MATHEMATICS MAGAZINE

THE MATHEMATICAL ASSOCIATION · OF AMERICA ·

Vol. 58 No. 1
JANUARY 1985

STRIKING A LOGO • TWO SQUARES AND TWO RINGS
CALENDAR GENERATOR • DESIGNING A SPRINKLER

Eminent Mathematicians and Mathematical Expositors speak to

STUDENTS and TEACHERS in . . .

The NEW MATHEMATICAL LIBRARY

An internationally acclaimed paperback series providing:

- stimulating excursions for students beyond traditional school mathematics.
- supplementary reading for school and college classrooms.
- valuable background reading for teachers.
- challenging problems for solvers of all ages from high school competitions in the US and abroad.

The New Mathematical Library is published by the MATHEMATICAL ASSOCIATION OF AMERICA. The volumes are paperbound.

NUMBERS: RATIONAL AND IRRATIONAL by Ivan Niven. NML-01

WHAT IS CALCULUS ABOUT? by W. W. Sawyer. NML-02

AN INTRODUCTION TO INEQUALITIES, by E. F. Beckenbach, and R. Bellman. NML-03

GEOMETRIC INEQUALITIES, by N. D. Kazarinoff. NML-04

THE CONTEST PROBLEM BOOK. Problems from the Annual High School Mathematics Examinations sponsored by the MAA, NCTM, Mu Alpha Theta, The Society of Actuaries, and the Casualty Actuarial Society. Covers the period 1950-1960. Compiled and with solutions by C. T. Salkind. NML-05

THE LORE OF LARGE NUMBERS, by P. J. Davis. NML-06

USES OF INFINITY, by Leo Zippin. NML-07

GEOMETRIC TRANSFORMATIONS, by I. M. Yaglom, translated by Allen Shields. NML-08

CONTINUED FRACTIONS, by C. D. Olds. NML-09

GRAPHS AND THEIR USES, by Oystein Ore. NML-10

HUNGARIAN PROBLEM BOOKS I and II, based on the Eötvos Competitions 1894-1905 and 1906-1928. Translated by E. Rapaport. NML-11 and NML-12

EPISODES FROM THE EARLY HISTORY OF MATHEMATICS, BY A. Aaboe. NML-13

GROUPS AND THEIR GRAPHS, by I. Grossman and W. Magnus. NML-14

THE MATHEMATICS OF CHOICE, by Ivan Niven. NML-15

FROM PYTHAGORAS TO EINSTEIN, by K. O. Friedrichs. NML-16

THE CONTEST PROBLEM BOOK II. A continuation of NML-05 containing problems and solutions from the Annual High School Mathematics Examinations for the period 1961-1965. NML-17

FIRST CONCEPTS OF TOPOLOGY, by W. G. Chinn and N. E. Steenrod. NML-18

GEOMETRY REVISITED, by H.S.M. Coxeter, and S. L. Greitzer. NML-19

INVITATION TO NUMBER THEORY, by Oystein Ore. NML-20

GEOMETRIC TRANSFORMATIONS II, by I. M. Yaglom, translated by Allen Shields. NML-21

ELEMENTARY CRYPTANALYSIS—A Mathematical Approach, by Abraham Sinkov. NML-22

INGENUITY IN MATHEMATICS, by Ross Honsberger. NML-23

GEOMETRIC TRANSFORMATIONS III, by I. M. Yaglom, translated by Abe Shenitzer. NML-24

THE CONTEST PROBLEM BOOK III. A continuation of NML-05 and NML-17, containing problems and solutions from the Annual High School Mathematics Examinations for the period 1966-1972. NML-25

MATHEMATICAL METHODS IN SCIENCE, by George Pólya. NML-26

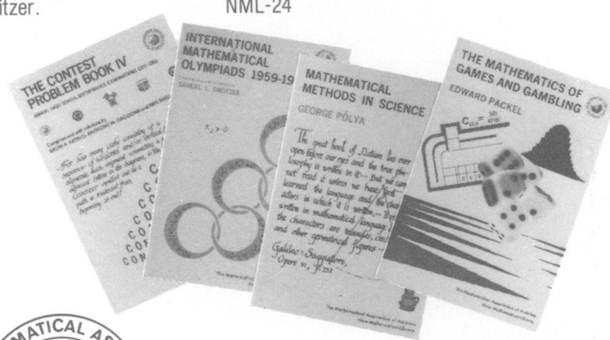
INTERNATIONAL MATHEMATICAL OLYMPIADS, 1959-1977. Problems, with solutions, from the first nineteen International Mathematical Olympiads. Compiled and with solutions by S. L. Greitzer. NML-27

THE MATHEMATICS OF GAMES AND GAMBLING, by Edward W. Packel. NML-28

THE CONTEST PROGRAM BOOK IV, Annual High School Mathematics Examinations 1973-1982. Compiled and with solutions by R. A. Artino, A. M. Gaglione and Niel Shell. NML-29

THE ROLE OF MATHEMATICS IN SCIENCE, by M. M. Schiffer and Leon Bowden. NML-30

For information regarding the prices of these publications, please contact The Mathematical Association of America at the address listed below.



Send Orders to:
The Mathematical Association of America
1529 Eighteenth St., N.W.
Washington, D.C. 20036

MATHEMATICS MAGAZINE

Volume 58, Number 1, January 1985
(ISSN—0025-570X)

COVER: After sixty years, an accurately drawn MAA logo. See pp. 12–18.

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

EDITORIAL POLICY

Mathematics Magazine is a journal which aims to provide inviting, informal mathematical exposition. Manuscripts for the *Magazine* should be written in a clear and lively expository style and stocked with appropriate examples and graphics. Our advice to authors is: say something new in an appealing way or say something old in a refreshing way. The *Magazine* is not a research journal and so the style, quality, and level of articles should realistically permit their use to supplement undergraduate courses. The editor invites manuscripts that provide insight into the history and application of mathematics, that point out interrelationships between several branches of mathematics and that illustrate the fun of doing mathematics.

The full statement of editorial policy appears in this *Magazine,* Vol. 54, pp. 44-45, and is available from the Editor. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, nor published by another journal or publisher.

Send new manuscripts to: Doris Schattschneider, Editor, Mathematics Magazine, Moravian College, Bethlehem, PA 18018. Manuscripts should be typewritten and double spaced and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should submit the original and one copy and keep one copy. Illustrations should be carefully prepared on separate sheets in black ink, the original without lettering and two copies with lettering added.

AUTHORS

Harley Flanders ("A Tale of Two Squares—and Two Rings") received his Ph.D. at the University of Chicago under O. F. G. Schilling and A. Weil. This was so long ago that the names of A. A. Albert and M. H. Stone, under both of whom he served as research assistant, may be unknown to some mathematicians of the present generation. Over the years he held various jobs, mainly at Cal. Tech. (Bateman Fellow), University of California, Berkeley (sub sub jr., sub jr., and jr. faculty), Purdue University, and Tel Aviv University. His professional service includes a stint as Editor-in-Chief of the *American Mathematical Monthly.* His metamorphosis through class field theory, differential geometry, and linear algebra towards things numerical and computational (recent book: *Scientific Pascal)* started with electrical network theory and has been long and painful. It is possibly in arrested development at the chrysalis stage.

ILLUSTRATIONS

Ellen Gerkin pictures the sprinkler designed by a puppy (p. 294).

All other illustrations were provided by the authors.

The MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, five times a year: January, March, May, September, and November.

The annual subscription price for the MATHEMATICS MAGAZINE to an individual member of the Association is $11 included as part of the annual dues. (Annual dues for regular members, exclusive of annual subscription prices for MAA journals, are $22. Student, unemployed and emeritus members receive a 50% discount; new members receive a 30% dues discount for the first two years of membership.) The non-member/library subscription price is $28 per year. Bulk subscriptions (5 or more copies) are available to colleges and universities for classroom distribution to undergraduate students at a 41% discount ($6.50 per copy—minimum order $32.50).

Subscription correspondence and notice of change of address should be sent to the Membership/Subscriptions Department, 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 Company, Middletown, CT 06457. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to Ms. Elaine Pedreira, Advertising Manager, The Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

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

Postmaster: Send address changes to Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

PRINTED IN THE UNITED STATES OF AMERICA

ARTICLES

A Tale of Two Squares—and Two Rings

Gaussian integers are used to obtain information about rational integers.

HARLEY FLANDERS
Florida Atlantic University
Boca Raton, FL 33431

One of the glorious chapters of elementary number theory, going back to Fermat and Euler, concerns the problem: *which elements of the ring $\boldsymbol{Z}$ of rational integers are sums of two squares*? From the point of view of arithmetic, the ring $\boldsymbol{Z}$ is truly the best of rings (however, algebraists often think $\boldsymbol{Z}$ the worst of rings, overwhelmingly devoid of structure). The tale I want to tell goes one step beyond elementary number theory by using the arithmetic of a second ring, the ring $\boldsymbol{G}$ of Gaussian integers, to solve the problem. I shall try to include enough background material and detail so that a motivated undergraduate could safely be unleashed upon this story, given an occasional assist in algebra.

Divisibility in $\boldsymbol{Z}$

We first review the basic multiplicative arithmetic of $\boldsymbol{Z}$, and do it in such a way that our later study of $\boldsymbol{G}$ will seem natural.

Recall that if a and b are in $\boldsymbol{Z}$, we say that a **divides** b (and write $a|b$) provided $a \neq 0$ and $ax = b$ has a (unique) solution in $\boldsymbol{Z}$. A **unit** is a number e in $\boldsymbol{Z}$ that divides 1; this of course means that $e = 1$ or $e = -1$. A **prime** is a non-unit p in $\boldsymbol{Z}$ such that a factorization $p = xy$ implies x is a unit or y is a unit. (Negative primes like -7 are allowed.)

There is an order relation in $\boldsymbol{Z}$, tied up with the relation $<$ in its various forms. We know when $n > 0$. We also know the absolute value function, and we shall write it as a norm function in the notation $N(a)$. Thus

$$N(0) = 0 \qquad N(-7) = 7 \qquad N(15) = 15.$$

Note that $N(ab) = N(a)N(b)$.

Let us start with an unconventional form of that most famous of all algorithms, Euclid's division algorithm.

THEOREM 1. (Division Algorithm) *If a and $b \neq 0$ are in $\boldsymbol{Z}$, then there exist q and r in $\boldsymbol{Z}$ such that*

$$a = bq + r \quad \textit{and} \quad N(r) \leqslant \frac{1}{2}N(b).$$

Proof. Let q be the integer nearest the rational a/b. (If there are two such, choose either one, it doesn't matter which.) Then $|a/b - q| \leqslant \frac{1}{2}$, that is,

$$N(a - bq) \leqslant \frac{1}{2}N(b).$$

Take $r = a - bq$. Done.

The remaining theorems of this section are important, and are standard in every elementary number theory text. See, for example, Chapter 1 of [5] for proofs. The next "algorithm" also goes back to Euclid.

THEOREM 2. *Let a and b be in* $\boldsymbol{Z}$, *not both* 0. *Then there exists a linear combination*

$$d = ax + by \qquad (x, y \in \boldsymbol{Z})$$

that divides both a and b. Any integer that divides both a and b divides d.

Even though d is only unique up to a unit multiplier, we write $d = (a, b)$ and refer to d as the **greatest common divisor** of a and b.

THEOREM 3. *If p is a prime and* $p|ab$, *then* $p|a$ *or* $p|b$.

THEOREM 4. (Fundamental Theorem of Arithmetic) *Each non-zero non-unit n of* $\boldsymbol{Z}$ *is a product of primes. The factorization is unique up to order provided that unit factors are ignored, that is, provided p and* $-p$ *are considered the same.*

Gaussian integers

If $n = x^2 + y^2$, then n can be written as a product

$$n = (x + yi)(x - yi)$$

of a complex number and its conjugate. This suggests that we study the set

$$\boldsymbol{G} = \{x + yi \,|\, x, y \in \boldsymbol{Z}\}$$

of **Gaussian integers**. It is easy to see that $\boldsymbol{G}$ is a commutative ring, indeed, $\boldsymbol{G}$ is a domain, that is, $\boldsymbol{G}$ has no zero divisors. It inherits this property by being a subring of the field $\boldsymbol{C}$ of complex numbers.

It is most important that **conjugation**, $\alpha \to \bar{\alpha}$, defined by

$$\bar{\alpha} = \overline{a + bi} = a - bi,$$

is an automorphism of $\boldsymbol{G}$. Obviously $\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta}$. Not quite so obvious is the preservation of multiplication; $\overline{\alpha\beta} = \bar{\alpha}\bar{\beta}$. Conjugation is a crucial arithmetic tool. We use it to define a **norm** function N: $\boldsymbol{G} \to \boldsymbol{Z}$ by

$$N(\alpha) = \alpha\bar{\alpha}.$$

In ordinary absolute value notation, $N(\alpha) = |\alpha|^2$. Clearly

$$N(\alpha\beta) = (\alpha\beta)(\overline{\alpha\beta}) = (\alpha\beta)(\bar{\alpha}\bar{\beta}) = (\alpha\bar{\alpha})(\beta\bar{\beta}) = N(\alpha)N(\beta).$$

Also, $N(a + bi) = a^2 + b^2$, so the norm function goes hand-in-glove with representing integers as sums of two squares.

The basic definitions concerning divisibility in $\boldsymbol{G}$ are analogous to those for $\boldsymbol{Z}$. First, α **divides** β if $\alpha \neq 0$ and $\alpha z = \beta$ has a (unique) solution in $\boldsymbol{G}$, that is, if the complex number β/α is in $\boldsymbol{G}$. If so, we write $\alpha|\beta$, we say that α is a **divisor** of β, and we say that β is a **multiple** of α. For instance $\alpha = 4 + 5i$ divides $\beta = 23 - 2i$ because $\alpha(2 - 3i) = \beta$. But $\gamma = 2 - 3i$ is not a multiple of $\bar{\gamma} = 2 + 3i$ because $\gamma/\bar{\gamma} = -\frac{5}{13} - \frac{12}{13}i$ is not in $\boldsymbol{G}$. The following result connects $\boldsymbol{G}$-divisibility with $\boldsymbol{Z}$-divisibility.

THEOREM 5. *If* $\alpha|\beta$ *then* $N(\alpha)|N(\beta)$.

Proof. If $\beta = \alpha z$, then $N(\beta) = N(\alpha)N(z)$ is a multiple of $N(\alpha)$ in $\boldsymbol{Z}$.

There is a slight ambiguity here. Suppose m and n are in $\boldsymbol{Z}$. Since $m = m + 0i$, the ring $\boldsymbol{Z}$ is a subring of $\boldsymbol{G}$. Suppose m divides n in $\boldsymbol{G}$. Is this the same as $m|n$ in $\boldsymbol{Z}$? Absolutely yes, because $n = m(a + bi)$ implies $n = ma$.

A **unit** of G is a number that divides 1 in G. A **prime** of G is a non-unit π such that $\pi = \alpha\beta$ implies α is a unit or β is a unit. In other words, if $\alpha | \pi$, then either α is a unit, or $\alpha = \pi\varepsilon$, where ε is a unit. One of our jobs later will be to determine all the primes of G. At the moment we go after the units.

Suppose ε is a unit of G. Then $\varepsilon | 1$, so by Theorem 5, we have $N(\varepsilon) | N(1) = 1$. This implies $N(\varepsilon) = 1$ (because $N(\varepsilon) > 0$). If $\varepsilon = a + bi$, then $a^2 + b^2 = N(\varepsilon) = 1$. The only solutions in Z are $a = \pm 1$, $b = 0$ and $a = 0$, $b = \pm 1$. Hence $\varepsilon = \pm 1$ or $\varepsilon = \pm i$. Clearly each of these four numbers is a unit. We have proved

THEOREM 6. *The units of* G *are* 1, -1, i, *and* $-i$.

We are ready for the division algorithm.

THEOREM 7. (Division Algorithm in G) *If* α *and* $\beta \neq 0$ *are in* G, *then there exist* σ *and* ρ *in* G *such that*

$$\alpha = \beta\sigma + \rho \quad \text{and} \quad N(\rho) \leqslant \frac{1}{2} N(\beta).$$

Proof. Mimicking the proof of Theorem 1, we let σ be an element of G nearest α/β in the complex plane. Since α/β lies in a unit square whose vertices are in G, there are at most four choices for the nearest σ, and we choose any one of them. Clearly $|\alpha/\beta - \sigma|$ is at most half the diagonal of the square, so $|\alpha/\beta - \sigma| \leqslant \frac{1}{2}\sqrt{2}$. This means

$$N(\alpha - \beta\sigma) \leqslant \frac{1}{2} N(\beta).$$

We set $\rho = \alpha - \beta\sigma$, and the proof is completed.

The standard proof of Theorem 2 for the ring Z goes through verbatim (except for notation) for G, so let us merely state the result.

THEOREM 8. *Let* α *and* β *be in* G, *not both* 0. *Then there exists a linear combination*

$$\delta = \alpha\mu + \beta\nu \qquad (\mu, \nu \in G)$$

that divides both α *and* β. *Any other element of* G *that divides both* α *and* β *divides* δ.

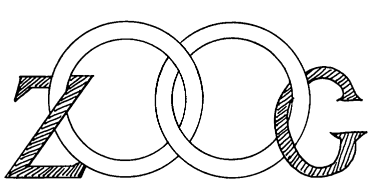
Even though δ is only unique up to a unit multiplier, we write $\delta = (\alpha, \beta)$ and refer to δ as the **greatest common divisor** of α and β. The critical tool in proving unique factorization in Z is Theorem 3. Its proof also goes through verbatim for G, so we may state the result.

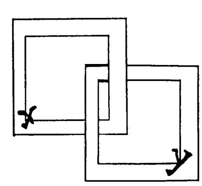
THEOREM 9. *If* π *is a prime in* G *and* $\pi | \alpha\beta$, *then* $\pi | \alpha$ *or* $\pi | \beta$.

This completes paving the way for the Fundamental Theorem for G. It is useful to work through the proofs for G of Theorems 7–10 in order to appreciate their validity in this non-Z setting.

THEOREM 10. (Fundamental Theorem of G-Arithmetic) *Each non-zero non-unit* α *of* G *is a product of primes. The factorization is unique up to order provided that unit factors are ignored, that is,* π *and* $\varepsilon\pi$ *are considered as the same whenever* ε *is a unit.*

We have reached the point where the multiplicative arithmetic of G ceases to look exactly like the multiplicative arithmetic of Z. Our job now is to survey the primes of G. We shall see that there is an intimate relation between primes of G and primes of Z.





The primes of G

Let π be a prime in G. Then $N(\pi)=\pi\bar{\pi}$ is a positive element of Z. Clearly $N(\pi)\neq 1$, because π is not a unit. By the Fundamental Theorem in Z,

$$N(\pi)=\pi\bar{\pi}=p_1 p_2 \cdots p_r \tag{1}$$

where the p_i are Z-primes.

LEMMA. *Let π be a G-prime. Then there exists a unique Z-prime p such that $\pi | p$. Furthermore, either $N(\pi)=p$ or $N(\pi)=p^2$.*

Proof. By Theorem 9 and equation (1), the G-prime π divides some Z-prime p. The prime p is unique, because if not, then $\pi | q$, where q is a Z-prime different from p. But then $1=(p,q)=px+qy$, where x and y are in Z, hence in G, and we have

$$\pi | (px+qy)=1,$$

clearly impossible. Next, from $\pi | p$ and Theorem 5 follows $N(\pi)|N(p)=p^2$, hence by the Fundamental Theorem in Z, we have $N(\pi)=p$ or $N(\pi)=p^2$. (Of course $N(\pi)\neq 1$.)

The unique Z-prime p determined by π is called its **associated Z-prime**. In the case $N(\pi)=p$, we have $\pi\bar{\pi}=p$, so p has exactly two prime factors in G. In the case $N(\pi)=p^2$, then the Fundamental Theorem in G applied to $\pi\bar{\pi}=p^2$ implies that p must be a G-prime, $\pi=\varepsilon p$, where ε is a unit.

Let us go at it from the opposite direction. Let p be a Z-prime, and factor p in G. Then surely p has at least one G-prime factor π. But then $\pi | p$ so either $N(\pi)=p$ and $p=\pi\bar{\pi}$, or $N(\pi)=p^2$ and p *is* a G-prime. Clearly we are faced with the problem of determining which p remain prime in G and which factor as $p=\pi\bar{\pi}$.

Suppose $N(\pi)=p$, and let $\pi=x+yi$. Then

$$p=N(\pi)=\pi\bar{\pi}=x^2+y^2,$$

so p is the sum of two squares. Clearly neither $x=0$ nor $y=0$, or else p would be a square, not a prime.

Conversely, suppose p is a Z-prime and $p=x^2+y^2$, where $x\neq 0$ and $y\neq 0$. Then

$$p=\alpha\bar{\alpha}, \quad \text{where } \alpha=x+yi.$$

Clearly α is not a unit. Let π be a G-prime factor of α. Then $N(\pi)|N(\alpha)=p$. Therefore p is the associated Z-prime of π and $\alpha=\varepsilon\pi$, $p=\pi\bar{\pi}$. Let us summarize these remarks.

THEOREM 11. *Let π be a G-prime. Then there is a unique Z-prime p such that $\pi | p$, and either $N(\pi)=p$ or $\pi=\varepsilon p$, where ε is a unit.*

Let p be a Z-prime. Then either p is a G-prime or $p=\pi\bar{\pi}$, where π and $\bar{\pi}$ are G-primes. This is the case if and only if

$$p=x^2+y^2$$

can be solved in Z. If so, then $N(\pi)=p$ and we may take $\pi=x+yi$.

Thus to survey the G-primes, we simply run over the Z-primes and test which are sums of squares. Let us try the first few.

$p=2$. Now $2=1^2+1^2=(1+i)(1-i)=\pi\bar{\pi}$. Something special happens here:

$$\bar{\pi}=1-i=-i(1+i)=-i\pi$$

so, up to a unit, 2 equals π^2. (Then 2 is said to be **ramified** in G, meaning that 2 is divisible by the square of a prime.)

$p=3$. It is easily checked that $3\neq x^2+y^2$, hence 3 is a G-prime.

$p = 5$. Now $5 = 2^2 + 1^2 = (2+i)(2-i) = \pi\bar{\pi}$.

$p = 7$ and $p = 11$. Both not sums of squares, hence both are $\boldsymbol{G}$-primes.

$p = 13 = 3^2 + 2^2$, so $p = \pi\bar{\pi}$ where $\pi = 3 + 2i$.

$p = 17 = 4^2 + 1^2$, so $p = \pi\bar{\pi}$ where $\pi = 4 + i$.

We have taken care of the only even prime, $p = 2$; we can temporarily restrict attention to odd primes. The fact that 3, 7, and 11 are $\boldsymbol{G}$-primes is part of a general pattern, but to explain it we must use congruences. If $x \in \boldsymbol{Z}$ then $x = 2n$ or $x = 2n+1$. In the first case $x^2 = 4n^2 \equiv 0 \bmod 4$ and in the second case

$$x^2 = (2n+1)^2 = 4n(n+1) + 1 \equiv 1 \bmod 4,$$

therefore, if x and y are in $\boldsymbol{Z}$, then $x^2 + y^2 \equiv 0$, 1, or $2 \bmod 4$. Thus a sum of two squares is never congruent to $3 \bmod 4$. Taken with Theorem 11, this implies the following result.

THEOREM 12. *Let p be a $\boldsymbol{Z}$-prime and suppose $p \equiv 3 \bmod 4$. Then p is a $\boldsymbol{G}$-prime.*

Every other odd prime is congruent to $1 \bmod 4$. We shall next prove that each such prime is a sum of two squares. This result, in a sense, is the crux of the matter, and it is by no means an easy result. To prove it, we shall make an excursion into finite fields and groups, preceded by a preliminary detour.

Finite groups in fields

Recall the Euler **totient** or ϕ function, defined on positive integers: $\phi(m)$ is the number of integers r in the interval $1 \leqslant r \leqslant m$ such that $(r, m) = 1$. We need to recall one result about this function besides its definition, the formula

$$\sum_{d|m} \phi(d) = m.$$

This is a standard result in number theory. For example, see Chapters 2 and 4 of [**5**].

The following theorem is a useful result with many applications.

THEOREM 13. *Let F be a field and $F^* = F \setminus \{0\}$ its multiplicative group. Let H be a finite subgroup of F^*. Then H is a cyclic group.*

Proof. Let $|H| = m$. If $h \in H$, then $\operatorname{ord}(h)|m$. The order function, from H into the set of divisors of m, partitions H, and we can write

$$m = \sum_{d|m} \psi(d) \quad \text{where} \quad \psi(d) = |\{h \in H | \operatorname{ord}(h) = d\}|.$$

Our problem is to prove that $\psi(m) > 0$. For then there is an element h of H of order m, that is, a generator of H, so H is a cyclic group. We shall do this by proving that $\psi(d) \leqslant \phi(d)$. If we can prove this, then comparison of

$$m = \sum_{d|m} \psi(d) \text{ with } m = \sum_{d|m} \phi(d)$$

implies that $\psi(d) = \phi(d)$ for each divisor d of m. In particular, $\psi(m) = \phi(m) > 0$.

So let $d|m$. If $\psi(d) = 0$, then surely $\psi(d) \leqslant \phi(d)$. Suppose $\psi(d) > 0$. Then there exists an element h_0 of the group H such that $\operatorname{ord}(h_0) = d$. The elements

$$1, h_0, h_0^2, \ldots, h_0^{d-1}$$

are then d distinct elements of F, and each is a zero of the polynomial

$$f(x) = x^d - 1.$$

But we are in a field, and in a field, a polynomial of degree d has *at most* d distinct zeros.

Therefore these powers of h_0 are *all* of the zeros of $f(x)$. Hence if $h \in H$ and $\operatorname{ord}(h) = d$, then $h = h_0^r$ for some r. But $\operatorname{ord}(h_0^r) = d$ if and only if $(r, d) = 1$. Therefore, in this case,

$$\begin{aligned}\psi(d) &= |\{h \in H | \operatorname{ord}(h) = d\}| \\ &= |\{h_0^r | 1 \leqslant r \leqslant d \quad \text{and} \quad (r, d) = 1\}| = \phi(d).\end{aligned}$$

So again $\psi(d) \leqslant \phi(d)$, completing the proof.

The case $p \equiv 1 \bmod 4$

We have reached one of our main objectives.

THEOREM 14. *Let p be a $\boldsymbol{Z}$-prime such that $p \equiv 1 \bmod 4$. Then p is a sum of two squares in $\boldsymbol{Z}$, and the factorization of p in $\boldsymbol{G}$ has the form $p = \pi\bar{\pi}$, where π and $\bar{\pi}$ are both $\boldsymbol{G}$-primes. We also have $N(\pi) = N(\bar{\pi}) = p$. Finally π and $\bar{\pi}$ are truly distinct primes; neither is a unit times the other.*

This is quite a bit to prove, and we shall take it in several steps.

LEMMA A. *Let $p \equiv 1 \bmod 4$. Then $x^2 \equiv -1 \bmod p$ has a solution in $\boldsymbol{Z}$.*

Proof. Let $F_p = \boldsymbol{Z}/\boldsymbol{Z}p$ denote the finite field of p elements, the ring $\boldsymbol{Z}$ modulo its maximal ideal $\boldsymbol{Z}p$. Let $H_p = F_p^*$, the multiplicative group of F_p. Then H_p is a finite group of $p - 1$ elements, and, by Theorem 13, is a cyclic group. Let h be a generator of H_p and set $k = h^n$, where $p - 1 = 4n$. The element k^2 has order 2 because k has order 4. But only $+1$ and -1 in F_p satisfy $x^2 = 1$, hence $k^2 = -1$. If x is any $\boldsymbol{Z}$-representative of the residue class k, then $x^2 \equiv -1 \bmod p$.

For example, if $p = 13$, then $h = 2$ in F_{13} generates H_{13}, and $k = h^3 = 8$ satisfies $k^2 = -1$ in F_{13}. Pulled back to $\boldsymbol{Z}$, this says $8^2 \equiv -1 \bmod 13$.

LEMMA B. *Let $p \equiv 1 \bmod 4$ and $x^2 \equiv -1 \bmod p$. Then*

$$p = \pi\bar{\pi} \quad \textit{where} \quad \pi = (x + i, p).$$

Proof. Set $\alpha = (x + i, p)$ so that $\bar{\alpha} = (x - i, p)$. On the one hand, $\alpha | p$ so that $\alpha\bar{\alpha} = N(\alpha) | N(p) = p^2$. On the other hand, α is a linear combination of $x + i$ and p and $\bar{\alpha}$ is a linear combination of $x - i$ and p. Multiplying, we deduce that $\alpha\bar{\alpha}$ is a linear combination of

$$(x + i)(x - i), \quad p(x + i), \quad p(x - i), \quad p^2,$$

each divisible by p. The first one because

$$(x + i)(x - i) = x^2 + 1 \equiv 0 \bmod p.$$

Hence $p | N(\alpha) | p^2$. Therefore either $N(\alpha) = p$ or $N(\alpha) = p^2$.

We shall rule out the second case. Suppose $N(\alpha) = p^2$. Then $\alpha\bar{\alpha} = p^2$. But $\alpha | p$ and $\bar{\alpha} | p$, so this forces $p = \alpha\varepsilon$ and $p = \bar{\alpha}\bar{\varepsilon}$, where ε is a unit. Thus

$$p | \alpha = (x + i, p) \quad \text{so that} \quad p | (x + i).$$

This is impossible because $(x/p) + (1/p)i$ is not in $\boldsymbol{G}$.

Consequently $N(\alpha) = p$. Since we know that either p is a $\boldsymbol{G}$-prime or $p = \pi\bar{\pi}$, the factorization $p = N(\alpha) = \alpha\bar{\alpha}$ implies the latter option: $p = \pi\bar{\pi}$, in fact, by the Fundamental Theorem, we can take $\pi = \alpha$, completing the proof of Lemma B.

Proof of Theorem 14. Only the last statement remains to be proved. Let $\pi = u + vi$ and suppose $\pi = \varepsilon\bar{\pi}$. Then

$$p = \pi\bar{\pi} | \pi^2 = (u^2 - v^2) + 2uvi, \quad \text{hence } p | 2uv.$$

If we can prove that $p \nmid u$ and $p \nmid v$, then $p | 2uv$ would imply $p = 2$, which is not the case. Suppose $p | u$. Then $\pi | u$, hence $\pi | (\pi - u) = vi$, so $\pi | v$. But then $p = \pi\bar{\pi} | v^2$, so $p | v$, hence $p | (u + vi) = \pi$, and

$$p^2 = N(p) | N(\pi) = p,$$

impossible. Similarly, $p|v$ is impossible, so the proof is complete.

Let us continue the example $p = 13$. We had $8^2 \equiv -1 \bmod 13$, hence

$$13 = \pi\bar{\pi} \quad \text{where} \quad \pi = (8+i, 13).$$

To express π in the form $x + yi$, we use Theorems 7 and 9 explicitly. First,

$$\frac{13}{8+i} = \frac{13(8-i)}{65} = \frac{8}{5} - \frac{13}{65} i.$$

The nearest element in $\boldsymbol{G}$ is 2, and we have

$$13 - 2(8+i) = -3 - 2i.$$

Consequently

$$\pi = (8+i, 2(8+i) - 3 - 2i) = (8+i, -3-2i) = (8+i, 3+2i).$$

Again,

$$\frac{8+i}{3+2i} = \frac{(8+i)(3-2i)}{13} = \frac{26-13i}{13} = 2 - i.$$

Therefore

$$\pi = ((3+2i)(2-i), 3+2i) = 3+2i,$$

so by cold logic we have found the factorization

$$13 = (3+2i)(3-2i) = N(3+2i).$$

Sums of two squares

We have reached the closing chapter of this tale, wherein is unveiled precisely which elements of $\boldsymbol{Z}$ are sums of two squares.

LEMMA C. *If* $p|(x^2+y^2)$ *in* $\boldsymbol{Z}$ *and* $p \equiv 3 \bmod 4$, *then* $p|x$ *and* $p|y$.

Proof. We know that p is a prime in $\boldsymbol{G}$ and

$$p|(x^2+y^2) = (x+yi)(x-yi).$$

By Theorem 9 it follows that $p|(x+yi)$ or $p|(x-yi)$. Either case implies $p|x$ and $p|y$.

THEOREM 15. *Suppose* $x > 0$ *and* $y > 0$ *in* $\boldsymbol{Z}$. *Then*

$$x^2 + y^2 = 2^e m^2 n$$

where n *is a product of* $\boldsymbol{Z}$*-primes congruent to* 1 *modulo* 4 *and where* m *is a product of* $\boldsymbol{Z}$*-primes congruent to* 3 *modulo* 4. *If* e *is even, then* $n > 1$.

Proof. We can extract those prime factors of $x^2 + y^2$ of the form $p \equiv 3 \bmod 4$ by Lemma C and lump them into m^2; each occurs an even number of times. Thus we may assume

$$x^2 + y^2 = 2^e m^2 n,$$

where $e \geqslant 0$ and where $n = 1$ or n is a product of primes congruent to 1 modulo 4. Suppose e is even and $n = 1$. Then

$$x^2 + y^2 = 2^{2f} m^2.$$

We know, by Lemma C applied several times, that $m|x$ and $m|y$. If we take out the factor m^2, the result is

$$u^2 + v^2 = 4^f \qquad (u > 0, v > 0).$$

This is impossible for many reasons, for instance, from $2 = \varepsilon\pi^2$, where $\pi = 1 + i$, we have

$$(u+vi)(u-vi)=\varepsilon^{2f}\pi^{4f},$$

hence

$$\pi^{2f}|(u+vi), \qquad 2^f|(u+vi), \qquad 2^f|u, \qquad 2^f|v,$$
$$(u/2^f)^2+(v/2^f)^2=1.$$

But $u/2^f \geqslant 1$ and $v/2^f \geqslant 1$, a contradiction.

The converse of Theorem 15 is our final item of business.

THEOREM 16. *Each element of* $\boldsymbol{Z}$ *of the form*

$$a=2^e m^2 n,$$

where n is a product of primes congruent to 1 *modulo* 4, *where m is a product of primes congruent to* 3 *modulo* 4, *and where either* $n>1$ *or e is odd, is a sum of two non-zero squares in* $\boldsymbol{Z}$.

Proof. Deleting a perfect square factor from a does not change either the hypotheses or the conclusion, so we can assume

$$a=p_1^{e_1}\cdots p_r^{e_r},$$

where the p_i are distinct $\boldsymbol{Z}$-primes, where $r\geqslant 1$, and where $p_i\equiv 1 \bmod 4$ for all i, except possibly $p_1=2$ and $e_1=1$. Now $p_i=\pi_i\bar{\pi}_i$, so if we set

$$\alpha=\pi_1^{e_1}\cdots\pi_r^{e_r},$$

then

$$\alpha=x+yi \quad \text{and} \quad a=\alpha\bar{\alpha}=x^2+y^2.$$

We are not quite finished yet because possibly $x=0$ or $y=0$. Suppose, for instance, that $y=0$. Then $a=x^2$. Since we are assuming that 2 divides a at most once, we immediately conclude that a is odd, hence x is odd.

Since x is real, $\bar{x}=x$, that is,

$$\bar{\pi}_1^{e_1}\cdots\bar{\pi}_r^{e_r}=\pi_1^{e_1}\cdots\pi_r^{e_r}.$$

This contradicts unique factorization in $\boldsymbol{G}$. (See the final statement in Theorem 14.)

The theorem is slightly more subtle than it appears. For instance, $841=29^2$ is a perfect square. But by Theorem 16, it is also a sum of two non-zero squares because $29\equiv 1 \bmod 4$. Let us work it out:

$$29=25+4=(5+2i)(5-2i)=\pi\bar{\pi},$$
$$\pi=5+2i, \qquad \pi^2=(5+2i)(5+2i)=21+20i,$$
$$841=29^2=\pi^2\bar{\pi}^2=21^2+20^2.$$

EXAMPLE. Find the smallest integer that is a sum of two non-zero squares in two different ways.

The answer is

$$50=7^2+1^2=5^2+5^2$$

and the reason is:

$$50=2\cdot 5^2, \qquad 2=(1+i)(1-i), \qquad 5=(2+i)(2-i),$$
$$(1+i)(2+i)^2=(1+i)(3+4i)=-1+7i$$
$$5(1+i)=5+5i.$$

Think about it.

The author is grateful to H. Alder, M. J. DeLeon, two referees, and the editor for many valuable suggestions.

References

The literature on number theory is so vast that it is hard to know where to start. We have barely touched on a vast subject, the arithmetic of quadratic number fields, itself a fragment of the theory of algebraic numbers. We list a few sources of relevant material.

[1] H. Cohn, A Second Course in Number Theory, Wiley, 1962 (extensive treatment of quadratic fields). Also reprinted under the name Advanced Number Theory, Dover, 1980.

[2] H. Halberstam, Gaps in integer sequences, this MAGAZINE, 56 (1983) 131–140, esp. 135–138.

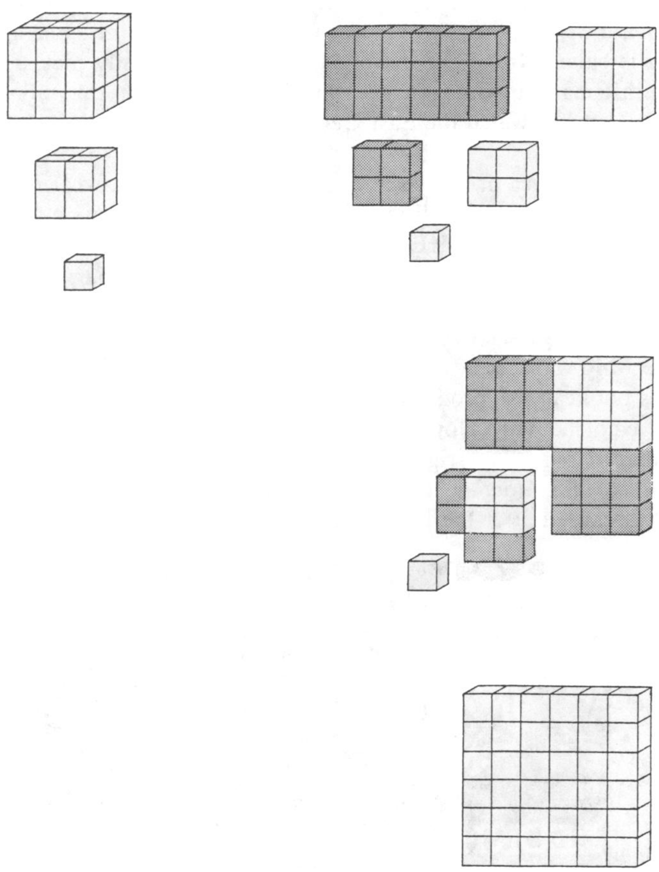
[3] G. H. Hardy and E. M. Wright, An Introduction to the Theory of Numbers, Oxford Univ. Press, 4th ed., 1960. (Chapter 15 on quadratic fields and Chapter 20 on sums of squares).

[4] K. Ireland and M. I. Rosen, Elements of Number Theory, Bogden & Quigley, 1972.

[5] I. Niven and H. S. Zuckerman, An Introduction to the Theory of Numbers, Wiley, 4th ed., 1980 (Section 5.10 and Chapter 9).

Proof without words: Sum of cubes

$$1^3+2^3+\cdots+n^3=(1+2+\cdots+n)^2$$



—ALAN L. FRY
University of North Carolina
at Greensboro

References

The literature on number theory is so vast that it is hard to know where to start. We have barely touched on a vast subject, the arithmetic of quadratic number fields, itself a fragment of the theory of algebraic numbers. We list a few sources of relevant material.

[1] H. Cohn, A Second Course in Number Theory, Wiley, 1962 (extensive treatment of quadratic fields). Also reprinted under the name Advanced Number Theory, Dover, 1980.

[2] H. Halberstam, Gaps in integer sequences, this MAGAZINE, 56 (1983) 131–140, esp. 135–138.

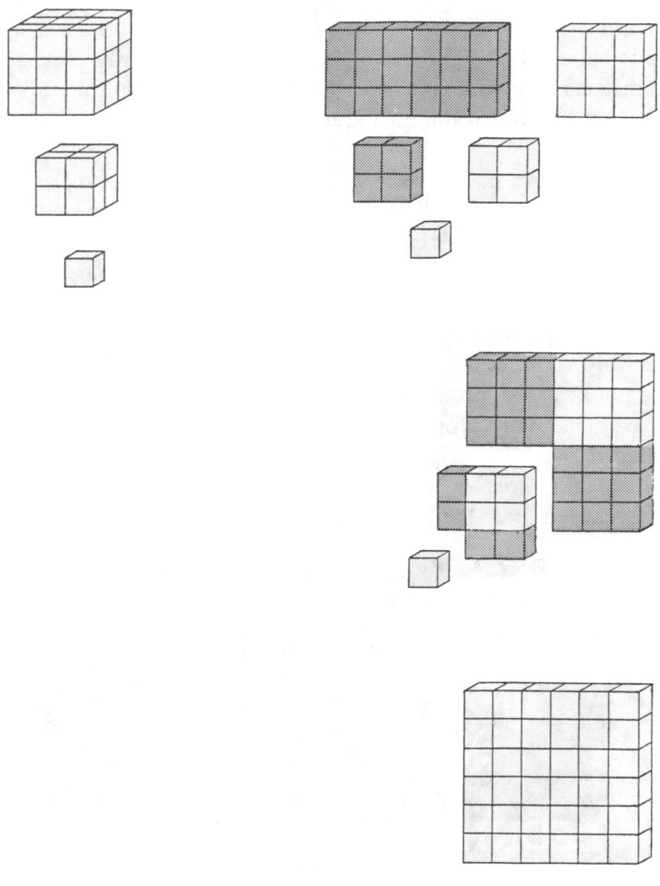
[3] G. H. Hardy and E. M. Wright, An Introduction to the Theory of Numbers, Oxford Univ. Press, 4th ed., 1960. (Chapter 15 on quadratic fields and Chapter 20 on sums of squares).

[4] K. Ireland and M. I. Rosen, Elements of Number Theory, Bogden & Quigley, 1972.

[5] I. Niven and H. S. Zuckerman, An Introduction to the Theory of Numbers, Wiley, 4th ed., 1980 (Section 5.10 and Chapter 9).

Proof without words: Sum of cubes

$$1^3+2^3+\cdots+n^3=(1+2+\cdots+n)^2$$



—ALAN L. FRY
University of North Carolina
at Greensboro

NOTES

Geometry Strikes Again

BRANKO GRÜNBAUM
University of Washington
Seattle, WA 98195

Last Sunday I was leisurely reading the May 1984 issue of MATHEMATICS MAGAZINE. Coming to the "Philatelists take note" item (on page 187) about the stamp issued by East Germany to mark the bicentennial of Euler's death, I remembered that the same stamp was used on a recent request for reprints I received from East Germany. So I retrieved that postcard and washed off the stamp. It turned out that I was lucky, and the stamp was much less marred by cancellation marks than the one reproduced in MATHEMATICS MAGAZINE. Contemplating the stamp I noticed that the drawing of the regular icosahedron shown in it (and I have no doubt that it was meant to show a *regular* icosahedron) is wrong. This is not a question of which kind of perspective or projection was used, it is just a logical error (repeated twice): if three of the five vertices of a plane pentagon project into one line, then the other two vertices must project into the same line. In FIGURE 1 the offending drawing is enlarged and the misaligned vertices are easily picked out. Recalling how highly precision draftsmanship used to be valued in European education, I could not help snickering at the low level to which the East Germans have fallen. Full of condescension I closed my issue of the MAGAZINE.

But the heavier outer cover of the issue opened by itself, and revealed to my horrified and unbelieving eyes another fallacious rendition of the regular icosahedron, in the upper left corner of the title page, IN THE LOGO OF THE MATHEMATICAL ASSOCIATION OF AMERICA!!! Under any reasonable projection, parallel segments either remain parallel or lie on concurrent lines—but in the logo three easily discernible pairs (see FIGURE 2) of such segments do neither. Quickly grabbing the recent issue of "Focus," and then the last issue of the *Monthly*, I saw that the drawing in the MAGAZINE is no isolated distortion. The official symbol of the Association, appearing on all its publications, proclaims louder than words that we have eyes and look but do not see, that those who shouted "Down with Euclid" achieved their goal much faster and much more thoroughly than they had any right to expect—that good old Geometry is dead! Thumbing through back issues of the *Monthly*, the time of death can be determined with precision. The previous logo was discarded after the December 1971 issue, and the defective new one has been

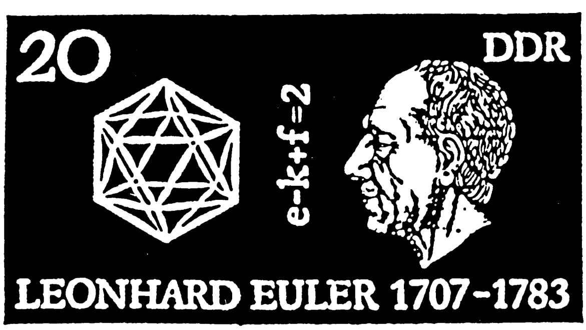


FIGURE 1. **An enlarged view of the East German stamp commemorating the bicentennial of the death of Leonhard Euler. Note the two regular pentagons which are seen "edge on," each of which shows two vertices not lying in the plane determined by its other three vertices.**

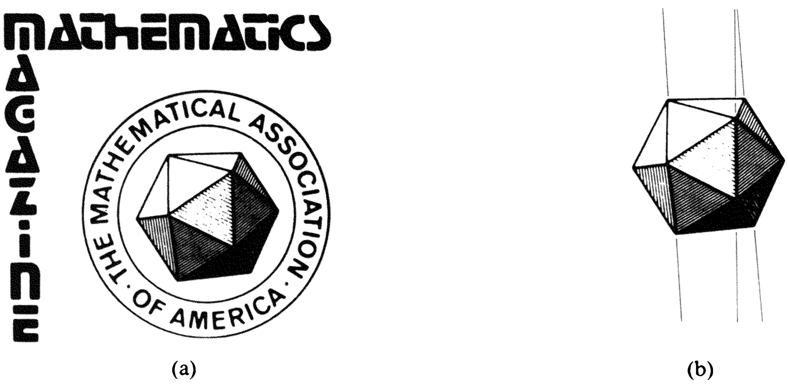


FIGURE 2. **(a) The emblem of the Mathematical Association of America, from the title page of the May 1984 issue of MATHEMATICS MAGAZINE. (b) The icosahedron from (a) with three lines determined by it; the lines should be parallel to each other (if parallel projection is used), or concurrent at one point (if the projection is central).**

gracing the *Monthly* since January 1972. (Possibly this can help in determining who was responsible for the foul deed—it certainly does not seem to me that the departure of the old dear was through natural causes, or accidental.)

Since revival seems impossible, at this stage I decided to try to extract at least a measure of revenge by writing this short note, and illustrating it with some of the gruesome pictures which I have collected over the last few years. They document, better than thousands of words could, the low level of common knowledge of Geometry and the lack of commonsense feeling for graphical rendition of spatial figures. In contrast to the examples given above—in which the lack of accuracy could be blamed on the graphic artist, who might have "corrected for better artistic effect" the original sketches—in the illustrations that follow the responsibility for the grotesque must squarely rest on the authors.

Since (for the uninitiated) looking at icosahedra can become tedious after the first few dozen, let me show here just two from my rich harvest of juicy "plums" (see FIGURE 3). From looking at the representations of icosahedra in various books one could easily conjecture a metatheorem to the effect that "regular icosahedra cannot be drawn correctly"; however, that conjecture is readily disproved (see p. 15). The truly frightening aspect of the situation is not that icosahedra are often grossly misrepresented and that hardly anybody notices, but that such fallacious images *pervade* our books—even those texts that are in other respects quite good. This goes for materials meant for elementary or high schools and their teachers, as much as for those intended for college students, research, or the general public (FIGURE 4)—and the number of examples could be increased almost indefinitely. I would like to stress that I have *not* been spending my time looking for errors, but just tried (with only partial success) to remember those that I encounter and notice.

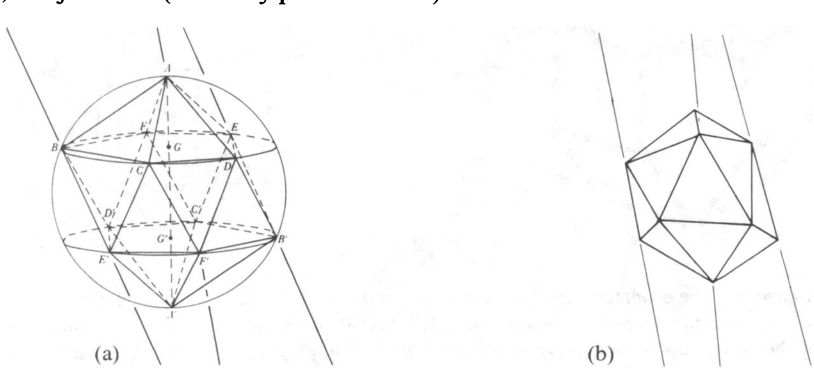


FIGURE 3. **The icosahedra from (a) Mueller [10], page 257, and (b) Firby & Gardiner [5], page 84. As in FIGURE 2, the lines outside the icosahedra should be either parallel, or else concurrent.**

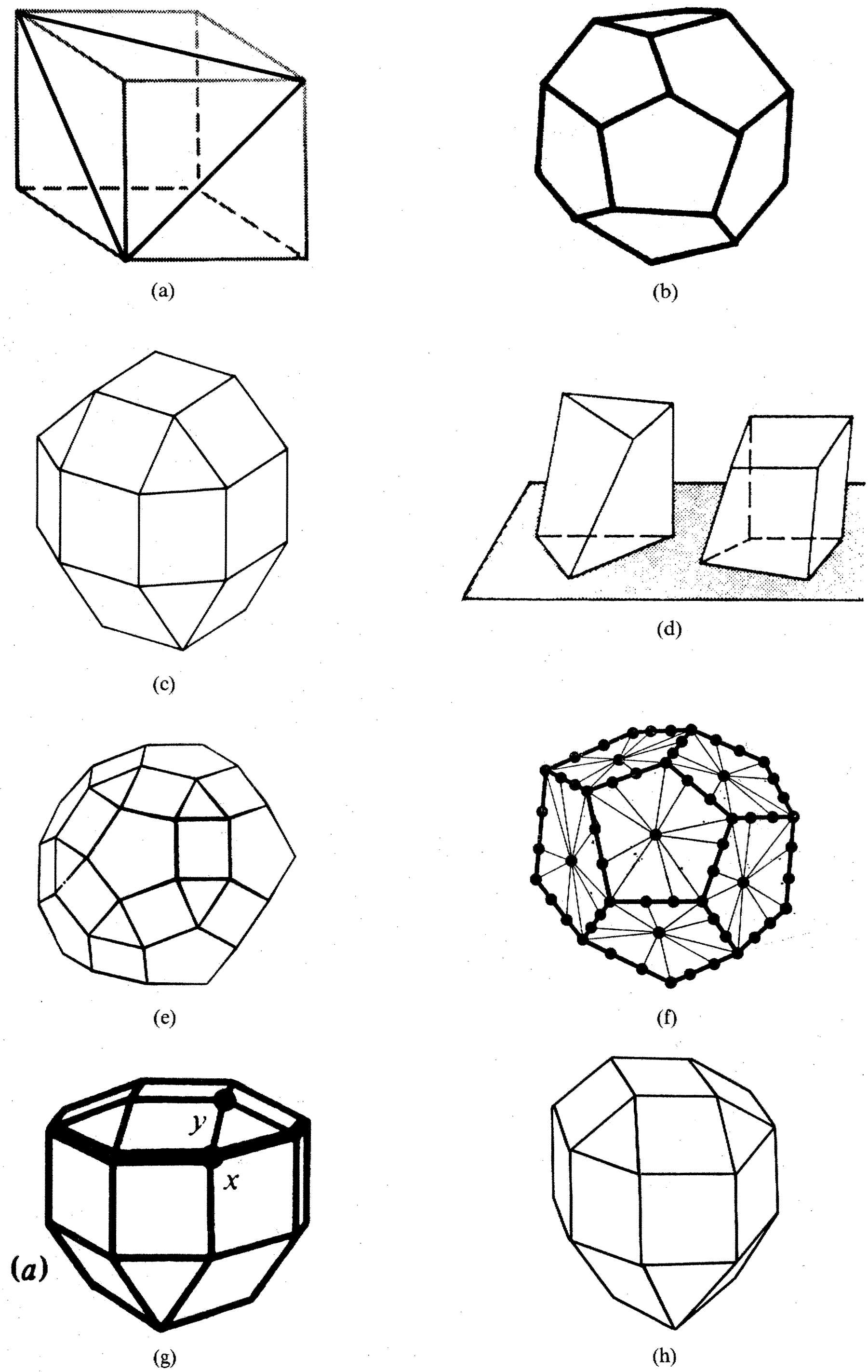
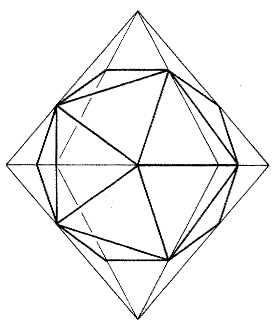


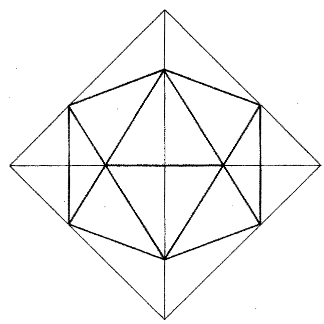
FIGURE 4. (a) A cube, and the equilateral triangle determined by three of its vertices. In the drawing, the edges of the triangle which are supposed to be foreshortened are actually longer than the edge seen full size. From Jacobs [7], page 85. (b) A view of the regular dodecahedron, from Young & Bush [14], page 101. (c) The rhombicuboctahedron, from O'Daffer & Clemens [11], page 141. (d) Two prisms, from Peterson [12], page 77. (e) The rhombicosidodecahedron, from Fejes Tóth [4], page 111. (f) A graph derived from that of the regular dodecahedron, from Lovász [8], page 434. (g) and (h) Two pseudo-rhombicuboctahedra, from Baglivo & Graver [1], page 74, and Martin [9], page 208.

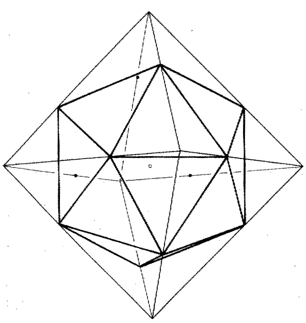
Drawing a Regular Icosahedron

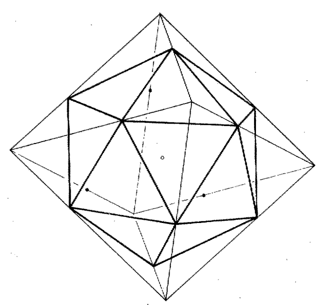
In her comments on a first version of this note, Prof. Schattschneider suggested that I include a *correct* drawing of the regular icosahedron. While it is not hard to find quite accurate drawings in various books, very few of them divulge how this can be done. My favorite method (which can easily be applied in various projections) is based on the observation that the vertices of a regular icosahedron can be placed one on each edge of a regular octahedron so as to divide the edge in the golden ratio $\tau:1$, where $\tau = (\sqrt{5}+1)/2 = 1.618034\ldots$. Since projections of the regular octahedron are easily drawn, this gives a handy means of finding various representations of the icosahedron. (It also can be used as an illustration of the various "dry" theorems of linear algebra, providing very useful practice material—especially if turned into a program on a programmable calculator or a computer, and connected to a graphics display or plotter.) The following are several views of the regular icosahedron obtainable as orthogonal projections in different directions. Skew parallel projections (which are frequently used in mathematical and other texts) can also be obtained in this manner. But although they are useful in various situations and easily drawn, I believe that they are only a poor substitute for a properly executed orthogonal projection.

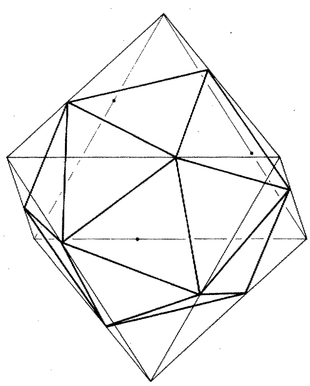
Editor's note: The view of the icosahedron on our cover was obtained from the orthogonal projection shown in the lower right corner, below. The simple directions to an artist for drawing it are: draw a regular hexagon and then locate the three points (such as B) for which the ratio OB/OA equals $\tau = 1.618\ldots$.

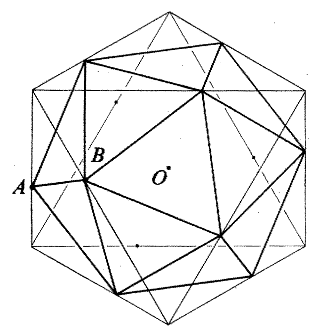




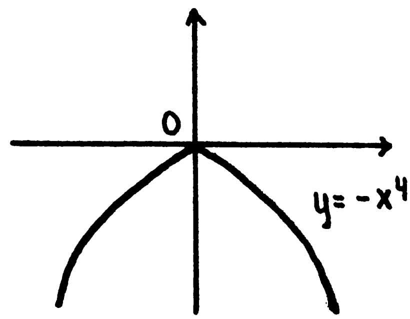


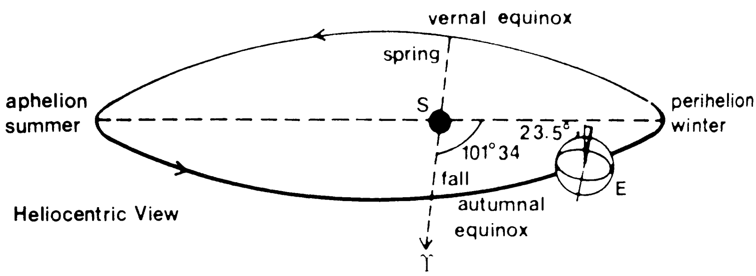






(a)





(b)

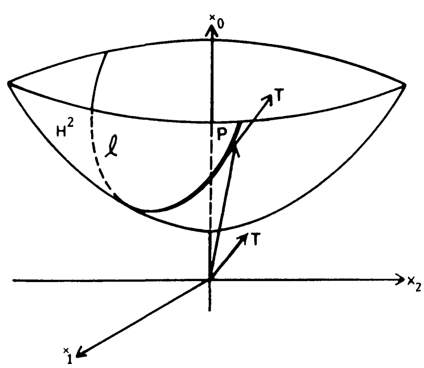


FIGURE 5. **(a) The graph of a very smooth curve, from Gill [6], page 75. (b) Two ellipses, from Faber [3], pp. 84, 263.**

It may be worth mentioning that it is not just polyhedra which get such cavalier treatment (although it is easiest to detect there). From beginning calculus through foundations of geometry to linear algebra and beyond—our texts are full of pictures (see FIGURE 5) that should not have been permitted to see the light of the day and should be banned from U.S. mails. (Some of the effects are almost amusing; had it been created on purpose, the illustration in FIGURE 5(d) would deserve a place with the Penrose bolts and Escher's staircase.)

Concerning a search for causes, I think that the pervasive phenomenon of such misleading illustrations can probably be best explained by assuming a conspiracy involving the adherents of axiomatics, and those who believe that geometry is only linear algebra clumsily done. They certainly have the motive—and now they can point to this little note and say: "See, even the dyed-in-the-wool geometers say you cannot rely on the pictures you see in books."

Time is running out. Unless there is a strong general outcry against the continued visual abuse of the few remaining specimens, the endangered species Homo Geometricus will surely vanish. That will be the end of a long era, and YOU will all be poorer for it... But meanwhile—if you plan to write a book with abominable illustrations, or if you aim to outsmart us by having your next geometry text contain no drawings—watch out: we may be on our last legs, but we are still kicking and will try to continue exposing visual skulduggery.

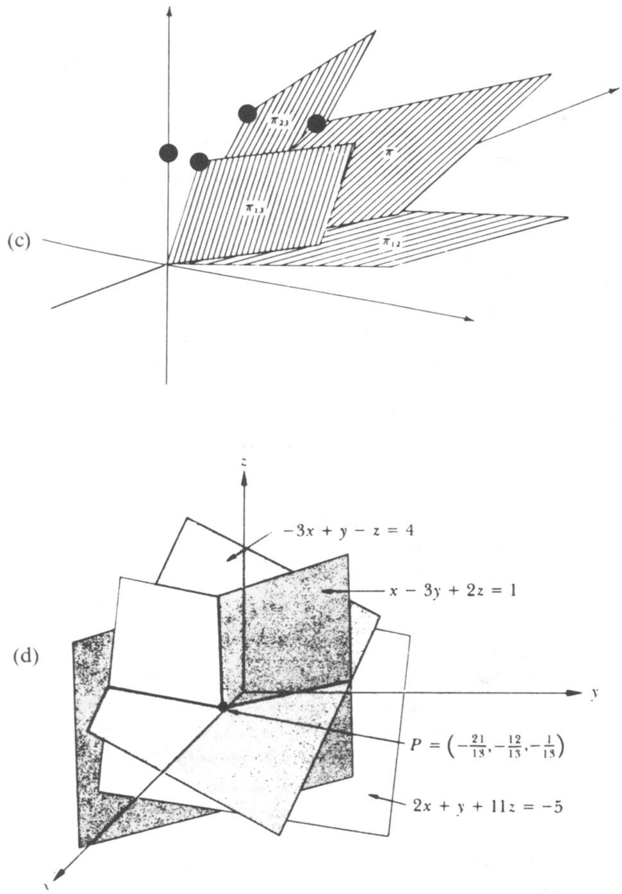


FIGURE 5. **(c) A parallelogram π spanned by two vectors in 3-space, and its parallel projections on the coordinate planes, from Banchoff & Wermer [2], page 123. I used the solid dots to mark points which should be the vertices of a parallelogram. (d) The remarkable intersection of three unremarkable planes, from Shields [13], page 6.**

References

[1] J. A. Baglivo and J. E. Graver, Incidence and Symmetry in Design and Architecture, Cambridge University Press, London & New York, 1983.

[2] T. Banchoff and J. Wermer, Linear Algebra Through Geometry, Springer-Verlag, New York, 1983.

[3] R. L. Faber, Foundations of Euclidean and Non-Euclidean Geometry, Marcel Dekker, New York, 1983.

[4] L. Fejes Tóth, Regular Figures, Pergamon Press, New York, 1964.

[5] P. A. Firby and C. F. Gardiner, Surface Topology, John Wiley & Sons, New York, 1982.

[6] G. S. Gill, Solutions Manual for "Calculus and Analytic Geometry" (5th ed.) by G. B. Thomas, Jr. and R. L. Finney, Addison-Wesley, Reading, Mass., 1979.

[7] H. R. Jacobs, A Teacher's Guide to Geometry, W. H. Freeman and Co., San Francisco, 1974.

[8] L. Lovász, Combinatorial Problems and Exercises, North-Holland, Amsterdam, 1979.

[9] G. E. Martin, Transformation Geometry, Springer-Verlag, New York, 1982.

[10] I. Mueller, Philosophy of Mathematics and Deductive Structure in Euclid's Elements, The MIT Press, Cambridge, Mass. and London, 1981.

[11] P. G. O'Daffer and S. R. Clemens, Geometry: An Investigative Approach, Addison-Wesley, Menlo Park, California, 1976.

[12] J. C. Peterson, Informal geometry in grades 7–14, Geometry in the Mathematics Curriculum, 36th Yearbook, National Council of Teachers of Mathematics, 1973, pp. 52–91.

[13] P. C. Shields, Elementary Linear Algebra, 3rd ed. Worth, New York, 1980.

[14] J. E. Young and G. A. Bush, Geometry for Elementary Teachers, Holden-Day, San Francisco, 1971.

The Mystery of the MAA Logo

Several questions naturally arose in the correspondence concerning the rendering of the icosahedron in the MAA logo: *When* was the logo (the seal on the contents page of this MAGAZINE) adopted? *Why* was the icosahedron chosen (and not, for instance, the pentagram of the Pythagoreans, or the cone and cylinder of Archimedes)? Was the icosahedron *ever* accurately rendered? Simple research has turned up virtually no answers—it would appear that mathematicians are uninterested in leaving records to satisfy our curiosity.

The MAA was founded on December 31, 1915, and was incorporated in September 1920. The reasons for its establishment, as well as its constitution, by-laws, and list of charter members can all be found in issues of the *American Mathematical Monthly* for the years which surround 1915. Kenneth May's history, *The Mathematical Association of America: Its's First Fifty Years*, MAA, 1972, and the special 50th anniversary issue of the *Monthly* (v. 74 II (1967)) are also good references, but none of these accounts mentions the adoption of the now-familiar logo with an icosahedron surrounded by a circular band containing the official name THE MATHEMATICAL ASSOCIATION OF AMERICA. (The November 1920 issue of the *Monthly* records the incorporation of the MAA, in Illinois, and states that the corporate seal of the Association "shall have inscribed thereon the name of the Association and the words 'Corporate Seal-Illinois'"—this is not the same as the MAA logo.)

The first occurrence of the logo with the icosahedron that we could find was on the cover and title page of the first Carus Monograph, *Calculus of Variations*, by G. A. Bliss, published in 1925 by the MAA. It continued to appear in that manner on all subsequent Carus Monographs. (A. Willcox and M. Callanan, at the Washington office of the MAA, searched early minutes of the meetings in which the Carus Monograph series was planned, but found no mention of the creation of the logo to appear on the title page.) Most of us think of the logo today as "always" being part of the cover of the *Monthly*, but it made its first appearance there in January, 1942, when Lester R. Ford assumed the editorship. These are the best answers that we could find to the "when?" question; there was not even a hint of an answer to the "why?" question.

Accuracy? Although the logo was redrawn around 1971 (because the old plates were in bad shape), and the rendering (by an artist) may have become 'worse' from the standpoint of descriptive geometry, careful scrutiny of an enlargement of the logo on the 1925 Carus Monograph reveals the same flaw pointed out by Grünbaum in FIGURE 2. The carefully executed representation of an icosahedron which is featured on our cover will now become the master for all new renderings of the MAA logo.

The best irony in our attempt to find answers to the various questions about the logo was to discover in the *Monthly* (v. 31, 1924, pp. 157–158), in the year in which the first Carus Monograph was prepared, the following account of a talk entitled "The nature and function of descriptive geometry", given by Professor W. H. Roever at the annual MAA meeting held in Cincinnati in 1923:

> In the paper by Professor Roever the principal purposes of the subject [descriptive geometry] were defined to be (1) representation of the objects of space by means of figures which lie in a plane (or upon a surface), and (2) solution of the problems of space by means of constructions which can be executed in the plane. It was shown that the requirement for a picture to be adequate, i.e., capable of producing a retinal image differing but little from that produced by the object itself, naturally leads to the use of central or parallel (orthographic or oblique) projection as a means of representation; and also that the requirement for an unambiguous correspondence between space and the plane results in the need for two projections or for one projection with information concerning the object (such, for instance, as the perpendicularity of edges).
>
> ...
>
> Finally, need for the production of good pictures in books on mathematics was stressed, and the value of the study of descriptive geometry as a means of developing the power of space visualization was emphasized. The reader interested in a brief account of the subject is referred to the author's paper entitled "Descriptive geometry and its merits as a collegiate as well as an engineering subject" published in the MONTHLY (1918, 145–159).
>
> Professor Bradshaw spoke of a recent book on calculus in which is found a so-called "Standard figure of the ellipsoid", a critical examination of which from the standpoint of descriptive geometry reveals three different directions of projection....

Plus ça change, plus c'est la même chose!

—DORIS SCHATTSCHNEIDER

Thirty Days Hath February

Bernard E. Rodden
Defence and Civil Institute of Environmental Medicine
Downsview, Ontario, Canada M3M 3B9

Algorithms pertaining to the Gregorian calendar system provide a pool of practical classroom problems for lectures in computation. Everyone has a common sense appreciation of the calendar, and associated algorithms have sufficient scope to be entertaining without being intellectually demanding. Unfortunately, it is sometimes difficult to assess the quality of a given algorithm, as criteria of "goodness" are typically unstated, and left to the reader's imagination (on the mistaken assumption that such criteria are self-evident).

A typical date algorithm is the production of a yearly calendar for any given year. Can reasonable people agree on what are *good* properties for a calendar algorithm? I suggest that the most obvious good property is the restriction to integer arithmetic. As the calendar is defined by integers, we should not have to leave the integer system to solve date problems. Computers work faster with integers; and integer algorithms are easily implemented in firmware. Not so obvious is the property that dates should be *generated* rather than be dependent on tabulated information. *Tables* in this sense include such computer language constructs as *arrays*, *structures*, *records*, *cases*, and *types*. Tables are usually relatively awkward and slow to manipulate, error prone, and difficult to transport from one system or language to another. More to the point, tables only assist in a mapping from one set of integers to another—an operation that can often be done more elegantly and efficiently by an algorithm. Many date algorithms presently in use apply only to a portion of the full period of the calendar. Since the full period is only 28,000 years, surely it is not too demanding to insist that date algorithms be effective over the entire cycle. The final good property of an algorithm is the ease with which it can be coded efficiently using modern computing languages.

In this Note, I shall describe a calendar generator based on ideas contained in an Algol procedure by R. A. Stone [**1**]. Stone used three particular integer parameters to map the days of a standardized year onto the days of the month. I use his three-parameter formulation, the standardized year, and his emphasis on tableless algorithms. In addition, I extend his three parameters to an infinite class, develop an inverse to his function, and show that the method can be used for any date problem.

The Gregorian calendar as it is today (with the exception of the 4000-year rule) was introduced in a papal bull dated March 1, 1582 [**2**]. If we extrapolate the current calendar backwards, this date would be March 11, because the *new style* calendar was to become effective on October 15 1582, replacing the *old style* date of October 5. The weekday sequence was unaltered. The only significant change from a mathematical point of view was the introduction of a rule excluding century years from the set of leap years, unless they were divisible by 400. The effect of these exclusions was to reduce the accumulation of error with respect to the solar year from one day every 128 years (approximately) to one day every 3323 years. This modification was fortuitous because a 400-year interval (146,097 days) is divisible by 7, and hence the period of calendar dates coincided with the approximation to the astronomic period. The most recent, and likely final, proposed correction fits the period of the earth's revolution to a 4000-year interval consisting of 1,460,969 days, which is not divisible by 7, and hence the calendar period is 28,000 years. The 4000-year rule reduces the error to one day every 20,000 years.

The (year, month, day) mixed radix system of the calendar is more complex than some other mixed systems, such as (pounds, shillings, pence), because the radix for the number of days is not fixed, but is a function of both month and year. An additional complication is the weekday cycle. There are 7 weekdays and two possible types of year, leap and non-leap, providing 14 distinct, yearly calendars.

A calendar can be generated for any year y by the following sequence of operations. Let (y,m,d) describe day d of month m of year y. By definition, $(y,1,1)$ is a valid date. The weekday corresponding to $(y,1,1)$ can be deduced from application of the weekday theorem, which we develop later. Let $(y,m,d)^*$ be the day subsequent to (y,m,d). If (y,m,d) is a valid date, then $(y,m,d)^*$ is the first valid member of the ordered list $\{(y,m,d+1),(y,m+1,1),(y+1,1,1)\}$. As the third member is always valid, there is always a valid subsequent date. Hence we only require a criterion for validating a given date. The criterion may be expressed as

$$(y,m,d) \textit{ is a valid date if and only if } 0<d\leqq L(m,q),$$

where m is a valid month number, year y is of type q, and $L(m,q)$ is the number of days in month m of year type q.

If we can find a functional form of L, along with the weekday function, we will have a tableless calendar generator.

The integer y represents a leap year if and only if y is a multiple of 4 but not of 100, or y is a multiple of 400 but not of 4000. Although this definition of a leap year seems a natural verbal description, it can lead to inefficient algorithms, since it does not take full advantage of the probabilistic ordering of the various conditions. Simplistic implementation results in the logical function

$$q(y)=[D_4(y)\cap\overline{D}_{100}(y)]\cup[D_{400}(y)\cap\overline{D}_{4000}(y)] \tag{1}$$

where $D_n(y)$ is *true* if y is a multiple of n, and *false* otherwise, and $\overline{D}_n(y)$ denotes the negation of $D_n(y)$. Hence q is *true* if y is a leap year, and *false* otherwise, as shown in the Venn diagram of FIGURE 1.

In this algorithm, the D calculations are computationally relatively expensive, and we shall show that despite partial optimization, the simplistic form (1) increases the expected number of D calculations about 50% with respect to a most efficient algorithm. We define a *most efficient* algorithm as one that requires the smallest expected number of D calculations for a year selected randomly from a full cycle of the calendar. To design such an algorithm, we take advantage of the sequence of logical implications,

$$D_{4000}\subset D_{400}\subset D_{100}\subset D_4$$

by so ordering the sequence of logical operations of the algorithm for $q(y)$ that those events that most probably determine q are calculated first. For example, the *truth* of D_{4000} establishes immediately the type of year, but D_{4000} is a rare event; the *falsity* of D_4 also establishes the type of year, and is the most probable event that does so. The efficient order of calculation is thus:

$$\overline{D}_4,\ \overline{D}_{100}|D_4,\ \overline{D}_{400}|D_{100},\ \overline{D}_{4000}|D_{400},$$

where $\overline{D}_m|D_n$ means $\overline{D}_m$ given D_n.

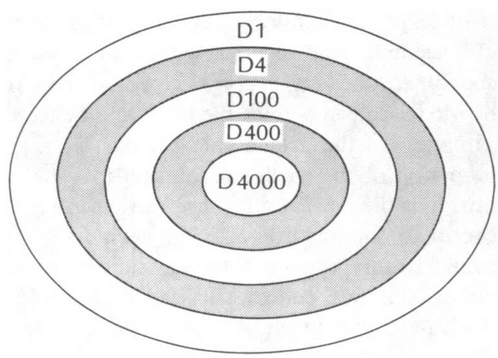


FIGURE 1. **Venn diagram of years.**

In a corresponding algorithm

$$q(y) = D_4 \cap [\overline{D}_{100} \cup (D_{400} \cap \overline{D}_{4000})] \tag{2}$$

calculation is from left to right, and ceases when the truth or falsity of the expression has been established. Hence for the determination of q, using (2), the expected number of D calculations is

$$\begin{aligned} p(\overline{D}_4)\cdot 1 + p(D_4)\cdot[p(\overline{D}_{100}|D_4)\cdot 2 + p(D_{100}|D_4)\cdot[p(\overline{D}_{400}|D_{100})\cdot 3 + p(D_{400}|D_{100})\cdot 4]] \\ = \frac{3}{4}\cdot 1 + \frac{1}{4}\cdot\left(\frac{24}{25}\cdot 2 + \frac{1}{25}\cdot\left(\frac{3}{4}\cdot 3 + \frac{1}{4}\cdot 4\right)\right) \\ = 1.2625, \end{aligned}$$

where $p(X|Y)$ means the conditional probability of event X given event Y.

The corresponding expansion for the simplistic form (1) gives an expected number of D calculations

$$\begin{aligned} &[p(D_4)\cdot p(\overline{D}_{100}|D_4) + p(\overline{D}_4)\cdot p(\overline{D}_{400}|\overline{D}_4)]\cdot 2 \\ &\quad + p(D_4)\cdot p(D_{100}|D_4)\cdot[p(\overline{D}_{400}|D_{100})\cdot 3 + p(D_{400}|D_{100})\cdot 4] \\ &= \left(\frac{1}{4}\cdot\frac{24}{25} + \frac{3}{4}\cdot 1\right)\cdot 2 + \frac{1}{4}\cdot\frac{1}{25}\cdot\left(\frac{3}{4}\cdot 3 + \frac{1}{4}\cdot 4\right) \\ &= 2.0125. \end{aligned}$$

Many computer languages do not interpret the left-to-right logical operations in the efficient manner described. For such languages, the coder must break down the algorithm in the obvious way, and branch out explicitly. Some languages, such as C and Modula, allow direct expression.

To calculate the day of the week on which an arbitrary date falls, we first compute the number of leap years between the given date and a standard date. To simplify the calculation of the number of leap year days between two arbitrary dates, we define **the number of leap year days** $n(y)$ to be the *algebraic* difference between year 0 and date $(y,1,1)$, as given by the following formulas.

$$\text{If } y > 0, \quad n(y) = \left\lfloor \frac{y-1}{4} \right\rfloor - \left\lfloor \frac{y-1}{100} \right\rfloor + \left\lfloor \frac{y-1}{400} \right\rfloor - \left\lfloor \frac{y-1}{4000} \right\rfloor \tag{3}$$

$$\text{If } y < 0, \quad n(y) = \left\lceil \frac{y}{4} \right\rceil - \left\lceil \frac{y}{100} \right\rceil + \left\lceil \frac{y}{400} \right\rceil - \left\lceil \frac{y}{4000} \right\rceil, \tag{4}$$

where $\lfloor x \rfloor$ is the floor function, that is, $\lfloor x \rfloor$ is the greatest integer not greater than the number x, and where $\lceil x \rceil$ is the ceiling function, that is, $\lceil x \rceil$ is the least integer not less than the number x. There is no computational complexity in using the floor function for positive quotients and the ceiling function for negative quotients, as this is the definition of integer division for most computer languages. As an example, the number of leap year days between the dates $(-4,1,1)$ and $(4,1,1)$ is $n(4) - n(-4) = 0 - (-1) = 1$.

We are now able to calculate the weekday corresponding to the first day of any year.

THE WEEKDAY THEOREM. *Year y begins on weekday w if*

$$w = (n + 365y)_7, \quad \text{or } w = (n + y)_7, \quad \text{or } w = (n + (y)_7)_7, \tag{5}$$

where $(x)_7$ is the member of the set $\{0,1,\ldots,6\}$ congruent modulo 7 to x, and where $n = n(y)$ is the number of leap years defined by (3) *and* (4).

The first formula of the weekday theorem expresses the definition of a weekday. The second formula saves an unnecessary multiplication by recognizing that $(365)_7 = 1$. The third formula is convenient for computations that use 16 bit registers and 2's complement arithmetic. For such computation, if y is restricted to the range $y < |28{,}000|$, then at the maximum value of y, $n = 6783$ and $y + n = 34{,}782$, which overflows, since the arithmetic is restricted to values less than 32,768, the 15th power of 2. The restriction on y does not affect the universality of the calendar generator, since any year outside the range can be reduced modulo 28,000 prior to computation.

To avoid the idiosyncracies of leap years, we propose, as a first approximation to the actual yearday, a **standardized yearday** s that is a function of (m, d) only,

$$s = S(m) + d - 1 \tag{6}$$

where S is a function of m only. In this definition, $(y, 1, 1)$ corresponds to standardized yearday 0 and actual yearday 0. Once s has been determined, the **actual yearday**, a, is given by

$$a = s + H(q, s), \tag{7}$$

where q is the kind of year, and $H(q, s)$ is a correction term. For computational purposes, we take $q = 0$ for non-leap years, and $q \neq 0$ for leap years.

A candidate for S is

$$S(m) = \left\lfloor \frac{\gamma(m-1) + \beta}{\alpha} \right\rfloor$$

where α, β, γ are integer parameters, selected so that

$$L(m, q) = S(m+1) - S(m)$$

for all months except $m = 2$ (February), whose length clearly depends on q, and must include the correction function H. The closest fit for $m = 2$ gives $S(3) - S(2) = 30$. Using this fit, a possible choice for H is

$$H(q, s) = -2 \cdot I_{\bar{q}} \cdot I_{s > 58} - I_q \cdot I_{s > 59},$$

where I is the indicator function; I_p is 1 if p is *true* and 0 if p is *false*.

We finally have a functional form for month length,

$$L(m, q) = S(m+1) - S(m) - I_{m=2}(2I_{\bar{q}} + I_q). \tag{8}$$

The fact that $m + 1$ is invalid for $m = 12$ merely imposes a restriction on the scope of S.

Like Stone, we have not been completely successful in our attempt to find a tableless method, as the indicator functions are equivalent to a table, albeit a small one. The introduction of two extra days in February is the only algorithmic trick; hence the title of this paper.

The inverse function of S is

$$M(s) = \left\lfloor \frac{as + b}{g} \right\rfloor,$$

where a, b, g are found in the same way as α, β, γ in the case of $S(m)$. $M(s)$ transforms any day of the year into the corresponding month. The inversion is not required for calendar generation, but is useful in many date problems.

It was mentioned previously that $S(13)$ is required for month length calculation. Hence S places 13 constraints on α, β, γ:

$$S(1) = 0; \qquad S(2) = 31; \qquad S(3) = 61; \ \ldots \ ; \ S(13) = 367.$$

Similarly, the inversion of S places 24 constraints on a, b, g:

$$M(0) = M(30) = 1; \qquad M(31) = M(60) = 2; \ \ldots \ ; \ M(336) = M(366) = 12.$$

Aided by a computer, we solve by simple enumeration. The smallest $[\alpha, \beta, \gamma]$ is [7,3,214], a solution that is also valid for the inverse. For languages and machines that allow shifting of operands, it is efficient to do integer multiplication and division by shifting when possible; hence we recommend values [16,8,489] for $[\alpha, \beta, \gamma]$, and [16,6,489] for $[a, b, g]$.

We now have all the theory required to find the weekday W corresponding to any given date: add the weekday corresponding to the first day of the year to the actual yearday and reduce the result modulo 7, giving

$$W(y, m, d) = (w(y, n) + a(s, q))_7, \tag{9}$$

where w, n, a, s, q have been defined in formulas (2) to (6).

As an example of the use of these formulas, we shall determine the day of the week corresponding to the fifteenth of October 1582, the date on which the new style calendar was inaugurated. Because dates are such a vital and familiar item of modern human activity, most people can at a glance deduce that the given date is valid, and that 1582 is not a leap year. A computer however must go through formal algorithms to do so, as discussed earlier. Having decided that the date is valid, the computer then determines the day of the week corresponding to the first day of the year. Substituting in formula (3), it finds that there have been 383 leap years since year 0. Substituting in (5), 1582 started on weekday 5, a Friday. Formula (6) gives 289 for the standardized day of the year. Formula (7) gives 287 for the actual day of the year. Finally, formula (9) gives the required weekday 5, also a Friday. Readers can check the result against the historic old style calendar used by England and her colonies. They will find that October 15, 1582 corresponds to a Monday, and subtracting the ten deleted days gives Friday, conforming to our result.

The preceding theory encourages exceedingly simple computer algorithms when the algorithms are expressed in a reasonably rich higher level computer language (such as C, Pascal, Algol). An obvious application is the generation of calendars, but more important is the efficient storage and retrieval of dates in data bases, as almost all data base entries are dated. Dates can be stored as *century day numbers* (cdn), with cdn 1 corresponding to the first day of this century, and cdn 36,524 corresponding to the final day. Hence storage requirement is reduced significantly, as the cdn is well within the capacity of a 16-bit register for all dates of this century. Alternatively we could use an *absolute day number* (adn), where adn 0 corresponds to the first day of century 0, and adn 10,226,782 corresponds to the final day of the 28,000-year cycle. Age calculations become remarkably simple, as age in days is the simple difference between two adn's or two cdn's. Absolute day numbers allow a simpler form of the weekday function, but require 24-bit arithmetic.

Various countries adopted the Gregorian calendar at different times. Dates for pre-adoptive eras differ from those of the extrapolated Gregorian calendar. Even for these cases, the suggested method is useful, as it provides a convenient standard that can easily be corrected to apply to any particular era.

References

[1] R. A. Stone, Algorithm 398, collected algorithms, CACM (1970).
[2] Encyclopaedia Britannica, 1968.

Measure Theory

We don't take up much space—
  $6 \times 3 \times 2$, roughly.

We don't take up much time—
  60—70—80...
  a miniscule span, as time goes.

But there is something within us
  exterior to space,
  anterior to time,
  of no recognizable cardinal number.

—KATHARINE O'BRIEN

where w, n, a, s, q have been defined in formulas (2) to (6).

As an example of the use of these formulas, we shall determine the day of the week corresponding to the fifteenth of October 1582, the date on which the new style calendar was inaugurated. Because dates are such a vital and familiar item of modern human activity, most people can at a glance deduce that the given date is valid, and that 1582 is not a leap year. A computer however must go through formal algorithms to do so, as discussed earlier. Having decided that the date is valid, the computer then determines the day of the week corresponding to the first day of the year. Substituting in formula (3), it finds that there have been 383 leap years since year 0. Substituting in (5), 1582 started on weekday 5, a Friday. Formula (6) gives 289 for the standardized day of the year. Formula (7) gives 287 for the actual day of the year. Finally, formula (9) gives the required weekday 5, also a Friday. Readers can check the result against the historic old style calendar used by England and her colonies. They will find that October 15, 1582 corresponds to a Monday, and subtracting the ten deleted days gives Friday, conforming to our result.

The preceding theory encourages exceedingly simple computer algorithms when the algorithms are expressed in a reasonably rich higher level computer language (such as C, Pascal, Algol). An obvious application is the generation of calendars, but more important is the efficient storage and retrieval of dates in data bases, as almost all data base entries are dated. Dates can be stored as *century day numbers* (cdn), with cdn 1 corresponding to the first day of this century, and cdn 36,524 corresponding to the final day. Hence storage requirement is reduced significantly, as the cdn is well within the capacity of a 16-bit register for all dates of this century. Alternatively we could use an *absolute day number* (adn), where adn 0 corresponds to the first day of century 0, and adn 10,226,782 corresponds to the final day of the 28,000-year cycle. Age calculations become remarkably simple, as age in days is the simple difference between two adn's or two cdn's. Absolute day numbers allow a simpler form of the weekday function, but require 24-bit arithmetic.

Various countries adopted the Gregorian calendar at different times. Dates for pre-adoptive eras differ from those of the extrapolated Gregorian calendar. Even for these cases, the suggested method is useful, as it provides a convenient standard that can easily be corrected to apply to any particular era.

References

[1] R. A. Stone, Algorithm 398, collected algorithms, CACM (1970).

[2] Encyclopaedia Britannica, 1968.

Measure Theory

We don't take up much space—
  6 × 3 × 2, roughly.

We don't take up much time—
  60—70—80...
  a miniscule span, as time goes.

But there is something within us
  exterior to space,
  anterior to time,
  of no recognizable cardinal number.

—KATHARINE O'BRIEN

Constructing a Minimal Counterexample in Group Theory

ARNOLD D. FELDMAN
Franklin and Marshall College
Lancaster, PA 17604

Desmond MacHale's intriguing article [**2**] neatly illustrates one mode of interplay between proof and example in group theory by describing the commonly used inductive technique of minimal counterexamples. In this Note, we use some ideas from group theory and linear algebra to lead the reader through a geometric and algebraic construction of one of the minimal counterexamples sought by MacHale.

Recall that a group is called **nilpotent** if each of its Sylow subgroups is normal. One of MacHale's list of known-to-be-false conjectures is this: if G is a finite group having a fixed-point-free automorphism ϕ, then G is nilpotent. The true theorem on which the conjecture is based assumes that ϕ is of prime order; it was established by Thompson in his dissertation in 1959 [**4**]. His proof so excited the mathematical community that even the New York Times reported his result and the reaction to it [**3**]. MacHale indicates that [**1**, p. 336] contains details of a counterexample to the conjecture with $|G| = 147 = 7^2 \cdot 3$ and $|\phi| = 4$. In our minimal counterexample to be constructed, $|\phi| = 6$ and $|G| = 48 = 2^4 \cdot 3$.

In what follows, we use some standard conventions of notation in group theory. In particular, we write operators on the right and denote both the action of a group homomorphism and conjugation using superscripts.

Recall that if H and K are subgroups of a group G such that $G = HK$, $H \cap K = 1$, and H is normal in G, then G is called the **semidirect product** of H by K. In this situation, since H is normal in G, each element of K acts via conjugation as an automorphism of H. Of particular interest here is the situation in which each nontrivial element of K acts as a nontrivial (i.e., nonidentity) automorphism on H. In this case K is isomorphic to a subgroup of Aut H, the group of all automorphisms of H.

A well-known geometric example leads directly to an important link between group theory and geometry. If A is the group of rotations of a regular tetrahedron, then A contains eight elements of order 3, three of order 2, and an identity element. Each element of order 3 is a 120°rotation of the tetrahedron about an axis passing through one of the 4 vertices and perpendicular to the opposite face. Each element of order 2 is a rotation of 180° about one of three axes joining the midpoints of two nonadjacent edges of the tetrahedron. (See FIGURE 1.) If the vertices are labeled, and each rotation is identified with the permutation of the labels it produces, this identification provides a natural isomorphism between A and A_4, the group of even permutations on 4 letters. Thus it is possible to use the algebra of permutations or geometry to analyze A.

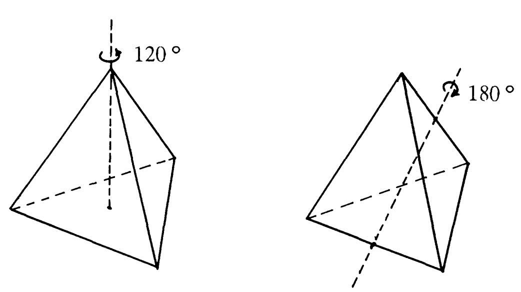


FIGURE 1

It is easy to see that A has a unique subgroup V of order 4, consisting of the identity element and the three elements of order 2. Thus if $x \neq 1$ is an element of V and σ is an element of order 3 in A, $x^{\sigma} = \sigma^{-1}x\sigma = y$ is in V and it is simple to check, using geometry or algebra, that $x \neq y$. Hence $V = \{1, x, y, xy\}$. Since conjugation by σ fixes only the identity element of V, we say that σ is a **fixed-point-free** automorphism of V. Using $\langle \sigma \rangle$ to denote the cyclic group of order 3 generated by σ, we have $A = V\langle \sigma \rangle$, the semidirect product of V by $\langle \sigma \rangle$.

We can learn more about the automorphism σ by rewriting V as an additive group and considering the 2×2 matrix M_{σ} associated with σ. This we do as follows:

Identify V with $(Z_2)^2$, the direct product of two additive groups of order 2. Identify x with $(1,0)$ and y with $(0,1)$, so that xy corresponds to $(1,1)$ and 1 corresponds to $(0,0)$. Thus in the semidirect product A, $(1,0)^{\sigma} = (0,1)$, $(0,1)^{\sigma} = (1,1)$, and $(1,1)^{\sigma} = (1,0)$. Of course, $(0,0)^{\sigma} = (0,0)$. We obtain the matrix M_{σ} associated with σ by using for its rows the images under σ of the "standard" basis elements $(1,0)$ and $(0,1)$ under conjugation by σ. Thus

$$M_{\sigma} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Since $v^{\sigma} = vM_{\sigma}$, $v^{\sigma} = v$ if and only if $v = vM_{\sigma}$. Thus the fact that σ is fixed-point-free is equivalent to the fact that there exists no element v of V other than $(0,0)$ such that $vM_{\sigma} = v$.

The Counterexample

We use more geometry to construct our minimal counterexample of a nonnilpotent group G having a fixed-point-free automorphism. First, let T_1 and T_2 be congruent regular tetrahedra, with the vertices of T_1 labeled 1, 2, 3, 4, and the vertices of T_2 labeled 5, 6, 7, 8. (See FIGURE 2a.) Also, consider the set S of 8 labeled points in 3-space depicted in FIGURE 2b; S consists of two subsets S_1 and S_2, each of which contains the vertices of a tetrahedron the same size as each of T_1 and T_2. For $i = 1$ or 2, if we place T_i on S_i in such a way that the vertices and labels match, we say T_i is in "home position". In FIGURE 2c, both T_1 and T_2 are in home position. In FIGURE 2d, only T_2 is in the home position.

There are 12 ways to place T_1 on S_1 with vertices on vertices, corresponding to the 12 elements of the group of rotations of a tetrahedron. Similarly, there are 12 ways to place T_2 on S_2. Thus there are $12 \cdot 12 = 144$ ways of placing T_1 on S_1 and T_2 on S_2. Similarly, there are 144 ways of placing T_2 on S_1 and T_1 on S_2. Thus there are 288 ways of placing the pair of tetrahedra T_1 and T_2 onto the framework formed by the subsets S_1 and S_2 of S. Each of these 288 positions for T_1 and T_2 corresponds to a permutation of the 8 elements of S, with the home position for both T_1 and T_2, depicted in FIGURE 2c, corresponding to the identity permutation. Any other position for T_1 and T_2 corresponds to the unique permutation of S required to take the vertices of T_1 and T_2 to this position from the home position. For example, the position of T_1 and T_2 in FIGURE 2d corresponds to the permutation (2 3 4).

The set U of permutations of S that corresponds to this set of 288 positions for T_1 and T_2 on S is clearly a subgroup of the group of permutations of S. We will show that U has a nonnilpotent subgroup G of order 48 and an element ϕ of order 6 which is a fixed-point-free automorphism of G via conjugation as an element of U. This G and ϕ provide the minimal counterexample described in the introduction. To construct that counterexample we need to introduce several subgroups of U.

Denote by A_1 the subgroup of elements of U that do not affect S_2. Thus if we start with T_1 and T_2 in home position as in FIGURE 2c and apply an element of A_1, we obtain a result in which T_1 is still on S_1 in one of 12 positions, and T_2 is still on S_2 in home position. Thus the elements of A_1 correspond to the rotations of the tetrahedron T_1, so A_1 is isomorphic to the group $A = V\langle \sigma \rangle$ described above. Similarly, denote by A_2 the subgroup of elements of U not affecting S_1. Thus the elements of A_2 correspond to the rotations of the tetrahedron T_2, and A_2 is also isomorphic to A. It is easy to see that A_1 and A_2 have trivial intersection and that the elements of A_1 commute with the elements of A_2, so U has a subgroup $A_1 \times A_2$. Since the index of $A_1 \times A_2$ in

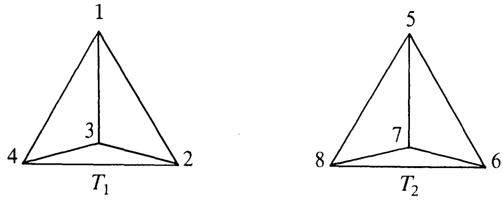


FIGURE 2a

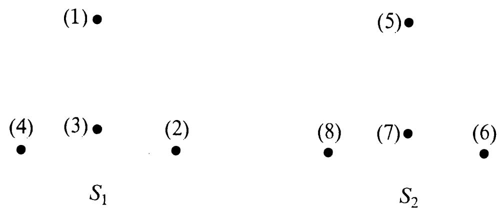


FIGURE 2b

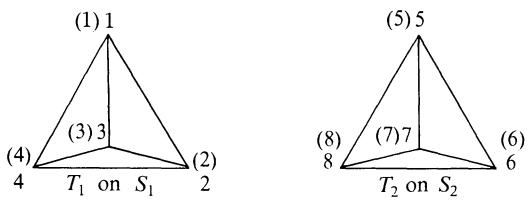


FIGURE 2c

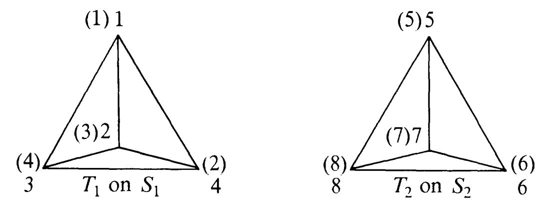


FIGURE 2d

U is 2, $A_1 \times A_2$ is a normal subgroup of U.

There is an element τ of U which corresponds to translating T_1 to the right and T_2 to the left from their respective home positions, so that T_1 is on S_2 and T_2 is on S_1. Thus, after application of τ, the vertex of T_1 labeled j will be at the point of S_2 labeled $(j+4)$, and the vertex of T_2 labeled k will be at the point of S_1 labeled $(k-4)$. Since $\tau^2 = 1$, $\langle \tau \rangle$ is a cyclic group of order 2. The elements of $A_1 \times A_2$ leave T_1 on S_1 and T_2 on S_2, so $A_1 \times A_2$ and $\langle \tau \rangle$ intersect trivially. Thus the semidirect product $(A_1 \times A_2)\langle \tau \rangle$ has order $144 \cdot 2 = 288$, so $(A_1 \times A_2)\langle \tau \rangle = U$.

Just as A has a unique subgroup V of order 4, each A_i has a unique subgroup V_i of order 4. Then $V_1 \times V_2$ is the unique subgroup of order 16 in $A_1 \times A_2$. This uniqueness implies $V_1 \times V_2$ is normal in U, since $A_1 \times A_2$ is normal in U. Therefore, each element of U acts on $V_1 \times V_2$ as an automorphism via conjugation.

Now consider V_1, V_2, and $V_1 \times V_2$ as additive groups, writing $V_1 \times V_2 = \{(x_1, x_2, x_3, x_4) \mid x_1, x_2, x_3, x_4 \in Z_2\}$, where $\{(1,0,0,0),(0,1,0,0)\}$ generates V_1 as a subgroup of $V_1 \times V_2$ and $\{(0,0,1,0),(0,0,0,1)\}$ generates V_2. Then we can represent automorphisms of $V_1 \times V_2$ by 4×4 matrices with entries in Z_2.

For example, let σ_1 be the element of A_1 analogous to the element σ of A described earlier; the position of T_1 and T_2 corresponding to σ_1 is depicted in FIGURE 2d. The analogy to $A = V\langle \sigma \rangle$ yields $(1,0,0,0)^{\sigma_1} = (0,1,0,0)$ and $(0,1,0,0)^{\sigma_1} = (1,1,0,0)$. Also, as an element of A_1, σ_1 commutes with each element of A_2, including the generators $(0,0,1,0)$ and $(0,0,0,1)$ of V_2. Thus $(0,0,1,0)^{\sigma_1} = (0,0,1,0)$ and $(0,0,0,1)^{\sigma_1} = (0,0,0,1)$. We use these images of $\{(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1)\}$ under σ_1 as the rows of the matrix M_{σ_1} for σ_1. Therefore,

$$M_{\sigma_1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Similarly, we let σ_2 be the element of A_2 analogous to σ in A, and obtain the matrix M_{σ_2} for σ_2:

$$M_{\sigma_2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Also $(x_1, x_2, x_3, x_4)^{\tau} = (x_3, x_4, x_1, x_2)$; this fact follows from the way in which τ interchanges the positions of the tetrahedra T_1 and T_2. Thus the matrix M_{τ} for the action of τ is

$$M_{\tau} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

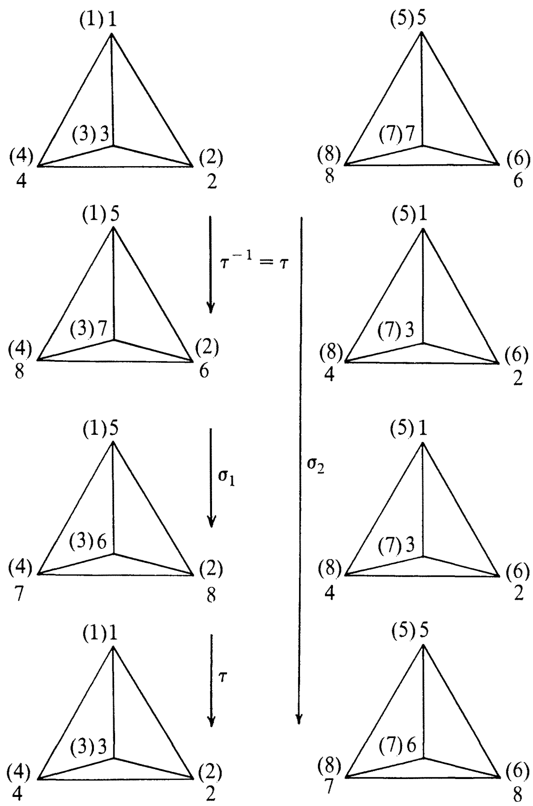


FIGURE 3

These matrices facilitate computation. For instance, the matrix for σ_1^{-1} equals $(M_{\sigma_1})^{-1}$, the inverse of the matrix for σ_1, and the matrix of a product is the product of the matrices: $M_{\sigma_1\sigma_2^{-1}} = (M_{\sigma_1})(M_{\sigma_2^{-1}}) = M_{\sigma_1}(M_{\sigma_2})^{-1}$. The facts we need about individual automorphisms could be verified geometrically, but instead we simply note that such use of matrices to study automorphisms of certain groups is a common technique of considerable power. See [**1**, 1.3.2, 2.6.1] for more detail.

Now we can define G and ϕ. First let $\alpha = \sigma_1\sigma_2^{-1}$ and let $\phi = \sigma_1\sigma_2\tau$. Define G as $(V_1 \times V_2)\langle\alpha\rangle$. G is a subgroup of U since $V_1 \times V_2$ is normal in U. Since G is of order $16 \cdot 3$, $V_1 \times V_2$ is a Sylow 2-subgroup of G and $\langle\alpha\rangle$ is a Sylow 3-subgroup of G. If G were nilpotent, then α would commute with each element v of $V_1 \times V_2$ for the following reason. The nilpotence of G would imply that $V_1 \times V_2$ and $\langle\alpha\rangle$ were normal in G. Thus for v in $V_1 \times V_2$, $(v^{-1}\alpha^{-1}v)\alpha = v^{-1}(\alpha^{-1}v\alpha)$ is in both $V_1 \times V_2$ and $\langle\alpha\rangle$. But the orders of these subgroups are relatively prime, so their intersection is the identity. Thus $v^{-1}\alpha^{-1}v\alpha$ is the identity, and $v\alpha = \alpha v$ as claimed.

If α did commute with each element of $V_1 \times V_2$, then conjugation by α would produce the trivial automorphism of $V_1 \times V_2$, so M_α would be the 4×4 identity matrix. But $\alpha = \sigma_1\sigma_2^{-1}$ implies $M_\alpha = M_{\sigma_1}(M_{\sigma_2})^{-1}$, which is

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

so G is not nilpotent.

Our first step in establishing that the automorphism $\phi = \sigma_1\sigma_2\tau$ fixes only the identity element of $G = (V_1 \times V_2)\langle\alpha\rangle$ is to show that if an element $v\alpha^i$ of G is fixed by ϕ, then α^i is trivial and therefore v in $V_1 \times V_2$ is fixed by ϕ. To do this, we note that $\sigma_1^\tau = \sigma_2$, for σ_2 does to the tetrahedron on S_2 just what σ_1 does to the tetrahedron on S_1. (See FIGURE 3.) Therefore, $\sigma_2^\tau = (\sigma_1^\tau)^\tau = \sigma_1$. Since σ_1 and σ_2 commute, $\alpha^\phi = (\sigma_1\sigma_2^{-1})^{\sigma_1\sigma_2\tau} = \sigma_1^\tau(\sigma_2^{-1})^\tau = \sigma_2\sigma_1^{-1} = \alpha^{-1}$. Thus if α^i is in $\langle\alpha\rangle$, $(\alpha^i)^\phi = \alpha^{-i}$.

Now suppose that g is an element of $G = (V_1 \times V_2)\langle\alpha\rangle$ fixed by ϕ, so $g = v\alpha^i$ and $g^\phi = g$. Then $v\alpha^i = (v\alpha^i)^\phi = v^\phi(\alpha^i)^\phi = v^\phi\alpha^{-i}$, so $v^{-1}v^\phi = \alpha^{2i}$. Since $V_1 \times V_2$ and $\langle\alpha\rangle$ have trivial intersection, $\alpha^{2i} = 1$ and $v^\phi = v$. Since α is of order 3, $\alpha^{2i} = 1$ implies $\alpha^i = \alpha^{4i} = 1$, so $g = v$, a fixed point for ϕ in $V_1 \times V_2$.

Now we show that ϕ has order 6, and show that if ϕ fixes v in $V_1 \times V_2$ then $v = (0,0,0,0)$. Note that $\phi^2 = (\sigma_1\sigma_2\tau)^2 = \sigma_1\sigma_2\tau\sigma_1\sigma_2\tau = \sigma_1\sigma_2(\sigma_1\sigma_2)^\tau$; but $(\sigma_1\sigma_2)^\tau = \sigma_2\sigma_1$, and σ_1 commutes with σ_2, so $\phi^2 = \sigma_1^2\sigma_2^2$. Therefore, $\phi^3 = \sigma_1^2\sigma_2^2\sigma_1\sigma_2\tau = \tau$, which has order 2. Hence ϕ is of order 6.

If ϕ fixes v in $V_1 \times V_2$, then surely ϕ^4 also fixes v. From the previous calculations, we have $\phi^4 = (\sigma_1^2\sigma_2^2)^2 = \sigma_1^4\sigma_2^4 = \sigma_1\sigma_2$, so $M_{\phi^4} = M_{\sigma_1\sigma_2} = M_{\sigma_1}M_{\sigma_2}$, which is

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Let $v = (x_1, x_2, x_3, x_4)$, so $vM_{\phi^4} = v$ implies

$$(x_1, x_2, x_3, x_4)\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} = (x_1, x_2, x_3, x_4),$$

where all entries are elements of F_2, the field of 2 elements. Clearly the only solution to this equation is $(0,0,0,0)$. Hence ϕ^4, and also ϕ, is a fixed-point-free automorphism of $V_1 \times V_2$, so ϕ is a fixed-point-free automorphism of G as claimed.

The proof that our counterexample is minimal with respect to the order of G is, unfortunately, too technical to allow its presentation here. The central idea is to use induction to determine as much as possible the *form* of a minimal counterexample, then investigate small groups of the appropriate form to see which is the one sought. Specifically, a candidate G must be of the form $G = WS$, where W is a normal subgroup of G that is the direct product of k copies of Z_r (a cyclic group of prime order r), while S is a Sylow s-subgroup of G for some prime s other than r. Note that in the actual minimal counterexample, $W = V_1 \times V_2$, $r = 2$, $k = 4$, $S = \langle \sigma \rangle$, and $s = 3$.

This description of the form of a minimal counterexample reminds us that minimality can be defined in a variety of ways. For instance, we might try to determine the pair G_1, ϕ_1 such that the number n of prime factors of the order of G_1, counting multiplicities, is minimal. In the counterexample minimizing the order of G, $n = 5$, since $48 = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 3$. However, the smallest possible value for n is actually 3; there exists a nonnilpotent group G_1 of order $75 = 5 \cdot 5 \cdot 3$ having a fixed-point-free automorphism ϕ_1 of order 4.

One of the main areas of interest in groups having fixed-point-free automorphisms is the analysis of the situation in which the group and the automorphism are of relatively prime order [**1**, Chapter 10]. Thus it is interesting to note that G_1 is also the nonnilpotent group of smallest order having a fixed-point-free automorphism of relatively prime order. Hence G_1, ϕ_1 provides a counterexample minimal with respect to two different criteria. An enterprising reader might wish to refer to MacHale's article [**2**] and examine more of his examples with respect to various definitions of minimality.

References

[1] D. Gorenstein, Finite Groups, Harper and Row, New York, 1968.

[2] D. MacHale, Minimum counterexamples in group theory, this MAGAZINE, 54 (1981) 23–28.

[3] New York Times, April 24, 1959, p. 29.

[4] J. Thompson, Finite groups with fixed-point-free automorphisms of prime order, Proc. Acad. Sci., 45 (1959) 578–581.

Design of an Oscillating Sprinkler

BART BRADEN
Northern Kentucky University
Highland Heights, KY 41076

The common oscillating lawn sprinkler has a hollow curved sprinkler arm, with a row of holes on top, which rocks slowly back and forth around a horizontal axis. Water issues from the holes in a family of streams, forming a curtain of water that sweeps back and forth to cover an approximately rectangular region of lawn. Can such a sprinkler be designed to spread water uniformly on a level lawn?

We break the analysis into three parts:

1. How should the sprinkler arm be curved so that streams issuing from evenly spaced holes along the curved arm will be evenly spaced when they strike the ground?
2. How should the rocking motion of the sprinkler arm be controlled so that each stream will deposit water uniformly along its path?
3. How can the power of the water passing through the sprinkler be used to drive the sprinkler arm in the desired motion?

The first two questions provide interesting applications of elementary differential equations. The third, an excursion into mechanical engineering, leads to an interesting family of plane curves which we've called curves of constant diameter. A serendipitous bonus is the surprisingly simple classification of these curves.

much as possible the *form* of a minimal counterexample, then investigate small groups of the appropriate form to see which is the one sought. Specifically, a candidate G must be of the form $G = WS$, where W is a normal subgroup of G that is the direct product of k copies of Z_r (a cyclic group of prime order r), while S is a Sylow s-subgroup of G for some prime s other than r. Note that in the actual minimal counterexample, $W = V_1 \times V_2$, $r = 2$, $k = 4$, $S = \langle \sigma \rangle$, and $s = 3$.

This description of the form of a minimal counterexample reminds us that minimality can be defined in a variety of ways. For instance, we might try to determine the pair G_1, ϕ_1 such that the number n of prime factors of the order of G_1, counting multiplicities, is minimal. In the counterexample minimizing the order of G, $n = 5$, since $48 = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 3$. However, the smallest possible value for n is actually 3; there exists a nonnilpotent group G_1 of order $75 = 5 \cdot 5 \cdot 3$ having a fixed-point-free automorphism ϕ_1 of order 4.

One of the main areas of interest in groups having fixed-point-free automorphisms is the analysis of the situation in which the group and the automorphism are of relatively prime order [**1**, Chapter 10]. Thus it is interesting to note that G_1 is also the nonnilpotent group of smallest order having a fixed-point-free automorphism of relatively prime order. Hence G_1, ϕ_1 provides a counterexample minimal with respect to two different criteria. An enterprising reader might wish to refer to MacHale's article [**2**] and examine more of his examples with respect to various definitions of minimality.

References

[1] D. Gorenstein, Finite Groups, Harper and Row, New York, 1968.

[2] D. MacHale, Minimum counterexamples in group theory, this MAGAZINE, 54 (1981) 23–28.

[3] New York Times, April 24, 1959, p. 29.

[4] J. Thompson, Finite groups with fixed-point-free automorphisms of prime order, Proc. Acad. Sci., 45 (1959) 578–581.

Design of an Oscillating Sprinkler

BART BRADEN

Northern Kentucky University

Highland Heights, KY 41076

The common oscillating lawn sprinkler has a hollow curved sprinkler arm, with a row of holes on top, which rocks slowly back and forth around a horizontal axis. Water issues from the holes in a family of streams, forming a curtain of water that sweeps back and forth to cover an approximately rectangular region of lawn. Can such a sprinkler be designed to spread water uniformly on a level lawn?

We break the analysis into three parts:

1. How should the sprinkler arm be curved so that streams issuing from evenly spaced holes along the curved arm will be evenly spaced when they strike the ground?
2. How should the rocking motion of the sprinkler arm be controlled so that each stream will deposit water uniformly along its path?
3. How can the power of the water passing through the sprinkler be used to drive the sprinkler arm in the desired motion?

The first two questions provide interesting applications of elementary differential equations. The third, an excursion into mechanical engineering, leads to an interesting family of plane curves which we've called curves of constant diameter. A serendipitous bonus is the surprisingly simple classification of these curves.

The following result, proved in most calculus textbooks, will play a fundamental role in our discussion.

LEMMA. *Ignoring air resistance, a projectile shot upward from the ground with speed v at an angle θ from the vertical, will come down at a distance* $(v^2/g)\sin 2\theta$. (*Here g is the acceleration due to gravity.*)

Note that $\theta = \pi/4$ gives the maximum projectile range, since then $\sin 2\theta = 1$. Textbooks usually express the projectile range in terms of the 'angle of elevation', $\pi/2 - \theta$; but since $\sin 2(\pi/2 - \theta) = \sin 2\theta$, the range formula is unaffected when the zenith angle is used instead.

The sprinkler arm curve

In FIGURE 1, a (half) sprinkler arm is shown in a vertical plane, which we take to be the xy plane throughout this section. Let L be the length of the arc from the center of the sprinkler arm to the outermost hole, and let $x = x(s)$, $y = y(s)$ be parametric equations for the curve, using the arc length s, $0 \leqslant s \leqslant L$, as parameter. Let $\alpha(s)$ denote the angle between the vertical and the outward normal to the arc at the point $(x(s), y(s))$.

We'll see that the functions $x(s)$ and $y(s)$ which define the curve are completely determined (once L, $\alpha(L)$ and $y(0)$ have been chosen) by the requirement that streams passing through evenly-spaced holes on the sprinkler arm should be uniformly spaced when they strike the ground.

If there were a hole at the point $(x(s), y(s))$ on the sprinkler arm, the direction vector of the stream issuing from this hole would be $N(s) = \langle \sin\alpha(s), \cos\alpha(s)\rangle$, and this stream would reach the ground at a distance

$$d(s) = \frac{v^2}{g}\sin 2\alpha(s).$$

The condition that evenly-spaced holes along the arm produce streams which are evenly spaced when they reach the ground is that $d(s)$ be proportional to s:

$$\frac{d(s)}{d(L)} = \frac{s}{L},$$

or equivalently,

$$\sin 2\alpha(s) = \frac{s}{L}\sin 2\alpha(L). \tag{1}$$

(We have made the assumption that the dimensions of the sprinkler are small in comparison to the dimensions of the area watered. This simplifies the calculations, and the errors introduced are not significant.)

The unit tangent vector to the sprinkler arm curve at $(x(s), y(s))$ is $T(s) = \langle x'(s), y'(s)\rangle$, so the unit outward normal vector (obtained by rotating $T(s)$ counterclockwise by $\pi/2$) is $N(s) = \langle -y'(s), x'(s)\rangle$. Comparing this with our earlier expression for $N(s)$, we have

$$x'(s) = \cos\alpha(s), \qquad y'(s) = -\sin\alpha(s).$$

Since $\sin 2\alpha(s) = 2\sin\alpha(s)\cos\alpha(s)$, equation (1) for the sprinkler arm curve becomes

$$-2x'(s)\,y'(s) = \frac{s}{L}\sin 2\alpha(L).$$

The value of $\alpha(L)$, the angle between the vertical and the outermost stream as it leaves the sprinkler, is a parameter under the designer's control; once it is chosen, the value of $\sin 2\alpha(L)$ is determined—call it k, where $0 < k \leqslant 1$. Our equation then becomes

$$x'(s)\,y'(s) = \frac{-k}{2L}s. \tag{2}$$

Since $N(s)$ is a unit vector, also

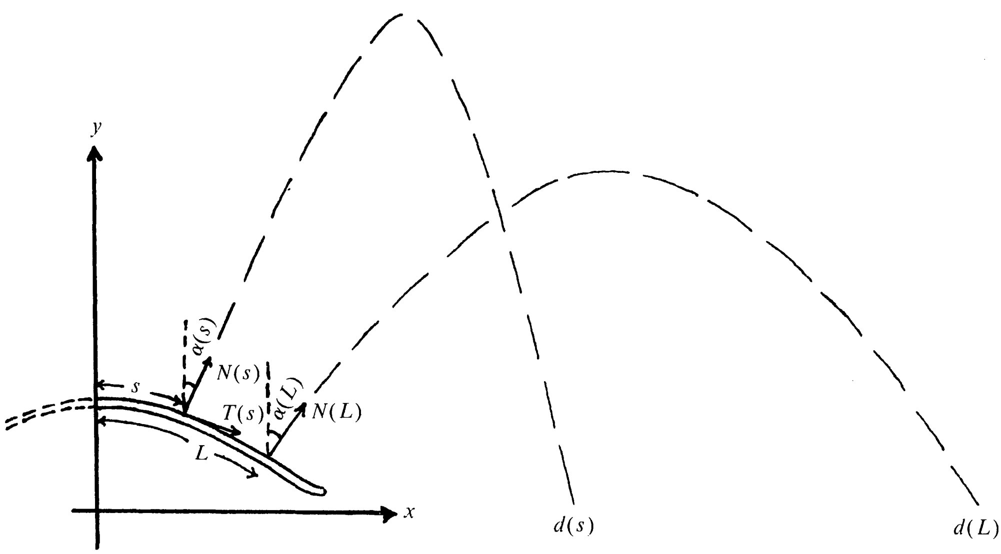


FIGURE 1

$$x'(s)^2 + y'(s)^2 = 1. \tag{3}$$

Fortunately, the pair of nonlinear differential equations (2) and (3) for $x(s)$ and $y(s)$ simplifies algebraically:

$$x'^2 + \left(\frac{-ks}{2Lx'}\right)^2 = 1,$$

or

$$x'^4 - x'^2 + \frac{k^2s^2}{4L^2} = 0.$$

Solving by the quadratic formula,

$$x'(s)^2 = \frac{1 \pm \sqrt{1 - \left(\frac{ks}{L}\right)^2}}{2}.$$

Since the sprinkler arm is horizontal at its midpoint, the unit tangent vector $T(0) = \langle x'(0), y'(0)\rangle$ is $\langle 1, 0\rangle$. Thus $x'(0) = 1$, which means we must use the + sign in the quadratic formula. Substituting

$$x'(s) = \frac{1}{\sqrt{2}}\left[1 + \sqrt{1 - \left(\frac{ks}{L}\right)^2}\,\right]^{1/2}$$

in equation (2) gives

$$y'(s) = \frac{-1}{\sqrt{2}}\left[1 - \sqrt{1 - \left(\frac{ks}{L}\right)^2}\,\right]^{1/2}$$

Since $x(0) = 0$ and $y(0)$ is arbitrary, we conclude that

$$x(s) = \frac{1}{\sqrt{2}}\int_0^s \left[1 + \sqrt{1 - \left(\frac{kt}{L}\right)^2}\,\right]^{1/2} dt$$

and

$$y(s) = y(0) - \frac{1}{\sqrt{2}} \int_0^s \left[1 - \sqrt{1 - \left(\frac{kt}{L} \right)^2} \right]^{1/2} dt.$$

These integrals can be evaluated in closed form, using the identity (kindly supplied by a reviewer)

$$\frac{1 \pm \sqrt{1 - \left(\frac{kt}{L} \right)^2}}{2} = \left[\frac{\sqrt{1 + \frac{kt}{L}}}{2} \pm \frac{\sqrt{1 - \frac{kt}{L}}}{2} \right]^2,$$

with the result

$$x(s) = \frac{L}{3k} \left[\left(1 + \frac{ks}{L} \right)^{3/2} - \left(1 - \frac{ks}{L} \right)^{3/2} \right],$$

$$y(s) = y(0) - \frac{2L}{3k} \left[\left(1 + \frac{ks}{L} \right)^{3/2} + \left(1 - \frac{ks}{L} \right)^{3/2} \right].$$

This sprinkler arm curve is drawn in FIGURE 2. Note that the curve is determined by the requirement that the streams be evenly spaced along the ground when the plane of the sprinkler arm is vertical. Later we indicate what happens when this plane makes an angle ϕ with the vertical.

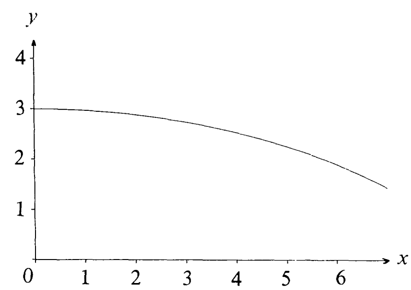


FIGURE 2

The rocking motion of the sprinkler arm

We wish the sprinkler arm to oscillate in such a way that each stream will deposit water uniformly along its path, or what is the same, the speed of the point of impact of the stream with the ground should be constant on each pass of the sprinkler. As it turns out, this condition cannot be satisfied by all the streams simultaneously, so we shall concentrate our attention on the central stream.

Henceforth, let's choose a coordinate system in space, as indicated in FIGURE 3, with the z-axis vertical and the axis of rotation of the sprinkler arm the y-axis, with the center of the arm on the positive z-axis. When the plane of the sprinkler arm makes an angle ϕ with the vertical, the central stream will reach the ground on the x-axis, at $x = (v^2/g)\sin 2\phi$. The oscillation of the sprinkler arm is described by the function $\phi(t)$, and the corresponding speed of the central stream over the lawn is the derivative $x'(t) = (2v^2/g)\cos 2\phi(t)\phi'(t)$. Setting $x'(t) = Kv^2/g$, a conveniently labelled constant, we see that uniform coverage by the central stream is equivalent to the requirement that $\phi(t)$ be a solution of the (separable) differential equation

$$2\cos 2\phi(t)\,\phi'(t) = K. \tag{4}$$

Integration of (4) gives the solution

$$\sin 2\phi(t) = Kt + c. \tag{5}$$

The parameters K and c have no apparent significance, so we next try to find an expression for the angular variation $\phi(t)$ in terms of two other constants which are easily interpreted.

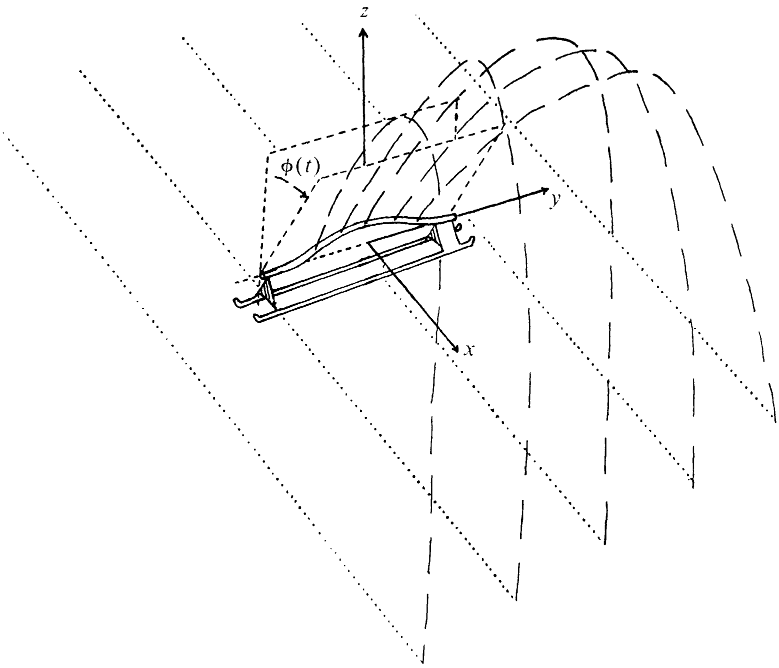


FIGURE 3

Suppose the sprinkler arm rocks back and forth in the range $-\phi_0 \leqslant \phi(t) \leqslant \phi_0$, where the maximum tilt, ϕ_0, is a design parameter in the range $0 < \phi_0 \leqslant \pi/4$. Let the time required for the sprinkler arm to rotate between the vertical and the maximum angle ϕ_0 be denoted by T. (Thus $2T$ is the time required for one pass of the sprinkler over the lawn, and $4T$ is the period of the complete oscillation.) If we measure time so that $\phi(0) = -\phi_0$, then setting $t = 0$ in equation (5) gives $c = -\sin 2\phi_0$. Since $\phi(2T) = \phi_0$, we then get $\sin 2\phi_0 = 2KT - \sin 2\phi_0$, or $K = \frac{1}{T}\sin 2\phi_0$. Thus

$$\sin 2\phi(t) = \frac{t-T}{T}\sin 2\phi_0,$$

or, since $-\pi/2 \leqslant 2\phi(t) \leqslant \pi/2$,

$$\phi(t) = \frac{1}{2}\arcsin\left[\frac{t-T}{T}\sin 2\phi_0\right]. \tag{6}$$

The oscillatory motion of the sprinkler arm is therefore uniquely determined (once choices of ϕ_0 and T have been made) by the requirement that the **central** stream cover the ground uniformly.

Remark: Since the maximum range of the central stream occurs when $\phi = \pi/4$, one might think the ideal value for ϕ_0 would be $\pi/4$. However, we will show later that the shape of the region covered by the sprinkler will be more nearly rectangular if ϕ_0 is somewhat smaller than $\pi/4$.

It remains to describe a mechanism which will produce the desired oscillation, given by (6). (It was by observing my own sprinkler, the Nelson 'dial-a-rain', which appears to use the design described below, that I was led to the questions considered in this paper.)

Mechanical design of the sprinkler

The stream of water entering the sprinkler from a hose can be used to turn an impeller (waterwheel), which is then geared down to turn a cam with a constant angular velocity ω. A cam follower linkage converts the uniform rotational motion of the cam into an oscillatory motion of

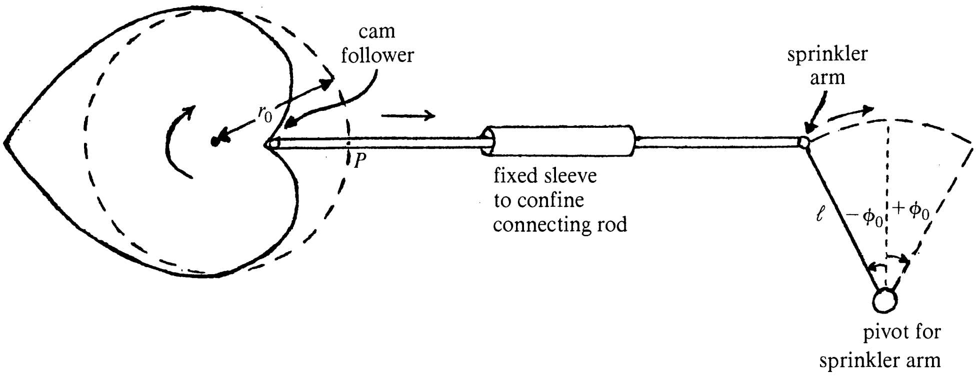


FIGURE 4. **A typical cam mechanism.**

the sprinkler arm—as the cam makes a half-revolution, the sprinkler arm is pushed from $\phi = -\phi_0$ to $\phi = +\phi_0$; and on the next half-revolution of the (bilaterally symmetric) cam, the sprinkler arm makes the return sweep.

What shape of cam will cause the oscillation of the sprinkler arm to be that given by (6)? We may describe the shape of a cam in polar coordinates by $r(\theta) = r_0 + f(\theta)$, where the function $f(\theta)$ describes the 'eccentricity' of the cam, i.e., its deviation from the circle $r(\theta) = r_0$. The pole of our coordinate system is placed at the center around which the cam rotates, so it is the eccentricity $f(\theta)$ which produces the motion of the sprinkler arm. As the cam follower moves to the right or left of the point labelled P in FIGURE 4, by the amount $f(\theta)$, the other end of the connecting rod moves the sprinkler arm the same distance along a circular arc of radius ℓ. Denoting the arclength by s and using the relation $s/\ell = \phi$, we have $f(\theta)/\ell = \phi$. Since $\theta(t) = \omega t$, we want

$$f(\omega t) = \frac{\ell}{2}\arcsin\left[\frac{t-T}{T}\sin 2\phi_0\right], \qquad 0 \leqslant t \leqslant 2T.$$

Our goal is to find a formula for the eccentricity $f(\theta)$, so we must express t and T in terms of θ and ω, using the relation $\theta = \omega t$. This is easy: as the cam turns a half-revolution, from $\theta = 0$ to $\theta = \pi$, the sprinkler arm moves from $\phi = -\phi_0$ to $\phi = \phi_0$; so equating the times required gives $\pi/\omega = 2T$. Thus

$$\frac{t-T}{T} = \frac{2\omega t}{\pi} - 1,$$

so

$$f(\omega t) = \frac{\ell}{2}\arcsin\left[\frac{2}{\pi}\left(\omega t - \frac{\pi}{2}\right)\sin 2\phi_0\right], \qquad 0 \leqslant t \leqslant 2T.$$

Replacing ωt by θ, we conclude that

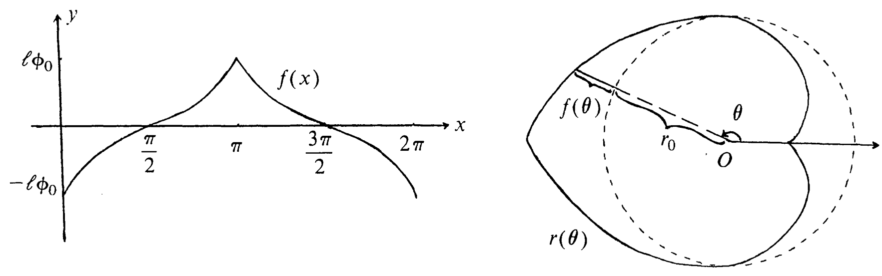
$$f(\theta) = \frac{\ell}{2}\arcsin\left[\frac{2}{\pi}\left(\theta - \frac{\pi}{2}\right)\sin 2\phi_0\right], \qquad 0 \leqslant \theta \leqslant \pi. \tag{7}$$

As θ goes from π to 2π we want the sprinkler arm to perform the same motion in reverse, i.e., the cam should be symmetric about the polar axis:

$$f(\theta) = f(2\pi - \theta) \quad \text{for } \pi \leqslant \theta \leqslant 2\pi. \tag{8}$$

The polar curve $r(\theta) = r_0 + f(\theta)$, where the eccentricity $f(\theta)$ is given by (7) and (8), is the cam shape which will produce the desired oscillatory motion of the sprinkler arm (see FIGURE 5). (Note that r_0 is arbitrary, provided $r_0 > \ell\phi_0$.) This curve has an interesting geometric property, described in the following definition.

DEFINITION. A simple closed curve C is said to be of **constant diameter** d if there is a point O inside C such that every chord of C through O has the same length, d. Any chord through this 'center' point O is called a **diameter** of C.



(a) (b)

FIGURE 5. **(a) The Cartesian graph of $f(x)$. (b) The polar graph of $r = r_0 + f(\theta)$.**

N.B. This class of curves should not be confused with 'curves of constant width', a family of **convex** curves which appears frequently in the literature, e.g. [**1**].

It is easy to verify using (7) and (8), that our cam curve $r(\theta) = r_0 + f(\theta)$ has constant diameter $2r_0$.

Proof. Let O be the pole of our coordinate system. Then the diameter of our curve which makes an angle θ with the polar axis is $r(\theta) + r(\theta + \pi)$, or $2r_0 + f(\theta) + f(\theta + \pi)$. Thus it must be shown that $f(\theta) + f(\theta + \pi) = 0$. Without loss of generality we may assume $0 \leqslant \theta \leqslant \pi$; then by (8),

$$\begin{aligned} f(\theta+\pi) &= f(2\pi - (\theta+\pi)) = f(\pi - \theta) \\ &= \frac{\ell}{2}\arcsin\left[\frac{2}{\pi}\left(\pi - \theta - \frac{\pi}{2}\right)\sin 2\phi_0\right] \\ &= \frac{\ell}{2}\arcsin\left[\frac{-2}{\pi}\left(\theta - \frac{\pi}{2}\right)\sin 2\phi_0\right] = -f(\theta). \end{aligned}$$

Examining this proof, we discover a simple construction for all curves of constant diameter. Given $d > 0$, take any continuous function $r(\theta)$ such that $r(0) + r(\pi) = d$ and $0 < r(\theta) < d$ for $0 \leqslant \theta \leqslant \pi$. If we extend the domain to $[\pi, 2\pi]$ by defining $r(\theta + \pi) = d - r(\theta)$, as in (8), the polar curve $r = r(\theta)$ will have constant diameter d.

For curves symmetric with respect to the polar axis, i.e., with $r(2\pi - \theta) = r(\theta)$ for $0 \leqslant \theta \leqslant \pi$, the constant diameter condition is simply that $r(\pi - \theta) = d - r(\theta)$ for $0 \leqslant \theta \leqslant \pi/2$. So $d = 2r(\pi/2)$. Thus an arbitrary continuous function $r(\theta)$ defined for $0 \leqslant \theta \leqslant \pi/2$, for which $0 < r(\theta) < 2r(\pi/2)$, can be extended uniquely to produce a simple closed curve of constant diameter $d = 2r(\pi/2)$ which is symmetric with respect to the polar axis.

The observation that the cam curve for our sprinkler has constant diameter $2r_0$ suggests a particularly simple mechanical design for the cam follower linkage: a post fixed to the center of the cam, sliding in a slot in the connecting rod, with rollers fixed on the rod separated by the distance $2r_0$. As the cam turns, the rollers remain in contact with it at opposite ends of a diameter, and the connecting rod is alternately pushed and pulled along the line of its slot (see FIGURE 6). If the cam did not have constant diameter, a more complicated mechanical linkage would be required to keep the cam follower in contact with the cam, and to confine the motion of the connecting rod to one dimension.

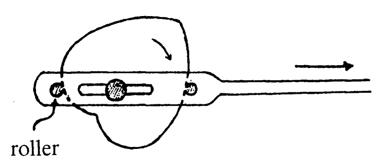


FIGURE 6

The complete sprinkler pattern

In determining the curve of the sprinkler arm we considered only the situation in which the arm is vertical, and found that the requirement of uniform spacing of the streams on the ground then determines the curve uniquely. Similarly, in analyzing the oscillation of the sprinkler arm we considered only the central stream, and found that the requirement of uniform coverage by this single stream along its path uniquely determines the motion $\phi(t)$. It remains to be seen whether the streams from the other holes will move along the lawn at constant speeds, and whether these streams will remain equally spaced as the sprinkler arm rocks back and forth.

Suppose there are $2n+1$ holes in the sprinkler arm: one in the center and n more spaced at equal intervals on each side. By symmetry we need only consider the streams from one half of the sprinkler arm. Using the coordinate system described earlier (see FIGURE 3), denote the angles between the vertical and the streams as they leave the sprinkler arm (when the plane of the arm is vertical) by $\alpha_0, \alpha_1, \dots, \alpha_n$, where $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_n \leqslant \pi/4$. The streams will strike the ground at distances $d_i = (v^2/g)\sin 2\alpha_i$, and since the streams are equally spaced along the ground, $d_i = (i/n)d_n$. That is,

$$\frac{v^2}{g}\sin 2\alpha_i = \frac{i}{n}\left(\frac{v^2}{g}\sin 2\alpha_n\right),$$

or

$$\alpha_i = \frac{1}{2}\arcsin\left[\frac{i}{n}\sin 2\alpha_n\right].$$

The direction vectors of the streams as they leave the sprinkler arm are $N_i = \langle 0, \sin\alpha_i, \cos\alpha_i\rangle$, $0 \leqslant i \leqslant n$.

When the plane of the sprinkler arm is tilted at an angle ϕ, the direction vectors N_i are rotated through the angle ϕ around the y-axis, so the streams issue from the holes in the directions $N_i(\phi) = \langle \cos\alpha_i \sin\phi, \sin\alpha_i, \cos\alpha_i \cos\phi\rangle$. The angle θ_i between $N_i(\phi)$ and the vertical is given by

$$\cos\theta_i = N_i(\phi)\cdot\langle 0,0,1\rangle = \cos\alpha_i \cos\phi,$$

so the ith stream strikes the ground at a distance

$$d_i(\phi) = \frac{v^2}{g}\sin 2\theta_i = \frac{2v^2}{g}\cos\theta_i \sin\theta_i = \frac{2v^2}{g}\cos\alpha_i\cos\phi\sqrt{1-\cos^2\alpha_i\cos^2\phi}\,.$$

The point of impact is $d_i(\phi)\overline{N}_i(\phi)$, where

$$\overline{N}_i(\phi) = \frac{\langle \cos\alpha_i\sin\phi, \sin\alpha_i, 0\rangle}{\sqrt{\cos^2\alpha_i\sin^2\phi + \sin^2\alpha_i}},$$

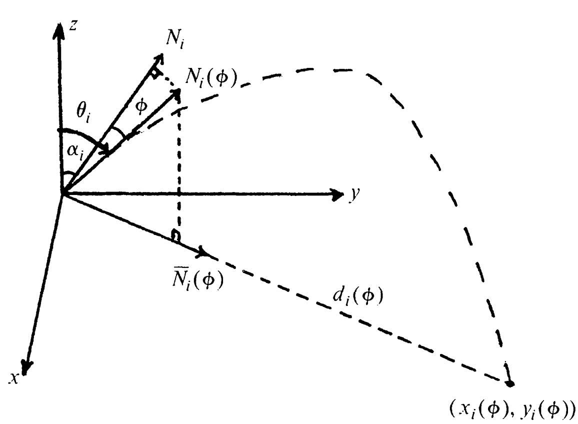


FIGURE 7

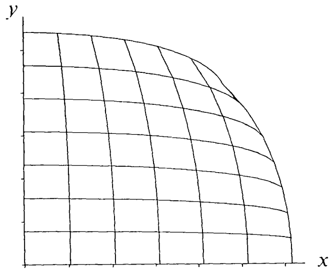


FIGURE 8

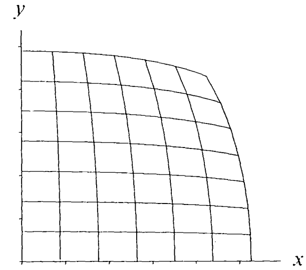


FIGURE 9

the unit vector in the direction of the projection of $N_i(\phi)$ on the xy plane. Thus parametric equations for the path of the ith stream as it moves over the lawn are

$$x_i(\phi) = \frac{2v^2\cos^2\alpha_i \cos\phi \sin\phi\sqrt{1-\cos^2\alpha_i\cos^2\phi}}{g\sqrt{\cos^2\alpha_i \sin^2\phi + \sin^2\alpha_i}},$$

$$y_i(\phi) = \frac{2v^2\cos\alpha_i \sin\alpha_i \cos\phi\sqrt{1-\cos^2\alpha_i\cos^2\phi}}{g\sqrt{\cos^2\alpha_i \sin^2\phi + \sin^2\alpha_i}}.$$

Evidently $y_i(\phi)$ is not constant, so the path of the ith stream of water is not the straight line parallel to the central stream which one might have expected.

To parametrize the paths by time, we can simply replace the parameter ϕ by the expression for $\phi(t)$ in (6). Computer plots of the resulting family of curves are shown in FIGURES 8 and 9. In each of these figures, the curves running approximately parallel to the x-axis are the paths of the streams from one side of the sprinkler arm. Time is indicated by the curves nearly parallel to the y-axis, which are polygonal arcs connecting the points on the eight stream paths at six equally spaced instants $T+(k/6)T$, $1 \leqslant k \leqslant 6$. (Recall that as t runs through the interval $T \leqslant t \leqslant 2T$, the plane of the sprinkler arm turns through the interval $0 \leqslant \phi \leqslant \phi_0$.) Thus each of the resulting 'squares' receives the same amount of water on each pass of the sprinkler.

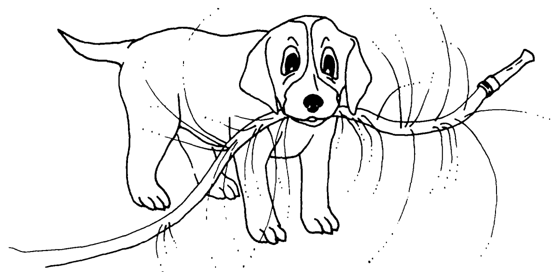
As seen in FIGURE 8, where $\phi_0 \approx 40°$ and $\alpha_7 \approx 30°$, the outer streams curve in significantly, making the 'squares' near the outside corner smaller in area. Since each 'square' receives the same amount of water per pass of the sprinkler, the sprinkler shown would overwater the four corners of the region it sprinkled.

In FIGURE 9, by reducing ϕ_0 to about 29° (and decreasing α_7 to 26° to keep similar proportions to the region watered), not only is the non-uniformity of coverage reduced, but at the same time the region covered is more nearly rectangular.

We conclude with some observations which could not be followed up here; their investigation is left to the proverbial interested reader.

1. No attempt has been made to define an optimal shape for the region watered. Evidently, decreasing the angle parameters ϕ_0 and $\alpha(L)$ will make the coverage more uniform, but at the cost of decreasing the area watered. An interesting question, suggested by a reviewer, might be to design a sprinkler to maximize the area covered without exceeding a stipulated amount of variation in the water applied per unit area. This would mean introducing some non-uniformity along the x and y axes (i.e., unequal spacing of the streams as they strike the ground when the plane of the sprinkler arm is vertical, and non-uniform speed of the central stream along its path), to compensate for the overwatering of the corners observed in FIGURES 8 and 9 with our sprinkler.

2. My Nelson 'dial-a-rain' sprinkler has an additional feature of interest. On the sprinkler arm support is a dial which, when turned, changes the radius ℓ of the arc on which the sprinkler arm moves. The effect of doubling ℓ, for instance, can be shown to be to cut in half the region watered. The coverage of this smaller area is slightly less uniform, however.



... of course, other sprinkler designs are possible.

I find it remarkable that not only are the curve of the sprinkler arm and the motion $\phi(t)$ of the arm unique, but even the mechanical design of the sprinkler is essentially determined by the requirement that water should be spread uniformly along the two coordinate axes. The wealth of mathematical questions raised in the analysis of this simple mechanism gives me a new respect for mechanical engineering, and greater confidence in the importance of classical mathematics to students in this field.

Reference

[1] G. D. Chakerian, A characterization of curves of constant width, Amer. Math. Monthly, 81 (1974) 153–155.

A Combinatorial Identity

ZIVIA S. WURTELE
Pan Heuristics, P.O. Box 9695
Marina del Rey, CA 90292

The theorem I prove here was motivated by the following problem. Suppose that N weapons are fired independently, and that each weapon hits at most one of $k+m$ targets, where $m \geqslant 0$. For each weapon, the probability of hitting target i is p_i, and the probability of missing all $k+m$ targets is

$$q = 1 - \sum_{i=1}^{k+m} p_i.$$

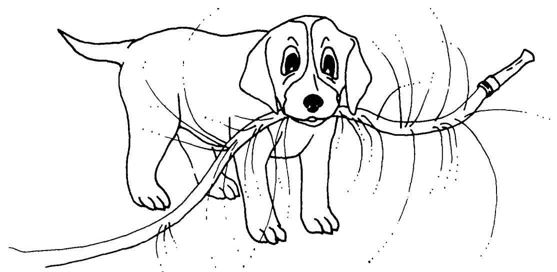
We are interested in computing the probability that all of a specified subset of the targets, say the first k targets, are hit at least once in the N firings and that the remaining m targets are not hit at all. When N is large relative to k, the theorem considerably simplifies the computation required to solve this problem. This result, however, is more general, for it does not assume that the p's are probabilities.

The following is the combinatorial identity to be proved:

THEOREM.

$$\sum_{\substack{n_i \geqslant 1 \\ N - \sum n_i \geqslant 0}} \frac{N! p_1^{n_1} \cdots p_k^{n_k} q^{N-n_1 \cdots -n_k}}{n_1! \cdots n_k!(N-n_1 - \cdots n_k)!} = \sum_{h=0}^{k} \sum_{\substack{\text{all } h \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (-1)^{k-h}(p_{i_1} + p_{i_2} \cdots + p_{i_h} + q)^N.$$

Each term on the left of the above sum equals the probability that in N firings the first target is



... of course, other sprinkler designs are possible.

I find it remarkable that not only are the curve of the sprinkler arm and the motion $\phi(t)$ of the arm unique, but even the mechanical design of the sprinkler is essentially determined by the requirement that water should be spread uniformly along the two coordinate axes. The wealth of mathematical questions raised in the analysis of this simple mechanism gives me a new respect for mechanical engineering, and greater confidence in the importance of classical mathematics to students in this field.

Reference

[1] G. D. Chakerian, A characterization of curves of constant width, Amer. Math. Monthly, 81 (1974) 153–155.

A Combinatorial Identity

ZIVIA S. WURTELE
Pan Heuristics, P.O. Box 9695
Marina del Rey, CA 90292

The theorem I prove here was motivated by the following problem. Suppose that N weapons are fired independently, and that each weapon hits at most one of $k+m$ targets, where $m \geqslant 0$. For each weapon, the probability of hitting target i is p_i, and the probability of missing all $k+m$ targets is

$$q = 1 - \sum_{i=1}^{k+m} p_i.$$

We are interested in computing the probability that all of a specified subset of the targets, say the first k targets, are hit at least once in the N firings and that the remaining m targets are not hit at all. When N is large relative to k, the theorem considerably simplifies the computation required to solve this problem. This result, however, is more general, for it does not assume that the p's are probabilities.

The following is the combinatorial identity to be proved:

THEOREM.

$$\sum_{\substack{n_i \geqslant 1 \\ N - \Sigma n_i \geqslant 0}} \frac{N!\, p_1^{n_1} \cdots p_k^{n_k} q^{N-n_1 \cdots - n_k}}{n_1! \cdots n_k!(N - n_1 - \cdots n_k)!} = \sum_{h=0}^{k} \sum_{\substack{\text{all } h \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (-1)^{k-h}(p_{i_1} + p_{i_2} \cdots + p_{i_h} + q)^N.$$

Each term on the left of the above sum equals the probability that in N firings the first target is hit n_1 times, the second target n_2 times,..., the kth target n_k times, and that no hits are scored against any of the targets in the remaining $N - \sum_{i=1}^{k} n_i$ firings. The summation is over all n_i's that are greater than or equal to 1 and for which the total number of hits of the k targets does not exceed N.

Proof. The left-hand side of the above equation equals $(\sum_{i=1}^{k} p_i + q)^N$ minus those terms in the expansion for which one or more of the exponents of the p_i's equals zero. Applying the inclusion-exclusion principle, we exclude first those terms where p_1^0 appears. Their sum is $(p_2 + p_3 + \cdots + p_k + q)^N$. Similarly for $p_2^0, p_3^0, \ldots, p_k^0$. The excluded sum is then

$$\sum_{\substack{\text{all } (k-1) \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (p_{i_1} + p_{i_2} + \cdots + p_{i_{k-1}} + q)^N.$$

Next we add all terms containing products of type $p_1^0 p_2^0$. The sum of all such terms is

$$\sum_{\substack{\text{all } (k-2) \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (p_{i_1} + p_{i_2} + \cdots + p_{i_{k-2}} + q)^N.$$

We then subtract all terms which contain products of type $p_1^0 p_2^0 p_3^0$. And so on, until finally we reach the last term q^N, which contains the product $p_1^0 p_2^0 \cdots p_k^0$.

The following are immediate consequences of the Theorem:

COROLLARY 1. *If $N < k$, then for all p_i and q,*

$$\sum_{h=0}^{k} \sum_{\substack{\text{all } h \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (-1)^{k-h} (p_{i_1} + p_{i_2} \cdots + p_{i_h} + q)^N = 0.$$

COROLLARY 2. *For all q,*

$$\sum_{h=0}^{k} \sum_{\substack{\text{all } h \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (-1)^{k-h} (p_{i_1} + p_{i_2} \cdots + p_{i_h} + q)^k = \sum_{h=0}^{k} \sum_{\substack{\text{all } h \text{ size} \\ \text{subsets of} \\ \{1,2,\ldots,k\}}} (-1)^{k-h} (p_{i_1} + p_{i_2} \cdots + p_{i_h})^k.$$

A simple example will illustrate the usefulness of the Theorem. Assume there are only two targets and that in a single firing the probability of hitting the first target is $p_1 = .01$ and the probability of hitting the second target is $p_2 = .10$. The Theorem tells us that the probability of hitting each of the targets at least once in N independent trials is equal to

$$(.89)^N - \left[(.01 + .89)^N + (.10 + .89)^N\right] + (.01 + .10 + .89)^N. \tag{1}$$

Suppose we wish to know how many weapons need to be fired so that this probability equals .95. Using (1), we obtain the following results.

N	Probability of hitting both targets
10	.06
100	.63
200	.87
300	.95
400	.98
500	.99

Thus the answer to our question is 300 weapons.

My original proof was by mathematical induction; I am grateful to a referee for suggesting this simpler proof based upon the inclusion-exclusion principle.

Convex Curves with Periodic Billiard Polygons

DUANE W. DETEMPLE
JACK M. ROBERTSON
Washington State University
Pullman, WA 99164

Let C be a piecewise-smooth closed-convex curve in the plane $\mathbb{R}^2$. A **billiard ball** is a point interior to C which moves with unit velocity until it hits a point $p \in C$. If C has a tangent at p the ball bounces in the direction determined by reflection off of the unique tangent line at p, again with unit velocity. The **billiard path** is then the well-determined orbit traced by the point in the course of time, where we shall neglect orbits which encounter any corner point of C. A billiard path is said to be **periodic** if the orbit is a closed polygon with finitely many vertices on C. The polygon need not be simple; for example, nonsimple star figures occur as billiard orbits in circles.

It is of interest to determine if a class of convex curves (or, more generally, a class of convex figures in $\mathbb{R}^d$) can be characterized by a corresponding property of the billiard paths within the figure. Indeed there are a number of results of this type, both recent and classical. The billiard characterizations of the conic sections have long been known. For an example of a more recent result, *a closed convex polygon P is regular if and only if P contains a periodic billiard path P' which is similar to P* (see [**1**]). Two billiard characterizations of curves of constant width are known, the more recent an elegant result found in [**3**]. The other characterization motivates the work of this paper and is obtained by simply restating the well-known double normal property: *a smooth closed convex curve C has constant width if and only if there is a periodic billiard 2-gon (i.e., double segment) at each point of C.*

This characterization of curves of constant width raises a rather intriguing question:

> *Given a positive integer n, $n \geqslant 3$, are there smooth closed convex curves C, other than circles, for which each point $p \in C$ is the vertex of a periodic billiard n-gon in C?*

If the word "each" is replaced by "some" then a result of George Birkhoff (see [**2**]) settles the question. All smooth convex curves contain at least one periodic billiard n-gon. In fact a compactness argument shows the existence of an m-gon P^*, with $m \leqslant n$, which has maximum perimeter among all polygons with n or fewer vertices on C, and it then follows by straightforward geometric reasoning that $m = n$ and P^* is a periodic billiard path.

The following construction, shown in FIGURE 1, produces a smooth convex curve AF_1BF_2 which "nearly" provides an affirmative answer to our question in the case $n = 3$. Here the points F_i are the foci of the parabolas P_i and P_i', with F_1F_2 perpendicular to the horizontal axes of the parabolas. The corners at F_i can be smoothed to give a horizontal tangent at F_i, say by the following construction: near F_1 locate points A_1 and B_1 on the respective parabolic arcs AF_1 and F_1B; at A_1 and B_1 extend tangent segments upward until they intercept the horizontal line through F_1; finally, round off the two angles where the tangents meet the horizontal line by small circular arcs. Similar smoothing can be done at A, F_2, and B. The reflection properties of the parabola show that periodic billiard triangles exist at all points of the smoothed convex curve AF_1BF_2, with the exception of the eight open smoothing arcs of arbitrarily small length.

Unlike the case for $n = 2$ the following theorem indicates that for $n \geqslant 3$ it is too much to hope that a periodic regular n-gon billiard path can be found in noncircular smooth convex curves, even if the size of the regular n-gons is allowed to vary.

THEOREM. *Let C be a twice differentiable closed convex curve which admits a periodic regular n-gon billiard path at each point of C, where $n \geqslant 3$. Then C is a circle.*

Proof. A portion of C is shown in FIGURE 2, which we view as lying in the complex plane. Two successive sides of a regular periodic billiard n-gon of side length ℓ are also shown, with

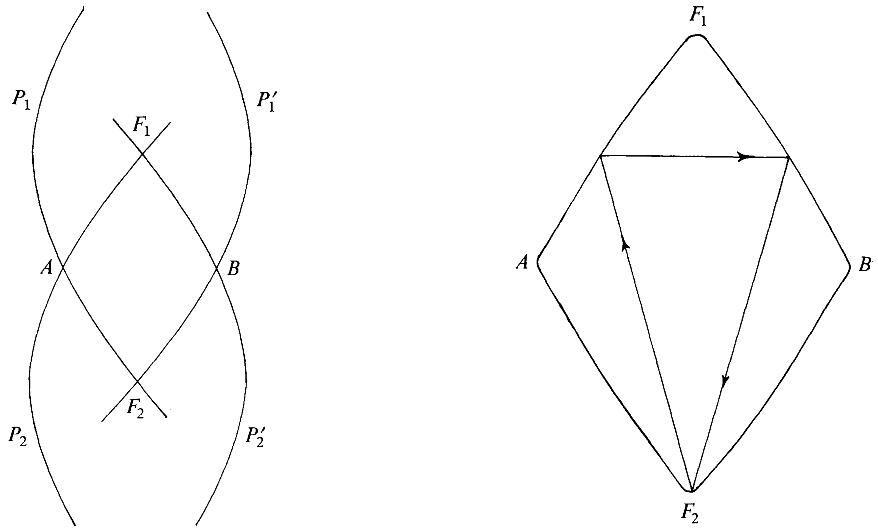


FIGURE 1

vertices ω, z, ζ, and unit tangent t at z. Since complex multiplication by $\rho \equiv e^{i\pi/n}$ is geometrically equivalent to counterclockwise rotation by π/n, the billiard bounce condition at z is expressed by

$$\zeta = z + \ell\rho t, \qquad \omega = z - \ell\bar{\rho}t. \tag{1}$$

We now let z vary on C by considering z as a twice differentiable function of arclength s measured counterclockwise from some fixed point on C. Using primes to denote differentiation with respect to s, we have

$$z' = t, \qquad t' = ikt, \tag{2}$$

where $k \geqslant 0$ is the curvature of C at z. The second equation in (2) is a Frenet formula. Although we omit the details, it can be shown that ℓ (and hence ζ and ω as well) has differentiable dependence on s. Differentiating equations (1) we obtain, in view of (2),

$$\zeta' = (1 + \ell\rho ki + \ell'\rho)t, \qquad \omega' = (1 - \ell\bar{\rho}ik - \ell'\rho)t. \tag{3}$$

Now ζ' must be tangent to C at ζ, and so the billiard condition at ζ can hold only if $\rho^2 t$ is parallel to ζ'. That is, $\rho^2 t$ must be a real multiple of ζ'. Similarly, the billiard condition at ω requires that $\bar{\rho}^2 t$ must be a real multiple of ω'. Since $\rho\bar{\rho} = 1$, it follows from (3) that

$$\bar{\rho}^2 + \ell\bar{\rho}ki + \ell'\bar{\rho} = \text{real}, \qquad \rho^2 - \ell\rho ik - \ell'\rho = \text{real}. \tag{4}$$

Adding the two quantities in (4) we deduce that $\ell' \operatorname{Im}\rho = 0$. Thus $\ell' = 0$, and so the side length ℓ of the regular periodic billiard n-gon is necessarily constant.

Returning to (4), and taking the imaginary part of either equation, we find

$$\sin(2\pi/n) - \ell\cos(\pi/n)k = 0. \tag{5}$$

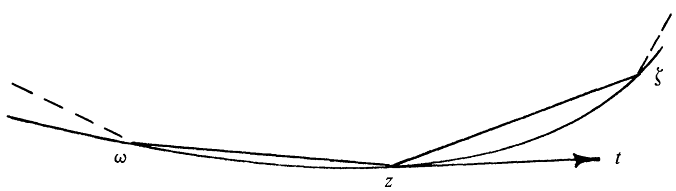


FIGURE 2

Since $n \geqslant 3$, we see that $\cos(\pi/n) > 0$, and so from equation (5) we learn

$$k = [2\sin(\pi/n)]/\ell. \tag{6}$$

As the curvature is constant, we conclude (solve the differential equations in (2)) that C is a circle.

It is interesting to examine the above proof in the case $n = 2$. Here $\rho = i$ and so $\omega = \zeta$ as expected. The derivation that ℓ is a constant remains valid and so we obtain a proof of the result we mentioned in the introduction: if there is a periodic billiard 2-gon at each point of a smooth convex curve C (i.e., if C has the double normal property), then C has constant width. Notice that in this case equation (5) gives no information about the curvature k, since $\cos(\pi/2) = 0$.

References

[1] D. W. DeTemple and J. M. Robertson, A billiard characterization of regular polygons, this MAGAZINE, 54 (1981) 73–75.

[2] Y. G. Sinai, Introduction to Ergodic Theory, Princeton University Press, 1976.

[3] R. Sine and V. Kreĭnovič, Remarks on billiards, Amer. Math. Monthly, 86 (1979) 204–206.

A Mean Value Property of the Derivative of Quadratic Polynomials—without Mean Values and Derivatives

J. ACZÉL

University of Waterloo

Waterloo, Ontario, Canada N2L 3G1

Quadratic polynomials $f(x) = ax^2 + bx + c$ have the remarkable property that on their graphs (parabolas) the abscissa of the point where the tangent is parallel to a chord is the arithmetic mean of the abscissas of the endpoints of the chord (see FIGURE 1). In formula, this translates to the following particular mean value property of the derivative of quadratic polynomials:

$$\frac{f(x)-f(y)}{x-y} = f'\left(\frac{x+y}{2}\right) \tag{1}$$

for all real $x \neq y$. In fact, this property is *characteristic* to parabolas; that is, *only* $f(x) = ax^2 + bx + c$ (with arbitrary constants a, b, c) satisfies (1) for all x, y. For a proof see, for instance, [**5**, p. 122], where f was supposed to be three times differentiable. Of course, for (1) to make sense, f has to be differentiable. Then, by (1), it is also twice, three times, actually any number of times differentiable.

On the other hand, (1) is a special case of the equation

$$\frac{f(x)-f(y)}{x-y} = h(x+y) \qquad (x \neq y;\, x, y \in \mathbb{R}) \tag{2}$$

which contains no derivative (and no mean value). So here it does not seem natural anymore to suppose that f (or h) is differentiable. Such equations, which serve to determine unknown functions and don't contain derivatives and integrals, are called *functional equations*. A remarkable feature of functional equations is that one equation can determine several unknown functions, in this case both f and h.

I proved *without any* differentiability or other *regularity conditions* some twenty years ago that the equation (2), containing two unknown functions, also has $f(x) = ax^2 + bx + c$ and $h(x) = ax + b$ as its only solutions. Having done this, I soon forgot about it, but both the result and the

Since $n \geqslant 3$, we see that $\cos(\pi/n) > 0$, and so from equation (5) we learn

$$k = [2\sin(\pi/n)]/\ell. \tag{6}$$

As the curvature is constant, we conclude (solve the differential equations in (2)) that C is a circle.

It is interesting to examine the above proof in the case $n = 2$. Here $\rho = i$ and so $\omega = \zeta$ as expected. The derivation that ℓ is a constant remains valid and so we obtain a proof of the result we mentioned in the introduction: if there is a periodic billiard 2-gon at each point of a smooth convex curve C (i.e., if C has the double normal property), then C has constant width. Notice that in this case equation (5) gives no information about the curvature k, since $\cos(\pi/2) = 0$.

References

[1] D. W. DeTemple and J. M. Robertson, A billiard characterization of regular polygons, this MAGAZINE, 54 (1981) 73–75.

[2] Y. G. Sinai, Introduction to Ergodic Theory, Princeton University Press, 1976.

[3] R. Sine and V. Kreĭnovič, Remarks on billiards, Amer. Math. Monthly, 86 (1979) 204–206.

A Mean Value Property of the Derivative of Quadratic Polynomials—without Mean Values and Derivatives

J. ACZÉL
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

Quadratic polynomials $f(x) = ax^2 + bx + c$ have the remarkable property that on their graphs (parabolas) the abscissa of the point where the tangent is parallel to a chord is the arithmetic mean of the abscissas of the endpoints of the chord (see FIGURE 1). In formula, this translates to the following particular mean value property of the derivative of quadratic polynomials:

$$\frac{f(x) - f(y)}{x - y} = f'\left(\frac{x+y}{2}\right) \tag{1}$$

for all real $x \neq y$. In fact, this property is *characteristic* to parabolas; that is, *only* $f(x) = ax^2 + bx + c$ (with arbitrary constants a, b, c) satisfies (1) for all x, y. For a proof see, for instance, [**5**, p. 122], where f was supposed to be three times differentiable. Of course, for (1) to make sense, f has to be differentiable. Then, by (1), it is also twice, three times, actually any number of times differentiable.

On the other hand, (1) is a special case of the equation

$$\frac{f(x) - f(y)}{x - y} = h(x+y) \qquad (x \neq y;\, x, y \in \mathbb{R}) \tag{2}$$

which contains no derivative (and no mean value). So here it does not seem natural anymore to suppose that f (or h) is differentiable. Such equations, which serve to determine unknown functions and don't contain derivatives and integrals, are called *functional equations*. A remarkable feature of functional equations is that one equation can determine several unknown functions, in this case both f and h.

I proved *without any* differentiability or other *regularity conditions* some twenty years ago that the equation (2), containing two unknown functions, also has $f(x) = ax^2 + bx + c$ and $h(x) = ax + b$ as its only solutions. Having done this, I soon forgot about it, but both the result and the

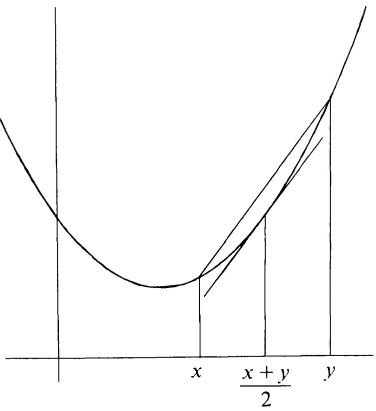


FIGURE 1

proof were preserved in [**1**]. The proof made use of the so-called Jensen functional equation

$$\phi\left(\frac{x+y}{2}\right)=\frac{\phi(x)+\phi(y)}{2}.$$

Without knowing of my result, Shigeru Haruki took up the subject and in 1979 [**4**] went somewhat further. He proved that the more general equation

$$\frac{f(x)-g(y)}{x-y}=h(x+y) \qquad (x\neq y;\, x,y\in\mathbb{R}), \tag{3}$$

containing three unknown functions, also has $f(x)=g(x)=ax^2+bx+c$ and $h(x)=ax+b$ as its only solutions, again without any regularity conditions. Also his proof went by reduction to the Jensen equation.

I was reminded of my 'forgotten' result by P. Volkmann's talk [**6**] at a meeting in August 1983. During the same meeting I found an even simpler way of solving (2), which applies also to (3) and for which the Jensen equation is not needed. This proof, which I want to share with you here, remains valid for fields of characteristic different from 2. You may wish to stay first with real functions of real variables and check later that our proof holds also on the more general structure.

Let us assume that f, g, h are defined for all elements of our field (in particular, for all real numbers, if you so prefer). The equation

$$\frac{f(x)-g(y)}{x-y}=h(x+y) \qquad (x\neq y) \tag{3'}$$

can be written as

$$f(x)-g(y)=(x-y)h(x+y), \tag{4}$$

which shows no trace of derivatives or mean values. Unlike (3′), this equation (4) would make sense also if $x=y$ were permitted, which would give an easy proof that $g(x)=f(x)$, but we can proceed further under the restriction $x\neq y$. By interchanging x and y in (3′) (or (4)), we obtain

$$f(x)-g(y)=g(x)-f(y),$$

so

$$f(x)-g(x)=g(y)-f(y).$$

Since the left-hand side depends only upon x and the right only upon y, both have to be constant, say c. So now $f(t)-g(t)=c$, but also $g(t)-f(t)=c$, thus $c=-c$, that is,

$$2c=0 \quad \text{and so} \quad c=0. \tag{5}$$

Therefore

$$g(x)=f(x) \qquad \text{for all } x. \tag{6}$$

Consequently, equation (4) becomes

$$f(x)-f(y)=(x-y)h(x+y) \tag{7}$$

which is also true for $x=y$. Of course, if f satisfies (7), so does $f+b$ (b is constant). Therefore we may suppose without loss of generality that $f(0)=0$. Put $y=0$ into (7) in order to get

$$f(x)=xh(x). \tag{8}$$

This equation may be used to transform (7) into

$$xh(x)-yh(y)=(x-y)h(x+y). \tag{9}$$

Again, if this is satisfied by h, it is satisfied also by $h+c$, so we may suppose that $h(0)=0$. Therefore, puting $x=-y$ into (9), we get

$$-yh(-y)=yh(y);$$

that is, h is an odd function. We take this into consideration when replacing y by $-y$ in (9), getting

$$xh(x)-yh(y)=(x+y)h(x-y).$$

Comparison with (9) gives $(x-y)h(x+y)=(x+y)h(x-y)$ and substituting

$$u=x+y, \quad v=x-y \tag{10}$$

produces the equation

$$vh(u)=uh(v)$$

for all u,v, thus (choosing $v=v_0$, a nonzero constant, and $a=h(v_0)/v_0$)

$$h(u)=au.$$

If we do not assume $h(0)=0$, we have in general

$$h(u)=au+b.$$

By (8) this gives $f(x)=x(ax+b)$ and, if we do not assume $f(0)=0$, but take (6) into consideration, then

$$f(x)=g(x)=ax^2+bx+c.$$

So we have indeed proved that all solutions of (3′) (and also of (4)) are of the form

$$f(x)=g(x)=ax^2+bx+c, \quad h(x)=ax+b, \tag{11}$$

where a,b,c are arbitrary constants, as asserted. Straightforward substitution shows that all functions of the form (11) satisfy (3′) and (4).

If you considered throughout the proof the variables to be real numbers, you may go through it again, replacing real numbers by elements of an arbitrary field. You will notice that everything works, except that in two steps we need that the characteristic of the field be different from 2. The first is in (5) where $2c=0$ was supposed to imply $c=0$ which is true only if the characteristic is different from 2. The other is the supposition that the system of equations (10) has solutions x,y for arbitrarily given u,v. These ($x=(u+v)/2$, $y=(u-v)/2$) again exist only if the characteristic is different from 2. But that is all that needs to be supposed. So we have proved the following.

THEOREM. *The general solutions of* (3′) *in a field of characteristic different from* 2 *are given by* (11), *where a,b,c are arbitrary constants in the field. The solutions are the same for equation* (4) *in which $x=y$ is allowed.*

For fields of characteristic 2 the theorem is not true. To see this, suppose that our variables move in the field $\mathbb{Z}_2$, the integers modulo 2. Then the functions $f(x)=x$, $g(x)=x+1$, $h(x)=x+1$ satisfy (3′) and (4) for $x\neq y$ (but not for $x=y$), while $f(x)=g(x)=h(x)=x$

satisfy (3′) (for $x \neq y$) and (4) (also for $x = y$), but neither is of the form (11). (The first example is due to Ulrich Daepp.)

REMARKS. There are several interesting generalizations of (1), which do involve both means and derivatives.

One question is: *what kind of mean values* (other than arithmetic means) *can appear on the right-hand side of equations like* (1)? In other words, for what mean values $M(x, y)$ do there exist differentiable functions f such that

$$\frac{f(x)-f(y)}{x-y}=f'[M(x,y)]. \tag{12}$$

R. Bojanić [**3**] has found necessary and sufficient conditions in the form of partial differential equations for M.

Another question arises in considering **quasiarithmetic means**, that is, those which can be represented in the form

$$M(x,y)=F^{-1}\left(\frac{F(x)+F(y)}{2}\right),$$

where F is continuous and strictly monotonic. (For the arithmetic mean, $F(x) = x$.) *When can* M *in* (12) *be a quasiarithmetic mean*? Geometrically, this question asks for what curves will the abscissa of the point where the tangent is parallel to a chord be a quasiarithmetic mean of the abscissas of the endpoints of the chord for every chord. G. Aumann proved [**2**], (see also [**3**]), that arcs of conic sections are exactly the curves with this property (on an interval) even if (12) is replaced by

$$\frac{f(x)-f(y)}{g(x)-g(y)}=\frac{f'[M(x,y)]}{g'[M(x,y)]}.$$

(We talk about *arcs* of conic sections because, for instance, a complete ellipse is not the graph of a *function*, since it is not single valued.) This is a nice generalization of the property of parabolas which served as our point of departure.

References

[1] J. Aczél, Bemerkungen 11–12, Tagungsbericht Funktionalgleichungen 7–11 Okt. 1963 (2. Tagung), Mathematisches Forschungsinstitut Oberwolfach, 1963, p. 14.

[2] G. Aumann, Über einen Zusammenhang zwischen Mittelwerten, Streckenrechnung und Kegelschnitten, Tōhoku Math. J., 42 (1936) 32–37.

[3] R. Bojanić, Sur la formule des acroissements finis, Acad. Serbe Sci. Publ. Inst. Math., 3 (1950) 219–226.

[4] Sh. Haruki, A property of quadratic polynomials, Amer. Math. Monthly, 86 (1979) 577–579.

[5] T. L. Saaty, Modern Nonlinear Equations, McGraw-Hill, New York, 1967.

[6] P. Volkmann, Eine Funktionalgleichung für dividierte Differenzen. The Twenty-first International Symposium on Functional Equations, August 6–13, Konolfingen, Switzerland, Aequationes Math., 26 (1984) 225–294, in part. pp. 252–253.

0 For A Solution!

The solution to my problem I sought,
When I hit on a brilliant thought.
By a method involved
The equation was solved
And the answer was: nought equals nought.

—MARTA SVED

satisfy (3′) (for $x \neq y$) and (4) (also for $x = y$), but neither is of the form (11). (The first example is due to Ulrich Daepp.)

REMARKS. There are several interesting generalizations of (1), which do involve both means and derivatives.

One question is: *what kind of mean values* (other than arithmetic means) *can appear on the right-hand side of equations like* (1)? In other words, for what mean values $M(x, y)$ do there exist differentiable functions f such that

$$\frac{f(x)-f(y)}{x-y}=f'[M(x,y)]. \tag{12}$$

R. Bojanić [3] has found necessary and sufficient conditions in the form of partial differential equations for M.

Another question arises in considering **quasiarithmetic means**, that is, those which can be represented in the form

$$M(x,y)=F^{-1}\left(\frac{F(x)+F(y)}{2}\right),$$

where F is continuous and strictly monotonic. (For the arithmetic mean, $F(x)=x$.) *When can M in* (12) *be a quasiarithmetic mean*? Geometrically, this question asks for what curves will the abscissa of the point where the tangent is parallel to a chord be a quasiarithmetic mean of the abscissas of the endpoints of the chord for every chord. G. Aumann proved [**2**], (see also [**3**]), that arcs of conic sections are exactly the curves with this property (on an interval) even if (12) is replaced by

$$\frac{f(x)-f(y)}{g(x)-g(y)}=\frac{f'[M(x,y)]}{g'[M(x,y)]}.$$

(We talk about *arcs* of conic sections because, for instance, a complete ellipse is not the graph of a *function*, since it is not single valued.) This is a nice generalization of the property of parabolas which served as our point of departure.

References

[1] J. Aczél, Bemerkungen 11–12, Tagungsbericht Funktionalgleichungen 7–11 Okt. 1963 (2. Tagung), Mathematisches Forschungsinstitut Oberwolfach, 1963, p. 14.

[2] G. Aumann, Über einen Zusammenhang zwischen Mittelwerten, Streckenrechnung und Kegelschnitten, Tōhoku Math. J., 42 (1936) 32–37.

[3] R. Bojanić, Sur la formule des acroissements finis, Acad. Serbe Sci. Publ. Inst. Math., 3 (1950) 219–226.

[4] Sh. Haruki, A property of quadratic polynomials, Amer. Math. Monthly, 86 (1979) 577–579.

[5] T. L. Saaty, Modern Nonlinear Equations, McGraw-Hill, New York, 1967.

[6] P. Volkmann, Eine Funktionalgleichung für dividierte Differenzen. The Twenty-first International Symposium on Functional Equations, August 6–13, Konolfingen, Switzerland, Aequationes Math., 26 (1984) 225–294, in part. pp. 252–253.

0 For A Solution!

The solution to my problem I sought,
When I hit on a brilliant thought.
By a method involved
The equation was solved
And the answer was: nought equals nought.

—MARTA SVED

PROBLEMS

LEROY F. MEYERS, Editor
G. A. EDGAR, Associate Editor
The Ohio State University

Proposals

To be considered for publication, solutions should be mailed before June 1, 1985.

1206. Let ABC be a triangle with sides a, b, and c and semiperimeter s. Let the side BC be subdivided using the points $B = P_0, P_1, \ldots, P_{n-1}, P_n = C$ in order. If r_i is the inradius of triangle $AP_{i-1}P_i$ for $i = 1, \ldots, n$, prove that

$$r_1 + \cdots + r_n < \frac{1}{2} h_a \ln \frac{s}{s-a},$$

where h_a is the length of the altitude from vertex A. [*Hüseyin Demir*, *Middle East Technical University*, *Ankara*, *Turkey*.]

1207. Prove that for each positive integer K there exist infinitely many even positive integers which can be written in more than K ways as the sum of two odd primes. [*Barry Powell*, *Kirkland*, *Washington*.]

1208. Prove that if a and b are positive, then

$$\prod_{k=1}^{n} (a^k + b^k)^2 \geqq (a^{n+1} + b^{n+1})^n.$$

[*Mihály Bencze*, *Săcele*, *Romania*.]

1209. Evaluate

$$\int_0^\infty \frac{\sqrt{x} \log x}{(1+x)^2} \, dx.$$

[*Themistocles M. Rassias*, *Athens*, *Greece*.]

1210. For a fixed integer $n \geqq 3$, consider the polynomials $f(x)$ with rational coefficients and degree less than n such that $|f(\omega)| = 1$ whenever ω is an nth root of unity. Must there be infinitely many such polynomials $f(x)$? [*J. Rosenblatt*, *The Ohio State University*.]

We invite readers to submit problems believed to be new. Proposals should be accompanied by solutions, if at all possible, and by any other information that will assist the editors. A problem submitted as a Quickie should have an unexpected, succinct solution. An asterisk (*) will be placed next to a problem number to indicate that the proposer did not supply a solution.

Solutions should be written in a style appropriate for Mathematics Magazine. Each solution should begin on a separate sheet containing the solver's name and full address. It is not necessary to submit duplicate copies.

Send all communications to the problems department to Leroy F. Meyers, Mathematics Department, The Ohio State University, 231 W. 18th Ave., Columbus, Ohio 43210.

Solutions

A Determinant Evaluated by Legendre Polynomials — March 1983

1165. For real numbers a, b with $a < b$ and positive integers n consider the matrix $A_n = (a_{ij})_{i,j=0,1,\ldots,n}$ where $a_{ij} = \int_a^b x^{i+j}\,dx$. Prove that

$$\det A_n = \frac{n!\cdot 2^n}{(2n+1)!}(b-a)^{(n+1)^2}\prod_{k=1}^{n}\binom{2k}{k}^{-2}.$$

[*Heinz-Jürgen Seiffert, student, Freie Universität Berlin.*]

Solution: Let $(f,g) = \int_a^b f(x)g(x)\,dx$ for all f, g in the space of real-valued functions continuous on $[a,b]$. The $(n+1)$-dimensional subspace of polynomials of degree at most n is generated by the orthonormal set of Legendre polynomials $p_0, p_1, \ldots, p_n$ suitably normalized, i.e., $(p_i, p_j) = \int_a^b p_i(x)p_j(x)\,dx = \delta_{ij}$ (Kronecker delta). Now define the lower triangular matrix $B_n = (b_{ij})_{i,j=0,1,\ldots,n}$ by $x^i = \sum_{k=0}^n b_{ik}p_k(x)$ for $i = 0,1,\ldots,n$. Note that the leading coefficient of p_i equals b_{ii}^{-1}. Then

$$a_{ij} = (x^i, x^j) = \sum_{k=0}^{\min(i,j)} b_{ik}b_{jk},$$

so that $A_n = B_nB_n^t$, where B_n^t is the transpose of B_n. Because B_n is triangular, we deduce

$$\det A_n = \prod_{k=0}^{n} b_{kk}^2.$$

Since the leading coefficient of p_k is (see Abramowitz and Stegun, *Handbook of Mathematical Functions*, pp. 773–775, and adjust interval and normalization)

$$\binom{2k}{k}(2k+1)^{1/2}(b-a)^{-k-1/2},$$

the required formula for $\det A_n$ now readily follows.

R. J. STROEKER
Erasmus University, The Netherlands

Also solved by H. G. Mushenheim, Paul J. Zwier, and the proposer.

An Easy Trigonometric Inequality — January 1984

1182. Prove that

$$x\cot\frac{x}{2} - x\tan^3\frac{x}{2} < 2 \quad \text{for } 0 < x < \frac{\pi}{2}.$$

[*R. S. Luthar, University of Wisconsin Center, Janesville.*]

Solution: Substituting $\theta = \frac{1}{2}x$ in the well known inequality $\tan\theta > \theta$ for $0 < \theta < \frac{1}{2}\pi$ yields $\tan\frac{1}{2}x > \frac{1}{2}x$ for $0 < x < \pi$. Hence

$$2 > x\cot\frac{x}{2} > x\cot\frac{x}{2} - x\tan^3\frac{x}{2} \quad \text{for } 0 < x < \pi.$$

In fact, the inequality holds for $0 < |x| < \pi$, and is equivalent to $x\cot x < \cos^2(x/2)$ and, for $0 < x < \frac{1}{2}\pi$, to $\sin x + \tan x > 2x$.

EDITORS' COMPOSITE

Fifty-six correct solutions (including the proposer's) were received, of which nine were by students. There were three incorrect solutions.

Most solvers used calculus to show that certain functions connected with the inequality were monotonic. A popular function was given by $f(x)=\sin x+\tan x-2x$ for $0\le x<\frac{1}{2}\pi$. The following strengthenings were submitted:

$$x\cot\frac{x}{2}-x\tan^3\frac{x}{2}<2-\frac{8}{5}\tan^4\frac{x}{2}\qquad \text{for } 0<|x|<\frac{\pi}{2}\qquad \text{(Erhard Braune, Austria);}$$

$$x\cot\frac{x}{2}-x\tan^3\frac{x}{2}<2-\frac{2}{3}\tan^2\frac{x}{2}\qquad \text{for } 0<|x|<\frac{\pi}{2}\qquad \text{(Padmini T. Joshi);}$$

$$x\cot\frac{x}{2}+\left(\frac{4}{\pi}-1\right)x\tan\frac{x}{2}<2\qquad \text{for } 0<|x|<\frac{\pi}{2}\qquad \text{(Murray S. Klamkin, Canada);}$$

$$\frac{3(1+\cos y)(4+\cos y)}{(19+11\cos y)\sin y}<\frac{1}{y}<\frac{1}{3}(\cot y+2\csc y)\quad \text{for } 0<y<\frac{\pi}{2} \text{ with } y=x/2\qquad \text{(Robert E. Shafer).}$$

Chance Solitaire — January 1984

1183. *Solitario de la suerte* (chance solitaire) is a Spanish game similar to "clock." (Cf. in this MAGAZINE, problem 1066, solved v. 53 (1980) 184–185, and the articles by Jenkins & Miller, v. 54 (1981) 202–208, and by Ecker, v. 55 (1982) 42–43.) A version using American cards can be described mathematically as follows.

Forty-eight cards from a well-shuffled deck are placed face down in the first twelve columns of a 4×13 rectangular array. The rows are named after the suits (clubs, diamonds, spades, hearts) and the columns are named after the ranks (ace, two,...,king). The remaining four cards of the deck form the *hand*. A card is chosen from the hand and placed face up in the array in its proper position according to its suit and rank, replacing the face-down card, which is now placed in the array according to its suit and rank, etc., until a king is placed. The process is repeated for the second, third, and fourth cards in the hand. When all cards have been placed, the remaining face-down cards are turned over. The game is won if all cards are in their proper places, and lost otherwise.

(a) What is the probability of winning the game?

*(b) If the hand is known to contain exactly i kings, where $0\leqslant i\leqslant 4$, what is the probability of winning? [*Julio Castiñeira, Segovia, Spain.*]

Solution I: (a) The game is unchanged if we place the cards face down in a single row. Consider the more general case of a deck of n cards, consisting of k kings and $n-k$ non-kings, where $n\ge 2k$ and the hand consists of k cards, the remaining $n-k$ cards forming the array.

If there are j face down cards remaining when the last king appears, then the probability of winning is $1/j!$, since there is only one winning order for these j cards. The probability that there are j face down cards remaining when the last king appears is $\binom{n-j-1}{k-1}\Big/\binom{n}{k}$, for $k\ge 1$ and $0\le j\le n-k$, since $k-1$ kings must occur in the first $n-j-1$ cards placed, but the k kings might have occurred anywhere in the array or in the hand. Hence $p(n,k)$, the probability of winning, is given by

$$p(n,k)=\begin{cases}1/n! & \text{if } k=0,\\[4pt] \displaystyle\sum_{j=0}^{n-k}\frac{1}{j!}\frac{\binom{n-j-1}{k-1}}{\binom{n}{k}} & \text{if } k\ge 1.\end{cases}$$

A classical argument, based on the situation after the first card is turned up, establishes the recursion

$$p(n,k)=\frac{k}{n}p(n-1,k-1)+\frac{n-k}{n}p(n-1,k)\quad \text{for } 1\le k<\frac{1}{2}n.$$

(b) If there are no kings in the hand, the probability that there will be j face down cards remaining when the last king appears is $\binom{n-k-j-1}{k-1}\Big/\binom{n-k}{k}$ for $k \geq 1$ and $0 \leq j \leq n-2k$, since all kings must appear in the array. Hence $p_0(n,k)$, the probability of winning when there are 0 kings in the hand, is given by

$$p_0(n,k) = \begin{cases} 1/n! & \text{if } k = 0, \\ \displaystyle\sum_{j=0}^{n-2k} \frac{1}{j!} \frac{\binom{n-k-j-1}{k-1}}{\binom{n-k}{k}} & \text{if } k \geq 1, \end{cases}$$

which is easily seen to equal $p(n-k,k)$.

If we think of a hand card as a "king annihilator", the kings in the hand annihilate themselves, and so $p_i(n,k)$, the probability of winning when there are i kings in the hand, satisfies

$$p_i(n,k) = p_0(n-i,k-i) = p(n-k,k-i) \quad \text{for } 0 \leq i \leq k.$$

VÍCTOR HERNÁNDEZ
Universidad Autónoma de Madrid, Spain

Solution II: (a) The game may be simulated by simply turning over the cards one at a time from the top of a shuffled deck, stopping when the fourth king appears. The first card turned corresponds to the card played initially from the hand, the second card turned corresponds to the card found at the position indicated by the first card, and so on. After a king is turned, the next card turned corresponds to the next card played from the hand. When the game is blocked by the fourth king, the unturned cards determine win or loss—a win only if the remaining cards are in their "proper" positions, only one possible arrangement being "proper."

There are $4!\binom{j-1}{3}\frac{48!}{(52-j)!}$ winning shuffles (out of 52!) with the fourth king in position j $(4 \leq j \leq 52)$. The binomial coefficient positions the preceding three kings, the 4! arranges the four kings, and the factorial quotient fills in the $j-4$ positions before the fourth king with non-kings. Hence the probability of winning is

$$\begin{aligned} p(52,4) &= \frac{4!}{52!} \sum_{j=4}^{52} \binom{j-1}{3} \frac{48!}{(52-j)!} \\ &= \frac{15893129561187559327146420289830606692788916475859332570444041791512}{8065817517094387857166063685640376697528950544088327782400000000000000} \\ &\approx 19.704\%. \end{aligned}$$

(b) If the hand is known to hold four kings (i.e., the top four cards in the simulation are kings), then the probability of winning is

$$\begin{aligned} p_4(52,4) &= \frac{1}{48!} = \frac{1}{12413915592536072670862289047373375038521486354677760000000000} \\ &\approx 8.0555 \cdot 10^{-60}\%. \end{aligned}$$

If the hand is known to hold i kings, where $0 \leq i \leq 3$, then the number of winning shuffles with the fourth king in position j is $4!\binom{j-5}{3-i}\binom{4}{4-i}\frac{48!}{(52-j)!}$. The $j-4$ non-kings among the first j cards form $4-i$ nonempty blocks, separated and terminated by $5-i$ blocks of kings. There are $\binom{j-5}{3-i}$ such arrangements of non-kings. Now $4-i$ of the king blocks must be nonempty, so that there are $\binom{4}{4-i}$ arrangements of the kings. The number of shuffles with i kings in the hand is $4!48!\binom{48}{4-i}\binom{4}{4-i}$, since the non-kings following the fourth king must be considered. Hence the probability of winning is

$$p_i(52,4) = \frac{4! \sum_{j=8-i}^{52} \binom{j-5}{3-i}\binom{4}{4-i} \frac{48!}{(52-j)!}}{4!48!\binom{48}{4-i}\binom{4}{4-i}}.$$

Calculations give:

$$\begin{aligned} p_3(52,4) &= \frac{3239474032820659927738641541588967443708779877198579838726215680}{5720332305040622286733342793029651217750700912235511808000000000} \\ &\approx 5.663\%, \\ p_2(52,4) &= \frac{22352370826462553501396626636963875361590581152670200887210888832}{201641713752681935607350333454295205425712207156301791232000000000} \\ &\approx 11.085\%, \\ p_1(52,4) &= \frac{335447536098579335517336331631537578796044156283912942300099643801 6}{2061226407249637563986247853088350988796169228708862754816000000000} \\ &\approx 16.274\%, \\ p_0(52,4) &= \frac{1231189101790431977703135206560400318376886032161630255873221024896 0}{5797199270389605648711322086810987155989225955743676497920000000000} \\ &\approx 21.238\%. \end{aligned}$$

Richard Parris
Phillips Exeter Academy

Also solved by Jordi Dou (Spain), Thomas S. Ferguson (part (a) only), Paul O'Hara (part (b) only), Robert S. Stacy (West Germany, partially), Anders Szepessy (Sweden), and the proposer.

By setting $S(n,k) = \binom{n}{k} p(n,k)$, Ferguson obtained the recursion

$$S(n,k) = S(n-1,k-1) + S(n-1,k) \quad \text{for } n \geq k \geq 1, \quad \text{and} \quad S(n,0) = 1/n!.$$

Dou, Ferguson, Szepessy, and the proposer obtained the asymptotic upper estimate ke/n for $p(n,k)$, which Ferguson improved to $ke^{(n-k)/(n-1)}/n$.

No Solutions in Integers — January 1984

1184. (a) Show, using results from elementary number theory only (such as congruence conditions), that the equations

$$x^2 + 5 = y^3, \qquad x^2 + 9 = y^3, \qquad x^2 - 71 = y^7, \qquad x^2 + 37 = y^{11}, \quad \text{and} \quad x^4 + 282 = y^3$$

have no solutions in integers x, y.

(b) Of what general results are the examples in (a) corollaries? [*Barry Powell, Kirkland, Washington.*]

Solution: Rewrite the first four given equations as follows: $x^2 + 4 \cdot 1^2 = y^3 - 1^3$, $x^2 + 4 \cdot 3^2 = y^3 - (-3)^3$, $x^2 + 4 \cdot 23^2 = y^7 - (-3)^7$, $x^2 + 4 \cdot 3^2 = y^{11} - 1^{11}$. In each case the following theorem applies.

THEOREM. *Consider the equation*

$$x^2 + 4a^2 = y^p - b^p, \tag{1}$$

where a, b, and p are specified integers such that $b \equiv 1 \pmod 4$, $p \equiv 3 \pmod 4$, p is prime, and if q is any prime divisor of a such that $q \equiv 3 \pmod 4$, then $q^p \nmid a^2$ and $p \nmid (q-1)$ (if $q = p$, then also $q | b$). Then (1) *has no solution in integers x, y.*

Proof. Suppose on the contrary that x, y satisfy (1). Since $y^p \equiv x^2 + 1 \pmod 4$, we conclude that $y \equiv 1 \pmod 4$. Hence $y^p - b^p = (y-b)B$, where $B = y^{p-1} + by^{p-2} + \cdots + b^{p-1} \equiv p \equiv$

3 (mod 4). Thus B has a prime factor $q \equiv 3 \pmod 4$. Since $q | (x^2 + 4a^2)$, by the Fermat two-square theorem q must divide a. [A prime congruent to 3 modulo 4 and dividing a sum of two squares must divide each square separately. This is elementary number theory.—*Ed.*] If q divides y or b, then it divides the other and, by the two-square theorem, $q^p | a^2$, contrary to the restrictions imposed on q. Hence $q \nmid yb$ and so $q \neq p$. Let u be chosen so that $yu \equiv 1 \pmod q$. Then $(bu)^p \equiv 1 \pmod q$, so that the order of $bu \pmod q$ is 1 or p. By Fermat's little theorem the order would have to be a divisor of $q - 1$, contrary to the restrictions on q unless $bu \equiv 1 \pmod q$, in which case $b \equiv y \pmod q$ and $0 \equiv B \equiv py^{p-1} \pmod q$ and thus $q | y$, again an absurdity.

Rewrite the fifth equation as $x^4 + 5^4 = y^3 + 7^3$ and suppose that a solution in integers x, y exists. Easily $y \equiv 3 \pmod 8$, so that $y^3 + 7^3 = (y + 7)C$ where $C \equiv 5 \pmod 8$. Then C has a prime factor $q \equiv \pm 5 \pmod 8$. Now $C \equiv (y - 1)^2 - 2 \not\equiv 0 \pmod 5$, so that $q \neq 5$. If v is chosen so that $5v \equiv 1 \pmod q$, then $(xv)^4 \equiv -1 \pmod q$, $(xv)^8 \equiv 1 \pmod q$, and, by Fermat's little theorem, $8 | (q - 1)$, a contradiction.

L. E. Dickson, *History of the Theory of Numbers*, v. 2, ch. XX, esp. pp. 534–535, mentions, among others, theorems of A. de Jonquières and T. Pepin on equations of the form $x^2 + a = y^3$. Siegel's theorem (see L. J. Mordell, *Diophantine Equations*, p. 264) implies that all of the proposed equations have but finitely many solutions.

LORRAINE L. FOSTER
California State University, Northridge

Also solved by Tony Costa (student, part (a) only), L. Kuipers (Switzerland, part (a) only) and the proposer. There were two incorrect solutions.

The proposer obtained the following results (and more). Let p be an odd prime. If $p \equiv 3 \pmod 4$, a is even, $p \nmid a$, $K \not\equiv 7 \pmod 8$, $(a, K) = 1$, and $a^2 - K = b^p$, then there are no integers x, y such that $x^2 + K = y^p$. On the other hand, if $p \equiv 3 \pmod 4$, $p \nmid a$, $K \equiv 2 \pmod 8$, $(a, K) = 1$, and $a^4 - K = b^p$, then there are no integers x, y such that $x^4 + K = y^p$. The existence of solutions of one equation implies the nonexistence of solutions of a similar equation.

A Very Slowly Converging Sequence — January 1984

***1185.** Set $a_0 = 1$ and for $n \geqslant 1$, $a_n = a_{n'} + a_{n''} + a_{n'''}$, where $n' = [n/2]$, $n'' = [n/3]$, and $n''' = [n/6]$. Find $\lim_{n \to \infty} a_n/n$. (Compare solution 1158, v. 57 (1984), p. 49) [*Daniel A. Rawsthorne, Wheaton, Maryland.*]

Solution: Let $(a(n))_{n=0}^{\infty}$ be the sequence defined by

$$a(0) = 1, \quad \text{and} \quad a(n) = a\left(\left[\frac{n}{2}\right]\right) + a\left(\left[\frac{n}{3}\right]\right) + a\left(\left[\frac{n}{6}\right]\right) \quad \text{for } n \geq 1,$$

where the brackets denote the greatest integer function. We show that $\lim_{n \to \infty} a(n)/n$ exists and equals $12/\log 432$.

First, let $A(x) = a([x])$ for $x > 0$. Since $[x/m] = [[x]/m]$ for all positive integers m, we have

$$A(x) = 1 \quad \text{for } 0 < x < 1, \quad \text{and } A(x) = A\left(\frac{x}{2}\right) + A\left(\frac{x}{3}\right) + A\left(\frac{x}{6}\right) \quad \text{for } x \geq 1.$$

Now let $f(x) = A(x)/x$ for $x > 0$ and $g(u) = f(e^u)$ for all real u. Then

$$g(u) = \begin{cases} e^{-u} & \text{if } u < 0, \\ \frac{1}{2}g(u - \log 2) + \frac{1}{3}g(u - \log 3) + \frac{1}{6}g(u - \log 6) & \text{if } u \geq 0. \end{cases} \tag{1}$$

Thus g satisfies a renewal equation

$$g(u) = h(u) + \int_0^u g(u - v)\, F\{dv\} \quad \text{for } u > 0, \tag{2}$$

where the probability measure $F\{dv\}$ has mass m^{-1} at $\log m$ for $m \in \{2,3,6\}$ and $h(u)$ is the discrepancy between the full recurrence (1) and that part which comes from the convolution in (2). It is not hard to see that

$$h(u) = \sum_{m \in \{2,3,6\}} e^{-(u - \log m)} \frac{1}{m} \chi_{[0, \log m)}(u) = \sum_{m \in \{2,3,6\}} e^{-u} \chi_{[0, \log m)}(u),$$

where χ_S denotes the characteristic function of the set S. The limiting properties of $g(u)$ as u goes to ∞ are well known; using, for example, formula (1.17) on p. 362 of W. Feller's *An Introduction to Probability Theory*, v. 2, §XI.1, 1971, we have

$$\lim_{n \to \infty} \frac{a(n)}{n} = \lim_{u \to \infty} g(u) = \frac{\int_0^\infty h(u)\,du}{\int_0^\infty yF\{dy\}} = \frac{\frac{1}{2}(2-1) + \frac{1}{3}(3-1) + \frac{1}{6}(6-1)}{\frac{1}{2}\log 2 + \frac{1}{3}\log 3 + \frac{1}{6}\log 6} = \frac{12}{\log 432}.$$

We are writing a paper inspired by this problem and its generalizations. Suppose $(a(n))_{n=0}^{\infty}$ is a sequence defined by

$$a(0) = 1, \quad \text{and} \quad a(n) = \sum_{i=1}^{s} r_i a\left(\left[\frac{n}{m_i}\right]\right) \quad \text{for } n \geq 1, \tag{3}$$

where $r_i > 0$ and the m_i's are integers greater than 1. Let τ be the unique real number defined by $\sum_{i=1}^{s} r_i / m_i^{\tau} = 1$, and let $p_i = r_i / m_i^{\tau}$, so that $p_i > 0$ and $\sum_{i=1}^{s} p_i = 1$. As before, let $A(x) = a([x])$ and $f(x) = A(x)/x^{\tau}$. We must distinguish two cases, depending on the denominators m_i. Either there are an integer d and integers v_i such that $m_i = d^{v_i}$ (the lattice case) or no such integers exist and $(\log m_j)/\log m_k$ is irrational for some (j, k) (the ordinary case). Using methods similar to those above, we show the following. In the ordinary case, if $\tau \neq 0$, then

$$\lim_{u \to \infty} f(u) = \lim_{n \to \infty} \frac{a(n)}{n^{\tau}} = \frac{\sum_{i=1}^{s} p_i m_i^{\tau} - 1}{\tau \sum_{i=1}^{s} p_i \log m_i}. \tag{4}$$

(If $\tau = 0$, then $\sum_{i=1}^{s} r_i = 1$, so that $a(n) = 1$ for all n, which is consistent with the limiting case of (4) and (5) below.) In the lattice case, $a(n)$ is constant for $d^k \leq n < d^{k+1}$, so there is no hope that the limit exists as in (4). However,

$$\lim_{k \to \infty} \frac{a(d^k)}{d^{k\tau}} = \frac{\sum_{i=1}^{s} p_i m_i^{\tau} - 1}{\sum_{i=1}^{s} p_i \log m_i} \cdot \frac{d^{\tau} \log d}{d^{\tau} - 1}. \tag{5}$$

We also consider the rate of convergence in (4). In particular, for the given sequence we have $f(432^k) - f(432^k - 1) \sim (5\pi k/6)^{-1/2}$ for large k, where $f(n) = a(n)/n$.

Finally, we may use (4) to answer problem 1158 (Jan. 1984, pp. 49–50) more explicitly. We have $s = 2$, $r_1 = r_2 = 1$, $m_1 = 2$, and $m_2 = 3$. Then $\tau \approx .788$ and $\lim_{n \to \infty} a(n)/n^{\tau} = ((2^{-\tau}\log 2 + 3^{-\tau}\log 3)\tau)^{-1} \approx 1.469$.

P. ERDŐS
Hungarian Academy of Sciences
A. ODLYZKO
AT&T Bell Laboratories

A. HILDEBRAND
P. PUDAITE
B. REZNICK
University of Illinois

Also solved by Noam Elkies (student), who used Dirichlet series and the residue theorem; and partially (under the assumption that the limit exists) by Don Coppersmith, who gave the explicit formula

$$a_n = 1 + 2\sum \frac{(r+s+t)!}{r!s!t!},$$

where the sum is extended over all triples (r, s, t) of nonnegative integers such that $2^r 3^s 6^t \leq n$.

REVIEWS

Paul J. Campbell, Editor
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of the mathematics literature. Readers are invited to suggest items for review to the editors.

Kolata, Gina, *Surprise proof of an old conjecture*, Science 225 (7 September 1984) 1006-1007.

The Bieberbach conjecture (1916) has been proved by L. de Branges (Purdue). The new theorem concerns one-to-one analytic functions on the unit disk which fix the origin. The theorem asserts that the kth coefficient in the power series of such a function is $\leq k$ in absolute value. In the past de Branges offered wrong proofs for other problems; all the same, it is quite disturbing that he could not get any serious hearing in the U.S. for his new proof: "the Soviets proved to be a patient audience." How would Galois fare today with the American mathematical establishment?

Kolata, Gina, *A fast way to solve hard problems*, Science 225 (21 September 1984) 1379-1380.

Five years ago L.G. Khachiyan discovered a polynomial-time algorithm for linear programming, but it proved to be far slower than the exponential-time simplex method. Now N. Karmarkar (AT&T Bell Labs) has found a practical polynomial-time algorithm *far* faster than the simplex method. This may be the algorithmic discovery of the decade!

Dorfman, Robert, *The discovery of linear programming*, Annals of the History of Computing 6 (1984) 283-295.

Linear programming was discovered independently at least twice, by L.V. Kantorovich in 1939 and by G.B. Dantzig (and several colleagues) in 1947. What is fascinating is that the two formulations differ substantially (though they are very closely related mathematically), and the differences reflect the different economic systems and cultural milieux of the Soviet Union and the U.S.

Kolata, Gina, *The proper display of data*, Science 226 (12 October 1984) 156-157.

Offers constructive suggestions for more-effective graphical presentation of data.

Peterson, Ivars, *In search of speedier searches*, Science News 126 (15 September 1984) 170-171.

Discusses pros and cons of two strategies for dynamic reorganization of frequently-searched sequential files, the "move-to-front" and the "frequency-count" strategies. Because of their dynamic nature, the efficiency of the resulting self-adjusting data structures is difficult to analyze. Current research centers on similar ideas for binary search trees.

Computer Software. Scientific American 251:3 (September 1984).

Special single-topic issue devoted to computer software, featuring articles on computer software in a variety of contexts, such as information management, natural language, process control, and intelligent systems.

Holden, Constance, *Will computers transform schools?*, Science 225 (20 July 1984) 296.

Survey of viewpoints of participants at a recent conference, who predicted "that the integration of home and school use of technology will profoundly affect basic education." This hopeful outlook for the future contradicted their assessments of the present: "this movement is all being driven by business interests other than education," the sexual and socioeconomic differentiation of children's computer usage continues, and "instructional fare generally has no appeal unless presented (like 'Math Blaster') in arcade game formats."

"Mathematical Sciences, Computer Research," NSF Annual Report for 1983, pp. 58-65.

Items reported on for 1983 include research in chaotic behavior, connections between dynamical systems and coding, nonparametric approaches to pattern recognition, concurrent programming languages, and symbolic and algebraic manipulation.

Peterson, Ivars, *Super problems for supercomputers*, Science News 126 (29 September 1984) 200-203.

Large-scale computing is becoming fundamental in many areas of research: "By 1988, new computers like the Cray-2 will allow designers to tackle a complete aircraft or a helicopter rotor by tracking what happens on a million-point grid... ." Will increasing speed and memory of computers also provoke a qualitative change in how science is done? Some researchers suggest that "the use of supercomputers and the resulting simulations and graphics already represent a third way of doing science, alongside theory and experience."

Koffman, Elliot B., *et al, Recommended curriculum for CS1*, 1984, Communications of the ACM 27 (1984) 998-1001.

Official report of the ACM Curriculum Committee Task Force on what should go into "a first course in computer science that emphasizes programming methodology and problem solving." The course is specifically designed for computer science majors; "it is not designed to be a computer literacy course or a service course for other disciplines." But are students in other disciplines the only ones who need to consider social and ethical implications of computer usage? Surely the future "high priests" of the programming craft require as much--some say even more--sensitization to such issues at an early stage and throughout their studies. The omission from the new CS1 of any consideration of the human dimensions of computer usage is a great misfortune; without such a component this course is unsuitable as "the general introductory course in computing," for potential majors or anyone else.

Ralston, Anthony, *The first course in computer science needs a mathematics corequisite*, Communications of the ACM 27 (1984 1002-1005.

Ralston details the specific topics in mathematics needed to support the first course in computer science. He suggests that the best way for students in that course to get the math they need is in "a one-year course in discrete mathematics pitched at an intellectual level equivalent to that of an introductory calculus course." [Of his list of topics 70% are still covered in good four-year non-calculus programs in high school, often at a higher intellectual level than the typical introductory college calculus course.]

Kac, Mark, *Marginalia: How I became a mathematician*, American Scientist 72 (1984) 498-499 (modified from Rehovot 9 [1981/82] 26-29).

Each of us who is a mathematician has a personal tale of how we came to the profession. Gauss's tale is well known; the story Kac tells here is also classic. We will not spoil the elegance of his story by recounting it here. Let it suffice to note the ingredients: an initial inclination to engineering, an acute obsession with a mathematical problem, and the crucial element of personal intervention and good guidance by a professional mathematician. One may rationalize past choices in many ways, but survival is its own justification: "Had I gone into engineering I would have, unquestionably, shared the fate of my family and six million others." This article is excerpted from Kac's forthcoming autobiography, Enigmas of Chance (Harper & Row).

Tierney, John, *"Paul Erdös is in town. His brain is open."* Science 84 (October 1984) 40-47.

"It seems strange, or at least unfair, that the public has no idea what Paul Erdös has done in the past 50 years -- or what any mathematician has done this century." Whether still another article on mathematics' best-known eccentric will do much to change that state of affairs is doubtful; in fact, this article is more likely simply to reinforce the stereotype of mathematician as weirdo.

Jaffe, Arthur M., *Ordering the universe: the role of mathematics*, Notices of the AMS 31 (October 1984) 589-608.

Striking essay reporting some developments at the forefront of mathematics and science. Jaffe emphasizes two themes: "Excellent mathematics, however abstract, leads to practical applications in nature... . It is impossible to predict just where an area of mathematics will be useful." He cites areas from Fourier analysis to computer-assisted proofs, from group theory to coding theory.

McDermott, Lillian C., *Research on conceptual understanding in mechanics*, Physics Today 37:7 (July 1984) 24-32.

Summarizes recent research on the conceptual obstacles students experience with the concepts of force, acceleration, and velocity. Students often reason on the basis of intuitive preconceptions which conflict with the concepts of physics "and have proved to be highly resistant to instruction." Since students bring the same misconceptions to the quantitative study of these concepts in *calculus* class, calculus teachers will benefit from being aware of the research results noted in this article.

Sorensen, Peter R., *Fractals: exploring the rough edges between dimensions*, Byte 9:10 (September 1984) 157-172.

Major popular article on fractals, replete with color illustrations and a Basic program to draw one family of fractals.

Renewing U.S. Mathematics: Critical Resource for the Future, Report of the Ad Hoc Committee on Resources for the Mathematical Science, The Commission on Physical Science, Mathematics, and Resources, National Research Council, National Academy Press, 1984 (available from Board on Mathematical Sciences, 2101 Constitution Ave., Washington, DC 20418); xiv + 207 pp. (P).

Known popularly as "the David report," after the chairman of the Authoring Committee, this document describes the disastrous decline in federal support for mathematical research over the past 15 years. It proposes an ambitious plan for renewal and specific budget proposals to Congress and the Administration. The appendices include A. Jaffe's notable essay, reviewed above.

Lam, L.Y., and Shen, S.K., *Right-angled triangles in ancient China*, Archive for History of Exact Sciences 30 (1984) 87-112.

Contrary to van der Waerden's hypothesis of transmission of a Pre-Babylonian mathematics from Central Europe to China and elsewhere, the authors assert that the perfection and high level of reasoning about right triangles in ancient China suggest an opposite conclusion: "a strong case to presume that fragments of these materials were transmitted westward."

Wagner, David L. (ed.), The Seven Liberal Arts in the Middle Ages, Indiana U. Pr., 1983; xii + 282 pp, $25.

Two chapters treat medieval mathematics: "Arithmetic" by Michael Masi and "Geometry" by Lon R. Shelby.

Arganbright, Dean E., Mathematical Applications of Electronic Spreadsheets, McGraw-Hill, 1985; x + 165 pp, $16.95 (P).

So your college administration has finally agreed that your students need more access to microcomputers; but instead of buying the Apple II with UCSD Pascal your department has been requesting for the past several years (or the VAX you really covet), the president has bought: 100 IBM PC's for the business department, with no compilers, just a word processor and the VisiCalc (or Lotus 1-2-3) spreadsheet. All is not lost!!! This book tells you how to use that "business software" to find Taylor polynomials, solve differential equations, find eigenvalues, draw contour graphs, do linear programming, and solve boundary-value problems. Best of all, according to the author "the use of the electronic spreadsheet in mathematics encourages the *doing* of mathematics".

Huck, Schuyler W., and Sandler, Howard M., Statistical Illusions: Problems, and Statistical Illusions: Solutions, Harper & Row, 1984; xiii + 175 pp, $10.95 (P), and ix + 174 pp, (P).

This pair of books features 100 exercises in statistics for which "commonsense" and "intuition" will lead the reader astray. The problems presented are paradoxes or illusions; the pedagogical appeal is to the reader's desire to know "what went wrong" when the true answer does not agree with the expected one.

Smith, John Maynard, Evolution and the Theory of Games, Cambridge U. Pr, 1982; viii + 224 pp.

The theory of games, first devised to analyze economic behavior, has been applied in the last 10 years to animal conflict and evolution. In the latter context the concept of human rationality is replaced by that of evolutionary stability. Smith includes both theory and applications to field data.

Climents, David L., An Introduction to Mathematical Models in Economic Dynamics, Polygonal, 1984; ii + 170 pp, $18.95.

This introduction to linear and non-linear deterministic models in economics, suitable for undergraduates familiar with differential equations, explains concepts from economics as they arise. References to other literature are scarce.

Bransford, John D., and Stein, Barry S., The IDEAL Problem Solver: A Guide for Improving Thinking, Learning, and Creativity, Freeman, 1984; xii + 150 pp, (P).

This book represents an attempt to convert the results of research into a methodical approach to the practical problems of daily life. The word "IDEAL" of the title is an acronym for Identify, Explore, Act, Look. No doubt this book will enhance general creativity, but other research suggests that there is unlikely to be any transfer to *mathematical* problem-solving skills unless a teacher makes connections explicit.

NEWS & LETTERS

MAA AWARDS

At the annual Business Meeting of the MAA, held January 12, 1985, in Anaheim, California, three individuals received special recognition.

Everett Pitcher, professor emeritus of Lehigh University, was honored for his long, devoted, and outstanding service to the American Mathematical Society and the larger mathematics community, and received the MAA Award for Distinguished Service to Mathematics.

Pitcher earned the MA (1933) and Ph.D. (1935) in mathematics at Harvard University, then went with his thesis advisor, Marston Morse, to the Institute for Advanced Study for a year. He returned to Harvard to teach for two years as a Benjamin Pierce Instructor, then accepted a position at Lehigh University. Pitcher spent active duty in the U.S. Army Ordinance Department 1942-45, with most of his time spent at the Ballistics Research Laboratory working with a distinguished group of mathematicians brought together by Oswald Veblen. After a year at the Institute for Advanced Study (1945-46), he returned to a permanent position at Lehigh, where, in addition to faculty assignments, he served as Department Chairman (1960-78) and consultant to the president (1978-present). Pitcher was Associate Secretary of the AMS (1959-66), and a founder of SIAM and member of the Board of Trustees (1961-63). He has served as Secretary of the AMS since 1967.

Charles Robert Hadlock, of Arthur D. Little, Inc., was awarded the MAA book prize for his 1978 Carus Monograph (no. 19), *Field Theory and its Classical Problems*. The following citation was provided by the committee on the MAA Book Prize (Paul Halmos, J. A. Seebach, Doris Schattschneider).

In the Preface the author writes about his work as follows. "I wanted to piece together carefully my own path through Galois theory, a subject whose mathematical centrality and beauty I had often glimpsed.... I approached this project as an inquirer rather than as an expert, and I hope to share some of the sense of discovery and excitement I experienced." The book provides a guide for a fascinating journey, avoiding both the sometimes narrow technical special cases of the 19th century and homological generalities of the 20th. It gives us the theorems and the examples, the Greek problems and the field extensions, the transcendence of π and the solvability of equations. In the opinion of the committee the author succeeded admirably in writing a scholarly and accurate but enjoyable and rewarding exposition.

Carl Pomerance, of the University of Georgia, was awarded the Chauvenet Prize for his paper "Recent Developments in Primality Testing", which appeared in *The Mathematical Intelligencer*, 3(1981) 97-105. The committee on the Chauvenet Prize (Gilbert Strang, Peter Hilton, and Larry Zalcman) describe the paper as follows.

It is a beautiful exposition of the basic number theory that has led to fast algorithms for deciding whether a number is prime. Professor Pomerance is himself a major contributor to the algorithms - probably he will know the longest prime. Or more precisely, he will know the largest known prime! The problem has found important applications in cryptography, and it remains at the same time one of the central and classical problems in mathematics - to which Pomerance and Adleman and Rumely and Lenstra and others are adding something important to the work of Fermat and Jacobi and Gauss.

EDITOR NAMED

The selection of Gerald Alexanderson, of the University of Santa Clara, as Editor of *Mathematics Magazine* for the term 1986-90 was approved by the MAA Board of Governors at their meeting in Eugene, Oregon, in August, 1984. Alexanderson, in addition to teaching, has been active in many facets of undergraduate mathematics. He is currently a Problems Editor for the *American Mathematical Monthly*, serves on the Putnam Committee, and is a member of CUPM. He has authored and edited several books, and is author of over 50 articles; two recent articles in this MAGAZINE can be found in 56(1983) 274-278, and 55(1982) 98-103.

STRENS COLLECTION ON RECREATIONAL MATHEMATICS

In 1983, Richard Guy purchased the Strens Collection of books and manuscripts on recreational mathematics and its history. Alan H. MacDonald, Director of Libraries of The University of Calgary, has installed it as a Special Collection, and cataloguing and classification are well under way.

Eugene Strens was an engineer, amateur mathematician and friend of the mathematical artist, M.C. Escher. He lived in Breda in The Netherlands and devoted much of his life to the lighter side of mathematics. Over the years he assembled a definitive collection of 2000 items, most of them books, but some of them periodicals, newspaper cuttings and manuscripts of his own compilation.

The next stage is for specialists to assist in the preparation of a database in the area, after which it would be easy to answer questions from the general public and to assist editors and authors with bibliographic details.

A conference in July 1986 is planned for a formal opening of the Collection. Meanwhile, donations and offers of help are most welcome, and may be addressed to "Strens Collection", Richard K. Guy, Department of Mathematics & Statistics, or to A.H. MacDonald, Director of Libraries, The University of Calgary, Calgary, Alberta, Canada, T2N 1N4.

CORRECTION TO QUICKIE 695

The statement of Quickie 695 (this MAGAZINE, November 1984, p. 299) is too restrictive since it implies that any two triangles in the chain are either congruent or skew-congruent. The statement should be:

Q 695. Triangles T and T' are said to be skew-congruent iff we can label their vertices A, B, C and A', B', C', respectively, in such a way that $\angle A = \angle A'$, $AB = A'B'$, and $BC = B'C'$. Suppose the triangles $T_0, T_1, \ldots, T_n$ are such that T_{k-1} is skew-congruent to T_k for $1 \leq k \leq n$. Is it possible for T_0 and T_n to be similar but not congruent? (The answer is unchanged.)

James Propp
U. of California, Berkeley

25TH INTERNATIONAL MATH OLYMPIAD SOLUTIONS

The solutions which follow have been especially prepared for publication in this MAGAZINE by Loren Larson and Bruce Hanson of St. Olaf College.

1. Prove that

$$0 \leq yz + zx + xy - 2xyz \leq \frac{7}{27},$$

where x, y, z are non-negative real numbers for which $x + y + z = 1$.

Sol. For the left inequality, $xy + xz + yz - 2xyz = xy(1 - z) + xz(1 - y) + yz \geq 0$. For the right inequality, we may assume that $0 \leq x \leq y \leq z \leq 1$. If $z \geq 1/2$ then

$$xy + xz + yz - 2xyz \leq xy + xz + yz - xy$$
$$= (x + y)z \leq \left(\frac{x + y + z}{2}\right)^2 = 1/4 < 7/27.$$

If $z < 1/2$, then

$$\left(\frac{1}{2} - x\right)\left(\frac{1}{2} - y\right)\left(\frac{1}{2} - z\right)$$

$$\leq \left(\frac{(1/2 - x) + (1/2 - y) + (1/2 - z)}{3}\right)^3$$

$$= \frac{1}{8 \times 27}$$

Expanding the left side, this is the same as

$$\frac{1}{8} - \frac{1}{4}(x + y + z) + \frac{1}{2}(xy + xz + yz) - xyz \leq \frac{1}{8 \times 27}$$

$$\frac{1}{2}(xy + xz + yz) - xyz \leq \frac{1}{8} + \frac{1}{8 \times 27} = \frac{7}{54},$$

$$xy + xz + yz - 2xyz \leq \frac{7}{27}.$$

2. Find one pair of positive integers a, b such that

(1) $ab(a+b)$ is not divisible by 7,

(2) $(a+b)^7 - a^7 - b^7$ is divisible by 7^7.

Justify your answer.

Sol. Since $(a+b)^7 - a^7 - b^7 = 7ab(a+b)(a^2+ab+b^2)^2$, it suffices to find a and b which satisfy (1) and such that $a^2 + ab + b^2$ is divisible by 7^3. If $a \equiv 2 \bmod 7$ and $b \equiv 1 \bmod 7$ then $a^2 + ab + b^2$ is divisible by 7. One way to choose a and b is to try $a = 2$ and $b = 1 + 7k$ for some integer k. Then $a^2 + ab + b^2 = 7(1 + 4k + 7k^2)$. When $k = 5$, $1 + 4k + 7k^2 = 7^2 \times 4$. Thus, $a = 2$, $b = 36$ is a solution.

3. In the plane two different points O,A are given. For each point X of the plane, other than O, denote by $a(X)$ the measure of the angle between OA and OX in radians, counterclockwise from OA ($0 \leq a(X) < 2\pi$). Let $C(X)$ be the circle with center O and radius of length $OX + \frac{a(X)}{OX}$. Each point of the plane is colored by one of a finite number of colors. Prove that there exists a point Y for which $a(Y) > 0$ such that its color appears on the circumference of the circle $C(Y)$.

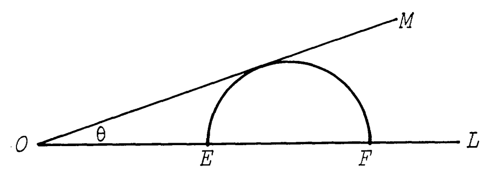
Sol. Let $S(C)$ denote the set of colors that appear on the circumference of circle C. Now consider the set of all circles with center O and radius r, $0 < r < \sqrt{2\pi}$. This is an infinite set, so there are two of them, say C_1 and C_2 with radii r_1 and r_2, $r_1 < r_2$, such that $S(C_1) = S(C_2)$. Let $y = r_1(r_2 - r_1)$. Note that $0 < y < (\sqrt{2\pi})^2 = 2\pi$. Let X be a point on C_1 such that $a(X) = y$. Then $C(X)$ has radius equal to

$$r_1 + \frac{y}{r_1} = r_1 + \frac{r_1(r_2 - r_1)}{r_1} = r_2.$$

By our choice of C_1 and C_2, the color of X must be on the circumference of $C_2 = C(X)$.

4. Let $ABCD$ be a convex quadrilateral such that the line CD is a tangent to the circle on AB as diameter. Prove that the line AB is a tangent to the circle on CD as diameter if and only if the lines BC and AD are parallel.

Sol. First, consider the following general situation: rays OL and OM intersect with acute angle θ, and E and F are two points on OL, with E between O and F (see figure). Let $C(E,F)$ denote the circle with diameter EF.



Then, by similar figures, $C(E,F)$ is tangent to OM if and only if $\frac{|EF|}{|OF|}$ is a fixed ratio, $K(\theta)$, depending only on θ.

Now consider our problem and look at the case in which AB and CD are not parallel. Take OL and OM to be rays passing through A,B and D,C respectively. Since $C(A,B)$ is tangent to OM, $\frac{|AB|}{|OB|} = K(\theta)$. From above, $C(D,C)$ is tangent to OL if and only if $\frac{|DC|}{|OC|} = K(\theta) = \frac{|AB|}{|OB|}$, and these ratios are equal if and only if BC is parallel to AD.

If AB and CD are parallel then $C(C,D)$ is tangent to AB if and only if $|AB| = |CD|$, and this holds if and only if BC and AD are parallel.

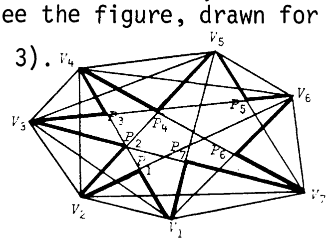
5. Let d be the sum of the lengths of all the diagonals of a plane convex polygon with n vertices ($n > 3$), and let p be its perimeter. Prove that

$$n - 3 < \frac{2d}{p} < \left[\frac{n}{2}\right]\left[\frac{n+1}{2}\right] - 2.$$

($[x]$ denotes the greatest integer not exceeding x.)

Sol. Label the vertices consecutively $V_1, V_2, \ldots, V_n$, and denote the diagonal between V_i and V_j by $d_{i,j}$. If n is even, let $m = \frac{n}{2} - 1$, and if n is odd, let $m = \frac{n-1}{2}$. For $k = 2,3,\ldots,m$, let $D_k = \{d_{i,i+k} : i = 1,2,\ldots,n\}$, where subscripts are taken modulo n.

Choose k between 2 and m, inclusive, and fix it, and let P_i denote the intersection of diagonals $d_{i,i+k}$ and $d_{i+1,i+1-k}$ (see the figure, drawn for $n = 7$ and $k = 3$).



The sum of the lengths of all the diagonals in D_k is $S_k \equiv \sum_{i=1}^{n} |d_{i,i+k}| \geq \sum_{i=1}^{n}\left(|V_iP_i| + |P_iV_{i+1}|\right) > \sum_{i=1}^{n} |V_iV_{i+1}| = p$. When n is even, $D_{n/2}$ contains only $n/2$ distinct diagonals, and a similar argument shows that $S_{n/2} \geq p/2$.

If n is odd, $d = \sum_{k=2}^{m} S_k > (m-2+1)p = \frac{n-3}{2}p$. If n is even, $d = \sum_{k=2}^{m} S_k + S_{n/2} > (m-2+1)p + p/2 = \frac{n-3}{2}p$. This proves the left inequality.

Again, fix k between 2 and m inclusive. Then

$$S_k = \sum_{i=1}^{n} |d_{i,i+k}| < \sum_{i=1}^{n}\left(|V_iV_{i+1}| + |V_{i+1}V_{i+2}| + \cdots + |V_{i+k-1}V_{i+k}|\right) = k\sum_{j=1}^{n} |V_jV_{j+1}| = kp.$$

If n is even, $2S_{n/2} = \sum_{i=1}^{n} |d_{i,i+n/2}| < \frac{n}{2}p$, so $S_{n/2} < \frac{n}{4}p$.

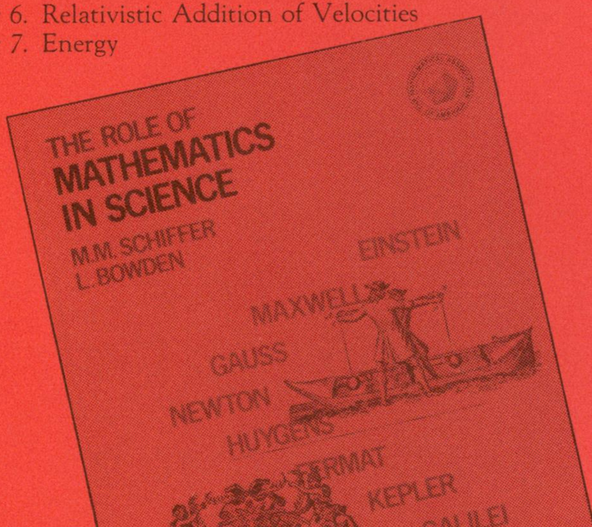
If n is odd,

$$d = \sum_{k=2}^{m} S_k < \sum_{k=2}^{m} kp = \left(\frac{[n/2][(n+1)/2]-2}{2}\right)p.$$

If n is even, $d = \sum_{k=2}^{m} S_k + S_{n/2} < \sum_{k=2}^{m} kp\ \frac{n}{4}p = \left(\frac{[n/2][(n+1)/2]-2}{2}\right)p$. This completes the proof.

6. Let a, b, c, d be odd integers such that $0 < a < b < c < d$ and $ad = bc$. Prove that if $a + d = 2^k$, $b + c = 2^m$ for some integers k and m, then $a = 1$.

Sol. It is easy to show that $k > m$. From $ad = bc$, $a+d = 2^k$, and $b+c = 2^m$, we find that $a(2^k-a) = b(2^m-b)$, or equivalently, $2^mb - 2^ka = (b-a)(b+a)$. It follows that 2^m divides $(b-a)(b+a)$. But $b-a$ and $b+a$ differ by $2a$, an odd multiple of 2, so either $b-a$ or $b+a$ is not divisible by 4. Hence, either 2^{m-1} divides $b+a$ or 2^{m-1} divides $b-a$. But $0 < b-a < b < 2^{m-1}$ so it must be that 2^{m-1} divides $b+a$. Since $0 < b+a < b+c = 2^m$, it follows that $b+a = 2^{m-1}$, or equivalently, $b = 2^{m-1} - a$. From this we find that $c = 2^{m-1} + a$, and therefore $ad = bc = (2^{m-1} - a)(2^{m-1} + a)$. This implies that $a(a+d) = 2^{2m-2}$, or, since $a+d = 2^k$, that $a = 2^{2m-2-k}$. But a is odd, and therefore $a = 1$.



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

MATHEMATICS MAGAZINE VOL. 58, NO. 1, JANUARY 1985