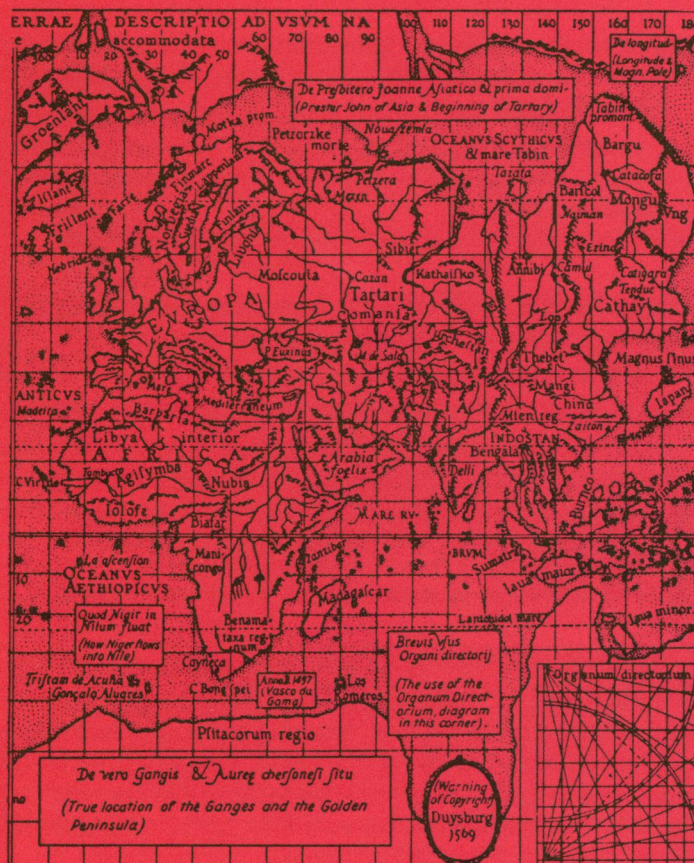


Δ G Δ Z i N E

CURVES OF CONSTANT WIDTH • MAPS
MURPHY'S LAW AND PROBABILITY • PI

Summer Reading

Tracking the Past and the Future

THE BOOK OF SQUARES

Leonardo Pisano (Fibonacci)

An annotated translation into modern English by L.E. Sigler

This collection of theorems on indeterminate analysis and equations of second degree is one of the most important mathematical treatises to have been written in the Middle Ages.

Now presented in English for the first time "this charming work ... should appeal to all readers interested in the history of number theory." – André Weil, The Institute for Advanced Study.

1986, 144 pp. \$19.95 Casebound/ISBN: 0-12-643130-2

DESCARTES' DREAM: The World According to Mathematics

Phillip J. Davis and Reuben Hersh

(Authors of *The Mathematical Experience* – a winner of the 1983 American Book Award in Science.)

"Complex, provocative, and highly original, this collection of essays calls attention to how mathematics, mostly through computers, has pervaded all aspects of modern life – technological, social, intellectual, artistic, even emotional. "

– Adult Editors' Choice 1986, Booklist

1986. Illustrated. 344 pp. \$19.95 Casebound/ISBN: 0-12-311601-5

Harcourt Brace Jovanovich, Publishers

THE SUPERCOMPUTER ERA

Sidney Karin and Norris Parker Smith

Supercomputers are the most powerful computers available and are rapidly coming into wider use all over the world. Written in a lively style, *The Supercomputer Era* is the first book to delve into the various developments in the field, providing interesting reading for both the non-specialist and the expert as well as being an essential reference.

Karin and Smith describe the machines, the software, and the multi-million-dollar systems that enable users to take full advantage of the enormous speed and power of supercomputers. They have assembled timely information needed by computer specialists: where supercomputers are being used, how different supercomputer centers are equipped, and what a researcher or industrialist needs to do to arrange use of a supercomputer.

July 1987, 208 pp. \$19.95 Casebound/ISBN: 0-12-311602-3

With an eight-page insert of colored photographs.

Harcourt Brace Jovanovich, Publishers.



ACADEMIC PRESS
Harcourt Brace Jovanovich, Publishers
Orlando, FL 32887-0510

Orlando San Diego New York Austin

44067

Credit card orders call toll free 1-800-321-5068.
Missouri, Hawaii, or Alaska 1-314-528-8110.

Prices are in U.S. Dollars and are subject to change.
Boston London Sydney Tokyo Toronto

EDITOR

Gerald L. Alexanderson
Santa Clara University

ASSOCIATE EDITORS

Donald J. Albers
Menlo College

Douglas M. Campbell
Brigham Young University

Paul J. Campbell
Beloit College

Lee Dembart
Los Angeles Times

Underwood Dudley
DePauw University

Judith V. Grabiner
Pitzer College

Elgin H. Johnston
Iowa State University

Loren C. Larson
St. Olaf College

Calvin T. Long
Washington State University

Constance Reid
San Francisco, California

William C. Schulz
Northern Arizona University

Martha J. Siegel
Towson State University

Harry Waldman
MAA, Washington, DC

EDITORIAL ASSISTANT

Mary Jackson

ARTICLES

- 131 Curves of Constant Width from a Linear Viewpoint,
by J. Chris Fisher.
- 141 Pi: Difficult or Easy?, *by G. A. Edgar.*

NOTES

- 151 A Curious Mixture of Maps, Dates, and Names,
by J. M. Sachs.
- 158 Proof without Words: The Harmonic Mean-Geometric Mean-Arithmetic Mean-Root Mean Square Inequality, *by Roger B. Nelsen.*
- 159 Murphy's Law and Probability or How to Compute Your Misfortune, *by Gene G. Garza.*
- 165 Proof without Words: $\pi^e < e^\pi$, *by Fouad Nakhli.*
- 166 The Fermat Last Theorem—A Brief, Elementary, Rigorous Proof, *by B. L. Schwartz.*
- 167 Reflections on the Ellipse, *by William C. Schulz and Charles G. Moore.*
- 169 Superexponentiation, *by Nick Bromer.*
- 174 Non-Associative Operations, *by N. J. Lord.*
- 177 Proof without Words: $\sum_{n=0}^{\infty} ar^n = a/(1-r)$, *by J. H. Webb.*

PROBLEMS

- 178 Proposals Numbers 1267–1271.
- 179 Quickies Numbers 722–723.
- 180 Solutions Numbers 1242–1246.
- 184 Answers to Quickies Numbers 722–723.

REVIEWS

- 185 Reviews of recent books and expository articles.

NEWS AND LETTERS

- 191 Announcements, Letters to the Editor

EDITORIAL POLICY

The aim of *Mathematics Magazine* is to provide lively and appealing mathematical exposition. This is not a research journal and, in general, the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for an article for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Articles on pedagogy alone, unaccompanied by interesting mathematics, are not suitable. Neither are articles consisting mainly of computer programs unless these are essential to the presentation of some good mathematics. Manuscripts on history are especially welcome, as are those showing relationships between various branches of mathematics and between mathematics and other disciplines.

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: G. L. Alexanderson, Editor, *Mathematics Magazine*, Santa Clara University, Santa Clara, CA 95053. 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

J. Chris Fisher ("Curves of Constant Width from a Linear Viewpoint") received his Ph.D. from the University of Toronto in 1972 and has been at the University of Regina—except for visiting positions at universities in Bologna, Munich, and Brussels—ever since. Although much of his fascination with geometry can be traced to H. S. M. Coxeter (his thesis supervisor), his interest in curves of constant width came from F. Bachmann's theory of n -gons. This theory touches geometry and algebra at many different levels and in many different ways. It led two Regina colleagues and the author through a seven-year adventure that they summarized in a 1985 *Monthly* article. It also led to a study of the geometric properties of Fourier series, which is the subject of the present article.

G. A. Edgar ("Pi: Difficult or Easy?") received his Ph.D. at Harvard University in 1973. His research has been in measure theory and nearby parts of functional analysis and probability. He has never been quite the same since he got his first computer in 1977. A computer-related article, "A compiler written in Prolog," was published in *Dr. Dobbs's Journal* (May 1985).

Acknowledgment. The quotes on pp. 140 and 150 were kindly provided by Ralph P. Boas.

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, bimonthly except July/August.

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. 20077-9564. 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.

Copyright © by the Mathematical Association of America (incorporated), 1987, including rights to this journal issue as a whole and, except where otherwise noted, rights to each individual contribution. Reprint permission should be requested from A. B. Willcox, Executive Director, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. General permission is granted to Institutional Members of the MAA for noncommercial reproduction in limited quantities of individual articles (in whole or part) provided a complete reference is made to the source.

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

Postmaster: Send address changes to *Mathematics Magazine* Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20077-9564.

PRINTED IN THE UNITED STATES OF AMERICA

Curves of Constant Width from a Linear Viewpoint

J. CHRIS FISHER

University of Regina

Regina, Canada S4S 0A2

Curves of constant width provide, among other things, a beautiful example of a vector space. Although the theory of Fourier series is always lurking in the background, only a passing acquaintance with that subject is actually required. The recurring theme here is how the innocent-looking phrase, “Let $z(\phi)$ equal its Fourier series expansion,” turns out to be a mathematical equivalent to “open sesame,” magic words that gain access to a great treasure cave. Section 2 provides a small detour designed to convince linear-algebra students that a vector in \mathbb{R}^3 need not be interpreted as an arrow; a new interpretation might lead to a different geometry. With such an interpretation we establish in Section 3 the characterization of constant-width curves that makes it so easy to study their properties in Section 4. Whereas the main tool throughout this study is elementary calculus, linear algebra serves to guide its development. Section 5 elaborates upon that theme, affording insight into the strengths and limitations of this approach. The final section surveys references and other points of view.

Those readers who prefer to avoid the use of complex numbers should note that there would be little lost in replacing my complex-valued functions by vector-valued functions; indeed, the references discussed in Section 5 all favor vector notation.

1. The vector space of ϕ -parametrizable curves

Let us call a smooth closed curve in the Euclidean plane **ϕ -parametrizable**, or more simply a **ϕ -curve**, if for each ϕ , $0 \leq \phi \leq 2\pi$, the curve has exactly one point where its tangent makes an

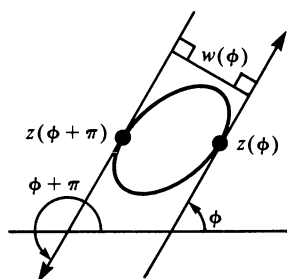


FIGURE 1A. The ϕ -parametrization.

angle of ϕ with the positive x -axis (FIGURE 1A). Such a curve is described by its ϕ -**parametrization** $z(\phi): \mathbb{R} \rightarrow \mathbb{C}$, where $z(\phi) = z(\phi + 2\pi)$ and the unit tangent vector is $e^{i\phi}$ for all $\phi \in \mathbb{R}$. In terms of the arc length s , $e^{i\phi} = dz/ds = z'(\phi)(d\phi/ds)$. For each ϕ one can define the **width** $w(\phi)$ to be the distance between the parallel tangents at $z(\phi)$ and $z(\phi + \pi)$. The curve has **constant width** if $w = w(\phi)$ is constant.

Our first goal is to obtain the Fourier series expansion of a ϕ -curve. Denote the radius of curvature of $z(\phi)$ by $R(\phi): [0, 2\pi] \rightarrow \mathbb{R}$. By definition $R(\phi) = ds/d\phi$ so that

$$R(\phi) = e^{-i\phi} z'(\phi). \quad (1.1)$$

Textbooks tell us that the k th Fourier coefficient of $R(\phi)$ will be

$$\rho_k = \frac{1}{2\pi} \int_0^{2\pi} R(\phi) e^{-ik\phi} d\phi,$$

so that under mild assumptions, soon to be explained,

$$R(\phi) = \sum_{k=-\infty}^{\infty} \rho_k e^{ik\phi}. \quad (1.2)$$

Since $R(\phi)$ is real-valued, $\rho_k = \bar{\rho}_{-k}$ for all k . Furthermore,

$$\rho_{-1} := \frac{1}{2\pi} \int_0^{2\pi} R(\phi) e^{i\phi} d\phi = \frac{1}{2\pi} \int_0^{2\pi} z'(\phi) d\phi = z(2\pi) - z(0) = 0,$$

and so $\rho_1 = \rho_{-1} = 0$. Consequently, as can be verified by differentiating, the Fourier series of $z(\phi)$ is

$$z(\phi) = \frac{\rho_0}{i} e^{i\phi} + \frac{1}{i} \sum_{k=2}^{\infty} \left(\frac{\rho_k}{1+k} e^{i(1+k)\phi} + \frac{\bar{\rho}_k}{1-k} e^{i(1-k)\phi} \right), \quad \rho_0 \in \mathbb{R}, \quad \rho_k \in \mathbb{C}. \quad (1.3)$$

Note that the integration constant (i.e., the coefficient of e^0) has been omitted; this amounts to centering the figure at the origin. Finally, abbreviate the above summand by $z_k(\phi)$ so that (1.3) becomes

$$z(\phi) = \rho_0 e^{i\phi}/i + \sum_{k \geq 2} z_k(\phi)/i. \quad (1.4)$$

Let's now turn things around and study curves that satisfy (1.3). Certainly there is no obvious reason why any interesting curve should be representable in this way. Indeed, historians seem to enjoy describing how Fourier was criticized by the mathematical community when he made the claim (around 1807) that all functions equal their Fourier expansions. It required almost a century to show that he was more or less correct. For instance, for equality to hold in (1.3) it is sufficient that a given function $z(\phi)$ be continuous and rectifiable [22, p. 164].

Curves that satisfy (1.3) form a vector space over \mathbb{R} . Examples of such curves are shown in FIGURE 1B: (i) boundaries of smooth convex sets, (ii) clockwise oriented curves with cusps (i.e., points where $R(\phi)$ changes sign), and (iii) hybrids of (i) and (ii) such as my family's coat of arms.

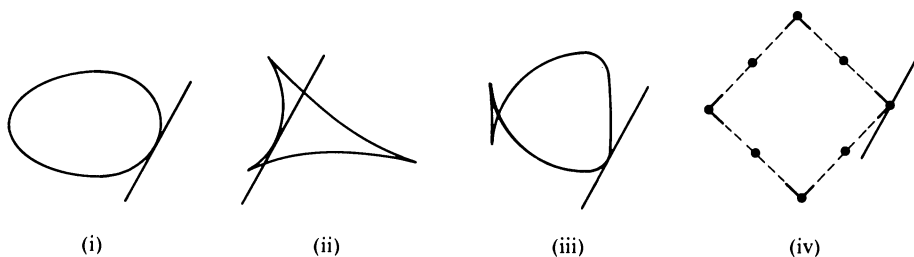


FIGURE 1B. Curves that satisfy (1.3).

Remark. One should follow a tangent vector around (iii) to convince himself that it is indeed a ϕ -curve. Begin with a horizontal tangent at the bottom pointing to the right, and let it roll counterclockwise around the fish's body over the top to the tail. Note that after it arrives at the cusp at the bottom of the tail, the point of tangency begins to move clockwise (with respect to the centroid of the fish) while the vector continues to rotate counterclockwise. This is the meaning of $R(\phi) < 0$. The point of tangency then resumes its counterclockwise motion ($R(\phi) > 0$) as it returns to the starting point from the top of the tail.

FIGURE 1B (iv) suggests how convex polygons fit into the theory. These are discontinuous examples of (1.3) that arise as limits of continuous ϕ -curves. The square, for example, is essentially the step function

$$z(\phi) = \begin{cases} i^k & \text{for } (2k+1)(\pi/4) < \phi < (2k+3)(\pi/4) \\ (1/2)(i^k + i^{k+1}) & \text{for } \phi = (2k+3)(\pi/4) \end{cases}$$

whose Fourier series is

$$z(\phi) = \frac{4}{\pi i \sqrt{2}} e^{i\phi} + \frac{1}{i} \sum_{j \geq 1} \left(\frac{4(-1)^j e^{(1+4j)\phi i}}{\sqrt{2}(1+4j)\pi} + \frac{4(-1)^j e^{(1-4j)\phi i}}{\sqrt{2}(1-4j)\pi} \right). \quad (1.5)$$

To extend the concept of a ϕ -parametrization to those curves such as (1.5) that have singular points, one must first generalize the notion of *tangent* to include lines that touch $z(\phi)$ at such points as corners and jump discontinuities and then interpret these lines in terms of (1.3). Instead of taking such a detour one can impose the condition that $R(\phi)$ be continuous and equal its Fourier expansion (1.2). In order to ensure such a condition conveniently (with only a small loss of generality) we shall restrict our attention to those curves that have a continuous second derivative. Summarizing the above discussion, *if $z(\phi)$ is a closed curve with center at 0 and a continuous second derivative, then it is ϕ -parametrizable if and only if it satisfies (1.3).*

The reader may now turn to the main theorem (Section 3) since it depends only on the above characterization of ϕ -curves. The present section concludes with two examples to illustrate how calculus can be used in combination with (1.3) to obtain properties of ϕ -curves.

(1.6) *If $z(\phi)$ bounds a convex region whose perimeter is L , then $L = 2\pi|\rho_0|$ where ρ_0/i is the coefficient of $e^{i\phi}$ in the Fourier expansion (1.3).*

Proof.

$$L = \int_0^L ds = \left| \int_0^{2\pi} \frac{ds}{d\phi} d\phi \right| = \left| \int_0^{2\pi} R(\phi) d\phi \right| = 2\pi|\rho_0|.$$

Note the role of convexity here. One can show that the convexity of a ϕ -parametrizable curve is equivalent to the condition that $R(\phi)$ does not change sign. This obviates the need for an absolute value sign inside the integral for L . Because negative values of ρ_0 are permitted and $L \geq 0$, one still requires the absolute value of the integral.

Convexity can be dropped if one is willing to introduce a notion of "signed arc length" analogous to the more familiar "signed area." The latter arises from Green's theorem, which in complex notation is just

$$A(z) := (\text{The area surrounded by } z(\phi)) = \frac{1}{2} \int_0^{2\pi} \text{Im}(-z(\phi) \bar{z}'(\phi)) d\phi.$$

(1.7) *When $z(\phi)$ is given by (1.3) the signed area of the region it surrounds is*

$$A(z) = \pi\rho_0^2 + \sum_{k \geq 2} \frac{2\pi|\rho_k|^2}{1-k^2}.$$

Proof. $e^{i\phi}$ and the functions $z_k(\phi)$ are mutually orthogonal in the sense that

$$\frac{1}{2} \int_0^{2\pi} -\operatorname{Im}(e^{ik\phi}(e^{-il\phi})') d\phi = \begin{cases} 0 & \text{if } k \neq l \\ k\pi & \text{if } k = l. \end{cases}$$

Thus $A(z) = \pi\rho_0^2 + \sum_{k \geq 2} A(z_k)$, where

$$\begin{aligned} A(z_k) &= A\left(\frac{\rho_k}{1+k} e^{i(1+k)\phi} + \frac{\bar{\rho}_k}{1-k} e^{i(1-k)\phi}\right) \\ &= A\left(\frac{\rho_k}{1+k} e^{i(1+k)\phi}\right) + A\left(\frac{\bar{\rho}_k}{1-k} e^{i(1-k)\phi}\right) \\ &= |\rho_k|^2 \pi \left(\frac{1}{1+k} + \frac{1}{1-k}\right) = \frac{2\pi|\rho_k|^2}{1-k^2}. \end{aligned}$$

Note that for $k \geq 2$ each z_k is a ϕ -curve that surrounds a nonpositive area; that is, $z_k(\phi)$ is always clockwise oriented. These curves are described more fully in the next section.

2. A familiar vector space of ϕ -curves

It might be helpful to postpone the main theorem until after having convinced oneself that, despite the formidable appearance of formula (1.3), there is nothing deep or difficult involved. Consider the special case in which $\rho_3 = 4a + 4bi$ and $\rho_k = 0$ for $k \neq 0, 3$. The resulting set of curves form a 3-dimensional vector space over \mathbb{R} . That is, for $\rho_0, a, b \in \mathbb{R}$,

$$z(\phi) = \rho_0(e^{i\phi}/i) + (a(e^{4i\phi} - 2e^{-2i\phi})/i) + b(e^{4i\phi} + 2e^{-2i\phi})$$

is isomorphic to an element of \mathbb{R}^3 , which we can denote by $\underline{z} := (\rho_0, a, b)$. In the Euclidean plane the parametric equations of the curve represented by $(\rho_0, a, 0)$ are

$$\begin{aligned} x(\phi) &= \rho_0 \sin \phi + 2a \sin 2\phi + a \sin 4\phi \\ y(\phi) &= -\rho_0 \cos \phi + 2a \cos 2\phi - a \cos 4\phi. \end{aligned}$$

Thus $(\rho_0, 0, 0)$ represents a circle of radius ρ_0 centered at the origin, while $(0, a, 0)$ is a deltoid [12, Chapter 8], a curve with three cusps much like (ii) in FIGURE 1B. Instructions for sketching $(\rho_0, a, 0)$ will be given in Section 4; depending on the ratio $\rho_0 : a$, these curves look something like (ii) of FIGURE 1B or like (i) or (ii) of FIGURE 4B. It is not hard to show that (ρ_0, a, b) is congruent to $(\rho_0, \sqrt{a^2 + b^2}, 0)$. (They are related by a rotation and a change of parameter). We shall soon see that (ρ_0, a, b) represents a curve whose width is the constant $2|\rho_0|$.

Remark. It is a straightforward exercise to show that, more generally, the curve $z_k(\theta)$ of (1.4) is a hypocycloid, the locus of a point attached to a circle of radius $k-1$ as it rolls (without slipping) on the inside of a fixed circle of radius $2k$. One sees at once that this hypocycloid is indeed a ϕ -curve from its alternative description as the envelope of a diameter of a circle of radius $2(k-1)$ rolling inside a fixed circle of radius $2k$ [12, p. 140]. When k is odd the curve has k cusps and wraps $(k-1)/2$ times about the center of the fixed circle. When k is even it has $2k$ cusps and a density of $k-1$.

The area formula discussed in (1.7) provides our vector space with a familiar quadratic form. Even though this form is not positive definite, it leads in the usual way to a bilinear form,

$$(\underline{z}, \underline{w}) := \frac{1}{2} \int_0^{2\pi} \operatorname{Im}(-z(\phi) \bar{w}'(\phi)) d\phi.$$

There is consequently an orthogonal basis composed of three vectors, one for which $(\underline{z}, \underline{z}) = 1$ and two for which the product is -1 :

$$\underline{e}_1 := e^{i\phi}/i\sqrt{\pi}, \quad \underline{e}_2 := (e^{4i\phi} - 2e^{-2i\phi})/2\sqrt{\pi}, \quad \underline{e}_3 := (e^{4i\phi} + 2e^{-2i\phi})/2\sqrt{\pi}.$$

If with respect to this basis $\underline{z} = (a_1, a_2, a_3)$ and $\underline{w} = (b_1, b_2, b_3)$, then

$$(\underline{z}, \underline{w}) = a_1 b_1 - a_2 b_2 - a_3 b_3.$$

This we recognize as the spacetime inner product of special relativity.

There are several types of linear transformations whose interpretation is of interest in the present context. The Lorentz group, described recently in the elementary paper [20], provides the area-preserving automorphisms of our 3-dimensional vector space of constant-width curves. More relevant to the present study, however, will be the projections $(a_1, a_2, a_3) \rightarrow (a_1, 0, 0)$, which maps $z(\phi)$ to the circle $(z(\phi) - z(\phi + \pi))/2$, and $(a_1, a_2, a_3) \rightarrow (0, a_2, a_3)$, mapping $z(\phi)$ onto $m(\phi) = (z(\phi) + z(\phi + \pi))/2$. According to (1.6) the component $(\underline{z}, \underline{e}_1) = a_1$ is proportional to the arc length of $z(\phi)$.

Although the information and notation contained in this section will not be required explicitly, they are certainly implicit in everything that follows.

3. The characterization of curves of constant width

Now for the proof of the main theorem (Hurwitz [9, §4]).

(3.1) **THEOREM.** *A closed curve $z(\phi)$ (with center at 0 and a continuous second derivative) has constant width $|w|$ if and only if there exist constants $w \in \mathbb{R}$ and $\rho_{2j+1} \in \mathbb{C}$ for which*

$$z(\phi) = \frac{w}{2i} e^{i\phi} + \frac{1}{i} \sum_{j \geq 1} \left(\frac{\rho_{2j+1}}{2+2j} e^{i(2+2j)\phi} - \frac{\bar{\rho}_{2j+1}}{2j} e^{-2ij\phi} \right). \quad (3.2)$$

Proof. Since by definition any curve of constant width is ϕ -parametrizable, let us suppose that $z(\phi)$ is an arbitrary ϕ -curve given by equation (1.4). Set $f(\phi)$ equal to the sum (3.2) above, namely, $f(\phi) = we^{i\phi}/2i + \sum_{j \geq 1} z_{2j+1}(\phi)/i$, and let $g(\phi) = z(\phi) - f(\phi)$. We are to show that the width of $z(\phi)$ is constant if and only if $g(\phi) = 0$ for all ϕ .

First note that because of the even powers of $e^{i\phi}$, $z_{2j+1}(\phi) = z_{2j+1}(\phi + \pi)$. Consequently $f(\phi) - f(\phi + \pi) = we^{i\phi}/i$ is a vector (in the Euclidean plane) of length $|w|$ that is perpendicular to $e^{i\phi}$. The tangent vectors $f'(\phi) = R(\phi)e^{i\phi}$ and $f'(\phi + \pi) = -R(\phi + \pi)e^{i\phi}$ (according to (1.1)) are parallel to $e^{i\phi}$. Thus the distance between any two parallel tangents of $f(\phi)$ is the constant $|w|$; that is, any curve satisfying (3.2) has constant width $|w|$.

For the converse assume $z(\phi)$ has constant width $w \geq 0$. Because $g(\phi)$ contains only odd powers of $e^{i\phi}$, $g(\phi) = -g(\phi + \pi)$. Since $z(\phi)$ is a ϕ -curve, its tangents at $z(\phi) = f(\phi) + g(\phi)$ and at $z(\phi + \pi) = f(\phi + \pi) + g(\phi + \pi) = f(\phi + \pi) - g(\phi)$ are parallel to $e^{i\phi}$ and so are perpendicular to the vector $f(\phi) - f(\phi + \pi)$. (See FIGURE 3A.) But we know that $|f(\phi) - f(\phi + \pi)| = w$. The only way for the vectors $g(\phi)$ and $g(\phi + \pi)$ to (a) point in opposite directions and (b) not change the distance w between the tangents is for $g(\phi)$ to be parallel to $e^{i\phi}$. But $g'(\phi)$ is also parallel to $e^{i\phi}$ (since $g(\phi)$ is a ϕ -curve). This occurs for all ϕ exactly when $g(\phi)$ is the zero function. (That is, $g(\phi) = s(\phi)e^{i\phi}$ for $s(\phi)$ real-valued implies $g'(\phi) = e^{i\phi}(s'(\phi) + is(\phi))$, which is a real multiple of $e^{i\phi}$ only if $s(\phi) = 0$). \square

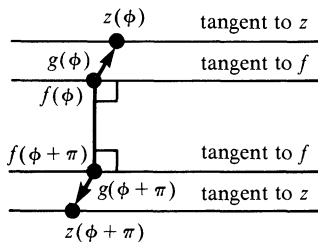


FIGURE 3A. Proving the converse of (3.1).

4. Consequences of the characterization theorem

Let $z(\phi)$ be a curve of constant width $w \geq 0$ whose center is 0 and second derivative is continuous. Starting from theorem 3.1 one easily obtains many of the familiar results concerning such curves. For example,

(4.1) *A line joining the opposite points $z(\phi)$ and $z(\phi + \pi)$ meets the curve at right angles* (since, as in the proof of (3.1), $z(\phi) - z(\phi + \pi) = we^{i\phi}/i$ while the tangent vectors are $R(\phi)e^i$ and $-R(\phi + \pi)e^{i\phi}$).

(4.2) *The length of any line segment joining opposite points $z(\phi)$ and $z(\phi + \pi)$ equals the width w* (since $|z(\phi) - z(\phi + \pi)| = |(w/i)e^{i\phi}| = w$ as in the proof of (3.1)).

The segment of (4.1) and (4.2) is called a **diameter**.

(4.3) (BARBIER'S THEOREM) *If $z(\phi)$ is convex then its perimeter is πw .* (This is 1.6.)

(4.4) *A circle is the only constant-width curve that is centrally symmetric.* (Centrally symmetric means that $z(\phi) = -z(\phi + \pi)$, hence $z_k(\phi) \equiv 0$ in (3.2) for all k .)

Associated with any constant-width curve is the curve $m(\phi)$ defined by

$$m(\phi) := \sum_{j \geq 1} z_{2j+1}(\phi)/i.$$

Because $m(\phi) = m(\phi + \pi)$ it is described twice as ϕ runs from 0 to 2π . Thus $m(\phi)$ is really a curve of zero width—its two parallel tangents coincide. Further, $m(\phi)$ has an odd number of cusps on $[0, \pi)$ (since the cusps are, by definition, precisely those points where the curvature changes sign, while the radius of curvature at $m(\phi) = m(\phi + \pi)$ is, by (1.1), $R(\phi) = m'(\phi)e^{-i\phi} = -m'(\phi + \pi)e^{-i(\phi + \pi)} = -R(\phi + \pi)$; that means that the sign of $R(\phi)$ must change an odd number of times as ϕ runs from 0 to π).

(4.5) *The radii of curvature at opposite points $z(\phi)$ and $z(\phi + \pi)$ have a constant sum w* (since $z(\phi) = (we^{i\phi}/2i) + m(\phi)$ these radii of curvature are $R(\phi) = z'(\phi)e^{-i\phi} = (w/2) + m'(\phi)e^{-i\phi}$ and $R(\phi + \pi) = (w/2) + m'(\phi + \pi)e^{-i(\phi + \pi)} = (w/2) - m'(\phi)e^{-i\phi}$).

(4.6) *$z(\phi)$ is convex if and only if $w/2 \geq m'(\phi)e^{i\phi}$ for all ϕ* (since for ϕ -curves convexity is equivalent to $R(\phi)$, as given in the proof (4.5), not changing sign).

A *vertex* of a convex curve (whose curvature is continuously differentiable) is a point where the curvature takes an extreme value. It follows from (4.5) that points of minimum curvature (where $m'(\phi)e^{-i\phi}$ is a maximum) occur opposite points of maximum curvature (where $m'(\phi)e^{-i\phi}$ is a minimum). Hence,

(4.7) *If $z(\phi)$ is convex and has a finite number of vertices, that number must be of the form $4j + 2$* (in fact, twice the number of times that $\frac{d}{d\phi}(m'(\phi)e^{-i\phi})$ changes sign on $[0, \pi)$, a number which we've seen to be always odd).

A connection between the Fourier series of $m(\phi)$ and the number of times $m'(\phi)e^{-i\phi}$ changes sign is given by a theorem of A. Hurwitz (discussed in [9, p. 537] and proved in [10, §9, Satz IX]): *For*

$$m(\phi) = \frac{1}{i} \sum_{j \geq N} z_{2j+1}(\phi)$$

with $z_{2N+1}(\phi)$ not identically zero, $m'(\phi)e^{-i\phi}$ changes sign at least $2N + 1$ times on $[0, \pi)$. Thus $m(\phi)$ has at least $2N + 1$ cusps. It follows that $z(\phi) = (we^{i\phi}/2i) + m(\phi)$ has at least $4N + 2 \geq 6$ vertices. Compare this with the theorem that a convex curve is in general guaranteed only 4 vertices [21; p. 48].

Roughly speaking, one can imagine $m(\phi)$ to be constructed from a semicircle that has been broken into an odd number of arcs $A_0A_1, A_1A_2, \dots, A_{2n}A_{2n+1}$ and stuck back together with the

arcs facing alternately in and out as in FIGURE 4A (with $A_{j-1}A_j$ and A_jA_{j+1} forming a cusp at A_j , while $A_{2n+1} = A_0$). Of course, these circular arcs can be replaced by quite general curves on which $R(\phi)$ does not change sign, on the condition that the tangent turns continuously through an angle of π as ϕ runs from 0 to π . (That is, $m(\phi)$ must be ϕ -parametrizable; see [4, p. 132] or [21, p. 50].)

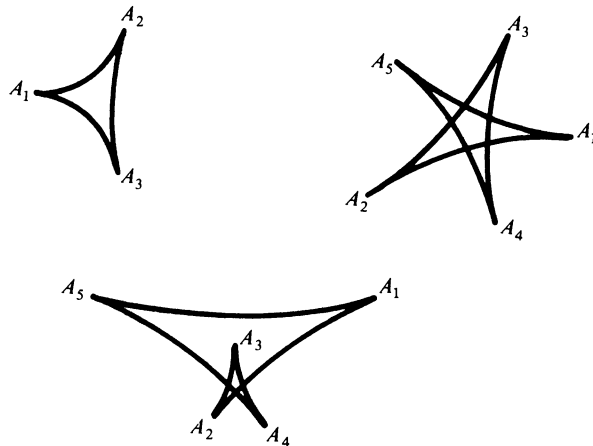


FIGURE 4A. Examples of the curve $m(\phi)$.

Because $m(\phi) = (1/2)(z(\phi) + z(\phi + \pi))$, it is the locus of the midpoints of the diameters of $z(\phi)$. This suggests an easy way to construct curves of constant width, an observation that seems to have first been made by A. Hurwitz [9, §4]. Starting with a constant $w \geq 0$ and the curve $m(\phi)$ described above—any ϕ -curve which has an odd number of cusps $A_1, A_2, \dots, A_{2n+1}$ and satisfies $m(\phi) = m(\phi + \pi)$ —a curve of constant width is obtained by adding for each ϕ the vector $(w/2i)e^{i\phi}$ to the vector $m(\phi)$. This construction is illustrated in FIGURE 4B where the n line

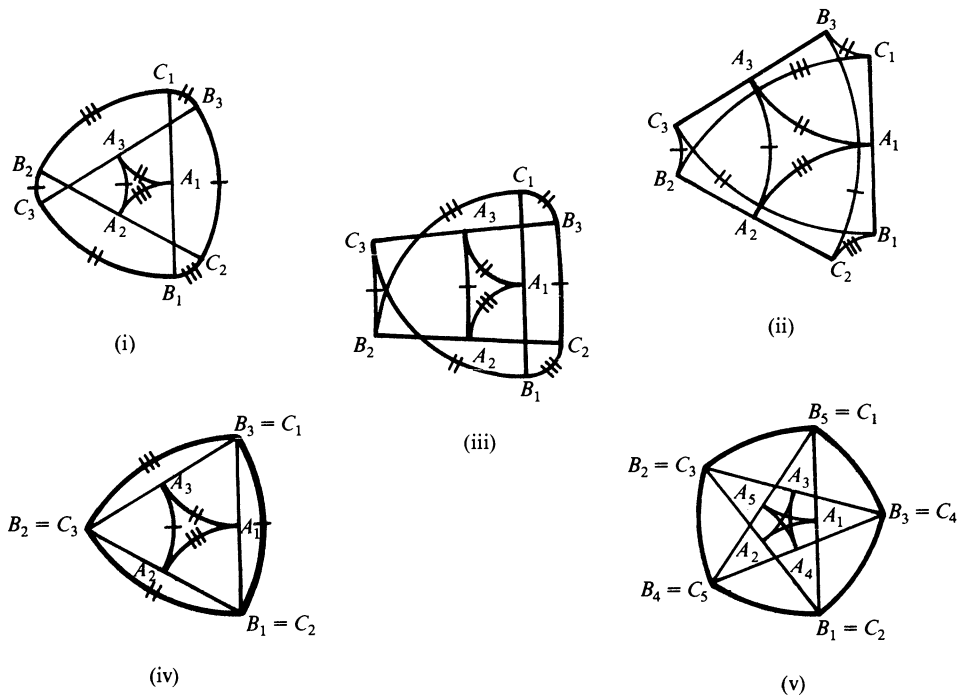


FIGURE 4B. The construction of curves of constant width.

segments $B_j C_j$ each have length w and are perpendicular to $m(\phi)$ at their midpoints A_j ; the curves joining B_j to C_{j+1} and C_j to B_{j+1} are parallel to that arc of $m(\phi)$ from A_j to A_{j+1} .

It follows from (4.6) that when w is chosen sufficiently large, namely, $w/2 \geq \sup \{m'(\phi)e^{i\phi}\}$, the curve constructed by the Hurwitz method will be convex. Should $A_1 A_2$, say, be a line segment (with an infinite radius of curvature), the resulting constant-width curve will contain line segments $B_1 C_2$ and $C_1 B_2$ but could never be convex.

The curves (iv) and (v) of FIGURE 4B are examples of the convex constant-width curves known as **Reuleaux polygons**. These curvilinear polygons are composed of $2n + 1$ circular arcs, the circle in each case having radius w and center at the vertex opposite the arc. For Reuleaux polygons, $B_j = C_{j+1}$ and $B_1 B_2 \cdots B_{2n+1}$ is an equilateral $(2n + 1)$ -gon that encloses its centroid n times.

Euler constructed his curves of constant width as involutes of curves such as $m(\phi)$, an approach often found in differential geometry texts such as [21, p. 50]. To see that this construction is different from (but equivalent to) the Hurwitz construction, let $z(\phi)$ be a curve of constant width. The **center of curvature** for a particular value of ϕ is the point on the normal that is $R(\phi)$ units inside the curve, namely, $z(\phi) + iz'(\phi)$. The locus of this point is called the **evolute**; by (3.2) the evolute of $z(\phi)$ is

$$z(\phi) + iz'(\phi) = i \sum_{j \geq 1} \left(\frac{(1 + 2j)\rho_{2j+1}}{2 + 2j} e^{i(2+2j)\phi} + \frac{(1 + 2j)\bar{\rho}_{2j+1}}{2j} e^{-2ij\phi} \right). \quad (4.8)$$

Because it is ϕ -parametrizable (each tangent is perpendicular to $e^{i\phi}$) and only even powers of $e^{i\phi}$ appear, this evolute has all the properties of the curve $m(\phi)$ discussed above. It coincides with $m(\phi)$ exactly when $\rho_k = 0$ for $k \neq 0$ (and $z(\phi)$ is a circle). In other words, for a fixed family of curves $\{z_{2j+1}(\phi) : j \geq 1\}$ and varying widths $w \geq 0$ one obtains a family of constant-width curves

$$\frac{w}{2i} e^{i\phi} + \frac{1}{i} \sum z_{2j+1}(\phi);$$

these are the **involute**s of (4.8). The unique involute of this family with $w = 0$ is $m(\phi)$.

5. Further properties of constant-width curves

Certainly not all properties of constant-width curves are so easily verified. In general, only properties that can be described naturally in terms of the vector space (1.3) qualify for treatment. Consider the theorem (of Blaschke):

If the smallest circumcircle of a curve of constant width has unit radius, then the radius of the largest incircle ranges from $\sqrt{3} - 1$ (for the Reuleaux triangle) to 1 (for the circle).

Although the proof in [4, p. 164] or [3] is not inordinately difficult, the vector space methods of this paper seem to be inappropriate. Among other difficulties, the set of *convex* constant-width curves is not a subspace but a cone. (In general only linear combinations involving *positive* multiples of convex curves yield convex curves.) More intractable is the notion of a “circle contained in the interior of a region bounded by a given curve,” something which seems to lack a workable vector-space description.

On the other hand it is sometimes possible to use linear algebra to simplify proofs. Take, for example, the following

THEOREM (Blaschke-Lebesgue). *Among all convex curves of constant unit width, the circle has the largest area and the Reuleaux triangle has the smallest.*

Remark. The elegant proof of this theorem by Lebesgue (reproduced in [4, §66] and [3]) is short, clever, and elementary. Other simple proofs are to be found in many references. Still, an approach via linear algebra is not entirely without interest.

Proof outline.

$$z(\phi) = e^{i\phi}/2i + \sum_{j \geq 1} z_{2j+1}(\phi)/i$$

has area

$$A = \frac{\pi}{4} + \sum_{j \geq 1} \frac{2\pi|\rho_{2j+1}|^2}{-4j(1+j)}$$

according to (1.7). Therefore the area is a maximum exactly when $\rho_{2j+1} = 0$ for $j \geq 1$. (Indeed, this argument provides for all ϕ -curves a proof of the isoperimetric inequality: The circle has the largest area of any curve, convex or not, having perimeter π .) To find the curve of minimum area requires more work. We must find the sequence $\{\rho_{2j+1}\}_{j \geq 1}$ that minimizes A subject to the condition that $z(\phi)$ be convex. Thus, recalling (4.6), we wish to maximize

$$\sum_{j \geq 1} 2\pi|\rho_{2j+1}|^2/4j(1+j),$$

the area of $m(-\phi)$, subject to the condition that

$$1/2 \geq m'(\phi)e^{-i\phi} = \sum (\rho_{2j+1}e^{(2j+1)i\phi} + \bar{\rho}_{2j+1}e^{-(2j+1)i\phi}).$$

The solution is not impossible by analytic means [11], but it is evidently quite involved. On the other hand, it is easy to show that the radius of curvature of $m(-\phi)$ should achieve its maximum absolute value of $1/2$ at all but finitely many points (since any arc where $|m'(\phi)e^{-i\phi}| < 1/2$ could be replaced by one that increases the area yet still satisfies the constraint). The problem is thus reduced to

Maximize the area (counting multiply enclosed regions more than once as in Green's theorem) surrounded by the arc of a semicircle that has been broken into an odd number of smaller arcs and reassembled to face alternately in and out, forming cusps at the points where two of these arcs are joined (as in FIGURE 4A).

This can be solved along the lines of the elementary arguments used by Pólya in his proof of the isoperimetric inequality [16]. The answer is to break the semicircle into three equal pieces; the resulting curve of constant width is the Reuleaux triangle ((iv) in FIGURE 4B).

6. Remarks on the literature

Curves of constant width have been used in countless books to illustrate principles of convexity theory ([4], [6], [13]), of differential geometry ([2], [21]), and of recreational mathematics ([5], [17]). (Note that these references are chosen for their accessibility; there are many others.) Perhaps the most thorough treatment of the topic is by Bonnesen and Fenchel [4, Chapter 15] where the history is traced from the beginning with Euler in 1778 up to the early 1930's; [1] continues the story to 1982 and includes a bibliography of some 250 items that have appeared since 1930.

New results continue to appear. For example, a series of papers by Hammer (see [8]) contains several interesting results and references. Also, the generalization to rotors has received considerable attention. A *rotor* is a curve which is inscribed in a fixed polygon and is free to turn through an angle of 2π while maintaining contact with all sides of the polygon (like a constant-width curve inscribed in a rhombus). Goldberg [7] provides numerous references to this topic.

The approach via Fourier series dates back to Hurwitz in 1902 [9], a paper that deals with a variety of interesting topics in addition to constant-width curves. A couple of years later Meissner [14], [15] extended the method to rotors and related problems. The technique has been revived occasionally; Schneider [19] exploited it to characterize completely n -dimensional rotors. He also provides a historical sketch and bibliography.

The role of Franz Reuleaux (1829–1905) in the development of the theory of constant-width curves is somewhat curious: he seems to have contributed nothing but his name. He was, according to the *Dictionary of Scientific Biography*, an influential engineer who, among other things, founded the science of kinematics; his system of analyzing and classifying machinery “was philosophical in scope and has proved remarkably durable.” No mention is made of any mathematical achievement (although his meter-preserving translation into German of Longfellow's *Hiawatha* rates a comment). In his important 1875 book [18] he analyzed the kinematical

properties of the *Bogendreieck* (now called the Reuleaux triangle) and mentioned the other Reuleaux polygons. Perhaps the greatest irony of all: of the 50 or so examples of rotary engines that he described, none involve the curvilinear triangle bearing his name, the shape of the rotor in the first practical rotary engine created by Felix Wankel around 1954.

References

- [1] G. D. Chakerian and H. Groemer, Convex bodies of constant width, *Convexity and its Applications*, ed. by Peter M. Gruber and Jörg M. Wills, Birkhäuser, Boston, 1983, 49–96.
- [2] Ludwig Bieberbach, *Differentialgeometrie*, Teubners Math. Leitfäden Band 31, Leipzig, 1932.
- [3] Christian Blatter, Über Kurven konstanter Breite, *Elem. Math.*, 36 (1981) 105–115.
- [4] T. Bonnesen and W. Fenchel, *Theorie der Konvexen Körper*, Chelsea, N.Y., 1971.
- [5] J. H. Cadwell, *Topics in Recreational Mathematics*, Cambridge University Press, 1970.
- [6] H. G. Eggleston, *Convexity*, Cambridge Tract, No. 47, Cambridge University Press, 1963.
- [7] M. Goldberg, Rotors in polygons and polyhedra, *Math. Comp.*, 14 (1960) 229–239.
- [8] Preston C. Hammer, Convex curves of constant Minkowski breadth, *Proc. Sympos. Pure Math.* VII, pp. 291–304, Amer. Math. Soc., Providence, R.I., 1963.
- [9] Adolf Hurwitz, Sur quelques applications géométriques des séries de Fourier, in *Mathematische Werke*, vol. 1, Basel, 1932, pp. 509–554. (Originally in *Ann. Ecole Norm. Sup.* (3), 19 (1902) 357–408.)
- [10] ———, Über die Fourierschen Konstanten integrierbarer Funktionen, in *Mathematische Werke*, vol. 1, Basel, 1932, pp. 555–576. (Originally in *Math. Ann.* 51 (1903) 425–446.)
- [11] R. Klötzler, Beweis einer Vermutung über n -Orbiformen kleinsten Inhalts, *Zeit. Angew. Math. Mech.*, 55 (1975) 557–570.
- [12] E. H. Lockwood, *A Book of Curves*, Cambridge U. Press, 1961.
- [13] L. A. Lyusternik, *Convex Figures and Polyhedra*, GITTL, Moscow, 1956; English translations, Dover, N.Y., 1963, and Heath, Boston, 1966.
- [14] E. Meissner, Über die Anwendung von Fourierreihen auf einige Aufgaben der Geometrie und Kinematik, *Vierteljahresschr. Naturf. Ges. Zürich*, 54 (1909) 309–329.
- [15] ———, Über die durch reguläre Polyeder nicht stützbaren Körper, *Vierteljahresschr. Naturf. Ges. Zürich*, 63 (1918) 544–551.
- [16] G. Pólya, *Induction and Analogy in Mathematics, Mathematics and Plausible Reasoning*, vol. 1, Princeton Univ. Press, 1954.
- [17] H. Rademacher and O. Toeplitz, *The Enjoyment of Mathematics*, Princeton University Press, 1957.
- [18] Franz Reuleaux, *The Kinematics of Machinery: Outline of a Theory of Machines*, Macmillan, London, 1876. (Translation into English of *Theoretische Kinematik*, Friedrich Vieweg und Sohn, Braunschweig, 1875).
- [19] Rolf Schneider, Gleitkörper in konvexen Polytopen, *J. Reine Angew. Math.*, 284 (1971) 193–220.
- [20] Frederick Solomon, A theorem concerning Lorentz transformations, *Amer. Math. Monthly*, 91 (1984) 638–641.
- [21] Dirk J. Struik, *Lectures on Classical Differential Geometry*, Addison-Wesley, Cambridge, Mass., 1950.
- [22] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 3rd ed., Cambridge University Press, 1920.

It is tempting to speculate as to how much progress the Greeks would have made in mechanics had they been acquainted with algebra, or how much more rapidly physics would have advanced in the seventeenth century had the infinitesimal calculus been discovered at its beginning instead of at its end. The moral for statesmen would seem to be that, for proper scientific ‘planning’, pure mathematicians should be endowed fifty years ahead of scientists.

R. B. Braithwaite, *Scientific Explanation*, Cambridge, 1953, p. 49.

Pi: Difficult or Easy?

Mathematical considerations for the multidigit computation of pi.

G. A. EDGAR*

The Ohio State University

Columbus, Ohio 43210

I want to consider how difficult it is to compute the decimal expansions of π and e to many places, using a computer, of course. The considerations to be discussed might interest mathematics students who want to do something unusual on a computer, or computer science students who need some motivation in their mathematics studies.

When I say “many” places, I mean it. This could be hundreds of places on a microcomputer, or hundreds of thousands on a mainframe. I will not emphasize the result itself, only the considerations that come up along the way. There are, in fact, few “practical” uses for a computation of π beyond 30 or 40 places. Kasner and Newman [6, p. 78] quote Simon Newcomb: “Ten decimal places are sufficient to give the circumference of the earth to the fraction of an inch, and thirty decimals would give the circumference of the whole visible universe to a quantity imperceptible with the most powerful telescope.”

The hardware requirement for what we will do is minimal. Practically any computer will suffice. (Even this might be reduced—a freshman class of mine in 1980 computed π to 50 places using TI-58 programmable calculators.) Most of my own computations were done on a 64K Z-80 computer that was built in 1977.

The software requirement is a bit more difficult. We will need to be able to do integer arithmetic involving integers of, say, 500 digits. This is not a capability included in your standard BASIC interpreter. (An interesting programming project might be to write the routines for such arithmetic. For a microcomputer, this would probably have to be done in assembly language in order to have acceptable speed.) I imagine most large computer installations will have one or more software packages available with the capability to do large integer computations. My own computations were done under CP/M using first MuMath (see [13]) and then Waltz Lisp. Under MS-DOS it should be possible to use certain versions of Lisp, Logo, and Prolog.

Easy as e

I will first discuss the calculations of e to many places, because it is easier than π . It also serves to illustrate the considerations that come up in the calculation of π .

Here is the formula that will be used (see, for example, [12, Theorem 2.58]):

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}. \quad (1)$$

Let d be a positive integer. I am interested in computing e correct to d decimals. I understand that to mean that the error is at most $0.5 \cdot 10^{-d}$. I have in mind values of d such as 50 or 100 or 500.

First, let's consider the error involved in approximating e by a partial sum of the series (1).

*Supported in part by N.S.F. grant DMS 84-01986

Let

$$S_n = \sum_{k=0}^n \frac{1}{k!}.$$

I need to estimate the error $|e - S_n|$. This can be done, for example, by comparison with a geometric series:

$$\begin{aligned} e - S_n &= \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \cdots \\ &= \frac{1}{(n+1)!} \left[1 + \frac{1}{n+2} + \frac{1}{(n+2)(n+3)} + \cdots \right] \\ &< \frac{1}{(n+1)!} \left[1 + \frac{1}{n+2} + \frac{1}{(n+2)^2} + \cdots \right] \\ &= \frac{1}{(n+1)!} \frac{n+2}{n+1} < \frac{1}{n!}. \end{aligned}$$

(Some of the estimates here, and below, are somewhat crude. It might be instructive to consider what effect the use of more careful estimates would have.)

Now suppose I want to arrange for this error to be smaller than some preassigned number ϵ . I must solve an inequality

$$\frac{1}{n!} < \epsilon$$

for n . This can be done by trial and error, but that may be slow. Another approach is discussed here. The estimates that will be useful are of this sort: if $n = u/\log u$, then $n!$ is approximately e^u . More precisely:

LEMMA. *If $c > 1$ is a constant, and $n = cu/\log u$ is an integer, then $n! > e^u$ for large enough n . (If we take $c = 2$, for example, then we will have $n! > e^u$ for $n \geq 3$.)*

To prove this, we start with Stirling's formula (see [12, Theorem 5.44]).

$$n! > \sqrt{2\pi n} \left(\frac{n}{e} \right)^n,$$

then substitute $n = cu/\log u$ and take logarithms. The result is

$$\begin{aligned} \log(n!) &> cu - \frac{cu \log \log u}{\log u} + \frac{cu}{\log u} \log(c/e) + \\ &\quad \left(\frac{1}{2} \right) \log u - \left(\frac{1}{2} \right) \log \log u + \left(\frac{1}{2} \right) \log(2\pi c). \end{aligned}$$

All the terms after cu grow more slowly than u (as $u \rightarrow \infty$); but $c > 1$, so $\log(n!) - u \rightarrow \infty$ as $u \rightarrow \infty$. Thus, for large n (and large u), we have $\log(n!) - u > 0$, or $n! > e^u$.

We will need to have $|e - S_n| > 0.5 \cdot 10^{-d}$, which will hold if $n! > e^u$ and $e^{-u} < 0.5 \cdot 10^{-d}$; that is, if $u \geq d \log 10 + \log 2$, and n is the least integer greater than $2u/\log u$. So n is approximately

$$\frac{2d \log 10 + 2 \log 2}{\log(d \log 10 + \log 2)}. \quad (2)$$

We can say, then,

$$n = O(d/\log d) \quad (3)$$

in order to single out just the order of magnitude of n , instead of the complicated formula (2). This "big O " notation means that there is a constant A with

$$n \leq A \frac{d}{\log d}$$

for all d sufficiently large.

I am now ready to formulate the algorithm to be used.

Algorithm E: multidigit computation of e . Let d be a positive integer, let n be given by (2), ($n \geq 3$). There will be four integers stored in memory: K must be able to contain integers from 0 to n ; A , B and E must be able to contain integers from 0 to $4n!$.

Step 1: $K \leftarrow 0$, $A \leftarrow 1$, $B \leftarrow 1$, $E \leftarrow 2 \cdot 10^d$.

Step 2: $K \leftarrow K + 1$, $A \leftarrow A \cdot K + 1$, $B \leftarrow B \cdot K$.

Step 3: if $B < E$, then go to Step 2.

Step 4: the result is A/B , stop.

The arrow " \leftarrow " denotes assignment: the value on the right is placed in the memory location named on the left. I claim that the result is the number required:

$$\left| \frac{A}{B} - e \right| < 0.5 \cdot 10^{-d}. \quad (4)$$

To show that this is true, what would be involved? One should prove the following facts about Algorithm E, using induction:

(a) At all times: K, A, B, E are non-negative integers.

(b) Whenever Step 2 begins: $0 \leq K < n$, $B = K!$,

$$\frac{A}{B} = \sum_{k=0}^K \frac{1}{k!}.$$

(c) Whenever Step 3 begins: $0 < K \leq n$, $B = K!$,

$$\frac{A}{B} = \sum_{k=0}^K \frac{1}{k!}.$$

(d) Whenever Step 4 begins: $0 < K \leq n$, $B = K! > 2 \cdot 10^d$,

$$\frac{A}{B} = \sum_{k=0}^K \frac{1}{k!}.$$

The proof is tedious but easy. Only a few remarks will be made here. The choice of n insures that $n! > 2 \cdot 10^d$. There are two ways to reach Step 2: from Step 1 or from Step 3. When Step 1 finishes, we have $K = 0$, $B = 1$, $A = 1$: thus the requirements of (b) are easy. When Step 3 finishes with a decision to go to Step 2, we have $B < E$, or $K! < 2 \cdot 10^d$, so that $K < n$.

(One of the referees for this paper remarked that this "induction" technique for proving the algorithm correct may be more natural for a mathematician, but that some computer literature uses a "loop-invariant" approach, as, for example, in [4, Chapter 11]. He went on to say that the essential details are the same as those given here.)

What do I know when (a)–(d) are established? When the algorithm stops (how do I know it will ever stop?), we have $K! > 2 \cdot 10^d$ and A/B is the partial sum S_K , so that (4) holds as required.

FIGURE 1 shows an implementation of this algorithm. (It is in Waltz Lisp for CP/M.) The computation of e to 500 decimals took 15 seconds on a 4 MHz Z-80, not including the time for printing out the result.

Space and time

How good is the algorithm that has just been described? One way to analyze such a question is to consider how the requirements for resources (space in the computer memory and time for execution) depend on the parameters of the problem (in this case, d , the size of the output).

```

; compute e to d decimals
; July 1, 1985

(defun ecalc (d)
  (prog (K A B E)
    step1
      (setq K 0)
      (setq A 1)
      (setq B 1)
      (setq E (times 2 (expt 10 d)))
    step2
      (setq K (add1 K))
      (setq A (add1 (times A K)))
      (setq B (times B K))
    step3
      (cond [(lessp B E)(go step2)])
    step4
      (princ "e, to ")(print d)
      (princ " decimals;")(terpr)
      (decimal-print A B d)
  )
)

(defun decimal-print (Num Denom Digits)
  (prog (Q)
    (cond [(minusp Num)(princ "-")(setq Num (minus Num))])
    (print (setq Q (div Num Denom)))(princ ".")(terpr)
    (do [(j 1 (add1 j))] [(greaterp j Digits)(terpr)]
      (setq Num (times 10 (sub Num (times Denom Q))))
      (print (setq Q (div Num Denom)))
      (cond
        ((zerop (mod j 1000))(terpr)(terpr))
        ((zerop (mod j 50))(terpr))
        ((zerop (mod j 10))(princ " "))
      )
    )
  )
)

->(ecalc 500)

e, to 500 decimals:
2.
7182818284 5904523536 0287471352 6624977572 4709369995
9574966967 6277240766 3035354759 4571382178 5251664274
2746639193 2003059921 8174135966 2904357290 0334295260
5956307381 3232862794 3490763233 8298807531 9525101901
1573834187 9307021540 8914993488 4167509244 7614606680
8226480016 8477411853 7423454424 3710753907 7744992069
5517027618 3860626133 1384583000 7520449338 2656029760
6737113200 7093287091 2744374704 7230696977 2093101416
9283681902 5515108657 4637721112 5238978442 5056953696
7707854499 6996794686 4454905987 9316368892 3009879312

@ nil

```

FIGURE 1

First, space. Our algorithm requires four items in memory: K , A , B , E . The memory space for K must be able to hold numbers from 0 to n . Now we know that n is of order of magnitude $O(d/\log d)$, and the space for storage is about $\log_{10} n$ digits, or $\log_2 n$ bits. So the space required for K is $O(\log d)$. The memory space for A , B , E must be able to hold integers up to about $2 \cdot 10^d$, which requires space $O(d)$. So the total storage requirements have order of magnitude $O(d)$. We might say Algorithm E has “linear” space requirements.

Now let’s consider time. How long will a computation take? Of course, that depends on the particulars of the computer and the software. But we can estimate the order of magnitude of the time requirements independently of the particular details.

(I assume we are using a “sequential” or “von Neumann” machine, however. Parallel processing may change things. If I have 500 processors, one for each digit, my estimates may be improved. See [7, Section 4.3.3 E].)

Step 1 is executed only once. Assignment of a number of size $O(d)$ digits takes time at most $O(d)$. What about computation of $2 \cdot 10^d$? In base 10, it is just counting up to d , or time $O(d)$. In base 2, it is at most time $O(d^2)$: multiply a d digit number by 10 at most d times. (Actually, it can be done in less time. See [7].)

Step 2 is executed at most $n = O(d/\log d)$ times. The computation of $K + 1$, where K has $O(\log d)$ digits, is done in time $O(\log d)$. Multiplications $A \cdot K$ and $B \cdot K$ involve numbers with $O(d)$ digits multiplied by numbers with $O(\log d)$ digits. This requires at most time $O(d \log d)$. So once through Step 2 takes time $O(d \log d)$; this is done at most $O(d/\log d)$ times, so the total time in Step 2 is at most $O(d^2)$.

Step 3 is executed at most $n = O(d/\log d)$ times. A comparison of two numbers with $O(d)$ digits each takes at most time $O(d)$. So the total time in Step 3 is at most $O(d^2/\log d)$.

Step 4 is executed once. The division A/B , including converting to d digits in decimal, is at most time $O(d^2)$. (Actually, division can be done faster than this: see [7, Section 4.3.3 D].)

The grand total for Algorithm E is time $O(d^2)$. This is an example of a “polynomial time” algorithm. The time is bounded by a polynomial $A + Bd^2$ in this case.

And now, pi

Consider first this formula of Leibniz:

$$\frac{\pi}{4} = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1}. \quad (5)$$

This is not a good formula to use for our purposes. To see why not, consider the running time of an algorithm based on it. The number n of terms required for the partial sum S_n to be correct to d decimals satisfies (roughly)

$$\frac{1}{2n+3} < 0.5 \cdot 10^{-d}$$

or $n = 10^d$, approximately. But this rate of growth, $n = O(10^d)$, is much worse than the $O(d^2)$ found above. Let’s suppose each term can be processed in 10^{-9} seconds. How long will it take to do 10^{500} terms? Only 10^{491} seconds, which is over 10^{483} years. I’m not going to wait for that.

Does this mean that computing 500 digits of π is hopeless? Not at all. It means only that formula (5) is not the way to do it.

I will use here the Maclaurin series

$$\arctan x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{2k+1}, \quad -1 < x \leq 1$$

(see [12, Theorem 5.33]), together with this identity, due to Machin:

$$\pi = 16 \arctan(1/5) - 4 \arctan(1/239)$$

(see [12, Theorem 5.34] or, better yet, try to prove it yourself).

We start with a positive integer d , and we want to find a number A so that $|A - \pi| < 0.5 \cdot 10^{-d}$. It will be enough to take $A = 16A_5 - 4A_{239}$, where A_5 and A_{239} satisfy

$$|A_5 - \arctan(1/5)| < \frac{1}{64} 10^{-d}$$

$$|A_{239} - \arctan(1/239)| < \frac{1}{16} 10^{-d}.$$

For these approximations we will use partial sums of the series above.

For the approximation of $\arctan(1/5)$, I need to find n so that

$$\left| \sum_{k=0}^n \frac{(-1)^k}{(2k+1)5^{2k+1}} - \arctan(1/5) \right| < \frac{1}{64} 10^{-d}.$$

The error in an alternating series is bounded by the absolute value of the first missing term. So it suffices to have

$$\frac{1}{(2n+3)5^{2n+3}} \leq \frac{1}{64} \cdot 10^{-d}.$$

For this, it suffices that $5^{2n} \geq 64 \cdot 10^d$, or

$$n \geq \frac{d \log 10 + \log 64}{2 \log 5}.$$

So the order of magnitude required is $n = O(d)$.

Now we could continue from here roughly as in the previous case. The number of digits in the numerator and denominator will exceed $2d$. Instead of proceeding exactly this way, I will make some further approximations, so that only about d digits are needed for each number stored. An easy way to do this, but still keep control of the error estimates, is to choose an integer p slightly larger than d , and compute with numbers of the form $A/10^p$, where A is an integer. (This is the same thing as using decimals with p places to the right of the decimal point.)

So in fact, n should be chosen so that

$$\left| \sum_{k=0}^n \frac{(-1)^k}{(2k+1)5^{2k+1}} - \arctan(1/5) \right| < \frac{1}{128} \cdot 10^{-d}$$

and then S computed so that

$$\left| \frac{S}{10^p} - \sum_{k=0}^n \frac{(-1)^k}{(2k+1)5^{2k+1}} \right| < \frac{1}{128} \cdot 10^{-d}.$$

A similar calculation must be done with denominator 239. The approximations are:

$$n_5 = \frac{d \log 10 + \log 128}{2 \log 5}$$

$$n_{239} = \frac{d \log 10 + \log 32}{2 \log 239}$$

both of order of magnitude $O(d)$. We will see below that the choice $p \geq d + \log_{10}(384(n_5 + 1))$ is enough. When $d = 500$, we have $p = 506$. Here is our algorithm.

Algorithm A: Approximate $\arctan(1/m)$ with error $\leq \epsilon$ (where $m \geq 2, \epsilon > 0$). First, let n satisfy

$$\left| \sum_{k=0}^n \frac{(-1)^k}{(2k+1)m^{2k+1}} - \arctan(1/m) \right| < \frac{1}{2} \epsilon,$$

then let p satisfy $p \geq \log_{10}(6(n+1)/\epsilon)$. The memory requirements are for four integers: N and

J capable of storing numbers from 0 to $2n+1$; S and T capable of storing numbers from -10^p to 10^p .

Step 1: $N \leftarrow 2n+1, J \leftarrow 1, S \leftarrow \text{quotient}(10^p, m), T \leftarrow S$.

Step 2: $J \leftarrow J+2, T \leftarrow -\text{quotient}(T, m^2), S \leftarrow S + \text{quotient}(T, J)$.

Step 3: if $J < N$, then go to Step 2.

Step 4: the result is $S/10^p$, stop.

The notation $\text{quotient}(T, J)$ denotes the integer quotient of T by J , that is, the greatest integer in T/J .

In order to show that Algorithm A does what is claimed, one would prove the following assertions by induction.

(a) At all times: N, J, S, T are integers.

(b) Whenever Step 2 begins: $J = 2k+1$ for some integer k with $0 \leq k < n$,

$$\left| T - 10^p \frac{(-1)^k}{m^{2k+1}} \right| < \frac{m^2}{m^2 - 1}, \quad (6)$$

$$\left| S - 10^p \sum_{j=0}^k \frac{(-1)^j}{(2j+1)m^{2j+1}} \right| < 3(k+1). \quad (7)$$

(c) Whenever Step 3 begins: $J = 2k+1$ for some integer k with $0 < k \leq n$, and (6), (7) hold.

(d) Whenever Step 4 begins: $J = 2n+1$ and

$$\left| \frac{S}{10^p} - \sum_{j=0}^n \frac{(-1)^j}{(2j+1)m^{2j+1}} \right| < \frac{3(n+1)}{10^p}.$$

The desired inequality follows from the choice of p so that

$$\frac{3(n+1)}{10^p} < \frac{1}{2}\epsilon.$$

The verifications of (a)–(d) are easy. Let us consider first (b). There are two ways to reach Step 2: from Step 1 and from Step 3. If Step 2 begins following Step 1, we have $J = 1$, $S = T$, and $|S - 10^p/m| < 1$. So the assertions in (b) are true. On the other hand, suppose Step 2 begins following Step 3. Then $J < N$, so $k < n$, and we have (6) and (7) from the corresponding inductive assumptions for Step 3.

Now let us consider (c). The previous information concerns Step 2. Let $J' = 2k' + 1$, S' , T' be the values of these variables at the beginning of the preceding Step 2. Then $J = J' + 2$, so $k = k' + 1$. Since $0 \leq k' < n$, we have $0 < k \leq n$. Since (6) holds for T' and k' , we deduce

$$\left| \frac{-T'}{m^2} - 10^p \frac{(-1)^{k'+1}}{m^{2k'+3}} \right| < \frac{1}{m^2 - 1}.$$

But also $T = -\text{quotient}(T', m^2)$, so

$$\left| T - 10^p \frac{(-1)^k}{m^2} \right| < 1.$$

Therefore,

$$\left| T - 10^p \frac{(-1)^k}{m^{2k+1}} \right| < 1 + \frac{1}{m^2 - 1} = \frac{m^2}{m^2 - 1},$$

so that (6) is satisfied when Step 3 begins. Next, we have

$$\left| \text{quotient}(T, J) - \frac{T}{2k+1} \right| < 1,$$

and

$$\left| \frac{T}{2k+1} - 10^p \frac{(-1)^k}{m^{2k+1}} \right| < \frac{m^2}{(2k+1)(m^2-1)} < 2.$$

By the inductive assertion from Step 2, we know that (7) holds for S' and k' . Combining these produces (7) for S and k .

The rest of the verification of (a)–(d) is left to the reader.

Now let's discuss how efficient the algorithm is. In the cases we are interested in, $m = 5$ or $m = 239$ is independent of d ; $n = O(d)$; $p = O(d)$. So the space required is of order $O(d)$.

The time for Step 1 is at most $O(d^2)$.

Step 2 involves dividing a number with $O(d)$ digits by one with $O(\log d)$ digits. The time for this is at most $O(d \log d)$. Step 2 is done at most n times, so the total time is $O(d^2 \log d)$.

The time for doing Step 3 once is $O(\log d)$, so the total time for Step 3 is $O(d \log d)$.

The time for Step 4, including conversion to decimal, is $O(d^2)$.

So everything together requires time at most $O(d^2 \log d)$. Computation of π involves multiplying two of these results by 16 and 4, and adding them together. The time required to compute d digits of π is thus at most $O(d^2 \log d)$.

FIGURE 2 shows an implementation of this computation (with minor changes). This time it is written for the MACSYMA computer algebra system. (See, for example, [10].) TABLE 1 shows some times used by this program on a DEC System 2060 computer.

```

picsalc(digits) :=
(
  d : digits,
  p : ev(d + 1 + entier(log(384*d)/log(10)),numer),
  tentop : 10^p,
  16*arctan(5) - 4*arctan(239)
) $

arctan(m) :=
BLOCK ([S, T, msquared],
  T : quot(tentop,m),
  S : T,
  msquared : m^2,
  FOR J : 3 STEP 2 DO
  (
    T : -quot(T, msquared),
    IF T=0 THEN RETURN (S),
    S : S + quot(T,J)
  )
) $

quot(Num,Den) := entier(Num/Den) $

```

FIGURE 2

TABLE 1		
Number of digits	Time (msec)	
d	t	t/d^2
100	2486	.249
200	5122	.128
300	8483	.094
400	10182	.064
500	13791	.055
600	17911	.050
700	22081	.045
800	26703	.042
900	31276	.039
1000	36844	.037

More and better

There are many related things that could be done. I will leave them to the interested student. They range from simple exercises to class projects to master's theses.

1. Some of the estimates used above are crude. Does Algorithm A really take time $O(d^2 \log d)$, or is it faster than that? The last column of TABLE 1 seems to suggest that even $O(d^2)$ is too pessimistic an estimate. Is it? What would happen if the partial sums of $\arctan(1/5)$ were computed using exact rational numbers? The least common denominator is the least common multiple of $5^{2n+1}, 1, 3, 5, \dots, 2n+1$. This has $O(n)$ digits. (See Chapter XXII of [5], especially Theorem 434.) But in order to remain this small, the rational numbers will have to be reduced to lowest terms. How will this affect the time required? (See [7, Section 4.5.2].)

2. Analyze the space and time requirements for implementation of a computation of π using another formula. Some possibilities:

(a) Wallis:

$$\frac{\pi}{2} = \frac{2 \cdot 2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdots (2n)(2n) \cdots}{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdots (2n-1)(2n+1) \cdots}$$

(b) Vieta:

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}}} \cdots$$

(c) Brouncker:

$$\frac{\pi}{4} = \frac{1}{1 + \frac{1^2}{2 + \frac{3^2}{2 + \frac{5^2}{2 + \cdots}}}}$$

(d) Other arctangent methods, see [9]. Or, Gauss's arithmetic-geometric mean, see [11], [3]. This one is supposed to do better than $O(d^2)$ time.

3. Compute many decimals for some other well-known constants. (See [7, Table 1].) For example:

- (a) Natural logarithms, such as $\log 2$.
- (b) Euler's constant

$$\gamma = \lim_{n \rightarrow \infty} \left[\sum_{k=1}^n \frac{1}{k} - \log n \right].$$

This particular formula converges too slowly to be used directly.

- (c) The Riemann zeta function, say,

$$\zeta(3) = \sum_{k=1}^{\infty} \frac{1}{k^3}.$$

This, too, converges slowly; so it must be rewritten to be practical to use for computation.

- (d) Catalan's constant

$$G = 1 - \frac{1}{3^2} + \frac{1}{5^2} - \frac{1}{7^2} + \cdots.$$

4. If a real number is rational, then d decimals can be computed using space $O(1)$ and time $O(d)$. Does this characterize the rational numbers? Are there real numbers whose decimal expansion cannot be computed at all? Are there real numbers whose decimal expansions can be computed, but not in polynomial time? (That is, in time $O(d^n)$ for some integer n .)

Finally, here are a few references that may be of related interest. Knuth [7, Sections 4.3 and 4.4] discusses implementation and timing of multidigit arithmetic. More material on proving programs (or algorithms) correct is in [1], [4], [8]. Other remarks on difficulty of computation of π , e , and $\sqrt{2}$ are in [2].

References

- [1] R. B. Anderson, *Proving Programs Correct*, Wiley, 1979.
- [2] L. Baxter, Are π , e , and $\sqrt{2}$ equally difficult to compute?, *Amer. Math. Monthly*, 86 (1979) 50–51.
- [3] J. M. Borwein and P. B. Borwein, A very rapidly convergent product expansion for π , *BIT*, 23 (1983) 538–540.
- [4] D. Gries, *The Science of Programming*, Springer-Verlag, 1981.
- [5] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Fifth Edition, Oxford, 1979.
- [6] E. Kasner and J. R. Newman, *Mathematics and the Imagination*, Simon and Schuster, 1940.
- [7] D. E. Knuth, *The Art of Computer Programming*, Volume 2, *Seminumerical Algorithms*, Addison-Wesley, 1969.
- [8] Z. Manna, *Lectures on the Logic of Computer Programming*, SIAM, 1980.
- [9] G. Miel, An algorithm for the calculation of π , *Amer. Math. Monthly*, 86 (1979) 694–697.
- [10] R. H. Rand, *Computer Algebra in Applied Mathematics: An Introduction to MACSYMA*, Pittman, 1984.
- [11] E. Salarmin, Computation of π using arithmetic-geometric mean, *Math. of Computation*, 30 (1976) 565–570.
- [12] K. R. Stromberg, *An Introduction to Classical and Real Analysis*, Wadsworth, 1981.
- [13] H. S. Wilf, The disk with the college education, *Amer. Math. Monthly*, 89 (1982) 4–8.

“If you ask mathematicians what it is that they do, they always give you the same answer. They think. They think about difficult and unusual problems. (They don’t think about ordinary problems; they just write down the answers.)”

—M. Evgrafov, A byl li brak?, *Literaturnaya Gazeta* (December 5, 1979), no. 49, p. 12.

A Curious Mixture of Maps, Dates, and Names

J. M. SACHS

Northeastern Illinois University
Chicago, Illinois 60625

While looking at an exhibit of maps, I noted the date 1569 for the Mercator Projection. I recalled that the equations for this mapping involved a logarithm, and I had a vague memory that John Napier had published his work on logarithms in the 17th century, not the 16th century. I found that *Mirifici Logarithmorum Canonis Descriptio* was published in 1614. Furthermore, deriving the equations for the projection requires the tools of the calculus applied in differential geometry. Newton and Leibniz were not born until about fifty years after Mercator's death; and Gauss, who was responsible for the development of the necessary differential geometry, was born in 1777.

Edward Wright, a Cambridge professor of mathematics and a navigational consultant to the East India Company, published *Errors in Navigation Detected and Corrected* in 1599, and an expanded edition in 1610. Wright described a mathematical way to construct the Mercator map, producing a better approximation than the original. Wright knew Napier, and some years after publishing the paper referred to above, translated Napier's work on logarithms from Latin into English [1].

What led Mercator to his map? How did he construct it and how did Wright improve upon it?



Mercator's Map of the World—1569

Mercator was concerned with the creation of a map that would be easy for navigators to use. A line of constant compass direction on the earth is a spiral unless that direction is N, S, E, or W. Mercator wanted a map in which these lines of constant direction—the way a ship would have to set its course—would be straight lines. Mercator was not the first to seek a map with this property. In 1511 and 1513, Erhard Etzlaub of Nuremberg, for example, made sundials on the cover of which there are tiny maps, Equator to Arctic Circle, and West Africa eastward for about sixty degrees, with, as in Mercator, the latitude spacing increasing as one moves north [2].

To see what Mercator did, consider using the center of the earth as the center of a projection. Project the surface of the sphere onto a cylinder tangent to the earth along the equator. When the cylinder is unrolled, the equator becomes a straight line and equally spaced meridians become a system of equally spaced straight lines perpendicular to the equator. Equally spaced parallels of latitude on the sphere become unequally spaced straight lines parallel to the equator; on the map the spacing between these lines increases as the distance from the equator increases.

A 16th- or 17th-century navigator steered a course by a magnetic compass. The most useful map would have a constant compass direction on the globe, called a loxodrome, mapping into a straight line. Unfortunately, on the perspective map described above, the constant compass direction maps into a rather complex curve. Mercator proposed a change in this map. He reasoned that on the perspective map a small arc of latitude away from the equator is enlarged as is an equal arc of longitude at the same place, but the increases are not equal. He proposed to keep the enlargement in longitude as in the perspective map and to make the increase in latitude for a small arc equal to the increase in longitude at that place. The following is a translation from the Latin of the last sentence of the legend on the Mercator Map of 1569:

In view of these things, I have given to the degree of latitude from the equator towards the poles, a gradual increase in length proportionate to the increase of the parallels beyond the length which they have on the globe, relative to the equator. [3]

Mercator produced his map with compass and protractor.

To see what this means, let us consider the parametric equation of a sphere with center at the origin: u = longitude in radians, v = latitude in radians, r = radius of sphere

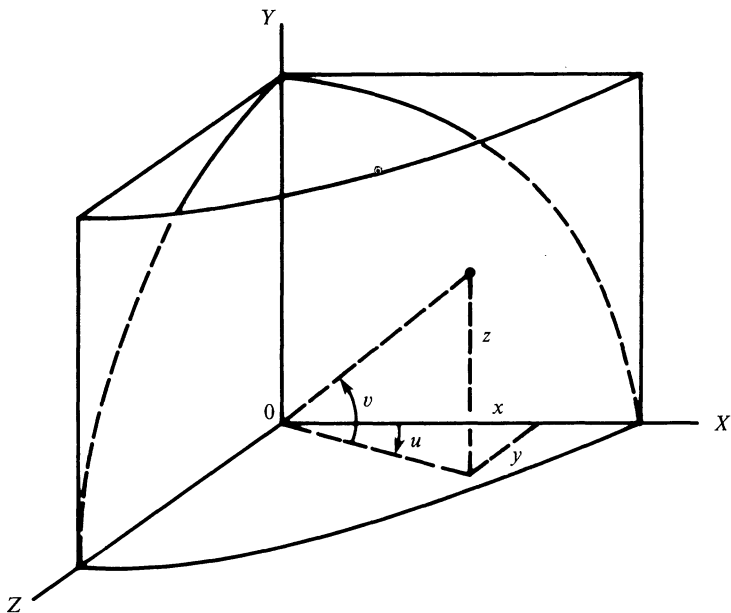


FIGURE 1

$$\begin{aligned}
 x &= r \cos u (\cos v) \\
 y &= r \sin u (\cos v) \\
 z &= r \sin v.
 \end{aligned}
 \tag{1}$$

If the unrolled tangent cylinder has a UV coordinate system, the equations of the Mercator Projection (to be derived later) are:

$$\begin{aligned}
 U &= -ru \\
 V &= r \ln(\sec v + \tan v).
 \end{aligned}$$

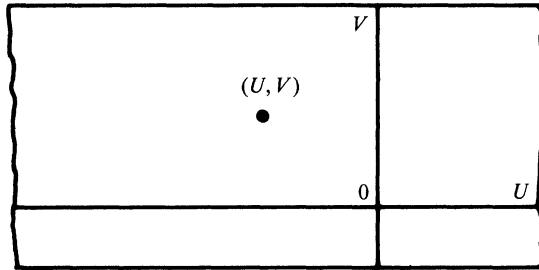


FIGURE 2

Imagine that we are working with small arc distances at latitude v on the surface of the sphere. One degree in longitude at this constant latitude is $1/360$ of a circle whose radius is $r \cos v$, where r is the radius of the earth. The circumference of this circle is $2\pi r \cos v$, so the length of one degree of arc is thus $2\pi r \cos v / 360$. Since the meridians on the map are equally spaced parallel straight lines, one degree of arc in longitude is everywhere equal to $2\pi r / 360$. Thus, the ratio of the earth length, E , of a small arc in longitude at the latitude, v , to the map length, M , of this small arc is $E/M = \cos v$, so that $E = M \cos v$ and $M = E \sec v$.

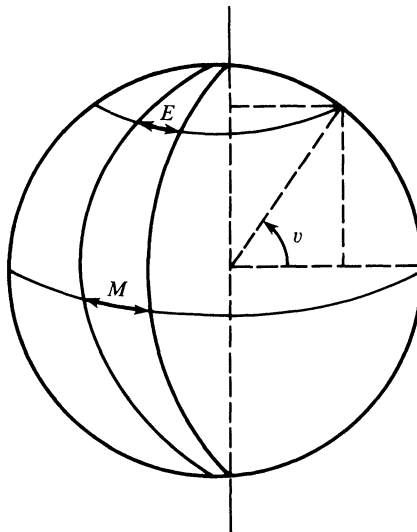


FIGURE 3

Thus, a map can be constructed by approximations using small increments in latitude and matching the stretch in latitude to the stretch in longitude.

Mercator proposed a way to approximate distance by using similar triangles on the map [4]. The difference in latitude of two points A and B on the map is noted in FIGURE 4. An arbitrary

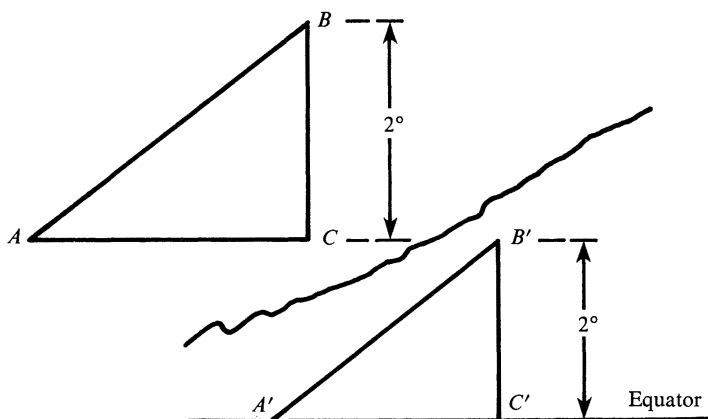


FIGURE 4

point C' is selected on the equator, and the difference in latitude is then marked off on a meridian through C' . This difference is $C'B'$. A parallel to the loxodrome joining A and B is constructed through B' . The point where this parallel intersects the equator is labeled A' . Then $A'B'$ is laid off along the equator. The number of degrees along the equator leads to an approximation for the distance AB . If in the example $A'B' = 3^\circ 12'$, the distance AB would be 192 nautical miles, since one minute of arc of a great circle on the earth is one nautical mile.

The Mercator map was not received with great acclaim by the sea captains in the late 16th century. Navigating was a tricky business and was more of an art than a science. Moreover, Mercator was an academic rather than a sailor so there was suspicion as well as a lack of understanding.

Edward Wright's treatise helped to popularize the loxodrome-into-straight-line map. Wright's work, dedicated to his patron, the Prince of Wales, was widely distributed to sea captains, particularly English ones. Wright provided a vivid and descriptive way to obtain the map as well as a careful mathematical development of how to space the parallels of latitude in making the map. He suggested that the sphere with meridians, parallels, loxodromes, etc. be visualized as a bladder surrounded by the tangent cylinder. (We would use a balloon for the bladder.) The bladder was inflated so that as it stretched it fit against the walls of the cylinder, the stretch in latitude being everywhere equal to the stretch in longitude. This description gave nonmathematicians a picture of what the chart did.

Wright made a table of secants with one-minute intervals. If the spacing of meridians along the map of the equator is at one-minute intervals and is taken as a unit, then the spacing of parallels along a meridian, one minute apart, can be found by multiplying that unit by the secant of the latitude. The drawing and table below show how this works. The intervals here are 10 degrees so that changes in the secant are larger, but the principle is the same.

sec	$10^\circ = 1.0154$
sec	$20^\circ = 1.0642$
sec	$30^\circ = 1.1547$
sec	$40^\circ = 1.3055$
sec	$50^\circ = 1.5557$
sec	$60^\circ = 2.0000$
Sum	$\overline{8.0955}$

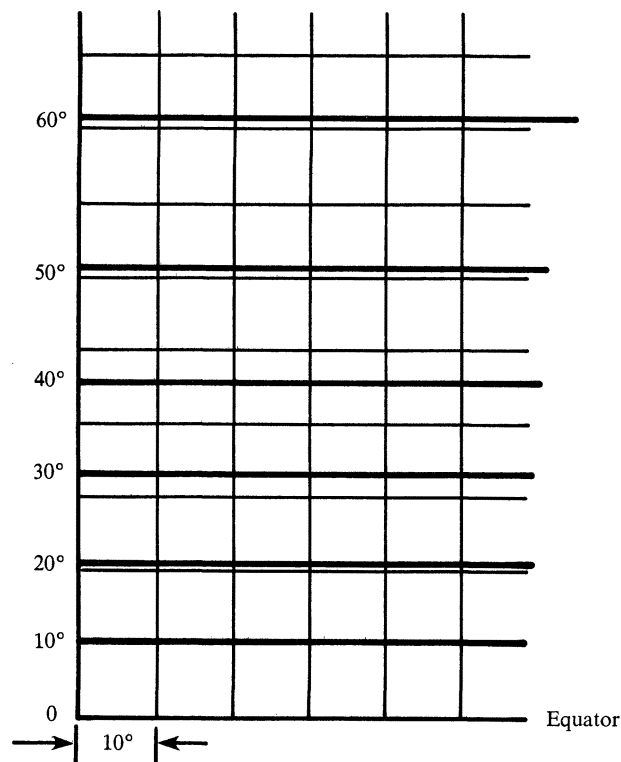


FIGURE 5

To correct for distance distortion, Wright made a table showing the change in distance scale, due to the stretch of the map, for each one-degree increase in latitude [5]. As shown by the equations accompanying FIGURE 3, this meant multiplying the map length by the cosine of the latitude. For short distances and for those involving only small changes in latitude, the approximations were reasonably good. Longer distances and those involving large differences in latitude could be approximated by dividing the loxodrome into small segments and then using the appropriate figure from the table for each segment.

Richard Hakluyt, who gives us invaluable accounts of exploration and discovery along with their economic and diplomatic backgrounds, included what was probably one of Wright's maps in his magnificent work written at the end of the 16th century [6], though he gave no one credit. The legend on the map reads:

Thou hast here (gentle reader) a true hydrographical description of so much of the world as hath been hitherto discovered and has come to our knowledge;... that all places herein set downe have the same positions... that they have in the globe, being therein placed in the same longitudes and latitudes which they have in this chart.... But to finde the distance if both places have the same latitude, see how many degrees of the meridian taken at that latitude are contained between the two places, for so many score leagues is the distance; if they differ in latitude see how many degrees of the meridian taken about the midst of the difference are conteyned between them and so many score leagues is the distance.

One minute in latitude is the equivalent of a nautical mile, and a league, in England at least, was three nautical miles. The drawings below show how Hakluyt would approximate distances on

a Mercator map, converting degrees and minutes into leagues and nautical miles.

How can a Mercator map now be constructed, taking advantage of the differential geometry of Gauss, who used the tools of the calculus to define curvature for surfaces and conformality for maps? (Conformal in mathematics refers to a transformation which preserves angles. More generally, conformal refers to a mapping which preserves local shape. A map projection which is conformal in the mathematical sense does preserve local shape.) Mercator wanted a constant compass direction to become a straight line on a map which carried the meridians into a family of parallel lines perpendicular to the equator, and the parallels of latitude into a family of straight lines parallel to the equator. If a cylindrical map is conformal, the constant compass directions on the earth will become straight lines on the map. Thus, in our terms, Mercator was seeking a cylindrical conformal map.

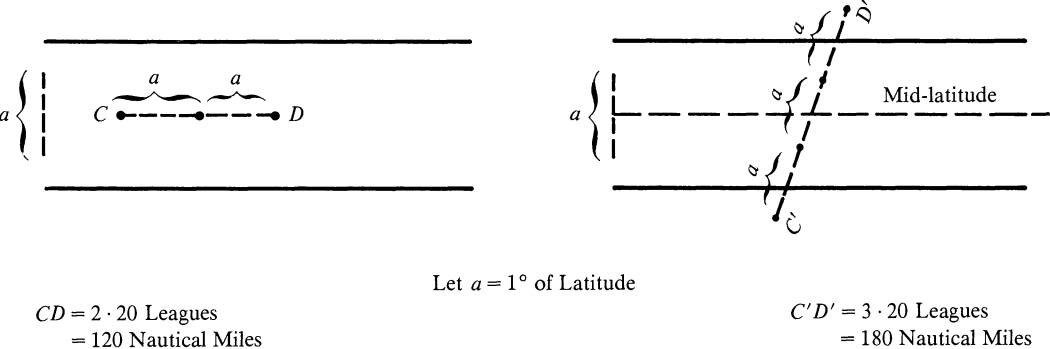


FIGURE 6

If we use equations (1) for the parametric representation of the sphere with the families $u = c$, meridians, and $v = k$, parallels of latitude, as the parametric net, the linear element or first fundamental form is given by

$$ds^2 = E du^2 + 2F du dv + G dv^2$$

where

$$E = \left(\frac{\partial x}{\partial u} \right)^2 + \left(\frac{\partial y}{\partial u} \right)^2 + \left(\frac{\partial z}{\partial u} \right)^2,$$

$$F = \frac{\partial x}{\partial u} \cdot \frac{\partial x}{\partial v} + \frac{\partial y}{\partial u} \cdot \frac{\partial y}{\partial v} + \frac{\partial z}{\partial u} \cdot \frac{\partial z}{\partial v},$$

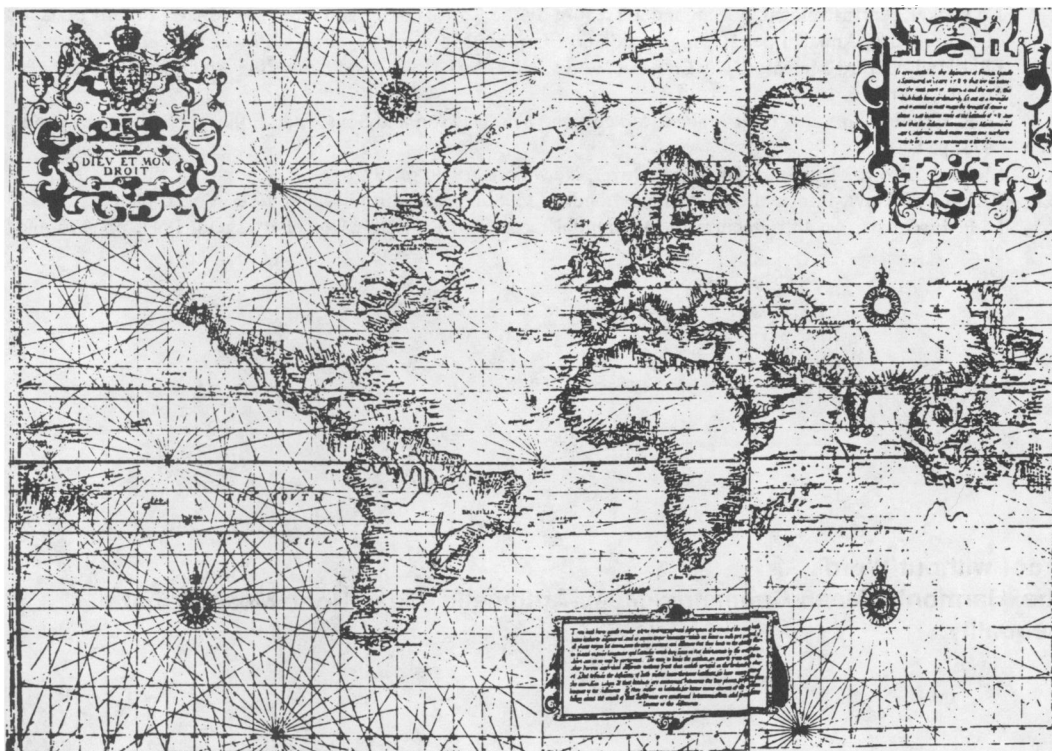
and

$$G = \left(\frac{\partial x}{\partial v} \right)^2 + \left(\frac{\partial y}{\partial v} \right)^2 + \left(\frac{\partial z}{\partial v} \right)^2 [7].$$

From (1) we get $E = r^2 \cos^2 v$, $F = 0$, and $G = r^2$ for the sphere. Thus,

$$ds^2 = r^2 (\cos^2 v du^2 + dv^2).$$

If the cylindrical map is to be conformal, then the first fundamental form for the map must be proportional to the first fundamental form for the sphere [7]. If dS^2 is the first fundamental form



**World Map from Hakluyt's Principall Navigations
(Attributed to Edward Wright, 1598.)**

for the map, then $dS^2 = t^2 ds^2$. Retaining the cylindrical form for the map, we have $U = -ru$ and we seek a mapping for V which will yield conformality. If we have

$$dS^2 = E' dU^2 + 2F' dUdV + G' dV^2,$$

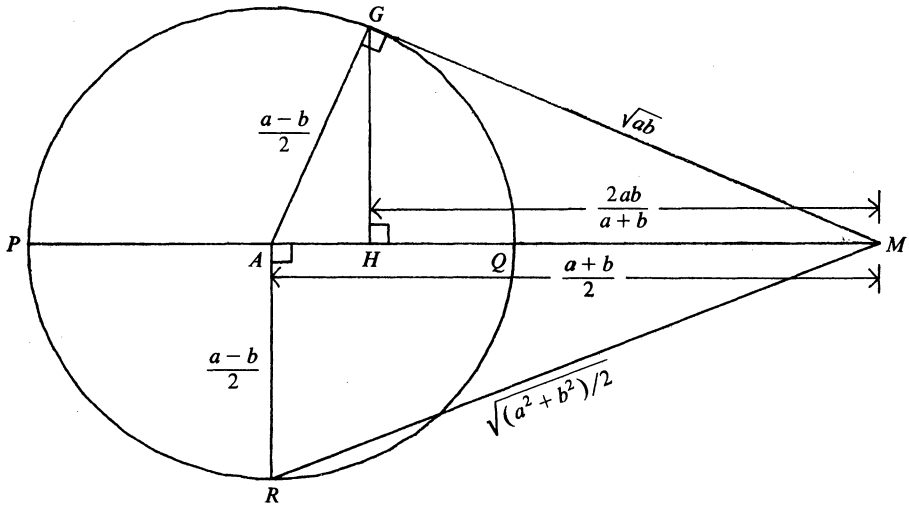
then $F' = 0$ since $F = 0$. This leads to the conclusions that $E'/E = G'/G$, V is a function of v only, and $dV/dv = r \sec v$. Integrating we get $V = r \ln(\sec v + \tan v)$.

In the 16th century, Mercator created his approximations based on making the N-S stretch equal to the E-W stretch on a tangent cylinder, thus mapping loxodromes into straight lines. In the late 16th and early 17th centuries, Wright used his table of secants to obtain a more accurate map with easily used tables for approximations for distance. A few years later, Napier wrote his treatise on logarithms and Wright translated it from Latin into English. Neither connected this new tool to the map refined by Wright. Late in the 17th century, Newton and Leibniz did the basic work on the calculus, and more than 100 years after that Gauss produced the differential geometry needed to derive the equations of the projection. The concept developed so imaginatively by Mercator and his predecessors had immense impact on navigation and exploration. The refinements by Wright made the maps even more useful. Some very unusual men had the vision that enabled them to find a practical solution to a serious problem when the theoretical tools were not yet available.

References

- [1] William M. McKinney, The Journal of Geography, Chicago, Vol. 68 no. 8, 472.
- [2] J. Drecker, An Instrument, a Map and a Treatise by the Cartographer and Compass Maker, Erhard Etzlaub of Nuremburg, Annalen der Hydrographie und Maritime Meteorology, June 1917.
- [3] Elial Hall, Gerard Mercator; His Life and Works, New York, 1878, read before the American Geographical Society, April 16, 1878.
- [4] G. R. Crone, Maps and Their Makers, 5th edition, Wm. Dawson and Sons, Ltd., Folkestone, Kent, England, 1978, p. 64.
- [5] Lloyd A. Brown, The Story Of Maps, Dover, New York, 1979, p. 137.
- [6] Richard Hakluyt, Principall Navigations, London, 1598, reprinted in Glasgow, 1903, Vol. I, plate 5.
- [7] L. P. Eisenhart, An Introduction to Differential Geometry, Princeton University Press, 1947, pp. 126 and 202.

Proof without Words: The Harmonic Mean-Geometric Mean-Arithmetic Mean-Root Mean Square Inequality.



$$PM = a, QM = b, a > b > 0$$

$$HM < GM < AM < RM$$

$$\frac{2ab}{a+b} < \sqrt{ab} < \frac{a+b}{2} < \sqrt{\frac{a^2+b^2}{2}}$$

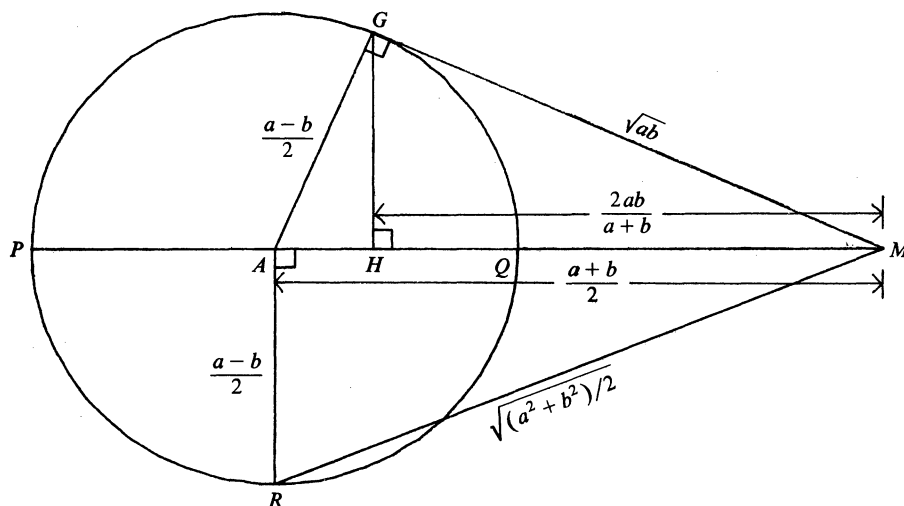
—ROGER B. NELSEN
Lewis and Clark College

References

- [1] William M. McKinney, The Journal of Geography, Chicago, Vol. 68 no. 8, 472.
- [2] J. Drecker, An Instrument, a Map and a Treatise by the Cartographer and Compass Maker, Erhard Etzlaub of Nuremburg, Annalen der Hydrographie und Maritime Meteorology, June 1917.
- [3] Elial Hall, Gerard Mercator; His Life and Works, New York, 1878, read before the American Geographical Society, April 16, 1878.
- [4] G. R. Crone, Maps and Their Makers, 5th edition, Wm. Dawson and Sons, Ltd., Folkestone, Kent, England, 1978, p. 64.
- [5] Lloyd A. Brown, The Story Of Maps, Dover, New York, 1979, p. 137.
- [6] Richard Hakluyt, Principall Navigations, London, 1598, reprinted in Glasgow, 1903, Vol. I, plate 5.
- [7] L. P. Eisenhart, An Introduction to Differential Geometry, Princeton University Press, 1947, pp. 126 and 202.

Proof without Words:

The Harmonic Mean-Geometric Mean-Arithmetic Mean-Root Mean Square Inequality.



$$PM = a, QM = b, a > b > 0$$

$$HM < GM < AM < RM$$

$$\frac{2ab}{a+b} < \sqrt{ab} < \frac{a+b}{2} < \sqrt{\frac{a^2+b^2}{2}}$$

—ROGER B. NELSEN
Lewis and Clark College

Murphy's Law and Probability or How to Compute Your Misfortune

GENE G. GARZA

The University of Montevallo
Montevallo, AL 35115

Have you ever complained to a friend that everything seems to go wrong at the same time, only to hear your friend say, sympathetically, "Well—that's Murphy's Law for you!" or "When it rains, it pours!" By the use of an appropriate mathematical model we can show that the clustering of unpleasant events in our daily lives is, in fact, a result of simple probability theory and has very little to do with what we normally call "bad luck." We will apply the model to Murphy's Law and then use it to compute some "Murphy" probabilities.

That probability should be mentioned with respect to Murphy's Law is, indeed, appropriate, since some versions of Murphy's Law are, in fact, simply restatements of certain well-known facts from probability theory. For example, consider the statement of Murphy's Law which says, "If something bad happens, then something else bad is going to happen." For those versed in probability, this is obviously nothing more than a statement of conditional probability. The way in which Murphy's Law affects our lives is already painfully obvious. For example, if you have two normally healthy children, then it is likely that when one gets sick the other will get sick—even if the second one is immune to whatever the first one has. Furthermore, it is likely that your spouse will also be ill at the same time. What we will soon discover is that events such as these tend to cluster together. The observation can then be made that good events should cluster together also. Of course, people do not complain when they are suddenly "plagued" by good fortune.

The model we need will represent "life and certain events called mishaps." It consists of twenty-four time periods, say, months, and twenty-four mishaps which occurred during those time periods. Notice that I said "occurred," so that we will, in effect, be looking at how the events clustered together rather than trying to predict how they will occur. We will assume that the mishaps are identical in the frustration they cause or in the amount by which they tend to upset us. Furthermore, we assume that each of these events is "equally likely" to occur during any given time period and is independent of the others in the sense that the occurrence of one has no effect on the occurrence of another. These assumptions may not be completely valid, but they at least make the problem tractable. For example, taking a child to the pediatrician may be much more upsetting than taking a car to the shop unless the child has only a cold while the car is in need of a new transmission!

Let's start by calculating the probability that some time period will have two or more mishaps. This probability is easy to compute if we do it in the right manner. The right manner is to consider the complementary event, that of no period having two or more mishaps. Since there are the same number of mishaps as time periods, this means there must be exactly one mishap per period. We can solve this by noting that the first mishap can be placed in any of twenty-four different time periods. Thus, with only one mishap the probability of no period having two or more mishaps is computed as

$$\begin{aligned} P(\text{"no period with two or more"}) &= \frac{\text{\# favorable choices}}{\text{\# possibilities}} \\ &= \frac{24}{24} = 1 \end{aligned}$$

which is exactly as we would expect. However, now suppose that we are ready to place the second mishap. The possible outcomes number 24×24 since the first mishap may go in any period and

the second mishap may go in any period. But, the number of possible choices is somewhat different. In order to be successful (i.e., to fail!) the first mishap may go into any of the twenty-four time periods; however, the second mishap must not go into the same period as the first, so that after we have chosen the first period we have only twenty-three alternatives for the second period. This means that the probability of failing to have two or more mishaps in one time period is computed as

$$P(\text{failure}) = \frac{24 \times 23}{24 \times 24} = \frac{23}{24} .$$

Thus the probability of “two or more in some period” with only two mishaps is

$$P(\text{success}) = 1 - P(\text{failure})$$

or
$$1 - \frac{23}{24} = \frac{1}{24} \approx 0.0416 .$$

If we add a third mishap the probability becomes

$$\begin{aligned} P(\text{failure}) &= \frac{24 \times 23 \times 22}{24 \times 24 \times 24} \\ &\approx .8785, \end{aligned}$$

so that

$$\begin{aligned} P(\text{success}) &\approx 1 - .8785 \\ &= .1215. \end{aligned}$$

Continuing in this manner we would find that the probability of “two or more in one time period” passes the 50% mark when the number of mishaps is only seven. The table below shows the probability of some period having at least two mishaps as being 1.0 (we are rounding off to five decimal places). The first column shows the number I of mishaps, the second shows the probability of a specific period having exactly I mishaps; the third, of having I or more mishaps; the fourth, of having $I + 1$ (or more) mishaps if it is known that I mishaps have already occurred; the fifth, of some period having at least I mishaps; the sixth, of some “two-month period” (not necessarily consecutive) having I or more mishaps; and, finally, the seventh shows the probability of some two-month period having $I + 1$ (or more) mishaps if it is known that I mishaps have already occurred. (We shall indicate how this table was obtained shortly.)

Table for 24 mishaps in 24 time periods						
I	Period Has Exactly I	Period Has I or More	Period Has $I + 1$ or More Given I	Some Period with at Least I	2 Periods Have I or More	2 Periods Have $I + 1$ or More Given I
0	.36008	1				
1	.37574	.63992	.41284	1		
2	.18787	.26418	.28889	1		
3	.05990	.07632	.21513	.92445	1	1
4	.01367	.01642	.16722	.35214	1	.92445
5	.00238	.00274	.13391	.06510	.92445	.76065
6	.00033	.00037	.10952	.00882	.70315	.40797
7	.00004	.00004	.09093	.00097	.28687	.28036
8	.00000	.00000	.07632	.00009	.08043	.19794

Anyone who has studied probability will recognize our last problem (that of calculating the probability of two or more mishaps in some time period) as being nothing more than a version of the “birthday problem” [1]. The birthday problem, for those who have not studied probability, asks “How many people should be present at any gathering for the probability of ‘two having the same birthday’ to be more than 50%?” The answer to the question is a surprisingly low twenty-three. Another way to interpret this would be to assume the birthdays are mishaps instead

(some people consider them to be mishaps already!). Then we could state that with only twenty-three mishaps in 365 days, there is a 50–50 chance that at least two will occur on the same day.

From the table we may extract many interesting facts. First of all, consider the entry which states that the probability of a month having zero mishaps is 0.36008. Multiplying this number by 24 we obtain 8.742, which is the number of periods that can be expected to have no mishap at all. Exactly what does that mean? Well, if we subtract this number from 24 (obtaining 15.258), we have what we shall call the expected number of periods in which a mishap is likely to occur. That is, all 24 mishaps can be reasonably expected to occur or to cluster within some 15 time periods (which we should not assume to be consecutive). We can expect these fifteen time periods to be marked by a mishap in the same manner that we can expect 10000 tosses of a coin to land heads 5000 times. That is, it is little more than an average.

Next, consider the probability that some month will have at least three mishaps. From the table this is seen to be 92%. This is followed by a probability of 35% that some month will have at least four mishaps. This clearly shows the clustering effect and allows us finally to understand “why everything seems to go wrong at the same time.”

Now let’s look at the probability of a specific month having, say, two or more mishaps. From the table we find 26%; however, from the next column we see the probability that a month will have two or more mishaps, if it is known that the month has had one mishap already, is 41%. Indeed, the probability of a third mishap after two have already occurred is (from the same column) 29%. Notice what happens as we move down the column. The numbers change very slowly. These numbers represent conditional probabilities, which, loosely interpreted would say, for example, that if four mishaps have occurred, then there is a 17% chance of another, and that if the fifth one does occur then there is still a 13% chance of a sixth mishap.

Perhaps the most illuminating figures lie in the second row. Note that the probability of a specific month having exactly one mishap is 38%. Now looking at the third column in the second row we see that the probability of a second mishap if a first has already occurred is 41%. This may loosely be restated as the following: The chances of a second mishap are greater than the chances of the first mishap. This statement is quite contrary to our intuition, but accounts nicely for the version of Murphy’s Law which says that if something goes wrong then something else is bound to go wrong. It is also responsible for the clustering of events. It, however, does not state that the probability of two mishaps is greater than that of one mishap. The reader should recall that we are dealing with conditional probabilities. The table provides other figures (especially the probabilities for events occurring within two time periods) which the reader may want to interpret for himself or herself.

Let us try to understand how these numbers were obtained. This is best done by using a simple example with, say, three mishaps in three time periods. These mishaps can occur in $3 \times 3 \times 3$ or 27 ways, since the first mishap could occur in any one of the three periods and, likewise, for the other two mishaps. The sample space could then be represented as follows:

(a, b, c) (a, c, b) (b, a, c)	[1, 1, 1]
(b, c, a) (c, a, b) (c, b, a)	
(a, a, b) (a, b, a) (b, a, a)	[2, 1, 0]
(b, b, a) (b, a, b) (a, b, b)	
(a, a, c) (a, c, a) (c, a, a)	
(c, c, a) (c, a, c) (a, c, c)	
(b, b, c) (b, c, b) (c, b, b)	
(c, c, b) (c, b, c) (b, c, c)	
(a, a, a) (b, b, b) (c, c, c)	[3, 0, 0]

By $[2, 1, 0]$, for example, we mean simply that some month (although not necessarily the first one) had two mishaps; some month had only one mishap; and the remaining month had no mishap. An example of this class would be (b, a, b) , which stands for the event in which the first mishap occurs in period b , the second occurs in period a , and the third occurs in time period b along with the first mishap, so that the third time period has no mishap occurring in it. This is, then, abbreviated by $[2, 1, 0]$. Notice that this listing does contain all twenty-seven possible ways in which three mishaps might occur in three time periods. These twenty-seven elements or “elementary events” have been grouped into classes labeled by: $[1, 1, 1]$, $[2, 1, 0]$ and $[3, 0, 0]$. These classes are the proper way to study these and other combinations of mishaps and time periods, especially our original with twenty-four mishaps in twenty-four months, which has a sample space consisting of 24^{24} elements. This is roughly a number with thirty-three zeros.

Certainly the events $[1, 1, 1]$, $[2, 1, 0]$, and $[3, 0, 0]$ are not equally likely, and we have a new problem—that of computing their probability. Let’s start with $[2, 1, 0]$. The proper way to compute the probability of this event is to try to compute the number of “elementary events” which this new event represents and then divide by $3 \times 3 \times 3$ (why?). One way is to list all 27 possibilities, as we did in our sample space, and determine how many lead to the outcome $[2, 1, 0]$. We see the answer is 18. Another way is to develop a method for counting directly how many elementary events correspond to the outcome $[2, 1, 0]$. This can be done as follows. With $[2, 1, 0]$ we shall first choose one month for the two mishaps: there are $C(3, 1)$ or 3 ways of doing this (the notation $C(m, n)$ stands for the number of combinations of m objects taken n at a time and is computed as $m!/(m-n)!n!$, where $x!$ represents the product of all integers less than or equal to x but greater than 0). Next we must choose a month in which the remaining mishap may occur (but not the month already chosen). Since there are now only two periods to choose from, this can be accomplished in $C(2, 1)$ or 2 ways. And, finally, we note that with three mishaps, there are $3!$ arrangements or ways in which this number of mishaps may occur, but if, as in this case, two months are the same then we must divide by $2!$. This gives a total of $3 \times 2 \times 3!/2!$ or 18 ways in which two mishaps may occur in a single month while the third mishap occurs in some other time period. For $[3, 0, 0]$ we obtain

$$C(3, 1) \times 3!/3!$$

or

$$3 \times 6/6 = 3.$$

For $[1, 1, 1]$ we obtain

$$C(3, 3) \times 3!$$

or

$$1 \times 6 = 6.$$

This gives the total of twenty-seven as we expected. Thus we have the following probabilities:

$[1, 1, 1]$	$6/27 = 22\%$
$[2, 1, 0]$	$18/27 = 67\%$
$[3, 0, 0]$	$3/27 = 11\%$

A slightly larger sample space would be that of four events in four months. In this case there would be $4 \times 4 \times 4 \times 4$ or 256 possible ways for all the mishaps to occur. This sample space would be a bit tedious to write out. Instead we use the smaller sample space $\{[1, 1, 1, 1], [2, 1, 1, 0], [2, 2, 0, 0], [3, 1, 0, 0], [4, 0, 0, 0]\}$. Then we compute the number of ways in which each of these can occur and divide by $4 \times 4 \times 4 \times 4$ (or 256) as follows:

"class"	"number of events"	"probability"
[1, 1, 1, 1]	$C(4, 4) \times 4!$.0937
[2, 1, 1, 0]	$C(4, 1) \times C(3, 2) \times 4!/2!$.5625
[2, 2, 0, 0]	$C(4, 2) \times 4!/2!/2!$.1406
[3, 1, 0, 0]	$C(4, 1) \times C(3, 1) \times 4!/3!$.1875
[4, 0, 0, 0]	$C(4, 1) \times 4!/4!$.0156

These classes and their respective "number of events" are obtained as before. For example, for [3, 1, 0, 0] we must first choose one period out of four to hold three mishaps and then we must choose one period out of the remaining three periods to hold the remaining mishap. This is done in $C(4, 1) \times C(3, 1)$ ways. Then we multiply by $4!$ (the number of possible arrangements) and divide by $3!$ since 3 of the periods are the same. For [2, 2, 0, 0] we choose two periods out of four to hold two mishaps each. This is done in $C(4, 2)$ ways. Then we multiply by $4!$ and divide by $2!$ twice (since we have two pairs of periods which have two mishaps each).

As an application of this case we might compute some "Murphy" probabilities as follows: Suppose four friends are walking along a street after a heavy rain. There are many puddles because of the rain, and for some reason the group is walking with the flow of traffic instead of against it. Also suppose the wind is blowing hard. Thus the group can neither hear traffic coming its way nor see it. Now suppose four cars come along and hit four puddles and splash one person each time. Now we ask ourselves, "What is the most likely way for these splashes to occur?" Looking back up at the case [2, 1, 1, 0] we see a probability of 56%. Therefore, we conclude that the chances are better than 50-50 that one person will be splashed twice while one person is not splashed at all. The next most likely occurrence is that of [3, 1, 0, 0], which says one person is splashed three times and two are not splashed at all. For the person who is splashed two or maybe even three times, how is he or she supposed to react? Well, most of us would probably blame not only the cars and the rain but also our friends. We could blame them because we had to walk in front or because we had to walk in back or because we had to walk wherever it was that we walked. However, the proper response would be to simply say, "Well, this was a cluster day for me" and then hope for sunshine tomorrow.

As another application of this last case, suppose that during the last four months four mishaps occurred which upset us very much. According to what we have just seen the chances are better than 50% that two of these occurred during the same month with the next most likely case being that of three mishaps in the same month!

For our original example of 24 and 24, even our sample of classes would consist of several hundred elements which would be quite tedious to calculate by hand. We, therefore, must eventually turn to a computer for those events and their probabilities.

Let's consider just two further examples. For the case of twenty-four mishaps in only twelve periods, the most likely event is that of two periods with four mishaps, two periods with three mishaps, three periods with two mishaps, and four periods with one mishap each. This leaves one period with no mishap. The second most likely event is almost the same but has one of the periods which had only one mishap giving it up to one of the periods which already had four. This now leaves two periods with no mishap even though the average number of mishaps per period is two! Also interesting is the probability that this occurs 47% of the time. That is, 47% of the time some period will have five mishaps, which is more than twice the number which one might expect. For twelve mishaps in twenty-four periods the most likely event is two periods with two mishaps each, eight periods with one each, and the remaining fourteen with no mishaps. For this event there is a probability of 29%. For the event that three periods have two mishaps each while six have one each the probability is 18%. Thus these two events account for almost half of all possible events.

The clustering effect can be used to explain, to some extent at least, the roulette wheel. When betting on a wheel with 36 numbers (not counting zero or double zero!) we should anticipate the

wheel stopping on certain numbers more frequently than others. In fact on a given day, there may be some numbers which are never hit. This will be true day after day. However, if the wheel is fair, the same number shouldn't be among those not hit day after day. This is one lesson the author learned by experience rather than by the less expensive way using probability theory.

In many cases such theories can be tested very easily. To test this one, take two dice, one red and one blue, and make a table as follows:

		blue die					
		1	2	3	4	5	6
red die	1	"	'	"	'	"	'
	2	"	'	"	'	"	'
	3	'	"	'	"	'	"
	4	"	'	"	'	"	'
	5	'	"	'	"	'	"
	6	"	'	"	'	"	'

(This table has already been filled in with the results of 36 tosses made by the author). After each toss, the row indicated by the red die and the column indicated by the blue die are used to locate and mark a "box." After an additional 36 tosses there remained four boxes which still were unmarked!

One use for this theory might be to test a "random number generator." For example, when a certain random number generator was tested it produced the class

[3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0].

This class actually tied for the honor of being the "most likely" case for our original problem of 24 mishaps in 24 periods. These data are consistent with the behavior we would expect if the random number generator is a good one. On three additional tests of the same random number generator, similar "highly likely" cases were obtained. This may, in fact, have some application to "queueing theory" since we may now check the randomness of our data by checking its "likeliness." This would be accomplished by comparing data produced by the random number generator against the probability of the data's "Murphy" class.

Conclusion

Exactly what does all this mean? The main thing it does is to provide an explanation for why certain events occur, so that in some cases at least, preparations can be made to avoid greater frustration or disaster. Some examples are:

- (1) Insurance companies should anticipate major claims to occur in bunches and plan their financial reserves accordingly.
- (2) Government officials should anticipate crises in groups and should have crisis-contingency teams that can coordinate and cooperate with one another.
- (3) College students should expect exam dates to coincide and should, therefore, either plan to be sick, or take only one course per semester, or study ahead.
- (4) Obstetric wards should have enough delivery rooms available to handle three or four times the average daily number of births.

Another possible use might be to explain the apparent variations in density that exist within the universe [2]. While stars cluster into galaxies and galaxies into supergalaxies, density, even if very large portions of space are measured, is not seen to be uniform. The reason that some astronomers might expect it to be uniform lies partially in their assumptions about the Big Bang theory. This theory states that during one very large explosion matter was sent flying through

space in all possible directions. Then the assumption is made that all directions are equally likely. All of this is fine until we then conclude that since all directions are equally likely, density in all directions should be the same. We now know this should, indeed, not be the case. Accordingly, we see that rules of probability or Murphy's Law, if you will, may have played a major role in the makeup of our universe as well as the makeup of our daily lives.

Other events that cluster together are major fires, train wrecks, plane crashes (indeed, the summer of 1985 was a good example!), phone calls from friends we haven't heard from in ages, oil spills, and even major scientific discoveries or breakthroughs. Thus a good argument could be made, as was suggested in [3], that our lives are completely controlled by probability—or to put it another way, “Murphy's Law is in fact a law of nature which, like those of physics discovered by Newton 300 years ago, cannot be broken.”

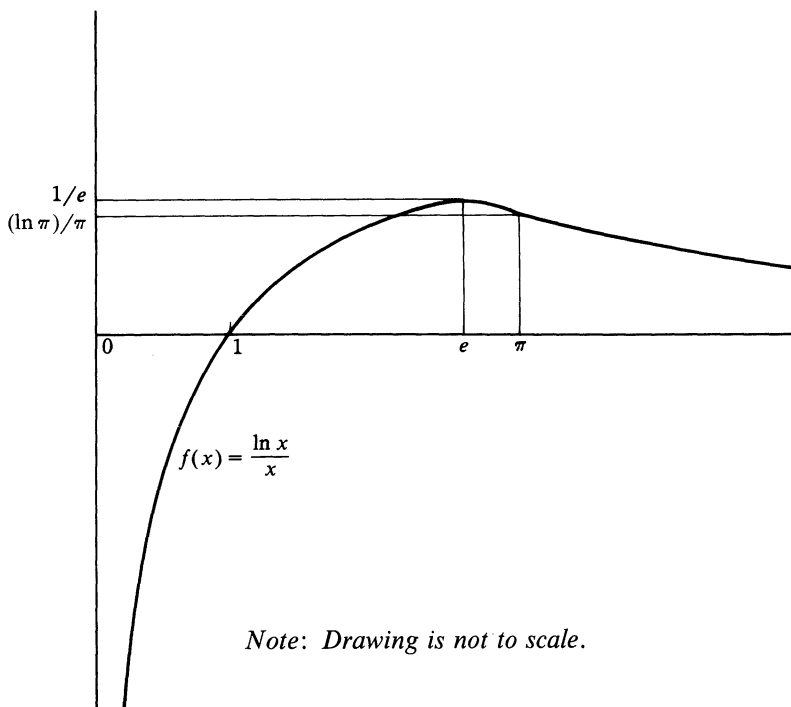
Note: Processes that exhibit this clustering behavior have been greatly studied. Such processes are called Poisson processes and are based on a special probability distribution called the Poisson distribution (see [4] or [5] or almost any text on probability).

References

- [1] M. Ecker, Betting on birthdays, Popular Computing, (July 1984) 189–191.
- [2] A new geometry of nature, Discover, (June 1982) 66–68.
- [3] James S. Trefil, So you think you can break the law of averages, Smithsonian, (September 1984) 66–75.
- [4] Bernard Harris, Theory of Probability, Addison-Wesley, 1966.
- [5] William Feller, An Introduction to Probability Theory and Its Applications, John Wiley & Sons, Inc., 1957.

Proof without Words:

$$\pi^e < e^\pi$$



—FOUAD NAKHLI
American University of Beirut

space in all possible directions. Then the assumption is made that all directions are equally likely. All of this is fine until we then conclude that since all directions are equally likely, density in all directions should be the same. We now know this should, indeed, not be the case. Accordingly, we see that rules of probability or Murphy's Law, if you will, may have played a major role in the makeup of our universe as well as the makeup of our daily lives.

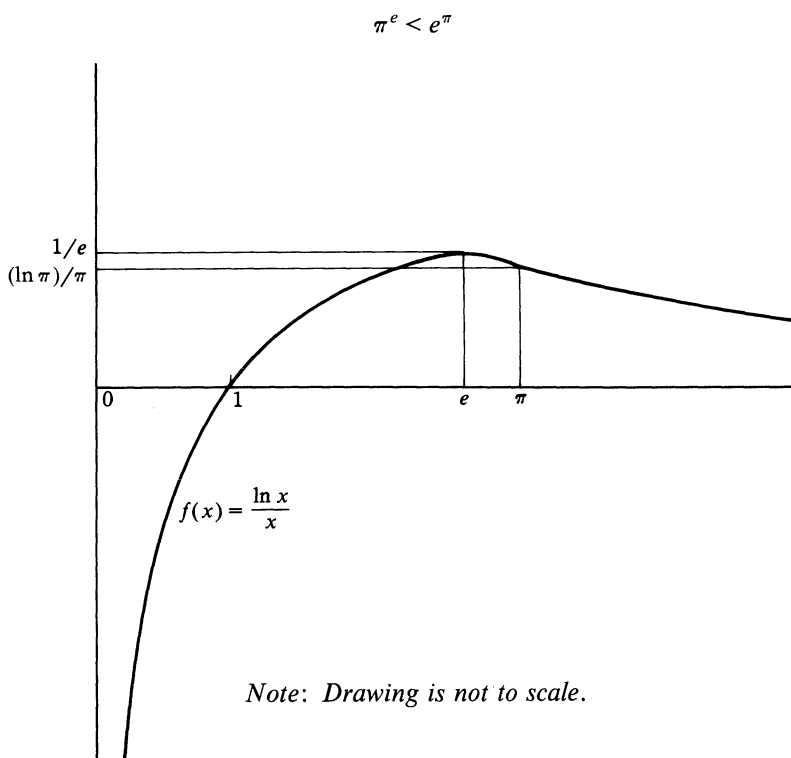
Other events that cluster together are major fires, train wrecks, plane crashes (indeed, the summer of 1985 was a good example!), phone calls from friends we haven't heard from in ages, oil spills, and even major scientific discoveries or breakthroughs. Thus a good argument could be made, as was suggested in [3], that our lives are completely controlled by probability—or to put it another way, "Murphy's Law is in fact a law of nature which, like those of physics discovered by Newton 300 years ago, cannot be broken."

Note: Processes that exhibit this clustering behavior have been greatly studied. Such processes are called Poisson processes and are based on a special probability distribution called the Poisson distribution (see [4] or [5] or almost any text on probability).

References

- [1] M. Ecker, Betting on birthdays, Popular Computing, (July 1984) 189–191.
- [2] A new geometry of nature, Discover, (June 1982) 66–68.
- [3] James S. Trefil, So you think you can break the law of averages, Smithsonian, (September 1984) 66–75.
- [4] Bernard Harris, Theory of Probability, Addison-Wesley, 1966.
- [5] William Feller, An Introduction to Probability Theory and Its Applications, John Wiley & Sons, Inc., 1957.

Proof without Words:



—FOUAD NAKHLI
American University of Beirut

The Fermat Last Theorem — A Brief, Elementary, Rigorous Proof

B.L. SCHWARTZ

216 Apple Blossom Ct.

Vienna, VA 22180

Henri Fermat (1588?–1631) was a seventeenth-century French shoemaker and amateur mathematician, no relation to his noble and better-known namesake, Pierre [8]. Among Parisian shoemakers of that period, it was well known that one could obtain a shape for the left shoe of a pair if one had the form (called in English a ‘last’) for the right foot. A mirror image of the right-foot last produced one for the left foot [2].

Some customers preferred not to provide the right-foot outline from which a last could be formed. A local superstition among a circle of Parisians held it to be bad luck to allow an image to be made of any part of the body on the right side [9], [10]. They, therefore, offered the left foot as a model when ordering shoes. This created confusion for most shoemakers, who were used to having a right-side form and using it to create the left-side one.

Only Henri appeared to realize that the mirror-image process could be used from left to right in such cases, rather than right to left. He never formulated this in exact form, nor offered proof. If he had, he would probably have encountered difficulty in getting it published, lacking the connections of his more prominent namesake. Nevertheless, the claim became popularly known among French artisans as the Fermat Last Theorem.

For unknown reasons, interest in the topic lagged for 3 1/2 centuries after his death. However, the present author has undertaken to give an exact formulation in modern notation, and a formal proof.

Let Γ be a rectifiable simple plane closed curve, defined parametrically by the real functions $f(t)$, $g(t)$ ($0 \leq t \leq 1$).

$$\Gamma = [(x, y) | x = f(t), y = g(t)]$$

where $f(0) = f(1)$, $g(0) = g(1)$.

We consider Γ to be the outline of a right foot. Define Γ' as the mirror image of Γ in the y -axis: formally, $\Gamma' = [(-x, y) | (x, y) \in \Gamma]$.

We consider Γ' to be the outline of a matching left foot. A pair (Γ_1, Γ_2) is called a Marcos* set if Γ_2 is the mirror image of Γ_1 .

Suppose now we are given a simple closed curve Γ , as the left-foot image of a client. The Fermat construction was to create the mirror image Γ' as the right foot. A Marcos set from that right-foot image would be (Γ', Γ'') . A pair of shoes from these lasts will fit the client if $\Gamma'' = \Gamma$. The Fermat Last Theorem is, therefore, formalized as $\Gamma'' = \Gamma$.

The formal proof follows. Let $(x, y) \in \Gamma$.

$(-x, y) \in \Gamma'$ by definition of the mirror image transformation, and similarly

$$(-(-x), y) \in \Gamma''.$$

It remains to prove that

$$-(-x) = x$$

for any real x . But this is a known result [4], [5].

The converse, $(x, y) \in \Gamma'' \rightarrow (x, y) \in \Gamma$ is proved similarly, and the Fermat Last Theorem

* Imelda Marcos was reported to be a Spanish immigrant in Paris at the period and mistress to Henri. She is said to have loved shoes [7].

follows.

We trust this unexpectedly brief demonstration will serve to close this chapter in mathematical history for those who have had an interest in this topic.

References

- [1] W. W. R. Ball, *Récréations Mathématiques et Problèmes des temps anciens et modernes*, Paris, 1909.
- [2] J. L. Coolidge, *A History of Geometrical Methods*, Oxford, 1940.
- [3] E. Fourrey, *Curosités Géométriques*, Vuibert et Nony, Paris, 1907.
- [4] Benjamin Greenleaf, A. M., *The Complete Arithmetic*, Norwood Press, Boston, 1881.
- [5] Felix Klein, *Elementary Mathematics from an Advanced Standpoint*, translated from the 1908 German edition by E. R. Hedrick and C. A. Noble, Dover Publications, NY (undated) p. 24.
- [6] J. E. Montucla, *Histoire des Mathématiques*, Paris, 1758.
- [7] Claude Mydorge, *Examen du livres des récréations mathématique et de ses problèmes*, Paris, 1630.
- [8] D. J. Struik, *A Concise History of Mathematics*, 3rd Ed., Dover Publications, NY, 1991.
- [9] The Holy Bible (various authors), Psalms 137:5, Harper and Row, NY, 1933.
- [10] ———, Matthew 5:29, 30.

Reflections on the Ellipse

WILLIAM C. SCHULZ

CHARLES G. MOORE

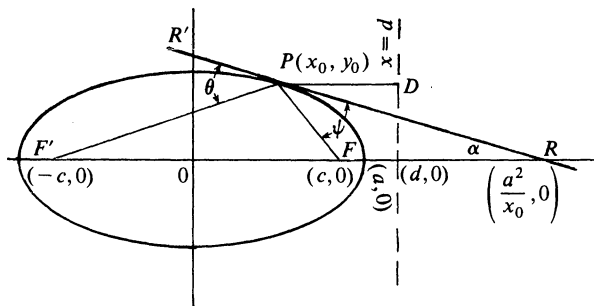
Northern Arizona University

Flagstaff, AZ 86001

The reflection property is an important and striking property of the ellipse. However, calculus-related proofs of the property are often tedious, so that a short and elegant proof using the derivative is desirable. Our proof is based on the surprisingly simple and quite useful formula:

$$FP = a - ex \quad (1)$$

for the distance from F , the focus on the positive x -axis, to a point $P(x, y)$ on the ellipse. See FIGURE.



follows.

We trust this unexpectedly brief demonstration will serve to close this chapter in mathematical history for those who have had an interest in this topic.

References

- [1] W. W. R. Ball, *Récréations Mathématiques et Problèmes des temps anciens et modernes*, Paris, 1909.
- [2] J. L. Coolidge, *A History of Geometrical Methods*, Oxford, 1940.
- [3] E. Fourrey, *Curosités Géométriques*, Vuibert et Nony, Paris, 1907.
- [4] Benjamin Greenleaf, A. M., *The Complete Arithmetic*, Norwood Press, Boston, 1881.
- [5] Felix Klein, *Elementary Mathematics from an Advanced Standpoint*, translated from the 1908 German edition by E. R. Hedrick and C. A. Noble, Dover Publications, NY (undated) p. 24.
- [6] J. E. Montucla, *Histoire des Mathématiques*, Paris, 1758.
- [7] Claude Mydorge, *Examen du livres des récréations mathématique et de ses problèmes*, Paris, 1630.
- [8] D. J. Struik, *A Concise History of Mathematics*, 3rd Ed., Dover Publications, NY, 1991.
- [9] The Holy Bible (various authors), Psalms 137:5, Harper and Row, NY, 1933.
- [10] ———, Matthew 5:29, 30.

Reflections on the Ellipse

WILLIAM C. SCHULZ

CHARLES G. MOORE

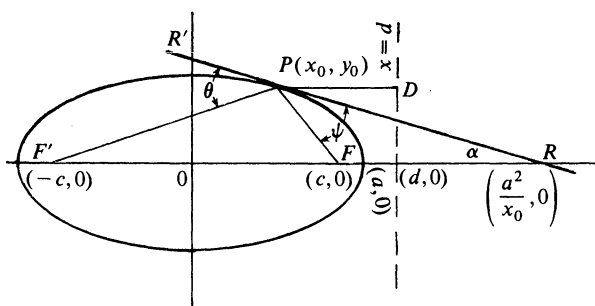
Northern Arizona University

Flagstaff, AZ 86001

The reflection property is an important and striking property of the ellipse. However, calculus-related proofs of the property are often tedious, so that a short and elegant proof using the derivative is desirable. Our proof is based on the surprisingly simple and quite useful formula:

$$FP = a - ex \quad (1)$$

for the distance from F , the focus on the positive x -axis, to a point $P(x, y)$ on the ellipse. See FIGURE.



Relation (1) is easily proved from the well-known relations

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad b^2 + c^2 = a^2, \quad \frac{c}{a} = e.$$

For we have

$$\begin{aligned} (FP)^2 &= (x - c)^2 + y^2 \\ &= x^2 - 2cx + c^2 + b^2 \left(1 - \frac{x^2}{a^2}\right) \\ &= \left(1 - \frac{b^2}{a^2}\right)x^2 - 2cx + c^2 + b^2 \\ &= \frac{c^2}{a^2}x^2 - 2cx + a^2 \\ &= e^2x^2 - 2aex + a^2 \\ &= (a - ex)^2. \end{aligned}$$

Observing that $x \leq a$ and $e < 1$ for an ellipse, we thus have equation (1). By changing the sign of c , we show in exactly the same way that

$$F'P = a + ex. \quad (2)$$

We may also derive (2) from (1) by noting that $FP + F'P = 2a$.

We note that, given the focus-directrix property of the ellipse, $FP/PD = e$, relation (1) is more easily proved as follows: we know, since $(a - c)/(d - a) = e$, that $d = a/e$. Thus $FP = ePD = e(d - x) = a - ex$, and, similarly, $F'P = a + ex$. Conversely, from the relation (1) we may derive the focus-directrix property. Indeed, setting $d = a/e$, $FP = a - ex = ed - ex = ePD$.

Considering (x_0, y_0) to be a fixed point on the ellipse and (x, y) to be a variable point on the tangent line, we now turn to the reflection property: $F'PR' = FPR$. Denoting the coordinates of P by (x_0, y_0) , the tangent line is given by

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1,$$

from which the coordinates of the point R are seen to be $(a^2/x_0, 0)$. We now calculate, using (1) and (2),

$$FR = \frac{a^2}{x_0} - c = \frac{1}{x_0}(a^2 - aex_0) = \frac{a}{x_0}(a - ex_0) = \frac{a}{x_0}FP, \quad (3)$$

$$F'R = \frac{a^2}{x_0} + c = \frac{1}{x_0}(a^2 + aex_0) = \frac{a}{x_0}(a + ex_0) = \frac{a}{x_0}F'P. \quad (4)$$

By the law of sines for the triangle FRP , we have, using (3),

$$\sin \psi = \frac{FR \sin \alpha}{FP} = \frac{a}{x_0} \frac{FP \sin \alpha}{FP} = \frac{a}{x_0} \sin \alpha,$$

and, similarly, for triangle $F'RP$ and using (4),

$$\sin \theta = \sin(\pi - \theta) = \frac{F'R \sin \alpha}{F'P} = \frac{a}{x_0} \frac{F'P \sin \alpha}{F'P} = \frac{a}{x_0} \sin \alpha.$$

Thus, $\sin \theta = \sin \psi$, and, noting that $0 \leq \theta + \psi < \pi$, we have $\theta = \psi$, as desired.

Superexponentiation

NICK BROMER

Bryn Mawr College

Bryn Mawr, PA 19010

The operations of addition, multiplication, and exponentiation are the first three of an infinite sequence of operations. In view of their importance in mathematics and physics, I will call them *basic* operations.

Generalizing from the first three, we see that each basic operation after the first is derived from the previous one according to the following definition:

Let $*$ be a basic operation which is defined on some number x , which I will call the *operand*. We can define a new, higher-order operation $**$ by

$$x ** n = x * x * x * \cdots * x,$$

where the number of x 's appearing on the right is a positive integer n , which I will call the *exponent*.

For example, x multiplied by 4 is equal to $x + x + x + x$. We can denote this, using the definition's symbolism, by $4 + + x$.

I will indicate the next higher order of operation after $**$ by $***$. Thus, 1.4 cubed can be denoted by $1.4 + + + 3$.

The arrow operation

Using the above definition, let us investigate superexponentiation, the next basic operation after raising to powers. Since "superexponentiation" is such a mouthful, I will call it the arrow operation, and denote it by the vertical arrow symbol first used by Donald E. Knuth [1].

As soon as we try to define the arrow operation, a problem arises. The sequence of operations bifurcates: two arrow operations are possible. The reason for this is that exponentiation is not commutative.

To make this clear, let us first examine the previous level. When we create exponentiation from multiplication, the association of the operands makes no difference, since multiplication is commutative. Thus

$$(x(x(x))) = (((x)x)x),$$

and so x^n has only one definition.

In the case of repeated exponentiation, parentheses make a difference:

$$x^{(x^x)} \quad \text{and} \quad (x^x)^x$$

have different values, in general.

I will call the form on the left above, in which operands are added onto the left side of the parentheses, the left mode of superexponentiation, and I will denote it with the symbol \uparrow . For example,

$$3 \uparrow 4 = 3^{[3^{(3^3)}]}.$$

Similarly, I define the right mode by the symbol \downarrow . Thus

$$3 \downarrow 4 = [(3^3)^3]^3.$$

Operations of orders higher than arrow may be indicated by conjunctions of arrows in the manner of the definition $(4 \downarrow \downarrow 3 = (4 \downarrow 4) \downarrow 4$, for example) with the added stipulation that the arrows are executed from left to right (that is, $(4 \uparrow 4) \uparrow 4 = 4 \uparrow \downarrow 3$, and $4 \downarrow (4 \downarrow 4) = 4 \downarrow \uparrow 3$). This convention is arbitrary—the arrows could as well have been read right to left.

As long as we restrict ourselves to positive integral operands, all these operations will be defined.

The numbers that result from these higher-order basic operations quickly become huge. For example, $10 \uparrow 4$, which also goes under the name “googolplex,” is a number so big that it is physically impractical to write down using exponential notation. (It seems that each order of basic operation allows one to write down numbers that are inconveniently large in the notation of the next-lowest level.) The arrow notation thus opens new realms of magnitude [1], [2], [3].

We can note that for any basic operation $*$, $2 * 2 = 4$. This assertion is readily proved: since $x * x = x ** 2$, $2 * 2 = 2 ** 2$. The statement follows by induction from $2 + 2 = 4$. Also note that for $*$ other than $+$, $x * 1 = x$ (by definition).

Arrow functions

Since $x \uparrow n$ and $x \downarrow n$ are defined for x positive real and n positive integral, we have a family of continuous functions. Graphs of the basic functions $f(x) = x \uparrow n$ and $f(x) = x \downarrow n$ for $n = 2, 3, 4, 5, 6$, and 7 are shown in FIGURES 1 and 2.

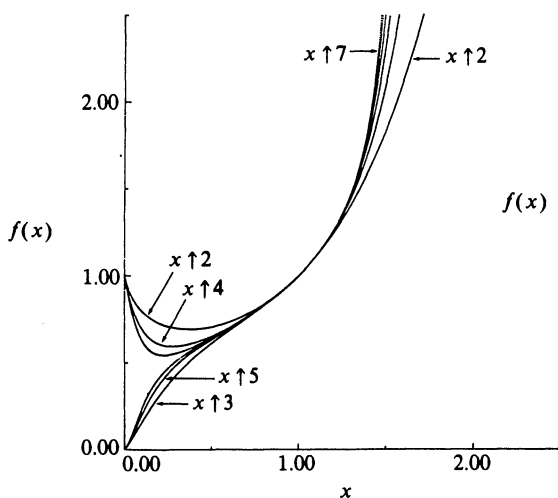


FIGURE 1

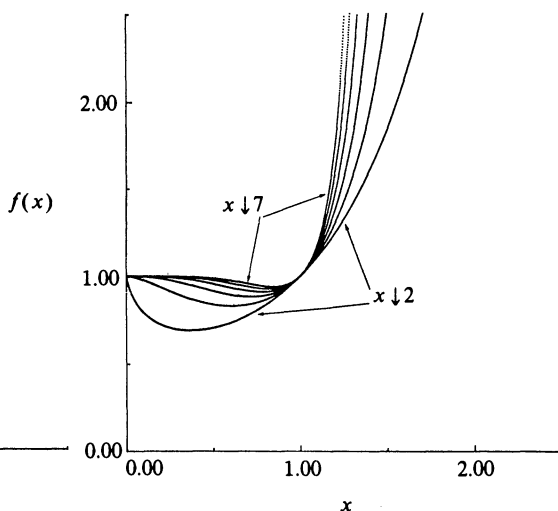


FIGURE 2

Only positive x -values are shown because x^x (that is, $x \uparrow 2$ or $x \downarrow 2$) is pathological in the negative x region (and, hence, so are the other functions). It is undefined for irrational x , and jumps about among imaginary, positive, and negative values for rational x less than zero.

The “wagging-tail” function $x \uparrow n$ has very interesting behavior as n goes to infinity. Its value quickly goes to infinity for x greater than e to the $1/e$ power ($1.44\dots$) and is less than e for x less than e to the $1/e$. For x less than $1/e$ to the e power ($0.0659\dots$) it bifurcates into two values which alternate as n is incremented. The value of the function at the bifurcation point is $1/e$. (All of these values were found empirically. Readers may be able to derive them.) To the best of my knowledge, the $x \uparrow n$ function was first investigated in the early 1970’s by A. Guyton (private communication).

The bifurcation is shown in FIGURE 3, in which a plot of $f(x) = x \uparrow 1000$ is superimposed on a plot of $g(x) = x \uparrow 1001$. (The two functions are virtually identical to the right of the conjunction point, so that only one line appears there.) The x -scale is logarithmic.

Arrow functions with negative exponents

Single-arrow functions are defined above for positive real operands and whole number exponents. The next step is to try negative integers for exponents. We can do this by figuring out

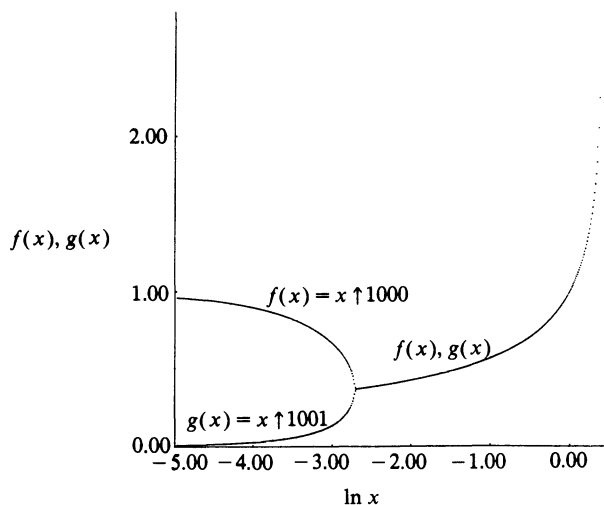


FIGURE 3

a process that will reduce the exponent of an arrow function by one: repeating the process, we will get negative exponents.

To change $x \uparrow n$ to $x \uparrow (n-1)$, we can take the logarithm to the base x of $x \uparrow n$, which gives $x \uparrow (n-1) \log x = x \uparrow (n-1)$. If we do this n times to the function $x \uparrow n$, the result is 1, which we can define as $x \uparrow 0$. Repeating the procedure, we get $x \uparrow (-1) = \log 1 = 0$, and $x \uparrow (-2) = \log 0 = \text{minus infinity}$. It would seem that there is no $x \uparrow n$ for n less than -2 .

With the \downarrow operation we have more success. The reducing operation here is raising to the $(1/x)$ power: $(x \downarrow n)$ to the $(1/x)$ power is $x \downarrow (n-1)$. This can be applied any number of times if x is positive real.

Some of the negative \downarrow functions are shown in FIGURE 4.

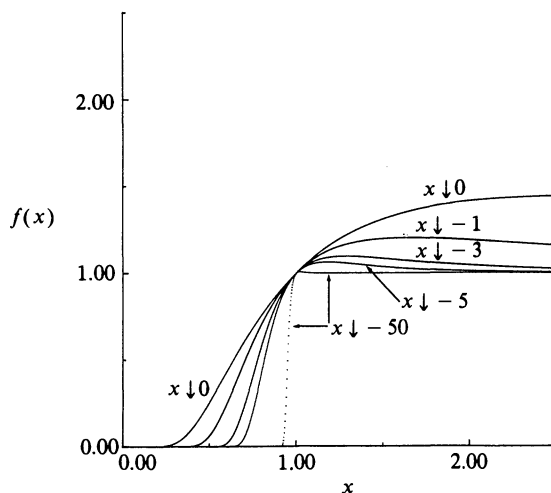


FIGURE 4

Inverse operations

Once an operation is generated, we can define an inverse operation which “undoes” the work of the operation. I’ll say that $\hat{*}$ is the inverse of $*$ if

$$(x \hat{*} n) * n = x.$$

The number $y = x \hat{*} n$, where n is a positive integer, I will call “the n th asterisk root of x ” (generalizing the term used for exponentiation’s inverse).

For example, -5 is the 8th additive root of 3; $1/5$ is the 15th multiplicative root of 3; $1.414\dots$ is the 2nd exponential root of 2; 3 is the 3rd superexponential (left mode) of 19683; and $1.559\dots$ is the 2nd superexponential root of 2.

The last-mentioned number is transcendental, by the following argument. If $x^x = 2$, then $x(\ln x) = \ln 2$ and so $x = (\ln 2)/(\ln x)$. Now, the Gelfond-Schneider Theorem [4] says that if $y = (\ln a)/(\ln b)$, where a and b are algebraic, then y is either transcendental or rational. So x must be either rational or transcendental. Suppose, for contradiction, that x is rational. Let it equal p/q , where p and q are relatively prime (the fraction is in lowest terms). Then $2 = (p/q)^{(p/q)}$ and so

$$p^p = 2^q q^p.$$

By the relative primeness of p and q , there are three cases: (1) both p and q are odd, which leads to a contradiction in the above equation (implying one side odd and the other even); (2) q is even and p is odd, which is similarly a contradiction; and (3) p is even and q odd. In this last case, there are integers m and r such that $p = 2^m r$ and r is odd. Equating exponents of 2, we get $mp = q$ or $m = q/p$. But this contradicts the assumption that p and q are relatively prime, since the ratio of relative primes cannot equal an integer. Each case leads to a contradiction. Therefore x is not rational, and so x is transcendental.

As we ascend the orders of basic operations, it seems that every new basic operation’s inverse generates roots which are either a new class of reals, or else a new type of number. Subtraction generates negative numbers, division creates fractions, and exponential roots give us algebraic numbers. Exponential roots of negatives create imaginaries.

The preceding two paragraphs lead me to this conjecture: that the arrow operations’ inverse operations generate a countable infinity of real superroots, which are, in general, transcendental. I propose to call these superalgebraic numbers.

There are also superroots which are apparently undefined. The equation $x \downarrow 2 = 1/2$, for instance, has no solution in the reals, and seems not to have a complex solution. (I have had no luck with the problematic question of complex solution, but readers may be able to find one, or prove impossibility.) Thus one might say: there is no square superroot of one-half. Historically, this sort of situation has always resulted in the creation of new types of numbers, whose names reflect the idea that they don’t make sense (“irrational”), don’t exist (“imaginary”), or are simply unpleasant (“negative”). I hope that if anyone ever discovers a consistent way of defining the unreal superroots, he will name them in this tradition.

Note that numbers less than 1 always have two right-mode superroots, and have two left-mode superroots if the root is even.

Fractional exponents and \downarrow

Let’s try to get a consistent way to define $x \downarrow (p/q)$, where p and q are positive integers. This would give, in the limit, a definition of $x \downarrow x$ for real x , which would immediately give us double-arrow continuous functions $f(x) = x \downarrow \downarrow n$ or $f(x) = x \downarrow \uparrow n$, n a whole number.

Here it is important to proceed by analogy to the established basic operations, lest we be led astray (as I was) by the following formula. It can be shown that all of the positive and negative integer exponent \downarrow functions are generated by

$$x \downarrow p = x^{[x^{(p-1)}]}.$$

This formula gives us values for any fraction p , not only integers. Are these values acceptable for defining $x \downarrow \downarrow n$? I now feel they are not, because this definition is not analogous to the way the lower-order operations define fractional exponents.

Let's look at how fractional exponents are handled in multiplication and exponentiation. When multiplying a number by a fraction p/q , we multiply by p and then take the q th multiplicative root (i.e., divide by q). Similarly, to raise a number to a fractional exponential power p/q , we raise the number to the p th power, and then take the q th exponential root.

To continue the process with the arrow operation, we interpret $y = x \downarrow (p/q)$ as $y = (x \downarrow p) \downarrow q$ —and we run into trouble. Here's why: if we raise a number to the superpower $2/3$, it is not the same as raising it to the superpower $4/6$; the two exponents yield different values. (These different values have nothing to do with the fact that there are double roots of numbers less than 1.) This means that the function $y = x \downarrow x$ is neither continuous nor single-valued, which in turn means that the next level of basic continuous functions is undefined.

Similar problems arise with \uparrow , with the added complications associated with the “wagging-tail” behavior of $f(x) = x \uparrow n$ for $0 < x < 1$.

The values of $x \downarrow (np/nq)$ seem to converge toward a limit as n goes to infinity (which is not equal to the value given by the formula above for $x \downarrow p$). This raises the interesting possibility that, using a limiting process, fractional exponents of the arrow functions can be defined in a way analogous to lower-level basic operations. To explain: if we pick a number m as the denominator of the exponential fraction, we will get a series of points. As m goes to infinity, the points become a (we hope) smooth curve. This method is difficult to investigate, because the arrow function causes overflow in a computer when only modestly large operands and exponents are inserted.

If the problems of the above program were overcome, perhaps the double-arrow functions could be graphed. Because $(x \downarrow p) \downarrow q \neq (x \downarrow q) \downarrow p$ in general, there would probably be two classes of such functions.

Conclusions

A hierarchy of operations on the positive integers can be denoted using the arrow symbol. These operations may be of interest to number theorists, as perfect squares, primes, and so on can be generalized into higher orders of basic operations.

The family of continuous arrow functions is analogous to linear functions at the multiplicative level and to quadratics, cubics, and so on at the exponential level. The arrow functions may find an application in the sciences. Nature uses the first three basic functions profusely: why should she not use arrow?

A new set of numbers, the real superroots, has been conjectured to be transcendental. If a continuous double-arrow function can be developed, another set should result.

It is not clear whether or not the sequence of families of basic continuous functions can be developed beyond the single arrow level. It is clear that more symmetries are lost, and more complications arise, with each new basic operation. While it is unlikely that we will be able to penetrate this mathematical thicket very far, the attempt should be interesting.

Acknowledgements. I wish to thank Neal Abraham, Teymour Darkosh, John Lavelle, and Rodica Simion for helpful conversations.

References

- [1] Donald E. Knuth, Mathematics and computer science: coping with finiteness, *Science*, 194 (December 1976) 1235–1242. Knuth's notation differs from that used in this paper in using only \uparrow , not \downarrow , and in using one more arrow (Knuth denotes exponentiation in the manner of the Basic computer language, by \uparrow , and left-mode superexponentiation by $\uparrow\uparrow$).
- [2] C. Smorynski, Some rapidly growing functions, *The Mathematical Intelligencer*, 2 (1980) 149–154.

- [3] Martin Gardner, *Mathematical Games*, Scientific American (November 1977) 28. Gardner discusses Knuth's notation.
- [4] A. O. Gelfond, *Doklady Akad. Nauk S.S.S.R.*, 2 (1934) 1–6; and Th. Schneider, *J. reine angew. Math.*, 172 (1935) 65–69. The Gelfond-Schneider Theorem is one of Hilbert's 23 problems. See also Ivan Niven, *Irrational Numbers*, Wiley, chap. 10, or C. L. Siegel and R. Bellman, *Transcendental Numbers*, Princeton, pp. 80–83.

Non-Associative Operations

N. J. LORD

Tonbridge School

Kent TN9 1JP, England

In many algebraic structures the associativity of an operation is postulated as an axiom. However, there are important non-associative operations (e.g., subtraction, division, vector multiplication) and the purpose of this article is to discuss the extent to which an operation is non-associative.

An operation $*$ defined on a set S is associative only if both ways of inserting parentheses in the product $a_1 * a_2 * a_3$, namely, $(a_1 * a_2) * a_3$ and $a_1 * (a_2 * a_3)$, give the same result for all a_1 , a_2 , and a_3 in S . For some non-associative operations on some sets the non-associativity is incomplete; that is, for products of four or more factors, two different ways of inserting parentheses ('bracketing') give equal results. For example, with the operation of subtraction on the set of real numbers, it is true that $a_1 - (a_2 - (a_3 - a_4)) = (a_1 - (a_2 - a_3)) - a_4$ for any real numbers a_1 , a_2 , a_3 , and a_4 .

It thus makes sense to say that subtraction of real numbers has "limited" non-associativity. Since one cannot find two ways of bracketing $a_1 - a_2 - a_3$ that lead to the same answer for all a_1 , a_2 , and a_3 , but one can find two bracketings of $a_1 - a_2 - a_3 - a_4$ that always lead to the same result, this leads to a characterization of the non-associativity of subtraction of real numbers as "having depth 4". More generally, the *depth* of non-associativity, $d(*)$, for a non-associative operation on a set S may be defined as:

$$d(*) = \begin{cases} \min\{n > 3: \text{there exist two bracketings of } a_1 * \cdots * a_n \text{ that give the same} \\ \text{result for every } a_1, \dots, a_n \text{ in } S\} \\ \infty \quad \text{if, for each } n \geq 3, \text{ there exist elements } a_1, \dots, a_n \text{ in } S \text{ for which all} \\ \text{bracketings of } a_1 * \cdots * a_n \text{ give different results.} \end{cases}$$

In the latter case, we will say that $*$ has *unlimited non-associativity*, abbreviated UNA.

For convenience, we shall denote by $N(n)$ the number of ways of inserting parentheses to define unambiguously $a_1 * \cdots * a_n$. So, for example, $N(4) = 5$ corresponding to the bracketings

$$(a_1 * (a_2 * a_3)) * a_4, ((a_1 * a_2) * a_3) * a_4, a_1 * ((a_2 * a_3) * a_4), \\ a_1 * (a_2 * (a_3 * a_4)), (a_1 * a_2) * (a_3 * a_4).$$

(For more information on $N(n)$, in particular a derivation of the formula

$$N(n) = \frac{(2n-2)!}{n!(n-1)!},$$

see [1] or [2].)

- [3] Martin Gardner, *Mathematical Games*, Scientific American (November 1977) 28. Gardner discusses Knuth's notation.
- [4] A. O. Gelfond, *Doklady Akad. Nauk S.S.S.R.*, 2 (1934) 1–6; and Th. Schneider, *J. reine angew. Math.*, 172 (1935) 65–69. The Gelfond-Schneider Theorem is one of Hilbert's 23 problems. See also Ivan Niven, *Irrational Numbers*, Wiley, chap. 10, or C. L. Siegel and R. Bellman, *Transcendental Numbers*, Princeton, pp. 80–83.

Non-Associative Operations

N. J. LORD

Tonbridge School

Kent TN9 1JP, England

In many algebraic structures the associativity of an operation is postulated as an axiom. However, there are important non-associative operations (e.g., subtraction, division, vector multiplication) and the purpose of this article is to discuss the extent to which an operation is non-associative.

An operation $*$ defined on a set S is associative only if both ways of inserting parentheses in the product $a_1 * a_2 * a_3$, namely, $(a_1 * a_2) * a_3$ and $a_1 * (a_2 * a_3)$, give the same result for all a_1 , a_2 , and a_3 in S . For some non-associative operations on some sets the non-associativity is incomplete; that is, for products of four or more factors, two different ways of inserting parentheses ('bracketing') give equal results. For example, with the operation of subtraction on the set of real numbers, it is true that $a_1 - (a_2 - (a_3 - a_4)) = (a_1 - (a_2 - a_3)) - a_4$ for any real numbers a_1 , a_2 , a_3 , and a_4 .

It thus makes sense to say that subtraction of real numbers has "limited" non-associativity. Since one cannot find two ways of bracketing $a_1 - a_2 - a_3$ that lead to the same answer for all a_1 , a_2 , and a_3 , but one can find two bracketings of $a_1 - a_2 - a_3 - a_4$ that always lead to the same result, this leads to a characterization of the non-associativity of subtraction of real numbers as "having depth 4". More generally, the *depth* of non-associativity, $d(*)$, for a non-associative operation on a set S may be defined as:

$$d(*) = \begin{cases} \min\{n > 3: \text{there exist two bracketings of } a_1 * \cdots * a_n \text{ that give the same} \\ \text{result for every } a_1, \dots, a_n \text{ in } S\} \\ \infty & \text{if, for each } n \geq 3, \text{ there exist elements } a_1, \dots, a_n \text{ in } S \text{ for which all} \\ & \text{bracketings of } a_1 * \cdots * a_n \text{ give different results.} \end{cases}$$

In the latter case, we will say that $*$ has *unlimited non-associativity*, abbreviated UNA.

For convenience, we shall denote by $N(n)$ the number of ways of inserting parentheses to define unambiguously $a_1 * \cdots * a_n$. So, for example, $N(4) = 5$ corresponding to the bracketings

$$(a_1 * (a_2 * a_3)) * a_4, ((a_1 * a_2) * a_3) * a_4, a_1 * ((a_2 * a_3) * a_4), \\ a_1 * (a_2 * (a_3 * a_4)), (a_1 * a_2) * (a_3 * a_4).$$

(For more information on $N(n)$, in particular a derivation of the formula

$$N(n) = \frac{(2n-2)!}{n!(n-1)!},$$

see [1] or [2].)

Some examples

In this section the depth of non-associativity of several operations will be investigated.

EXAMPLE 1. $S = \mathbb{R}$, $k \in \mathbb{R}$ fixed, $a * b = a + kb$.

If $k = 0$, then $*$ is associative; $k = 1$ gives addition and $k = -1$ subtraction.

However, every other value of k gives rise to an operation having UNA. Certainly $*$ is then non-associative—because $a_1 * (a_2 * a_3) = a_1 + ka_2 + k^2a_3$ and $(a_1 * a_2) * a_3 = a_1 + ka_2 + ka_3$, which are different numbers if $a_3 \neq 0$. Similarly, for $n > 3$, a bracketing of $a_1 * \cdots * a_n$ multiplies out to give an expression $a_1 + ka_2 + k^{r_3}a_3 + \cdots + k^{r_n}a_n$ for some n -tuple of powers of k : $(0, 1, r_3, \dots, r_n)$. Moreover, distinct bracketings give rise to distinct n -tuples of powers of k . Indeed, given such an n -tuple, the bracketing that it must have come from can be reconstructed by the following algorithm:

- (i) Locate the first entry, r_i say, from the right such that $r_i = 1$. (Such exists because r_2 is always 1.)
 - (ii) Since a_i is the first entry from the right to have been in the right-hand position of the composition only once, there must be a bracketing $(a_1 * \cdots * a_{i-1}) * (a_i * \cdots * a_n)$.
 - (iii) By repeating this process, the bracketings of $a_1 * \cdots * a_{i-1}$ and $a_i * \cdots * a_n$ can be reconstructed from the tuples $(0, 1, \dots, r_{i-1})$ and $(r_i - 1, r_{i+1} - 1, \dots, r_n - 1)$.
- (For example, the 6-tuple $(0, 1, 2, 2, 1, 2)$ unfolds in the following manner:

$$\begin{array}{rcl}
 (0, 1, 2, 2, 1, 2) & & a_1 * a_2 * a_3 * a_4 * a_5 * a_6 \\
 \downarrow \quad \searrow & & \\
 (0, 1, 2, 2), (0, 1) & & (a_1 * a_2 * a_3 * a_4) * (a_5 * a_6) \\
 \downarrow \quad \searrow & & \\
 (0), (0, 1, 1) & & (a_1 * (a_2 * a_3 * a_4)) * (a_5 * a_6) \\
 \downarrow \quad \searrow & & \\
 (0, 1), (0) & & (a_1 * ((a_2 * a_3) * a_4)) * (a_5 * a_6)
 \end{array}$$

By the hypothesis on k ($\neq 0, -1, 1$), the $N(n)$ vectors $\mathbf{v}_i = (1, k, k^{r_3}, \dots, k^{r_n}), 1 \leq i \leq N(n)$, corresponding to each bracketing are distinct, and the results of applying each bracketing to the real numbers a_1, \dots, a_n are given by the scalar products $\mathbf{a} \cdot \mathbf{v}_i, 1 \leq i \leq N(n)$, where \mathbf{a} is the vector (a_1, \dots, a_n) .

The operation $*$ will have UNA if a specific choice of \mathbf{a} can be made such that $\mathbf{a} \cdot \mathbf{v}_i \neq \mathbf{a} \cdot \mathbf{v}_j$, i.e., $\mathbf{a} \cdot (\mathbf{v}_i - \mathbf{v}_j) \neq 0$, for all $1 \leq i, j \leq N(n), i \neq j$. Now, since $\mathbf{v}_i \neq \mathbf{v}_j$ for $i \neq j$, $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \cdot (\mathbf{v}_i - \mathbf{v}_j) = 0\}$ is a hyperplane in \mathbb{R}^n and the union of these hyperplanes for $1 \leq i, j \leq N(n), i \neq j$, cannot constitute the whole of \mathbb{R}^n .

The choice of any \mathbf{a} in \mathbb{R}^n but outside this union then suffices to show that $*$ has UNA.

By taking logarithms, we see that $a * b = ab^k$ (with the same restrictions on k) has UNA on \mathbb{R}^+ .

EXAMPLE 2. If, in Example 1, \mathbb{R} is replaced by \mathbb{C} and k is a primitive m th root of unity, an analogous argument shows that there will exist a_1, \dots, a_n in \mathbb{C} for which all bracketings of $a_1 * \cdots * a_n$ give different results provided that all of the vectors \mathbf{v}_i are different. The maximum disparity in powers of k between corresponding entries in two such vectors is $n - 2$, attained by the bracketing $(a_1 * (a_2 * (a_3 * (\cdots (a_{n-2} * a_{n-1}) \cdots)))) * a_n$ with associated vector $(1, k, k^2, \dots, k^{n-2}, k)$ and the bracketing $a_1 * (a_2 * (a_3 * \cdots (a_{n-1} * a_n) \cdots))$ with associated vector $(1, k, k^2, \dots, k^{n-2}, k^{n-1})$. These n -tuples therefore coincide once $m = n - 2$ and thus $*$ is an operation of depth precisely $m + 2$. (In particular this shows that there are operations of each possible depth.)

EXAMPLE 3. $S = \mathbb{N}$, $a * b = a^b$.

This operation has UNA. Indeed, if p_1, p_2, p_3, \dots denotes the sequence of prime numbers, then all bracketings of $p_1 * \cdots * p_n$ for all values of n all give different results.

Thus, for $n = 3$, the two possible bracketings of $p_1 * p_2 * p_3$ produce

$$p_1^{p_2^{p_3}} \text{ and } p_1^{p_2 p_3}$$

which uniqueness of prime factorization shows to be different numbers. And, in general, uniqueness of prime factorizations will prove the assertion once it has been established that different bracketings give rise to different arrays of exponents of p_1 . Fortunately, we can adapt part of the argument in Example 1 to do this by noticing that if each p_i is assigned its 'level' r_i in the array of exponents of p_1 (so, for example, in $p_1^{p_2^{p_3}}$, p_1 has level 0, p_2 has level 1 and p_3 has level 2), any bracketing of $p_1 * \cdots * p_n$ generates precisely that n -tuple $(0, 1, r_3, \dots, r_n)$ associated with the same bracketing in Example 1.

EXAMPLE 4. $S = \mathbb{R}$, $a * b = a^2 + b^2$.

This is a commutative operation having UNA. This will follow in a way similar to Example 1 as soon as it has been shown that different bracketings of $a_1 * \cdots * a_n$ generate different polynomials in a_1, \dots, a_n . Equivalently, given such a polynomial, can the bracketing that it came from be uniquely reconstructed? To see that it can, notice that the degree of each monomial term, necessarily of the form

$$a_i^{2^{s_i}},$$

involves s_i , the number of times that the operation has been applied to a_i (or to a bracket containing a_i). This information is sufficient to reconstruct the bracketing using the following algorithm:

- (i) Identify s_1, \dots, s_n .
 - (ii) Starting with the largest s_i 's, pair off adjacent equal s_i 's successively, reducing the common s_i of each pair by one after pairing them off, and treating a pair as a single element in each succeeding stage.
 - (iii) The pairings obtained in (ii) correspond to the brackets.
- (For example, the polynomial

$$a_1^2 + a_2^4 + 2a_2^2 a_3^4 + 4a_2^2 a_3^2 a_4^2 + 2a_2^2 a_4^4 + a_3^8 + 4a_3^6 a_4^2 + 6a_3^4 a_4^4 + 4a_3^2 a_4^6 + a_4^8$$

has 1, 2, 3, 3 as a string of s_i 's which the algorithm processes as follows:

$$\begin{array}{ll} \underline{1, 2, 3, 3} & a_1 * a_2 * a_3 * a_4 \\ \underline{1, 2, 2} & a_1 * a_2 * (a_3 * a_4) \\ \underline{1, 1} & a_1 * (a_2 * (a_3 * a_4)) = a_1^2 + (a_2^2 + (a_3^2 + a_4^2)^2)^2. \end{array}$$

As a second example, the string 2, 2, 2, 2 corresponds to the bracketing

$$(a_1 * a_2) * (a_3 * a_4) = (a_1^2 + a_2^2)^2 + (a_3^2 + a_4^2)^2.$$

So the $N(n)$ polynomials $p_i(a_1, \dots, a_n)$, $1 \leq i \leq N(n)$, corresponding to each bracketing are distinct. But \mathbb{R}^n cannot be expressed as the union of 'zero-sets' of a finite number of real (non-zero) polynomials in n variables. (For a proof of this generalization of the fact about hyperplanes appealed to in Example 1, see [3].) In particular, the union of the zero-sets

$$\{(a_1, \dots, a_n) \in \mathbb{R}^n : p_i(a_1, \dots, a_n) - p_j(a_1, \dots, a_n) = 0\},$$

$$1 \leq i, j \leq N(n), \quad i \neq j,$$

does not exhaust \mathbb{R}^n , and there thus exists (a_1, \dots, a_n) such that the values $p_i(a_1, \dots, a_n)$ of each polynomial at (a_1, \dots, a_n) are different real numbers; i.e. all $N(n)$ bracketings of $a_1 * \cdots * a_n$ give different results.

It is interesting to note the parallels in the proofs in Example 1 and Example 4: both operations are defined on \mathbb{R} , and the proofs of existence of real numbers a_1, \dots, a_n for which all $N(n)$ bracketings of $a_1 * \cdots * a_n$ are different both proceed by establishing that agreement of the result of any two bracketings of $a_1 * \cdots * a_n$ is equivalent to the point (a_1, \dots, a_n) belonging

to a subset of \mathbb{R}^n of a certain geometric type, whereas \mathbb{R}^n cannot be expressed as the union of finitely many subsets of such a type.

Open problems

We close with some open problems that the reader is invited to consider.

1. Does vector multiplication ($S = \mathbb{R}^3$, $a * b = a \times b$) have UNA?
2. Does there exist a commutative operation of each possible depth?
3. An alternative characterization of "depth" for a non-associative operation $*$ might be:

$$d'(*) = \begin{cases} \min\{n > 3: & \text{for every } a_1, \dots, a_n \text{ in } S \text{ there are two bracketings of} \\ & a_1 * \dots * a_n \text{ (perhaps depending on } a_1, \dots, a_n) \text{ that give the} \\ & \text{same result}\} \\ \infty & \text{otherwise.} \end{cases}$$

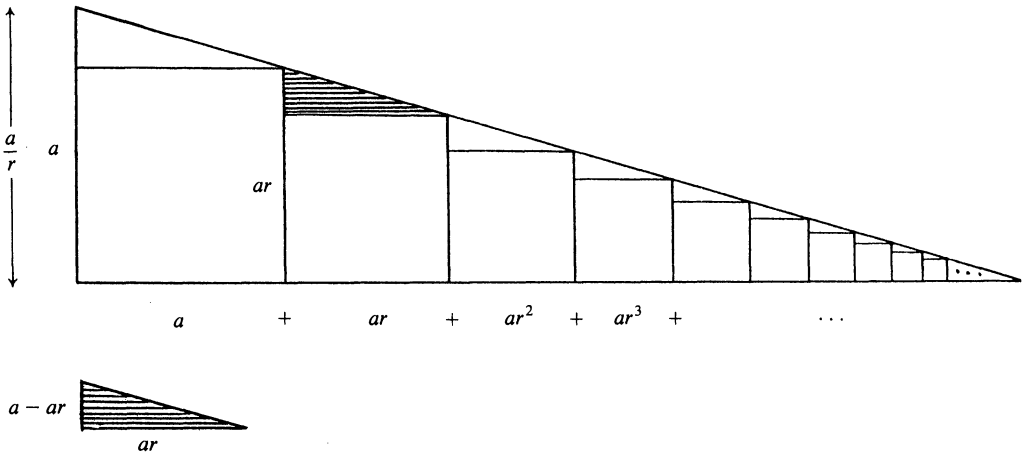
Clearly, $d(*) \geq d'(*)$; can they ever be different?

References

- [1] N. Jacobson, Lectures in Abstract Algebra I, Springer (reprint), 1975, pp. 18–19.
- [2] D. E. Knuth, Fundamental Algorithms, 2nd ed., Addison-Wesley, 1981, p. 531.
- [3] S. Lang, Algebra, Addison-Wesley, 1977, pp. 257–258.

Proof without Words:

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$$



$$\begin{aligned} \frac{a - ar}{ar} &= \frac{\frac{a}{r}}{a + ar + ar^2 + ar^3 + \dots} \\ &\Rightarrow a + ar + ar^2 + ar^3 + \dots \\ &= \frac{a}{1 - r}. \end{aligned}$$

J. H. Webb
University of Cape Town

to a subset of \mathbb{R}^n of a certain geometric type, whereas \mathbb{R}^n cannot be expressed as the union of finitely many subsets of such a type.

Open problems

We close with some open problems that the reader is invited to consider.

1. Does vector multiplication ($S = \mathbb{R}^3$, $a * b = a \times b$) have UNA?
2. Does there exist a commutative operation of each possible depth?
3. An alternative characterization of "depth" for a non-associative operation $*$ might be:

$$d'(*) = \begin{cases} \min\{n > 3: & \text{for every } a_1, \dots, a_n \text{ in } S \text{ there are two bracketings of} \\ & a_1 * \dots * a_n \text{ (perhaps depending on } a_1, \dots, a_n) \text{ that give the} \\ & \text{same result}\} \\ \infty & \text{otherwise.} \end{cases}$$

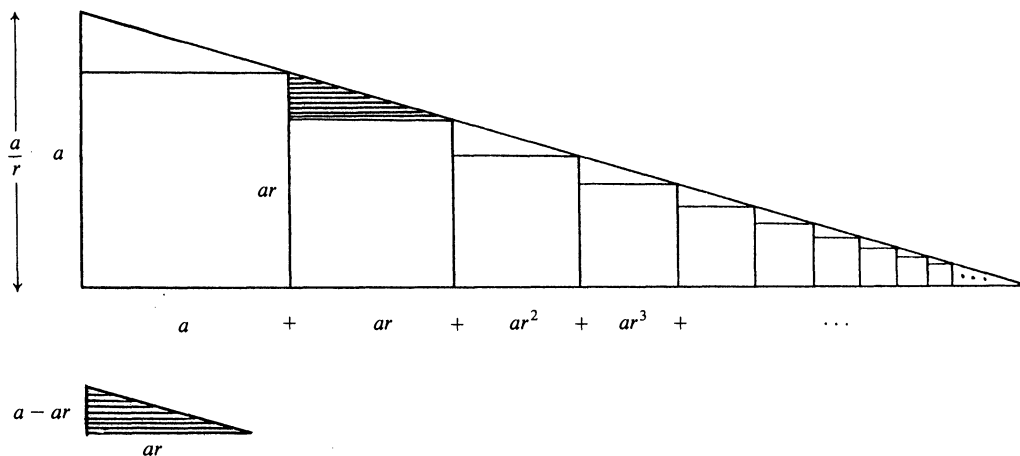
Clearly, $d(*) \geq d'(*)$; can they ever be different?

References

- [1] N. Jacobson, Lectures in Abstract Algebra I, Springer (reprint), 1975, pp. 18–19.
- [2] D. E. Knuth, Fundamental Algorithms, 2nd ed., Addison-Wesley, 1981, p. 531.
- [3] S. Lang, Algebra, Addison-Wesley, 1977, pp. 257–258.

Proof without Words:

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$$



$$\begin{aligned} \frac{a - ar}{ar} &= \frac{\frac{a}{r}}{a + ar + ar^2 + ar^3 + \dots} \\ &\Rightarrow a + ar + ar^2 + ar^3 + \dots \\ &= \frac{a}{1-r}. \end{aligned}$$

J. H. Webb
University of Cape Town

PROBLEMS

LOREN C. LARSON, Editor
BRUCE HANSON, Associate Editor
St. Olaf College

Proposals

To be considered for publication, solutions should be received by December 1, 1987.

Correction to **1259** (February 1987 issue, p. 40).

The term a_{k+1}^2 should read $a^{2(k+1)}$.

1267. *Proposed by Ronald Graham, AT&T Bell Laboratories, Murray Hill, New Jersey.*

Let $X = \{1, 2, \dots, n\}$ and let S be any nonempty collection of subsets of X . Define S' to be the collection of all subsets of X that are subsets of an odd number of elements of S . For example, if $X = \{1, 2, 3\}$ and $S = \{\{3\}, \{1, 2\}, \{1, 3\}\}$, then $S' = \{\emptyset, \{2\}, \{1, 2\}, \{1, 3\}\}$. Prove that $(S')' = S$.

1268. *Proposed by Lawrence Stout, Illinois Wesleyan University, Bloomington.*

An experiment with probability p of success is repeated. Each time a failure occurs, \$1 is put into a kitty. Each time a success occurs, the contents of the kitty are paid out as a jackpot.

a. What is the expected waiting time until the kitty reaches \$ n for the first time starting from an empty kitty?

b. What is the expected waiting time until a jackpot of exactly \$ n is won?

1269. *Proposed by L. Matthew Christophe, Jr., Wilmington, Delaware.*

For each nonnegative real number x , let

$$a_n(x) = \frac{\prod_{k=1}^{n-1} (k+x) \prod_{k=2}^n (k+x)}{(n!)^2}$$

for $n = 1, 2, 3, \dots$. Evaluate $\lim_{n \rightarrow \infty} a_n(x)$ as a function of x .

ASSISTANT EDITORS: CLIFTON CORZATT and THEODORE VESSEY, *St. Olaf College*. We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals should be accompanied by solutions, if at all possible, and by any other information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. An asterisk (*) next to a problem number indicates that neither the proposer nor the editors supplied 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.

Solutions and new proposals should be mailed in duplicate to Loren C. Larson, Department of Mathematics, St. Olaf College, Northfield, MN 55057.

1270. *Proposed by Roger B. Eggleton, The University of Newcastle, Australia.*

Let $n \geq 1$ and $a \geq 2$ be integers. For which values of a and n is $(n+1)!$ a multiple of a^n ?

1271. *Proposed by Skip Thompson, Oak Ridge National Laboratory, Oak Ridge, Tennessee.*

A ball is released from rest at position (x_0, y_0) , where $0 \leq x_0 < 1$ and $x_0 + y_0 > 1$. Its acceleration due to gravity is the constant vector $(0, -g)$. The ball repeatedly bounces on the ramp $x + y = 1$ ($0 \leq x \leq 1$). Suppose that every bounce has the effect of “switching” the velocity vector from (u, v) just before the collision to $(-kv, -ku)$ just after the collision, where k is a constant between 0 and 1. (The case of $k = 1$ corresponds to a perfectly elastic collision.) Determine all values of (x_0, y_0) for which the x -coordinate of the ball will never exceed 1, i.e., the ball will not go beyond the bottom of the ramp.

Quickies

Solutions to Quickies appear on p. 184.

Q722. *Proposed by George E. Andrews, Pennsylvania State University.*

Prove

$$\sum_{j=0}^n (-1)^j 2^{n-j} \binom{x}{j} = \sum_{j=0}^n \binom{n-x}{j}.$$

Q723. *Proposed by M. S. Klamkin, University of Alberta, Canada.*

$ABCD$ is a quadrilateral inscribed in a circle. Prove that the four lines, each passing through a midpoint of one of the sides of $ABCD$ and perpendicular to the opposite side, are concurrent.

Solutions

Bijection Between Integers and Composites

June 1986

1242. *Proposed by Ronald Graham, AT&T Bell Laboratories, Murray Hill, New Jersey.*

For each positive integer n , let $f(n)$ be the smallest integer r for which there is an increasing sequence of integers $n = a_1 < a_2 < \cdots < a_k = r$ such that the product $a_1 a_2 \cdots a_k$ is a perfect square. For example, $f(2) = 6$, $f(3) = 8$, $f(4) = 4$, $f(8) = 15$. Prove that f is a one-to-one function.

Solution by Michael Reid (student), Harvard College.

Note that for $n > 1$, $f(n)$ is composite. We prove the following stronger claim. Let $C = \{4, 6, 8, 9, 10, \dots\}$ be the set of composites, and let $B = \{2, 3, 4, 5, \dots\}$ be the set of integers greater than 1. Then $f: B \rightarrow C$ is a bijection.

To simplify notation, let $\Pi(S)$ denote $\prod_{n \in S} n$, where $S \subseteq \mathbb{N}$ is finite, and $\Pi(\emptyset) = 1$.

Suppose $a < b$ and $f(a) = f(b) = r$. Then from the definition of f , there are sets

$$\{a, r\} \subseteq S \subseteq \{a, a+1, \dots, r\}$$

$$\{b, r\} \subseteq T \subseteq \{b, b+1, \dots, r\},$$

where $\Pi(S)$ and $\Pi(T)$ are both perfect squares. Let U be the symmetric difference of S and T . Then $\Pi(U)[\Pi(S \cap T)]^2 = \Pi(S)\Pi(T)$, whence $\Pi(U)$ is a perfect square. Also

$$a \in U \subseteq \{a, a+1, \dots, r-1\},$$

so $f(a) < r$, a contradiction. Thus, f is injective.

If r is composite, then $r = p_1 p_2 \cdots p_j t^2$, where $p_1 < p_2 < \cdots < p_j$ are primes. Thus we have the decreasing sequence $r > p_j > \cdots > p_1$, with $r p_j \cdots p_2 p_1$ a perfect square. For composite r , define $g(r)$ as the largest integer n for which there is a decreasing sequence $r = a_1 > a_2 > \cdots > a_k = n$, such that the product $a_1 a_2 \cdots a_k$ is a perfect square. Clearly, $f(g(r)) \leq r$, so suppose for some r , we have $f(g(r)) = s < r$. As before, there are sets

$$\{r, g(r)\} \subseteq S \subseteq \{r, r-1, \dots, g(r)\},$$

$$\{g(r), r\} \subseteq T \subseteq \{g(r), g(r)+1, \dots, s\},$$

such that $\Pi(S)$ and $\Pi(T)$ are perfect squares. Again, if U is the symmetric difference of S and T , then $\Pi(U)$ is a perfect square, and

$$r \in U \subseteq \{r, r-1, \dots, g(r)+1\},$$

contradicting the definition of g . Thus, $f(g(r)) = r$, and f is surjective.

Also solved by John O. Bennett, May Beresin and Eugene Levine, Aaron Bertram (student), Edwin Buchman, Sydney Bulman-Fleming (Canada), A. A. Diwan (India), Roger B. Eggleton (Australia), David C. Flaspohler, Zachary Franco, Sharon Kunoff, Oxford Running Club (University of Mississippi), John P. Robertson, Nimish Shah (student; India), Robert Simon (student), Paul Smith (Canada), Patrick Touhey (student), Robert L. Taylor, and Douglas H. Underwood.

1243. Proposed by Edwin Buchman, California State University, Fullerton.

How many different numbers can be represented among the six expressions:

$$\begin{aligned} e_1: \lim_{x \rightarrow a} g(f(x)) & \quad e_4: g(f(a)) \\ e_2: g(\lim_{x \rightarrow a} f(x)) & \quad e_5: \lim_{y \rightarrow f(a)} g(y) \\ e_3: \lim_{x \rightarrow a} \lim_{y \rightarrow f(x)} g(y) & \quad e_6: \lim_{y \rightarrow \left(\lim_{x \rightarrow a} f(x) \right)} g(y). \end{aligned}$$

I. Solution by Peter Collinge, Brockport, New York.

Consider

$$f(x) = \begin{cases} 0 & x = 0 \\ x + 1 & \text{otherwise,} \end{cases} \quad \text{and } g(y) = \begin{cases} 2 & y = 1 \\ -1 & y = 0 \\ y & \text{otherwise.} \end{cases}$$

Then, for $a = 0$, $e_1 = 1$, $e_2 = 2$, $e_3 = 1$, $e_4 = -1$, $e_5 = 0$, and $e_6 = 1$. Thus, at least four different numbers can be represented.

We will show that if all six limits exist, then at most four different limits are possible, and these will be found as e_2 , e_3 , e_4 , and e_5 . To simplify notation, let $L = \lim_{x \rightarrow a} f(x)$.

A careful look shows that if e_3 and e_6 both exist, then $e_3 = e_6$. To see this, note that $e_6 = \lim_{y \rightarrow L} g(y)$. Then, since $y \rightarrow L$ as $y \rightarrow f(x)$ and $x \rightarrow a$, we know that $\lim_{x \rightarrow a} \lim_{y \rightarrow f(x)} g(y) = e_3$ must be equal to e_6 .

Finally, we show that if e_1 , e_2 , and e_6 exist, then $e_1 = e_2$ or $e_1 = e_6$. Suppose that $f(x) \equiv L$ for all x within some neighborhood of a . Then $e_1 = \lim_{x \rightarrow a} g(f(x)) = g(L) = e_2$. Otherwise, each neighborhood of a contains values of x such that $f(x) \neq L$. Thus there exists a sequence (x_n) with $\lim_{n \rightarrow \infty} x_n = a$ and $x_i \neq a$, $f(x_i) \neq L$ for all i . Now, $\lim_{n \rightarrow \infty} f(x_n) = L$ since (x_n) converges to a , so $\lim_{n \rightarrow \infty} g(f(x_n)) = e_6$. Similarly, since $\lim_{n \rightarrow \infty} x_n = a$, then

$$\lim_{n \rightarrow \infty} g(f(x_n)) = e_1.$$

Then $e_1 = e_6$. So $e_1 = e_2$ or $e_1 = e_6$.

II. Solution by Nimish Shah (student), St. Xavier's College, India.

It is possible to get five distinct values if one of the limits fails to exist. For example, define $g: \mathbf{R} \rightarrow \mathbf{R}$ by $g(x) = \sin(1/x)$, if $x \neq 0$, $2/\pi, 1/k\pi$ for $k = \pm 1, \pm 2, \pm 3, \dots$, $g(0) = -1/2$, $g(2/\pi) = -1$, $g(1/k\pi) = 1/2$ for $k = \pm 1, \pm 2, \pm 3, \dots$. Define $f: \mathbf{R} \rightarrow \mathbf{R}$ by $f(0) = 2/\pi$, and for $|x| \in (1/2^k, 1/2^{k-1}]$, $k = 1, 2, 3, \dots$, $f(x) = 1/k\pi$ if $x > 0$, and $f(x) = -1/k\pi$ if $x < 0$, and for any other x , set $f(x) = 0$.

Then it is easy to check that

$$\begin{aligned} e_1: \lim_{x \rightarrow 0} g(f(x)) &= \frac{1}{2}, & e_4: g(f(0)) &= -1, \\ e_2: g(\lim_{x \rightarrow 0} f(x)) &= -\frac{1}{2}, & e_5: \lim_{y \rightarrow f(0)} g(y) &= 1, \\ e_3: \lim_{x \rightarrow 0} \lim_{y \rightarrow f(x)} g(y) &= 0, & e_6: \lim_{y \rightarrow \left(\lim_{x \rightarrow 0} f(x) \right)} g(y) &\text{does not exist.} \end{aligned}$$

Here, e_1 , e_2 , e_3 , e_4 , e_5 are different, but e_6 does not exist.

Also solved by Zachary Franco, mathematics teachers participating in the 1986 Houston Mathematics and Science Improvement Consortium Program, and the proposer.


1244. Submitted by the Problem Solving Class, University of California, Berkeley.

Define a boomerang to be any nonconvex quadrilateral. Prove that it is impossible to tile any convex polygon with a finite number of (not necessarily congruent) boomerangs.

Solution by Allen J. Schwenk, Western Michigan University.

A convex n -gon C has vertices with angles totaling $(n-2)\pi$. Suppose it has been tiled with k boomerangs. Each boomerang has an angle exceeding π which must lie in the interior of C . Since the large angles prevent two of these vertices being placed at the same point, they comprise k interior points with angles of $2k\pi$ to be covered. Thus, the angles of the boomerangs must cover angles totaling $(2k+n-2)\pi$, but the boomerangs only contain angles totaling $2k\pi$. That is, no tiling is possible.

Also solved by Irl Bivens and L. R. King, Edwin Buchman, Nick Martin (student), and Dennis White.

Berry Kercheval seized upon the strict interpretation of the word "any" to give a counterexample. Consider the quadrilateral: . Two of these, one rotated by 90 degrees, will tile a unit square.

A Logarithmic Inequality

June 1986

1245. Proposed by Fouad Nakhli (student), American University of Beirut, Lebanon.

For each number x in the open interval $(1, e)$ it is easy to show that there is a unique number y in (e, ∞) such that $(\ln y)/y = (\ln x)/x$. For such an x and y , show that $x + y > x \ln y + y \ln x$.

I. Solution by the Chico Problem Group, California State University, Chico.

We have $\ln y - \ln x = \int_x^y (1/t) dt$. Because the graph of the function $f(t) = 1/t$ is concave upward, the integral is less than its trapezoidal approximation. Therefore,

$$\ln y - \ln x < (y - x) \left(\frac{1}{y} + \frac{1}{x} \right) / 2, \text{ or equivalently, } 2 \frac{\ln y - \ln x}{y - x} < \frac{1}{y} + \frac{1}{x}.$$

The straight line through the three points $(0, 0)$, $(x, \ln x)$, $(y, \ln y)$ has slope

$$\frac{\ln y - \ln x}{y - x} = \frac{\ln x}{x} = \frac{\ln y}{y}.$$

Thus,

$$\frac{\ln x}{x} + \frac{\ln y}{y} < \frac{1}{y} + \frac{1}{x},$$

or

$$y \ln x + x \ln y < x + y.$$

II. Solution by N. J. Lord, Tonbridge School, Surrey, England.

Let $x/y = 1 - t$, $0 < t < 1$. Then $(\ln y)/y = (\ln x)/x$ is the same as $(\ln y)/y = ((\ln y)(1 - t))/(y(1 - t))$, and this is equivalent to $(1 - t)\ln y = \ln y + \ln(1 - t)$. It follows that

$$\begin{aligned} \ln y &= -\frac{\ln(1 - t)}{t} \\ &= 1 + \frac{1}{2}t + \frac{1}{3}t^2 + \dots \end{aligned}$$

$$\begin{aligned}
&< 1 + \frac{1}{2}t + \frac{1}{2}t^2 + \dots \\
&= 1 + \frac{t/2}{1-t} \\
&= \frac{y+x}{2x}.
\end{aligned}$$

Hence $x \ln y < (1/2)(y+x)$. But $x \ln y = y \ln x$, so $x \ln y + y \ln x < 2 \cdot (1/2) \cdot (y+x) = y+x$.

III. *Solution by Allen J. Schwenk, Western Michigan University.*

We shall verify the stronger inequality

$$\frac{1}{x} + \frac{1}{y} \geq \frac{2}{e}. \quad (1)$$

To see that this is indeed stronger, first examine $(\ln x)/x$ to see that the maximum value $1/e$ is achieved only at $x=e$. Thus,

$$\frac{2}{e} = \frac{1}{e} + \frac{1}{e} > \frac{\ln x}{x} + \frac{\ln y}{y} \quad \text{for } x < e < y. \quad (2)$$

Now combining (1) and (2) and multiplying by xy produces the inequality posed.

To verify (1), we use the Lagrange multiplier technique to minimize $f(x, y) = 1/x + 1/y$ subject to the constraint $g(x, y) = (\ln x)/x - (\ln y)/y = 0$. Now $\nabla f = \lambda \nabla g$ implies

$$\left(\frac{-1}{x^2}, \frac{-1}{y^2} \right) = \lambda \left(\frac{1 - \ln x}{x^2}, \frac{\ln y - 1}{y^2} \right). \quad (3)$$

Eliminating λ gives $2 = \ln x + \ln y$, or $xy = e^2$ at any extremum. But then the arithmetic-geometric mean inequality requires that

$$\frac{1}{2} \left(\frac{1}{x} + \frac{1}{y} \right) \geq \sqrt{\frac{1}{xy}} = \frac{1}{e}. \quad (4)$$

This verifies (1) at extrema for $f(x, y)$ on $g(x, y) = 0$. Now as $x \rightarrow \infty$ on $g(x, y) = 0$ we have $y \rightarrow 1$ and vice-versa. Hence,

$$\lim_{x \rightarrow \infty} \frac{1}{x} + \frac{1}{y} = \lim_{y \rightarrow \infty} \frac{1}{x} + \frac{1}{y} = 1 > \frac{2}{e}.$$

That is, $f(x, y) \geq 2/e$ at each extremum and in the limit as $x \rightarrow \infty$ or $y \rightarrow \infty$. Thus, by continuity, (1) is true for all (x, y) on $g(x, y) = 0$. Finally, we observe that equality occurs in (4) if and only if $x = y = e$, so in fact the inequality in (1) is strict for x in $(1, e)$.

Also solved by Barry Brunson, Edwin Buchman, George Crofts, Harry D'Souza, Thomas P. and Joseph B. Dence, Mordechai Falkowitz, Ismor Fischer (student), Zachary Franco, Dennis Hamlin (student), Robert Heller, Francis M. Henderson, Benjamin G. Klein, L. Kuipers (Switzerland), Kee-wai Lau (Hong Kong), Roger B. Nelsen, Stephen Noltie, Beno Onbel (Israel), H.-J. Seiffert (two solutions; West Germany), Robert E. Shafer, Nimish Shah (student; India), J. M. Stark, Michael Vowe (Switzerland), Western Maryland College Problems Group, Yan-loi Wong, and the proposer. There were two incorrect solutions.

A Convergence Test?

June 1986

1246. *Proposed by Albert Wilansky, Lehigh University.*

Is this a test for convergence of $\sum_{n=1}^{\infty} a_n$? For each $\varepsilon > 0$ there exists a sequence $(b_n)_{n=1}^{\infty}$ with $|1 - b_n| < \varepsilon$ for all n such that $\sum_{n=1}^{\infty} a_n b_n$ is convergent.

Solution by Edwin Buchman, California State University, Fullerton.

Of course, every convergent series passes this test. But in general the "test" is not valid.

Let (a_n) be the sequence

$$1 + \frac{1}{\sqrt{1}}, -\frac{1}{\sqrt{1}}, \frac{1}{2} + \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, \frac{1}{3} + \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \dots$$

The even-numbered partial sums of $\sum a_n$ are the partial sums of the harmonic series, so $\sum a_n$ diverges.

Let (c_n) be the sequence

$$\frac{\sqrt{1}}{1 + \sqrt{1}}, 1, \frac{\sqrt{2}}{1 + \sqrt{2}}, 1, \frac{\sqrt{3}}{1 + \sqrt{3}}, 1, \dots$$

Note that $c_n \rightarrow 1$ as $n \rightarrow \infty$, so that for every $\varepsilon > 0$, a sequence (b_n) exists such that $|b_n - 1| < \varepsilon$, for all n , and $b_n = c_n$ for all but a finite number of values of n . Except for these values of n , the sequence $(a_n b_n)$ is identical to $(a_n c_n)$, which is the sequence

$$1, -1, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \dots$$

Since the series $\sum a_n c_n$ converges, so does $\sum a_n b_n$.

Also solved by Chico Problem Group, George Crofts, Roger B. Eggleton (Australia), Mordechai Falkowitz, Ole Jørsboe (Denmark), Richard Katz, N. J. Lord (England), Richard Neidinger, Stephen Noltie, Allen J. Schwenk, Nimish Shah (student; India), and Dennis White. There were two incorrect solutions.

Several readers noted that the test is valid if convergence of $\sum a_n b_n$ is replaced by absolute convergence of $\sum a_n b_n$.

Answers

Solutions to the Quickies which appear on p. 179.

A722. Both sides are polynomials in x of degree n . For $x = 0, 1, \dots, n$ the right side is 2^{n-x} and the left side is $2^n(1 - \frac{1}{2})^x = 2^{n-x}$. Two polynomials of degree n in x that agree for $n + 1$ values are identical.

A723. Let **A**, **B**, **C**, **D** be vectors from the center of the circle to the respective vertices, A , B , C , D . The four lines will intersect at the point P given by $\mathbf{P} = (\mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D})/2$. Note that the vector from the midpoint of AB to P is $(\mathbf{C} + \mathbf{D})/2$, and this is perpendicular to CD since $|\mathbf{C}| = |\mathbf{D}|$, and similarly for the other segments BC , CD , and DA . It also holds for the diagonals AC and BD , so that there are six concurrent lines.

This result was given by R. E. Lester, *Mathematical Gazette*, 46 (1962) p. 147. For other interesting properties of the point P , see R. A. Johnson, *Advanced Euclidean Geometry*, Dover, New York, 1960, p. 252.

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, Mathematicians look to SDI for research funds, *Science* 234 (7 November 1986) 665-666.

Why are mathematicians always so late in getting on the gravy train? At the instigation of The Conference Board on the Mathematical Sciences, at last a group of mathematicians has met with directors of the Strategic Defense Initiative to discuss funding of related mathematics research. At stake is \$1-2 million in 1987, just crumbs from the table of our computer science colleagues. Morality being a function of poverty, more adequate funding of mathematical research by NSF would make it easier to say "no thanks" to war work.

Peitgen, H.-O., and P. H. Richter, *The Beauty of Fractals: Images of Complex Dynamical Systems*, Springer-Verlag, 1986; x + 199pp, \$29.50.

Documentation in English, replete with color photos, of the exhibition "Schönheit in Chaos/Frontiers of Chaos." The mathematical explanations are here, too, at an upper undergraduate level, along with guest essays by Mandelbrot and others. But it is the beautiful images that are the lure, and we should show them to all, far and wide.

Baum, Joan, *The Calculating Passion of Ada Byron*, Shoe String Press, 1986; xix + 133 pp, \$21.50.

Despite dust-jacket testimonials by Martin Gardner and Peter Hilton, this biography does not dig as deeply into the details of Ada's life and circumstance as Dorothy K. Stein's *Ada: A Life and a Legacy* (1985). Author Baum does not seem to have benefited from Stein's scholarship (their writing may have been concurrent). True, Baum does not "force tenuous conclusions about this driven personality" (as the dust-jacket obliquely damns Stein's book). Like Stein, Baum imposes her own interpretation on Ada's life; but appreciative as it is, it seems to go little beneath the surface facts. Baum, a professor of English, seems overly impressed at the naming of a computer language after Ada; perhaps we will not have a definitive biography until a historian trained in mathematics and computing essays the task.

Coxeter, H.S.M., et al. (eds.), *M.C. Escher: Art and Science*, North-Holland, 1986; xiv + 402 pp.

As long as there is geometry, individuals will be attracted to it because of Escher's prints. The 35 essays here were presented at an interdisciplinary congress in Rome in 1985. Most of the well-known Escher-loving geometers are represented, along with people from other disciplines. Topics vary from "What Escher might have done" (G.C. Shephard) to "Creating hyperbolic Escher patterns" (D.J. Dunham).

Acton, John R., and Patrick T. Squire, *Solving Equations with Physical Understanding*, Adam Hilger, 1985; x + 219 pp (distributed in the U.S. by International Publishers Service, \$26.00).

Featuring "approximation with understanding as its goal," this book offers a "back of the envelope" approach to differential equations. Designed to be studied *after* a conventional course in differential equations, the book will benefit readers of all degrees of sophistication.

Kolata, Gina, Math proof refuted during Berkeley scrutiny, *Science* 234 (19 December 1986) 1498-1499.

For most of 1986 there was reason to believe that the 3-D Poincaré conjecture had finally been proved, after 80 years. The claim by Colin Rourke (Warwick) and Eduardo Rego (Oporto) was first announced publicly in *Nature* in March 1986. Mathematicians don't read *Nature*, and there wasn't even any discussion on the grapevine at the International Congress of Mathematicians in August. Without the splash in the *New York Times* at the end of September, we would have had to confront a classic philosophical problem: If a conjecture is proved in the "hinterlands" but no mathematician hears about it, is the mathematics sound? At Rourke's November seminar at Berkeley, the audience pointed out a gap in the proof that he could not fill; and now colleagues are crawling out of the woodwork to attest that they never really believed the proof. Is it so embarrassing for mathematicians to have the public be told of a result, only to have a gap emerge? After all, most of the public would never have heard of the Poincaré conjecture otherwise; and aren't physicists and others always announcing new theories, many of which are later refuted? Isn't it more embarrassing that communication even *within* the mathematical community is so poor that most mathematicians heard nothing from their professional associations about all this?

Albers, Donald J., G. L. Alexanderson and Constance Reid, *International Mathematical Congresses: An Illustrated History 1893-1986*, Springer-Verlag, 1986; 62 pp, \$29.95.

Delightful souvenir commemorative volume, published on the occasion of the 1986 International Congress at Berkeley. Includes lists of all the plenary lectures and photos and one-sentence accounts of the work of Fields Medalists.

Lowe, John W. G., *The Dynamics of Apocalypse: A Systems Simulation of the Classic Maya Collapse*, University of New Mexico Press, 1985; vii + 275 pp, \$22.50.

The author, with a B.S. in engineering physics, an M.S. in physics, and a Ph.D. in anthropology, brings a new background to contemplation of the collapse of the Classic Maya civilization a thousand years ago. After considering the available data (77 sites with dated monuments), the agreed-upon history, and previous theories, he formulates an epidemiological model for influence on sites by food shortages and population growth. The three ordinary differential equations are subjected to identification of parameters and both formal and computer analysis (he's an engineer, so it's completely undocumented Fortran). Sensitivity analysis shows that the revealed threshold effect is independent of exact specification of parameters: "...the social and demographic catastrophe...was the nearly inevitable outcome of a collision of constraints."

Peterson, I., Pi wars: Dueling supercomputers, *Science News* 131 (21 February 1987) 118.

"The story of computing digits of pi is no longer a story of great practicality. It hasn't been a story of great practicality since maybe the 16th century ..., but it is a problem that has captured many, many people's imaginations." So says Peter Borwein (Dalhousie University), who with his brother Jonathan developed the algorithms for the recent record-setting computation of pi to 134,217,700 digits by Yasumasa Kanada (University of Tokyo) on an NEC SX-2 supercomputer. On each iteration, Borwein's algorithm gives four times as many correct digits as at the end of the preceding iteration. Borwein asserts that his algorithm is close to best possible, so that he conjectures that no one will ever know the 10^{1000} th digit of pi.

McKnight, Curtis, et al., *The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective* (A National Report on the Second International Mathematics Study), Stipes Publishing Company (10-12 Chester St., Champaign, IL 61820), 1987; xiv + 127 pp, \$8 (P) + \$1 postage.

How bad is the mathematics achievement of U.S. students? At age 13 (8th grade), U.S. students swamped Swaziland; at age 17 (12th grade), they topped Thailand. But U.S. achievement levels are barely one-half to two-thirds those of students in Japan and Hong Kong. How come? This report considers and dismisses five "explanations" as "deceptive": less instruction time (44% more in U.S.), larger class size (44 in Hong Kong, 26 in U.S.), education for the many (92% of 17-year-olds in Japan are in school, 82% in U.S.), relatively untrained or inexperienced teachers (not true), poor quality of teaching (little difference). So, what's to do? "Much could be accomplished ... by the expectation that students can accomplish far more than they now do, and by clear goals for a reorganized and revitalized mathematics curriculum." There's more; everyone concerned about mathematics education in the U.S. should read this report.

Holmes, Peter, *The Best of Teaching Statistics: A Series of Articles for Teachers of Pupils Aged 9 to 19*, Teaching Statistics Trust (Centre for Statistical Education, 25 Broomgrove Road, Sheffield S10 2TN, England), 1986; 184 pp (P), U.S. \$11.00, CDN \$15.00.

For many year British mathematics teaching has emphasized probability and statistics; now American schools are adopting integrated curricula for grades 7-12. This book collects valuable source materials, selected from the first seven volumes of the journal *Teaching Statistics*, and including several prize-winning articles. College teachers of the ubiquitous elementary statistics course will find useful material here.

Douglas, Ronald G., *Toward a Lean and Lively Calculus*, MAA, 1986; xxvi + 249 pp, \$12 (P).

Proceedings of a conference of curriculum leaders, who (surprisingly) agreed that the calculus syllabus should contain fewer topics but have more conceptual depth, numerically and geometrically. Lynn Steen's "Twenty questions for calculus reformers" is concise, succinct, and thought-provoking; and the remaining essays provide partial answers. One common theme is that a major reason why the teaching of calculus in the U.S. is unsuccessful is that the students in calculus classes are "not intellectually ready to understand the central ideas of calculus." (In an era of desktop publishing, however, editors should demand that authors submit "camera-ready" contributions in letter-quality print, not dot-matrix on a worn ribbon.)

Curcio, Frances R., *Teaching and Learning: A Problem-Solving Focus*, NCTM, 1987; viii + 116 pp, \$12 (P).

Despite the uninformative title, this is actually a memorial volume to George Pólya. It includes his famous essay "On learning, teaching, and learning teaching," a bibliography of his work in mathematics education, and a biographical sketch, together with other essays furthering the spirit of his approach to problem solving.

The Spode Group, *Realistic Applications in Mechanics*, Oxford University Press, 1986; viii + 76 pp, \$19.95 (P).

Seventeen case studies, each starting from a real-life mechanics problem in sports, driving, or other areas. The solutions freely use calculus and concepts from mechanics (e.g., impulse, moment of inertia). With one chapter per group member, the book was drafted in a single weekend! A few rough spots remain: some square roots fly away (bottom, p. 11); the discussion about disk brakes is garbled (bottom, p. 43); several studies seem closely derived from other sources; more references are needed, especially for the empirical constants that are quoted, as well as for further reading (e.g., readers of "Miracle at Mexico City?" might be glad to know of the fuller treatment in the pages of this *MAGAZINE*); and finally, the U.S. price is high (but explicit permission is given for teachers to reproduce individual pages for class use).

Peterson, Ivars, Zeroing in on chaos, *Science News* 131 (28 February 1987) 137-139.

"Plain old" Newton's method, when coupled to computer graphics and applied to finding complex roots, reveals still new secrets. Each point in the plane is assigned a color corresponding to the root to which the method converges when started at that point, with shading used to indicate how quickly the method converges from there. "The result is a glowing tapestry"—and, of course, fractals turn up! Points that lead to no root comprise a Julia set. Here are discoveries that were not made before computers (gee, expensive ones, with color displays, even!) appeared on a few mathematicians' desks. How much longer will it be before we realize that when it comes to computers, "the computer scientists stole our lunch" (in the words of Louis Rall (Wisconsin)), and it's time we made our case for funding mathematics with more than chalk?

Jensen, Roderick V., Classical chaos, *American Scientist* 75 (March-April 1987) 168-181.

Excellent exposition of the mathematics of chaos and the consequences in physics ("Is physics conquering chaos, or chaos undermining physics?").

Brief U.S. suppression of proof stirs anger, *New York Times* (17 February 1987).

In January Adi Shamir (Weizmann Institute), one of the world's best-known cryptologists, was ordered by the U.S. government to recover and destroy all preprints and material related to a breakthrough he had made concerning "unforgeable ID cards" based on zero-knowledge proofs. The secrecy order originated with the U.S. Army, which was contacted by the Patent Office when Shamir sought a patent on his method. Within days, the government was forced by public furor among mathematicians to rescind its order; it did so on the narrow ground that secrecy could not be imposed on a non-citizen. Whether the National Security Agency had any role in the secrecy order or its rescinding is unclear; associates of Shamir assert that he had earlier submitted his work to NSA, which had not requested suppression of publication.

Gleick, James, A new approach to protecting secrets is discovered, *New York Times* (17 February 1987) 17-18.

Adi Shamir (Weizmann Institute) and two co-workers, Amos Fiat and Uriel Feige, have designed a way to incorporate zero-knowledge proofs into electronic chips that could be imbedded in "smart" credit cards. Such a card could identify the cardholder to a merchant, without giving the merchant the card number or other information that could let the merchant make unauthorized purchases or forge a new card. The advantage of Shamir's technique is speed. Similar systems could be used by the military for unforgeable "friend or foe" recognition signals or for emergency destruct signals from ground controllers to missiles—and even the signals to launch warheads.

Stewart, Ian, Topology: The three-sphere strikes back, *Nature* 325 (12 February 1987) 579-580.

Stewart traces the flaw in the prematurely-announced "proof" of the Poincaré conjecture, which flaw in fact was pointed out by one of its authors.

Children's Television Workshop, Square One TV, showing on public television since February; 75 half-hour shows, \$16 million.

Aimed at ages 8 to 12, this series by the creators of Sesame Street is supposed to make math fun and interesting. As Richard Zoglin of *Time* remarked, "The mathematics is so well hidden as to be nearly invisible." The shows' takeoffs on classic TV and films (Dragnet, the original Saturday Night Live, The Paper Chase, Casablanca) no doubt amused the writers and producers; but the allusions are likely to be lost on the intended audience, even though, as Zoglin remarks, what children know best is other TV shows. The shows, however, are not designed to teach any mathematics, merely to lure viewers away from competing fare to show them that mathematics can be useful. The message may well be lost in the medium: what television reinforces is television.

Stewart, Ian, Number theory: Geometry finds factors faster, *Nature* 325 (15 January 1987) 199.

H. Lenstra (Amsterdam) has developed a new method for efficient factorization of integers, using elliptic curves. A key feature of the new algorithm—apart from the unexpected connection between factoring and elliptic curves—is that the running time depends on the size of the prime factors.

Ascher, Marcia, and Robert Ascher, *Ethnomathematics, History of Science* 24 (1986) 125-144.

"A fair number of statements about nonliterate peoples are found in the mathematical literature ... Most of the presentations are theoretically and factually flawed." The authors offer an anthropological perspective to juxtapose against the cultural chauvinism projected by standard histories of mathematics.

Littlewood, J.E., *Littlewood's Miscellany*, edited by Béla Bollobás, Cambridge U Pr, 1986, 200 pp, (P).

After 20 years out of print, Littlewood's *A Mathematical Miscellany* is now available again, with other new material from Littlewood's latter 25 years. The book exhibits "the gaiety of genius," in its amusing anecdotes about mathematics and mathematicians, by a man regarded by G.H. Hardy as "the greatest mathematician he has even known." Bollobás has added a biographical foreword.

Devaney, Robert L., *An Introduction to Chaotic Dynamical Systems*, Benjamin Cummings, 1986; xiv + 319 pp, \$31.95.

Splendid introduction to nonlinear dynamical systems, accessible to students with advanced calculus and linear algebra background (no previous differential geometry, measure theory, or topology required). (For only \$500-\$1000 more in production costs, the author and publishers could have rendered the text in high-density laser printer and avoided fuzzy letters and symbols.)

Schroeder, M.R., *Number Theory in Science and Communication, with Applications in Cryptography, Physics, Digital Information, Computing, and Self-Similarity*, 2nd enl. ed., Springer-Verlag, 1986; xix + 374 pp, \$24.50 (P).

A welcome early appearance of a second edition of a marvelous book. Featured additions include sections on quasicrystals and "silver ratios," self-similarity and fractals, deterministic chaos, Hensel codes, and applications of the Zech logarithm.

Hargittai, István (ed.), *Symmetry: Unifying Human Understanding*, Pergamon; xi + 1045 pp, \$115.00.

Symmetry is a perennial topic, and the 65 essays here celebrate its manifestations in chemistry, art, music, country dancing, and everything else under the sun. Among the authors are many of the world's eminent geometers; all but one of the contributors come from Europe and North America. (The volume also serves as Vol. 12B of the journal *Computers and Mathematics with Applications*.)

Traub, Joseph F., et al. (eds.), *Annual Review of Computer Science*, Vol. 1, 1986, Annual Reviews Inc., 1986; xii + 459 pp, \$39.

Collection of 14 topical essays, on various topics ranging from advances in compiler technology to dataflow architectures, from information-based complexity to natural-language interfaces. Only a couple (by contributors from Columbia) deviate from the objective of offering a wide survey to discuss instead favorite work of their own. Despite the title of this new serial, the essays do not relate specifically to achievements in computer science in the preceding year. All the same, it's time mathematics had an annual in this series, which embraces 26 fields of knowledge.

Dewdney, A.K., and I.R. Lapidus (eds.), *The Second Symposium on Two-Dimensional Science and Technology*, Turing Omnibus, Inc. (P.O. Box 1456, London, Ontario, Canada N6A 5M2), 1986; ii + 125 pp.

More contributions to the possible features of a real two-dimensional world, ranging from an investigation of the periodic table to 2-D clocks, from plasma physics to art in the "planiverse."

Walker, Jearl, *The Amateur Scientist: Methods for going through a maze without becoming lost or confused*, *Scientific American* 255:6 (December 1986) 132-139, 152.

Enjoyable exposition of the mathematics behind threading mazes, including the use of graphs and their corresponding matrices.

Kocak, Huseyin, *Differential and Difference Equations Through Computer Experiments with Diskettes Containing PHASER; An Animator/Simulator for Dynamical Systems for IBM Personal Computers*, Springer-Verlag, 1986; xv + 224 pp + 2 diskettes, \$44.

"Phase space—the final frontier. These are the voyages of the diskette PHASER, her user's mission to explore strange new attractors, to seek out new equations and new dynamics, to boldly go where no P(oin)C(aré) has gone before." Part text, mostly manual and tutorial, the book provides excellent accompaniment to a versatile program package for difference and differential equations, including dynamical systems.

Conrad, Steven R., and Daniel Flegler, *The 1st High School Math League Problem Book*, Steven R. Conrad (117 Manhasset Ave., Manhasset, NY 11030), 1986; ii + 66 pp, \$12.95 (P).

Collection of math league contest problems, all original, with solutions; the problems require knowledge only of high-school mathematics (excluding calculus) and are designed to be solved without calculators or graph paper.

Sloyer, Cliff, *Fantastiks of Mathematics: Applications of Secondary Mathematics*, Janson Publications, 1986; xi + 143 pp, \$27.50, \$13.95 (P).

Collection of interesting and stimulating illustrations of potential uses of particular mathematical concepts. Some of the punch of reality, though, is lost in the use of imaginary companies and made-up data.

Feuer, Lewis S., America's first Jewish professor: James Joseph Sylvester at the University of Virginia, *American Jewish Archives* 36:2 (November 1984) 151-201; a shortened version appears as Sylvester in Virginia, *Mathematical Intelligencer* 9:2 (1987) 13-19.

Less well known than Sylvester's five-year visit to Johns Hopkins is his four-month tenure at Virginia thirty-five years earlier. This well-researched article investigates the suppressed reasons for his sudden resignation at Virginia and the unfortunate effects of the turn of events on his production of mathematics, and details the extreme conditions at what was "during the pre-Civil War years the most lawless and prone to extreme violence of all American universities."

La Brecque, Mort, Fractal applications, *Mosaic* 17:4 (Winter 1986) 34-48.

Fractal rainfall, protein surfaces, protein behavior, seismic faults, cluster of galaxies: there are fractals everywhere, and the National Science Foundation is funding a lot of non-mathematicians to look into such applications.

Jaffe, A.J., and Herbert F. Spirer, *Misused Statistics: Straight Talk for Twisted Numbers*, Dekker, 1987; xi + 237 pp.

Informative guide to both deliberate and inadvertent misuses of statistics, with examples from news stories, journals, and government reports.

NEWS & LETTERS

MATHEMATICS AND STATISTICS CONFERENCE

The Fifteenth Annual Mathematics and Statistics Conference at Miami University, Oxford, Ohio, will be held October 9 and 10, 1987. The theme for this year's conference will be "Computers and Mathematics". Featured speakers will include Robert E. Tarjan, AT&T Bell Laboratories and Princeton University; Anthony Ralston, SUNY at Buffalo; and A.K. Dewdney, University of Western Ontario and editor of the "Computer Recreations" column of *Scientific American*. There will be contributed paper sessions which should be suitable for a diverse audience of mathematicians, statisticians, and students. Abstracts should be sent by June 15, 1987 to Professor Zevi Miller, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056. Information regarding preregistration and housing may also be obtained from Professor Miller.

STUDENT CONFERENCE - CALL FOR PAPERS

The Ohio Delta Chapter of Pi Mu Epsilon will hold its thirteenth annual Student Conference October 9 and 10, 1987. Undergraduate mathematics and statistics students are invited to contribute papers and should send abstracts by September 25, 1987 to Professor Milton Cox, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056.

LETTERS TO THE EDITOR

Dear Editor:

I just received my copy of *Mathematics Magazine*, Vol. 60, No. 2 (April, 1987). While looking at the article by Danzer, Grünbaum and Shephard titled "Equitransitive Tilings, or How to Discover New Mathematics," I noticed an error. (It is a small error but I thought I should point it out anyway.)

Figure 10 gives an example of a tiling for each of the 17 wallpaper groups. On page 76, the tiling given for the group

pg is really a tiling for the group pgg . (The example actually given for pgg is correct. It is another tiling for pgg .)

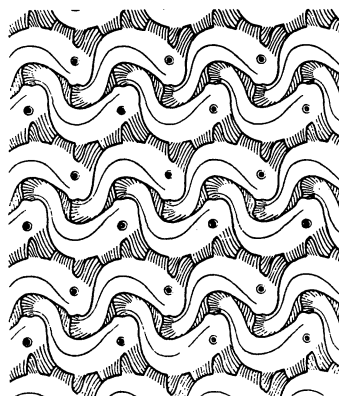
The tiling given for pg has two-fold rotational symmetry. For example, take as a center of rotational symmetry the center of one of the tiles. Also, the tiling has no reflections and perpendicular glide-reflection axes. Only the pgg wallpaper group has these properties.

Richard E. Stone
AT&T Bell Laboratories
Room HO 3K-328
Crawfords Corner Road
Holmdel, NJ 07733

Dear Editor:

Yes, Stone is right. We (in fact, I) goofed in selecting the tilings. The one purporting to illustrate the group pg is actually used in G.C. Shephard's and my "Tilings and Patterns" (Freeman 1986) to illustrate the group pgg . With apologies for the error, may I bring to your reader's attention the attached tiling with symmetry group pgg (checked and verified). [See below.] It is not a tiling by polygons, but it is somewhat remarkable since it was designed by the Viennese artist Koloman Moser about 1900, when M.C. Escher was 2 or 3 years old.

Branko Grünbaum
University of Washington GN-50
Seattle, WA 98195



JOURNAL OF THE INTERNATIONAL SOCIETY LEONARDO FOR THE ARTS SCIENCES AND TECHNOLOGY

Executive Editor: **Roger F. Malina**

Managing Editor: **Pamela Grant-Ryan**

Main Editorial Office: **LEONARDO**, 2020 Milvia, Suite 310, Berkeley, California 94704, U.S.A.

LEONARDO is a unique international journal for artists and others interested in the contemporary arts. Particularly concerned with the interaction between the arts, sciences and technology, LEONARDO has no restriction on artistic tendency, content or medium. We feature articles written by artists about their own work, discussions of new concepts, materials and techniques, and subjects of general artistic interest.

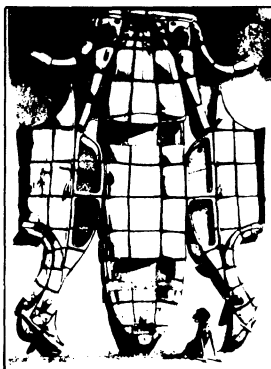
Forthcoming Special Issues

20TH ANNIVERSARY ISSUE—
"Art of the Future; the Future of Art"

Contributors include:

Arthur C. Clarke, Rudolf Arnheim, Frank Dietrich, Herbert Franke, Gyorgy Kepes, Benoit Mandelbrot, Cyril Stanley Smith and Murray Gell-Mann.

Photo: Dustin Shuler



VISUAL ART/SOUND/MUSIC/TECHNOLOGY

Contributors include:

Charles Ames,
Francois and Bernard Baschet,
Bill Fontana,
Larry Polanski,
Ernest Robson,
Lilliane Lijn,
Tom DeWitt
and Rudolf Arnheim.

Subscription Information

Published Quarterly 360 pp.
Volume 20, 1987
Individual subscription rate (1987) US\$ 40.00
Annual library subscription (1987) US\$170.00
Two-year library
subscription (1987/88) US\$323.00

Free Sample Copy Available Upon Request.

PERGAMON JOURNALS

USA, Central and South America:

Maxwell House, Fairview Park, Elmsford, New York 10523 Telephone: (914) 592-7700

UK and all other countries:

Headington Hill Hall, Oxford OX3 0BW, England Telephone: (0865) 64881

A member of the BPCC plc Group of Companies



The Mathematical Association of America CORPORATE MEMBERS

*The Mathematical Association of America gratefully acknowledges the
interest and support of the following:*

CORPORATE MEMBERS

Addison-Wesley Publishing Company, Inc.

American Mathematical Society, Inc.

Birkhäuser Boston, Inc.

Digital Equipment Corporation

International Business Machines Corporation

Macmillan Publishing Company

Marcel Dekker, Inc.

The Mitre Corporation

Pergamon Press, Inc.

Prindle Weber and Schmidt, Inc.

Saunders College Publishing, Inc.

Scott Foresman & Company, Inc.

Springer-Verlag New York, Inc.

John Wiley & Sons, Inc.

CONTRIBUTING CORPORATE MEMBER

Bell Laboratories, Inc.

Studies in Mathematical Economics

Volume 25 in the MAA Studies in Mathematics

Edited by Stanley Reiter

420 pp. Hardbound

ISBN-0-88385-027-X

List: \$42.00

MAA Member: \$31.00

*"For the mathematician desiring
to become familiar with modern
mathematical, microeconomic theory,
this volume is indispensable."*

Robert Rosenthal
SUNY, Stony Brook
Department of Economics

Stanley Reiter, as editor, has brought together a distinguished group of contributors in this volume, in order to give mathematicians and their students a clear understanding of the issues, methods, and results of mathematical economics. The range of material is wide: game theory; optimization; effective computation of equilibria; analysis of conditions under which economies will move to the greatest possible efficiency under various forces, and the requirements for the flow of information needed to achieve efficient markets.

The material is interesting at all mathematical levels. For example, the initial article shows how even mathematically simple, concrete, two-person, nonzero sum games present us with the complexities and dilemmas of choices in real life. At the other extreme, the final article, by Debreu, begins by using the power of Kakutani's fixed point theorem to prove the existence of economic equilibria. In between, the reader will find beautiful uses of calculus, topology, combinatorial topology, and other topics.

The chapters of this volume can be read independently, although they are related. The book begins with Meyerson's chapter on game theory and its theoretic foundations. The second chapter, by Simon, starts with the familiar criteria for maxima from calculus and goes on to develop more general tools of mathematical economics,

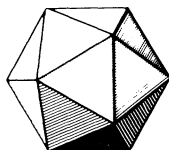
including the Kuhn-Tucker and related conditions. The third contribution, by Mas-Collell, uses the tools of differential topology, including Sard's theorem, to study the competitive equilibria of whole families of economies using a differentiable point of view. Next Kuhn, building on the work of Scarf, shows how methods based on Sperner's lemma can be used to compute equilibria.

The next two chapters by Reiter and Hurwicz explore the properties of systems that are not purely competitive. They bring analytical and topological tools to bear to determine what conditions on the exchange of information are needed to allow such markets to become optimally efficient.

Radner addresses one consequence of what Herbert Simon calls "bounded rationality." Managers neither know all the facts nor do they have unlimited ability to calculate. How should they allocate their time? The tools used to answer this question are fittingly probabilistic.

In the final chapter, Debreu gives four examples of mathematical methods in economics. These four examples alone give a sense of the breadth and nature of the field.

In this study, Reiter and his other contributors show the reader the subtlety and complexity of the subject along with the precision and clarity that mathematics bring to it.



ORDER FROM

The Mathematical Association of America
1529 Eighteenth Street, NW
Washington, DC 20036

for your library

**The William Lowell Putnam
Mathematical Competition, 1965-1984,**

compiled by Gerald L. Alexanderson,
Leonard Klosinski and Loren Larson
151 pp., Cloth, ISBN 0-88385-441-4
List: \$24.00 MAA Member: \$18.00

The Putnam Competition has since 1928 been providing a challenge to the gifted college mathematics student. This volume contains problems with their solutions, for the years 1965-1984. Corrections to the solutions have been made, and additional solutions are presented. Included is an essay on recollections of the first Putnam Exam by Herbert Robbins, as well as appendices listing the winning teams and students from 1965 through 1984.

The earlier collection of Putnam problems, **The William Lowell Putnam Mathematical Competition, 1938-1964** was a scholarly approach to the problems which gave many different solutions for each problem. The authors gave us a background, in depth, on each problem and showed us, whenever possible, how the problems stimulated further questions. The current volume of Putnam problems does not attempt to duplicate that effort, but it does offer the problem solver an enticing sample of challenging problems and their solutions.



Order from:
The Mathematical Association of America
1529 Eighteenth Street, NW
Washington, DC 20036

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036
MATHEMATICS MAGAZINE VOL. 60, NO. 3, JUNE 1987