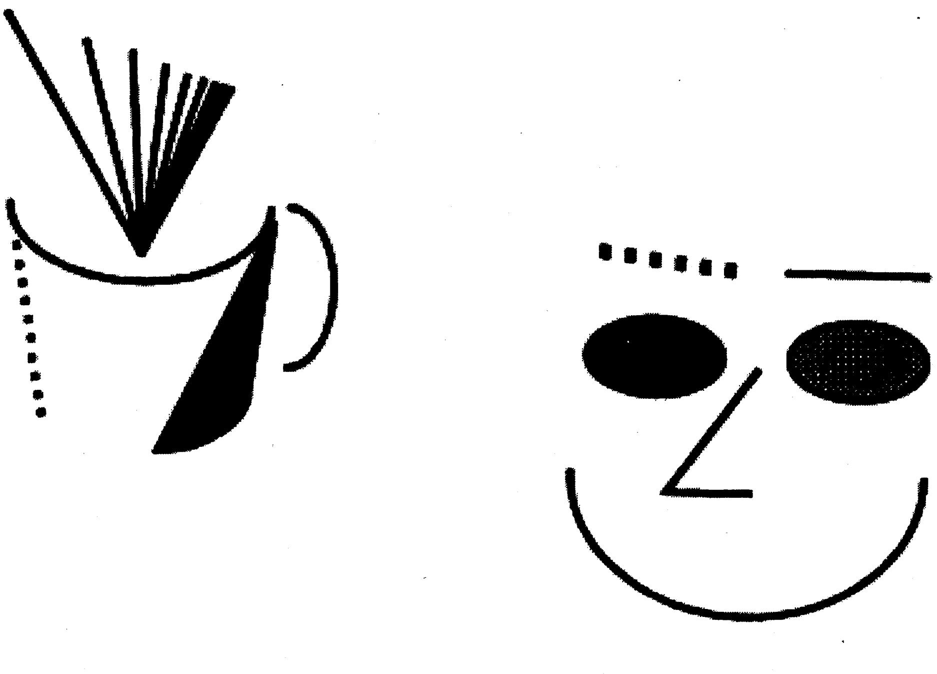
Vol. 71, No. 1 February 1998

MATHEMATICS MAGAZINE



Are they 2-homothetic?

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims 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. Manuscripts on history are especially welcome, as are those showing relationships between various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 71, pp. 76–78, and is available from the Editor. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Send new manuscripts to Paul Zorn, Editor, Department of Mathematics, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098. Manuscripts should be laser-printed, with wide line-spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should submit three copies and keep one copy. In addition, authors should supply the full five-symbol Mathematics Subject Classification number, as described in *Mathematical Reviews*, 1980 and later. Copies of figures should be supplied on separate sheets, both with and without lettering added.

AUTHORS

Robert M. Corless is the editor of the ACM SIGSAM Bulletin (Communications in Computer Algebra), and the author of *Essential Maple* (Springer-Verlag, 1994). He got his Ph.D. in Mechanical Engineering under the supervision of G.V. Parkinson, in the topic of mathematical modeling of flow-induced vibration. Since his Ph.D., he has "galloped madly off in all directions," to paraphrase Stephen Leacock (for a list of his favorite authors, see RMC's home page (at `http://pineapple.apmaths.uwo.ca/~rmc`), studying computer algebra, numerical analysis, dynamical systems, and special functions (including the Lambert W function, which, alas, forms no part of this present article). The present article was written in a perhaps misguided attempt at enriching first year calculus at the University of Western Ontario. While some students showed up for the extra talk, and seemed to enjoy it, enrichment is taking a different path this year, in that fewer lectures and more teamwork and writing will be tried. If all goes well, a future article will describe the success; failure will not be heard from.

Dick Darst received a B.S. from IIT in 1957 and a Ph.D. from LSU in 1960. After two years at MIT he was hired by Purdue, where lots of people were active in analysis. At IIT he enjoyed reading old issues of *Fundamenta* and working on analysis problems, so Purdue was a good place for him and he liked it. In 1971 he moved to CSU. When fractals became popular years later, numerous problems in analysis became "interesting" again. (At IIT he graded for Karl Menger, who told him about *Fundamenta*; Menger could have given a very insightful course in the 50s.) Darst enjoyed trying to figure out and explain the interplay among some basic ideas in algebra, analysis, and geometry involved in drawing the kinds of fractals described in this paper. He funds his own research institute at Crystal Lakes in the mountains.

Victor Klee's 1949 Ph.D. is from the University of Virginia, which attracted him because of an initial interest in point-set topology. While there, he became interested also in functional analysis and convex geometry. After the move to Seattle, in 1953, his interests broadened to include combinatorics, optimization, and computational complexity. These days, he likes to work in a variety of fields in order to spread his mistakes more thinly. He is a co-author, with Stan Wagon, of the MAA book *Old and New Unsolved Problems in Plane Geometry and Number Theory*, and he was MAA President in 1971–73.

William McWorter received his doctorate from Ohio State University in 1964. He taught first at the University of British Columbia. A bout of homesickness brought him back to Ohio State in 1966, where he remained, retiring in 1994. He likes abstract algebra, especially finite group theory, even though, as the applied mathematician Greg Forrest puts it, "pure mathematics has no bearing on the life of Joe Sixpack." The current article marks the first time he has published more papers after receiving his doctorate (4) than he published as a graduate student (3).

Judith Palagallo received her Ph.D. from Colorado State University, under the direction of Richard Darst. Since 1978 she has been at the University of Akron, where she met her husband and co-author Thomas Price. The present article in fractal geometry was written during a recent faculty leave they spent at CSU. She enjoys teaching analysis courses. Her hobbies include playing the piano, home decorating, and Italian cooking.

Thomas Price, a graduate of the University of Georgia, is a member of the Applied Mathematics Division of Mathematical Sciences at the University of Akron. He enjoys doing research in approximation theory and calculus reform. He has published papers with his wife, Judith Palagallo, but this article is the result of his first collaboration with Dick Darst. He welcomed this occurrence for many reasons, not the least of which was the reduction of his Erdős number to two.

John Reay studied music at Pacific Lutheran University and mathematics at the University of Washington, where Victor Klee directed his 1963 Ph.D. thesis. He now teaches at Western Washington University, and plays in the Whatcom Symphony Orchestra. This paper on 2-homothetic sets grew out of a talk he prepared for the Visiting Lecturer Program of the MAA, and the goading of friends who wanted a written version. The talk was based on earlier lectures of Klee.

Vol. 71, No. 1, February 1998

MATHEMATICS MAGAZINE

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 $16 included as part of the annual dues. (Annual dues for regular members, exclusive of annual subscription prices for MAA journals, are $64. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 40% dues discount for the first two years of membership.) The nonmember/library subscription price is $68 per year.

Subscription correspondence and notice of change of address should be sent to the Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

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

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

Postmaster: Send address changes to Mathematics Magazine, Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

PRINTED IN THE UNITED STATES OF AMERICA

ARTICLES

A Surprising but Easily Proved Geometric Decomposition Theorem

VICTOR KLEE
University of Washington
Seattle, WA 98195-4350

JOHN R. REAY
Western Washington University
Bellingham, WA 98225-9063

Introduction

The whole is the sum of its parts—what might those parts look like? If we have two very different-looking sets in the plane, when can their corresponding separate parts look alike? It is a question with some surprising answers.

In FIGURE 1, two closed sets A and B are composed of disjoint subsets—$A = A_1 \cup A_2$ and $B = B_1 \cup B_2$—in such a way that A_1 is similar to B_1 and A_2 is similar to B_2. For the "summands" to be truly disjoint, we must also account for the boundaries. To obtain the desired similarities, we assign the bottom edge of the square A_1 to the rectangle A_2 and the top edge of the square B_1 to the rectangle B_2. Could the same sort of decomposition be obtained if, say, the set A was replaced by a circular disk? A glance ahead to FIGURE 3 might affect your answer. And look at FIGURE 4—can each of those sets be partitioned into two disjoint subsets so that the corresponding parts of each set look alike? How would you bet?

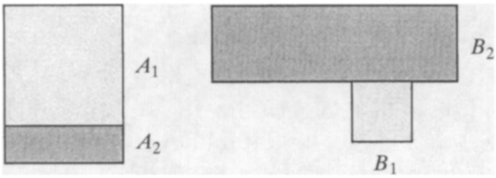


FIGURE 1

A remarkable result

Two sets A and B in the plane are *homothetic*, denoted $A \sim B$, if they are similar and similarly oriented. For example, in FIGURE 2, the sets A, B, and C are homothets of each other, but not of set D (even though D is congruent to A) because "similarly oriented" does not permit rotations or reflections. Thus in Figure 1, with A_1 missing its bottom edge and B_1 missing its top edge, the sets A_1 and B_1 are similar but they are not homothetic because the similarity mapping A_1 onto B_1 involves a 180° rotation. A *homothetic transformation* (or *homothety*) of the plane onto itself is a mapping of the form $f(\mathbf{v}) = k\mathbf{v} + \mathbf{a}$, where $\mathbf{a}$ is a constant vector and k is a positive scalar constant. When $k = 1$, f is a *translation*. When $\mathbf{a} = 0$ and $k = 1$, f is the identity mapping. When $\mathbf{a} = 0$ and $k \neq 1$, f is a *contraction* toward the origin or an *expansion* about the origin, according as $k < 1$ or $k > 1$. When $k \neq 1$, we may set $m = 1/(1-k)$ and note that

$$f(m\mathbf{a} + (\mathbf{v} - m\mathbf{a})) = f(\mathbf{v}) = k\mathbf{v} + \mathbf{a} = m\mathbf{a} + k(\mathbf{v} - m\mathbf{a}),$$

thus representing f as a contraction toward or expansion about the point m**a**.

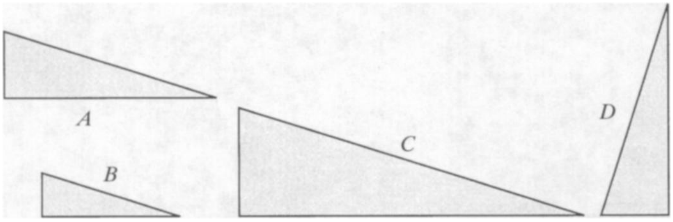


FIGURE 2

Using the definition, the reader will readily verify that the composition of two homotheties is again a homothety, that the inverse of a homothety is a homothety, and that each line is mapped by a homothety onto a parallel line.

We will say that two sets A and B are *2-homothetic*, denoted $A \approx B$, if each of them can be partitioned into two disjoint sets ($A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ with $A_1 \cap A_2 = \varnothing = B_1 \cap B_2$) in such a way that $A_1 \sim B_1$ and $A_2 \sim B_2$.

In Figure 1, if square B_1 were on top of rectangle B_2 rather than below, then A and B would be 2-homothetic, since the bottom edges of squares A_1 and B_1 could be assigned to A_2 and B_2 respectively, and then no forbidden rotation would be needed to establish the similarities. But when B_1 is tacked onto the bottom of B_2, as in Figure 1, it becomes an interesting exercise to try to show that A and B are 2-homothetic by finding the required partitions, remembering to take care of the boundaries.

Another example is found in Figure 3, which suggests an infinite nesting of inscribed squares and disks that might show the square and the disk to be 2-homothetic!!! Of course, we must always be careful of what is happening on the boundaries of the subsets. Is it really true that a square and a disk can be built from the same two pieces if we are allowed just expansions and contractions?

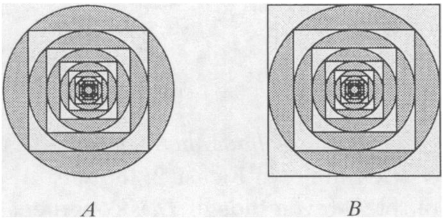


FIGURE 3

It is certainly not obvious that the two sets in FIGURE 4 are 2-homothetic, since the sets include isolated points, whiskers, random curves, components that may not be Lebesgue measurable (the shaded eye in B), and are generally as badly behaved as we could draw them. However, their 2-homotheticity is a consequence of the following remarkable result.

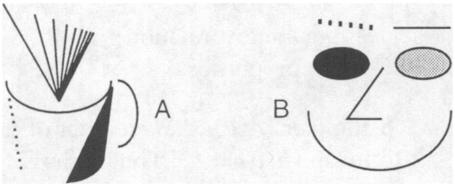


FIGURE 4

THEOREM 2HOM. *Two sets in the plane are 2-homothetic provided each of them is bounded and has nonempty interior.*

Although Theorem 2HOM seems surprising, it turns out to be an easy corollary of the following strengthened form of the famous Cantor–Bernstein theorem, and thus is a nice example to show the geometric power of abstract set theory.

THEOREM CBB. *If f: $A \to B$ is a function that maps a set A one-to-one* ***into*** *a set B (i.e., onto a subset of B) and g: $B \to A$ is a function that maps B one-to-one* ***into*** *A, then there are partitions $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ such that $f(A_1) = B_1$ and $g(B_2) = A_2$. Setting $h(a) = f(a)$ for all $a \in A_1$, and $h(a) = g^{-1}(a)$ for all $a \in A_2$, we have a one-to-one mapping h of A* ***onto*** *B.*

Proof of Theorem 2HOM. Suppose that A and B are both bounded, and each has an interior point. Since A has an interior point, A contains an entire circular disk C, and since B is bounded, a sufficiently great expansion of C about its center produces a larger disk D that contains B. The inverse of this expansion is a contraction (hence a homothety) that maps B into A. Similarly, there is a contraction that maps A into B. Since these contractions are clearly one-to-one, an application of Theorem CBB immediately yields the stated conclusion. ∎

Under the hypotheses of Theorem 2HOM, there are infinitely many contractions that pull set A into set B, and infinitely many that pull B into A, so there are infinitely many partitions $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ for showing that A and B are 2-homothetic. Nevertheless, it is an interesting exercise to try to draw (or even imagine) such a partition in specific cases such as the one provided by FIGURE 4.

The original Cantor–Bernstein theorem asserts the existence of a one-to-one mapping h of A onto B, without specifying the relationship of h to the original mappings f and g. According to Fraenkel **[8]**, the stronger form stated above is due to Banach **[1]**, so we think of it as the Cantor–Bernstein–Banach (CBB) theorem. (The name of Schröder is often associated with the Cantor–Bernstein theorem. However, according to **[8]**, the theorem was conjectured by Cantor, the first complete published proof was due to Bernstein, and an independent proof of Schröder turned out to be defective.) See **[5]** for an extension of the CBB theorem.

The first explicit statement of Theorem 2HOM may have been the one in **[12]**, but Banach in **[1]** had already mentioned the possibility of geometric applications of the CBB theorem, and Theorem CBB was used in **[2]** to establish the famous Banach–Tarski paradox (see (7) below).

Two proofs of the CBB theorem

With such a strong corollary, you might expect that CBB has a difficult proof, but the classic proof of Banach [1] (found also in Birkhoff and MacLane [4]) is short and easy. It is the second proof below. Another nice proof of the CBB theorem uses a fixed-point theorem of Birkhoff [3]. To set this up, we need a quick review of complete lattices.

A *partial order* for a set S is a binary relation $\leq$ on S (i.e., a subset of the Cartesian product $S \times S$) with these properties:

1) *Reflexivity*: For each $a \in S$ the pair (a, a) is an element of the subset $\leq$ of $S \times S$. (We usually write $a \leq b$ to mean $(a, b) \in \leq$. Thus reflexivity is the condition that $a \leq a$ for all $a \in S$.)
2) *Anti-symmetry*: If $a \leq b$ and $b \leq a$ then $a = b$.
3) *Transitivity*: If $a \leq b$ and $b \leq c$ then $a \leq c$.

The pair $(S, \leq)$ is called a *partially ordered set*, or *poset*. For example, the real numbers form a poset with their usual ordering. But the reals have the additional property that every two elements are comparable, and hence we say that they form a *totally ordered set*. However, in posets it may happen that two elements a and b are not comparable—i.e., neither $a \leq b$ nor $b \leq a$ is true.

An element $s \in S$ is a *lower bound* for the set $T \subseteq S$ if $s \leq t$ for each $t \in T$. Similarly $u \in S$ is an *upper bound* for T if $t \leq u$ for each $t \in T$. The (necessarily unique) *least upper bound* for a subset T is an upper bound m for T such that $m \leq u$ for every upper bound u. *Greatest lower bounds* are similarly defined. A *lattice* is a non-empty poset in which each set of two elements (and hence each nonempty finite subset) has a least upper bound and a greatest lower bound. A *complete lattice* is a lattice in which every nonempty subset has a least upper bound and a greatest lower bound. The upper bound for the whole set S is usually denoted 1 and the lower bound for S is denoted 0.

Some examples might help.

Example 1. Let L denote the integer lattice in the Cartesian plane—the set of all points with both coordinates integers. If we define $(x, y) \leq (u, v)$ to mean $x \leq u$ and $y \leq v$ (in the usual sense) then $(L, \leq)$ is a poset. Some pairs of points, such as $(5, 8)$ and $(9, 6)$, are not comparable. But the point $(5, 6)$ is the greatest of all their lower bounds. The finite part of L shown in FIGURE 5 is a complete lattice. The upper right point is the upper bound and the lower left point is the lower bound for the whole subset shown. However, the infinite set L is a lattice but not a complete lattice.

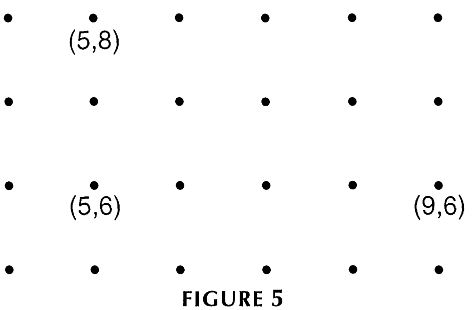


FIGURE 5

Example 2. Let $\mathscr{P}(R^2)$ denote the collection of all subsets of the plane R^2. Then $(\mathscr{P}(R^2), \subseteq)$ is a complete lattice. For any nonempty collection $\mathbb{C}$ of elements of $\mathscr{P}(R^2)$, the least upper bound (resp. greatest lower bound) of $\mathbb{C}$ is the union (resp. intersection) of all elements of $\mathbb{C}$.

When a function f maps a set S into itself, a point $a \in S$ is a *fixed point* for f if $f(a) = a$. Fixed-point theorems are among the most interesting and useful tools in mathematics. Theorem FP below is an all-time favorite that will be used to give a proof of the CBB theorem. A mapping f of a poset $(S, \leq)$ into a poset $(W, \preccurlyeq)$ is *order-preserving* if $x \leq y$ in S implies $f(x) \preccurlyeq f(y)$ in W.

THEOREM FP [3]. *Every order-preserving function f of a complete lattice $(S, \leq)$ into itself has a fixed point.*

Proof of Theorem FP. Let $T = \{a \in S \mid a \leq f(a)\}$. Clearly $0 \in T$ so $T \neq \varnothing$. Let m be the least upper bound of T. Since $t \leq m$ for every $t \in T$, and f is order-preserving, $t \leq f(t) \leq f(m)$, so $f(m)$ is also an upper bound of T. Hence $m \leq f(m)$ because m is the least upper bound of T. Thus $f(m) \leq f(f(m))$, so $f(m) \in T$ and $f(m) \leq m$. Since $m \leq f(m)$ and $f(m) \leq m$, it follows from anti-symmetry that $f(m) = m$ and m is the desired fixed point. ■

Fixed-point Proof of the CBB Theorem. Assuming without loss of generality that the sets A and B are disjoint, we will use the given one-to-one into functions $f\colon A \to B$ and $g\colon B \to A$ to define a function φ from the complete lattice $(\mathscr{P}(A), \subseteq)$ into itself. For each subset of A, let $C' = \{a \in A \mid a \notin C\}$ denote the *complement* of C in A. Similarly if $D \subseteq B$ let D' denote the complement of D in B. Then for each $C \subseteq A$ define $\varphi(C) = g((f(C'))')$. That is, we take the complement of C in A, map it into B by f, take the complement in B, and map this complement back into A by g. Since $C_1 \subseteq C_2$ implies $f(C_1) \subseteq f(C_2)$ and $C_1' \supseteq C_2'$, it is easily seen that φ is an order-preserving mapping of $\mathscr{P}(A)$ into itself. Hence by Theorem FP, φ has a fixed point. Call this fixed point A_2, set $A_1 = A_2'$, and set $B_1 = f(A_1) = B_2'$. Then the restrictions $f\colon A_1 \to B_1$ and $g\colon B_2 \to A_2$ are one-to-one and onto, and the partitions $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ are the ones desired for the CBB Theorem. ■

Classic Proof of the CBB Theorem [1, 4]). We again assume that the sets A and B are disjoint. A point $x \in A \cup B$ is a *parent* of a point $y \in A \cup B$ if $x \in A$ and $f(x) = y$, or $x \in B$ and $g(x) = y$. Since A and B are disjoint and the mappings f and g are one-to-one, each point of $A \cup B$ has at most one parent. That parent (if it exists) has at most one parent, etc. This sequence of parents forms the *ancestral chain* of y. The sequence may be empty, as would be the case if $y \in B \setminus f(A)$ or $y \in A \setminus g(B)$. It may be infinite, as would be the case if $y = g(f(y))$ or $y = f(g(y))$. If the ancestral chain is neither empty nor infinite, it terminates in a point that has no parent. (See FIGURE 6).

Now let A_{even}, A_{odd}, and A_∞ denote the points of A for which the length of the ancestral chain is respectively even, odd, or infinite. This partitions A, and B has a similar partitioning. It is clear that f maps A_∞ into B_∞, A_{even} into B_{odd}, and A_{odd} into B_{even}. Further, since each point of $B_\infty \cup B_{\text{odd}}$ has a parent, the first two mappings are onto; that is, $f(A_\infty) = B_\infty$ and $f(A_{\text{even}}) = B_{\text{odd}}$. Similarly, $g(B) = A_\infty$ and $g(B_{\text{even}}) = A_{\text{odd}}$. Setting

$$A_1 = A_{\text{even}} \cup A_\infty, \quad A_2 = A_{\text{odd}}, \quad B_1 = B_{\text{odd}} \cup B_\infty, \text{ and } B_2 = B_{\text{even}},$$

we have the partitions whose existence is asserted by the CBB Theorem. ■

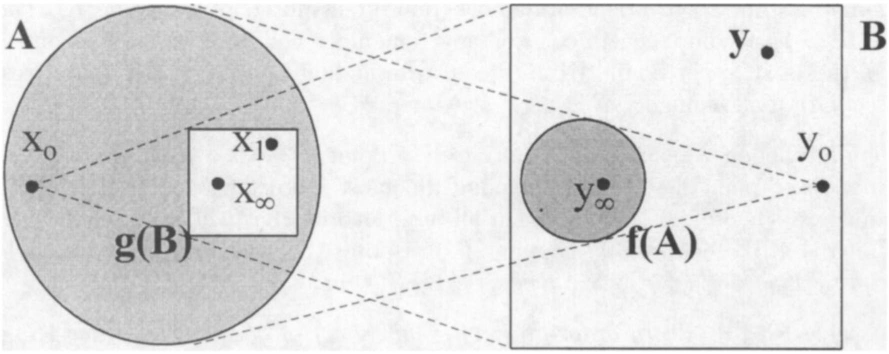


FIGURE 6

EXAMPLE 3. In FIGURE 6, the contraction f about the point y_o in the interior of B maps set A homothetically and one-to-one into set B. Similarly, the contraction g about the point $x_o \in A$ is a homothety which maps B one-to-one into A. Clearly, each of the points x_o, y_o, and y is an *orphan* (i.e., has no parent). Thus the ancestral chain of $x_1 = g(y)$ is just $\{y\}$, of (odd) length 1. Since $x_\infty = g(y_\infty) = g(f(x_\infty)$, the ancestral chain of $y_\infty \in B$ is $\{x_\infty, y_\infty, x_\infty, y_\infty, \ldots\}$, of infinite length.

Remarks and open problems

1) The setting for Theorem 2HOM was the plane R^2, but the definitions (2-homothetic, bounded, interior) and the proof of Theorem 2HOM are all valid in an arbitrary (even infinite-dimensional) normed vector space.
2) When two subsets A and B of d-space are 2-homothetic and are both geometrically "nice" in some sense, it is interesting to ask how nice their summands (the sets A_1, A_2, B_1, B_2 in the partitions) can be made. Of course, niceness is in the eye of the beholder, and in any case the answer must depend on geometric or topological properties of the sets A and B. In particular, if the set A is connected, then it is impossible for A_1 and A_2 both to be closed (or both to be open) relative to A unless one of A_1 or A_2 is empty. However, one might hope to have A_1 closed and A_2 open relative to A, and then of course B_1 closed and B_2 open relative to B. FIGURE 7 shows that this can happen in some cases. In FIGURE 7a, the sets A and B are both bounded and convex, but neither is compact. In FIGURE 7b, the sets A and B are both compact, but neither is convex. However, it seems that the following problems are open for each $d > 2$:
   (a) Is there an example of two d-dimensional compact convex subsets A and B of d-space such that A and B are not homothetic but they are 2-homothetic by means of convex summands, $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$, in such a way that the sets A_1 and B_1 are not only convex but also closed?
   (b) If A and B are both d-dimensional compact convex sets in Euclidean d-space, must they be 2-homothetic by means of summands A_i and B_i that are connected?

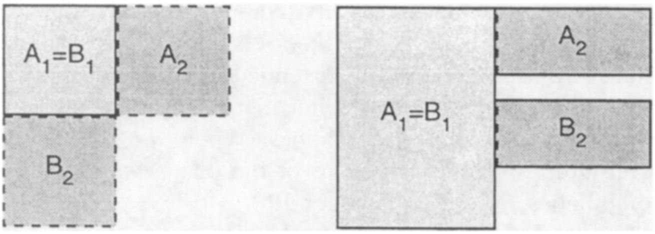


FIGURE 7
In both cases, A_1 and B_1 are closed relative to the sets $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ respectively. In 7a, A and B are convex but not compact, and in 7b they are compact but not convex.

3) In connection with problem 2(a), note that if A and B are compact subsets of d-space, each with nonempty interior, and $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ are the partitions constructed in the proof of Theorem 2HOM, then each A_i and each B_i is both an F_σ-set (the union of countably many closed sets) and a G_δ-set (the intersection of countably many open sets). This follows from Banach's proof of the CBB Theorem. For let f and g be homotheties which, respectively, carry A into B and B into A. Define $A_0 = A$, $B_0 = B$, and having defined A_i and B_i, set $A_{i+1} = g(B_i)$ and $B_{i+1} = f(A_i)$. Then each A_i and each B_i is compact, hence is a G_δ set, and

$$A_0 \supseteq A_1 \supseteq \cdots, \qquad B_0 \supseteq B_1 \supseteq \cdots .$$

It follows that each set $A_i \setminus A_{i+1}$ is σ-compact, as is each set $B_i \setminus B_{i+1}$. Now define

$$\begin{aligned} A_{\text{even}} &= (A_0 \setminus A_1) \cup (A_2 \setminus A_3) \cup \cdots \cup (A_{2j-2} \setminus A_{2j-1}) \cup \cdots \\ A_{\text{odd}} &= (A_1 \setminus A_2) \cup (A_3 \setminus A_4) \cup \cdots \cup (A_{2j-1} \setminus A_{2j}) \cup \cdots \\ A_\infty &= A_0 \cap A_1 \cap \cdots , \end{aligned}$$

and

$$\begin{aligned} B_{\text{even}} &= (B_0 \setminus B_1) \cup (B_2 \setminus B_3) \cup \cdots \cup (B_{2j-2} \setminus B_{2j-1}) \cup \cdots \\ B_{\text{odd}} &= (B_1 \setminus B_2) \cup (B_3 \setminus B_4) \cup \cdots \cup (B_{2j-1} \setminus B_{2j}) \cup \cdots \\ B_\infty &= B_0 \cap B_1 \cap \cdots . \end{aligned}$$

Then each of the sets A_∞ and B_∞ is compact, each of the sets A_{even}, A_{odd}, B_{even}, B_{odd} is σ-compact, and we have already seen that the desired partition is obtained by setting

$$A_1 = A_{\text{even}} \cup A_\infty, \quad A_2 = A_{\text{odd}}, \quad B_1 = B_{\text{odd}} \cup B_\infty, \quad \text{and} \quad B_2 = B_{\text{even}} .$$

Since the disjoint sets A_1 and A_2 are both F_σ-sets and their union is the compact set A, A_1 and A_2 are both also G_δ-sets. Similarly, B_1 and B_2 are both F_σ-sets and G_δ-sets.

4) It is an easy exercise to show that for any two homotheties f and g, the *commutator* $fgf^{-1}g^{-1}$ is merely a translation. Thus, although the group of

homotheties is not commutative, its first commutator subgroup is commutative. This (the fact that the group of homotheties is *solvable*) is a key to showing that Lebesgue measure in d-space can be extended to a finitely additive measure that is defined for all bounded sets and is not merely invariant under translation but multiplies properly under all homotheties. When $d = 2$, a similar conclusion applies to the group of transformations of the plane generated by the rotations and the homotheties. (See **[20]**, Chapter 10.)

5) It is easy to see that the homothety relation $\sim$ is reflexive, symmetric, and transitive. In particular, if $B = kA + \mathbf{a}$ and $C = mB + \mathbf{b}$, then $C = (km)A + (m\mathbf{a} + \mathbf{b})$, so $A \sim C$. The 2-homothety relation $\approx$ is reflexive and symmetric, but it is not transitive. FIGURE 8 shows sets A, B, and C, made up of parallel half-open intervals in the plane, with $A \approx B$ and $B \approx C$, but it is not true that $A \approx C$.

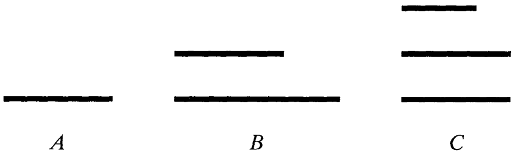


FIGURE 8

6) For any integer r with $2 \leq r \leq |A| = |B|$ we may define sets A and B to be *r-homothetic* in the obvious way: there exist partitions $A = A_1 \cup \cdots \cup A_r$ and $B = B_1 \cup \cdots \cup B_r$ and homotheties $f_i(\mathbf{x}) = k_i\mathbf{x} + \mathbf{a_i}$ such that $f_i(A_i) = B_i$ for each i. If, in addition, each A_i and each B_i has at least two points and the scalars $k_1, \ldots, k_r$ are all different, we say that the sets A and B are *nontrivially r-homothetic*. In FIGURE 8, A is nontrivially 3-homothetic to C but A and C are not 2-homothetic. Other aspects of r-homothety make easy exercises.

7) A new family of problems arises when the group of homothetic transformations is replaced by some other group of transformations such as the rigid motions. The most famous result in this direction is the *Banach–Tarski paradox* **[2]**, asserting that if $d \geq 3$ and A and B are subsets of d-space each of which is bounded and has nonempty interior, then A and B are *equidecomposable* in the sense that for some finite n, A can be partitioned into n sets $A_1, \ldots, A_n$ and B can be partitioned into n sets $B_1, \ldots, B_n$ such that A_i is congruent to B_i for $1 \leq i \leq n$. See **[18]** and **[9]** for expositions of some aspects of the Banach–Tarski result, and see Wagon's book **[20]** for an extensive study of the "paradox" and related material.

8) In connection with the questions in 2), see **[17]** and **[11]** for some results and problems that involve decomposing two convex sets into a finite number of respectively congruent convex parts. And see **[6]** for a proof that in partitioning a ball of unit radius (in 3-space) into five sets that can be rearranged to form a partition of the union of two such balls, it can be arranged that each of the five sets is both connected and locally connected (of course, they cannot all be measurable).

9) Because of the measure-extension result mentioned in 4), if two subsets of the plane are both bounded and Lebesgue measurable, they cannot be equidecomposable unless they have the same measure. In 1925, Tarski **[19]** posed the

following modern version of the problem of squaring the circle: If D is a circular disk and S is a square of the same area, are D and S equidecomposable? Dubins, Hirsch, and Karush **[7]** showed that a circle and a square cannot be decomposed into respectively congruent parts that could (intuitively speaking) be cut out with a pair of scissors. However, Tarski's question did not restrict the nature of the sets in the partitions, and a brilliant affirmative solution to the question was given by Laczkovich **[15]** in 1990. His partitions involve a very large number of sets, but he requires only translations rather than the full group of rigid motions to move these sets from a disk-filling position to a square-filling position. For an excellent exposition of his work, see the article by Gardner and Wagon **[10]**. See also **[14]** and the 1994 survey article by Laczkovich **[16]**.

10) Even though Theorem CBB has the remarkable decomposition result Theorem 2HOM as an easy consequence, neither proof of CBB used the axiom of choice. This is in contrast to the situation for the measure-extension result mentioned in 4), for the Banach–Tarski paradox in 7), and for the theorem of Laczkovich in 9).

Acknowledgment. We are indebted to Jack Robertson for suggesting that we write this article and to him, Stan Wagon, Richard Gardner, and Kevin Short for helpful comments. A shorter version appeared in [13].

REFERENCES

1. S. Banach, Un theorème sur les transformations biunivoque, *Fund. Math.*, **6** (1924), 236–239.
2. S. Banach and A. Tarski, Sur la decomposition des ensembles de points en parties respectivement congruents, *Fund. Math.*, **6** (1924), 244–277.
3. G. Birkhoff, *Lattice Theory, revised ed.*, Amer. Math. Soc. Coll. Publ. 25, 1967.
4. G. Birkhoff and S. MacLane, *A Survey of Modern Algebra, 4th ed.*, Macmillan, New York, NY, 1977.
5. R. Brualdi, An extension of Banach's mapping theorem, *Proc. Amer. Math. Soc.*, **20** (1969), 520–526.
6. T. J. Dekker and J. de Groot, Decompositions of a sphere, *Fund. Math.*, **43** (1956), 185–194.
7. L. Dubins, M. Hirsch, and J. Karush, Scissor congruence, *Israel J. Math.*, **1** (1963), 239–247.
8. A. Fraenkel, *Abstract Set Theory, 4th revised ed.*, Amer. Elsevier Pub., Boston, MA, 1976.
9. R. M. French, The Banach–Tarski theorem, *Math. Intelligencer*, **10** (no. 4) (1988), 21–28.
10. R. J. Gardner and S. Wagon, At long last, the circle has been squared, *Amer. Math. Soc. Notices*, **36** (1989), 1338–1343.
11. R. J. Gardner, A problem of Sallee on equidecomposable convex bodies, *Proc. Amer. Math. Soc.*, **94** (1985), 329–332.
12. V. Klee, Some unsolved problems in plane geometry, this MAGAZINE, **52** (1979), 131–145.
13. V. Klee and J. Reay, 2-homothets, complete lattices, & other nice stuff, *Washington State University Mathematics Notes*, **39** (no. 1) (1996), 1–4.
14. V. Klee and S. Wagon, *Old and New Unsolved Problems in Plane Geometry and Number Theory*, Dolciani Mathematical Expositions - No. 11, Math. Assoc. of Amer., Washington, DC, 1991.
15. M. Laczkovich, Equidecomposability and discrepancy, A solution of Tarski's circle-squaring problem, *J. reine u. angewandte Mathematik*, **404** (1990), 77–117.
16. M. Laczkovich, Paradoxical decompositions: a survey of recent results, *First European Congress of Mathematics, Vol. II (Paris* 1992) 159–184, Progr. Math., **120** Birkhäuser, Basel, 1994.
17. G. T. Sallee, Are equidecomposable plane convex sets convex equidecomposable?, *Amer. Math. Monthly*, **76** (1969), 926–927.
18. K. Stromberg, The Banach–Tarski paradox, *Amer. Math. Monthly*, **86** (1979), 151–161.
19. A. Tarski, Problème 38, *Fund. Math.*, **7** (1925), 3817.
20. S. Wagon, *The Banach–Tarski Paradox*, Cambridge Univ. Press, New York, NY, 1985.

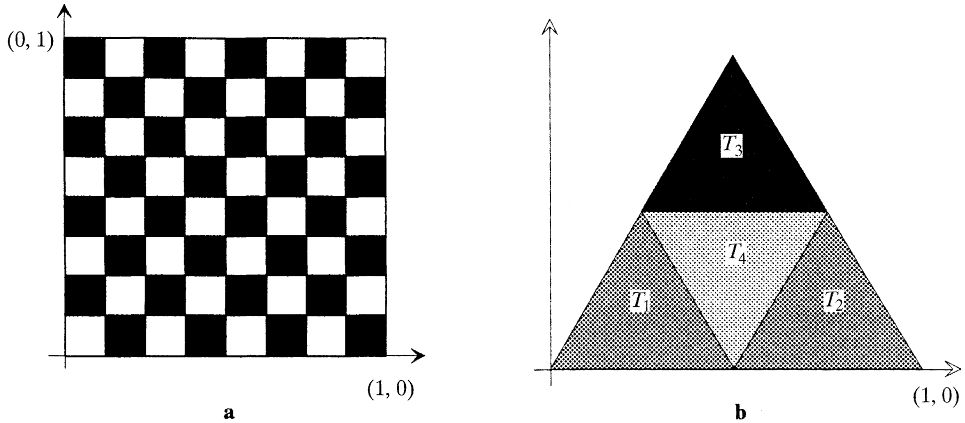
Fractal Tilings in the Plane

RICHARD DARST
Colorado State University
Fort Collins, CO 80523

JUDITH PALAGALLO*
THOMAS PRICE*
University of Akron
Akron, OH 44325

Introduction

Tilings have appeared in human activity since prehistoric times. They are used in the design of floor and wall coverings for cathedrals, commercial buildings, and personal dwellings. Mathematicians study the geometric structure of tilings. A checkerboard is an elementary example of a *similarity tiling*, one that is composed of smaller tiles (*rep tiles*) of the same size, each having the same shape as the whole. Each rep tile in the checkerboard is the scaled and translated image of the entire board. For the checkerboard in FIGURE 1a, the lower left tile is the image of the checkerboard under the mapping



a
Checkerboard tiling

b
Triangular tiling

FIGURE 1

$$f := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} \frac{1}{8} & 0 \\ 0 & \frac{1}{8} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

where $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ denotes any point on the checkerboard. The whole checkerboard can be formed using translates of this smaller image. One goal of this paper is to investigate the properties of linear mappings that generate similarity tilings.

[*] Research was conducted while on professional leave from the University of Akron.

In a more general setting the mappings may also involve rotations. For instance, the equilateral triangle T in FIGURE 1b is a similarity tiling since it is composed of four smaller equilateral triangles T_1, T_2, T_3, T_4. Consider the mappings defined by

$$f_1 := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$f_2 := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} + \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$f_3 := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} 0.25 \\ \sqrt{3}/4 \end{bmatrix} + \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$f_4 := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} 0.75 \\ \sqrt{3}/4 \end{bmatrix} + \begin{bmatrix} -.5 & 0 \\ 0 & -.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Then $T_j = f_j(T)$, $j = 1, \dots, 4$. Notice that the matrices "shrink" the tiling T to a rep tile, and the last matrix also performs a rotation. Adding the fixed vectors to the corresponding rep tiles translates them to their appropriate locations.

The checkerboard and triangle tiles have straight edges. In this paper we are interested in generating tilings with tiles (*fractiles*) whose boundaries are fractal curves. (Various definitions of fractal curves are given in **[2]**, **[11]**, and **[12]**. However, as Barnsley states in [**2**, p. 33], fractals are best explained by the many pictures and contexts that refer to them.) We will use an iterative process, involving repeated compositions of two or more functions, to generate these fractal tilings. The functions are constructed from translates of the inverse of a linear transformation $g(z) = Mz$, where M is an invertible 2×2 matrix with integer entries. (Geometric properties used later in this article require that M be an integer matrix.) The inverse transformation $g^{-1}(z)$ is the "shrinking function" that maps the entire fractal tiling onto a fractile. Before outlining the underlying mathematics, we briefly describe the basic algorithm and illustrate it with some examples.

Examples of fractal tilings

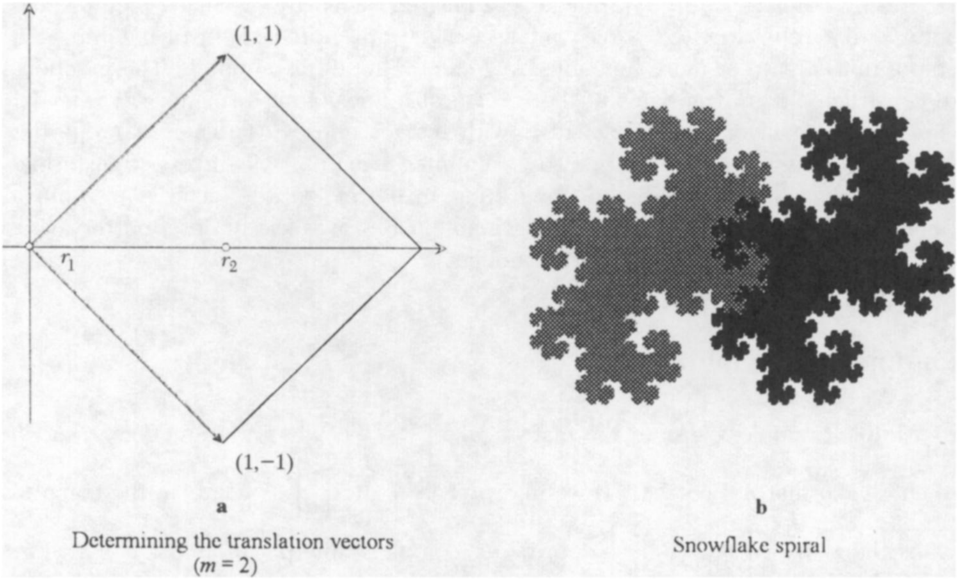
As a simple example, consider the matrix $M = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, where the integers a and b are chosen so that $a^2 + b^2 > 1$. If we interpret $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\begin{bmatrix} a \\ b \end{bmatrix}$ as points in the complex plane, then $M\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} ax_1 - bx_2 \\ ax_2 + bx_1 \end{bmatrix}$ represents complex multiplication of $x_1 + ix_2$ by $a + ib$. Next, we choose a collection of vectors to translate copies of the fractile so that they are positioned correctly in the tiling. Notice that the unit square, determined by the vectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, is mapped by M onto the square S, of area $m = a^2 + b^2$, spanned by vectors $\mathbf{v}_1 = \begin{bmatrix} a \\ b \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} -b \\ a \end{bmatrix}$. Define the set $\mathscr{E} = \{\mathbf{r}_j\}$ of vectors with integer coordinates that lie in or on S but not on the two outer edges that do not have the origin as a vertex. Then $\mathscr{E}$ contains exactly m vectors; we will use them for the translation vectors $\{\mathbf{r}_j\}$. The tiling can be drawn by a computer using the iterative procedure illustrated in Examples A and B.

Example A. Let $M = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$; then $m = 2$. From the square in FIGURE 2a we determine the two translation vectors $\mathbf{r}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\mathbf{r}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Then $\mathscr{E} = \{\mathbf{r}_1, \mathbf{r}_2\}$, and for $z = (x_1, x_2)$, we define the mappings $f_j(\mathbf{z}) := \mathbf{r}_j + M^{-1}(\mathbf{z})$ for $j = 1, 2$. That is,

$$f_1 := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} .5 & -.5 \\ .5 & .5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$f_2 := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} .5 & -.5 \\ .5 & .5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

To initiate the iteration process we randomly choose any point $\mathbf{z}_0$ in the plane and evaluate $f_1(\mathbf{z}_0)$ and $f_2(\mathbf{z}_0)$. Then for $n \geq 1$, we choose recursively and randomly $\mathbf{z}_n \in \{f_1(\mathbf{z}_{n-1}), f_2(\mathbf{z}_{n-1})\}$. After a few iterations the generated points lie near the tiling. So for $n \geq 100$, plot the points as they are generated. FIGURE 2b shows the result of several thousand iterations. (Increasing the number of iterations may improve the quality of the computed image. Special purpose software for drawing fractal tilings is described at the end of this paper.) The boundary of the spiral, the snowflake curve, is an example of a fractal curve. Mandelbrot [12] showed that the distance along the boundary between any two points is infinite.



Determining the translation vectors ($m = 2$)

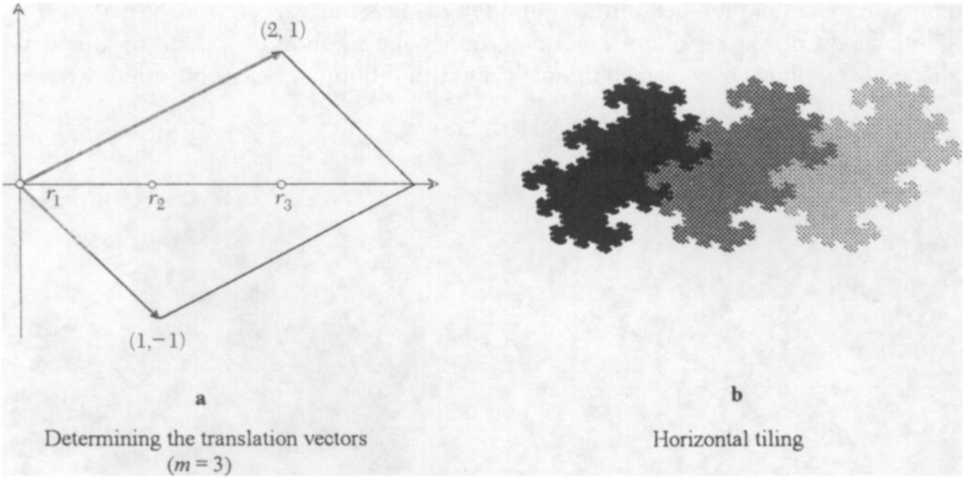
Snowflake spiral

FIGURE 2

The collection of functions $\{f_j\}$ is called an *iterated function system*. The set A of randomly generated points that results from this process is called the *attractor*. Notice that $A = A_1 \cup A_2$, where $A_j = f_j(A)$ and the A_j's have disjoint interiors.

Using M from Example A, the reader can explore the four different tilings generated when $\mathbf{r}_2$ is replaced by any of $\begin{bmatrix} -1 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ -1 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, and $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Observe that the shape of the tiling may change with the choice of translation vector.

Example B. Let $M = \begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}$. As shown in the parallelogram in FIGURE 3a, let $\mathbf{r}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{r}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{r}_3 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$. This choice of vectors produces a tiling with the three tiles stacked horizontally as shown in FIGURE 3b. As an exercise, we recommend that the reader choose translation vectors and generate a tiling for the matrix $M = \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix}$.



Determining the translation vectors ($m = 3$)

Horizontal tiling

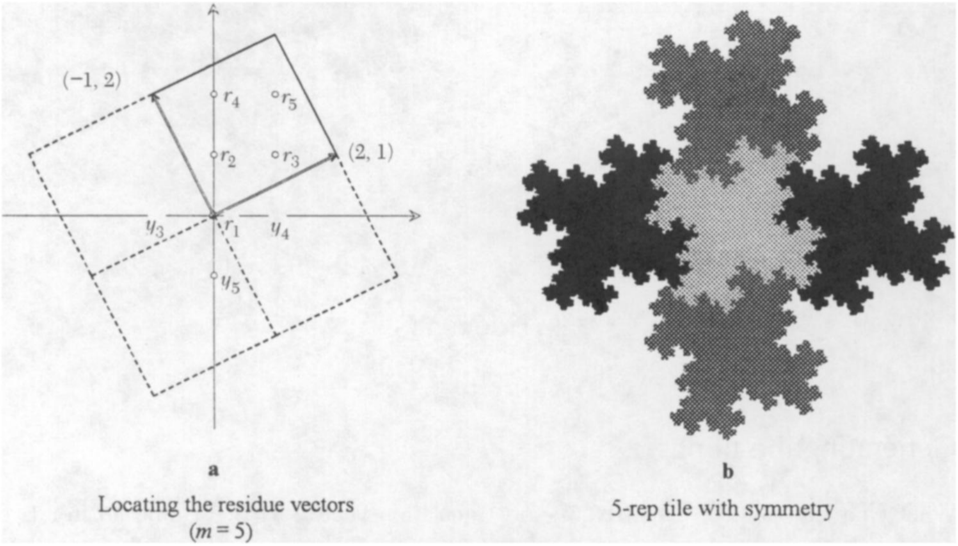
FIGURE 3

Generating the tilings

What characteristics of a matrix M and translation vectors $\mathbf{r}_j$ determine an iterative process that produces a desired tiling? We want the invertible integer matrix $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to be an *expansive map*; i.e., all the eigenvalues of M have modulus larger than 1. A property of expansive maps is that, for some $n > 0$, M^{-n} is a *contraction mapping*; i.e., $|M^{-n}\mathbf{z}| < |\mathbf{z}|$. This ensures that iteration using the collection of functions $f_j = \mathbf{r}_j + M^{-1}\mathbf{z}$, $j = 1, \dots, m$, produces the attractor regardless of the choice of translation vectors $\{\mathbf{r}_j\}$ or of initial value $\mathbf{z}_0$. (The definition of convergence to an attractor, using an iterated function system, requires a knowledge of the Hausdorff metric. For more details, see chapter 2 of Barnsley **[2]**.)

How are translation vectors chosen to produce tilings? For a matrix M as given above, $|\det(M)| = |ad - bc| = m$ is the area of the parallelogram P spanned by the vectors $\mathbf{v}_1 = \begin{bmatrix} a \\ c \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} b \\ d \end{bmatrix}$. Recall that the vectors in $\mathscr{E}$ are those with integer coordinates that lie in or on P, but not on the two edges that do not contain $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$. We shall call these vectors the *principal residue vectors*. Let L denote the lattice of all points in the complex plane with integer coordinates. (These points are known as *Gaussian integers*.) Note that $\mathscr{E} \subseteq L$. For $j = 1, \dots, m$, define $L_j := \{\mathbf{r}_j + M\mathbf{x} \colon \mathbf{x} \in L\}$. The vectors $\{\mathbf{r}_j\}$ are said to form a *complete residue system* for M, because $L = \bigcup_{j=1}^{m} L_j$ and $L_j \cap L_k = \varnothing$ whenever $j \neq k$. (For more on complete residue systems, see Gilbert **[4]**.)

The corner points of the parallelogram P are members of L and are linear combinations (with coefficients either 0 or 1) of the columns of M. All integer linear combinations of the columns of M form a subset G of L. In the plane, the subset G forms a grid of parallelograms, each congruent to P. Each parallelogram contains m points of L, just as P does. (We count the points in the congruent parallelograms with the same conventions as in P. See, for example, FIGURE 4a.) Each point $\mathbf{r}_j$ of L inside P is equivalent to one point $\mathbf{y}_j$ (and we write $\mathbf{r}_j \approx \mathbf{y}_j$) inside each of these congruent parallelograms. In general, as long as $\mathbf{y}_1 = \mathbf{r}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\mathbf{y}_j \approx \mathbf{r}_j$ for $j = 2, \ldots, m$, then the collection $\{\mathbf{y}_j\}$ will also form a complete residue system for the matrix M. Following Example A, each of the four additional vectors was equivalent to vector $\mathbf{r}_2$. The location of the residue vectors determines the location of the fractiles, and the shape of the tilings may change dramatically with different choices of residue systems.



Locating the residue vectors ($m = 5$)

5-rep tile with symmetry

FIGURE 4

We can summarize these ideas as follows:

PROPOSITION. *Suppose that M represents an expansive map, $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ is a complete residue system for M, and $f_j(\mathbf{z}) = \mathbf{y}_j + M^{-1}\mathbf{z}$. Then the attractor set $A = \cup_{j=1}^{m} A_j$ is the union of m tiles A_j that have disjoint interiors and satisfy $A_j = f_j(A)$. Such tiles are called m-rep tiles. (This result is Theorem 1 in* **[1]**.)

We illustrate the construction of a tiling using m-rep tiles in the following example.

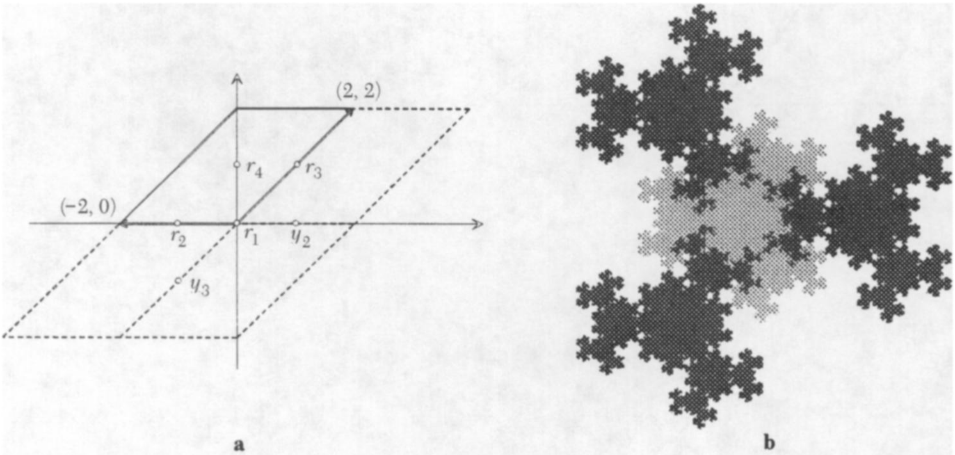
Example C. Let $M = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$; then $m = 5$. As FIGURE 4a illustrates, the principal residue vectors are $\mathbf{r}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{r}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\mathbf{r}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\mathbf{r}_4 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, and $\mathbf{r}_5 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. For a more symmetric tiling, we choose equivalent vectors for our residue system. FIGURE 4b is generated by setting $\mathbf{y}_1 = \mathbf{r}_1$, $\mathbf{y}_2 = \mathbf{r}_2$, $\mathbf{y}_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \approx \mathbf{r}_3$, $\mathbf{y}_4 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \approx \mathbf{r}_4$, and $\mathbf{y}_5 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \approx \mathbf{r}_5$. Note that the vectors $\{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\}$, considered as complex numbers, are the fourth roots of unity and are symmetric about $\mathbf{r}_1$.

Constructing tiles with radial symmetry

In Example B, we can replace $\begin{bmatrix}2\\0\end{bmatrix}$ with the equivalent vector $\begin{bmatrix}-1\\0\end{bmatrix}$ and use $\begin{bmatrix}0\\0\end{bmatrix}$, $\begin{bmatrix}1\\0\end{bmatrix}$, and $\begin{bmatrix}-1\\0\end{bmatrix}$ as residue vectors. Observe that $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}-1\\0\end{bmatrix}$ are symmetrically located about $\mathbf{r}_1 = \begin{bmatrix}0\\0\end{bmatrix}$. Can we always find residue vectors $\mathbf{r}_2, \dots, \mathbf{r}_m$ that are symmetrically located about r_1? That is, can we construct a tiling that exhibits radial symmetry about $\mathbf{r}_1$? (See, for example, FIGURE 4b.) This turns out to be possible only when $m = 2$, 3, 4, 5, and 7. (For an algebraic proof, see **[8]**.) The cases $m = 2$, 3, and 5 were illustrated in previous examples. The cases $m = 4$ and $m = 7$ require more analysis.

For $m = 4$, the matrix $M = \begin{bmatrix}a & b\\ c & d\end{bmatrix}$ must represent an expansive map and have determinant 4. These conditions mean that $(a - \lambda)(d - \lambda) - bc = 0$ has roots λ with $|\lambda| > 1$ and that $ad - bc = 4$. Since $\lambda = \frac{1}{2}[(a + d) \pm \sqrt{(a + d)^2 - 16}\,]$, we have two cases: (i) If $(a + d)^2 - 16 < 0$, then λ is complex, $|a + d| < 4$ and $\lambda\bar{\lambda} = 4$. (ii) If λ is real, then $|a + d| \geq 4$, and $|\lambda| > 1$ implies that $|a + d| - \sqrt{(a + d)^2 - 16} > 2$, or $|a + d| < 5$. Thus $|a + d| = 4$. From (i) and (ii) we conclude that M must be chosen so that $|a + d| \leq 4$. In Example D, $a + d = 2$.

Example D. Let $M = \begin{bmatrix}2 & -2\\ 2 & 0\end{bmatrix}$, with principal residue vectors as shown in FIGURE 5a. To achieve radial symmetry we want to generate a tiling using three vectors symmetrically located about $\mathbf{r}_1 = \begin{bmatrix}0\\0\end{bmatrix}$. The vectors $\mathbf{y}_1 = \mathbf{r}_1$, $\mathbf{y}_2 = \begin{bmatrix}1\\0\end{bmatrix} \approx \mathbf{r}_2$, $\mathbf{y}_3 = \begin{bmatrix}-1\\-1\end{bmatrix} \approx \mathbf{r}_3$ and $\mathbf{y}_4 = \mathbf{r}_4 = \begin{bmatrix}0\\1\end{bmatrix}$ form a complete residue system for M, but they are not symmetric about $\mathbf{r}_1$. We observe that the complex third roots of unity $\mathbf{v}_1 = \begin{bmatrix}-1/2\\ \sqrt{3}/2\end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix}1\\0\end{bmatrix}$, and $\mathbf{v}_3 = \begin{bmatrix}-1/2\\ -\sqrt{3}/2\end{bmatrix}$ are symmetrically located about $\mathbf{r}_1$, but that the vectors $\mathbf{v}_1$ and $\mathbf{v}_3$ are not Gaussian integers. If we apply the linear transformation represented



a

Locating equivalent residue vectors ($m = 4$)

b

4-rep tile with symmetry

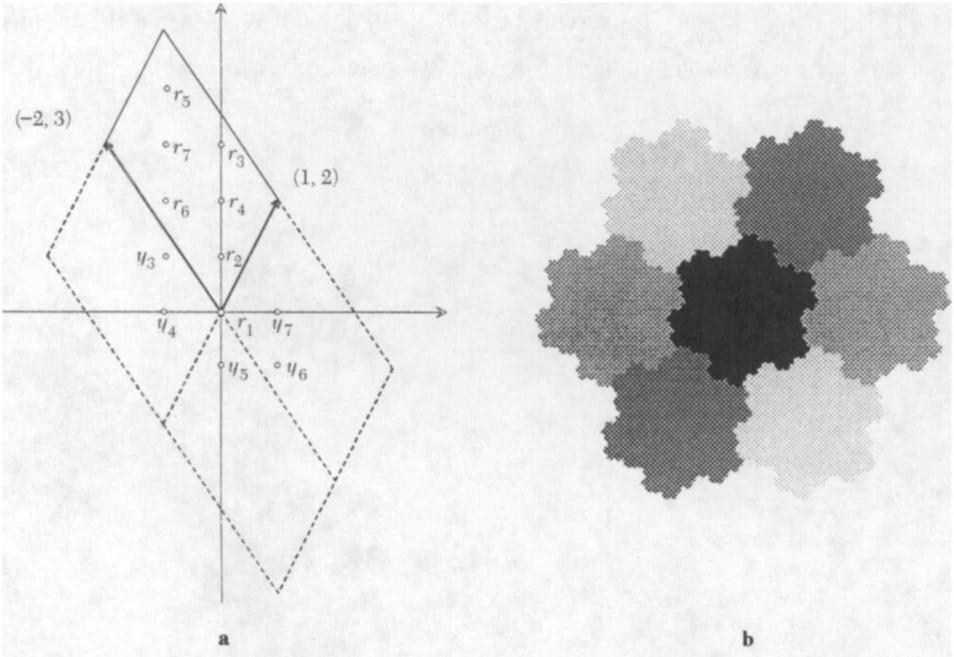
FIGURE 5

by $B = \begin{bmatrix} 1 & -1/2 \\ 0 & \sqrt{3}/2 \end{bmatrix}$, then we get $B\mathbf{y}_1 = \mathbf{y}_1$, $B\mathbf{y}_2 = \mathbf{v}_2$, $B\mathbf{y}_3 = \mathbf{v}_3$, and $B\mathbf{y}_4 = \mathbf{v}_1$. The iteration process can now be performed using the functions $f_j(\mathbf{z}) = B\mathbf{y}_j + h^{-1}(\mathbf{z})$, where $h = BMB^{-1}$ and $h^{-1} = \begin{bmatrix} 1/4 & \sqrt{3}/4 \\ -\sqrt{3}/4 & 1/4 \end{bmatrix}$. (Note that h^{-1} represents multiplication by the complex number $1/4 - i\sqrt{3}/4$.) The resulting tiling is shown in FIGURE 5b. The transformation B is the change of basis matrix that converts the lattice formed by $\mathscr{E}$ into one formed by $\mathbf{v}_1$ and $\mathbf{v}_2$. The reader is encouraged to try the same procedure for other matrices M with $|a + d| \leq 4$.

To see that, even after a change of basis, the transformations produce a tiling, set $A_j := f_j(A) = \mathbf{y}_j + M^{-1}(A)$. Then $A = \cup_{j=1}^{m} A_j$ and the A_j's have disjoint interiors. Set $K_j := B(A_j)$ for each j and $K := B(A)$. Then $K_j = B(\mathbf{y}_j + M^{-1}(A)) = B\mathbf{y}_j + BM^{-1}B^{-1}B(A) = B\mathbf{y}_j + h^{-1}K$. It follows that $K = \cup_{j=1}^{m} K_j$ and that the K_j's have disjoint interiors.

In the next example, we will construct a tiling with six unit vectors symmetrically located on the unit circle. We use the complex sixth roots of unity, $\mathbf{v}_j = \exp(i\pi j/3)$, $1 \leq j \leq 6$. In this case, $\det(M) = 7$; for reasons discussed prior to Example D, we must restrict $|a + d| \leq 7$.

Example E. Let $M = \begin{bmatrix} 1 & -2 \\ 2 & 3 \end{bmatrix}$, with principal residue vectors $\{\mathbf{r}_j\}$ as shown in FIGURE 6a. Note that we can choose the residue vectors $\mathbf{y}_1 = \mathbf{r}_1$, $\mathbf{y}_2 = \mathbf{r}_2$, $\mathbf{y}_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \approx \mathbf{r}_3$, $\mathbf{y}_4 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \approx \mathbf{r}_4$, $\mathbf{y}_5 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \approx \mathbf{r}_5$, $\mathbf{y}_6 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \approx \mathbf{r}_6$ and $\mathbf{y}_7 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \approx \mathbf{r}_7$. We set $B = \begin{bmatrix} 1 & 1/2 \\ 0 & \sqrt{3}/2 \end{bmatrix}$ and note that $B\mathbf{y}_{j+1} = \mathbf{v}_j$, $1 \leq j \leq 6$. We now iterate using the functions



a Locating equivalent residue vectors ($m = 7$)

b Gosper snowflake

FIGURE 6

$f_j(\mathbf{z}) = B\mathbf{y}_j + h^{-1}(\mathbf{z})$, $j = 1, \ldots, 7$, where $h^{-1} = B^{-1}M^{-1}B = \begin{bmatrix} 2/7 & \sqrt{3}/7 \\ -\sqrt{3}/7 & 2/7 \end{bmatrix}$. Again we note that the matrix h^{-1} represents multiplication by the complex number $2/7 - i\sqrt{3}/7$. The resulting tiling, called the Gosper snowflake, is shown in FIGURE 6b. The Gosper snowflake changes the regular hexagon "just enough" to allow a subdivision into seven similar fractiles.

Similarity maps

A similarity map g satisfies $|g(\mathbf{x}) - g(\mathbf{y})| = r|\mathbf{x} - \mathbf{y}|$, for $r > 0$ and all $\mathbf{x}, \mathbf{y}$ in the plane. Geometrically, a similarity map is a composition of any collection of the four simple mappings: scaling by a positive factor r, rotation about the origin, translation, and reflection. If each mapping in our collection $\{f_j\}$ is a similarity map with $0 < r < 1$, then the resulting tiling (attractor) A is *self-similar*. That is, A is the union of m smaller copies of itself. In this case, the attractor has the same shape as each of its m-rep tiles. This phenomenon appears in FIGURES 2b, 4b, 5b and 6b. Each of these rep tiles arises from a special sort of linear map: multiplication by a complex number.

Multiplication by a complex number

As we have seen, each complex number $q = \alpha + i\beta$ corresponds to the matrix

$$h = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix};$$

then the matrix operations correspond to ordinary arithmetic operations on complex numbers. If $|q| > 1$, the map h is expansive, since the eigenvalues of h have modulus $\alpha^2 + \beta^2 = |q|^2$. Also, h is a similarity map, since if $\mathbf{z} = (x_1, x_2)$, then

$$|h(x_1, x_2)|^2 = (\alpha x_1 - \beta x_2)^2 + (\beta x_1 + \alpha x_2)^2 = (\alpha^2 + \beta^2)(x_1^2 + x_2^2) = |q|^2|x|^2.$$

FIGURE 2b is derived from the complex number $1 - i$ and FIGURE 4b from $2 + i$.

If α and β are not integers, but $|q| > 1$, a change of basis can still produce a similarity map h from an integer matrix M. For example, define the change of basis matrix

$$B := \begin{bmatrix} 1 & \alpha \\ 0 & -\beta \end{bmatrix};$$

set

$$M = \begin{bmatrix} 2\alpha & \alpha^2 + \beta^2 \\ -1 & 0 \end{bmatrix};$$

and let $h = BMB^{-1}$. If α and β are chosen so that M is an integer matrix, all earlier methods can be applied. We illustrate the process in the next example.

Example F. Let $q = \frac{1}{2} + i\sqrt{15}/2$. With $B = \begin{bmatrix} 1 & 1/2 \\ 0 & -\sqrt{15}/2 \end{bmatrix}$ we find that $M = \begin{bmatrix} 1 & 4 \\ -1 & 0 \end{bmatrix}$, and $\mathbf{r}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{r}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{r}_3 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, and $\mathbf{r}_4 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$. A tiling can be derived using the functions $f_j = B\mathbf{y}_j + h^{-1}(\mathbf{z})$, where $\mathbf{y}_1 = \begin{bmatrix}1\\-1\end{bmatrix} \approx \mathbf{r}_1$, $\mathbf{y}_2 = \mathbf{r}_2$, $\mathbf{y}_3 = \begin{bmatrix}-1\\-1\end{bmatrix} \approx \mathbf{r}_3$, $\mathbf{y}_4 = \begin{bmatrix}-1\\0\end{bmatrix} \approx \mathbf{r}_4$ and $h = BMB^{-1}$.

Developing similarity maps

In Examples D and E, a change of basis applied to M produced similarity mappings and attractive tilings. This process works—i.e., we can choose a matrix B such that $h = BMB^{-1}$ is a similarity mapping—if M either has (i) two real eigenvalues with equal modulus and independent eigenvectors or (ii) a pair of complex conjugate eigenvalues.

In case (i), let λ_1 and λ_2 ($\lambda_1 = \pm\lambda_2$) be real eigenvalues for M with corresponding eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Let $B^{-1} = [\mathbf{v}_1, \mathbf{v}_2]$ be the matrix with column vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Then

$$MB^{-1} = M[\mathbf{v}_1, \mathbf{v}_2] = [\lambda_1\mathbf{v}_1, \lambda_2\mathbf{v}_2] = B^{-1}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}$$

so that $h := BMB^{-1} = \begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}$ is a similarity map. Then h can be used to generate the tiling with functions of the form $f_j = B\mathbf{y}_j + h^{-1}(\mathbf{z})$, where $\{\mathbf{y}_j\}$ are residue vectors for M. (Note that h depends only on the choice of eigen*values* while the translation vectors, and the tiling, vary with the choice of eigen*vectors*.) Example G illustrates this case.

Example G. The matrix $M = \begin{bmatrix}2 & 2\\ 1 & -2\end{bmatrix}$ has determinant -6 and eigenvalues $\lambda = \pm\sqrt{6}$. Associated eigenvectors of M are $\mathbf{v}_1 = \begin{bmatrix}1\\ (\sqrt{6}-2)/2\end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix}1\\ (\sqrt{6}+2)/2\end{bmatrix}$ so $B^{-1} = \begin{bmatrix}1 & 1\\ (\sqrt{6}-2)/2 & (-\sqrt{6}+2)/2\end{bmatrix}$ and the similarity map is $h = \begin{bmatrix}\sqrt{6} & 0\\ 0 & -\sqrt{6}\end{bmatrix}$. FIGURE 7 shows the tiling generated using the functions $f_j(\mathbf{z}) = B\mathbf{r}_j + h^{-1}(\mathbf{z})$ with principal residue vectors

$$\mathbf{r}_1 = \begin{bmatrix}0\\0\end{bmatrix}, \mathbf{r}_2 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{r}_3 = \begin{bmatrix}2\\0\end{bmatrix}, \mathbf{r}_4 = \begin{bmatrix}2\\-1\end{bmatrix}, \mathbf{r}_5 = \begin{bmatrix}1\\-1\end{bmatrix}, \mathbf{r}_6 = \begin{bmatrix}3\\-1\end{bmatrix}.$$

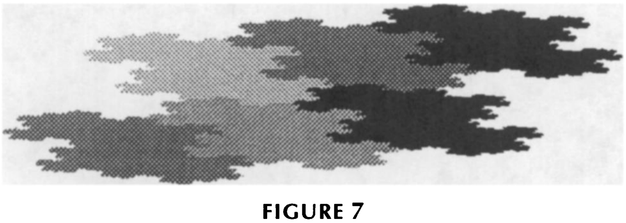


FIGURE 7
Similarity tiling using a change of basis ($m = 6$)

In case (ii), suppose that $M = \begin{bmatrix}a & b\\ c & d\end{bmatrix}$ has complex conjugate eigenvalues $\lambda = \alpha + i\beta$

and $\bar{\lambda} = \alpha - i\beta$, $\beta \neq 0$. Let $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_{11} + iv_{12} \\ v_{21} + iv_{22} \end{bmatrix}$ be an eigenvector associated with λ. Then $B^{-1} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$ is the inverse of an appropriate change of basis matrix for obtaining a similarity transformation h. To see this, keep in mind that the real and imaginary parts of $M\mathbf{v}$ and $\lambda\mathbf{v}$ must be the same, and use the same strategy as in case (i) to obtain $MB^{-1} = B^{-1}\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}$. Therefore, $h = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}$, which represents multiplication by the complex number $\bar{\lambda}$. Since B^{-1} is defined using an eigenvector of M, the matrix B can have many forms. For example, if we choose $v_1 = 1 + iv_{12}$ and $v_2 = 0 + iv_{22}$, we find that

$$B^{-1} = \begin{bmatrix} 1 & \dfrac{1}{\beta}(\alpha - a) \\ 0 & \dfrac{-c}{\beta} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & \dfrac{(\alpha - a)}{c} \\ 0 & \dfrac{-\beta}{c} \end{bmatrix}$$

as illustrated in Example H. If, as in Example I, we choose $v_1 = 1 + i0$, we have

$$B^{-1} = \begin{bmatrix} 1 & 0 \\ \dfrac{\alpha - a}{b} & \dfrac{-\beta}{b} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 0 \\ \dfrac{\alpha - a}{\beta} & \dfrac{-b}{\beta} \end{bmatrix}.$$

In both cases, the transformation $h = BMB^{-1}$ is a similarity mapping.

Example H. Let $M = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}$; then M has determinant 3 and eigenvalues $\frac{3}{2} \pm i\sqrt{3}/2$. With $B = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & -\sqrt{3}/2 \end{bmatrix}$, the similarity mapping $h = BMB^{-1}$ is given by the matrix $\begin{bmatrix} \frac{3}{2} & \sqrt{3}/2 \\ -\sqrt{3}/2 & \frac{3}{2} \end{bmatrix}$. If we choose the residue vectors $\mathbf{y}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{y}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and $\mathbf{y}_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and iterate using the functions $f_j = B\mathbf{y}_j + h^{-1}(\mathbf{z})$, $j = 1, 2, 3$, we obtain the so-called terdragon shown in FIGURE 8. (For comparison, see again the 3-rep tile of Example B, and note the subtle changes in the matrices.)

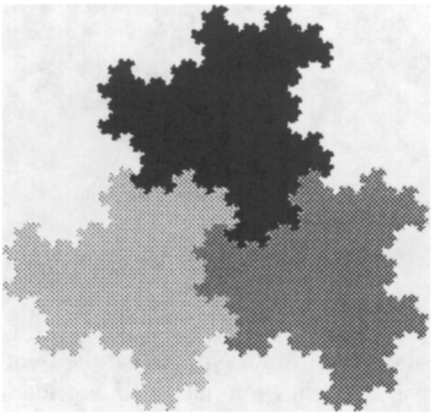


FIGURE 8
The Terdragon (a 3-rep tile)

Example I. Let $M = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix}$; then $\det(M) = 5$ and M has eigenvalues $\lambda = 2 \pm i$. The change of basis matrix $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ yields the similarity transformation $h = BMB^{-1} = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$. With the residue vectors

$$\mathbf{y}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{y}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \mathbf{y}_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{y}_4 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \mathbf{y}_5 = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

we obtain the 5-rep tile shown in FIGURE 4b.

Variations

Once one begins to generate tilings as above, ideas for modifying the figures abound. We demonstrate a few variations below; readers are encouraged to experiment further.

Example J. What happens if one of the functions is removed from the iteration process? Using the similarity transformation from Example E, we generated the wreath in FIGURE 9 by omitting the function with residue vector $\mathbf{r}_1$. (Compare FIGURE 6b.)

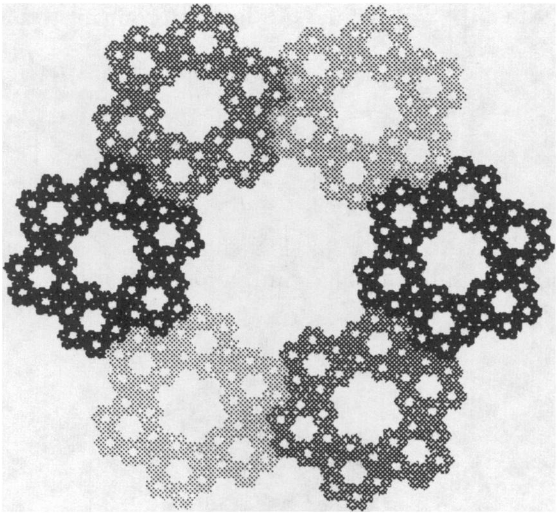


FIGURE 9
Wreath (modified snowflake)

Example K. Let $M = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ and use the residue vectors $\mathbf{y}_1$, $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ from Example D. FIGURE 10a shows the result. Notice that the fractiles appear to overlap. In fact, they do not. FIGURE 10b shows a modified version of the generated tiling, omitting the piece A_1 associated with the $\mathbf{y}_1$ residue vector. This shows that the fractiles $\{A_j\}$ are not simply connected.

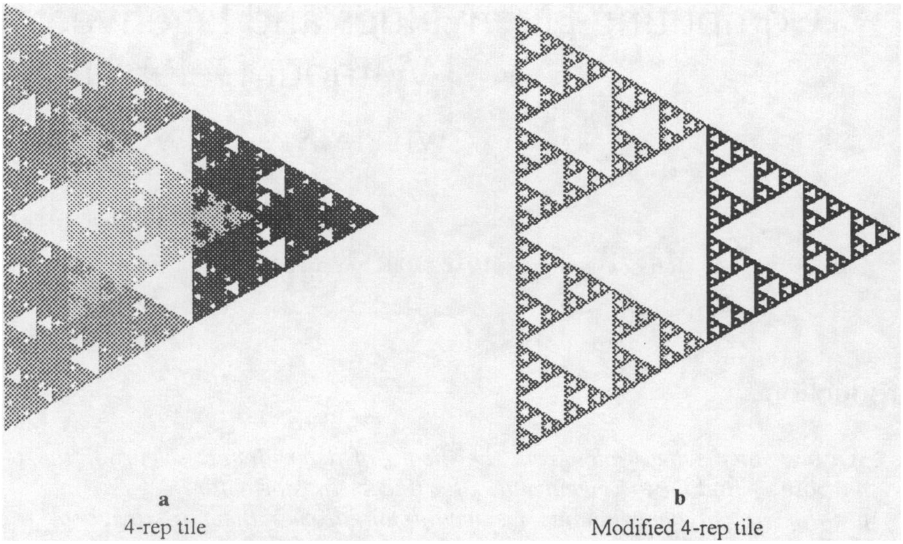


FIGURE 10

Note. Many computer resources are available for generating fractals. The Random Iteration Algorithm, presented by Barnsley in [2], can be used to generate pictures like those in this paper. *Fractal Attraction* [11] is another useful tool for investigating these ideas. FRACTINT, used to generate the fractals in this article, is freeware, available from `http://spanky.triumf.ca/www/fractint/getting.html`. Generating fractals in color presents even more dramatic pictures.

Acknowledgment. The authors wish to thank Frank DeMeyer, the referees, and the editor for many suggestions that improved this article.

REFERENCES

1. C. Bandt, Self-similar sets 5. Integer matrices and fractal tilings of $\mathbb{R}^n$, *Proceedings of the AMS* 112 (1991), 549–562.
2. M. F. Barnsley, *Fractals Everywhere*, Academic Press, New York, NY, 1993.
3. G. A. Edgar, *Measure, Topology, and Fractal Geometry*, Springer-Verlag, New York, NY, 1990.
4. W. J. Gilbert, Fractal geometry derived from complex bases, *The Mathematical Intelligencer* 4 (1982), 78–86.
5. D. Goffinet, Number systems with a complex base: a fractal tool for teaching topology, *Amer. Math. Monthly* 98 (1991), 249–255.
6. B. Grünbaum and G. C. Shephard, *Tilings and Patterns*, Freeman, New York, NY, 1987.
7. B. Grünbaum and G. C. Shephard, Tilings by regular polygons, this MAGAZINE 50 (1977), 227–247.
8. T. W. Hungerford, *Algebra*, Springer-Verlag, New York, NY, 1974.
9. J. E. Hutchinson, Fractals and self-similarity, *Indiana Univ. Math. J.* 30 (1981), 713–747.
10. D. E. Knuth, *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, Addison-Wesley, Reading, MA, 1981.
11. K. Lee and Y. Cohen, *Fractal Attraction*, Academic Press, Boston, MA, 1992.
12. B. B. Mandelbrot, *Fractals, Forms, Chance and Dimension*, Freeman, San Francisco, CA, 1977.
13. H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals*, Springer-Verlag, New York, NY, 1992.
14. D. Schattschneider, Will it tile? Try the Conway criterion!, this MAGAZINE 53 (1980), 224–233.
15. Sherman K. Stein and Sandor Szabo, *Algebra and Tiling*, Math. Assoc. of America, Washington, DC, 1994.

Computing Eigenvalues and Eigenvectors Without Determinants

WILLIAM A. MCWORTER, JR.
LEROY F. MEYERS[1]
Ohio State University
Columbus, OH 43210

Hans Zassenhaus (1912–1991) in memoriam

Introduction

> *We introduce some topics from the theory of determinants solely for the purpose of finding the eigenvalues of a linear transformation. Were it not for this use of determinants we would not discuss them in this book.*
> —E. Nering [**12**]

Who but Simon Legree would demand that a student use a determinant to compute by hand the eigenvalues and eigenvectors of the matrix

$$A=\begin{bmatrix} 3 & -1 & -6 & 1\\ -1 & 3 & 4 & -1\\ 1 & -1 & -2 & 1\\ -1 & 1 & 4 & 1\end{bmatrix}?$$

The student would first have to compute the 4×4 determinant $\det(tI-A)$, whose entries are polynomials, then find all the zeros of the resulting polynomial of degree 4, and finally, as is the case with this particular matrix A, find the null spaces of three 4×4 matrices.

To replace the computation of polynomial determinants and unwieldy null spaces, this paper describes a faster way, of greater educational value because it requires *understanding* of the concepts of eigenvalue and eigenvector. The algorithm, a modification of McWorter [**11**], uses only fundamental concepts of linear algebra, especially linear dependence and independence, for the exact computation of eigenvectors, and is easily extended to yield generalized eigenvectors and a Jordan basis. The algorithm produces the eigenvectors of the above matrix A in a few minutes on less than half a sheet of paper.

This algorithm has been used in our elementary linear algebra classes for over ten years. One student comment seems to say it all, "I know it is faster, but with the determinant you don't have to think" (Denise Sayre, with permission).

The first author, reading Nering's words quoted above while he was a fresh Ph.D., was motivated to eliminate the necessity for an excursion into determinant theory even to obtain eigenvalues. The underlying idea behind his ultimate approach is not new. A related determinant-free theoretical procedure was developed nearly 80 years ago by Kowalewski (1917) to find the invariant factors of a matrix. Bennett [**2**] in 1931

[1] Professor Meyers died on November 8, 1995; he had been a member of the Ohio State University mathematics department since 1953.

partially remedied the lack of an explicit construction in Kowalewski's procedure. Krylov [9] in 1931, while investigating systems of linear differential equations, showed how to simplify the computation of the characteristic polynomial $\det(tI - A)$ for certain matrices A, and Danilevskiĭ [5] in 1937 extended Krylov's algorithm to all square matrices. (Clearer expositions of Krylov's and Danilevskiĭ's algorithms are given in Faddeev & Faddeeva [6], pp. 263–273, 285–295; (1963), pp. 231–241, 251–260.)

The present paper generalizes the algorithms of Kowalewski, Krylov, and Danilevskiĭ by providing an elementary and efficient symbolic algorithm for the exact computation of eigenvectors. Section 1 illustrates the algorithm for finding eigenvalues and eigenvectors, Section 2 provides a justification for that algorithm, and Section 3 describes, justifies, and illustrates the extension of the algorithm to find generalized eigenvectors.

1. First example: eigenvalues and eigenvectors

We begin with definitions.

DEFINITIONS. *An eigenvector of the matrix A over the field* $\mathbf{F}$ *for the eigenvalue λ in* $\mathbf{F}$ *is a nonzero vector* $\mathbf{x}$ *such that* $A\mathbf{x} = \lambda\mathbf{x}$. *The eigenspace for λ consists of all vectors* $\mathbf{x}$ *such that* $A\mathbf{x} = \lambda\mathbf{x}$.

Given an $n \times n$ matrix A over an algebraically closed field $\mathbf{F}$ (such as the complex numbers), the algorithm described here produces equations of the form $(A - \lambda I)\mathbf{x} = \mathbf{o}$, with $\mathbf{x}$ an eigenvector and $\mathbf{o}$ the zero vector.

The algorithm begins by following a procedure used in proving that every $n \times n$ matrix over an algebraically closed field $\mathbf{F}$ has at least one eigenvalue and corresponding eigenvector. (See Faddeev & Faddeeva [6], Cater [3], and Axler [1].) Let $\mathbf{u}$ be any nonzero vector in $\mathbf{F}^n$. Since $\mathbf{F}^n$ has finite dimension n, the $n+1$ vectors $\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^n\mathbf{u}$ are linearly dependent. Let k be the smallest positive integer such that $a_0\mathbf{u} + a_1 A\mathbf{u} + a_2 A^2\mathbf{u} + \cdots + a_k A^k\mathbf{u} = \mathbf{o}$, for some $a_0, \ldots, a_k$ in $\mathbf{F}$ with $a_k \neq 0$. Algebraic closure ensures that the polynomial $a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k$ in $\mathbf{F}[t]$ is factorable as $(t - \lambda)Q(t)$ for some λ in $\mathbf{F}$ and some polynomial $Q(t)$ in $\mathbf{F}[t]$. Hence $(A - \lambda I)Q(A)\mathbf{u} = \mathbf{o}$. The minimality of k implies that the vector $Q(A)\mathbf{u}$ is nonzero and so is an eigenvector of A for the eigenvalue λ.

Very little modification of the procedure just described is needed to find *every* eigenvalue and a basis for each eigenspace of A. Indeed, we can illustrate the algorithm right now on the 4×4 matrix A above. We will justify the algorithm in the next section.

As in the procedure above, begin by choosing $\mathbf{u}$ to be *any* nonzero column vector in $\mathbb{C}^4$, say, for perversity, $\mathbf{u} = [0 \quad 1 \quad -1 \quad 1]^{\mathrm{T}}$, denoted by $(0, 1, -1, 1)$ in running text. The vector $\mathbf{u}$ is called a *seed* because other vectors grow from it. Then compute $A\mathbf{u}$, $A^2\mathbf{u} = A(A\mathbf{u})$, $A^3\mathbf{u} = A(A^2\mathbf{u})$, etc., by successive left multiplication by A until, *for the first time*, $A^k\mathbf{u}$ is a linear combination of the vectors $\mathbf{u}, A\mathbf{u}, \ldots, A^{k-1}\mathbf{u}$. Only

$$\mathbf{u} = (0, 1, -1, 1), \quad A\mathbf{u} = (6, -2, 2, -2), \quad \text{and} \quad A^2\mathbf{u} = (6, -2, 2, -2)$$

need be computed, since $A^2\mathbf{u}$ is the *first* of the generated vectors to be linearly dependent on previous generated vectors. One obvious dependence relation among the generated vectors is

$$A^2\mathbf{u} - A\mathbf{u} = \mathbf{o}. \tag{1}$$

This dependence relation alone yields two eigenvalues and their corresponding eigenvectors. Equation (1) can be put into the desired form $(A-\lambda I)\mathbf{x}=\mathbf{o}$ by factoring in two ways:

$$(A-1I)(A\mathbf{u})=\mathbf{o} \quad \text{and} \quad (A-0I)(A\mathbf{u}-\mathbf{u})=\mathbf{o}.$$

The first factorization says that $A\mathbf{u}$ belongs to the eigenspace for the eigenvalue 1, and the second factorization says that $A\mathbf{u}-\mathbf{u}$ belongs to the eigenspace for the eigenvalue 0. The vectors $A\mathbf{u}$ and $A\mathbf{u}-\mathbf{u}$ are nonzero because they are nonzero linear combinations of the linearly independent vectors $\mathbf{u}$ and $A\mathbf{u}$ (because the generation of vectors stopped at the *first* occurrence of linear dependence). Since $\mathbf{u}$ and $A\mathbf{u}$ have already been computed, the two eigenvectors can be given explicitly with little further work:

$$A\mathbf{u}=(6,-2,2,-2) \text{ for } 1; \quad \begin{aligned} A\mathbf{u}-\mathbf{u}&=(6,-2,2,-2)-(0,1,-1,1)\\ &=(6,-3,3,-3) \text{ for } 0. \end{aligned}$$

At this point, the vectors $\mathbf{u}$, $A\mathbf{u}$, and $A^2\mathbf{u}$ generated so far, as well as the eigenvectors constructed from them, span a 2-dimensional subspace of $\mathbb{C}^4$. Additional independent eigenvectors, if they exist, must lie outside this subspace. Continuing the generation with seeds outside this subspace will get any remaining eigenvectors.

Reseed with a new vector *linearly independent of the vectors generated so far*, say with $\mathbf{v}=(0,0,1,0)$. Then compute $A\mathbf{v}$, $A^2\mathbf{v}$, etc., until *for the first time* a vector $A^1\mathbf{v}$ is a linear combination of previously generated vectors. Only $\mathbf{v}=(0,0,1,0)$ and $A\mathbf{v}=(-6,4,-2,4)$ need be computed, since the set $\{\mathbf{u}, A\mathbf{u},\mathbf{v}\}$ is linearly independent, but $\{\mathbf{u}, A\mathbf{u},\mathbf{v}, A\mathbf{v}\}$ is linearly dependent. This dependence can be expressed by the equation

$$A\mathbf{v}-2\mathbf{v}+A\mathbf{u}-2\mathbf{u}=\mathbf{o}. \tag{2}$$

This dependence relation (as well as the check that $\{\mathbf{u}, A\mathbf{u},\mathbf{v}\}$ is linearly independent) can be found by the usual methods; however, we are in a classroom situation and so we can make the dependence checks succumb to inspection, as in this example.

Equation (2) can be put into the desired form $(A-\lambda I)\mathbf{x}=\mathbf{o}$ as follows:

$$(A-2I)(\mathbf{v}+\mathbf{u})=\mathbf{o}.$$

This shows that

$$\mathbf{v}+\mathbf{u}=(0,0,1,0)+(0,1,-1,1)=(0,1,0,1)$$

is an eigenvector for 2. This eigenvector is linearly independent of those produced earlier, because it involves the vector $\mathbf{v}$, which is outside the subspace spanned by the others.

At this point, the vectors $\mathbf{u}$, $A\mathbf{u}$, $\mathbf{v}$, $A\mathbf{v}$ generated so far, as well as the eigenvectors constructed from them, span a 3-dimensional subspace of $\mathbb{C}^4$. To obtain further independent eigenvectors, if any, reseed with yet another vector linearly independent of the vectors generated so far, say with $\mathbf{w}=(0,1,0,0)$. The set $\{\mathbf{u}, A\mathbf{u},\mathbf{v},\mathbf{w}\}$ is linearly independent, but the set $\{\mathbf{u}, A\mathbf{u},\mathbf{v},\mathbf{w}, A\mathbf{w}\}$, is necessarily linearly dependent, being a set of 5 vectors in a 4-dimensional space. One dependence relation is

$$6A\mathbf{w}-12\mathbf{w}+A\mathbf{u}-4\mathbf{u}=\mathbf{0}. \tag{3}$$

(Standard basis vectors may always be used as seeds, not necessarily in turn; they sometimes simplify testing linear independence.)

Equation (3) cannot be put into the desired form $(A - \lambda I)\mathbf{x} = \mathbf{o}$. But not to worry; we can combine this equation with equations (1) and (2) to get what we need. Equation (3) insists that $\lambda = 2$. In fact, adding (1) to (3) produces

$$6A\mathbf{w} - 12\mathbf{w} + A^2\mathbf{u} - 4\mathbf{u} = (A - 2I)(6\mathbf{w} + A\mathbf{u} + 2\mathbf{u}) = \mathbf{o},$$

from which it follows that

$$6\mathbf{w} + A\mathbf{u} + 2\mathbf{u} = 6(0,1,0,0) + (6,-2,2,-2) + 2(0,1,-1,1) = (6,6,0,0)$$

is an eigenvector for 2. This eigenvector is linearly independent of the eigenvectors obtained earlier because $\mathbf{w}$ occurs in it with nonzero coefficient and is outside the subspace spanned by the others.

A mechanical way to find the right linear combination of equations (1), (2), and (3) involves putting these equations in quotient-remainder form:

$$\begin{aligned}(A - 2I)(A\mathbf{u} + \mathbf{u}) &+ 2\mathbf{u} = \mathbf{o},\\ (A - 2I)(\mathbf{v} + \mathbf{u}) \quad &+ \ \mathbf{o} = \mathbf{o},\\ (A - 2I)(6\mathbf{w} + \mathbf{u}) &- 2\mathbf{u} = \mathbf{o}.\end{aligned}$$

By choosing, if possible, a nontrivial linear combination of remainders that adds up to $\mathbf{o}$, the corresponding linear combination of these equations allows $A - 2I$ to be factored out, and then the corresponding linear combination of quotients $A\mathbf{u} + \mathbf{u}$, $\mathbf{v} + \mathbf{u}$, and $6\mathbf{w} + \mathbf{u}$ is an eigenvector for 2. Since $1\cdot(-2\mathbf{u}) + \mathbf{o} + 1\cdot(2\mathbf{u}) = \mathbf{o}$, the corresponding eigenvector is $1\cdot(6\mathbf{w} + \mathbf{u}) + \mathbf{o} + 1\cdot(\mathrm{A}\mathbf{u} + \mathbf{u}) = 6\mathbf{w} + \mathrm{A}\mathbf{u} + 2\mathbf{u} = (6,6,0,0)$.

The computation is finished. Since the generated vectors span $\mathbb{C}^4$, there can be no further seeds. The eigenspaces for the eigenvalues 0, 1, and 2 have respective dimensions 1, 1, and 2. Hence no further independent eigenvectors are possible. (If the remainder in equation (3) could not be eliminated, then there would be no additional eigenvector for 2 and no basis for $\mathbb{C}^4$ consisting of eigenvectors of A. Section 3 shows how easy it is to complete a basis consisting of generalized eigenvectors.)

The assertion made in the Introduction that the calculations take less than half a sheet of paper is confirmed by the compact display below, followed by the short calculation above for eigenvectors.

$$\begin{array}{cccc|ccc|cc|cc|}
 & & A & & \mathbf{u} & A\mathbf{u} & A^2\mathbf{u} & \mathbf{v} & A\mathbf{v} & \mathbf{w} & A\mathbf{w} \\
3 & -1 & -6 & 1 & 0 & 6 & 6 & 0 & -6 & 0 & -1 \\
-1 & 3 & 4 & -1 & 1 & -2 & -2 & 0 & 4 & 1 & 3 \\
1 & -1 & -2 & 1 & -1 & 2 & 2 & 1 & -2 & 0 & -1 \\
-1 & 1 & 4 & 1 & 1 & -2 & -2 & 0 & 4 & 0 & 1 \\
\hline
 & & & & 0 & -1 & 1 & & & & \\
\hline
 & & & & -2 & 1 & & -2 & 1 & & \\
\hline
 & & & & -4 & 1 & & 0 & & -12 & 6
\end{array}$$

At the left is the matrix A. To the right of A are the vectors generated by the algorithm, labeled on top. A vertical rule is drawn to the right of each vector linearly dependent on those to its left. The numbers in the rth row under the generated vectors are the coefficients of the vectors that occur in the rth dependence relation. The relations are used to construct eigenvectors.

A change from the standard basis for $\mathbb{C}^4$ to $\{\mathbf{u}, A\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ (those generated vectors that are not dependent on previous vectors) transforms the matrix A to Frobenius form

$$\left[\begin{array}{cc|c|r} 0 & 0 & 2 & 4/6 \\ 1 & 1 & -1 & -1/6 \\ \hline 0 & 0 & 2 & 0 \\ \cline{3-4} 0 & 0 & 0 & 2 \end{array}\right],$$

a block upper triangular matrix in which the diagonal blocks are companion matrices. The rightmost columns of the diagonal blocks exhibit the coefficients in the dependence relations (1), (2), and (3) divided by the negative of the coefficient of the vector giving rise to the dependence relation. If the seeds for the Frobenius matrix are taken to be the first, third, and fourth standard basis vectors for $\mathbb{C}^4$, then the dependence relations are the same as those for A.

As is evident from the above procedure, the only places where explicit entries in the matrix and vectors are used are in finding the dependence relations among the generated vectors and in writing the eigenvectors explicitly. Otherwise, linear combinations of the generated vectors are treated formally without regard to their values as vectors in $\mathbb{C}^4$.

2. Description and justification of the method

Let A be an $n \times n$ matrix over an algebraically closed field $\mathbf{F}$. The first phase of the algorithm constructs a list of vectors called *generated vectors*, which span $\mathbf{F}^n$, together with a set of dependence relations among the vectors in the list in the following way. The first vector in the list, called a *seed*, is any nonzero vector. Suppose that the first $k > 1$ vectors constructed are $\mathbf{v}_1, \ldots, \mathbf{v}_k$. If $\mathbf{v}_k$ is not a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$, then set $\mathbf{v}_{k+1} = A\mathbf{v}_k$. If $\mathbf{v}_k$ *is* a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$, then record one such linear combination and set $\mathbf{v}_{k+1}$ equal to any vector not a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$, also called a seed, provided such a vector exists. The algorithm must end because the two cases can happen at most n times each.

The generated vectors linearly independent of previously generated vectors form a basis for $\mathbf{F}^n$ and are called *independent generated vectors*. The remaining vectors are called *dependent generated vectors*.

As the first example shows, the algorithm uses vectors as they are *and* as they are expressed as linear combinations of generated vectors. For example, the first eigenvector constructed in the first example arose first as the linear combination $A\mathbf{u} - \mathbf{u}$ of the generated vectors $\mathbf{u}$, $A\mathbf{u}$, $A^2\mathbf{u}$, $\mathbf{v}$, $A\mathbf{v}$, $\mathbf{w}$, and $A\mathbf{w}$. It was then evaluated as the vector $(6, -3, 3, -3)$. We will call a linear combination of generated vectors an *expression* and the set of all such linear combinations $\mathbf{E}$. The set $\mathbf{E}$ forms a vector space under obvious rules for addition and scalar multiplication of expressions. A basis for $\mathbf{E}$ is the set of generated vectors regarded as expressions and so the dimension of $\mathbf{E}$ is $n + m$, where n is the number of independent generated vectors and m is the number of dependent generated vectors regarded as expressions. The integer m is also the number of dependence equations generated by the algorithm and the number of seed vectors. The evaluation of an expression as a vector in $\mathbf{F}^n$ is called the *value* of the expression.

We need to distinguish several subsets of $\mathbf{E}$. Expressions whose value is the zero vector of $\mathbf{F}^n$ are called *null expressions*. Expressions that are linear combinations of only independent generated vectors are called *lean* expressions. The lean expressions form an n-dimensional subspace of $\mathbf{E}$. The m dependence expressions constructed by the algorithm are linearly independent because each has a dependent generated vector with nonzero coefficient where all the others have that coefficient equal to zero. Hence the subspace of all null expressions has dimension at least m. Since no nonzero lean expression has the zero vector as its value, no nonzero lean expression can equal a null expression. Hence the set of all null expressions has dimension precisely m; so the dependence expressions form a basis for all null expressions.

Let $\mathbf{g}_1, \dots, \mathbf{g}_n$ be the independent generated vectors regarded as expressions, and let λ be a scalar. Then the expressions $(A - \lambda I)\mathbf{g}_i$, for $i = 1, \dots, n$, are linearly independent because each has a nonzero coefficient for a generated vector where all others have a zero (i.e., the coefficient of $A\mathbf{g}_i$ in $(A - \lambda I)\mathbf{g}_i$ is 1, while the corresponding coefficient in the other expressions is 0). Let $\mathbf{s}_1, \dots, \mathbf{s}_m$ be the seeds regarded as expressions. Then $\{(A - \lambda I)\mathbf{g}_1, \dots, (A - \lambda I)\mathbf{g}_n, \mathbf{s}_1, \dots, \mathbf{s}_m\}$ is a basis for all expressions. Any nonzero linear combination of the $(A - \lambda I)\mathbf{g}_i$ has a nonzero coefficient for some generated vector which is not a seed, whereas a nonzero linear combination of seeds has that coefficient equal to zero. Hence every expression $\mathbf{x}$ can be written uniquely in quotient-remainder form $\mathbf{x} = (A - \lambda I)\mathbf{q} + \mathbf{r}$, where $\mathbf{r}$ is a linear combination of seeds regarded as an expression.

Every eigenvector for the eigenvalue λ can be expressed uniquely as a linear combination of independent generated vectors regarded as a lean expression $\mathbf{x}$. Hence $(A - \lambda I)\mathbf{x}$ can be regarded as an expression, indeed a null expression since the value of $\mathbf{x}$ is an eigenvector. We need the fact that lean expressions are linearly independent if and only if their multiples by $A - \lambda I$ are linearly independent. To that end, suppose $\mathbf{v}$ is a nonzero lean expression. Let $\mathbf{g}_i$ be the latest generated vector in $\mathbf{v}$ with nonzero coefficient. Then $(A - \lambda I)\mathbf{v}$ is an expression with the coefficient of $\mathbf{g}_{i+1}$ nonzero. Now assume that $\mathbf{v}_1, \dots, \mathbf{v}_p$ are linearly independent lean expressions and that $\Sigma a_i(A - \lambda I)\mathbf{v}_i$ is the zero expression, for some a_i, not all zero. Then $(A - \lambda I)\Sigma a_i \mathbf{v}_i$ is the zero expression and $\Sigma a_i \mathbf{v}_i$ is a lean expression. Hence $\Sigma a_i \mathbf{v}_i$ is the zero expression, contradicting the fact the $\mathbf{v}_i$ are linearly independent. Conversely, assume that the expressions $(A - \lambda I)\mathbf{v}_1, \dots, (A - \lambda I)\mathbf{v}_p$ are linearly independent expressions, with the $\mathbf{v}_i$ lean expressions. Suppose further that $\Sigma a_i \mathbf{v}_i$ is the zero expression, with not all of the a_i equal to zero. Then $(A - \lambda I)\Sigma a_i \mathbf{v}_i = \Sigma a_i(A - \lambda I)\mathbf{v}_i$ is the zero expression, contradicting the assumption that the expressions $(A - \lambda I)\mathbf{v}_i$ are linearly independent. Thus lean expressions are linearly independent if and only if their multiples by $A - \lambda I$ are linearly independent. Therefore, to find a basis for the eigenspace for λ, it suffices to find a basis for the subspace of all null expressions of the form $(A - \lambda I)\mathbf{z}$ and then factor out the expressions $\mathbf{z}$.

A basis for the subspace of all null expressions of the form $(A - \lambda I)\mathbf{z}$ can be constructed out of a basis for the space of all m-tuples $(c_1, \dots, c_m)$ such that $\Sigma_{j=1}^{m} c_j \mathbf{r}_j = \mathbf{o}$ (the zero expression), where $(A - \lambda I)\mathbf{q}_j + \mathbf{r}_j$ is the quotient-remainder form of the dependence expression $\mathbf{x}_j$. This basis can be constructed exactly the same way the dependence expressions were constructed, with the $\mathbf{r}_j$ playing the role of the generated vectors.

Let $\mathbf{x}_{m+1}, \dots, \mathbf{x}_k$ be the basis for the null expressions of the form $(A - \lambda I)\mathbf{z}$ and let $\mathbf{x}_i = (A - \lambda I)\mathbf{q}_i$, for $i = m + 1, \dots, k$, be their quotient-remainder forms. Then the values of the expressions $\mathbf{q}_i$, form a basis for the eigenspace for λ.

We have yet to address from where the eigenvalues come. They are the zeros of certain polynomials derived from the dependence expressions. Let $\mathbf{d}$ be a dependence expression and let $\mathbf{s}$ be the latest seed such that the generated vector $A^i\mathbf{s}$ has nonzero coefficient in $\mathbf{d}$, for some i. If all such generated vectors are taken together, they can be written in the form $P(A)\mathbf{s}$, for some nonzero polynomial in $\mathbf{F}[x]$. We show that the roots of these polynomials, one polynomial for each dependence expression, are the eigenvalues of A.

Let $\mathbf{v}$ be an eigenvector with eigenvalue λ. Express $\mathbf{v}$ as a linear combination of independent generated vectors and regard it as an expression. Then $(A-\lambda I)\mathbf{v}$ is a null expression. As such,

$$(A-\lambda I)\mathbf{v}=\sum_{1}^{m}a_i\mathbf{d}_i,$$

where the $\mathbf{d}_i$ are the dependence expressions and the a_i are scalars. Let p be the largest integer such that $a_p \neq 0$. Now each dependence expression $\mathbf{d}_i$ involves with nonzero coefficient only the first i seeds. Since $(A-\lambda I)\mathbf{v}$ is a null expression, the coefficients of all seeds in this expression are zero. Thus, since only the p-th dependence expression can involve the p-th seed with nonzero coefficient, the coefficient of the p-th seed in the dependence expression $\mathbf{d}_p$ must be zero. Hence the polynomial associated with the p-th dependence expression must have the factor $A-\lambda I$.

Conversely, if λ is a root of one of the polynomials associated with the dependence expressions, let p be the least index such that the polynomial associated with the p-th dependence expression has λ as a root. Then, for each $i=1,\dots,p-1$, $\mathbf{d}_i=(A-\lambda I)\mathbf{q}_i+\mathbf{r}_i$ (the quotient-remainder form), with each $\mathbf{r}_i\neq\mathbf{o}$. Moreover, for each $i=1,\dots,p-1$, the remainder $\mathbf{r}_i$ involves with nonzero coefficient the i-th seed but no later seeds. Hence these $p-1$ remainders are linearly independent and span the subspace spanned by the first $p-1$ seeds. Now, since the polynomial associated with the p-th dependence expression has λ as a root, its remainder $\mathbf{r}_p$ does not involve the seed $\mathbf{s}_p$ and so is an element of the subspace spanned by the seed expressions $\mathbf{s}_1,\dots,\mathbf{s}_{p-1}$. Hence $\mathbf{d}_p$ plus an appropriate linear combination of the first $p-1$ dependence expressions has quotient-remainder form with zero remainder; that is, $\mathbf{d}_p$ plus some linear combination of the first $p-1$ dependence expressions is a null expression whose quotient expression evaluates to a nonzero eigenvector for λ.

3. Generalized eigenvectors and second example

DEFINITION. *A generalized eigenvector of positive integer order q (for short, a q-eigenvector) of the square matrix A for the scalar λ is a vector $\mathbf{x}$ such that*

$$(A-\lambda I)^{q-1}\mathbf{x}\neq\mathbf{o}\quad \textit{but}\quad (A-\lambda I)^{q}\mathbf{x}=\mathbf{o}.$$

In particular, a 1-eigenvector is an ordinary eigenvector. The *generalized eigenspace of order q* (or *q-eigenspace*) of A for λ consists of all vectors $\mathbf{x}$ such that $(A-\lambda I)^q\mathbf{x}=\mathbf{o}$. The *generalized eigenspace* for λ is the union of the generalized eigenspaces of all orders for λ.

The algorithm for producing a basis for the generalized eigenspace for λ extends that described in the preceding section for eigenvectors. This time we look for equations of the form $(A-\lambda I)\mathbf{z}=\mathbf{y}$, where $\mathbf{y}$ is a generalized eigenvector, not just the zero expression. The algorithm first constructs a basis for the 1-eigenspace as in the preceding section, then, using this basis, it constructs a basis for the 2-eigenspace, and so on until a basis for the entire generalized eigenspace is constructed.

Let λ be an eigenvalue of the $n \times n$ matrix A over the algebraically closed field $\mathbf{F}$. The construction builds a sequence $\mathbf{x}_1, \dots, x_m, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p$ of expressions as follows. $\mathbf{x}_1, \dots, \mathbf{x}_m$ are the dependence expressions. As in the previous section, find a basis for all linear combinations of $\mathbf{x}_1, \dots, \mathbf{x}_m$ which have the form $(A - \lambda\ I)\mathbf{z}$. Set $\mathbf{x}_{m+1}, \dots, \mathbf{x}_k$ equal to the quotients from this basis. These expressions, as vectors, form a basis for the 1-eigenspace for λ. The set $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is linearly independent as expressions because no nonzero lean expression can equal a null expression. Next, find a basis for all linear combinations of the expressions $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ which have the form $(A - \lambda I)\mathbf{z}$. This basis can be chosen so as to include the basis found above. Set $\mathbf{x}_{k+1}, \dots, \mathbf{x}_r$ equal to the additional quotients, if any, that occur. The expressions $\mathbf{x}_1, \dots, \mathbf{x}_r$ are all linearly independent because the dependence expressions are linearly independent, the lean expressions which as vectors are generalized eigenvectors are linearly independent, and no nonzero lean expression can equal a null expression.

Continue in this way until no new expressions result. Then the values of the expressions $\mathbf{x}_{m+1}, \dots, \mathbf{x}_p$ form a basis for the generalized eigenspace for λ.

Let's illustrate this algorithm with an example. We begin with the display for a 7×7 matrix A.

			A				$\mathbf{u}$	$A\mathbf{u}$	$A^2\mathbf{u}$	$\mathbf{v}$	$A\mathbf{v}$	$A^2\mathbf{v}$	$\mathbf{w}$	$A\mathbf{w}$	$A^2\mathbf{w}$	$A^3\mathbf{w}$
5	1	−28	1	−17	34	39	1	5	22	0	1	−1	0	1	6	11
1	6	−29	−1	−45	66	61	0	1	6	1	6	25	0	−1	−2	−21
5	−2	−33	3	51	−32	−1	0	5	30	0	−2	−27	0	3	16	65
4	−2	−26	3	49	−37	−8	0	4	24	0	−2	−24	1	3	11	44
2	−1	−13	2	23	−13	−2	0	2	12	0	−1	−12	0	2	10	39
−2	1	13	0	−26	23	5	0	−2	−12	0	1	12	0	0	−1	−9
6	−3	−39	3	71	−52	−9	0	6	36	0	−3	−36	0	3	16	69
							8	−6	1							
							−8	3		8	−6	1				
							3	1		−10	5		−8	12	−6	1

The dependence expressions (see the bottom three rows of the display) are as follows:

$$\mathbf{x}_1 = 8\mathbf{u} - 6A\mathbf{u} + A^2\mathbf{u},$$

$$\mathbf{x}_2 = 8\mathbf{u} + 3A\mathbf{u} + 8\mathbf{v} - 6A\mathbf{v} + A^2\mathbf{v},$$

$$\mathbf{x}_3 = 3\mathbf{u} + A\mathbf{u} - 10\mathbf{v} + 5A\mathbf{v} - 8\mathbf{w} + 12A\mathbf{w} - 6A^2\mathbf{w} + A^3\mathbf{w}.$$

The polynomials $P_1(t) = t^2 - 6t + 8 = (t-4)(t-2)$, $P_2(t) = t^2 - 6t + 8 = (t-4)(t-2)$, and $P_3(t) = t^3 - 6t^2 + 12t - 8 = (t-2)^3$, obtained from the coefficients of the latest seed in each dependence expression, shows that the eigenvalues of A are 2 and 4. We treat the eigenvalue 4 first.

The quotient-remainder forms of the dependence expressions $\mathbf{x}_i$ are as follows:

$$\mathbf{x}_1 = (A - 4I)(A\mathbf{u} - 2\mathbf{u}) \qquad + \mathbf{o},$$

$$\mathbf{x}_2 = (A - 4I)(A\mathbf{v} - 2\mathbf{v} + 3\mathbf{u}) \qquad + 4\mathbf{u},$$

$$\mathbf{x}_3 = (A - 4I)(A^3\mathbf{w} - 2A\mathbf{w} + 4\mathbf{w} + 5\mathbf{v} + \mathbf{u}) + 8\mathbf{w} + 10\mathbf{v} + 7\mathbf{u}.$$

Since the first remainder is zero, the corresponding quotient $\mathbf{q}_1 = \mathrm{A}\mathbf{u} - 2\mathbf{u}$ from $\mathbf{x}_1$ is, as a vector, an eigenvector. The second remainder, $\mathbf{r}_2 = 4\mathbf{u}$, is not a linear combination of the first remainder, so we skip to the next remainder $\mathbf{r}_3 = 8\mathbf{w} + 10\mathbf{v} + 7\mathbf{u}$. It is not a linear combination of the first two remainders. We now have a basis for the 1-eigenspace for 4, namely, $\{\mathbf{q}_1 = A\mathbf{u} - 2\mathbf{u}\}$, with $\mathbf{q}_1$ as a vector.

Here is where the extension to generalized eigenvectors kicks in. Extend the list of expressions $\mathbf{x}_i$ by appending the expression $\mathbf{x}_4 = A\mathbf{u} - 2\mathbf{u}$ and compute its quotient and remainder:

$$\mathbf{x}_4 = A\mathbf{u} - 2\mathbf{u} = (A - 4I)\mathbf{u} + 2\mathbf{u}.$$

Now check whether this fourth remainder $\mathbf{r}_4 = 2\mathbf{u}$ is a linear combination of the first three remainders. It is: $2\mathbf{r}_4 - \mathbf{r}_2 = \mathbf{o}$. The corresponding linear combination of $\mathbf{x}_i$, namely $2\mathbf{x}_4 - \mathbf{x}_2$ has the form $(A - 4I)\mathbf{z}$ and lies in the 1-eigenspace for 4. Hence its quotient $2\mathbf{q}_4 - \mathbf{q}_2 = 2\mathbf{u} - (A\mathbf{v} - 2\mathbf{v} + 3\mathbf{u}) = -A\mathbf{v} + 2\mathbf{v} - \mathbf{u}$, as a vector, is in the 2-eigenspace for 4. Now append the expression $\mathbf{x}_5 = -A\mathbf{v} + 2\mathbf{v} - \mathbf{u}$ to the list and compute its quotient $\mathbf{q}_5 = -\mathbf{v}$ and remainder $\mathbf{r}_5 = -2\mathbf{v} - \mathbf{u}$. Test if this new remainder is a linear combination of the other remainders; it is not. Thus a basis for the generalized eigenspace for the eigenvalue 4 has the two vectors $\mathbf{x}_4 = (3, 1, 5, 4, 2, -2, 6)$ and $\mathbf{x}_5 = (2, 4, -2, -2, -1, 1, -3)$.

For the eigenvalue 2 the computation is longer, because the generalized eigenspace for 2 must have dimension $7 - 2 = 5$. The computation is summarized in the table below. To emphasize that the algorithm is the same for the eigenvalue 2 as it was for the eigenvalue 4, we use the same notation.

The $\mathbf{x}_j$	where from	value	quotient $\mathbf{q}_j$	remainder $\mathbf{r}_j$
$\mathbf{x}_1$		$= A^2\mathbf{u} - 6A\mathbf{u} + 8\mathbf{u}$	$A\mathbf{u} - 4\mathbf{u}$	$\mathbf{o}$
$\mathbf{x}_2$		$= A^2 - 6A\mathbf{v} + 8\mathbf{v} + 3A\mathbf{u} - 8\mathbf{u}$	$A\mathbf{v} - 4\mathbf{v} + 3\mathbf{u}$	$-2\mathbf{u}$
$\mathbf{x}_3$		$= A_3\mathbf{w} - 6A^2\mathbf{w} + 12A\mathbf{w} - 8\mathbf{w}$	$A^2\mathbf{w} - 4A\mathbf{w} + 4\mathbf{w}$	
		$+5A\mathbf{v} - 10\mathbf{v} + A\mathbf{u} + 3\mathbf{u}$	$+5\mathbf{v} + \mathbf{u}$	$5\mathbf{u}$
$\mathbf{x}_4$	$= \mathbf{q}_1$	$= A\mathbf{u} - 4\mathbf{u}$	$\mathbf{u}$	$-2\mathbf{u}$
x_5	$= 2\mathbf{q}_3 + 5\mathbf{q}_2$	$= 2A^2\mathbf{w} - 8A\mathbf{w} + 8\mathbf{w}$		
		$+5A\mathbf{v} - 10\mathbf{v} + 17\mathbf{u}$	$2A\mathbf{w} - 4\mathbf{w} + 5\mathbf{v}$	$17\mathbf{u}$
$\mathbf{x}_6$	$= \mathbf{q}_2 - \mathbf{q}_4$	$= -A\mathbf{v} - 4\mathbf{v} + 2\mathbf{u}$	$\mathbf{v}$	$-2\mathbf{v} + 2\mathbf{u}$
$\mathbf{x}_7$	$= 2\mathbf{q}_5 + 17\mathbf{q}_2$	$= 4A\mathbf{w} - 8\mathbf{w} + 17A\mathbf{v} - 58\mathbf{v} + 51\mathbf{u}$	$4\mathbf{w} + 17\mathbf{v}$	$-24\mathbf{v} + 51\mathbf{u}$
$\mathbf{x}_8$	$= 2\mathbf{q}_7 - 24\mathbf{q}_6 + 27\mathbf{q}_2$	$= 8\mathbf{w} + 27A\mathbf{v} - 98\mathbf{v} + 81\mathbf{u}$	$27\mathbf{v}$	$8\mathbf{w} - 44\mathbf{v} + 81\mathbf{u}$

The jth row of the table contains the constructed expressions $\mathbf{x}_j$, together with their quotients and remainders on division by $A - 2I$. Except for the first three rows, which contain the dependence expressions, each $\mathbf{x}_j$ is accompanied by its derivation as a linear combination of quotients. For example, since $2\mathbf{r}_5 + 17\mathbf{r}_2 = \mathbf{o}$, the combination $\mathbf{x}_7 = 2\mathbf{q}_5 + 17\mathbf{q}_2$ is adjoined to the table. A horizontal rule separates the dependence expressions from the expressions which evaluate to generalized eigenvectors and further rules separate expressions which, as generalized eigenvectors, have different orders.

The table shows that a basis for the generalized eigenspace for the eigenvalue 2 consists of two 1-eigenvectors, $\mathbf{x}_4 = (1, 1, 5, 4, 2, -2, 6)$ and $\mathbf{x}_5 = (2, 4, -2, -2, -1, 1, -3)$; two 2-eigenvectors, $\mathbf{x}_6 = (3, 2, -2, -2, -1, 1, -3)$ and $\mathbf{x}_7 = (72, 40, -22, -30, -9, 17, -39)$; and one 3-eigenvector $\mathbf{x}_8 = (108, 64, -54, -46, -27, 27, -81)$. There are no further generalized eigenvectors for the eigenvalue 2 because the last remainder $\mathbf{r}_8$ is not a linear combination of earlier remainders; so the table cannot be extended.

4. Concluding remarks

Who killed determinants?—May [**10**]

In most developments of the eigenvalue problem, the characteristic polynomial of the $n \times n$ matrix A is defined as $\det(tI - A)$ and used to obtain the eigenvalues of A, which numbers were then used to obtain eigenvectors. Here the characteristic polynomial plays no role. However, Cater [**4**] and Axler [**1**] show that the characteristic polynomial can be defined without determinants, as $\prod_\lambda (t-\lambda)^{d(\lambda)}$, where the product is extended over the distinct eigenvalues λ of A and $d(\lambda)$ is the dimension of the generalized eigenspace for λ, which is equal to the dimension of the null space of $(A - \lambda I)^n$.

The characteristic polynomial can, however, be obtained directly from the dependence relations without first finding the eigenvalues. Except for a nonzero scalar factor to make it monic, it is the product of the polynomials $P_r(t)$ introduced at the end of Section 2. Details can be found in McWorter [**11**].

REFERENCES

1. Sheldon Axler, Down with determinants, *Amer. Math. Monthly* **102** (1995), pp. 139–154.
2. Albert A. Bennett, Construction of a rational canonical form for a linear transformation, *Amer. Math. Monthly* **38** (1931), pp. 377–383.
3. F. S. Cater, An elementary development of the Jordan canonical form, *Amer. Math. Monthly* **69** (1962), pp. 391–393.
4. F. S. Cater, *Lectures on Real and Complex Vector Spaces*, Saunders, Philadelphia, PA, 1966.
5. A. M. Danilevskiĭ, On the numerical solution of the secular equation (Russian), *Matem. Sb.* (*Rec. Math.*), vol. 2, 44 (1937), pp. 159–162.
6. D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra* (Russian), Fizmatgiz, Moscow, 1960; Eng. transl.: W. H. Freeman, San Francisco, CA, 1963.
7. G. Kowalewski, *Einführung in die Determinantentheorie*, 3rd ed., Walter de Gruyter, Berlin, 1942; reprinted by Dover Pubs., New York, NY, 1948. See esp. pp. 298–314.
8. G. Kowalewski, Natürliche Normalformen linearer Transformationen, *Ber. Sächs. Ges. d. Wissenschaften*, Math–Phys. Kl., vol. 69 (1917), pp. 325–335.
9. A. N. Krylov, On the numerical solution of the equation which in technical questions determines the oscillation frequencies of material systems (Russian), *Izv. Akad. Nauk SSR*, *Otdel. Mat. i Estest. Nauk*, (1931), pp. 491–539.
10. O. May, Who killed determinants?, talk given at MAA session on Jan. 28, 1966; abstract in *Amer. Math. Monthly* 73 (1966), p. 440. Also available as film.
11. W. A. McWorter, Jr., An algorithm for the characteristic polynomial, this MAGAZINE 56 (1983), pp. 168–175.
12. E. Nering, *Linear Algebra and Matrix Theory*, Wiley, New York, NY, 1963; 2nd ed., 1970.

Variations on a Theme of Newton

ROBERT M. CORLESS
University of Western Ontario
London, Ontario
Canada N6A 5B7

Introduction

We use a particularly simple example function[1], and the computer algebra system Maple, to try to learn something about Newton's method. The discussion here presumes only a minimal amount of calculus—including the standard introduction to Newton's method, such as is found in [**2**, Sec. 2.10]—and some algebraic fluency. This discussion, though aimed at undergraduate students, contains surprises (perhaps even for instructors), items not found in the usual calculus course, and pointers to many more such items. The intention is to provoke or reinforce an interest in pure and applied mathematics. If this works, *everyone* will take something new away.

Newton's method

Newton's method is for approximately solving nonlinear equations $f(x)=0$. Applied mathematics problems usually lead to nonlinear equations—we cannot rely on everything being linear. Some examples of applied problems requiring Newton's method or an equivalent are:

- so-called "implicit" numerical methods for the solution of ordinary differential equations.
- practically any engineering design problem, where instead of being asked to calculate the behavior of a machine or system as given, you are asked to calculate the design parameters that will make the system behave in a certain desired way. For example, many problems in robotic control fall into this category.
- computer-aided design uses piecewise polynomials to model physical objects. Calculating their intersection points requires the solution of systems of polynomial equations. Even if initial approximations to the solutions are arrived at by other means, Newton's method can be used to "polish" the roots.

The basic idea behind Newton's method is that if you can't solve $f(x)=0$ for x, replace f with a simpler function F, namely, the best linear approximation to $f(x)$ near some initial guess point x_0. This approximation is $F(x)=f(x_0)+f'(x_0)(x-x_0)$, and we *can* solve $F(x)=0$ to get $x_1=x_0-f(x_0)/f'(x_0)$, provided $f'(x_0)\neq 0$. Repeating this with the new approximation x_1 to get x_2 and so on gives us the iterative formula

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}.$$

[1] Our example function $f(x)=x^2-a$ is indeed particularly simple, and this is important: if it were not so simple, we wouldn't be able to go anywhere near as far as we do. Hold on to your seat!

We will explore this formula with an extremely simple nonlinear function, namely $f(x)=x^2-a$, in order to learn something about Newton's method (and some computer tools). It is clear that the zeros of $f(x)=x^2-a$ are just $x=\sqrt{a}$ and $x=-\sqrt{a}$, so Newton's method is not really required for this problem. Even worse, we are later going to specify $a=1$, so we will be using Newton's method to find the square root of 1! Our iteration is, for general a,

$$x_{n+1}=x_n-\frac{x_n^2-a}{2x_n} \tag{1}$$

or, mathematically equivalent but slightly less numerically stable,

$$x_{n+1}=\frac{1}{2}\left(x_n+\frac{a}{x_n}\right).$$

A Maple program The following program, written in the computer algebra language Maple (see [1], for example, for an accelerated introduction to Maple), will be used to compute iterates of Newton's method for the rest of this discussion. The routine normal just simplifies expressions.

```
Newton := proc(a, x0, n)  local xn;
  xn := x0;
  to n do
    xn := normal(xn- (xn ^2 - a) / (2*xn))
  od
end:
```

Numerical tests If we choose $a=2$, then our function is $f(x)=x^2-2$ and we are looking for $\sqrt{2}$. Choosing an initial guess of $x_0=1$, the program Newton produces Table 1.

TABLE 1 Newton iterates of $f(x)=x^2-2$.

n	x_n	error
0	1	-1.0
1	3/2	$2.5\cdot 10^{-1}$
2	17/12	$6.0\cdot 10^{-3}$
3	577/408	$6.0\cdot 10^{-6}$
4	665857/470832	$4.5\cdot 10^{-12}$
5	886731088897/627013566048	$2.5\cdot 10^{-24}$

REMARKS

1. The error reported in the above table is the so-called "residual" error $r_n=f(x_n)$. If r_n is zero, then of course x_n is a root; if r_n is "small," then, in some sense, x_n is "close" to a root. This type of measure of accuracy is always available, even when the exact answer is not known. For "well-conditioned" problems it gives the same information as the difference between the approximate answer and the true answer; this problem is well-conditioned because $x_n-\sqrt{a}=(x_n^2-a)/(x_n+\sqrt{a})\approx r_n/(2\sqrt{a})$ and so the *relative* error here ($a=2$) is about $(x_n-\sqrt{a})/\sqrt{a}\approx r_n/4$.
2. Exact arithmetic *costs a lot*. We notice that the *length* of the answer approximately doubles each time; a quick calculation shows that the answer after 30

iterations would take a few gigabytes of memory to store. This is why people instead use arithmetic with a fixed number of decimals (*i.e.*, floating-point).

3. We can simplify our problem by the *nondimensionalization*[2] $u_n = x_n/\sqrt{a}$, at least for the purpose of understanding what is happening. Of course, for actual calculations we can't nondimensionalize by $\sqrt{a}$ which we don't know. If we use this conceptual scaling, then the Newton iteration becomes

$$u_{n+1} = \frac{1}{2}\left(u_n + \frac{1}{u_n}\right).$$

This is exactly the same iteration but with $a = 1$. Thus the scaled iteration uses Newton's method to compute the square root of 1. But the relative error in x_n is $(x_n - \sqrt{a})/\sqrt{a} = u_n - 1$ and so this iteration really does tell us something about Newton's method, and we will keep it in mind. It is easy to see that if $u_n = 1 + e_n$ where e_n represents the error after n iterations, then

$$e_{n+1} = \frac{e_n^2}{2(1+e_n)} \approx \frac{1}{2}e_n^2.$$

This is called *quadratic convergence.* Using this formula shows that after about 30 iterations we will have about 1 billion digits of $\sqrt{a}$ correct, if we start with roughly one correct digit.

4. If we convert these rational numbers to "continued fraction form" (using the Maple routine `convert(17/12, confrac)`) where a *continued fraction* is something of the form

$$n_0 + \cfrac{1}{n_1 + \cfrac{1}{n_2 + \cfrac{1}{\ddots}}}$$

we see the quite remarkable patterns

$$1 = 1$$

$$3/2 = 1 + \frac{1}{2}$$

$$17/12 = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2}}}$$

$$577/408 = 1 + \cfrac{1}{2 + \cfrac{1}{\ddots + \cfrac{1}{2}}}$$

where the length of the continued fraction is 2^n, and every entry is 2. This is the beginning of an interesting foray into number theory.

[2] If a has units, say square meters, this scaling removes them.

Symbolic initial guess If the program Newton given earlier had been written in C or FORTRAN, then calling it with a symbol (say g) for the initial guess would generate an error message. But here,

```
> Newton(1, g, 1);
```

in Maple, returns $(g^2+1)/(2g)$. We can ask Maple to continue, giving the results in Table 2.

TABLE 2 Newton iterates for x^2-1 with a symbolic initial guess, g.

n	x_n
0	g
1	$\frac{g^2+1}{2g}$
2	$\frac{1}{4}\frac{g^4+6g^2+1}{g(g^2+1)}$
3	$\frac{1}{8}\frac{g^8+28g^6+70g^4+28g^2+1}{g(g^4+6g^2+1)(g^2+1)}$

In FIGURE 1 we plot the first few results from Newton. We see that these rational functions are trying to approximate a step function; as n increases, we see clear evidence that these functions converge. The moral of this section is that the error message that FORTRAN would have given us would have concealed an insight, namely that the result of n iterations of Newton's method is a rational function of the initial guess g. Further, we have learned that this rational function looks (for large n) rather like a step function with heights $\pm\sqrt{a}$. Note that the graph in FIGURE 1 works for all a because the axes are scaled—the horizontal axis is the $g/\sqrt{a}$ axis and the vertical axis is the $x_n/\sqrt{a}$ axis.

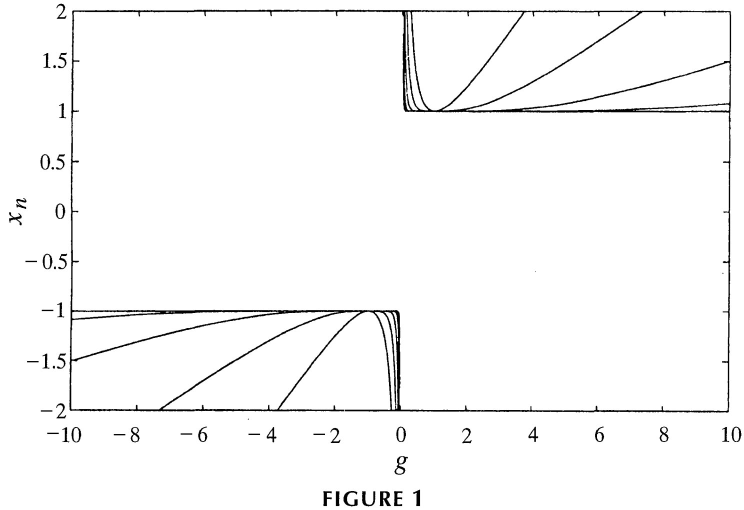


FIGURE 1

Newton iterates with a symbolic initial guess, plotted together. As n increases we must have $x_n/\sqrt{a} \to \pm 1$, and we can see that the convergence is rapid near $g/\sqrt{a} = \pm 1$, as we expect.

Symbolic a Now let us choose instead $x_0 = 1$ (we will discuss this choice of initial guess in a moment) and look at the results from Maple if we input a symbolic a to the program. The first few of these are presented in Table 3.

TABLE 3 Rational approximations obtained by using a symbolic *a*.

n	x_n
0	1
1	$\frac{1}{2}+\frac{1}{2}a$
2	$\frac{1}{4}\frac{1+6a+a^2}{1+a}$
3	$\frac{1}{8}\frac{1+28a+70a^2+28a^3+a^4}{(1+6a+a^2)(1+a)}$

When we plot these[3], we get a sequence of rational (in a) approximations to $\sqrt{a}$, as is quite evident in FIGURE 2.

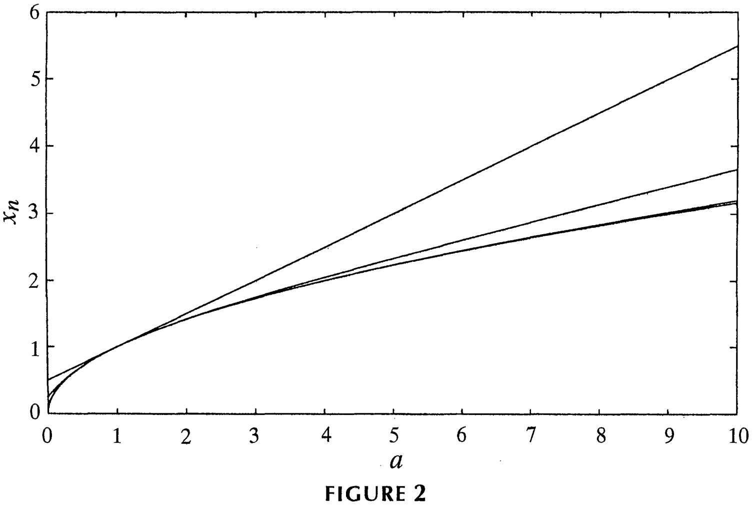


FIGURE 2
The first few iterates of Newton's method on $f(x) = x^2 - a$ with symbolic a give quite good rational approximations to $\sqrt{a}$.

REMARKS

1. Nondimensionalization shows that choosing $x_0 = 1$ is perfectly general. Put $x_n = x_0 v_n$ in equation 1, and simplify to get

$$v_{n+1} = \frac{1}{2}\left(v_n + \frac{a/x_0^2}{v_n}\right).$$

[3] Both FIGURE 1 and FIGURE 2 were actually prepared using Matlab, not Maple, because Matlab plots look slightly nicer; moreover, the graphs were generated by giving a *vector* of g values and a *vector* of a values to a Matlab implementation of Newton's method, much like the Maple symbolic version.

This is just the iteration for finding the square root of a/x_0^2. Therefore, the graph in FIGURE 2 differs from the graph of the approximations we would get with some other initial guess (say $x_0 = 2$) *only in the scale of the axes*—in particular the shape of the curves remains the same. If we label the y-axis with x_0 where 1 is now, and likewise $2x_0$ for 2 and so on, and label the x-axis with x_0^2 where 1 is now, etc., then FIGURE 2 represents the first few iterates of the general case. That is, all the curves with general initial guess *collapse onto the same curve*. This shows the true power of nondimensionalization.

2. We can replace `normal` in the routine `Newton` with a call to Maple's `series` command, and execute Newton's method in the domain of power series. Quadratic convergence in this domain means that the number of correct terms in the power series doubles each time.
3. We can show with Maple that the error in our rational functions of a above are proportional to $(a-1)^{2^k}$; for example, after three iterations the error is

$$f_3(a)^2 - a = \frac{1}{64}\frac{(a-1)^8}{(1+6a+a^2)^2(1+a)^2}.$$

   As before the difference between $f_3(a)$ and $\sqrt{a}$ will be about $1/(2\sqrt{a})$ times this.
4. We can convert the rational approximations in Table 3 to continued fraction form; indeed these approximations are one step towards *approximation theory* which underlies much of scientific computing.
5. Again FORTRAN would give us an error message if we tried this in that language. We begin to suspect that whenever a language gives us an error message, there is something to learn.

Chaotic dynamics

Now we choose $a = -1$ and see what happens. We are trying to find an x such that $x^2 + 1 = 0$, and if we start with a real x_0 we are doomed to failure. However, the failure is very interesting.

A few experiments show us that some initial guesses ($x_0 = 0$, $x_0 = 1$, $x_0 = 1 - \sqrt{2}$, etc.) lead to division by zero. We ignore these minor annoyances. A few more experiments show that most initial guesses don't lead (immediately) to division by zero, but rather wander all over the x-axis, without showing any kind of pattern.

Since the x_n appear random in this case, we consider looking at a *frequency distribution* of them. We divide the axis up into bins—the bins are chosen according to a rule given by an advanced theory, namely a rule depending on the theoretical probability density function—and count the number of x_n that appear in each bin. The results appear in Table 4.

To explain the theoretical probability density function would take us to the boundaries of *ergodic theory*, which is a "main artery," if you will, of statistical mechanics, dynamical systems, and indeed probability theory.

Symbolic n If we call the Maple program not with symbolic a or x_0 but rather with symbolic n, the number of iterations, we get the error message

```
Error, (in Newton) unable to execute for loop.
```

TABLE 4 Frequency distribution for x_n where $x_{n+1} = (x_n - 1/x_n)/2$ and $x_0 = 0.4$ (10,000 iterates). The bin boundaries b_k, $0 \le k \le 10$ are chosen so that $b_0 = -\infty$ and $\int_{b_{k-1}}^{b_k} 1/(\pi(x^2+1))\,dx = 1/10$. According to theory, there should be roughly the same number of x_n in each bin.

k	b_k	number of x_n in (b_{k-1}, b_k)
1	−3.0777	1001
2	−1.3764	999
3	−0.7265	1006
4	−0.3249	1000
5	0.0000	986
6	0.3249	1000
7	0.7265	1007
8	1.3764	980
9	3.0777	986
10	∞	1035

As we have discovered, an error message indicates that we have something to learn. Maple might not be able to do this problem for a symbolic n, *but we can* (in this case). Assume first that $u_0 > 1$ (this corresponds to $x_0 > \sqrt{a}$). Put $u_n = \coth\theta_n$. (The *hyperbolic functions* $\sinh\theta = (\exp(\theta) - \exp(-\theta))/2$, $\cosh\theta = (\exp(\theta) + \exp(-\theta))/2$, $\tanh\theta = \sinh\theta/\cosh\theta$ and so on, are strongly related to the ordinary trig functions.) We have

$$\begin{aligned}\coth\theta_{n+1} = u_{n+1} &= \frac{1}{2}\left(\frac{\cosh\theta_n}{\sinh\theta_n} + \frac{\sinh\theta_n}{\cosh\theta_n}\right)\\ &= \frac{\cosh 2\theta_n}{\sinh 2\theta_n}\\ &= \coth 2\theta_n\end{aligned}$$

where we have used $\cosh^2\theta + \sinh^2\theta = \cosh 2\theta$ and $2\sinh\theta\cosh\theta = \sinh 2\theta$ to simplify. Taking $\coth^{-1}$ of both sides, we see $\theta_{n+1} = 2\theta_n$, which is easily solved to get

$$\theta_n = 2^n\theta_0.$$

Therefore $u_n = \coth(2^n\theta_0)$, if $u_0 > 1$.

For the case when $0 < u_0 < 1$, we note that we will immediately have $u_1 = (u_0 + 1/u_0)/2 > 1$ (for example, by elementary calculus we see the minimum of u_1 occurs when $u_0 = 1$). Thereafter the previous analysis applies. The case of $u_0 < 0$ is symmetric to the positive case. So we can say $u_n = \coth 2^{n-1}\theta_1$, regardless of what u_0 is.

Similarly, it is an elementary exercise to show in the complex case, with $a = -1$, that $u_n = \cot(\theta_n)$ gives $\theta_{n+1} = 2\theta_n$ or

$$u_n = \cot(2^n\theta_0).$$

This lays bare all of the chaotic dynamics of this iteration in the complex case. See [3] for more discussion of this case.

Remarks

1. Now we have the solution for symbolic n, we can answer the question "What do you get if you do half a Newton iteration?" For this problem, we get $u_{3/2} = \coth(\sqrt{2}\,\theta_1)$ (by definition). This doesn't have any apparent application, but in more complicated dynamical systems finding such an interpolation is very useful indeed.

2. The *Lyapunov exponent* in the chaotic case is ln 2. The formula (3) also tells us how to find the theoretical probability density function alluded to earlier.
3. No "fractals" appear in this problem, unless it is on the imaginary axis (where the chaos is). However, looking at Newton's method for solving $f(x) = x^3 - 1 = 0$, we get fractals in $\mathbb{C}$ immediately. See **[3]**, and the other papers in that same issue of the *College Mathematics Journal.*
4. The "asymptotics" of $\coth 2^n \theta_0$ tell us how quickly the iterates approach 1. By Maple,

$$u_n = 1 + 2e^{-1 \cdot 2^n \theta_1} + 2e^{-2 \cdot 2^n \theta_1} + O(e^{-3 \cdot 2^n \theta_1})$$

which tells us everything about how fast u_n approaches 1 (and by extension how fast x_n approaches $\sqrt{a}$).

Concluding remarks

In this discussion we have stepped outside the normal route to mathematics. By asking just slightly different questions about Newton's method than is usual in a calculus class —using a very simple example, just trying to understand it better—we have used or discovered links to nondimensionalization, numerical analysis, complexity theory, continued fractions, approximation theory, series algebra, asymptotics, ergodic theory, and dynamical systems (chaos and fractals). One hopes the student will be stimulated to search out other references on these subjects (one might begin with the references in **[3]**, and the other papers in that same issue of the *College Mathematics Journal*).

The discussion in this paper also suggests that it might have been premature to drop Newton's method (for computing the square root) from the high-school curriculum, as it has been dropped in some districts, merely because calculators can compute square roots with the press of a button. The important thing may not ever have been to compute a square root, but rather to provide a nice introduction to Newton's method, from which "central trunk" we may move on to other significant areas of modern mathematics.

Probably the most significant concept used in this discussion is nondimensionalization. From a practical viewpoint, it is an invaluable tool in the management of large numbers of variables; from the pure mathematical viewpoint it is an overture to the theory of symmetry, itself a vigorous and powerful branch of modern mathematics.

But even just on its own, Newton's method is an extremely important and well-studied tool in applied mathematics, used every day for the solution of systems of nonlinear equations. It is surprising how easy it is to find new questions to ask about it.

Acknowledgment. Many of these ideas are due to Charles M. Patton, and I first heard them in his workshop at the 4th International Conference on Technology in Education, in Portland, Oregon, 1991. This paper also benefited from discussions with Peter Poole, David Jeffrey, and Bob Bryan.

REFERENCES

1. Robert M. Corless, *Essential Maple*, Springer-Verlag, New York, NY, 1994.
2. J. Stewart, *Calculus*, Brooks/Cole, Pacific Grove, CA, 1991.
3. Gilbert Strang, A chaotic search for *i*, *College Math. Journal* 22 (1991), pp. 3–12.

NOTES

Differentiating Among Infinite Series

RICK KREMINSKI
Texas A&M University–Commerce
Commerce, TX 75429

Introduction Calculus students often spend a lot of time deciding whether or not a series like $\Sigma_1^\infty 1/j^{3/2}$ converges. But relatively little time is spent investigating the numerical *values* of such (convergent) series. Can't we just use a computer, many students wonder, and keep adding more and more terms until we "see" what the limit is? For many (rapidly converging) series, this logic is, of course, essentially valid. But there are also many series whose partial sums converge very slowly. For $\Sigma_1^\infty 1/j^{3/2}$, for instance, the 6-digit accuracy we will get below from one of our estimation formulae would be attained by a partial sum only after 160 billion terms were added. As a more dramatic example, we will consider the excruciatingly slow convergence of $\Sigma_2^\infty 1/(j\,(\ln j)^2)$; not even the addition of 10^{10000} terms would match the five digits of accuracy we will obtain by our first, most basic method. Using our formulae, we can accurately estimate the values of such slowly convergent series, provided we have a minute's time and adequate computing power—a generic scientific calculator will do just fine. Our formulae, moreover, are based almost completely on something calculus students are familiar with (**[1]**, **[2]**, **[3]**):

$$\frac{F(j+h)-F(j-h)}{2h} \approx F'(j). \tag{1}$$

Many students are surprised to learn that (1), something useful in estimating derivatives, can be used to estimate the value of certain series (including many alternating series). In this note, we first deduce some series estimation formulae, then illustrate their use in a few examples, and then find other formulae based on generalizations of (1). Next, we give a brief discussion of error bounding. Finally, we include some comments on approximating Euler's γ constant.

Derivation of some summation estimation formulae Consider a convergent series of the form $\Sigma f(j)$, where we assume that f is continuous and integrable on some interval of the form $[m,\infty)$. Let a_j denote $f(j)$ and let f denote an antiderivative of f. Of course, f is only well-defined up to a constant (for instance, if we consider $\Sigma 1/j^2$, then $f(j) = 1/j^2$ and $F(j) = -1/j + C$). Now $F'(j) = f(j) = a_j$ is approximated by (1); setting $h = 1$ and letting j be $k+1$, $k+2$, etc., we have

$$\begin{array}{rrrrrrrl}
 & & & & F(k+2) & - F(k) & & \approx 2a_{k+1} \\
 & & & F(k+3) & - F(k+1) & & & \approx 2a_{k+2} \\
 & & F(k+4) & - F(k+2) & & & & \approx 2a_{k+3} \\
 & F(k+5) & - F(k+3) & & & & & \approx 2a_{k+4} \\
F(k+6) & - F(k+4) & & & & & & \approx 2a_{k+5} \\
 & & & & & & \cdots\cdots\cdots\cdots &
\end{array} \tag{2}$$

Observe all the cancellation that occurs when we add the first p approximate equations. The result is

$$F(k+p+1)+F(k+p)-F(k+1)-F(k)\approx 2(a_{k+1}+a_{k+2}+\cdots+a_{k+p}).$$

Note that the left side is indeed well-defined, even though F is only defined up to a constant. Taking the limit as p tends to infinity (which really amounts to adding up *all* the approximate equations in (2)), we see that a quantity like $F(\infty)+F(\infty)$ remains on the left hand side of the expression above. To avoid cluttering up our results, we agree to choose the constant of integration in F so that $F(\infty)=0$ (this is possible since f is integrable). Therefore (2) ultimately yields

$$-F(k+1)-F(k)\approx 2(a_{k+1}+a_{k+2}+a_{k+3}+\cdots). \tag{3}$$

Equation (3) provides us with a way to approximate the "tail" of the series $\Sigma\, a_j$. Let S denote $\Sigma_1^\infty a_j$ and s_k denote the partial sum $\Sigma_{j\le k} a_j$, so that $S-s_k=a_{k+1}+a_{k+2}+a_{k+3}+\cdots$ is the truncation error (or "tail") when s_k is taken as an estimate for S. Rearranging (3) and dividing by 2 leads to our first summation formula:

$$S\approx s_k-\frac{F(k+1)+F(k)}{2}, \tag{4}$$

for F vanishing at ∞.

Before supplying examples, we observe that a similar approximation method can be obtained as follows. Using (1), but this time with $h=1/2$ instead of $h=1$, gives

$$\begin{aligned}
&&&F(k+3/2)-F(k+1/2)\approx a_{k+1}\\
&&F(k+5/2)-F(k+3/2)&\qquad\approx a_{k+2}\\
&F(k+7/2)-F(k+5/2)&&\qquad\approx a_{k+3}\\
F(k+9/2)-F(k+7/2)&&&\qquad\approx a_{k+4}\\
&&&\qquad\cdots\cdots\cdots\cdots
\end{aligned}$$

Again, we add these approximate equations and get almost complete cancellation on the left hand side. Heuristically, $F(\infty)-F(k+\frac{1}{2})\approx S-s_k$; this gives our second approximation formula: For F vanishing at ∞,

$$S\approx s_k-F(k+1/2). \tag{5}$$

Example. We estimate $S=\Sigma_1^\infty 1/j^2$. For simplicity, we'll use $k=20$; any reader with a calculator can implement both (4) and (5). Then

$$s_k=s_{20}=1+\frac{1}{4}+\frac{1}{9}+\cdots+\frac{1}{400}=1.596163\ldots\,.$$

To avoid further use of ellipses in truncated numerical values, we'll use "(ad)" to denote "accurate to all digits displayed"; so $s_{20}=1.596163$(ad). Now since $f(j)=1/j^2$, we have $F(j)=-1/j$, so equation (4) yields

$$S\approx s_{20}-\frac{-1/21+-1/20}{2}=1.64497276\,(\text{ad}).$$

Similarly, (5) yields

$$S\approx s_{20}-\frac{-1}{20+1/2}=1.6449437\,(\text{ad}).$$

How good are these approximations? It is well known from Fourier series or complex analysis (or by purely elementary means involving trigonometric identities, as in [4]) that $\Sigma_1^\infty 1/j^2 = \pi^2/6 = 1.64493406(\text{ad})$. (We chose this series as our first example because, for students, we can view the exercise as really trying to estimate π.) From the sum's true value we find that the absolute error in our use of (4) with $k = 20$ is roughly 3.9×10^{-5}; the absolute error using $k = 20$ in (5) is roughly 9.7×10^{-6}. (We shall see below that error analysis predicts that for a generic f, the absolute error in (4) is expected to be about four times the absolute error in (5).) A standard argument using the integral test (exercise!) indicates that for a partial sum of $S = \Sigma_1^\infty 1/j^2$ to be within 9.7×10^{-6} of S, one would have to add over 100,000 terms. Implementing (5), we achieved this accuracy with only 20 terms.

Generalizations using more accurate estimates of derivatives Since (1) concerned first derivatives, we can ask whether other, more accurate estimates for the first derivative also lead to summation schemes. (Later, we will generalize in a different way, by considering numerical estimates for higher order derivatives instead of the first derivative.)

Consider the following more accurate estimate for $F'(k)$, analogous to (1) (for further details, see either **[6]** or Section 4.2 of **[5]**).

$$\frac{-F(j+2h) + 8F(j+h) - 8F(j-h) + F(j-2h)}{12h} \approx F'(j) \tag{6}$$

Letting $h = 1$ and $j = k+1, k+2, \dots,$ we have

$$\begin{array}{lllllllll}
 & & & & -F(k+3)+8F(k+2) & & & -8F(k)+F(k-1) & \approx 12a_{k+1} \\
 & & & -F(k+4)+8F(k+3) & & & -8F(k+1)\ +F(k) & & \approx 12a_{k+2} \\
 & & -F(k+5)+8F(k+4) & & & -8F(k+2)+\ F(k+1) & & & \approx 12a_{k+3} \\
 & -F(k+6)+8F(k+5) & & & -8F(k+3)+\ F(k+2) & & & & \approx 12a_{k+4} \\
-F(k+7)+8F(k+6) & & & -8F(k+4)+\ F(k+3) & & & & & \approx 12a_{k+5} \\
 & & & & & & \cdots\cdots & &
\end{array}$$

Adding the approximate equations and again choosing F to vanish at infinity leads to

$$F(k+2) - 7F(k+1) - 7F(k) + F(k-1) \approx 12(a_{k+1} + a_{k+2} + \cdots),$$

or, equivalently,

$$S \approx s_k + \frac{F(k+2) - 7F(k+1) - 7F(k) + F(k-1)}{12}. \tag{7}$$

Applying (7) to $\Sigma_1^\infty 1/j^2$ with $k = 20$ yields $S \approx 1.64493384(\text{ad})$; the absolute error is roughly 2.2×10^{-7}. To suggest how the error is affected as k increases, we observe that with $k = 60$, (7) yields $S \approx 1.6449340658(\text{ad})$, representing an absolute error of roughly 9.9×10^{-10}. Summing 60 terms and adding some correction terms achieves an accuracy that a partial sum alone would achieve only after summing more than one billion terms.

An alternating series estimate Let A denote the series $\Sigma_1^\infty (-1)^{j+1} a_j = a_1 - a_2 + a_3 - \dots$ and let A_k denote the partial sum $\Sigma_1^k (-1)^{j+1} a_j$. Assume the sequence (a_j) tends to zero. Let f denote a continuous function, vanishing at infinity, such that $f(j) = a_j$. (Interestingly, we no longer need f to be integrable on some interval of the form $[m, \infty)$, even though this was crucial to the derivations of (4), (5), and (7).) Let F

be an antiderivative of f. Begin with the system of approximate equations in (2), and change the signs in every other equation:

$$\begin{array}{rrrrrrcr}
 & & & F(k+2) & & -F(k) & \approx & 2a_{k+1} \\
 & & -F(k+3) & & +F(k+1) & & \approx & -2a_{k+2} \\
 & F(k+4) & & -F(k+2) & & & \approx & 2a_{k+3} \\
-F(k+5) & & +F(k+3) & & & & \approx & -2a_{k+4} \\
 & & & & & \cdots\cdots & \cdots & \cdots\cdots
\end{array}$$

Adding p of these approximate equations, where p is odd, gives $F(k+p+1)-F(k+p)+F(k+1)-F(k)$ on the left hand side. But by the mean value theorem $F(m+1)-F(m)=F'(\xi_m)$ for $\xi_m \in [m, m+1]$. Since $F'=f$ and f tends to zero at infinity, $F(k+p+1)-F(k+p)$ must tend to zero as p gets large. Taking the limit as p tends to infinity leaves $F(k+1)-F(k) \approx 2(a_{k+1}-a_{k+2}+a_{k+3}-a_{k+4}+\cdots)$. But this means that $F(k+1)-F(k) \approx 2(A-A_k)$, provided k is even. Hence, for even k,

$$A \approx A_k + \frac{F(k+1)-F(k)}{2}. \tag{8}$$

Before implementing this approximation scheme, we will obtain a more accurate summation scheme for alternating series using the more accurate estimate of F', discussed above. In the spirit of (7), we have

$$\begin{array}{rrrrrrrrrcr}
 & & & & -F(k+3) & +8F(k+2) & & -8F(k) & +F(k-1) & \approx & 12a_{k+1} \\
 & & & F(k+4) & -8F(k+3) & & +8F(k+1) & -F(k) & & \approx & -12a_{k+2} \\
 & & -F(k+5) & +8F(k+4) & & -8F(k+2) & +F(k+1) & & & \approx & 12a_{k+3} \\
 & F(k+6) & -8F(k+5) & & +8F(k+3) & -F(k+2) & & & & \approx & -12a_{k+4} \\
-F(k+7) & +8F(k+6) & & -8F(k+4) & +F(k+3) & & & & & \approx & 12a_{k+5} \\
 & & & & & & & \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots
\end{array}$$

Once again, adding the approximate equations yields much cancellation, leaving

$$F(k-1)-9F(k)+9F(k+1)-F(k+2) \approx 12(a_{k+1}-a_{k+2}+a_{k+3}-a_{k+4}+\cdots).$$

Hence we obtain, for even k,

$$A \approx A_k + \frac{F(k-1)-9F(k)+9F(k+1)-F(k+2)}{12}. \tag{9}$$

Example. The sum $1-1/3+1/5-1/7+\cdots$ is exactly $\pi/4$, or .7853981633(ad). Here $f(x)=1/(2x-1)$, which is clearly not integrable (so an antiderivative F could not be chosen to vanish at infinity). Nevertheless, (8) with $k=20$ yields $S \approx$ .785408(ad), for an error of approximately 1.0×10^{-5}. (The partial sums would not have this accuracy until over 3000 terms were summed.) Similarly, (9) with $k=20$ yields $S \approx .7853981009$(ad), for an absolute error of roughly 6.2×10^{-8}; achieving this accuracy with partial sums alone would require more than 4 million terms.

Generalizations using higher-order derivatives We use the following numerical estimate for the second derivative of G:

$$\frac{G(k+h)-2G(k)+G(k-h)}{h^2} \approx G''(k). \tag{10}$$

One could motivate this to a calculus student as a double use of (1):

$$G''(x) \approx \frac{G'(x+H)-G'(x-H)}{2\,H} \approx \frac{\dfrac{G(x+2H)-G(x)}{2H}-\dfrac{G(x)-G(x-2H)}{2H}}{2H}$$
$$\approx \frac{G(x+2H)-2G(x)+G(x-2H)}{4H^2}.$$

With $H=h/2$, (10) follows. (A less *ad hoc* derivation of (10), along with error term, is given below, but it requires familiarity with Taylor series.)

Consider then a convergent series of the form $\Sigma f(k)$, where we assume f is continuous and, once again, integrable on $[m,\infty)$ for some m. As before, we denote $f(k)$ by a_k. Since f is integrable, it has an antiderivative, F, vanishing at infinity. Now let G be an antiderivative of F, and let $b_k=G(k)$. (Of course, $G(k)$ is defined only up to a constant.) Now $G''(k)=f(k)=a_k$ is approximated by (10), and, with $h=1$, we have the following approximate identities:

$$\begin{array}{llllll} b_k & -2b_{k+1} & +b_{k+2} & & & \approx a_{k+1} \\ & b_{k+1} & -2b_{k+2} & +\ b_{k+3} & & \approx a_{k+2} \\ & & b_{k+2} & -2b_{k+3} & +\ b_{k+4} & \approx a_{k+3} \\ & & & b_{k+3} & -2b_{k+4}\ +b_{k+5} & \approx a_{k+4} \\ & & & & \cdots\cdots\cdots\cdots\cdots & \end{array}$$

Adding these infinitely many approximate equations, again we obtain almost complete cancellation on the left side. (The terms of the form "$b_{\infty+1}-b_\infty$" indeed disappear. Namely, $b_{m+1}-b_m=G(m+1)-G(m)$, which by the mean value theorem is $G'(\xi_m)$, which in turn is $F(\xi_m)$; and this vanishes at infinity by hypothesis.) We obtain $b_k-b_{k+1}\approx a_{k+1}+a_{k+2}+a_{k+3}+\cdots$, providing yet another way to approximate the "tail" of Σa_k. This simplifies to $b_k-b_{k+1}\approx S-s_k$, or

$$S \approx s_k+b_k-b_{k+1}. \tag{11}$$

Comparing the methods To compare our methods we apply them to two sample problems.

Problem 1. Find $\Sigma_1^\infty 1/j^{3/2}$ to 3 digits past the decimal point.

Solution. After fifteen or twenty seconds of furious computation, a programmable calculator can conclude that the partial sum s_{1000} is 2.549145(ad). Unfortunately, as an approximation to the infinite series, this is not even correct to 1 digit past the decimal. In fact, to attain 3-digit accuracy by direct summation would require over 28 million terms. Using (11) with $k=10$ yields 2.612725(ad). The true value is $2.612\underline{375}3$(ad), so (11) led to an absolute error of approximately 3.5×10^{-4}. With $k=20$, (11) yields $2.612\underline{441}$(ad), (4) yields $2.612\underline{506}$(ad), (5) yields $2.612\underline{408}$(ad), and (7) yields $2.612\underline{374}7$(ad). (We have underlined various portions of the values for ease in comparison.)

Problem 2. Find $\Sigma_2^\infty 1/j(\ln(j))^2$, to 4 digits past the decimal point.

Solution. Direct summation will not achieve this accuracy until more than 10^{10000} terms have been added; this is a *very* slowly convergent series. (At this point in class we observe that there are considerably fewer than 10^{100} quarks and neutrinos, i.e., fewer than 10^{100} "things," in the observable universe.) The true value is $2.1097\underline{42}801$(ad). Using (11) with $k=30$ yields $2.1097\underline{54}$(ad), (5) yields $2.1097\underline{48}$(ad), (4) yields $2.1097\underline{67}$(ad), and (7) yields $2.1097\underline{42}7$(ad).

Error bounding We now develop an error bound for (4), and leave as exercises the analogous derivations of error bounds for our other formulae. Beyond standard ideas from calculus, we require only the (Lagrange form of the) Taylor series error formula, which can be found in almost any calculus text (see, e.g., [**2**] or [**3**]).

Assuming that F''' exists, we have

$$F(j+h)=F(j)+F'(j)h+\frac{F''(j)}{2!}h^2+\frac{F'''(\xi_1)}{3!}h^3 \tag{12}$$

and

$$F(j-h)=F(j)-F'(j)h+\frac{F''(j)}{2!}h^2-\frac{F'''(\xi_2)}{3!}h^3 \tag{13}$$

where $\xi_1 \in [j, j+h]$ and $\xi_2 \in [j-h, j]$. Subtracting equation (13) from (12) and dividing by $2h$ yields

$$\frac{F(j+h)-F(j-h)}{2h}=F'(j)+\frac{F'''(\xi_1)h^3+F'''(\xi_2)h^3}{3!2h}. \tag{14}$$

(Note in passing that if we instead add the equations, we essentially arrive at (10), but now with a precise expression for the error incurred in its use.) Since derivatives satisfy the intermediate value property, the average of two F''' values is another F''' value. Thus we can rewrite (14) as

$$\frac{F(j+h)-F(j-h)}{2h}=F'(j)+\frac{F'''(\xi)h^2}{3!}. \tag{15}$$

Therefore the error "true first derivative − estimate in (1)" is $-F'''(\xi)h^2/6$. (We also report, for future reference, that the analogous error for (6) is $+F^{(5)}(\xi)\mathrm{h}^4/30$. For details, see Section 4.1 of [**5**].)

Now we apply (15), our formula for the error in using (1), to each of the approximate equations in (2). We can replace each " $\approx$ " by " $=$ ", provided that $-2F'''(\xi_m)/6$ is attached to the left side of the m^{th} equation; here $\xi_m \in [k+m-1, k+m+1]$. Then the error in using (4) can be expressed as

$$S-\left(s_k-\frac{F(k+1)+F(k)}{2}\right)=-\sum_{m=1}^{\infty}\frac{f''(\xi_m)}{6}. \tag{16}$$

Before simplifying (16), we make two remarks.

- From (16) alone, we see that if $f''>0$ on $[k,\infty)$ (as when $f(j)=1/j^p$ for $p>1$) the error is negative, so the approximations all exceed S. Furthermore, we see that *as k increases, the estimates from* (4) *approach S monotonically from above.* (And whenever $f^{(4)}>0$ on $[k,\infty)$, the error formula for (6) that we mentioned above implies that *as k increases, the estimates from* (7) *should approach S monotonically from below.*) For examples illustrating these phenomena, look back to any of the series where (4) or (7) were implemented.
- Had we kept the dependence on h, (16) would have a factor of h^2 on the right side. This explains why (4) and (5) had errors differing by a factor of about four in the numerical example on page 44, we used $h=1$ in arriving at (4), but used $h=1/2$ in deriving (5). (Of course, the error in using (4) will not be precisely four times the error in (5), and will depend somewhat on f; the ξ's arising in (16) will in general not be the same as those arising in the analogous sum for the error in (5).)

We now sketch how one could bound the sum in (16). (A similar discussion, but with more details, appears in [7].) Consider the sum in (16) as *two* Riemann sums, one for $\int_k^\infty f''/12$ and the other for $\int_{k+1}^\infty f''/12$. (Two Riemann sums arise naturally, with rectangle width 2, one sum for m even and the other for m odd.) For situations where f'' is positive and decreases, as in $\Sigma_1^\infty 1/j^p$ for $p > 1$, the two Riemann sums are less than right sums for the integrals $\int_{k-2}^\infty f''/12$ and $\int_{k-1}^\infty f''/12$ respectively. These integrals are readily evaluated, and we conclude that one (crude) bound for the error in using (4) is

$$\left|S - \left(s_k - \frac{F(k+1)+F(k)}{2}\right)\right| \le \left|\frac{f'(k-2)+f'(k-1)}{12}\right|$$

for the situation where f'' is positive and decreasing on the interval $[k-2, \infty)$. This bound can be improved; for one approach, see how the analogous error term in [7] is treated. We leave bounds for (5), (7), (8), (9), and (11) as exercises for the reader.

The key to our estimation formulae is that both (1) and (6) express F' in terms of a weighted sum of F-values at a finite number of equally spaced points. The error bounding of (4) that we just completed provides hints for the general case. Omitting details, if the error in the approximation for F' is proportional to $F^{(m)}$, then the error in the associated summation formula will essentially be proportional to $f^{(m-2)}(k)$. From the usual point of view in numerical analysis, one differentiation approximation method for F' is often considered "better" than another if its error term depends on a larger power of the stepsize h, since h is usually chosen to be a fixed number close to zero. But from the viewpoint of developing summation approximations, one differentiation method is "better" than another if its error depends on $F^{(m)}$ for a larger value of m. In this case, the error in the associated summation formula is proportional to$f^{(m-2)}$, and for many slowly converging series, higher derivatives of f tend to zero much more rapidly than lower order derivatives. In this sense (6) is a better differentiation method than (1), since the respective errors depend on $F^{(5)}$ and $F^{(3)}$. Still better differentiation methods can be obtained by applying Richardson extrapolation to certain Taylor series expansions, as described in Chapter 4 of [5]; see also [6]. It is a routine matter to develop corresponding summation approximation schemes for any of these better differentiation methods.

A final example: Euler's constant Let $\gamma_m = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m} - \ln m$. Euler's constant γ is defined as $\lim_{m\to\infty}\gamma_m$. Calculus methods can be used to show that γ exists and is less than 1. (See, e.g., [8].) The constant γ arises, among other places, in infinite product formulas in complex analysis, including in the Γ function [9]. Computing its value from the definition is notoriously ineffective; $\gamma_{20} = 0.60200738$(ad), $\gamma_{1000} = 0.57771558$(ad) and (over 50 seconds later on a 120Mhz Pentium, running *Mathematica*) $\gamma_{100000} = 0.57722066$(ad). We will see that even this last value is barely within 5×10^{-6} of γ. To apply our methods to speed the convergence of the γ_m, begin with

$$\gamma_{p+k} = \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}\right) + \frac{1}{k+1} + \frac{1}{k+2} + \cdots + \frac{1}{k+p} - \ln(k+p).$$

Now consider (2) in the situation where $a_j = f(j) = 1/j$ and $F(j) = \ln j$. Adding the corresponding p approximate equations in (2) gives

$$\frac{\ln(k+p+1) + \ln(k+p) - \ln(k+1) - \ln(k)}{2} \approx \frac{1}{k+1} + \frac{1}{k+2} + \cdots + \frac{1}{k+p}.$$

So, after some algebra,

$$\gamma_{p+k} \approx \gamma_k + \frac{\ln\left(\dfrac{k}{k+1}\right)}{2} + \frac{\ln(k+p+1) - \ln(k+p)}{2}.$$

Taking the limit as $p \to \infty$ yields

$$\gamma \approx \gamma_k + \frac{\ln\left(\dfrac{k}{k+1}\right)}{2}. \qquad (17)$$

Using $k = 20$ we get $\gamma \approx 0.57761230$(ad). In fact, $\gamma = 0.57721566$(ad); so while γ_{20} differs from γ already in the first digit past the decimal, our estimate is accurate to within 0.0004.

We can do better. Using the approximate system that led to (5), the reader can check that $\gamma \approx \gamma_k + \ln(k) - \ln(k + 1/2)$. (Coincidentally, exactly the same approximation method was analyzed in **[10]**.) For $k = 20$, this method yields $\gamma \approx 0.57731477$(ad), for an error of about 0.0001. This represents approximately one fourth the error we obtained from using $k = 20$ in (17), as we expect from earlier discussion. Finally, using the system just prior to (7), we obtain

$$\gamma \approx \gamma_k + \ln(k) + \frac{\ln(k+2) - 7\ln(k+1) - 7\ln(k) + \ln(k-1)}{12}. \qquad (18)$$

Using $k = 20$ in (18), we get $\gamma \approx 0.57721452$(ad), with an error of about 10^{-6}. Notice how this estimates γ more accurately than does γ_{100000}.

Final remarks The error bounds that our methods produce, of the form of a constant times $f^{(m-2)}(k)$, are comparable to error bounds that occur in summation estimation based on the Euler–Maclaurin summation formula (cited in **[7]**). Such summation formulae require values of f's derivative as well as f's antiderivative in producing their estimates of the sum—whereas our formulae only require knowledge of f's derivatives in error bounding (and not in the summation formulae themselves). Still, it is likely that there is some underlying relation between our methods and Euler–Maclaurin-based methods—if, in the latter, derivatives are replaced by finite differences. The precise relation between the two approaches remains to be explored.

Bibliographic remarks Since this article is based on numerical *differentiation*, it complements **[7]**, where $\Sigma f(j)$ is estimated using numerical *integration*. Readers who wish to examine other, recently proposed, methods of accelerating the convergence of series (and methods of estimating Euler's γ constant) might begin with the references in **[7]**. The present approach also shares something with the methods in **[11]** and **[12]**. For another interesting approach to accelerating convergence of alternating series, see **[13]**. In another direction, **[14]** discusses *nonlinear* methods of accelerating series convergence; all the methods we have discussed have been linear in the terms a_k of the series.

Acknowledgment. Thanks to the anonymous referees, who suggested references **[11]**, **[12]**, and **[14]**. Thanks also to *Mathematica*, without which the author doubts that he would have discovered any of the approximation schemes discussed here.

REFERENCES

1. J. Callahan and K. Hoffman, *Calculus in Context*, W. H. Freeman, New York, NY, 1995, pp. 116–117.
2. R. Ellis and D. Gulick, *Calculus with Analytic Geometry*, 5th edition, Harcourt Brace, New York, NY, 1994, p. 137.
3. E. Swokowski, M. Olinick, D. Pence, and J. Cole, *Calculus*, 6th edition, PWS-Kent, Boston, MA, 1994, p. 172.
4. I. Papadimitriou, A simple proof of the formula $\Sigma_{k=1}^{\infty} k^{-2} = \pi^2/6$, *Amer. Math. Monthly* **80** (1973), pp. 424–425.
5. R. Burden and J. Faires, *Numerical Analysis*, PWS-Kent, Boston, MA, 1989.
6. M. Abramowitz and I. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, NY, 1972, p. 914.
7. R. Kreminski, Using Simpson's rule to approximate sums of infinite series, *College Math. Journal* **28** (1997), pp. 368–376.
8. G. Strang, *Calculus*, Wellesley-Cambridge Press, Wellesley, MA, 1991, p. 381.
9. J. Conway, *Functions of One Complex Variable*, second ed., Springer-Verlag, New York, NY, 1978, pp. 176–178.
10. D. DeTemple, A quicker convergence to Euler's constant, *Amer. Math. Monthly* **100** (1993), pp. 468–470.
11. G. M. Phillips, Gregory's method for numerical integration, *Amer. Math. Monthly* **79** (1972), pp. 270–274.
12. S. Libeskind, Summation of finite series—a unified approach, *Two-Year College Math Journal* **12** (1981), pp. 41–50.
13. J. Harper, Estimating the sum of alternating series, *College Math. Journal* **19** (1988), pp. 149–154.
14. D. Shanks, Non-linear transformations of divergent and slowly convergent sequences, *J. Math. and Phys.* 34 (1955), pp. 1–34.

Uniqueness of the Decomposition of Finite Abelian Groups: A Simple Proof

F. S. CATER
Portland State University
Portland, OR 97207

We shall use additive notation for abelian groups. The order of a group G is denoted $|G|$ and the cyclic subgroup generated by $c \in G$ is denoted (c). Let $\mathbb{Z}_n$ denote the additive group of integers modulo n.

Among the results included in many first courses in abstract algebra is the Fundamental Theorem of Abelian Groups:

THEOREM 1. *Let G be a finite abelian group. Then there exist cyclic groups $P_1, P_2, \ldots, P_r$, of respective orders $m_1, m_2, \ldots, m_r > 1$, such that m_j divides m_{j-1} for $j = 2, \ldots, r$ and $G = P_1 \oplus P_2 \oplus \cdots \oplus P_r$.*

There is a variety of proofs of Theorem 1; for example, see references **[1]**, **[2]**, **[3]**, **[4]**, **[6]**, and **[7]**. Actually, there is more to the Fundamental Theorem:

THEOREM 2. *The integers r and $m_1, m_2, \ldots, m_r$ in Theorem 1 are uniquely determined. That is, if it is also the case that $G = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_s$, where the Q_j are cyclic subgroups such that $|Q_j|$ divides $|Q_{j-1}|$ for $j = 2, \ldots, s$, then $r = s$ and $|P_j| = |Q_j|$ for $j = 1, 2, \ldots, r$.*

REFERENCES

1. J. Callahan and K. Hoffman, *Calculus in Context*, W. H. Freeman, New York, NY, 1995, pp. 116–117.
2. R. Ellis and D. Gulick, *Calculus with Analytic Geometry*, 5th edition, Harcourt Brace, New York, NY, 1994, p. 137.
3. E. Swokowski, M. Olinick, D. Pence, and J. Cole, *Calculus*, 6th edition, PWS-Kent, Boston, MA, 1994, p. 172.
4. I. Papadimitriou, A simple proof of the formula $\Sigma_{k=1}^{\infty} k^{-2} = \pi^2/6$, *Amer. Math. Monthly* **80** (1973), pp. 424–425.
5. R. Burden and J. Faires, *Numerical Analysis*, PWS-Kent, Boston, MA, 1989.
6. M. Abramowitz and I. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, NY, 1972, p. 914.
7. R. Kreminski, Using Simpson's rule to approximate sums of infinite series, *College Math. Journal* **28** (1997), pp. 368–376.
8. G. Strang, *Calculus*, Wellesley-Cambridge Press, Wellesley, MA, 1991, p. 381.
9. J. Conway, *Functions of One Complex Variable*, second ed., Springer-Verlag, New York, NY, 1978, pp. 176–178.
10. D. DeTemple, A quicker convergence to Euler's constant, *Amer. Math. Monthly* **100** (1993), pp. 468–470.
11. G. M. Phillips, Gregory's method for numerical integration, *Amer. Math. Monthly* **79** (1972), pp. 270–274.
12. S. Libeskind, Summation of finite series—a unified approach, *Two-Year College Math Journal* **12** (1981), pp. 41–50.
13. J. Harper, Estimating the sum of alternating series, *College Math. Journal* **19** (1988), pp. 149–154.
14. D. Shanks, Non-linear transformations of divergent and slowly convergent sequences, *J. Math. and Phys.* 34 (1955), pp. 1–34.

Uniqueness of the Decomposition of Finite Abelian Groups: A Simple Proof

F. S. CATER
Portland State University
Portland, OR 97207

We shall use additive notation for abelian groups. The order of a group G is denoted $|G|$ and the cyclic subgroup generated by $c \in G$ is denoted (c). Let $\mathbb{Z}_n$ denote the additive group of integers modulo n.

Among the results included in many first courses in abstract algebra is the Fundamental Theorem of Abelian Groups:

THEOREM 1. *Let G be a finite abelian group. Then there exist cyclic groups $P_1, P_2, \dots, P_r$, of respective orders $m_1, m_2, \dots, m_r > 1$, such that m_j divides m_{j-1} for $j = 2, \dots, r$ and $G = P_1 \oplus P_2 \oplus \cdots \oplus P_r$.*

There is a variety of proofs of Theorem 1; for example, see references [1], [2], [3], [4], [6], and [7]. Actually, there is more to the Fundamental Theorem:

THEOREM 2. *The integers r and $m_1, m_2, \dots, m_r$ in Theorem 1 are uniquely determined. That is, if it is also the case that $G = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_s$, where the Q_j are cyclic subgroups such that $|Q_j|$ divides $|Q_{j-1}|$ for $j = 2, \dots, s$, then $r = s$ and $|P_j| = |Q_j|$ for $j = 1, 2, \dots, r$.*

Theorem 2 demonstrates, for example, that $\mathbb{Z}_{32} \oplus \mathbb{Z}_8 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_4$ and $\mathbb{Z}_{32} \oplus \mathbb{Z}_8 \oplus \mathbb{Z}_8 \oplus \mathbb{Z}_2$ are not isomorphic, even though they have the same number of elements. Likewise, $\mathbb{Z}_{27} \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_3$ is not isomorphic to $\mathbb{Z}_{27} \oplus \mathbb{Z}_{27}$.

Unfortunately, the proof—and sometimes even the statement—of Theorem 2 is often omitted from first courses, because the usual proofs depend on developing a great deal of machinery. The main purpose of this note is to provide a simple proof of Theorem 2, one that depends only on results usually proved in a first course in abstract algebra. Along the way, we prove some other useful facts about finite abelian groups.

The key to proving Theorem 2 is the following theorem, which does not seem to be included in most abstract algebra texts (see **[5]**).

THEOREM 3. *Let H be a subgroup of a finite abelian group G. Let $G = P_1 \oplus P_2 \oplus \cdots \oplus P_r$ and $H = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_s$ be the decompositions of G and H described in Theorem 1. Then $s \leq r$ and $|Q_j|$ divides $|P_j|$ for $j = 1, 2, \dots, s$.*

Theorem 3 shows, for example, that the group $\mathbb{Z}_4 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_4$ cannot be isomorphic to a subgroup of $\mathbb{Z}_8 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Likewise the group $\mathbb{Z}_{27} \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_3$ cannot be isomorphic to a subgroup of $\mathbb{Z}_{27} \oplus \mathbb{Z}_{27} \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3$.

The proof of Theorem 3 uses the following facts, whose proofs can be found in most beginning abstract algebra texts (such as **[1]** and **[9]**).

Fact 1. Any subgroup of a cyclic group is a cyclic group.

Fact 2. Let C be a finite cyclic group, let n be a positive integer, and let $nC = \{ng | g \in C\}$. Then nC is a subgroup of C; moreover, $nC = (0)$ if and only if $|C|$ divides n.

Fact 3. If n divides the order of the cyclic group C, then C has a subgroup of order n.

Fact 4. If $G_0 = G_1 \oplus G_2 \oplus \cdots \oplus G_n$, m is a positive integer, and $mG_i = \{mg | g \in G_i\}$, then $mG_0 = (mG_1) \oplus (mG_2) \oplus \cdots \oplus (mG_n)$.

We will eventually use Theorem 3 to prove Theorem 2.

Proof of Theorem 3. The proof is by contradiction. Either let there be an index j for which $|Q_j|$ does not divide $|P_j|$, or let $s > r$.

Case 1. Let J be an index for which $|Q_J|$ does not divide $|P_J|$. We will define a subgroup we will call G_1, and make estimates of its order. From these estimates the required contradiction will emerge. Put $n = |P_J|$, and $m = |nQ_J|$. Then $m > 1$ by Fact 2. Put $G_1 = \{x \in nG | mx = 0\}$. Clearly, G_1 is a subgroup of G. The strategy is to find two inequalities involving $|G_1|$ that lead to an inconsistency. Now

$$nG = (nP_1) \oplus (nP_2) \oplus \cdots \oplus (nP_r),$$

by Fact 4. But $nP_j = (0)$ if $j \geq J$, by Fact 2. It follows that

$$nG = (nP_1) \oplus (nP_2) \oplus \cdots \oplus (nP_{J-1}). \tag{1}$$

Let x be any element of G_1. Say $x = x_1 + x_2 + \cdots + x_{J-1}$ where $x_j \in nP_j$. Then

$$0 = mx = mx_1 + mx_2 + \cdots + mx_{J-1}$$

and hence $mx_1 = mx_2 = \cdots = mx_{J-1} = 0$ because the nP_j form a direct sum. Thus $x_j \in G_1$ for $j = 1, 2, \dots, J-1$. But x was arbitrary in G_1, so

$$G_1 \subset ((nP_1) \cap G_1) \oplus ((nP_2) \cap G_1) \oplus \cdots \oplus ((nP_{J-1}) \cap G_1). \tag{2}$$

Moreover, $(nP_j) \cap G_1$ is cyclic, by Fact 1. For a generator g of $(nP_j) \cap G_1$ we have $mg = 0$, so $|(nP_j) \cap G_1| \leq m$. From (2) we conclude that

$$|G_1| \leq m^{J-1}. \tag{3}$$

Each subgroup Q_i $(i = 1, 2, \ldots, J)$ contains a subgroup T_i isomorphic to Q_J by Fact 3. Thus $nT_i \cong nQ_J$. But $|nQ_J| = m$, so $m(nT_i) \cong m(nQ_J) = (0)$ by Fact 2, and so each $nT_i \subset G_1$. Hence

$$(nT_1) \oplus (nT_2) \oplus \cdots \oplus (nT_J) \subset G_1. \tag{4}$$

Each nT_i has order m, so from (4) we deduce

$$|G_1| \geq m^J. \tag{5}$$

From (3) and (5) we deduce that $m^J \leq m^{J-1}$, which is inconsistent with the fact that $m > 1$. This completes the proof of Case 1.

Case 2. Let $r > s$. The argument is like the proof for Case 1, but with $J = r + 1$ and $n = 1$. We leave the rest to the reader. This completes the proof of Theorem 3.

Proof of Theorem 2. Put $H = G$ in Theorem 3. Then $r \geq s$ and $|Q_j|$ divides $|P_j|$ for $j = 1, 2, \ldots, s$. If we reverse the roles of the two decompositions of G, we find that $s \geq r$ and that $|P_i|$ divides $|Q_i|$ for $i = 1, 2, \ldots, r$. Finally, $r = s$ and $|Q_j| = |P_j|$ for $j = 1, 2, \ldots, s$.

Remark. Observe that prime integers did not enter our arguments, although they do enter the proof of Theorem 2 in most of our references.

REFERENCES

1. John B. Fraleigh, *A First Course in Abstract Algebra*, 5th ed., Addison-Wesley, Reading, MA, 1994.
2. Joseph A. Gallian, *Contemporary Abstract Algebra*, 3rd ed., D.C. Heath, Lexington, MA, 1994.
3. I. N. Herstein, *Topics in Algebra*, 2nd ed., John Wiley & Sons, New York, NY, 1975.
4. Thomas W. Hungerford, *Algebra*, Springer-Verlag, New York, NY, 1980.
5. A. C. Kurosh, *Theory of Groups*, *Volume I*, Chelsea, New York, NY, 1955.
6. Serge Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1992.
7. Walter Ledermann, *Introduction to the Theory of Finite Groups*, Interscience, New York, NY, 1957.
8. D. B. Surowski, The uniqueness aspect of the fundamental theorem of finite groups, *Amer. Math. Monthly*, 102 (1995), 162–163.
9. R. C. Thompson and Adil Yaqub, *Introduction to Abstract Algebra*, Scott, Foresman, Glenview, IL, 1970.
10. Elbert A. Walker, *Introduction to Modern Algebra and Analysis*, Random House, New York, NY, 1987.

Swapping Hats: A Generalization of Montmort's Problem

GABRIELA R. SANCHIS
Elizabethtown College
Elizabethtown, PA 17022-2298

Montmort's Matching Problem The following problem was first proposed by the mathematician Pierre Rémond de Montmort [7] in *Essay d'Analyse sur les Jeux de Hazard*, his 1708 treatise on the analysis of games of chance:

> Suppose you have a deck of N cards, numbered $1, 2, 3, \dots, N$. After shuffling, you draw one card at a time, without replacement, counting out loud as each card is drawn: "$1, 2, 3, \dots$". What is the probability that there will be no coincidence, i.e., no drawing of a card bearing the number just called out?

In Montmort's version of the problem, the deck had 13 cards, so the game was called *Treize*, French for thirteen. The game has also been called *Rencontres* (Coincidences), or Montmort's Matching Problem.

Montmort discusses a generalized version of this problem in his correspondence with Nicholas Bernoulli (1687–1759) from 1710 to 1712; these letters are included in the second edition of Montmort's work on gaming [8]. In the generalization, N cards are drawn from a deck of Ns cards; there are s cards bearing each number from 1 to N. Again, one seeks the probability of at least one coincidence, for which Montmort and Bernoulli find a formula.

Other mathematicians who have generalized and discussed this problem include de Moivre [6], Euler [1], Lambert [4], Laplace [5], and Waring [11]. For a more extensive account of the history of this problem, see [2, pp. 326–345] and [10].

Calculation of $P_m(N)$ Montmort's Matching Problem is often posed in the following more amusing form: N men, attending a banquet, check their hats. When each man leaves he takes a hat at random. What is the probability that at least one man gets his own hat?

If there are no such coincidences, the next best thing might be a two-way swap. So one might ask for the likelihood of no matches but at least one swap, or the likelihood of no matches and no swaps but at least one three-way swap. More generally, one is interested in the probability that for any m from 1 to N, m is the size of the smallest subset of N men who exchange hats among themselves.

We let $P_m(N)$ denote the probability that among N men, m is the size of the smallest subset of men that swap hats. $P_1(N)$, then, is the probability of at least one match, which is Montmort's original problem. The usual way of calculating $P_1(N)$ is to let E_i be the event that the ith man gets his own hat back. Then we use the inclusion-exclusion principle to calculate the probability of at least one match, as follows:

$$\begin{aligned}P_1(N) &= P(\cup_i E_i) = \sum_i P(E_i) - \sum_{i<j} P(E_i E_j)\\ &\quad + \sum_{i<j<k} P(E_i E_j E_k) - \cdots + (-1)^{N+1} P(E_1 E_2 \dots E_N)\\ &= \sum_i \frac{(N-1)!}{N!} - \sum_{i<j} \frac{(N-2)!}{N!} + \sum_{i<j<k} \frac{(N-3)!}{N!} - \cdots + (-1)^{N+1} \frac{1}{N!}\\ &= N\frac{(N-1)!}{N!} - \binom{N}{2}\frac{(N-2)!}{N!} + \binom{N}{3}\frac{(N-3)!}{N!} - \cdots + (-1)^{N+1}\frac{1}{N!}\\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + \frac{(-1)^{N+1}}{N!}.\end{aligned}$$

The series converges to $1 - 1/e \approx 0.63$ as N tends to infinity.

Let us now calculate $P_m(N)$ for some small values of N. If $N = 3$, there are $3! = 6$ ways of distributing the hats. In fact, the sample space is just S_3, the group of permutations of 3 elements. Let (ijk) indicate that the first man gets hat i, the second hat j, and the third hat k. Then our sample space becomes $S_3 = \{(1\,2\,3), (1\,3\,2), (2\,1\,3), (2\,3\,1), (3\,1\,2), (3\,2\,1)\}$. Then

$P_1(3) = P(\text{at least one match}) = P(\{(1\,2\;\;3), (1\,3\,2), (2\,1\,3), (3\,2\,1)\}) = 2/3$
$P_2(3) = P(\text{no matches but at least one swap}) = P(\emptyset) = 0$
$P_3(3) = P(\text{no matches, no swaps, but at least one 3-way swap})$
$= P(\{(2\,3\,1), (3\,1\,2)\}) = 1/3.$

With four men and four hats, there are $4! = 24$ sample points. Let $(i_1\, i_2\, i_3\, i_4)$ represent the outcome where the jth man gets hat i_j. We know that the probability of at least one match is

$$P_1(4) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} = \frac{15}{24} = \frac{5}{8}.$$

In how many ways can we distribute the four hats so that nobody gets his own, but at least one pair of men swaps? Notice that if two men swap, the other two must also swap (since no matches are allowed). Hence we want to count the number of ways of dividing four men into two pairs. This is $\frac{1}{2}\binom{4}{2} = 3$, so $P_2(4) = \frac{3}{24} = \frac{1}{8}$.

In any three-way swap, the fourth man gets his own hat back, so $P_3(4) = 0$. The last possibility is of a four-way swap; it has probability $P_4(4) = 1 - \frac{5}{8} - \frac{1}{8} = \frac{1}{4}$.

We now give a general formula for $P_m(N)$:

THEOREM 1. $P_m(N) = \displaystyle\sum_{k=1}^{\lfloor N/m \rfloor} \frac{(-1)^{k+1}}{m^k\, k!}\left(1 - \sum_{i=1}^{m-1} P_i(N - mk)\right).$

Proof. Let $E_{i_1 i_2 \dots i_m}$ be the event where men $i_1, \dots, i_m$ exchange hats among themselves and no smaller subset of men exchange hats. Then, by the inclusion-exclusion principle,

$$P_m(N) = \sum_{k=1}^{\lfloor N/m \rfloor} (-1)^{k+1} \sum P(E_{i_{11}\dots i_{m1}} E_{i_{12}\dots i_{m2}} \cdots E_{i_{1k}\dots i_{mk}}),$$

where the second summation is taken over all possible choices of disjoint subsets $\{i_{11},\dots,i_{m1}\},\dots,\{i_{1k},\dots,i_{mk}\}$ of $\{1,2,\dots,N\}$.

For a specific choice of these subsets, let us calculate the probability that each subset of men exchange hats among themselves and no smaller subset of these men exchange hats among themselves. This means that the men in $\{i_{1j},\dots,i_{mj}\}$ exchange hats in some cyclical manner. There are $(m-1)!$ such cyclical permutations of each subset $\{i_{1j},\dots,i_{mj}\}$. For a specific choice, the probability of the chosen cyclical permutations occurring among the members of the subsets, with no j-way swaps $(j<m)$ occurring among the remaining $N-mk$ men, is $\frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-mk+1}$ $(1-\sum_{i=1}^{m-1}P_i(N-mk))$.

Therefore

$$P_m(N)=\sum_{k=1}^{\lfloor N/m\rfloor}(-1)^{k+1}\sum((m-1)!)^k\frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-mk+1}$$
$$\times\left(1-\sum_{i=1}^{m-1}P_i(N-mk)\right)$$
$$=\sum_{k=1}^{\lfloor N/m\rfloor}(-1)^{k+1}\sum((m-1)!)^k\frac{(N-mk)!}{N!}\left(1-\sum_{i=1}^{m-1}P_i(N-mk)\right),$$

where the second summation is over all possible choices of disjoint subsets $\{i_{11},\dots,i_{m1}\},\dots,\{i_{1k},\dots,i_{mk}\}$ of $\{1,2,\dots,N\}$. The number of choices of such disjoint subsets is

$$\binom{N}{m}\binom{N-m}{m}\cdots\binom{N-mk+m}{m}\frac{1}{k!}=\frac{N!}{k!(m!)^k(N-mk)!}.$$

Therefore, finally,

$$P_m(N)=\sum_{k=1}^{\lfloor N/m\rfloor}(-1)^{k+1}\frac{N!}{k!(m!)^k(N-mk)!}((m-1)!)^k$$
$$\times\frac{(N-mk)!}{N!}\left(1-\sum_{i=1}^{m-1}P_i(N-mk)\right)$$
$$=\sum_{k=1}^{\lfloor N/m\rfloor}\frac{(-1)^{k+1}}{m^k k!}\left(1-\sum_{i=1}^{m-1}P_i(N-mk)\right).$$

This completes the proof. ■

The limiting value of $P_m(N)$ Next we find a general formula for the limit of $P_m(N)$ as N tends to infinity. Recall that for $m=1$, this reduces to Montmort's problem, so that $\lim_{N\to\infty}P_1(N)=1-1/e$. Theorem 2, which gives a formula for evaluating $P_m\equiv\lim_{N\to\infty}P_m(N)$, uses the following standard lemma of real analysis (see, e.g., [**9**, pp. 73–74]):

LEMMA. *Let $\{a_k\}$ and $\{b_k\}$ be sequences such that a_k converges to a and $\sum b_k$ converges absolutely and the sum is b. Let $c_l=\sum_{k=0}^{l}b_k a_{l-k}$. Then $\lim_{l\to\infty}c_l=ab$.*

THEOREM 2. *Let $P_m = \lim_{N\to\infty} P_m(N)$, $m = 1, 2, 3, \ldots, N$. Then P_m exists, and*

$$P_m = e^{-}\sum_{k=1}^{m-1}\frac{1}{k} - e^{-}\sum_{k=1}^{m}\frac{1}{k}. \tag{1}$$

Proof. The proof is by complete induction on m. We know (1) is true for $m = 1$. Now assume $m > 1$ and whenever $i = 1, \ldots, m-1$ then P_i exists and

$$P_i = e^{-}\sum_{k=1}^{i-1}\frac{1}{k} - e^{-}\sum_{k=1}^{i}\frac{1}{k}.$$

To show (1), we will first show that

$$P_m = (1 - e^{-1/m})\left(1 - \sum_{i=1}^{m-1} P_i\right). \tag{2}$$

It is enough to show that for each value of $r = 0, 1, 2, \ldots, m-1$, $P_m(mq + r)$ approaches $(1 - e^{-1/m})(1 - \sum_{i=1}^{m-1} P_i)$ as $q \to \infty$. To this end, fix r and apply the lemma with $a_k = 1 - \sum_{i=1}^{m-1} P_i(mk + r)$ and $b_k = \frac{(-1)^k}{m^k k!}$. Then $a = 1 - \sum_{i=1}^{m-1} P_i$ and $b = e^{-1/m}$. Therefore as $q \to \infty$,

$$\sum_{k=0}^{q} \frac{(-1)^k}{m^k k!}\left(1 - \sum_{i=1}^{m-1} P_i(mq + r - mk)\right) \to e^{-1/m}\left(1 - \sum_{i=1}^{m-1} P_i\right).$$

Now we have, as $q \to \infty$,

$$\begin{aligned}
P_m(mq + r) &= \sum_{k=1}^{q} \frac{(-1)^{k+1}}{m^k k!}\left(1 - \sum_{i=1}^{m-1} P_i(mq + r - mk)\right)\\
&= 1 - \sum_{i=1}^{m-1} P_i(mq + r) - \sum_{k=0}^{q} \frac{(-1)^k}{m^k k!}\left(1 - \sum_{i=1}^{m-1} P_i(mq + r - mk)\right),\\
&\to 1 - \sum_{i=1}^{m-1} P_i - e^{-1/m}\left(1 - \sum_{i=1}^{m-1} P_i\right)\\
&= (1 - e^{-1/m})\left(1 - \sum_{i=1}^{m-1} P_i\right),
\end{aligned}$$

which proves (2).

By the induction hypothesis, we know that

$$\sum_{i=1}^{m-1} P_i = \sum_{i=1}^{m-1}\left(e^{-}\sum_{k=1}^{i-1}\frac{1}{k} - e^{-}\sum_{k=1}^{i}\frac{1}{k}\right) = 1 - e^{-}\sum_{k=1}^{m-1}\frac{1}{k},$$

from which we obtain

$$P_m = (1 - e^{-1/m})\left(1 - \sum_{i=1}^{m-1} P_i\right) = (1 - e^{-1/m})\left(e^{-}\sum_{k=1}^{m-1}\frac{1}{k}\right) = e^{-}\sum_{k=1}^{m-1}\frac{1}{k} - e^{-}\sum_{k=1}^{m}\frac{1}{k}.$$

This completes the proof. ■

We can now state our main result:

THEOREM 3. *The probability that the size of the smallest subset of N men that exchange hats among themselves exceeds m approaches* $e^{-\sum_{k=1}^{m}\frac{1}{k}}$ *as* $N \to \infty$.

Proof. Theorem 3 follows immediately from Theorem 2, since the probability in Theorem 3 is simply $1-\sum_{i=1}^{m} P_i$. ∎

This result is not new. For instance, Kolchin [3] shows that if α_r is the number of cycles of length r in a random permutation of n elements, then

$$P(\alpha_{r_1}=k_1, \alpha_{r_2}=k_2, \dots, \alpha_{r_s}=k_s) = \frac{1}{k_1!k_2!\dots k_s!}e^{-\frac{1}{r_1}-\cdots-\frac{1}{r_s}}+o(1)$$

as $n \to \infty$. In particular,

$$P(\alpha_1=0, \alpha_2=0, \dots, \alpha_m=0) = e^{-1-\frac{1}{2}-\frac{1}{3}-\cdots-\frac{1}{m}}+o(1),$$

from which Theorem 3 follows. The proof given by Kolchin uses sophisticated tools, including local limit theorems in probability and integrals of complex-valued functions. The proof given above is relatively elementary and illustrates the method of inclusion-exclusion.

REFERENCES

1. L. Euler, Calcul de la probabilité dans le jeu de rencontre, *Mém. Acad. Sci. Berlin* 7 (1753), 255–270.
2. A. Hald, *A History of Probability and Statistics and their Applications Before* 1750, Wiley, New York, NY, 1990.
3. V. F. Kolchin, *Random Mappings*, Optimization Software Inc., New York, NY, 1986.
4. J. H. Lambert, Examen d'une espèce de superstition ramenée au calcul des probabilités, *Nouveau Mém. Acad. Roy. Sci. et Belle-Lettres de Berlin* (1773), 411–420.
5. P. S. de Laplace, *Théorie Analytique des Probabilités*, Paris, 1812.
6. A. de Moivre, *The Doctrine of Chances*, Third Edition, Millar, London, 1756. Reprinted by Chelsea, New York, NY, 1967.
7. P. R. de Montmort, *Essay d'Analyse sur les Jeux de Hazard*, Quillau, Paris, published anonymously, 1708.
8. P. R. de Montmort, *Essay d'Analyse sur les Jeux de Hazard*, Seconde Edition, Revûe et augmentée de plusieurs Lettre, Quillau, Paris, 1713. Reprinted 1714. Published anonymously. Reprinted by Chelsea, New York, NY, 1980.
9. K. R. Stromberg, *An Introduction to Classical Real Analysis*, Wadsworth, Inc., Belmont, CA, 1981.
10. L. Takács, The problem of coincidences, *Arch. Hist. Ex. Sci.* 21 (1980), 229–244.
11. E. Waring, *An Essay on the Principles of Human Knowledge*, Cambridge Univ. Press, Cambridge, UK, 1794.
12. W. Weaver, *Lady Luck: The Theory of Probability*, Dover, New York, NY, 1982.

What Can Be Learned From $n < n!$?

ANDREW LENARD
Indiana University
Bloomington, IN 47405

In his famous introduction to analysis, *A Course of Pure Mathematics*, the great English mathematician G. H. Hardy at one point writes "This is almost obvious" in the text, and then appends the following footnote:

> *There is a certain ambiguity in this phrase which the reader will do well to notice. When one says 'such and such a theorem is almost obvious' one may mean one or other of two things. One may mean 'it is difficult to doubt the truth of the theorem,' 'the theorem is such as common sense intuitively accepts,' as it accepts, for example, the truth of the propositions '$2+2=4$' or 'the base angles of isosceles triangles are equal'. That a theorem is 'obvious' in this sense does not prove that it is true, since the most confident of the intuitive judgments of common sense are often found to be mistaken; and even if the theorem is true, the fact that it is also 'obvious' is no reason for not proving it, if a proof can be found. The object of mathematics is to prove that certain premises imply certain conclusions; and the fact that the conclusions may be as 'obvious' as the premises never detracts from the necessity, and often not even from the interest of the proof.*
>
> *But sometimes (as for the example here) we mean by 'this is almost obvious' something quite different from this. We mean 'a moment's reflection should not only convince the reader of the truth of what is stated, but should also suggest to him the general lines of a rigorous proof'. And often, when a statement is 'obvious' in this sense, one may well omit the proof, not because the proof is unnecessary, but because it is a waste of time to state in detail what the reader can easily supply for himself.*[1]

A good example to illustrate these remarks is the inequality $n < 2^n$ $(n = 0, 1, 2, \ldots)$. It is "obvious," indeed so much so that children often become aware of it at an early age. And the common sense proof

$$n + 1 < 2^n + 1 \leq 2^n + 2^n = 2^{n+1}$$

is an excellent textbook case for introducing the student to proof by induction. But there is more to it than that. What is the *meaning* of the inequality? It is this: A finite set has more subsets than elements. Why? Because if we assign to every element x of the set S a subset of S, say A_x, in any manner whatsoever, there is still at least one more subset of S that has not been so assigned. Namely, the subset

$$R = \{x \in S \colon x \notin A_x\}$$

has the property that $R = A_y$ is impossible for every element y in S. (Assuming it is possible, just ask whether y is an element of R or not!)

[1] Hardy credits his colleague and collaborator J. E. Littlewood for the substance of these observations.

This beautiful proof of the "obvious" inequality $n < 2^n$ has the virtue that it works for every set, not just for finite sets: The cardinal number of any set, finite or infinite, is less than the cardinal number of the family of its subsets. Thus we see that—as Hardy tells us—there is indeed interest in proving an "obvious" theorem. The examination of such a proof may for instance, as in the present case, lead to a significant generalization.

The purpose of this note is to examine from this point of view another obvious inequality, namely $n < n!$ $(n = 3, 4, 5, \dots)$. Yes, it too is obvious; and in both senses of the Hardy–Littlewood remark. It is obvious in the second sense, and so we can trust any reader who is so inclined to construct the easy induction proof. But again, there is more depth here than meets the eye. The *meaning* of the inequality is this: A finite set has more permutations than the number of its elements, provided only that it has at least three elements. And in this formulation the immediate question arises whether the theorem is also true for infinite sets. It is; and the proof follows. It is patterned closely on the proof just given for $n < 2^n$. One cannot expect the present proof to be quite as simple though, for somewhere the hypothesis that the set has at least three elements must be used.

Let us assume then that S is a set (finite or infinite) with at least three elements. Let π be any mapping of S into the set of permutations of S. The permutation of S assigned by π to the element x of S shall be denoted π_x, and $\pi_x(y)$ shall denote the element of S into which π_x sends the element y.

Our aim is to exhibit a permutation σ of S that is not in the range of π. Once it is shown that this can be done for a truly arbitrary π, it becomes clear that it is impossible to have a one-to-one correspondence between S, or any subset of S, and the set of *all* permutations of S. Therefore the cardinal number of the set of permutations of S is revealed as strictly larger than the cardinal number of the set S itself.

An element x of S and the corresponding permutation π_x shall be called *self-fixing* if $\pi_x(x) = x$. We distinguish four mutually exclusive and exhaustive cases.

(1) There are no self-fixing elements in S.

(2) There is exactly one self-fixing element in S, and there is also a transposition[2] not in the range of π that interchanges the unique self-fixing element with some other element.

(3) There is exactly one self-fixing element in S, but every transposition that interchanges this unique self-fixing element with some other element is in the range of π.

(4) The number of self-fixing elements in S is at least two.

In case (1), σ may be chosen as the identity permutation. For then $\sigma(x) = x \neq \pi_x(x)$ for all x in S. Thus σ is not in the range of π, as required.

In case (2), σ may be chosen to be the transposition whose existence is assumed.

In case (3), let w be the unique self-fixing element of S. For any $y \neq w$ in S, let τ_y denote the transposition interchanging w and y. By hypothesis, it is of the form π_z for some z in S (depending on y). This z is certainly not w, for π_w fixes w but τ_y does not. Therefore, since w is the only self-fixing element, $z \neq \pi_z(z) = \tau_y(z)$. But τ_y being the transposition specified, this shows that $z = y$. We conclude then that in case (3) the range of π consists of one self-fixing permutation π_w, and otherwise only of transpositions that interchange w with another element. Let now y and z be two

[2] Recall that a transposition is a permutation that interchanges two elements of S but fixes all the rest.

elements of S, distinct and also distinct from w (remember that, by hypothesis, there are such!). Then the cyclic permutation σ defined by

$$\sigma(w)=z, \qquad \sigma(z)=y, \qquad \sigma(y)=w,$$

and fixing all other elements of S, neither has w as a fixed point nor is a transposition, and so it is not of the form π_x for any x in S, as required.

Finally, in case (4) we may choose σ to be any permutation whose set of fixed elements is precisely the complement of the set of self-fixing elements of S. If x is a self-fixing element of S then $\pi_x(x)=x\neq\sigma(x)$, and if x is not a self-fixing element of S then $\pi_x(x)\neq x=\sigma(x)$. Thus $\sigma\neq\pi_x$ for all x in S, as required.

Acknowledgment. The author wishes to acknowledge that the ideas expressed in this note originated in conversations with his friend, the late George J. Minty, at one time Professor of Mathematics at Indiana University. The writer is also indebted to his colleague W.H. Wheeler for pointing out that, intuitive as it may be, the statement that there exists a permutation that moves *every* point of a set, requires for its proof the Axiom of Choice in case the set is infinite. What one must do in that case is to partition the set into two subsets of equal cardinality with a one-to-one correspondence between them, and define the permutation as the one that interchanges corresponding points. It is the existence of such a partitioning that depends on set theory *with the Axiom of Choice.*

Math Bite: Why $2+2$ Equals 2×2

When I was in grade school, I wondered why $2+2=2\times 2$. Later, I discovered that $2+2=2\times 2=2^2$. Why?

Addition is a repeated application of the successor function, multiplication is repeated addition, exponentiation is repeated multiplication. It is natural to define recursively an operation $\langle n\rangle$, where $a\langle 1\rangle b$ is the successor of a taken b times, and where $a\langle n\rangle b$ is defined as repeated application of the operation $\langle n-1\rangle$. Thus $\langle 1\rangle$ is addition, $\langle 2\rangle$ is multiplication, and $\langle 3\rangle$ is exponentiation. Because exponentiation is neither commutative nor associative, we need the usual convention when $n>2$: group right. For example, $a\langle n\rangle 3=a\langle n-1\rangle(a\langle n-1\rangle a)$.

Now I know why $2+2=2\times 2$. It is a special case of $2\langle n\rangle 2=2\langle m\rangle 2$, for all natural numbers m and n.

I wonder how we would want to define $e\langle 4\rangle\pi\langle 4\rangle i$?

—Rick Norwood
East Tennessee State University
Johnson City, TN 37614

elements of S, distinct and also distinct from w (remember that, by hypothesis, there are such!). Then the cyclic permutation σ defined by

$$\sigma(w)=z, \qquad \sigma(z)=y, \qquad \sigma(y)=w,$$

and fixing all other elements of S, neither has w as a fixed point nor is a transposition, and so it is not of the form π_x for any x in S, as required.

Finally, in case (4) we may choose σ to be any permutation whose set of fixed elements is precisely the complement of the set of self-fixing elements of S. If x is a self-fixing element of S then $\pi_x(x)=x\neq\sigma(x)$, and if x is not a self-fixing element of S then $\pi_x(x)\neq x=\sigma(x)$. Thus $\sigma\neq\pi_x$ for all x in S, as required.

Acknowledgment. The author wishes to acknowledge that the ideas expressed in this note originated in conversations with his friend, the late George J. Minty, at one time Professor of Mathematics at Indiana University. The writer is also indebted to his colleague W.H. Wheeler for pointing out that, intuitive as it may be, the statement that there exists a permutation that moves *every* point of a set, requires for its proof the Axiom of Choice in case the set is infinite. What one must do in that case is to partition the set into two subsets of equal cardinality with a one-to-one correspondence between them, and define the permutation as the one that interchanges corresponding points. It is the existence of such a partitioning that depends on set theory *with the Axiom of Choice.*

Math Bite: Why 2 + 2 Equals 2 × 2

When I was in grade school, I wondered why $2+2=2\times 2$. Later, I discovered that $2+2=2\times 2=2^2$. Why?

Addition is a repeated application of the successor function, multiplication is repeated addition, exponentiation is repeated multiplication. It is natural to define recursively an operation $\langle n\rangle$, where $a\langle 1\rangle b$ is the successor of a taken b times, and where $a\langle n\rangle b$ is defined as repeated application of the operation $\langle n-1\rangle$. Thus $\langle 1\rangle$ is addition, $\langle 2\rangle$ is multiplication, and $\langle 3\rangle$ is exponentiation. Because exponentiation is neither commutative nor associative, we need the usual convention when $n>2$: group right. For example, $a\langle n\rangle 3=a\langle n-1\rangle(a\langle n-1\rangle a)$.

Now I know why $2+2=2\times 2$. It is a special case of $2\langle n\rangle 2=2\langle m\rangle 2$, for all natural numbers m and n.

I wonder how we would want to define $e\langle 4\rangle\pi\langle 4\rangle i$?

—Rick Norwood
East Tennessee State University
Johnson City, TN 37614

When is $(xy+1)(yz+1)(zx+1)$ a Square?

KIRAN S. KEDLAYA
Princeton University
Princeton, NJ 08544

To cut the suspense, let's start with the surprising answer to the title question.

Theorem. *If x, y, z are positive integers, then $(xy+1)(yz+1)(zx+1)$ is a perfect square if and only if $xy+1$, $yz+1$, and $zx+1$ are all perfect squares.*

The purpose of this note is to prove this result using Fermat's method of infinite descent, to provide historical context, and to investigate (and eventually refute) a possible generalization.

For t a positive integer, a *P_t-set* is a set of positive integers, the product of any two distinct elements of which is t less than a perfect square. (The positivity restriction is sometimes relaxed, but we will impose it throughout.) Classical examples of P_t-sets include the P_{256}-set $\{1, 33, 68, 105\}$ found by Diophantos and the P_1-set $\{1, 3, 8, 120\}$ found by Fermat.

A sizable literature exists addressing the existence or nonexistence of P_t-sets of certain forms; some early examples are chronicled in [**3**, Chap. XIX, pp. 513–520]. A little experimentation shows that P_t-sets become nontrivial to construct when they must have four or more elements; Euler found a general construction of four-element P_1-sets which includes Fermat's example. Since this construction is essential for the proof of the theorem, we state it as a lemma (following [**5**]).

Lemma. *If $\{p, q, r\}$ is a P_1-set, then so is $\{p, q, r, s\}$ for*

$$s = p + q + r + 2pqr \pm 2\sqrt{(pq+1)(qr+1)(rp+1)}\,, \tag{1}$$

as long as $s > 0$. (Note that s is necessarily an integer.)

Proof. The values of s defined in (1) are the roots of the quadratic equation

$$p^2 + q^2 + r^2 + s^2 - 2(pq + pr + qr + ps + qs + rs) - 4pqrs - 4 = 0, \tag{2}$$

which can be rewritten in the following ways:

$$(p+q-r-s)^2 = 4(pq+1)(rs+1)$$
$$(p+r-q-s)^2 = 4(pr+1)(qs+1)$$
$$(p+s-q-r)^2 = 4(qr+1)(ps+1).$$

Since $rs+1$ is an integer which is the quotient of two perfect squares, it is also a square, as are $ps+1$ and $qs+1$ by the same argument. Thus $\{p, q, r, s\}$ is a P_1-set.

Not surprisingly, constructing five-element P_t-sets is substantially harder. Euler gave a general construction, and a number of additional examples are also known; however, it is not known whether there exist infinitely many five-element P_t-sets for any particular values of t, or whether there exist any at all for $t = 1$.

The first significant nonexistence result for P_t-sets is due to Baker and Davenport [**1**]; using Baker's theory of linear forms in logarithms of algebraic numbers, they showed that Fermat's P_1-set $\{1, 3, 8\}$ can only be extended by adding 120. Their method was later refined by Grinstead [**4**] and Brown [**2**] and applied to other P_t-sets. An elementary approach to such questions is given by Kangasabapathy and Ponnudurai [**6**] and by Mohanty and Ramasamy [**8**]; a systematic presentation and a more complete bibliography appear in [**7**].

The above theorem does not directly apply to studying the existence or nonexistence of P_t-sets, but it does give an interesting characterization of three-element P_1-sets; after the proof, we will see that this phenomenon is (almost) unique to the case $t = 1$.

Proof of the Theorem. Suppose there exist triples p, q, r of positive integers (where we might as well assume $p \leq q \leq r$) such that $(pq + 1)(qr + 1)(rp + 1)$ is a perfect square, but not all of $pq + 1$, $qr + 1$, $rp + 1$ are squares. Choose a triple that minimizes $p + q + r$, and define s as in (1) using the negative square root. We will show that $0 < s < r$ and that $(pq + 1)(qs + 1)(sp + 1)$ is a square, but that not all of $pq + 1$, $qs + 1$, $sp + 1$ are squares, contradicting the minimality of $p + q + r$.

By the equivalent forms of (2), we know that

$$\begin{aligned} &16(pq+1)^2(pr+1)(qs+1)(qr+1)(ps+1) \\ &\quad = (pq+1)^2(p+r-q-s)^2(p+s-q-r)^2 \end{aligned}$$

is a perfect square; since $(pq + 1)(qr + 1)(rp + 1)$ is a square, so then is $(pq + 1)(qs + 1)(sp + 1)$. Moreover, $ps + 1$ is a square if and only if $qr + 1$ is a square, and $pr + 1$ is a square if and only if $qs + 1$ is a square, so not all of $pq + 1$, $qs + 1$, $sp + 1$ are squares.

We also have

$$rs + 1 \frac{(p+q-r-s)^2}{4(pq+1)} \geq 0$$

and so $s \geq -1/r$. Note that $r = 1$ implies (by our assumption that $p \leq q \leq r$) that $p = q = r = 1$, in which case $(pq + 1)(qr + 1)(rp + 1)$ is not a square, a contradiction. Hence $r > 1$ and so $s \geq 0$. Moreover, if $s = 0$, then we have

$$4(pq+1) = (p+q-r)^2, \quad 4(qr+1) = (q+r-p)^2, \quad 4(rp+1) = (r+p-q)^2,$$

contradicting the assumption that not all of $pq + 1$, $qr + 1$, and $rp + 1$ are squares. Therefore s is a positive integer.

If s' is the other root of (2) (which is to say, s' satisfies (1) using the positive square root), then we have

$$\begin{aligned} ss' &= p^2 + q^2 + r^2 - 2pq - 2pr - 2qr - 4 \\ &< r^2 - p(2r - p) - q(2r - q) \\ &< r^2. \end{aligned}$$

Since s is the smaller of the two roots, $s^2 \leq ss'$ and so we conclude $s < r$, yielding the desired contradiction. ■

Does an analogous characterization of P_t-sets exist for $t > 1$? In other words, is $(pq + t)(qr + t)(rp + t)$ a square if and only if $pq + t$, $qr + t$, $rp + t$ are all squares? The proof above does not work in general; the natural analogue of (2) would be

$$t(p^2 + q^2 + r^2 + s^2) - 2t(pq + pr + qr + ps + qs + rs) - 4pqrs - 4t^2 = 0, \quad (3)$$

whose equivalent forms are

$$4(pq+t)(rs+t)=t(p+q-r-s)^2$$

and so on, but two obstructions arise. If t is not a perfect square, then $\{p,q,r\}$ can be a counterexample even if $\{p,q,s\}$ is not. Even if t is a perfect square, though, if $t>4$, we cannot ensure that s is an integer.

Neither obstruction arises for $t=4$, and indeed the reader may check that the natural analogue of the theorem holds in this case with essentially the same proof. However, we will now show that this analogue does not hold for $t\neq 1,4$.

We first construct a counterexample $\{p,q,r\}$ where t is not a perfect square. Put $p=1$, $q=a^2-t$, where q is not a perfect square (which certainly holds if $t<2a+1$); we shall find r such that $r+t=tb^2$, $qr+t=tc^2$, which is equivalent to solving

$$c^2-qb^2=1-q.$$

Indeed, $b=c=1$ is a solution, but it yields $r=0$, which is not a positive integer. Nonetheless it is useful! To produce a nontrivial solution, let (u,v) be a solution in positive integers of the Pell equation

$$u^2-qv^2=1,$$

and put

$$\left(c+b\sqrt{q}\right)=\left(1+\sqrt{q}\right)\left(u+v\sqrt{q}\right).$$

Now $r=t[(u+v)^2-1]$ yields a counterexample. For example, if $t=a=q=2$, the solution $(3,2)$ of the Pell equation gives the set $\{p,q,r\}=\{1,2,48\}$.

On the other hand, if $t=d^2$ for $d>2$, we can write $t=a^2-p^2$ for some positive a,p, and a similar argument starting from the bogus counterexample p,p,r (r arbitrary) yields an actual counterexample.

Acknowledgment. Thanks to George Berzsenyi for providing ideas for my entry in the 1992 Westinghouse Science Talent Search, where the above result first appeared.

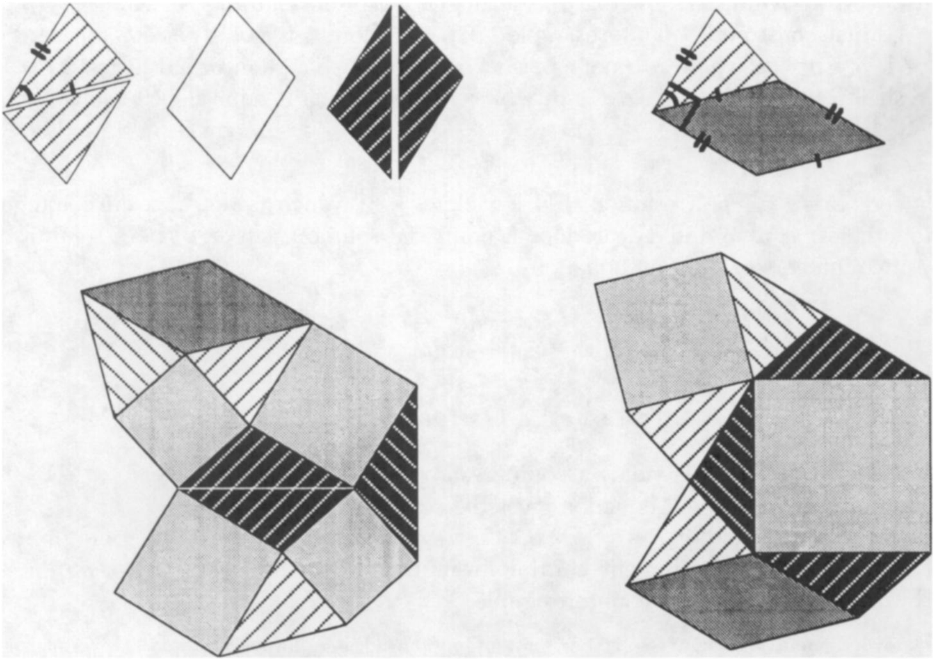
REFERENCES

1. A. Baker and H. Davenport, The equations $3x^2-2=y^2$ and $8x^2-7=z^2$, *Quart. J. Math. Oxford* (2) **20** (1969), 129–137.
2. E. Brown, Sets in which $xy+k$ is always a square, *Math. Comp.* **45** (1985), 613–620.
3. L. E. Dickson, *History of the Theory of Numbers, Volume* 2, Carnegie Institute, Washington, DC, 1920.
4. C. M. Grinstead, On a method of solving a class of Diophantine equations, *Math. Comp.* **32** (1978), 936–940.
5. P. Heichelheim, The study of positive integers (a,b) such that $ab+1$ is a square, *Fibon. Quart.* **17** (1979), 269–274.
6. P. Kangasabapathy and T. Ponnudurai, The simultaneous Diophantine equations $y^2-3x^2=-2$ and $z^2-8x^2=-7$, *Quart. J. Math. Oxford* (3) **26** (1975), 275–278.
7. K. S. Kedlaya, Solving constrained Pell equations, *Math. Comp.*, to appear.
8. S. P. Mohanty and A. M. S. Ramasamy, The simultaneous diophantine equations $5y^2-20=x^2$ and $2y^2+1=z^2$, *J. Number Theory* **18** (1984), 356–359.

Proof Without Words: A Generalization from Pythagoras

THEOREM. *The sum of the areas of two squares, whose sides are the lengths of the two diagonals of a parallelogram, is equal to the sum of the areas of four squares, whose sides are its four sides.*

Proof.



COROLLARY. *Pythagoras's theorem* (when the parallelogram is a rectangle).

Nelsen [1] reproduces a famous proof that uses tessellation similarly.

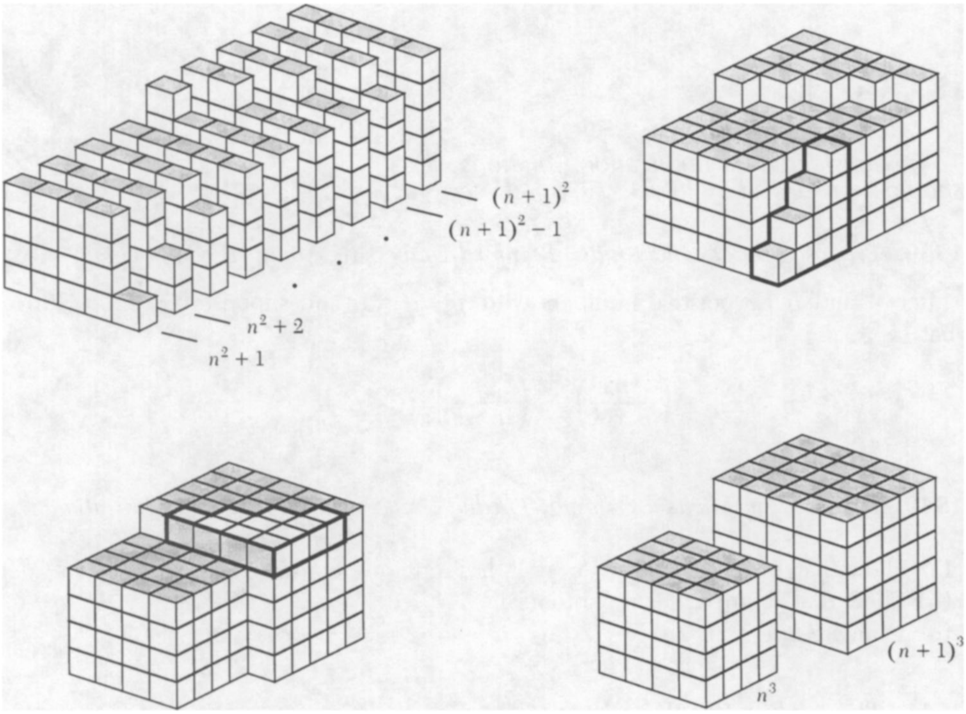
REFERENCE

1. R. B. Nelsen, *Proofs Without Words*, Math. Assoc. of America, Washington, DC, 1993, p. 3.

—DAVID S. WISE
INDIANA UNIVERSITY
BLOOMINGTON, IN 47405-4101

Proof Without Words: Sums of Integers as Sums of Cubes

$$\begin{aligned} 2+3+4 &= 1+8 \\ 5+6+7+8+9 &= 8+27 \\ 10+11+12+13+14+15+16 &= 27+64 \\ &\vdots \\ (n^2+1)+(n^2+2)+\cdots+(n+1)^2 &= n^3+(n+1)^3 \end{aligned}$$



—Roger B. Nelsen
Lewis and Clark College
Portland, OR 97219

PROBLEMS

GEORGE T. GILBERT, *Editor*
Texas Christian University

ZE-LI DOU, KEN RICHARDSON, and SUSAN G. STAPLES, *Assistant Editors*
Texas Christian University

Proposals

To be considered for publication, solutions should be received by July 1, 1998.

1539. *Proposed by Donald Knuth, Stanford University, Stanford, California.*

Let p and q be positive numbers with $p+q=1$, and suppose $0 \leq \epsilon < q$. Prove that

$$\left(\frac{p}{p+\epsilon}\right)^{p+\epsilon}\left(\frac{q}{q-\epsilon}\right)^{q-\epsilon} < e^{-2\epsilon^2}.$$

1540. *Proposed by Michael Golomb, Purdue University, West Lafayette, Indiana.*

(a) Show that $x^n + (x-1)^n - (x+1)^n$ has a unique non-zero real root r_n.
(b) Show that r_n increases monotonically.
(c) Evaluate $\lim_{n\to\infty} r_n/n$.

1541. *Proposed by Wu Wei Chao, He Nan Normal University, Xin Xiang City, He Nan Province, China.*

Assume $a_1 > 1$ and define $a_{n+1} = 1/a_n + a_1 - 1$ for $n = 1,2,3,\ldots$. Evaluate

$$\lim_{n\to\infty} |a_{n+1} - a_n|^{1/n}.$$

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE. *Each solution should begin on a separate sheet containing the solver's name and full address.*

Solutions and new proposals should be mailed to George T. Gilbert, Problems Editor, Department of Mathematics, Box 298900, Texas Christian University, Fort Worth, TX 76129, or mailed electronically (ideally as a LATEX file) to g.gilbert@tcu.edu. *Readers who use e-mail should also provide an e-mail address.*

1542. *Proposed by Jerrold W. Grossman and Barry Turett, Oakland University, Rochester, Michigan.*

Sam and Joe (names favored by the late Paul Erdös) play an infinite game on the real number line. They start at distinct initial positions and alternate turns. At each turn a player must move to some point strictly between the players' current positions. Being monotonic and bounded, the sequence of positions for each player converges. A player wins the game if his limit is rational and loses if his limit is irrational.

(a) Show that Joe can force Sam to lose.
(b) Find a strategy by which Joe will win with probability 1 if Sam plays randomly (i.e., at each turn, Sam chooses a point in the gap between the players, independent of previous choices, based on the uniform distribution).
(c) Does the result in (a) hold if the winning set is an arbitrary set of measure zero?

(Obviously they can play cooperatively and end up with a win/win situation. Furthermore, either player can unilaterally guarantee that the results for both players are identical by forcing the gap between them to vanish.)

1543. *Proposed by Michael Golomb, Purdue University, West Lafayette, Indiana.*

Let S be a given n-dimensional simplex with centroid C. A hyperplane through C divides the simplex into two regions, one or both of which are simplexes. Find the extrema of the volumes of those regions which are simplexes.

Quickies

Answers to the Quickies are on page 73.

Q874. *Proposed by Matt Baker, graduate student, University of California at Berkeley, Berkeley, California.*

Find all integer solutions to $x^2 + 6y^2 = 2z^2 + 3w^2$.

Q875. *Proposed by Hoe Teck Wee, Lengkok Bahru, Singapore.*

Given a list of $3n$ not necessarily distinct elements of a set S, determine necessary and sufficient conditions under which these $3n$ elements can be divided into n triples, none of which consist of three distinct elements.

Q876. *Proposed by Mihàly Bencze, Braşov, Romania.*

Let A and B be $n \times n$ matrices with integer entries such that $A + kB$ has an inverse with integer entries for $k = 0, 1, \dots, 2n$. What is the determinant of B?

Solutions

Subsets Whose Elements Sum to a Multiple of a Prime — February 1997

1514. *Proposed by Hoe Teck Wee, Lengkok Bahru, Singapore.*

Let p be an odd prime and k be a natural number. Find the sum of the elements of the subsets of $\{1, 2, \dots, kp\}$, the sum of whose elements is divisible by p. (For instance, when $p = 3$ and $k = 1$, the relevant subsets are $\{1, 2\}$, $\{3\}$, and $\{1, 2, 3\}$, and the required sum is 12.)

(This generalizes problem 6 of the 36th International Mathematical Olympiad, held in July 1995.)

Solution by Lior Pachter, Massachusetts Institute of Technology, Cambridge, Massachusetts.

The sum of such elements is $k(kp+1)(2^{kp}+2^k(p-1))/4$.

More generally, let $N_{n,k}$ denote the number of subsets of $\{1,2,\dots,kn\}$, the sum of whose elements is divisible by n, and let $S_{n,k}$ denote the sum of the elements of all such subsets. We first show that

$$N_{n,k}=\frac{1}{n}\sum_{\substack{d\mid n\\ d \text{ odd}}} 2^{kn/d}\phi(d),$$

where $\phi(d)$ is the number of positive integers that are at most d and relatively prime to d. Our argument follows the derivation for $k=1$ found in R. P. Stanley, *Enumerative Combinatorics*, vol. 1, p. 59. Consider the polynomial

$$P(x)=(1+x)(1+x^2)\cdots(1+x^{kn})=\sum_{j\geq 0} a_j x^j.$$

Then a_j counts the number of ways to express j as the sum of the elements of subsets of $\{1,2,\dots,kn\}$. Let $\zeta := e^{2\pi i/n}$. For any integer j, $\Sigma_{m=1}^{n}\zeta^{mj}$ equals n if n divides j and 0 otherwise. Therefore, we have

$$\frac{1}{n}\sum_{m=1}^{n} P(\zeta^m)=\sum_{j\geq 0} a_{jn}=N_{n,k}.$$

Setting $d=n/(n,m)$, we have that ζ^m is a primitive dth root of unity. Setting $x=-1$ in the identity

$$x^d-1=(x-\zeta^m)(x-\zeta^{2m})\cdots(x-\zeta^{dm})$$

yields

$$(1+\zeta^m)(1+\zeta^{2m})\cdots(1+\zeta^{dm})=\begin{cases}2 & \text{if } d \text{ is odd},\\ 0 & \text{if } d \text{ is even}.\end{cases}$$

Since there are $\phi(d)$ values of m for which ζ^m is a primitive dth root of unity, we obtain

$$N_{n,k}=\frac{1}{n}\sum_{m=1}^{n} P(\zeta^m)=\frac{1}{n}\sum_{\substack{d\mid n\\ d \text{ odd}}} 2^{kn/d}\phi(d).$$

Note that the sum of the elements of the set $\{1,2,\dots,kn\}$ is $kn(kn+1)/2$. For n odd or k even, if the elements of a subset S of $\{1,2,\dots,kn\}$ sum to a multiple of n, so do the elements of $\{1,2,\dots,kn\}-S$. This pairing yields that the mean of the sum of the elements of subsets summing to a multiple of n is $kn(kn+1)/4$. This implies that

$$S_{n,k}=\frac{kn(kn+1)}{4}N_{n,k}=\frac{k(kn+1)}{4}\sum_{\substack{d\mid n\\ d \text{ odd}}} 2^{kn/d}\phi(d).$$

The formula for n an odd prime p follows.

Also solved by Thomas Jager, Kee-Wai Lau (Hong Kong), Peter W. Lindstrom, and the proposer. There were one incorrect solution and one incomplete solution.

A 3-Dimensional Heron-Type Formula **February 1997**

1515. *Proposed by Isaac Sofair, Fredericksburg, Virginia.*

The edges of a parallelepiped emanating from one vertex are given by the vectors **a**, **b**, and **c**, of lengths a, b, and c, respectively. If α, β, and γ are the angles between **b** and **c**, **c** and **a**, and **a** and **b**, respectively, and $\sigma = (\alpha+\beta+\gamma)/2$, show that the volume of the parallelepiped is

$$2abc\sqrt{\sin\sigma\sin(\sigma-\alpha)\sin(\sigma-\beta)\sin(\sigma-\gamma)}\ .$$

Solution by Reza Akhlaghi, Prestonsburg Community College, Prestonsburg, Kentucky, and Fary Sami, Harford Community College, Bel Air, Maryland.

Without loss of generality, we may assume that $\mathbf{a} = a\mathbf{i}$, $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j}$, and $\mathbf{c} = c_1\mathbf{i} + c_2\mathbf{j} + c_3\mathbf{k}$. Furthermore, reflecting if necessary, we may assume that b_2 and c_3 are positive. Since the angle between **a** and **b** is γ, we get $b_1 = b\cos\gamma$ and $b_2 = b\sin\gamma$. Similarly, $c_1 = c\cos\beta$. From

$$bc\cos\alpha = \mathbf{b}\cdot\mathbf{c} = bc\cos\gamma\cos\beta + bc_2\sin\gamma,$$

we obtain $c_2 = c(\cos\alpha - \cos\beta\cos\gamma)/\sin\gamma$ and finally

$$c_3 = \sqrt{c^2 - c_1^2 - c_2^2} = c\frac{\sqrt{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma}}{\sin\gamma}.$$

The volume of the tetrahedron is given by

$$\begin{aligned}|\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})| &= \left|\det\begin{pmatrix} a & 0 & 0\\ b_1 & b_2 & 0\\ c_1 & c_2 & c_3\end{pmatrix}\right| = ab_2c_3\\ &= abc\sqrt{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma}\,.\end{aligned}$$

We must derive a trigonometric identity to complete the solution. Beginning with the desired result and using well-known identities for products of sines and cosines of sums, we have

$$\begin{aligned}&4\sin\sigma\sin(\sigma-\alpha)\sin(\sigma-\beta)\sin(\sigma-\gamma)\\ &= \left(2\sin\frac{\alpha+\beta+\gamma}{2}\sin\frac{\alpha+\beta-\gamma}{2}\right)\left(2\sin\frac{\gamma-\alpha+\beta}{2}\sin\frac{\gamma+\alpha-\beta}{2}\right)\\ &= (\cos\gamma-\cos(\alpha+\beta))(\cos(\alpha-\beta)-\cos\gamma)\\ &= \cos\gamma(\cos(\alpha-\beta)+\cos(\alpha+\beta)) - \cos(\alpha+\beta)\cos(\alpha-\beta) - \cos^2\gamma\\ &= \cos\gamma(2\cos\alpha\cos\beta) - (\cos^2\alpha\cos^2\beta - \sin^2\alpha\sin^2\beta) - \cos^2\gamma\\ &= 1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma.\end{aligned}$$

Comment. Can Minh reports that the problem of computing the volume of a parallelepiped in terms of its sides and the angles between its sides appeared as problem 27.11 in *The Mathematical Spectrum* **27:3** (1994/5). The answer published in **28:2** (1995/6) was the above expression involving cosines.

Also solved by Anchorage Math Solutions Group, Rich Bauer, J. C. Binz (Switzerland), Mangalam R. Gopal, S. A. Greenspan, John G. Heuver, Thomas Jager, Hengli Jiao, Murray S. Klamkin (Canada), Victor Y. Kutsenok, Neela Lakshmanan, Ralph Merrill, Can A. Minh (graduate student), William A.

Newcomb, P. E. Nüesch (Switzerland), Karel A. Post (the Netherlands), I. A. Sakmar, Volkhard Schindler (Germany), Michael Vowe (Switzerland), Peter Y. Woo, Robert L. Young, Paul J. Zwier, and the proposer.

A Sum Representing an Integral **February 1997**

1516. *Proposed by David Doster, Choate Rosemary Hall, Wallingford, Connecticut.*

Let $S_n = \Sigma_{k=1}^{n}\sqrt{4n^2 - k^2}$. Find the unique value of c for which $\lim_{n\to\infty}(cn - S_n/n)$ exists, and evaluate the limit for this value of c.

Solution by William A. Newcomb, Walnut Creek, California.

The unique value of c is $\sqrt{3}/2 + \pi/3$. In this case the limit equals $1 - \sqrt{3}/2$.

Let $f(x) = \sqrt{4 - x^2}$. If c exists, then from the definition of the Riemann sum it must satisfy

$$c = \lim_{n\to\infty}\frac{S_n}{n^2} = \lim_{n\to\infty}\sum_{k=1}^{n} f(k/n)\frac{1}{n} = \int_0^1 f(x)\,dx = \frac{\sqrt{3}}{2} + \frac{\pi}{3}.$$

The value of the integral follows from substitution or from interpreting the integral as the area of a triangle and a sector of a circle of radius 2. For this value of c, the mean value theorems for integrals and derivatives imply that there exist ζ_k and ξ_k in the interval $((k-1)/n, k/n)$ such that

$$\begin{aligned} n\int_{(k-1)/n}^{k/n}(f(x) - f(k/n))\,dx &= n\int_{(k-1)/n}^{k/n}\frac{f(x) - f(k/n)}{x - k/n}(x - k/n)\,dx \\ &= n\frac{f(\zeta_k) - f(k/n)}{\zeta_k - k/n}\int_{(k-1)/n}^{k/n}(x - k/n)\,dx \\ &= -\frac{1}{2}f'(\xi_k)\frac{1}{n}. \end{aligned}$$

Thus,

$$\begin{aligned} \lim_{n\to\infty}(cn - S_n/n) &= \lim_{n\to\infty}\sum_{k=1}^{n} n\int_{(k-1)/n}^{k/n}(f(x) - f(k/n))\,dx \\ &= \lim_{n\to\infty} -\frac{1}{2}\sum_{k=1}^{n} f'(\xi_k)\frac{1}{n} = -\frac{1}{2}\int_0^1 f'(x)\,dx \\ &= \frac{1}{2}(f(0) - f(1)) = 1 - \frac{\sqrt{3}}{2}. \end{aligned}$$

Alternatively, both limits are obtainable from the trapezoidal rule. For some ζ in $(0, 1)$, we may write the trapezoidal rule in the form

$$\frac{1}{2}\sum_{k=1}^{n} f(k/n) = \int_0^1 f(x)\,dx + \frac{1}{2n}(f(1) - f(0)) - \frac{f''(\zeta)}{12n^2}.$$

Multiplication by n leads to the desired limit. Additional terms in this expansion are obtainable from the Euler–Maclaurin summation formula.

Comment. Several readers pointed out that a more general derivation appears as Problem 10 of G. Pólya and G. Szegö, *Problems and Theorems in Analysis*, vol. 1, Springer, New York, 1976, p. 49.

Also solved by Robert A. Agnew, Reza Akhlaghi and Fary Sami, Anchorage Math Solutions Group, Michael H. Andreoli, Michel Bataille (France), Rich Bauer, Paul Bracken (Canada), Hongwei Chen, John Christopher, C. Coker, Charles Diminnie and Trey Smith and Roger Zarnowski, Daniele Donini (Italy), Robert L. Doucette, Mordechai Falkowitz (Canada), Russell A. Gordon, Michael J. Hoffman and Richard

Katz, Anne L. Hudson, Paul L. Irwin, Thomas Jager, Parviz Khalili, Peter W. Lindstrom, Kandasamy Muthuvel, Karel A. Post (the Netherlands), Heinz-Jürgen Seiffert (Germany), Nicholas C. Singer, Isaac Sofair, Dave Trautman, West Chester University Calculus II Class (Angeline Cremins, Mark Lamplugh, Kyung Lee, Wes Peoples, Craig Walter), Western Maryland College Problems Group, Yan-Loi Wong (Singapore), Paul J. Zwier, and the proposer. There were five incorrect solutions.

The Determinant of a Spiral February 1997

1517. *Proposed by Charles Vanden Eynden, Illinois State University, Normal, Illinois.*

Let M_n be the $n \times n$ matrix with entries the integers from 1 to n^2 spiraling clockwise inwardly, starting in the first row and column. For example

$$M_4 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 12 & 13 & 14 & 5 \\ 11 & 16 & 15 & 6 \\ 10 & 9 & 8 & 7 \end{pmatrix}.$$

Evaluate the determinant of M_n.

Solution by G. R. Miller, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

First observe that we get the same determinant if the clockwise spiraling of entries starts in the last row and column, since a matrix of the first form is transformed into one of the second by exchanging the pairs of rows and the pairs of columns indexed by 1 and n, 2 and $n-1$, 3 and $n-2$, and so forth.

More generally, write

$$M_n(x) = \begin{pmatrix} x & x+1 & \cdots & x+n-2 & x+n-1 \\ x+4n-5 & x+4n-4 & \cdots & x+5n-7 & x+n \\ & & \ddots & & \\ x+3n-3 & x+3n-4 & \cdots & x+2n-1 & x+2n-2 \end{pmatrix},$$

and let $A_n(x)$ denote the matrix formed by replacing the first row of $M_n(x)$ with a row of ones. We seek $\det M_n(1)$.

By subtracting x times the first row of $A_n(x)$ from each of its other rows, we see that $a_n := \det A_n$ is independent of x. Also, set $m_n(x) := \det M_n(x)$. For $n \geq 2$, adding the last row of $M_n(x)$ to its first row, we see that

$$m_n(x) = (2x + 3n - 3)a_n.$$

For $n \geq 3$, to get a recursion for a_n, add the 2nd and the last rows of $A_n(0)$ to its first row, obtaining

$$a_n = \det\begin{pmatrix} 7n-7 & 7n-7 & \cdots & 7n-7 & 3n-1 \\ 4n-5 & 4n-4 & \cdots & 5n-7 & n \\ & & \ddots & & \\ 3n-3 & 3n-4 & \cdots & 2n-1 & 2n-2 \end{pmatrix}.$$

Using linearity of the determinant in its first row along with our initial observation, we find that

$$\begin{aligned} a_n &= (7n-7)a_n + (-1)^n(4n-6)m_{n-1}(2n-1) \\ &= (7n-7)a_n + (-1)^n(4n-6)(7n-8)a_{n-1}. \end{aligned}$$

Therefore, $a_n = (-1)^{n-1}(4n-6)a_{n-1}$. Starting with $a_2 = -1$, it is easy to verify that

$$a_n = (-1)^{(n-1)n/2}2^{n-2}(2n-3)(2n-5)\cdots 3 = (-1)^{(n-1)n/2}(2n-3)!/(n-2)!.$$

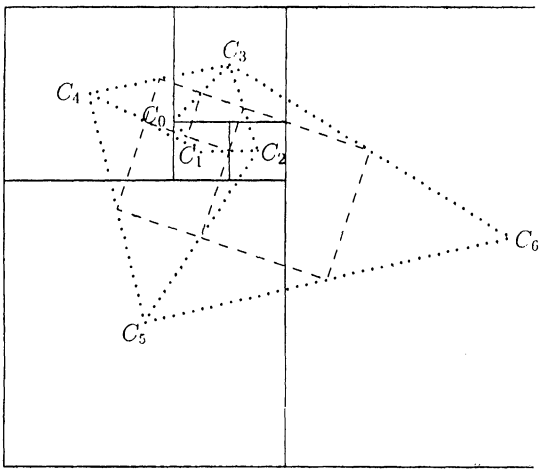
It follows immediately that $m_n(x) = (2x + 3n - 3)(-1)^{(n-1)n/2}(2n-3)!/(n-2)!$ and that $m_n(1) = (-1)^{(n-1)n/2}(3n-1)(2n-3)!/(n-2)!$.

Also solved by Sue Ackermann (graduate student), Anchorage Math Solutions Group, Rich Bauer, J. C. Binz (Switzerland), Darrah Chavey, John Chavez, John Christopher, C. Coker, Thelma W. Hedgepeth, Parviz Khalili, Norman F. Lindquist, Nicholas C. Singer, Irving C. Tang, Western Maryland College Problems Group, Michael Woltermann, Yan-Loi Wong (Singapore), and the proposer. There was one incomplete solution.

Spirals of Squares February 1997

1518. *Proposed by Edward Kitchen, Santa Monica, California.*

Let C_n be the center of a square whose side-length is F_n, $n \geq 0$, where (F_n) is the Fibonacci sequence $0, 1, 1, 2, 3, \ldots$. Place the squares side-by-side in a spiral as in the diagram below. For $n > 0$ join the midpoints of adjacent sides of each quadrangle $C_{n-2}\,C_{n-1}\,C_{n+2}\,C_{n+1}$ (where $C_{-1} = C_1$ by convention). Prove that the resulting pattern is another sequence of squares whose side-lengths are a constant multiple of the Fibonacci sequence.



Composite of solutions due to J. C. Binz, University of Bern, Bern, Switzerland, and the Editors.

Express the centers C_n as complex numbers with $C_0 = 0$. By considering each of the four possible remainders when n is divided by 4 as a separate case, it is easy to verify that

$$C_{n+1} = C_n + \frac{1+i}{2} \cdot i^{n+2}(F_n + F_{n+1}i).$$

The midpoints of the adjacent sides of any quadrangle $C_{n-2}\,C_{n-1}\,C_{n+2}\,C_{n+1}$ form a parallelogram. Thus, to show they form a square, it suffices to show that

$$\frac{C_{n-1} + C_{n+2}}{2} - \frac{C_{n-2} + C_{n-1}}{2} = i\left(\frac{C_{n+1} + C_{n-2}}{2} - \frac{C_{n-2} + C_{n-1}}{2}\right),$$

or that

$$\frac{C_{n+2} - C_{n-2}}{2} = i\left(\frac{C_{n+1} - C_{n-1}}{2}\right).$$

Now

$$\frac{C_{n+1}-C_{n-1}}{2}=\frac{C_{n+1}-C_n}{2}+\frac{C_n-C_{n-1}}{2}$$
$$=\frac{1+i}{4}\left[i^{n+2}(F_n+F_{n+1}i)+i^{n+1}(F_{n-1}+F_ni)\right]$$
$$=\frac{1+i}{4}i^{n+1}(-F_n+2F_ni)=\frac{-3+i}{4}i^{n+1}F_n.$$

Similarly,

$$\frac{C_{n+2}-C_{n-2}}{2}=\frac{C_{n+2}-C_n}{2}+\frac{C_n-C_{n-2}}{2}=\frac{-3+i}{4}i^{n+2}F_{n+1}+\frac{-3+i}{4}i^nF_{n-1}$$
$$=\frac{-3+i}{4}i^{n+2}F_n.$$

The claim follows. In addition, the proof shows that new squares have sides of length $F_n\sqrt{10}/4$ and that the original orientation has been rotated counterclockwise by an angle with tangent $-1/3$, or approximately 161.6 degrees.

Also solved by Neela Lakshmanan, Karel A. Post (the Netherlands), Volkhard Schindler (Germany), Joel Schlosberg (student), Stephen Swiniarski, and the proposer.

Answers

Solutions to the Quickies on page 67.

A874. The only solution is $x=y=z=w=0$. It clearly suffices to prove there are no solutions with x, y, z, w nonnegative and not all 0. Suppose otherwise, and let $(x, y, z, w)=(a, b, c, d)$ be a solution with $a+b+c+d$ minimal. Note that $a^2\equiv 2c^2$ (mod 3), and because 0 and 1 are the only squares modulo 3, we conclude that $a\equiv c\equiv 0$ (mod 3). Thus, $a=3m$ and $c=3n$ with $m+n<a+c$. (Otherwise, $a=c=0$, implying $\sqrt{2}b=d$, hence $b=d=0$.) But then $3m^2+2b^2=6n^2+d^2$, so the 4-tuple (d, n, b, m) satisfies the original equation. However, $d+n+b+m<a+b+c+d$, contradicting the minimality of $a+b+c+d$.

A875. Given $s\in S$, let $k_s\geq 0$ denote the number of times s appears in the list. The condition is that not more than n of the k_s are odd. To prove necessity of the condition, consider a division into triples satisfying the hypothesis. Observe that if k_s is odd, then there must exist some triple containing an odd number of s's. However, each triple includes exactly one element that occurs an odd number of times in the triple. Since there are n triples, k_s is odd for at most n distinct s. To prove sufficiency, assume that k_s is odd for at most n elements $s\in S$. For each s with k_s odd, begin a triple with one s in an empty group. Every s appears with even multiplicity in the remainder of the list. Thus, we may form n pairs of identical elements from these remaining elements. Place one pair in each of the n triples, thus ensuring that no triple consists of three distinct elements. Finally, put one of the remaining elements in each of those triples with only two elements.

A876. The determinant of $A+xB$ is a polynomial in x of degree at most n, with det B the coefficient of x^n. What is given implies that $\det(A+kB)=\pm 1$ for $k=0,1,\ldots,2n$. Thus, the polynomial $\det(A+xB)$ takes on the same value for $n+1$ values of x, hence must be a constant. Therefore, det $B=0$.

REVIEWS

PAUL J. CAMPBELL, *editor*
Beloit College
1997–98: University of Augsburg,
Germany

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 mathematics literature. Readers are invited to suggest items for review to the editors.

Singh, Simon, and Kenneth A. Ribet, Fermat's last stand, *Scientific American* (November 1997) 68–73.
Mauldin, R. Daniel, A generalization of Fermat's Last Theorem: The Beal Conjecture and Prize Problem, *Notices of the American Mathematical Society* 44 (December 1997); also available at `http://www.ams.org/publications/notices/199722/beal.html` .
Devlin, Keith, Devlin's Angle: Move over Fermat, now it's time for Beal's problem, `http://www.maa.org/devlin/devlin_12_97/html` .
Peterson, Ivars, Prize offered for solving number theory conundrum, *Science News* 152 (15 November 1997) 310; also available at `http://www.sciencenews.org/sn_arc97/11_15_97/.fob2.htm` .
Peterson, Ivars, The amazing ABC conjecture, `http://www.maa.org.mathland/mathtrek_12_8.html` .
Darmon, H., and A. Granville, On the equations $z^m = F(x,y)$ and $Ax^p + By^p = cZ^r$, *Bulletin of the London Mathematical Society* 27 (1995) 513–543.

In 1908 physician Paul Wolfskehl established a prize for the proof of Fermat's Last Theorem (FLT), and Andrew Wiles received its $50,000 last year. With FLT resolved, what's left to do? Now a Texas banker, Andrew Beal, is offering max{$50,000, $5,000 × ($Y$ − 1996)} for the resolution in year Y of the more general *Beal's conjecture*: *The equation* $x^m + y^n = z^r$ *has no solutions with* x, y, z *coprime for integers* $m, n, r > 2$. In other words, apart from squares, no two powers of integers sum to another power, unless they have a common factor (e.g., $2^3 + 2^3 = 2^4$ or $3^3 + 6^3 = 3^5$). FLT is the special case $m = n = r$. In 1995 Darmon and Granville showed that for fixed m, n, r, the equation has only finitely many solutions. (The related ABC conjecture, discussed previously in this column and also by Peterson, implies that $1/p + 1/q + 1/r > 1/2$, hence that there are no solutions for exponents sufficiently large.) Darmon and Granville also investigated related open problems and formulated what they call the *Fermat-Catalan conjecture*: *There are only finitely many solutions with* x, y, z *coprime when* $1/p + 1/q + 1/r < 1$. Mauldin and Peterson give further references.

Babai, László, Carl Pomerance, and Péter Vértesi, The mathematics of Paul Erdős, *Notices of the American Mathematical Society* 45 (January 1998) 19–31. Babai, László, and Joel Spencer, Paul Erdős (1913–1996), *ibid.* 64–73.

These commemorative articles offer an overview of Erdős's mathematical achievements and an account of his life and unusual work- and lifestyle. In addition, a recent film about Erdős, "*N* Is a Number—A Portrait of Paul Erdős" is available from the MAA, and more references are at the Web page `http://www.cs.uchicago.edu/groups/theory/erdos.html` .

Schulz, Andreas S., David B. Shmoys, and David P. Williamson, Approximation algorithms, *Proceedings of the National Academy of Sciences of the USA* 94 (November 1997) 12734–12735; also available at `http://www.pnas.org/cgi/content/full/94/24/12734` .

This paper is a brief nontechnical survey by leading experts of recent progress in approximation algorithms in the context of applications. An *α-approximation algorithm* is one that efficiently computes a solution whose value is within a factor α of optimal; we want α to be as close to 1 as possible. Randomization is one technique for approximation algorithms for NP-complete problems. The authors describe new randomized approaches that solve the maximum cut problem with an expected cut weight of $\alpha = 0.878$ of optimal, the routing problem in a communication network (minimizing congestion) with $\alpha = 1 + \epsilon$, and the problem of efficiently drilling holes in circuit boards (with Euclidean metric) with $\alpha = 1 + \epsilon$.

The 1997 Nobel Prize in Economics, `http://www.nobel.se/announcement-97/economy1997.html` .
Ferreyra, Guillermo, The mathematics behind the 1997 Nobel Prize in Economics, `http://www.ams.org/new-in-math/black-scholes-ito.html` .
Rubashs, Kevin, A study of option pricing models, `http://bradley.bradley.edu/%7Earr/bsm/model.html` .
Devlin, Keith, Devlin's Angle: A Nobel formula, `http://www.maa.org/devlin_11_97.html` .

The 1997 Nobel Prize in Economics was awarded to Robert C. Merton (Harvard) and Myron S. Scholes (Stanford) "for a new method to determine the value of [financial] derivatives" (such as stock options). That method involved the formulation of a model with appropriate assumptions and solution of the resulting stochastic differential equation (a differential equation whose solution is a stochastic process). The articles cited give a glimpse into the situation modeled and into the mathematics used.

Hayes, Brian, Square knots, *American Scientist* 85 (November-December 1997) 506–510.

What if space were (is?) discrete—that is, topologically equivalent to $\mathbf{Z}^3$ instead of to $\mathbf{R}^3$? This article investigates what knot theory in such a space is like. For instance, there is a smallest nontrivial knot, as measured by its length. Also, pursuing a three-dimensional random walk that otherwise avoids intersecting itself until it returns to its starting point produces a knot; the probability that this knot is the unknot tends to 0 as the number of steps increases.

Devlin, Keith, Making the invisible visible, `http://www.maa.org/features/invisible.html` .

"How do we set about rectifying the result of hundreds of years of bad press" for mathematics? So asks the former editor of MAA's *Focus* newsletter in a commencement address last spring. His answer, in all seriousness: "Sound bites ... the only way we have of changing public opinion .... I don't think there is much of a case to be made in favor of trying to [increase public understanding of mathematics]. What I want to change is the public *perception* of mathematics." He suggests two such sound bites: *Mathematics is the science of patterns*, and *Mathematics makes the invisible visible*, and gives several examples to show how each can encompass mathematics.

NEWS AND LETTERS

Guidelines for Authors

The following Guidelines have been updated in several respects from their earlier form, which appeared in the February 1996 issue of this MAGAZINE. *For instance, these Guidelines address some aspects of electronic submission of manuscripts and figures. These Guidelines (and additional information) are also available on the World Wide Web, at* `http://www.maa.org/pubs/mathmag.html` .

General information MATHEMATICS MAGAZINE is an expository journal of undergraduate mathematics.

Both adjectives in the preceding sentence are important. Articles submitted to the MAGAZINE should be written in a clear, lively, and inviting *expository* style. The MAGAZINE is not a research journal, so papers written in the "theorem-proof-corollary-remark" style are usually unsuitable for publication. The best contributions contain examples, applications, historical background, and illustrations.

Every article should contain interesting mathematics, readably presented. Originality and freshness of approach are essential. Original research results in pure mathematics are, as a rule, outside our purview, although description and exposition on research results may be appropriate.

We especially welcome papers that include a historical element, and papers that draw connections among various branches of the mathematical sciences, and between mathematics and other disciplines. Papers with educational or pedagogical content are welcome, and are held to the same high standards of mathematical content, exposition, and general interest as are other submissions. Educationally-focused papers that touch on other subjects, bring in history, or appeal to advanced undergraduates are more likely to be accepted.

Audience The MAGAZINE is an *undergraduate* journal in the broad sense that it addresses both teachers and students of collegiate mathematics. Among the intended uses of the MAGAZINE is to supplement and enliven undergraduate mathematics courses, especially at the upper undergraduate level. Articles, therefore, should be inviting and accessible to non-specialists, including well-prepared undergraduates. To this end, references should be provided generously, since we aim to invite readers to pursue ideas further. Bibliographies may contain suggested reading along with sources actually used or cited. Whenever possible, references should cite readily available sources, in their most recent editions.

What makes a good article? MATHEMATICS MAGAZINE is responsible first to its readers (most of whom are mathematical generalists), and then to its authors. A manuscript's publishability depends, therefore, as much on its quality of exposition as on its "pure" mathematical significance. Our general advice is simple: Say something new in an appealing way, or say something old in a refreshing way. But say it clearly and straightforwardly, assuming a minimum of technical background.

Good exposition in our sense is vigorous and informal, written in the active voice, and rich with helpful examples. Minimize computation; stress motivation, insight, and illustration. Illustrate your ideas with visually appealing graphics, including figures, diagrams, tables, drawings, and photographs.

First impressions are especially important. Titles should be short, descriptive, and attractive. The opening sentences should clearly summarize the paper's scope and aims. A successful introduction should aim to enlarge rather the paper's audience, rather than limit it to a few specialists.

Many useful references on good mathematical style and exposition are available; several are listed at the end of these notes. Some of these references may be especially helpful for writers who use computer writing environments.

Types of papers Most papers in the MAGAZINE are published either as Articles or as Notes. Articles have a broader scope than Notes, and usually run longer than 2000 words. Articles should be divided into a few subsections, each with a carefully chosen subtitle. Typical Notes are shorter, more narrowly focused, and less formally sectioned—a few paragraph headers should suffice. In addition to expository pieces, we publish Proofs without Words, Math Bites, and (in very limited quantity) poems, cartoons, and other mathematical miscellanea. See any issue of the MAGAZINE for examples of these genres.

Style and format Manuscripts should be clearly typewritten or laser-printed, with wide margins and line-spacing. The title, author, and author's address should appear at the top of the first page. Pages should be numbered.

References should be listed either alphabetically or in the order cited in the text; in either case, consistency is essential. Please adhere very closely to the MAGAZINE's style for capitalization, use of italics, etc. See any issue (and the references below) for examples. In particular, journal titles should be abbreviated as in *Mathematical Reviews.*

Figures and illustrations Figures may be either interspersed with text or appended to the end of a paper. (If the paper is accepted, *separate* copies of all figures must also be supplied, both with and without added lettering.) All figures should be numbered, and must be referred to by number in the text. Authors themselves are responsible for providing figures of publishable quality; the MAGAZINE has no "art department."

Submitting manuscripts As a rule, papers should be submitted to the MAGAZINE in physical form. Please submit three copies; keep another copy as protection against possible loss. Electronic submission is possible in limited circumstances, but we cannot guarantee any response to electronic submissions in formats that are obscure or unfamiliar to us. For details, contact mathmag@stolaf.edu .

Manuscripts and other correspondence should be mailed to

Paul Zorn, Editor, MATHEMATICS MAGAZINE, St. Olaf College, 1520
St. Olaf Avenue, Northfield, Minnesota 55057-1098.

Please include an e-mail address, if available.

Our referees are asked to check for mathematical accuracy, but also to give detailed suggestions on stylistic matters. In practice, almost all papers require some revision before being accepted for publication. After acceptance, papers are copy-edited in our office.

Electronic manuscripts Although original submissions will normally be in physical form, we appreciate receiving revisions and final versions in electronic form—ideally, in some variant of TeX or LaTeX, but any electronic form is better than none. Figures, if supplied electronically, should be saved to PostScript or Encapsulated PostScript (EPS) form.

Simple LaTeX "template" files are available for Articles and Notes; they can be had either by sending an e-mail request to `mathmag@stolaf.edu` or, via the Web, at `http://www.maa.org/pubs/mathmag.html` . These templates produce rough approximations to the appearance of Articles and Notes in the MAGAZINE.

REFERENCES

1. American Mathematical Society, *A Manual for Authors of Mathematical Papers*, 8th edition, Amer. Math. Soc., Providence RI, 1984.

2. R.P. Boas, Can we make mathematics intelligible? *Amer. Math. Monthly* 88 (1981), 727–731.

3. Harley Flanders, Manual for *Monthly* authors, *Amer. Math. Monthly* 78 (1971), 1–10.

4. Leonard Gillman, *Writing Mathematics Well*, Math. Assoc. of America, Washington DC, 1987.

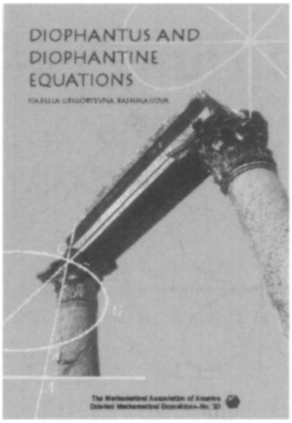
5. Andrew Hwang, Writing in the age of LaTeX, *AMS Notices* 42 (1995), 878–882.

6. D.E. Knuth, T. Larrabee, and P.M. Roberts, *Mathematical Writing*, MAA Notes #14, Math. Assoc. of America, Washington DC, 1989.

7. Steven K. Krantz, *A Primer of Mathematical Writing*, Amer. Math. Soc., Providence, RI, 1991.

8. N. David Mermin, *Boojums All the Way Through*, Cambridge Univ. Pr., Cambridge, UK, 1990.

THE MATHEMATICAL ASSOCIATION OF AMERICA



Diophantus and Diophantine Equations

I. G. Bashmakova

Updated by Joseph H. Silverman
Translated by Abe Shenitzer

Series: Dolciani Mathematical Expositions

Most readers associate the mathematics of antiquity with Euclid's *Elements* and the works of Archimedes and Apollonius. This wonderful little book will introduce the reader to a new aspect of the mathematics of antiquity in the works of Diophantus. The object of this book is to present the work of Diophantus, focusing on Diophantus' methods of obtaining rational solutions of indeterminate equations of the second and third order.

The first part of the book presents the elementary facts of algebraic geometry essential to understanding the rest of it. The book clears up the misconception that Diophantus relied on clever tricks rather than general methods to solve problems. Professor Bashmakova shows that in modern theorems, Diophantus used general methods to find rational points on algebraic curves of genus 0 and 1.

The second half of the book considers the evolution of the theory of Diophantine equations from the Renaissance to the middle of the 20th century. In particular, the book includes substantial descriptions of the relevant contributions of Viéte, Fermat, Euler, Jacobi, and Poincaré. The book ends with Joseph Silverman's survey of Diophantine analysis during the last twenty years in which he mentions the proof of the Mordei conjecture and of Fermat's Last Theorem.

The book is intended for a broad audience. It can be enjoyed by teachers as well as students at all levels.

Table of Contents: Introduction; Diophantus; Numbers and symbols; Diophantine equations; Evaluation of Diophantus' methods by historians of science; Indeterminate quadratic equations; Indeterminate cubic equations; Diophantus and number theory; Diophantus and the mathematicians of the 15th and 16th centuries; Diophantus' methods in the works of Viète and Fermat, Diophantine equations in the works of Euler and Jacobi; The geometric meaning of the operation of addition of points; The arithmetic of algebraic curves; Conclusion; Supplement: The role of concrete numbers in Diophantus' Arithmetic; Bibliography.

Catalog Code: DOL-20/JR
104 pp., Paperbound, 1997
ISBN-88385-526-7
List: $21.95 Member Price: $17.50

Phone in Your Order Now! ☎ *1-800-331-1622*

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________________

Address ______________________________

City ____________ State _____ Zip _______

Phone ______________________________

Qty.	Catalog Code	Price	Amount
_______	DOL-20/JR	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

TOTAL _______

Payment □ Check □ VISA □ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________

THE MATHEMATICAL ASSOCIATION OF AMERICA

Magic Tricks, Card Shuffling, and Dynamic Computer Memories

S. Brent Morris

Series: Spectrum

Magic Tricks, Card Shuffling, and Dynamic Computer Memories is a book that explores the fascinating interconnections between these seemingly unrelated topics. It is written for undergraduate mathematics, computer science, and electrical engineering majors, but it is accessible to motivated high school math students and magicians who want to understand the mathematics of card shuffling. It is a fun book that stands alone, but it could nicely supplement classes in discrete mathematics, combinatorics, algorithms, and computer networks. This book looks at the mathematics of the perfect shuffle and develops the algorithms for controlling dynamic memories (and doing some clever card tricks).

Each chapter begins with the description of a card trick and ends with its explanation, usually using mathematics developed in the chapter. The book itself is designed as a prop for a trick, but you don't need to use or understand any of its mathematics to do some good magic.

Read what reviewers have said about this book!

***Magic Tricks, Card Shuffling, and Dynamic Computer Memories** is essential reading for any magic buff who can faro shuffle or who wishes to acquire this unusual skill. The book will also be of great interest to computer scientists and to mathematicians working in the field of combinatorics ... anyone can read it with enjoyment and profit who is curious about the art and mathematics of card magic, or about the unexpected application of perfect shuffles to the storage and retrieval of computer information.*

—Martin Gardner

Provides a fascinating mix of history, mathematics and great magic tricks. I learned something on every page.

—Ron Graham, Chief Scientist, AT&T

....a tour de force for Morris... This is a most unusual and effective way to learn the concepts embodied in the interconnections of today's parallel computers.

—Harold Stone, NEC Research, Princeton

Table of Contents: **The Perfect Shuffle:** The Origins of the Perfect Shuffle; The Faro Dealer's Shuffle; The Mathematical Model of the Perfect Shuffle; The Stay-Stack Principle; Trick 1.2 (The Seekers, by Paul Swinford); **The Order of Shuffles:** The Product of Shuffles; Moving a Card in a Deck; Trick 2.9 (A Spelling Bee); **Shuffle Groups:** Randomizing a Deck of Cards; Shuffles and Cuts in Even Decks; Shuffles and Cuts in Odd Decks; Out- and In-Shuffles in an Even Deck; Trick 3.8 (A Challenge Poker Deal); **Generalizing the Perfect Shuffle:** Out-Shuffling Several Packets of Cards; Looking for a Neat Formula; Permutation Matrices; Generalizing Theorems; Generalized Shuffle Groups; Generalizing the In-Shuffle; Trick 4.7 (The Triple Seekers); **Dynamic Computer Memories:** The Shift-Register Memory; The Perfect Shuffle Memory; The Shift-Shuffle Memory; Details, Details, Details; The Perfect-Shuffle Memory for $N=2n$: Sequential Accessing in a Perfect-Shuffle Memory of Size $N=2n$; Properties of Tours; Epilogue; Trick 5.20. (Unshuffled by Paul Gertner); **Appendix 1:** The Order of Shuffles; **Appendix 2:** How to do the Faro Shuffle; The Double or Ordinary Faro Shuffle; The Triple Faro Shuffle; Appendix 3: Tours on Decks of Size 8, 16, 32, and 64; **Bibliography:** Selected Perfect Shuffle References.

Catalog Code: CARDS/JR
150 pp., Paperbound, 1998
ISBN 0-88385-527-5
List: $28.95 MAA Member: $22.50

Phone in Your Order Now! ☎ *1-800-331-1622*

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders:** $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________________

Address ______________________________

City ________________ State _____ Zip ________

Phone ______________________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	**CARDS/JR**	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

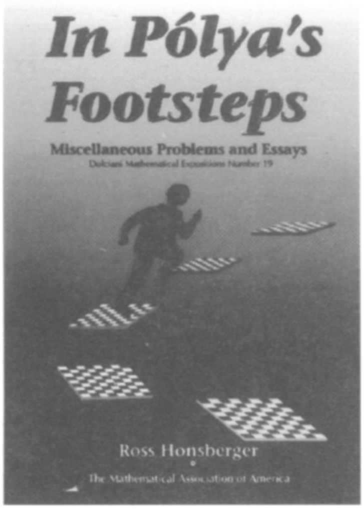
TOTAL _______

Payment □ Check □ VISA □ MasterCard

Credit Card No. ______________________ Expires ___/___

Signature ______________________________

THE MATHEMATICAL ASSOCIATION OF AMERICA


In Pólya's Footsteps
Miscellaneous Problems and Essays
Dolciani Mathematical Expositions Number 19
Ross Honsberger
The Mathematical Association of America

In Pólya's Footsteps

Miscellaneous Problems and Essays

Series: Dolciani Mathematical Expositions

Ross Honsberger

Another elegant collection of problems from Ross Honsberger

The study of mathematics is often undertaken with an air of such seriousness that it doesn't always seem to be much fun at the time. However, it is quite amazing how many surprising results and brilliant arguments one is in a position to enjoy with just a high school background. This is a book of miscellaneous delights, presented not in an attempt to instruct but as a harvest of rewards that are due good high school students and, of course, those more advanced — their teachers, and everyone in the university mathematics community. Admittedly, they take a little concentration, but the price is a bargain for such gems.

A half dozen essays are sprinkled among some hundred problems, most of which are the easier problems that have appeared on various national and international Olympiads. Many subjects are represented — combinatorics, geometry, number theory, algebra, probability, ... The sections may be read in any order. The book concludes with twenty-five exercises and their detailed solutions.

Something to delight will be found in every section — a surprising result, an intriguing approach, a stroke of ingenuity — and the leisurely pace and generous explanations make them a pleasure to read.

The inspiration for many of the problems came from the Olympiad Corner of *Crux Mathematicorum,* published by the Canadian Mathematical Society.

Catalog Code: DOL-19/JR
328 pp., Paperbound, 1997
ISBN 0-88385-326-4
List: $28.95 MAA Member: $23.00

Phone in Your Order Now! ☎ 1-800-331-1622

Monday – Friday 8:30 am – 5:00 pm FAX (301) 206-9789

or mail to: The Mathematical Association of America, PO Box 91112, Washington, DC 20090-1112

Shipping and Handling: Postage and handling are charged as follows: **USA orders (shipped via UPS):** $2.95 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 10 days of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas orders**: $3.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $7.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit Card orders are accepted for all customers.

Name ______________________________

Address ______________________________

City ________________ State _____ Zip ________

Phone ______________________________

QTY.	CATALOG CODE	PRICE	AMOUNT
_______	DOL-19/JR	_______	_______

All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Shipping & handling _______

TOTAL _______

Payment □ Check □ VISA □ MasterCard

Credit Card No. ____________________ Expires ___/___

Signature ______________________________

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, D.C. 20036