of fixing one subspace. Let x be a fixed vector and choose another vector y so that x and y are linearly independent. The matrix of T relative to x, y is the form $\begin{pmatrix} 1 & \lambda \\ 0 & \beta \end{pmatrix}$ where $\beta = 1$ as $\det T = 1$. Multiplying this matrix by itself yields $T^2 = I$, as required.

At this point, it is worth noting that the involutions play an important part in the investigation of $GL(2, r_n)$. This is the meaning of our second result.

PROPOSITION 2. *A linear transformation T is a member of $SL(2, r_n)$ if and only if it can be written as the composition of two involutions.*

Proof. We again suppose that T is not the identity transformation as the statement would then be trivially true. We may assume there are linearly independent vectors x and y with $T(x) = y$ since T, a member of $SL(2, r_n)$, is not in the center of $GL(2, r_n)$ and therefore does not map each subspace into itself (cf. the discussion of $Z(GL(2, 4))$). If we write $T(y) = \alpha x + \beta y$ then $\alpha = 1$ if $\det T$ is to be 1. We have now to produce two involutions G, F with $T = GF$.

Using Proposition 1, we define F by $F(x) = x$ and $F(y) = \beta x + y$ and let G be the involution which interchanges x with y. Hence; $GF(x) = y$ and $GF(y) = G(\beta x + y) = \beta G(x) + G(y) = \beta y + x$ and therefore $T = GF$ as required.

For the converse statement we assume T is the composition of two involutions. The multiplication property of determinants implies that the determinant of T is the product of the determinants of the individual involutions. By Proposition 1 the determinant of an involution is 1 and hence the determinant of T itself is also 1. Thus $T \in SL(2, r_n)$ and the proof of Proposition 2 is complete.

Although we are primarily concerned with a vector space over the field $J_2(x_n)$, it is convenient to digress here in order to remark on the generality of the proposition just proved. With minor modifications to the argument given, we can show that a linear transformation of any two-dimensional linear space with determinant one is expressible as the composition of two involutions. But the converse statement is only true if the underlying field K is of characteristic two. If K has characteristic other than two, an involution which interchanges at least one pair of subspaces had determinant -1. Recall that we may represent such an involution by the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. If this involution is composed with an involution of determinant $+1$, then the determinant of the composition is -1, and therefore the composition is not a member of the special linear group $SL(2, K)$.

We shall now apply the results of Propositions 1 and 2 to describe the effect of a linear transformation on the one-dimensional subspaces of $V(x_n)$.

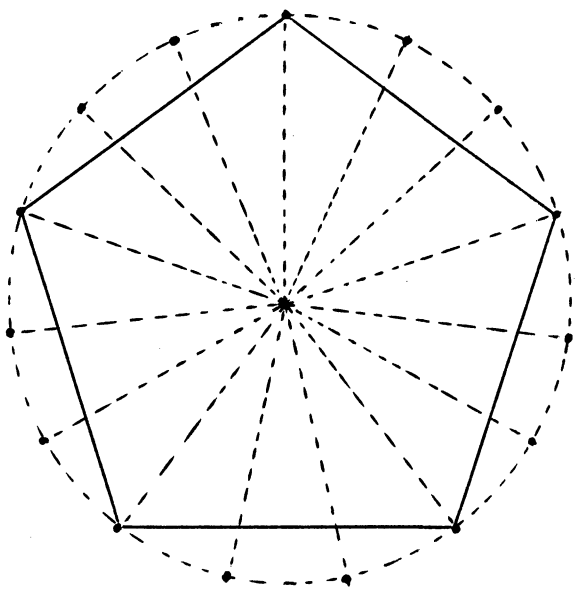
PROPOSITION 3. *For $n \geqslant 1$, the permutation of subspaces of $V(x_n)$ induced by any $T \in GL(2, r_n)$ is an even permutation.*

Proof. Without loss of generality we may consider $T \in SL(2, r_n)$ because any transformation is the composition of a transformation from the special linear group with one from the center of $GL(2, r_n)$ and the latter transformation does not permute subspaces. First we consider the action of an involution F on the q_n subspaces of $V(x_n)$. By Proposition 1, F keeps one subspace fixed and interchanges the remaining $r_n = 2^{2^n}$ in pairs. Therefore, F determines the composition of an even number ($\frac{1}{2} r_n$ is even if $n \geqslant 1$) of interchanges of subspaces and thus an even permutation of the subspaces. By Proposition 2, $T \in SL(2, r_n)$ may be written as the composition of two involutions and therefore T determines an even permutation of the subspaces as the composition of even permutations is even.

Proposition 3 suggests how the linear groups under discussion might be linked to the classical alternating groups. Recalling that the space $V(x_n)$ has q_n subspaces, the homomorphism θ (which takes each linear transformation to the induced transformation on subspaces) maps $GL(2, r_n)$ onto a subgroup of the alternating group A_{q_n} on q_n symbols. Notice that the kernel of θ is the center of $GL(2, r_n)$. In fact, for $n \geqslant 1$, the group $SL(2, r_n)$ is isomorphic to a subgroup of A_{q_n}. For if θ is restricted to $SL(2, r_n)$ the kernel of this restriction is the identity as $SL(2, r_n) \cap Z(GL(2, r_n)) = \{I\}$. Hence, this restriction is an isomorphism. In the case $n = 1$, the order of $SL(2, 4)$ is, as noted above, 60; therefore $SL(2, 4)$ is isomorphic to A_5.

Similarly, for $n \geqslant 1$, the group $HL(2, r_n)$ is isomorphic to a subgroup of A_{r_n}. This group keeps one vector fixed so that θ restricted to this subgroup maps it into $A_{q_n - 1} = A_{r_n}$. Once again the appropriate kernel is the identity and the restriction of θ to $HL(2, r_n)$ is an isomorphism. In the case $n = 1$, the order of $HL(2, 4)$ is 12 and therefore $HL(2, 4)$ is isomorphic to A_4.

We close with some observations about eigenvectors and dihedral groups, the symmetry groups of regular polygons. We show here that $SL(2, r_n)$ contains a dihedral group and rather surprisingly that the corresponding polygon appears in the circular diagram representing $V(x_n)$. In the case $n = 1$, the appropriate polygon is the pentagon (FIGURE 3). In studying the effect of a linear transformation of $V(x_n)$ it is convenient to make use of its eigenvectors. Of course, these may not exist in $V(x_n)$ but even if this is the case, they may exist in the 'higher' space $V(x_{n+1})$.



The pentagon in $V(a)$.

FIGURE 3.

This is analogous to rotation of the real plane through a right angle where the eigenvectors exist in the complex plane. The analogy is more evident when we consider the rotation R_n of the space $V(x_n)$: it has characteristic polynomial $\lambda^2+\lambda+x_n$ where the symbol x_{n+1} is introduced as a root of this polynomial. By direct substitution it can be seen that $x_n x_{n+1}^{-1}$ is another root of the polynomial and corresponding to these eigenvalues the respective eigenvectors in $V(x_{n+1})$ are $x_n u + x_{n+1} v$ and $x_{n+1} u + v$. Relative to this eigenvector basis for $V(x_{n+1})$, R_n is represented by the matrix

$$\begin{pmatrix} x_{n+1} & 0 \\ 0 & x_n x_{n+1}^{-1} \end{pmatrix}$$

and relative to the same basis, the reflection S of $V(x_n)$ is represented by the matrix

$$\begin{pmatrix} 0 & x_{n+1}^{-1} \\ x_{n+1} & 0 \end{pmatrix}.$$

By multiplication of the matrices and the identity $x_n^{p_n}=1$ we have $SR^{p_n}S=R_n^{-p_n}$, and on putting $Q=R_n^{p_n}$ we have $SQS=Q^{-1}$ and $Q^{q_n}=I$. These relations between the transformations S and Q show that $SL(2,r_n)$ contains the dihedral group of the q_n-sided polygon with $q_n=3$, 5, 17, 257, 65537,.... In the circular diagram for $V(x_n)$ these regular polygons appear symmetrically placed with respect to the vertical axis of symmetry. The numbers q_n are known as the Fermat numbers and arise in the celebrated constructibility problems associated with Gauss [3]; they have many interesting properties as well as a long history [4].

In this note the initial point of departure was the equilateral triangle viewed as a vector space of two dimensions. It may be interesting to generalise this to the n-dimensional case in which the vector space over the field with two elements is representable as a 2^n-1 sided regular polygon.

References

[1] W. W. Sawyer, A Concrete Approach to Abstract Algebra, W. H. Freeman, San Francisco, 1959.
[2] L. E. Dickson, Linear Groups (1900), Dover Reprint, New York, 1958.
[3] F. Klein, Famous Problems of Elementary Geometry (1895), Dover Reprint, New York, 1956.
[4] O. Ore, Number Theory and its History, McGraw-Hill, New York, 1948.

Locks, Keys and Majority Voting

Donald McCarthy
St. John's University
Jamaica, NY 11439

The following combinatorial problem appears in the excellent book of Yaglom and Yaglom [4; p. 5, Problem 9] and is used as an example in at least one major text on combinatorics [1; pp. 8–9, Example 1–11].

> A group of 11 scientists are working on a secret project, the materials of which are kept in a safe. They want to be able to open the safe only when a majority of the group is present. Therefore the safe is provided with a number of different locks, and each scientist is given the keys to certain of these locks. How many locks are required, and how many keys must each scientist have?

The amusing answer given is that 462 locks are required, and that each scientist must carry 252 keys. The significance of these particular numbers is that $462=\binom{11}{5}$ and $252=\binom{10}{5}$. In

general, as remarked in the solution presented in [4, p. 49], "if there are n scientists and it is required that the safe can be opened if and only if at least m of them are present, then the minimum number of locks is $\binom{n}{m-1}$, and the minimum number of keys held by each scientist is $\binom{n-1}{m-1}$. This indicates that such a system of safekeeping is impractical, since even for comparatively small values of m and n, opening the safe would require a whole day of fumbling with keys."

In presenting this cute result before a class, I encountered resistance from one student who said he felt sure it somehow could be done with fewer locks. Although his initial objections seemed to be based on little more than pragmatic skepticism, it turns out that he was right! Several abortive attempts to design ingenious lock arrangements eventually led to the germ of a rather neat and surprisingly simple solution to the problem which does indeed involve far fewer locks and keys. Before revealing this "practical" solution, it seems worth presenting the nice line of argument which yields the "impractical" results cited above. Knowing that a counterexample will follow should provide incentive for keeping on the lookout for gaps in the reasoning. To this end, the argument will be presented a little more formally than in [1] or [4].

Assume as above that we have a group of n scientists, and a safe which can be opened precisely when at least m scientists are present. (To dispense with vacuous cases, assume $1 \leqslant m \leqslant n$.) Let L be the minimal number of locks required for this, and let K be the minimal number of keys per scientist. We shall show that $L=\binom{n}{m-1}$ and $K=\binom{n-1}{m-1}$.

Let $\mathcal{L}$ be a set of locks which meets the stipulated condition and let $\mathcal{C}$ denote the collection of all $(m-1)$-element subsets of the set S of scientists. We begin by showing that $L \geqslant \binom{n}{m-1}$. Suppose we can establish the existence of a 1-1 function f from $\mathcal{C}$ into $\mathcal{L}$. Then $|\mathcal{L}| \geqslant |\mathcal{C}| = \binom{n}{m-1}$ as desired. Such a function f can be obtained as follows. Let $A \in \mathcal{C}$; since $|A| < m$ there must exist a lock which cannot be opened by any member of A. Take $f(A)$ to be one such lock. This defines a function f from $\mathcal{C}$ to $\mathcal{L}$. To see that f is 1-1, let $A, B \in \mathcal{C}$ with $A \neq B$; we show that $f(A) \neq f(B)$. Since $A \neq B$, $A \cup B$ properly contains A, hence $|A \cup B| \geqslant m$. Thus the safe can be opened by $A \cup B$. This implies, in particular, that some scientist $s \in A \cup B$ must be able to open lock $f(A)$. Since no member of A can open $f(A)$, we have $s \in B$. Thus s can open $f(A)$ but not $f(B)$, hence the two locks are distinct; i.e. $f(A) \neq f(B)$. This completes the proof that $L \geqslant \binom{n}{m-1}$.

Next we show that $K \geqslant \binom{n-1}{m-1}$. This requires, of course, the assumption that there are no "master" keys; explicitly, we assume that each key opens only one lock. For each $s \in S$, let $\mathcal{K}(s)$ be the set of all keys possessed by s, and $\mathcal{C}(s)$ the set of all $(m-1)$-element subsets of S which do not contain s. As before, it suffices to show the existence of a 1-1 function g from $\mathcal{C}(s)$ into $\mathcal{K}(s)$, and thus obtain $|\mathcal{K}(s)| \geqslant |\mathcal{C}(s)| = \binom{n-1}{m-1}$. If $A \in \mathcal{C}(s)$, then $|A \cup \{s\}| = m$, hence the safe can be opened by $A \cup \{s\}$. It follows that s must possess a key to the lock $f(A)$; let $g(A)$ be such a key. This defines a function g from $\mathcal{C}(s)$ to $\mathcal{K}(s)$, and it is easy to see that g is 1-1. (For if $g(A) = g(B)$, this key unlocks both $f(A)$ and $f(B)$; it follows that $f(A) = f(B)$ hence $A = B$.)

So far we have shown that $L \geqslant \binom{n}{m-1}$ and $K \geqslant \binom{n-1}{m-1}$. To see that equality holds, we need only exhibit an acceptable configuration of locks and keys with $|\mathcal{L}| = \binom{n}{m-1}$ and $|\mathcal{K}(s)| = \binom{n-1}{m-1}$ for each $s \in S$. This is easily done. For each $A \in \mathcal{C}$, attach a lock $L(A)$ to the safe and give keys for $L(A)$ to just those members of S in the complement of A. Let $\mathcal{L}$ be the resulting set of locks: one for each $A \in \mathcal{C}$. Note that $L(A)$ cannot be opened by a set B of scientists if and only if $B \subseteq A$. Since this last condition requires $|B| < m$, we see that all locks in $\mathcal{L}$ can be opened by B precisely when $|B| \geqslant m$. We see also that $\mathcal{L}$ is in 1-1 correspondence with $\mathcal{C}$; thus we have an acceptable set of locks with $|\mathcal{L}| = \binom{n}{m-1}$. Note, too, that each scientist s receives keys to precisely those $L(A)$ for which $A \in \mathcal{C}(s)$. Hence, for each $s \in S$, $|\mathcal{K}(s)| = \binom{n-1}{m-1}$. This completes the proof that the minimal values cited for K and L are attained.

Now we give an example of an assignment of locks and keys which satisfies the conditions of the problem, yet involves only n locks, with each scientist carrying but a single key. Whenever $m < n$, this clearly conflicts with the results of the previous section. The precise source of the conflict will be revealed after the example.

For explicitness, we describe a concrete physical situation. Consider a safe whose door is held closed by a large bolt. If the bolt is slid to the left at least m inches, the door can swing open freely, but not otherwise. The movement of the bolt to the left is impeded, however, by the presence of n locks located on the track along which the bolt slides. The crucial feature of this arrangement is that the locks are not of the padlock variety, attached at a fixed position on the track, but rather are the sort which can be used to lock a telephone dial, say. When "unlocked" such a lock can be removed entirely from the track; when "locked" it cannot be so removed, but is otherwise free to slide along the track. We want each of our locks to be just one inch in width, so that the removal of any k locks will permit the bolt to slide precisely k inches to the left. Thus the safe can be opened when and only when at least m locks have been removed. By providing each of the n scientists with the key to exactly one lock (with different scientists assigned to different locks), we have accomplished our objective.

This apparently shows that $L \leqslant n$ and $K \leqslant 1$; also, it is quite easy to verify that $L \geqslant n$ and $K \geqslant 1$ (the interested reader is invited to fill in the details). Thus we obtain $L = n$ and $K = 1$ rather than $L = \binom{n}{m-1}$ and $K = \binom{n-1}{m-1}$ as claimed earlier. Wherein lies the source of this conflict? Precisely what was wrong with the earlier proof?

The answer is that our proof made use of an assumption which, as the above example demonstrates, was not warranted by the actual statement of the problem. The hidden assumption is this: in order to open the safe, all the locks must be opened. This seemingly innocuous assumption was used crucially in the proof that f and g are 1-1. If we add this condition (along with the hypothesis that each key opens only one lock) as a requirement of the problem, then the earlier values for L and K are indeed correct. Without the condition that all locks must be removed, the correct values are $L = n$ and $K = 1$.

Once spelled out, the conflict between the "impractical" theorem and the "practical" example above may seem transparent and perhaps even trivial or petty. Nevertheless, in a classroom setting this simple example can serve as a good illustration of larger issues: e.g., the subversive role played by hidden assumptions in the context of creative problem solving or mathematical modelling; or, in a different vein, acknowledgement of the potential gap between mathematical idealizations and the pragmatic demands of "real" situations. Another point worth illustrating is that attempts to resolve difficulties related to such gaps between theory and reality often lead to further mathematical investigation or broader applicability of results—sometimes in tangential directions.

Consider, for example, an alternative way of implementing the sliding bolt solution. Rather than using a sliding bolt with moveable locks, we can utilize a safe with an "intelligent" door—one controlled by a simple computer. Each of the n scientists is again provided with a single key which can be used to turn a switch. The positions of the n switches are monitored by the computer, operating under a trivial program which will release the safe door precisely when at least m switches have been turned to the proper position. (As a challenge, the reader might like to investigate the possibility of dispensing with the computer in this scheme, e.g., by designing an appropriate electrical network which could accomplish the same purpose.) This implementation facilitates a conceptual shift in the character of the original problem and suggests further applications. Namely, rather than envisioning a situation where scientists must spend their time fumbling with keys, we can think of them simply being present to "vote" their preference as to whether or not the safe should be opened. If at least m votes are cast in favor of opening the safe, it is opened; otherwise it remains closed. From this viewpoint, the locks and keys serve primarily as a mechanism for registering individual votes. Hence we are led naturally from our lock and key problem to ponder the formulation of a more abstract voting problem.

Suppose we have an assembly comprised of n members who are to vote their preference among a set of k alternatives $A_1, A_2, \ldots, A_k$. When $k = 2$ we have described a device which will provide immediate results of a ballot (if desired, a secret ballot) where alternative A_1 wins if it receives at least m votes, and A_2 wins otherwise. Note that here the criterion for A_i winning is

not generally the same for $i=1$ as for $i=2$. In many situations it is natural to treat the alternatives homogeneously, so that the same criteria are applied uniformly for all $A_i, 1 \leqslant i \leqslant k$. When $k=2$, this homogeneity condition is satisfied in our earlier example only when $m=[n/2]+1$. In this case, the winning alternative is the one which receives a majority of votes, provided that an abstention is interpreted as a vote for one particular alternative. In situations where abstentions are discounted, we can declare the winning alternative to be the one which receives the largest number of votes cast. This last procedure is clearly homogeneous and can be applied whenever $k \geqslant 2$. When $k>2$, however, it has certain deficiencies which lie not in implementation but rather in the degree to which the winning alternative can be regarded as faithfully representing the overall preference of the voters. (E.g., a sizeable majority of voters may actually be strongly opposed to the winning alternative.) Moreover, it turns out that such deficiencies are not confined to this particular procedure, but are indicative of a more deeply rooted problem in the theory of social choice among multiple alternatives.

Here the problem is to come up with a satisfactory procedure for ranking the alternatives $A_1, A_2, \ldots, A_k$ in a manner which reflects the overall preference of the voters, each of whose individual rankings has been supplied. There is a theorem of Kenneth Arrow (a Nobel laureate in economics) which states that if certain fairly reasonable conditions are to be met, then, when $k>2$, no such satisfactory procedure exists.

An excellent introduction to the topic of social choice (also known as the theory of elections) can be obtained by reading Chapter 10 of [2] followed by Chapter 7 of [3]. The discussion in [3] just after the statement of Arrow's nonexistence theorem (Section 7.2.2) is especially relevant to the point at hand. It provides good evidence of the variety and extent of mathematical activity which can be engendered when a theoretical result runs counter to the practical demands of a motivating application.

The author gratefully acknowledges the support of the National Science Foundation (Grant No. SM177-17448) and that of St. John's University in the form of a Research Leave.

References

[1] C. L. Liu, Introduction to Combinatorial Mathematics, McGraw-Hill, New York, 1968.
[2] J. Malkevitch and W. Meyer, Graphs, Models and Finite Mathematics, Prentice-Hall, Englewood Cliffs, N.J., 1974.
[3] F. S. Roberts, Discrete Mathematical Models with Applications to Social, Biological and Environmental Problems, Prentice-Hall, Englewood Cliffs, N.J., 1976.
[4] A. M. Yaglom and I. M. Yaglom, Challenging Mathematical Problems with Elementary Solutions, Vol. 1, Holden-Day, San Francisco, 1964.

A Useful Characterization of a Normal Subgroup

FRANCIS E. MASAT
Glassboro State College
Glassboro, N.J. 08028

In the following note we develop an alternate definition of a normal subgroup of a group which does not seem to be widely known, and then provide some examples of its efficiency and usefulness. This new definition comes from semigroup theory, and yields insight into the more traditional approaches to normal subgroups, factor groups, and homomorphisms.

The usual definition states that a subgroup H of a group G is a **normal** subgroup of G if for each $g \in G, gHg^{-1} \subseteq H$. Our alternative is based on the following related idea: a subgroup H of a group G will be called **reflexive** if for $a, b \in G, ab \in H$ implies that $ba \in H$. Informally, we call this the "flip-flop" property. These two definitions, it turns out, are equivalent: *If H is a subgroup of a group G, then H is normal if and only if H is reflexive.* To see this, let H be reflexive. Since $g^{-1}(gh) = h \in H$ for any $g \in G$, then by "flip-flopping," we have $(gh)g^{-1} \in H$, and therefore H is normal. Conversely, if H is normal and $ab \in H$ for $a, b \in G$, then $ba = a^{-1}(ab)a \in H$. So H is reflexive.

Let H be a subgroup of a group G and define the relation R on G by aRb if $ab^{-1} \in H$. The relation R is an equivalence relation on G and is usually called "congruence modulo H." It is precisely the reflexiveness of H which is necessary and sufficient in order that R be a congruence relation on G, i.e., an equivalence relation which respects group multiplication. In other words, *R is a congruence on G if and only if the subgroup H is reflexive (normal).*

To prove this we recall first that R is an equivalence relation on G. If aRb and H is reflexive, then $ab^{-1}(g^{-1}g) \in H$ implies that $g(ab^{-1})g^{-1} \in H$; so $ga(gb)^{-1} \in H$. Moreover, for any $g \in G, ag(bg)^{-1} = a(gg^{-1})b^{-1} = ab^{-1} \in H$. Thus R is a congruence relation on G. Conversely, if $ab \in H$, then $ab = a(b^{-1})^{-1}$ says that aRb^{-1}. Since R is a congruence relation, multiplying on the left by b yields $baRbb^{-1}$. This implies that $ba \in H$, and therefore H is reflexive.

The importance of this result is that if you have defined a congruence relation on a group, then you have at the same time defined a homomorphism on the group; and conversely. In particular, when H is normal (reflexive), the cosets of H *coincide* with the congruence classes of R. Thus G/H, the factor group of G modulo H, is exactly the same as G/R, the factor semigroup of G modulo R.

We now present some examples which illustrate the reflexive viewpoint and its ease of use. These examples involve subgroups, generation of normal subgroups, and homomorphisms. We shall use $<$ to denote "is a subgroup of" and $\lhd$ to denote "is a normal subgroup of."

Example 1. Let G be a group with $H \lhd G, K < G$.
(i) If $K \lhd G$, then $H \cap K \lhd G$.
(ii) If $K < G$, then $H \cap K \lhd K$.
(iii) If $H < K < G$, then $H \lhd K$.

For (i), let $a, b \in G$. Then $ab \in H \cap K$ implies $ab \in H$ and $ab \in K$, so $ba \in H$ and $ba \in K$, since each is normal. Hence $ba \in H \cap K$; i.e., $H \cap K \lhd G$. For (ii), if $i, j \in K$ such that $ij \in H \cap K$, then $ij \in H$ and $ij \in K$. Thus $ji \in H$ (since $H \lhd G$) and $ji \in K$ (since $K < G$); i.e., $ji \in H \cap K$. Item (iii) follows similarly (try it!).

Example 2. If C denotes the commutator subgroup of G (the subgroup generated by elements of the form $aba^{-1}b^{-1}$), then $C \lhd G$. To show that this is true, let $ab \in C$, such that $a, b \in G$. By the definition of C, we have that $bab^{-1}a^{-1} \in C$. It follows that $(bab^{-1}a^{-1})ab \in C$, so $ba(b^{-1}a^{-1}ab) = ba \in C$, and therefore $C \lhd G$.

Example 3. Let θ be a homomorphism of a group G onto a group H.
(i) If $A \lhd G$, then $A\theta \lhd H$.
(ii) If $B \lhd H$, then $B\theta^{-1} \lhd G$.

For (i), let $x, y \in H$ such that $xy \in A\theta$. There then exist $p, q \in G$, $r \in A$, and $k \in \text{Ker } \theta$ such that $x = p\theta$, $y = q\theta$, $r\theta = xy$, and $r = pqk$. Since A is reflexive, $p(qk) \in A$ implies that $(qk)p \in A$ and hence $(qk)p\theta = yx \in A\theta$. For (ii), let $a, b \in G$ such that $ab \in B\theta^{-1}$. There then exists $h, k \in H$ such that $a\theta = h, b\theta = k$. Thus, $ab\theta = hk$, so $hk \in B$. But $B \lhd H$ implies $kh \in B$, and hence $ba\theta = kh$; i.e., $ba \in B\theta^{-1}$.

Other examples and exercises (e.g., the center of a group is normal, all subgroups of an abelian subgroup are normal, normality is preserved under direct products) may also be used to illustrate the ease of use of the property of reflexiveness. This is not to say that the more

traditional approach of manipulating inverses, i.e., showing that $ghg^{-1} \in H$ when $H \lhd G$ and $g \in G$, should be avoided. Rather, coupling the "flip-flop" property of reflexiveness with the traditional approach provides simpler and more understandable proofs and solutions.

In a more general setting, it is shown by R. R. Stoll in [5] that a semigroup similarly contains a normal subsemigroup corresponding to each group homomorphism of the semigroup. Such normal subsemigroups have to satisfy the following three properties, where H is a subsemigroup of the semigroup G.

(A) For each $g \in G, gG$ and Gg meet H.
(B) For $a, b \in G, a$ and $ab \in H$ imply that $b \in H$.
(C) For $a, b \in G, ab \in H$ implies that $ba \in H$.

Note that any non-empty subset H of a group satisfies (A). Moreover, it is quite easy to show that the following are equivalent:

(1) H is a subgroup of S.
(2) H satisfies property (B).
(3) If $a, b \in H$ then $a^{-1}b \in H$.

So by elimination the remaining property, (C), is the one which is necessary and sufficient in order that a subgroup be the kernel of a homomorphism, and thus a normal subgroup.

References

[1] A. H. Clifford and G. B. Preston, The Algebraic Theory of Semigroups, Vol. 1, Chapter 1, Sections 1.4 and 1.5, American Mathematical Society, R. I., 1967.
[2] F. E. Masat, Right group and group congruences on a regular semigroup, Duke Math. J., 40 (1973) 393–402.
[3] J. T. Moore, Elements of Abstract Algebra, Chapter 3, Macmillan, N. Y., 1962.
[4] W. R. Scott, Group Theory, Chapter 2, Sections 2.1 and 2.2, Prentice-Hall, N. J., 1964.
[5] R. R. Stoll, Homomorphisms of a semigroup onto a group, Amer. J. Math., 73 (1951) 475–481.

The Inverse of a Sum Can Be the Sum of the Inverses

THOMAS E. ELSNER
General Motors Institute
Flint, MI 48502

Those who have experienced the frustrations of teaching algebra at any level frequently encounter student errors that are the result of mixing identities of related operations. Students tend (especially during the stress of testing) to ascribe the truth of one pattern to a different operation where it is not valid. The author has been particularly betrayed by the obviousness and simplicity of the transpose identity $(A+B)^T \equiv A^T + B^T$ in beginning courses in matrix algebra for engineering students. Many students "fall" for the same structure in creating nonidentities for other unary operations such as $|A+B| = |A| + |B|$ and especially $(A+B)^{-1} = A^{-1} + B^{-1}$. Undoubtedly, there are similar examples in many subjects. One way to foster an appreciation for the incorrectness of a nonidentity is to try and solve it as an equation. This note

considers the equation $(a+b)^{-1}=a^{-1}+b^{-1}$ in general and specifically in the case of matrix rings. We begin by considering the necessary relationship between a and b in such equations which will show the way to the rest of our results.

THEOREM. *Suppose a, b and $a+b$ are invertible elements of a ring R with identity element* 1. *Then $(a+b)^{-1}=a^{-1}+b^{-1}$ if and only if there is some p in R such that $b=ap$ and $p^2+p+1=0$.*

Proof. Suppose $(a+b)^{-1}=a^{-1}+b^{-1}$. The natural candidate for p is $a^{-1}b$. Direct substitution shows that $p^2+p+1=a^{-1}ba^{-1}b+a^{-1}b+1=a^{-1}(ba^{-1}b+b)+1$. That $ba^{-1}b+b=-a$ follows by multiplying the equation $(a+b)^{-1}=a^{-1}+b^{-1}$ by $a+b$ on the left and by b on the right to obtain $b=2b+a+ba^{-1}b$.

Conversely, suppose $b=ap$ and $p^2+p+1=0$. The latter implies that $p^{-1}=p^2$ and $(1+p)^{-1}=1+p^2$ and so $(a+b)^{-1}=(a+ap)^{-1}=(a(1+p))^{-1}=(1+p)^{-1}a^{-1}=(1+p^2)a^{-1}=a^{-1}+p^{-1}a^{-1}=a^{-1}+b^{-1}$. This completes the proof.

Because of this result, construction of examples depends on finding the roots of the polynomial p^2+p+1 in the ring R. For instance if R is the reals we find that the equation $(a+b)^{-1}=a^{-1}+b^{-1}$ has no solutions. However, if R is the complex field, then there are an infinity of pairs (a,b), where $b=a(-\frac{1}{2}\pm\frac{\sqrt{3}}{2}i)$, that solve the equation. For noncommutative rings the number of roots for p^2+p+1 may be infinite. This follows since if p satisfies the equation $p^2+p+1=0$ then so does any conjugate $q^{-1}pq$, since $(q^{-1}pq)^2+(q^{-1}pq)+1=q^{-1}(p^2+p+1)q=0$. The condition $p^2+p+1=0$ severely limits the fields in which the "inverse of a sum is the sum of the inverses" is an identity.

COROLLARY. *The fields in which $(a+b)^{-1}=a^{-1}+b^{-1}$ holds whenever a, b and $a+b$ are invertible are the fields with* 2, 3, *or* 4 *elements.*

Proof. The polynomial p^2+p+1 has at most 2 roots in a field. Given an invertible element, a, there are at most two other invertible elements $b=ap$ that meet the conditions of the theorem. This limits the number of invertible elements to 3 and so the field has at most 4 elements. By direct examination the three possible fields have the identity stated.

Let us return to our particular interest, the ring R of matrices over a field F. Recall that the companion matrix $P=\begin{bmatrix}0 & -1\\ 1 & -1\end{bmatrix}$ of the polynomial x^2+x+1 satisfies as its characteristic equation $P^2+P+I=0$ where I is reserved for the identity matrix. Hence for a given invertible 2×2 matrix A there are an infinite number of invertible 2×2 matrices B such that $(A+B)^{-1}=A^{-1}+B^{-1}$. These may be constructed as $B=AQ^{-1}PQ$ where Q is any invertible 2×2 matrix. As an example with integer entries, for $A=\begin{bmatrix}5 & 3\\ 2 & 1\end{bmatrix}$ let $B=AP=\begin{bmatrix}3 & -8\\ 1 & -3\end{bmatrix}$. For larger matrices of even order $2n\times2n$ let C be the companion matrix of $x^{2n}+x^n+1$ and then let $P=C^n$ to get $P^2+P+I=0$.

The problem that motivated this investigation was that of finding a simple solution to the equation $(A+B)^{-1}=A^{-1}+B^{-1}$ for 3×3 matrices. Our theorem implies that this is not possible if real entries are required. Clearly, we must find a 3×3 matrix P with minimum polynomial x^2+x+1. Most discussions [1] of the minimum polynomial show that it must have as factors every linear factor of the characteristic polynomial of the matrix. This means that the characteristic polynomial of our desired matrix P over the complexes must have complex coefficients and so some entries in P must not be real. For completeness we note that a diagonal matrix P exists with the required properties. The diagonal entries are and include the two roots of x^2+x+1.

The author expresses sincere appreciation to the referee for his suggestions on strengthening the theorem and on its further uses as shown above.

Reference

[1] Franz E. Hohn, Elementary Matrix Algebra, 3rd ed., Macmillan, New York, 1973, p. 414.

Circular Coordinates and Computer Drawn Designs

ELLIOT A. TANIS
LEE KUIVINEN
Hope College
Holland, MI 49423

An interesting system of coordinates called circular coordinates was introduced in *The Mathematics Teacher* in 1971 [1]. A definition of this system, given in that article, follows: "It consists of two perpendicular axes, called, say U and V. Each point on the axes is the center of a circle that passes through the origin. The point (u,v) is defined to be the intersection of the circle whose center is at u on the horizontal axis with the circle whose center is at v on the vertical axis."

Given a function f, it is interesting to compare the graph of $y=f(x)$ in the rectangular (Cartesian) system with the graph of $v=f(u)$ in the circular system. Furthermore, interesting artistic designs can be created using circular graphs of fairly simple functions.

In order to depict the circular graph of $v=f(u)$, it is convenient to define s and t parametrically as functions of u and plot the set of ordered pairs (s,t). For a given point $(u,f(u))=(u,v)$, the corresponding point (s,t) is the intersection of the circles

$$(s-u)^2+t^2=u^2 \tag{1}$$

and

$$s^2+(t-v)^2=v^2. \tag{2}$$

Solving these two equations simultaneously for s and t yields

$$s=g(u)=\frac{2uv^2}{u^2+v^2}=\frac{2u[f(u)]^2}{u^2+[f(u)]^2} \tag{3}$$

and

$$t=h(u)=\frac{2u^2v}{u^2+v^2}=\frac{2u^2f(u)}{u^2+[f(u)]^2}. \tag{4}$$

Thus to depict the circular graph of $v=f(u)$, we use equations 3 and 4 to define $s=g(u)$ and $t=h(u)$, respectively. The set of ordered pairs (s,t) is then plotted.

It was pointed out by Trask [1] that intercepts $(a,0)$ or $(0,b)$ in the rectangular system will be at the origin in the circular system. Furthermore, for corresponding values of the independent variable, either x or u, the graph will lie in the same quadrant in both systems. The following two propositions relate asymptotes in the rectangular system to intercepts in the circular system.

PROPOSITION 1. *If the graph of $y=f(x)$ has a vertical asymptote at $x=a$ in the rectangular system, then the graph of $v=f(u)$ has an intercept at $u=2a$ in the circular system.*

Proof. Suppose that the function f is such that $\lim_{w\to a} f(w)=\infty$ or $\lim_{w\to a} f(w)=-\infty$, where w may approach a from the right or the left. In the rectangular system the graph of $y=f(x)$ would have a vertical asymptote at $x=a$. To determine the characteristics of the graph of $v=f(u)$ near $u=a$ in the circular system, we consider the limits of $s=g(u)$ and $t=h(u)$ as u approaches a. From equation 3 we have

$$\lim_{u\to a} g(u)=\lim_{u\to a}\frac{2u[f(u)]^2}{u^2+[f(u)]^2}=\lim_{u\to a}\frac{2a}{u^2/[f(u)]^2+1}=2a$$

and from equation 4 we have

$$\lim_{u\to a} h(u) = \lim_{u\to a} \frac{2u^2 f(u)}{u^2+[f(u)]^2} = \lim_{u\to a} \frac{2u^2}{u^2/f(u)+f(u)} = 0.$$

Thus we see that, for $v=f(u)$, as u approaches a, the circular graph of $v=f(u)$ approaches $(2a,0)$.

PROPOSITION 2. *If the graph of $y=f(x)$ has a horizontal asymptote at $y=b$ in the rectangular system, then the graph of $v=f(u)$ has a limiting intercept at $v=2b$ in the circular system.*

Proof. Suppose that $\lim_{w\to\infty} f(w)=b$. In the rectangular system the graph of $y=f(x)$ would have a horizontal asymptote at $y=b$. Considering limits of $s=g(u)$ and $t=h(u)$ as u increases without bound, we have from equation 3

$$\lim_{u\to\infty} g(u) = \lim_{u\to\infty} \frac{2[f(u)]^2/u}{1+[f(u)]^2/u^2} = 0$$

and from equation 4

$$\lim_{u\to\infty} h(u) = \lim_{u\to\infty} \frac{2f(u)}{1+[f(u)]^2/u^2} = 2b.$$

Thus the circular graph of $v=f(u)$, as u increases without bound, approaches $(0,2b)$.

A similar argument holds if we take $\lim_{w\to-\infty} f(w)=b$.

To illustrate these relationships between intercepts and asymptotes for the rectangular graph of $y=f(x)$ with the origin and intercepts for the circular graph of $v=f(u)$, respectively, we give some examples.

Example 1. Consider the rational function

$$f(w) = \frac{2(w+4)(w+1)(w-3)}{(w-2)(w+2)^2}.$$

Graphs of $y=f(x)$ for $-6\leqslant x\leqslant 6$ and $v=f(u)$ for $-100\leqslant u\leqslant 100$ are given in FIGURES 1 and 2, respectively.

In these figures the points $(-4,0)$, $(-1,0)$, $(3,0)$, and $(0,3)$ in FIGURE 1 correspond to the point $(0,0)$ in FIGURE 2. The vertical asymptotes at $x=-2$ and $x=2$ in FIGURE 1 correspond to the intercepts $(-4,0)$ and $(4,0)$ in FIGURE 2. Note the effect of squaring $(w+2)$. The horizontal asymptote at $y=2$ in FIGURE 1 corresponds to the point $(0,4)$ in FIGURE 2. Furthermore for

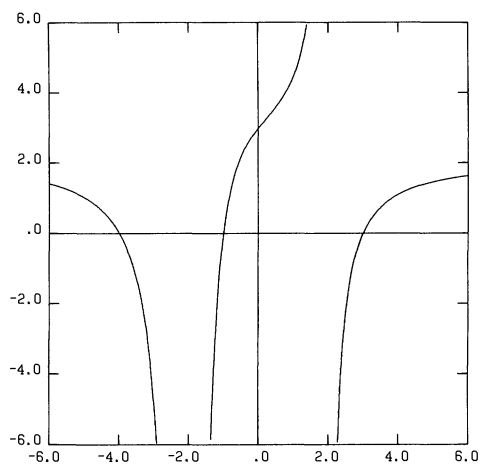


FIGURE 1.

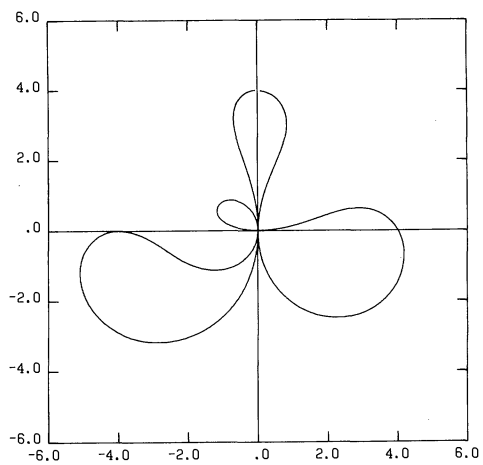


FIGURE 2.

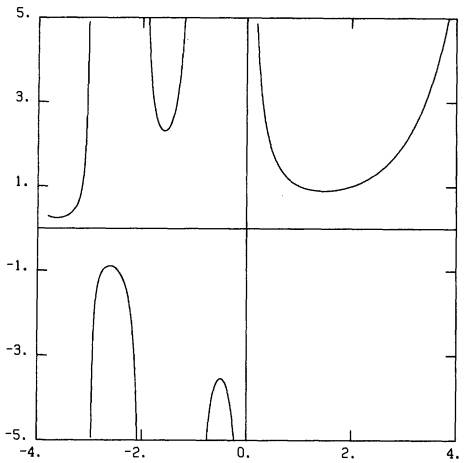


FIGURE 3.

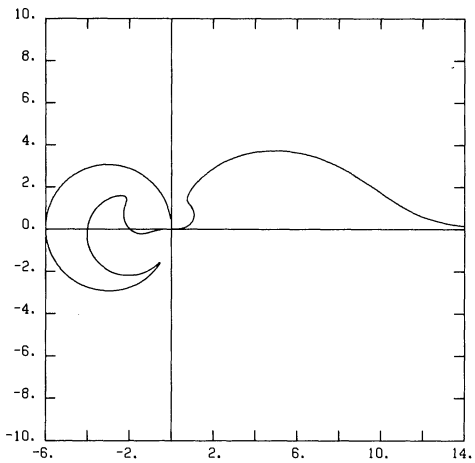


FIGURE 4.

those values of x that cause the graph of $y=f(x)$ to lie in a particular quadrant, the same values of u cause the graph of $v=f(u)$ to lie in the corresponding quadrant.

Example 2. An interesting function which has vertical asymptotes at each non-positive integer is the gamma function defined by $f(w)=\int_0^\infty t^{w-1}e^{-t}\,dt$. Graphs of $y=f(x)$, $-3.8\leqslant x\leqslant 4$, and $v=f(u)$, $-3.8\leqslant u\leqslant 14$, are given in FIGURES 3 and 4, respectively.

The gamma function can be used for an interesting artistic design. This is described in the next example.

Example 3. In FIGURE 5 we have plotted a family of curves $v=f(u)+c$, $-2.999\leqslant u\leqslant 9$. The function f is the gamma function defined in Example 2. The constant c takes on integer values from -10 to 10, inclusive.

The graph of a linear function $v=m\cdot u+b$ was discussed by Trask [1]. In the next example we show how the graph of this simple function can be used for artistic purposes.

Example 4. Consider the family of linear functions $v=f(u)=u+c$, $-10\leqslant u\leqslant 10$, where c takes on integer values from -10 to 10 inclusive. This family of curves is plotted in FIGURE 6 using the circular coordinate system.

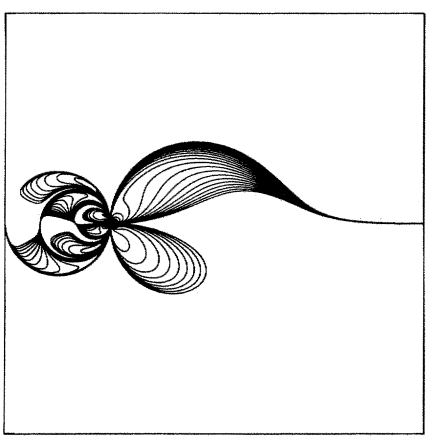


FIGURE 5.

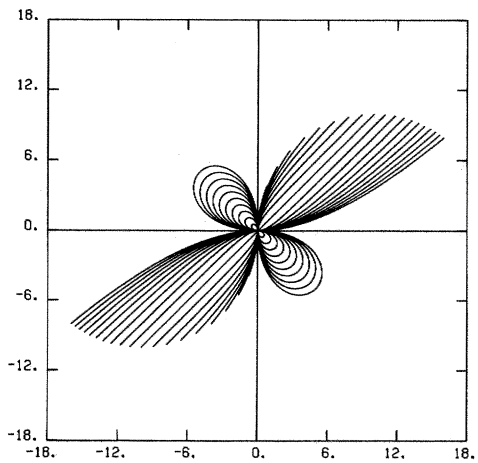


FIGURE 6.

It is possible to use the family of linear functions defined in Example 4 along with the family of linear functions $v=g(u)=-u+c$, $-10\leqslant u\leqslant 10$, where c takes on integer values from -10 to 10, inclusive, for artistic purposes. The reader is encouraged to use translations of these two families of linear functions for making their own designs.

This paper grew out of a senior independent study project. The project was one in which the student involved was able to generate the enthusiasm and interest of several other students in what he was doing. The results of such a project are esthetically pleasing. In addition a project of planning computer drawn designs provides an effective vehicle for a student to gain some understanding of the graphs of mathematical functions.

Reference

[1] F. K. Trask, III, Circular Coordinates: A Strange New System of Coordinates, Math. Teacher, LXIV (1971) 402–408.

Convergence and Divergence of $\Sigma_{n=1}^{\infty}1/n^p$

TERESA COHEN
Pennsylvania State University
University Park, PA 16802

WILLIAM J. KNIGHT
Indiana University at South Bend
South Bend, IN 46615

There is a little known proof of the divergence of the harmonic series that goes as follows. Suppose $1/1+1/2+1/3+1/4+\cdots$ converges to a number S. Then the even numbered terms clearly converge to $\frac{1}{2}S$. But this means that the odd numbered terms must converge to the other half of S, which is impossible because

$$\tfrac{1}{1}>\tfrac{1}{2},\ \tfrac{1}{3}>\tfrac{1}{4},\ \tfrac{1}{5}>\tfrac{1}{6},\ldots.$$

Thus the series must diverge.

This proof is so simple that it can be used with high school students. The purpose of this note is to show how the idea of the proof can be modified to establish the convergence of $\Sigma_{n=1}^{\infty}1/n^p$ when $p>1$.

Write S_N for the Nth partial sum of the series. Then

$$\begin{aligned}S_{2N+1}&=1+\left[\frac{1}{2^p}+\frac{1}{4^p}+\cdots+\frac{1}{(2N)^p}\right]+\left[\frac{1}{3^p}+\frac{1}{5^p}+\cdots+\frac{1}{(2N+1)^p}\right]\\&<1+\left[\frac{1}{2^p}+\frac{1}{4^p}+\cdots+\frac{1}{(2N)^p}\right]+\left[\frac{1}{2^p}+\frac{1}{4^p}+\cdots+\frac{1}{(2N)^p}\right]\\&=1+\frac{1}{2^p}S_N+\frac{1}{2^p}S_N<1+2^{1-p}S_{2N+1}\end{aligned}$$

because $S_N<S_{2N+1}$. Thus $(1-2^{1-p})S_{2N+1}<1$. Since $p>1$ the factor $1-2^{1-p}$ is positive, and so we have $S_{2N+1}<(1-2^{1-p})^{-1}$ for all N. Since $S_{2N}<S_{2N+1}$ we see that the increasing sequence $\{S_N\}$ is bounded above by $(1-2^{1-p})^{-1}$. Hence it converges.

Finally, note that when $p<1$ the series $\Sigma 1/n^p$ diverges by comparison with the harmonic series.

PROBLEMS

Dan Eustice, Editor
Leroy F. Meyers, Associate Editor
The Ohio State University

Proposals

To be considered for publication, solutions should be mailed before December 1, 1979.

1072. The professor is preparing her final exam for calculus. She wants to include the problem: "Find the relative maxima, relative minima, and points of inflection of the following function." The function should be a polynomial $P(x)$ of degree 4 with three distinct relative extrema and two distinct points of inflection. In order to solve the problem, the students must be able to factor $P'(x)$ and $P''(x)$. But the typical calculus student in her class can factor a quadratic polynomial correctly only if its roots are integers between -20 and 20, and the student can factor a cubic polynomial only if it is x times a quadratic which the student can factor. Help the professor find such a polynomial. [*Peter Ørno, The Ohio State University.*]

1073. Let A and B be the unique nondecreasing sequences of odd integers and even integers, respectively, such that for all $n \geqslant 1$, the number of integers i satisfying $A_i = 2n-1$ is A_n, and the number of integers i satisfying $B_i = 2n$ is B_n. That is, $A = (1,3,3,3,5,5,5,7,7,7,9,9,9,9,9,\ldots)$ and $B = (2,2,4,4,6,6,6,6,8,8,8,8,\ldots)$. Is the difference $|A_n - B_n|$ bounded? [*James Propp, student, Harvard College.*]

Quickies

Solutions to Quickies appear at the conclusion of the Problems section.

Q660. Find the ratio of the area of an ellipse to the area of the largest inscribed rectangle. [*Alan Wayne, Pasco-Hernando Community College.*]

Q661. Our old friend Professor Umbugio wishes to add up the integral parts of the numbers $\ln n$ for the first 10^9 *positive* integers n. He has appealed to us for help since he does not have access to a sufficiently fast computer. Show the professor how to evaluate his sum without using a high-speed computer at all. [*J. Phipps McGrath, Wellfleet, Massachusetts.*]

Solutions

Concurrent Perpendiculars — November 1977

1028. Let ABC be a triangle and P_1, P_2, and P_3 be arbitrary points in the plane of ABC. Let arbitrary lines perpendicular to AP_i, BP_i, and CP_i determine triangles $A_iB_iC_i$ for $i=1,2,3$. Now, let A_0, B_0, and C_0 be the respective centroids of triangles $A_1A_2A_3$, $B_1B_2B_3$, and $C_1C_2C_3$. Show that the perpendiculars from A, B, and C on the sides of triangle $A_0B_0C_0$ concur. [*Leon Gerber, St. John's University.*]

Solution: Let us denote by T the known theorem: *If two triangles are such that the perpendiculars from the vertices of one on the sides of the other are concurrent, then the perpendiculars from the vertices of the second on the sides of the first are also concurrent.* (This theorem is easily established, *e.g.*, by the trigonometric form of Ceva's theorem.)

By T, for each $i=1,2,3$, the perpendiculars from A_i, B_i, C_i on the sides of triangle ABC are concurrent in a point Q_i. Let A_m, B_m, C_m be the respective midpoints of A_1A_2, B_1B_2, C_1C_2. Then the perpendiculars from A_m, B_m, C_m on the sides of triangle ABC are concurrent at Q_m, the midpoint of Q_1Q_2. Now A_0, B_0, C_0 are the points one third the way from A_m to A_3, B_m to B_3, and C_m to C_3, respectively. It follows that the perpendiculars from A_0, B_0, C_0 on the sides of triangle ABC are concurrent at the point Q_0 one third the way from Q_m to Q_3. Finally, then, by T, the perpendiculars from A, B, C on the sides of triangle $A_0B_0C_0$ concur.

HOWARD EVES
University of Maine

Also solved by H. R. van der Vaart and the proposer.

Erdős and the Computer — January 1978

1029.* Does there exist any prime number such that if any digit (in base 10) is changed to any other digit, the resulting number is always composite? [*Murray S. Klamkin, University of Alberta.*]

Solution: We prove a slightly stronger result and in the end make some comments and state a few more problems.

For every $k>k_0$ there are primes

$$p=\sum_{i=0}^{k} a_i 10^i, a_0>0, a_k>0, 0\leqslant a_i\leqslant 9,$$

so that all the integers

$$p+t\cdot 10^i, |t|\leqslant 10, 0\leqslant i\leqslant k \tag{1}$$

are composite. In fact, there are $21(k+1)$ integers of the form (1). Put $x=10^{k+1}$. We will determine $p \bmod q_i$ where the q's will be suitably chosen primes whose product is less than x^ε where $\varepsilon>0$ is small but fixed. Then by Linnik's theorem (the smallest $p=a \pmod b$ is less than b^c for an absolute constant c if $(a,b)=1$) there is a $p<x$ which satisfies all these congruences. The congruences will be chosen so that all the numbers (1) will be multiples of one of the q's; thus they are all composite.

By a well-known theorem of Bang-Birkhoff-Vandiver, there always is a q_j so that $10^j \equiv 1 \pmod{q_j}$ and $10^i \not\equiv 1 \pmod{q_j}$ $1 \leqslant i < j$. (The theorem states: for every a and j [except $2^6 - 1$] there is a q_j so that $a^j - 1 \equiv 0 \pmod{q_j}$ and $a^i - 1 \not\equiv 0 \pmod{q_j}$ for every $1 \leqslant i < j$.) Consider these primes so that

$$\prod_{j=1}^{r} q_j \leqslant x^{\varepsilon/2} < \prod_{j=1}^{r+1} q_j. \tag{2}$$

since $\prod_{j=1}^{r} q_j < 10^{r^2}$, r can be chosen as $[\varepsilon\sqrt{\log x}\,]$. Now we determine the congruences. Suppose that the congruences

$$p \equiv u_m \pmod{q_m}, 1 \leqslant m \leqslant j-1 \tag{3}$$

have already been determined. Let $b_1, \ldots, b_{s_{j-1}}$ be the integers of the form $t \cdot 10^i, |t| \leqslant 10, 0 \leqslant i \leqslant k$ which do not satisfy any of the congruences

$$t \cdot 10^i \equiv -u_m \pmod{q_m}, 1 \leqslant m \leqslant j-1. \tag{3'}$$

The numbers $t \cdot 10^i$ determine at most $21j$ residues mod q_j (since 10^i takes exactly j distinct values by $10^j \equiv 1 \pmod{q_j}$). Therefore there is a u_j for which $b_L \equiv -u_j \pmod{q_j}$, $1 \leqslant L \leqslant j$ is satisfied by at least $\{b_{s_{j-1}}/21j\}$ values of L where $\{N\}$ denotes the least integer $\geqslant N$. Put $p \equiv u_j \pmod{q_j}$. This determines the congruences

$$p \equiv u_j \pmod{q_j}, \quad 1 \leqslant j \leqslant r = \left[\varepsilon\sqrt{\log x}\,\right]. \tag{4}$$

The number of integers $t \cdot 10^j, |t| \leqslant 10, 1 \leqslant j \leqslant k$ for which $p + t \cdot 10^j$ (p satisfying the congruences (4)) is not a multiple of one of the $q_j, 1 \leqslant j \leqslant r$ is at most

$$21(k+1) \prod_{j=2}^{r} \left(1 - \frac{1}{j}\right) < \frac{21 \log x}{r} < \frac{21\sqrt{\log x}}{\varepsilon}.$$

Let $v_1, v_2, \ldots, v_x, x \leqslant \dfrac{21 \log x}{\varepsilon}$ be these integers of the form $t \cdot 10^j$. Let $Q_1, Q_2, \ldots, Q_x$ be the consecutive primes which are not q's. Put

$$p \equiv -v_i \pmod{Q_i}, i = 1, \ldots, x. \tag{4'}$$

There are $r + x$ congruences (4) and (4'). The product of the moduli equals

$$\prod_{j=1}^{r} q_j \prod_{i=1}^{x} Q_i < x^{\varepsilon/2} x^{\varepsilon/2} = x^{\varepsilon}.$$

($\Pi Q_i < x^{\varepsilon/2}$ is trivial from the prime number theorem or a much more elementary result.) This completes our proof since the primes p satisfying (4) and (4') satisfy (1) and as stated by Linnik there are p's $< x$.

Denote by $l_a(p)$ the exponent of a mod p.

I can prove

$$\sum_{p<x} \frac{1}{l_a(p)} > x^c \tag{5}$$

and in (5) probably c can be taken to be $1 - \varepsilon$, but this seems very difficult.

From (5) we can deduce by the methods used here that there are infinitely many primes $p, 10^k < p < 10^{k+1}$, so that if we simultaneously alter $(\log k)^e$ digits we always get a composite number.

Is it true that if $a_m > c^m/m, c > 1, m = 1, 2, \ldots$ then there always are primes, in fact infinitely many of them, so that all the numbers $p + a_m$, $a_m < p$ are composite? I do not know.

PAUL ERDŐS
Hungarian Academy of Science

Editor's Comment. Individual computer searches by Allan Wm. Johnson, Harry L. Nelson, and Stanley Rabinowitz found six-digit primes which provide a solution. The complete list of such six-digit primes supplied by Nelson is: 294 001, 505 447, 584 141, 604 171, and 971 767. Rabinowitz supplied the table below which shows a divisor for each number that can be formed from 294 001 by changing one digit.

Digit Changed	Replacement Digit									
	0	1	2	3	4	5	6	7	8	9
100000	23	3	1	47	3	73	7	3	587	239
10000	7	173	3	29	17	3	227	7	3	1
1000	3	397	29	3	1	7	3	43	11	3
100	1	19	3	7	83	3	151	11	3	29
10	1	41	3	29	11	3	157	409	3	7
1	2	1	2	3	2	5	2	7	2	3

Signs and cosines **March 1978**

1033. For given positive integers $n_1, n_2, \ldots, n_k$, when is

$$\int_0^{2\pi} \cos n_1\theta \cos n_2\theta \cdots \cos n_k\theta d\theta$$

different from zero and what is its value? [*H. Kestelman, University College, London.*]

Solution: The problem is a generalization of the orthogonality property for $k=2$, namely

$$\int_0^{2\pi} \cos n_1\theta \cos n_2\theta d\theta = \frac{1}{2}\int_0^{2\pi} [\cos(n_1+n_2)\theta + \cos(n_1-n_2)\theta]d\theta = \pi\delta_{n_1n_2},$$

where $\delta_{n_1n_2}$ is the Kronecker delta. It can be attacked in the same way, and broken down into a sum of terms of the form

$$\frac{1}{2^{k-1}}\int_0^{2\pi} \cos(n_1 \pm n_2 \pm n_3 \pm \cdots \pm n_k)\theta d\theta,$$

where the signs on $n_2, n_3, \ldots, n_k$ go through all possible permutations. The value of this sum will depend on how many of the integers $n_u \pm n_2 \pm n_3 \pm \cdots \pm n_k$ are zero. The following can also be shown: (No proof supplied. See reference below—Ed.)

If $k=2l$, i.e., k is even, then the maximum number of zeros occurs when the n_i's are all equal and exactly l of the signs are negative. The total number of such terms is $\binom{2l-1}{l}$, since n_1 is always positive. If $k=2l+1$, then the maximum number of zeros occurs when $n_2, n_3, \ldots, n_k$ are all equal (to n, say) and $l+1$ of the signs are negative, with $n_1 = -2n$. The total number of such terms is $\binom{2l}{l+1}$. It is possible to choose the n_i's in such a way that there are any number of zero coefficients up to this maximum value. Hence the answer to the problem is that the given integral is zero if none of the coefficients $n_1 \pm n_2 \pm n_3 \pm \cdots \pm n_k$ is zero, and otherwise it can take on any of the values

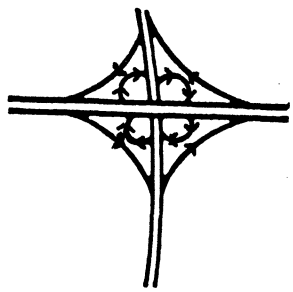
$$\frac{\pi}{2^{k-2}}, 2\frac{\pi}{2^{k-2}}, 3\frac{\pi}{2^{k-2}}, \ldots, \binom{k-1}{k-\left[\frac{k}{2}\right]}\frac{\pi}{2^{k-2}},$$

depending on the number of zero coefficients.

ZANE C. MOTTELER
Michigan Technological University

Also solved by B. D. Aggarwala & C. Nasim (Canada), Anders Bager (Denmark), Richard Beigel, Michael Ecker, Thomas Elsner, Daniel S. Freed, Michael Gilpin, Eli L. Isaacson, M. S. Klamkin (Canada), Peter W. Lindstrom, Viktors Linis (Canada), Graham Lord (Canada), Edwin P. McCravy, J. M. Metzger, William Myers, Boon-Yian Ng (Malaysia), J. Pfaendtner (Germany), J. M. Stark, Michael Vowe (Switzerland), and the proposer. M. S. Klamkin & A. Liu have a generalization of this problem which they have submitted for publication.

Cloverleaves **March 1978**



1034. We are familiar with the standard clover-leaf interchange [CLI] which has, inside the four ramps for making right-hand turns, the arrangement whereby left-hand turns are achieved by turning right into lanes which outline the four leaf clover. Your car approaches the CLI from the south. A mechanism has been installed so that at each point where there exists a choice of directions, the car turns to the right with fixed probability r.

a. If $r=1/2$, what is your chance of emerging from the CLI going west?

b. Find the value of r which maximizes your chance of westward departure.

[*Marlow Sholander, Case Western Reserve University.*]

Solution: In order to emerge from the CLI going west, the car must go straight at the first point of decision, then make $4n+1$ right turns in the CLI, and finally go straight a second time to leave the CLI. The probability, $P(r)$, of this occurring is:

$$P(r)=\sum_{n=0}^{\infty}(1-r)^2r^{4n+1}=\frac{r(1-r)^2}{1-r^4}=\frac{1}{1+r^2}-\frac{1}{1+r},$$

if $0\leqslant r<1$, but $P(1)=0$. So $P(1/2)=2/15$.

In order to maximize $P(r)$, I will solve $P'(r)=0$. Now $P'(r)=-2r/(1+r^2)^2+1/(1+r)^2$. Hence $P'(r)=0$ if and only if $(r+1/r)^2-2(r+1/r)-4=0$, or $r+1/r=1+\sqrt{5}$ (the negative root is rejected), or $r=(1+\sqrt{5})/2-\sqrt{(1+\sqrt{5})/2}$ (the other root is greater than 1).

Call this number R. R is approximately 0.346. Since $P(r)$ is continuous, it must have a maximum value either at an endpoint of its domain or at a point where $P'(r)=0$. R is the unique root of $P'(r)=0$ on the interval $(0,1)$. $P(0)=P(1)=0$ and $P(R)>0$; therefore, $P(r)$ has its maximum value at R. This is approximately 0.15.

RICHARD BEIGEL
Wilton, Connecticut

Comment. A related problem is perhaps even more interesting. Consider a pure cloverleaf interchange (PCLI), i.e., one in which there are no ramps for right-hand turns but only the two intersecting straight highways with cloverleaves for left-hand turns. Now, we can pose the same question as for the CLI problem. The probabilities of emerging north, west, south, and east are $f_i(r)=r^i/(1+r)(1+r^2)$ when i is 0, 1, 2, 3, respectively, if $r\neq 1$, but 0 if $r=1$. Now,

$$\lim_{r\to 1^-} f_i(r)=.25,$$

and there is no value of r which maximizes the probability of heading east or south since f_i is strictly increasing on $[0,1)$ if $i=2$ or 3. Intuitively, we see that as r gets larger (close to 1), then the car tends to stay in the intersection taking right turns until an event with very small probability $(1-r)$ occurs. This will tend to make all directions equally likely.

This anomaly did not appear in the CLI case due to the presence of the right turn ramps. It is interesting to note that in the early days of modern highway design, pure cloverleaf interchanges

were actually built. This was especially true in California in the late 50's. However, a noted Polish probabilist, J. R. Clopenski, was involved in an incident which convinced the state of California to cease building PCLI's. In 1958 Professor Clopenski spent his sabbatical leave at UCLA. During his stay he acquired a number of lady friends in the Southern California area and he would visit one of them every night on his way home from the University. It just so happened that there was a central freeway interchange which Clopenski approached from the south and he had one friend located west of the interchange, one located north, and one located east. This interchange was a PCLI and being a sporting man, he would choose r each night which would be his probability of making a right turn whenever a choice existed. Toward the end of Clopenski's sabbatical, the friend to the west had evolved as his favorite and while Clopenski did not wish to change his strategy, he did wish to maximize the probability of heading west out of the interchange. Clopenski made the necessary calculations and determined that $r \approx .657$ maximized his chances of heading west. (If Clopenski ever came out of the interchange headed south, then he returned to the University to prepare his lesson plans.) Well, one evening as he reached the interchange, Clopenski realized that he had forgotten his paycheck back at his office. Rather than simply turn around and head back to UCLA, he decided to proceed through the interchange as before and choose r to maximize his probability of heading south. Unfortunately, while Clopenski was a prodigious mental calculator he was not too concerned with limit points. He had correctly calculated the formula $r^2/(1+r)(1+r^2)$ and then determined that this expression was maximum at $r=1$.

In his memory, the California Department of Transportation agreed to cease building pure cloverleaf interchanges.

STANLEY J. BENKOSKI
Paoli, Pennsylvania

Also solved by John M. Atkins, S. F. Barger, Beloit College Solvers, Stanley J. Benkoski, Jordi Dou (Spain), Michael W. Ecker, Milton P. Eisner, Nick Franceschine III, Daniel S. Freed, W. W. Funkenbusch, Furman University Problem Group, P. K. Garlick, John A. Gillespie, Clifford H. Gordon, Lee O. Hagglund, George C. Harrison, Dale T. Hoffman, Eli L. Isaacson, Ralph Jones, Erwin Kronheimer (England), J. P. Lambert, Peter W. Lindstrom, Edwin P. McCravy, James P. McGill, Kirby M. McMaster, J. M. Metzger, Carl Moore, Susan Jane Morris, Zane C. Motteler, William Myers, Roger B. Nelson, Boon-Yian Ng (Malaysia), J. Pfaendtner (Germany), Howard W. Pullman, Stanley Rabinowitz, Daniel Mark Rosenblum, Seattle Pacific University Problems Seminar, James T. Smith, Scott Smith, Richard Stark, Daniel L. Stock, Roger G. Tishler, James D. Watson, Ann E. Watkins, Robert C. Williams, Harald Ziehms (Germany), and the proposer.

Answers

Solutions to the Quickies which appear near the beginning of the Problems section.

Q660. Since there is an affine map preserving ratio of areas and taking the ellipse with largest inscribed rectangle into a circle with inscribed square, the ratio must be $\pi/2$.

Q661. Consider the rectangle $R: 1 \leqslant x \leqslant 10^9$, $1 \leqslant y \leqslant 20$, which contains 2×10^{10} lattice points. The required sum is just the number of lattice points in R and below the curve $y = \ln x$. The number of lattice points above this curve and in R is $\sum_{k=1}^{20} [e^k]$ (add them up horizontally instead of vertically). Hence the required sum is $2 \times 10^{10} - \sum_{k=1}^{20} [e^k]$, which is easily found, from a good table of the exponential function or by pocket calculator, to be 1 92324 79973.

REVIEWS

PAUL J. CAMPBELL, Editor
Beloit College

PIERRE MALRAISON, Editor
Control Data Corp.

Assistant Editor: Eric S. Rosenthal, Princeton University. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of the mathematics literature. Some reviews of books are adapted from the Telegraphic Reviews in the American Mathematical Monthly.

Stockmeyer, Larry J. and Chandra, Ashok K., *Intrinsically hard problems*, Scientific American 240 (May 1979) 140-159.

A discussion of efficiency of algorithms, and of problems whose most efficient solution still requires astronomically long times. It is a little odd to read an article on this subject which doesn't have more of a discussion of NP-complete problems.

Kolata, Gina Bari, *Continuation methods: New ways to solve equations*, Science 204 (4 May 1979) 488-489.

Applications of topological methods to the problem of finding solutions to sets of linear equations. Once again, applied topology!

Morris, S.J., *Mod art: The art of mathematics*, Technology Review 81 (March-April 1979) 47-51.

Using colored tiles to represent the integers mod n, addition and multiplication tables yield "mod art" patterns.

Miller, Julie Ann, *Food web*, Science News 115 (14 April 1979) 250-251.

A report on the work of biologist Joel Cohen on using food webs to analyze niche spaces. He has shown that most food webs are in fact interval graphs, i.e., can be represented by intervals on the real line, one for each node, with edges corresponding to non-empty intersection. In biological terms, this implies that niche space is basically one-dimensional.

Kolata, Gina Bari, *Gian-Carlo Rota and combinatorial math*, Science 204 (6 April 1979) 44-45.

A look at applications and trends of combinatorial mathematics by one of the Grand Old Men (at age 47) of the field.

Kolata, Gina Bari, *Frederick Mosteller and applied statistics*, Science 204 (27 April 1979) 397-398.

A look at one of the premier statisticians and his work, which ranges from proving who wrote the Federalist papers to exploratory data analysis on cancer in rats.

Gardner, M., *Mathematical games: In which players of tick-tack-toe are taught to hunt bigger game*, Scientific American 240 (April 1979)

18-28.

In tick-tack-toe, the goal is to get three points in a row, i.e., color a triomino with "your" color. Harary tick-tack-toe (named after Frank Harary) is the game where the object is to color a polyomino or "animal" with your color. Gardner discusses which animals admit a winning strategy for the first player and on what size boards.

Begle, E.C., Critical Variables in Mathematical Education, NCTM/MAA, 1979; xxvii + 165 pp, $8.

Ed Begle was director of SMSG during the late fifties and early sixties; he was always a strong advocate of empirical research in mathematical education. This survey, published posthumously, summarizes his research on existing empirical studies, over factors ranging from physical variables in the classroom (high temperatures decrease ability to do summations), to the effects of different kinds of tests and the influence of basic curriculum decisions made at the outset. An essential reference source for the researcher in mathematical education and recommended reading for all teachers.

Chace, Arnold Buffum, The Rhind Mathematical Papyrus, NCTM, 1979; xi + 147 pp, $15.

Abridged reprint of 1927-1929 two-volume edition. The lengthy bibliographies have been omitted in favor of a list of more recent (and more available) sources, and only a representative selection--about one-fourth--of the plates has been reproduced. The result is a splendid scholarly treasure at a popular price.

Fairley, William B. and Mosteller, Frederick (eds), Statistics and Public Policy, Addison-Wesley, 1977; xiii + 397 pp, $14.95.

This collection of essays offers statistical analyses of intriguing issues in public policy. Should we seed hurricanes? Does air pollution shorten lives? Does vehicle inspection reduce auto accident mortality? The analyses are largely equation-free, and the result is a book that is mostly readable without knowledge of algebra or statistics. The analyses lean toward "exploratory data analysis" espoused by Tukey in his book of that title (1977).

Garey, Michael R. and Johnson, David S., Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, 1979; x + 338 pp, $10 (P); $18.50.

A wide variety of commonly encountered problems in mathematics, computer science, and operations research has been shown in recent years to be *NP-complete*--no known algorithm can solve such a problem in an amount of time that is a polynomial function of the "size" of the problem. Examples include the travelling salesman problem, graph coloring problems, job-shop scheduling, integer programming, determination of natural winner in generalized Hex, and many more. In addition to a definitive treatment of the general theory of NP-complete problems and their classification, this book includes an extensive (100 page) list of problems known to be NP-complete and an exhaustive (35 page) list of references.

Kovalevskaya, S., A Russian Childhood, transl: Beatrice Stillman, Springer-Verlag, 1978; xii + 250 pp, $14.80.

This book includes a short autobiographical sketch and an article on Kovalevskaya's mathematics by P.Y. Kochina, but consists primarily of the author's memoirs of her first fifteen years. Not many mathematicians are known for literary talents--Kovalevskaya was not only a good enough mathematician to win the Prix Borodin, but an able writer, as demonstrated by this enjoyable book.

Kimble, Gregory A., How to Use (and Misuse) Statistics, Prentice-Hall, 1978; xi + 290 pp, $14.95; $7.95 (P).

Attractive presentation of basic ideas of statistics *sans* formulas. The book is dedicated "To LLK and HP-65"!

Kline, Morris (ed), Mathematics: An Introduction to Its Spirit and Use, Freeman, 1979; viii + 249 pp, $14; $7.95 (P).

Of the 40 selections, 14 appeared in Kline's earlier compendium *Mathematics in the Modern World* (1968). A new feature is the inclusion of some 13 of Martin Gardner's columns. Still, apart from one of the latter, everything here is from 1971 or (usually much) earlier.

Oden, Teresa and Thompson, Christine, Computers and Public Policy: Proceedings of the Symposium on Man and the Computer, Kiewit Computation Center, Dartmouth College; v + 78 pp, (P).

Enlightening essays on the present status of computing, presented at the celebration of the 10th anniversary of the Kiewit Computation Center.

Sharron, Sidney and Reys, Robert (eds), Applications in School Mathematics, 1979 Yearbook, NCTM, 1979; viii + 247 pp, $12.

Although some of the contributions are not relevant at the college level, others are, such as "Applications through direct quote word problems," "Some everyday applications of the theory of interest," "The mathematics of finance revisited through the hand calculator," "Developing classroom applications based on socioeconomic problems," and "Mathematical modeling and cool buttermilk in the summer." Also included is a valuable bibliography.

Williams, Bill, A Sampler on Sampling, Wiley, 1978; xv + 254 pp, $15.95.

An attempt to convey to the nonmathematically-inclined an understanding of the principles of statistical sampling. The author proceeds to generalizations from numerical examples, and no prior knowledge of statistics is needed. Suitable as a supplementary text in an introductory statistics course, particularly since most texts for such a course by and large neglect sampling.

Wanamaker, John L., *Computers and scientific jury selection: A calculated risk*, J. Urban Law 55 (1978) 345-370.

Scientific jury selection methods employ sociological and psychological research, in conjunction with computer analysis, to determine what a prospective juror is likely to believe. The article explains the survey techniques used, but not the ensuing mathematics, and speculates on the fairness to opposing counsel and to the prospective jurors themselves.

Baker, Lillian F. and Schattschneider, Doris J., The Perceptive Eye: Art and Math, Allentown Art Museum, 1979; 69 pp, $9 (P).

"With a perceptive eye, mathematics can be used to analyze various works of art; with a deliberate hand, mathematics can be used to create art." This special monograph explains and catalogues an exhibit (at the Allentown Art Museum, April 1-June 7, 1979) which uses objects in the Museum's collection to illustrate two-dimensional geometric patterns: point, line, rotational, kaleidoscopic, translation and glide-reflection symmetry; periodic designs; ratio and proportion; perspective. Also available: supplementary notes (loose bound) on projects and bibliographies in art-math activities for school children (pre-school to college), and examples of computer-generated periodic ornamental design. A rich, imaginative scheme that could well serve as a model for similar cooperative ventures throughout the country.

Clark, Colin W., *Mathematical models in the economics of renewable resources*, SIAM Review 21 (January 1979) 81-99.

In the context of a fishery, Clark discusses three classes of models of renewable resource economics: profit-maximizing, competitive equilibrium, and cooperative equilibrium. A variety of complexities and practical implications are pursued, including the suggestion that simple models may be appropriate even when they seem to seriously misrepresent nature.

Schot, Steven H., *Jerk: The time rate of change of acceleration*, Amer. J. Physics 46:11 (November 1978) 1090-1094.

Jerk, the third derivative of position with regard to time, is analyzed into its planar vector components, with examples for pendulum and central force motions.

Sussmann, Héctor J. and Zahler, Raphael S., *Catastrophe theory as applied to the social and biological sciences: A critique*, Synthese 37 (1978) 117-216.

Detailed damnation of "applications" to date of catastrophe theory in the social and biological sciences.

Garey, M.R., Graham, R.L. and Johnson, D.S., *Performance guarantees for scheduling algorithms*, Operations Research 26:1 (Jan.-Feb. 1978) 3-21.

Investigates known multiprocessor algorithms to find schedules guaranteed to be "near-optimal."

Model Building Seminar, available from Charles Heuer, Math. Dept., Concordia College, Moorhead, MN 56560, 1978; 233 pp, $2.50 (P).

Collection of 14 presentations by lecturers from industry and higher education at MAA North Central Section summer seminar at Bemidji, MN, in 1977. Topics include solid waste management, computerized highway control, guard-scheduling at a prison, group insurance premium-setting, and pension planning.

Rork, Alvin E. and Murnighan, J. Keith, *Equilibrium behavior and repeated plays of the Prisoner's Dilemma*, J. Math. Psychology 17 (April 1978) 189-198.

Describes experiments to vary the nature of the Nash equilibria of the game.

Ghyka, Matila, The Geometry of Art and Life, Dover, 1977; xvii + 176 pp, $2.75 (P).

Corrected republication of 1946 original which grasps toward a universal harmony of art, architecture and nature through proportion, especially the Golden Ratio.

Lockwood, E.H. and Macmillan, R.H., Geometric Symmetry, Cambridge U Pr, 1978; x + 228 pp, $24.50.

Splendid and comprehensive presentation of both the descriptive facts (for the nonmathematical reader) and the mathematical structure of symmetries in the first three dimensions: frieze patterns, wallpaper patterns, color symmetry, and more. The layout is very attractive, and the book is printed in two colors in 8" x 12" format.

Vinik, A., Silvey, L. and Hughes, B. (eds), Mathematics and Humor, NCTM, 1978; vi + 58 pp, $4 (P).

Enough jokes to get through a semester. Some new, some hoary.

NEWS & LETTERS

INTERNATIONAL CONGRESS ON MATHEMATICAL EDUCATION: FIRST ANNOUNCEMENT

The first announcement concerning the 1980 International Congress on Mathematical Education is scheduled to be distributed early this spring. (The Congress will take place at the University of California, Berkeley, August 10-16, 1980.) If you wish to receive a copy of the first announcement, please make your request to: Office of Mathematical Sciences, National Research Council, 2101 Constitution Avenue, NW, Washington, D.C. 20418.

ERDÖS PROBLEMS

A collection of problems which Paul Erdös has assigned to others is to be assembled. Mathematicians interested in cooperating in this project may submit copies of the correspondence containing the problem and supplement it with the following information: date problem given, monetary value (if any), source of problem (from Erdös alone, with someone else, or relayed by Erdös), comments on notation or other matters useful for the general reader. Material should be sent to Pat Faudree, 2983 Elgin, Memphis, TN 38118.

OPERATIONS RESEARCH: MATHEMATICS AND MODELS

The American Mathematical Society, in conjunction with its eighty-third summer meeting in Duluth, Minnesota, will continue its short course series with a course entitled "Operations Research: Mathematics and Models." The one and one-half day course will be held on Sunday and Monday, August 19 and 20, 1979, in Bohannan Hall, Room 90, on the University of Minnesota, Duluth campus.

MAA SUMMER CONFERENCE

The Department of Mathematical and Computer Sciences at Michigan Technological University and the Michigan Section, MAA, are co-sponsoring a summer mathematics conference during July 16-20, 1979. Professor Harry Pollard will present five lectures on "Some Faces of Analysis," as will Professor Donald G. Saari on "Mathematics and the Social Sciences." There is no need to be a specialist in either field to benefit from the lectures. Informal seminars will also be organized by and for the participants on the topics of the lectures.

Further information can be obtained from Dr. Zane C. Motteler, Department of Mathematical and Computer Sciences, Michigan Technological University, Houghton, MI 49931 (Phone: 906-487-2068).

PRINCETON MATHEMATICIAN RECEIVES WATERMAN AWARD

A 33-year-old mathematician who introduced new geometrical methods that have revolutionized mathematics and solved important long-standing problems has been selected by the National Science Board to receive the fourth annual Alan T. Waterman Award.

Dr. William P. Thurston, professor of mathematics at Princeton University, was selected from among 87 nominees to receive the award which carries with it a medal and a grant of up to $50,000 a year for each of the three years for research or advanced study. He is the second Princeton mathematician to win the award. The first was Dr. Charles L. Fefferman, who was 28 when he was honored in 1976 when the initial award was made.

The award, authorized by the Congress in 1975, was presented at a dinner cere-

mony on May 17 at the National Academy of Sciences. Leaders of the scientific and educational communities, Members of the Congress, and other Washington officials concerned with scientific research attended the dinner ceremony.

Thurston was selected for the award "in recognition of his achievements in introducing revolutionary new geometrical methods in the theory of foliation, function theory, and topology." His work has already greatly influenced a number of fields of mathematics.

The tools developed by Thurston are among the most promising for understanding the complex mathematical problems involved in three-dimensional manifolds. His works have provided key ideas and techniques which settled a long-standing mathematical problem involving transformation of a three-dimensional sphere--one of topology's most puzzling questions.

Thurston has already influenced profoundly a number of fields of mathematics. His meteoric career has been highlighted by his introduction of revolutionary new geometrical methods in qualitative theory of differential equations, function theory, and topology.

Born October 30, 1946 in Washington, D.C., Thurston earned a bachelor of arts degree in 1967 at New College, Sarasota, Florida. Five years later he was awarded a doctor of philosophy degree in mathematics at the University of California at Berkeley. He did research at the Institute for Advanced Studies in Princeton, New Jersey from 1972 to 1973, and was named assistant professor of mathematics at the Massachusetts Institute of Technology in 1973.

In 1974, Thurston was awarded an Alfred P. Sloan Fellowship for research in mathematics. In that same year he was named professor of mathematics at Princeton. He has also received the American Mathematical Society's Oswald Veblen Prize in Geometry for his contributions in that field.

Dr. Richard C. Atkinson, Director of the National Science Foundation said in announcing the award: "When criteria were established for the Alan T. Waterman Award, one of the main ones was that the recipient show 'outstanding capability and exceptional promise for significant future achievement.' Those who established the criteria four years ago might well have had a person like Dr. Thurston in mind. His capability for exploring new paths in mathematics will contribute significantly to advances in science."

The award is named in honor of the first Director of the NSF. Dr. Waterman was appointed to that post by President Truman in 1951 and was reappointed by President Eisenhower in 1957. Widely respected as an administrator, he continued as NSF Director at the request of President Kennedy until his retirement in 1963. He died in 1967.

Other recipients of the Alan T. Waterman Award have been Dr. J. William Schopf, a University of California at Los Angeles (UCLA) paleontologist, in 1977; and Dr. Richard A. Muller, a physicist at the Lawrence Berkeley Laboratory of the University of California, in 1978.

SHORT AND TO THE POINT

Recently P.W. Harley in "A Countable Nowhere First Countable Hausdorff Space," *Canad. Math. Bull.* 16 (1973) 441-442 and R. Willmott in "Countable Yet Nowhere First Countable" (this *Magazine*, January 1979, pp. 26-27) have published examples of countable topological spaces which are nowhere first countable. I would like to call to your attention a much earlier example given by S. Mrowka in "On the Potency of Compact Spaces and the First Axiom of Countability," *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.* 6 (1958) 7-9.

> Let X be the set of all polynomials with rational coefficients, with the topology induced by the product topology on R^R (that is, the topology of pointwise convergence for functions from R to R). Then X is countable but nowhere first countable.

This example has several advantages over the others: it is more elementary, more easily described, and has a significant role in analysis.

Paul R. Meyer
Herbert H. Lehman College
Bronx
New York 10468

ERROR DISCOVERED

It is amusing to note that, in all of the correspondence (this *Magazine* News and Notes, September and November 1977, March and September 1978) regarding Manvel's article "Counterfeit Coin Problems" (this *Magazine*, March 1977, pp. 90-92), it has not been pointed out that the formulae given in Theorems 2 and 3 are incorrect! The error was discovered in my Liberal Arts Math class when it was determined that only 2 weighings are required to determine the strange coin among four.

The problem is in the proof of Theorem 2. The "anchor" in the induction proof alleges that $M(1) = 1$ and $M(2) = 4$ when, in fact, $M(1) = 2$ and $M(2) = 5$. This increases the maximum size of S by 1 so that $M(n) = (3^n+1)/2$. This then increases the maximum size of S by 1 in Theorem 3.

The correct inequalities for Theorems 2 and 3, respectively, are:

$$\frac{3^{n-1}+1}{2} < |S| \leq \frac{3^n+1}{2}$$

and

$$\frac{3^{n-1}-1}{2} < |S| \leq \frac{3^n-1}{2} .$$

Richard A. Gibbs
Fort Lewis College
Durango
Colorado 81301

GIVING CREDIT...

I regret that the excerpt of my letter--the one that appeared on page 60 of the January 1979 issue of *Mathematics Magazine*--was so abbreviated as to give a false impression. It would have been most accurate to leave it as I had originally phrased it: the paper "On a Class of Relatively Prime Sequences" describes work of Erdös and Pomerance that I was helping Pomerance with. I can take credit for only a small part of the proof of the theorem mentioned. Most of the work was done by my two coauthors. The name of Carl Pomerance should not have been omitted.

David E. Penney
The University of Georgia
Athens
Georgia 30602

1979 U.S.A. MATHEMATICAL OLYMPIAD

The eighth annual U.S.A. Mathematical Olympiad, which took place on May 1, 1979, consisted of the following five problems. These problems were prepared by a committee consisting of Murray Klamkin, Chairman, Samuel Greitzer, Tom Griffiths, and Cecil Rousseau.

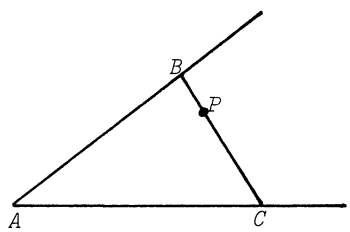
1. Determine all non-negative integral solutions $(n_1,n_2,\ldots,n_{14})$ if any, apart from permutations, of the Diophantine equation

$$n_1^4 + n_2^4 + \ldots + n_{14}^4 = 1,599 .$$

2. A great circle E on a sphere is one whose center is the center O of the sphere. A pole P of the great circle E is a point on the sphere such that OP is perpendicular to the plane of E. On any great circle through P, two points A and B are chosen equidistant from P. For any spherical triangle ABC (the sides are great circle arcs) where C is on E, prove that the great circle arc CP is the angle bisector of angle C.

3. Given three identical n-faced dice whose corresponding faces are identically numbered with arbitrary integers. Prove that if they are tossed at random, the probability that the sum of the top three face numbers is divisible by three is greater or equal to 1/4.

4. Show how to construct a chord BPC of a given angle A through a given point P within the angle A such that $1/BP + 1/PC$ is a maximum.



5. A certain organization has n numbers $(n \geq 5)$ and it has $n + 1$ three member committees, no two of which have identical membership. Prove that there are two committees which share exactly one member.

WOMEN SCIENTISTS TO VISIT U.S. HIGH SCHOOLS

The Research Triangle Institute will conduct a Visiting Women Scientists program through May 1979, to encourage young women to consider careers in science and technology. Women scientists will visit approximately 135 junior and senior high schools in the following areas: Los Angeles, Philadelphia, and Minneapolis-St. Paul. The program is in need of women scientists of disadvantaged ethnic minority group backgrounds and women from industry or government in the specific geographic areas. Please write Ms. Carol Place, Project Director, Research Triangle Institute, Box 12194, Research Triangle Park, NC 27709.

MATHEMATICS MEDIA REVIEW

The mathematics department at Adelphi University intends to launch a biannual journal devoted to reviews of non-print media in mathematics. The journal, to appear April 1980, will be called *Adelphi University Reviews: Mathematics Media*. The journal will also consider articles bearing upon the use on production of non-print media for use in undergraduate mathematics.

The editors would be glad to hear from anyone willing to write reviews for the new journal. Please address all correspondence to: *Adelphi University Reviews: Mathematics Media*, Department of Mathematics, Adelphi University, Garden City, New York 11530.

1979 NSF-CBMS REGIONAL RESEARCH CONFERENCES

The National Science Foundation has granted through the Conference Board of the Mathematical Sciences ten Regional Research Conferences for college and university mathematics teachers for the summer and fall of 1979. Each Regional Conference features ten lectures by a distinguished guest expert; subsequently a monograph by the principal lecturer based on his Regional Conference lectures normally appears in the Regional Conference Series in Mathematics published by the American Mathematical Society or in the Regional Conferences Series in Applied Mathematics published by the Society for Industrial and Applied Mathematics.

Date	Host Institution	Subject	Lecturer
May 7-12	University of Georgia	Algebraic and Analytic Geometry	P. Griffiths
June 4-8	Iowa State University	Mathematical Population Genetics	J.F. Kingman
June 11-15	SUNY, Albany	Ergodic Theory	B. Weiss
June 18-22	St. Olaf College	Ramsey Theory	R.L. Graham
June 18-22	University of Tennessee	Finite Elasticity	M. Gurtin
June 25-29	University of Missouri	Recent Advances in Reliability	F. Proschan
June 25-29	University of Montana	Approximation of Population Processes	T.G. Kurtz
June 25-29	Oakland University	Recent Progress in Operator Algebras	E. Effros
Aug. 6-10	George Mason University	Analytic Methods in Commutative Algebra	M. Hochster
Aug. 13-17	Tufts University	Recent Developments in Celestial Mechanics	R. McGehee

Approximately 25 mathematicians can attend each conference; travel and subsistence allowances are paid by NSF. Inquiries about particular conferences and requests for application forms should be addressed to the departments of mathematics at the host institutions.

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036

MATHEMATICS MAGAZINE VOL. 52, NO.3 MAY 1979