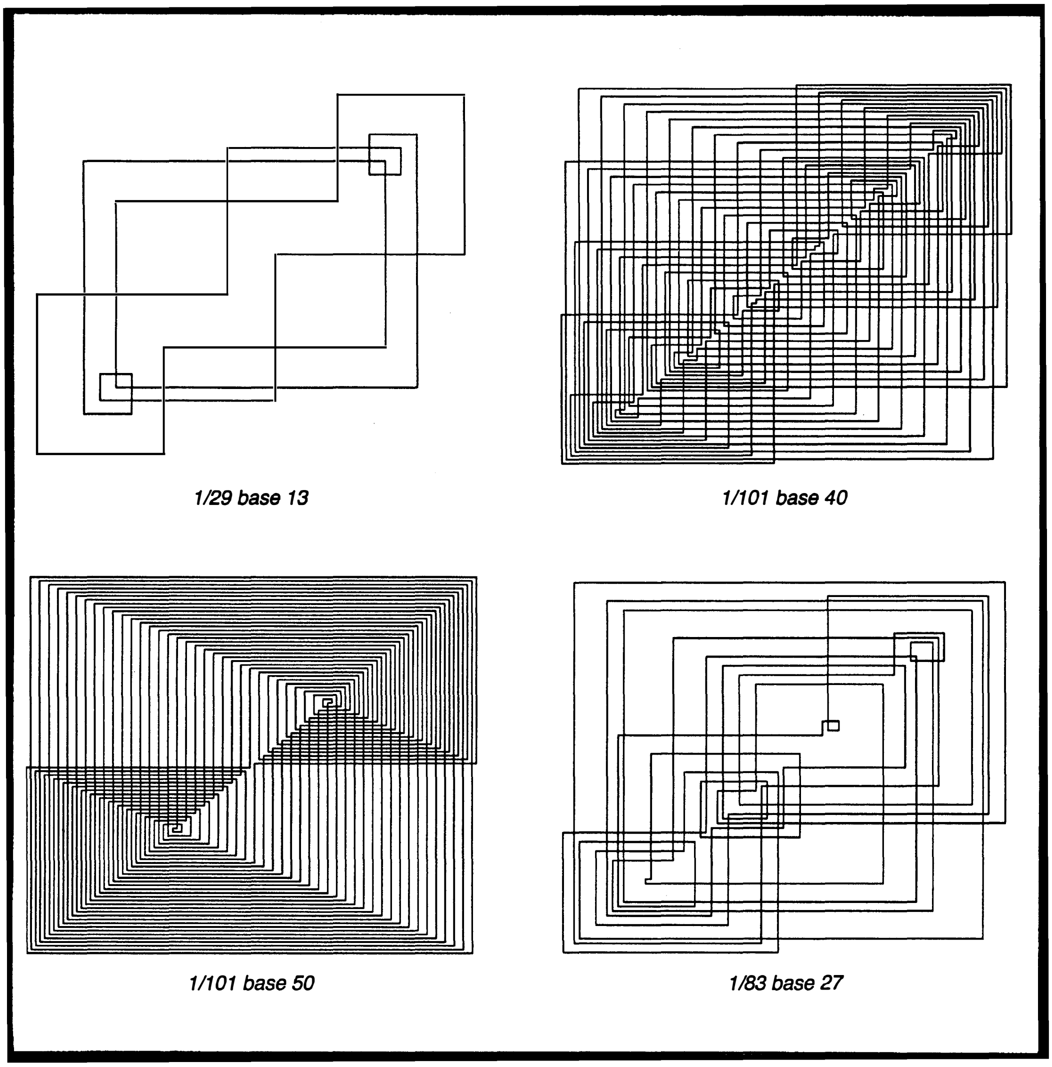
Vol. 73, No. 2, April 2000

MATHEMATICS MAGAZINE



1/29 base 13

1/101 base 40

1/101 base 50

1/83 base 27

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate 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 among 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 Frank Farris, Editor-Elect, Department of Mathematics and Computer Science, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0290. 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.

Cover illustration: The figures illustrate the graphical analysis of $F(x) = nx \pmod b$ for different values of n and b. The function $F(x)$ is tied closely to the fraction $1/n$ and the base b, so each figure is essentially a picture of $1/n$ in base b. See p. 83 ff.

AUTHORS

Arthur Benjamin is Associate Professor of Mathematics at Harvey Mudd College in Claremont, California, where he has taught since 1989, after earning his PhD from Johns Hopkins. His research interests are in combinatorics and game theory. He recently received the 2000 Deborah and Franklin Tepper Haimo Award for Distinguished Teaching from the MAA. He dedicates this paper to his baby daughter, Laurel, who is just learning to count.

Trygve Breiteig started his professional career as a high school teacher. His enthusiasm for mathematics pushed him to a master's degree at the University of Oslo in 1970. His interest in teaching and learning led him to the field of teacher education, in which he has been employed at Agder College, Kristiansand, Norway, since 1971. He has taught mathematics education and mathematics, and has developed curricula and textbooks for schools and for teacher education. His special interests include educational approaches to number theory and geometry.

Peter Hamburger was born in Mexico, grew up in Hungary, and is now an American. He received his PhD in Budapest. His research interests are combinatorics, graph theory, geometry, and set theory. His monograph, *Set Theory*, with A. Hajnal (Cambridge University Press) was published last year. He has taught at Indiana University— Purdue University Fort Wayne since 1989. He is the proud father of two yorkshire terriers, Alex and Samy.

Rafe Jones received a bachelor's degree in mathematics and French from Amherst College in 1998. After a one-year interlude as a visiting student at the Ecole Normale Supérieure in Paris, he began his graduate studies at Brown University in 1999. He expects Algebra and Number Theory to occupy much of his attention in the coming years. His involvement in the world of fractions stems from an informal summer research project with Prof. Pearce. Throughout the lengthy project, he has examined the graphs of hundreds of fractions, and delighted in discovering the often unexpected personality of each image.

Jan Pearce has been teaching mathematics and computer science at Berea College since 1992. She received her Ph.D. in mathematics from the now-infamous mathematics department at the University of Rochester, and her B.A. from Augustana College in Rock Island, Illinois, with majors in computer science, mathematics, and physics. Her research interests range widely, from environmental modeling research to number theoretic work like that explored in the present article. She enjoys involving students in research—which is, in fact, how this article came into being. She and Rafe Jones began these explorations while Rafe Jones was a high school student taking mathematics courses at Berea College. During the scant hours in which she is not working, Jan Pearce enjoys participating in traditional music and dance.

Raymond E. Pippert received his Ph.D. in mathematics under the supervision of S.M. Shah in 1965, having started as a physics major at the University of Kansas in Lawrence, where he was born and reared. His initial research was in classical analysis, but he changed to graph theory and combinatorics soon after he began teaching at Indiana University— Purdue University Fort Wayne, also in 1965. He spent three years in Malaysia as a professor and as a consultant to the Malaysian Ministry of Education. He is an avid sailor and traveler, especially in the Far East.

Jennifer Quinn is Associate Professor of Mathematics at Occidental College in Los Angeles. She earned her BA, MS, and PhD from Williams College; the University of Illinois, Chicago; and the University of Wisconsin, Madison, respectively. Her primary research interests are in combinatorics and graph theory. At a talk on combinatorial proofs of generalized Fibonacci numbers, a student asked about combinatorial connections between Fibonacci and continued fraction identities. Chris, here's your answer!

Francis Edward Su earned his Ph.D. in 1995 from Harvard, where his interest in continued fraction was sparked by a connection to a random walk he studied in his thesis. He is an assistant professor of mathematics at Harvey Mudd College and an MAA Project NExT "blue dot." In addition to probability, his research includes work in mathematical economics. In his spare time, he enjoys songwriting and teaches a mathematical enrichment course for 8th graders at a local middle school.

Steve Wassell received a bachelor's degree in architecture in 1984, a doctorate in mathematics in 1990, and a master's degree in computer science in 1999, all from the University of Virginia. He is currently the chair of the Department of Mathematical Sciences at Sweet Briar College, where he joined the faculty in 1990. After having published several articles in mathematical physics, he is now involved in three areas of research: the relationships between architecture and mathematics, computer science (genetic algorithms and physical design), and undergraduate-level topics in dynamical systems. Steve's overall aim is to explore and extol the mathematics of beauty and the beauty of mathematics.

Vol. 73, No. 2, April 2000

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.

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

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

Printed in the United States of America

A Postmodern View of Fractions and the Reciprocals of Fermat Primes

RAFE JONES
Brown University
Providence, RI 02912

JAN PEARCE
Berea College
Berea, KY 40404

Introduction and preliminaries

In America's visually-oriented, quantitatively illiterate culture, images have a great deal of power, so if a picture is today worth a thousand words, it must be worth at least a billion numbers. This power of the image is a hallmark of the postmodern era, in which the critical role of the observer has come to be recognized, and an understanding of the viewpoint has become inseparable from that of the object.

In some ways, the blossoming of chaos theory marked the arrival of mathematical postmodernism. Not so long ago, mathematical ideas were virtually unseen in American popular culture, and it took the enthralling fractal images of chaos theory to change that: the studies of chaos and fractals became some of the most widely discussed mathematical topics ever, and pictures of fractal images such as the

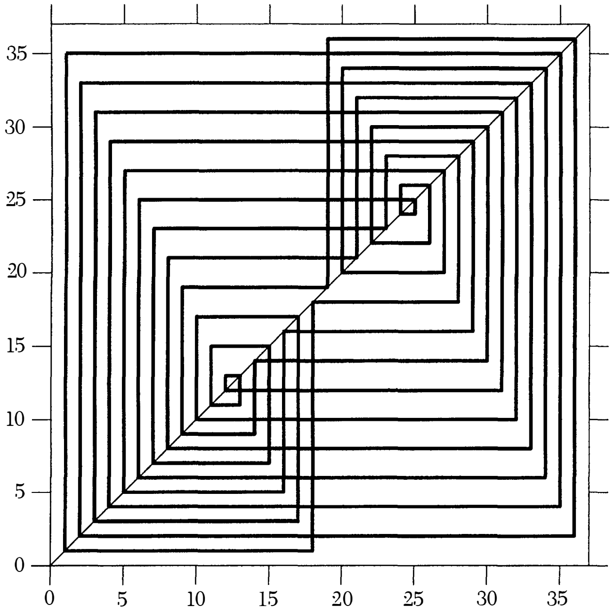


FIGURE 1
Graphical analysis of 1/37 base 35.

Mandelbrot set began cropping up on T-shirts and posters selling in American malls. The power of an image is difficult to underestimate, particularly when it comes to creating interest in a topic widely regarded as bland. Perhaps we could fuel a greater excitement in traditionally underappreciated areas of mathematics if only we could present them in a flashier graphical fashion. Take fractions, for instance, which to many people appear to be merely seas of numbers; after all, infinitely many fractions have infinitely long strings of digits as their decimal expansions. Wouldn't it be nice if we could see complicated fractions, like $\frac{1}{37}$ base 35, as simple images? Wouldn't it be even nicer if, as for the Mandelbrot set, those graphical images exposed something about the inherent mathematical structure that the concise algebraic expression only implied?

In this paper, we apply to the study of certain fractions the same graphical techniques used to transform the Mandelbrot set from algebra to image. This will enable us to turn arcane algebraic objects into eye-catching designs, such as the one pictured in FIGURE 1. What's more, the mathematics behind this metamorphosis is not very hard to describe. We begin by describing a somewhat unusual method of representing a fraction, which will be useful for our purposes. Fractions can be viewed in a number of ways, many of which are base-dependent: reduced or unreduced, as pieces of a pie, expanded into decimal, binary, octal, etc. The method we adopt is quite base-dependent, and relies upon the remainders generated at each stage of the long-division process in base b. Consider $\frac{1}{7}$, which has a base 10 (decimal) expansion of $0.\overline{142857}$. We can calculate this using the usual long division process in base 10 as follows:

$$\begin{array}{r}
0.142857\\
7\overline{)1.000000}\\
\underline{7}\\
30\\
\underline{28}\\
20\\
\underline{14}\\
60\\
\underline{56}\\
40\\
\underline{35}\\
50\\
\underline{49}\\
1
\end{array}$$

We can equivalently represent $\frac{1}{7}$ base 10 by writing the sequence of remainders produced in the above long division: $1 \to 3 \to 2 \to 6 \to 4 \to 5 \to 1$, a cycle that repeats infinitely. Note that what makes this a base ten long division is that we multiply the dividend by ten at every step; we could easily make it into a base b long division by multiplying by b at each step. This new long division would yield the sequence of remainders for $\frac{1}{7}$ base b; in fact, one can find the sequence of remainders for any fraction in any base simply by performing the appropriate long division. However, the laboriousness and iterative nature of long division make it desirable to have a simpler, more concise method of finding sequences of remainders. Happily, such a method exists, and it is simply the evaluation of the following function:

DEFINITION. Let a, b, and n be positive integers with $(n, a) = 1$ and $b > 1$. If r_i is the remainder produced at step i of the base b long division of $\frac{a}{n}$, the remainder produced at the $(i+1)$st step is given by $r_{i+1} = F_{b:n}(r_i) = b \times r_i (\text{mod } n)$. We call $F_{b:n}$ the *remainder function*, since if we begin with $r_0 = a$, iteration of $F_{b:n}$ yields the sequence of remainders of $\frac{a}{n}$ long divided in base b.

Note that a and n are relatively prime, so $\frac{a}{n}$ is a reduced fraction; we will assume throughout that all fractions are reduced. We can see the remainder function in action with the fraction used above, $\frac{1}{7}$ base 10. We begin with $r_0 = 1$. Next we have $r_1 = F_{10:7}(r_0) = 10 \times 1 \pmod 7 = 3$, followed by $r_2 = F_{10:7}(r_1) = 10 \times 3 \pmod 7 = 2$, $r_3 = F_{10:7}(r_2) = 10 \times 2 \pmod 7 = 6$, $r_4 = F_{10:7}(r_3) = 10 \times 6 \pmod 7 = 4$, $r_5 = F_{10:7}(r_4) = 10 \times 4 \pmod 7 = 5$, and $r_6 = F_{10:7}(r_5) = 10 \times 5 \pmod 7 = 1$.

Since $r_6 = r_0 = 1$, the sequence repeats. Note that each iteration of the remainder function simply multiplies by b and mods by n. Then, since r_0 is a, we can calculate the ith remainder directly using the formula $r_i = ab^i \pmod n$. This compact formula simplifies many arguments involving sequences of remainders, and you will see it often in the pages to come.

In the analysis above, our friend $\frac{1}{7}$ base 10 displays some surprising qualities. For example, $r_i + r_{i+3} = 7$ for all i. Moreover, if we let d_i represent the digit of the decimal expansion that is i places to the right of the decimal point, then in this example $d_i + d_{i+3} = 9$ for all i. These symmetries, as we shall see, have more than a numerical significance.

Before moving on to graphical topics, it will serve us well to discuss the three kinds of behavior a sequence of remainders (as well as the corresponding expansion) can exhibit. Each of these behaviors corresponds to a particular kind of graphical analysis graph, a concept we introduce in detail below. First, the sequence of remainders of $\frac{a}{n}$ in base b (as well as the corresponding base b expansion) may terminate; this happens if each remainder (and digit) is zero after some point, and such a fraction will have a graphical analysis graph that begins at some point and ends at some different point. This is the case if and only if every prime factor of n is also a prime factor of b. Second, the sequence may have a repeating cycle, but one that begins only after some initial string of remainders that never reappears. In this case, the graphical analysis graph will be an infinitely repeated figure, but with a tail created by the initial unrepeated string of remainders. This happens if and only if n has some factors that divide b and some that do not. Thirdly, the sequence may have only repeated cycles with no initial unrepeated string of remainders; this occurs if and only if n and b are relatively prime. This sort of fraction produces the neatest graphical analysis graph: a figure that retraces itself infinitely, with no unrepeated points.

The remainder function described above will allow us to work more easily with sequences of remainders. That it is a function also makes it a nice candidate for a graphical technique we will now introduce.

Graphical analysis

Graphical analysis or graphical iteration [2] gives us a visual way to explore function iteration. To graphically analyze a function $F(r)$, one does the following: Let r_0 be some number. Then, beginning with $i = 0$, draw a vertical line from (r_i, r_i) to the point $(r_i, F(r_i)) = (r_i, r_{i+1})$. (See FIGURE 2) From there, draw a horizontal line to $(F(r_i), F(r_i)) = (r_{i+1}, r_{i+1})$. Then increase i by one iteratively and repeat the preceding steps. Here, we will apply graphical analysis to our function $F_{b:n}(r)$. In order to avoid minor difficulties, we will say that if the remainder becomes zero at r_n, we stop the process at r_{n-1}. Although graphical analysis works only on functions, the remainder function associated with a given fraction is so closely tied to the fraction that we will refer to the graphical analysis of $F_{b:n}(r_i) = b \times r_i \pmod n$, with $r_0 = a$, as the graphical analysis of $\frac{a}{n}$ in base b.

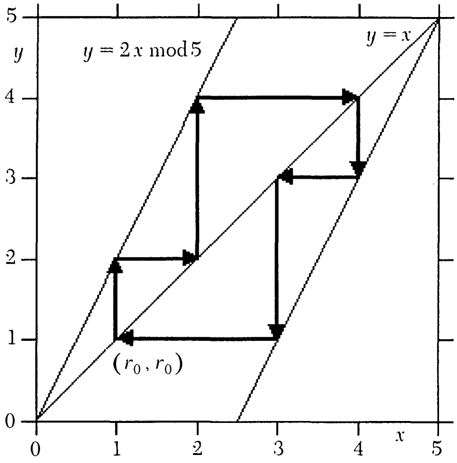


FIGURE 2

Graphical analysis of $\frac{1}{5}$ in base 2.

Note that the remainder function $F_{2:5}(x) = 2x \pmod 5$ plays a crucial role in Figure 2. However, you may have noticed that $F_{35:37}(x)$ does not appear in the graphical analysis graph of $\frac{1}{37}$ in base 35 (see Figure 1). The reason is that for so complex a picture, the slanted parallel lines of the remainder function become so dense as to obscure the image. Thus, despite their importance, for the sake of clarity we will omit them in the images to come.

Also, although the remainder function is theoretically important, one may graphically analyze a fraction without drawing the graph of the remainder function itself. In effect, the graphical analysis begins at the point (r_0, r_0), proceeds first vertically then horizontally to (r_1, r_1), then moves vertically then horizontally again to (r_2, r_2), and continues in this fashion. Hence in practice one can graphically analyze a fraction in a given base as follows: Compute the sequence of remainders; for each remainder, draw the appropriate dot on the line $y = x$; then connect the dots (following the order of the sequence of remainders), moving vertically then horizontally. Thus *the sequence of remainders entirely determines the graphical analysis graph of the fraction.* So when proving certain properties of graphical analysis graphs, such as various symmetries, we need not consider the entire image, but only the distribution of remainders.

Since the graphical analysis of a fraction varies from base to base, one might wonder how many distinct graphical pictures exist for a given fraction $\frac{a}{n}$. Bases zero and one are exempt from consideration. If b_1 and b_2 are bases such that $b_1 \equiv b_2 \pmod n$, then $ab_1{}^m \equiv ab_2{}^m \pmod n$, so $\frac{a}{n}$ will generate identical sequences of remainders in both bases. Thus, we only have to consider for our bases a single representative from each congruence class modulo n. This means, of course, that at most n bases may produce distinct graphs. Further narrowing the field is the fact that if b is a base such that $b \equiv 0 \pmod n$ or $b \equiv 1 \pmod n$, the pictures are not very interesting: in the former case, all remainders save the first are zero, so the graphical analysis graph is merely a single point, since the analysis ends with the last nonzero remainder. In the latter case, if m is a positive integer, then $\frac{1}{n}$ written in base $mn + 1$ is $0.\overline{1}$, and the sequence of remainders is an infinite string of ones; again, the graphical analysis graph is a single point. We will exclude bases in the 0 congruence class in many later considerations. However, we will often be interested in all bases in which a fraction has a repeating expansion, and thus we will include bases in the 1 congruence class in spite of their graphical shortcomings.

The various graphs of a fraction in different bases often bear some relation to one another. The following definition will help us relate some of them to others.

Rotational graph pairs

DEFINITION 1. $\frac{a_1}{n_1}$ and $\frac{a_2}{n_2}$ are *rotational graph pairs* if the graphical analysis graph of $\frac{a_1}{n_1}$, when rotated 180° about the point $\left(\frac{n}{2}, \frac{n}{2}\right)$, produces the graphical analysis graph of $\frac{a_2}{n_2}$.

These pictures exemplify rotational graph pairs:

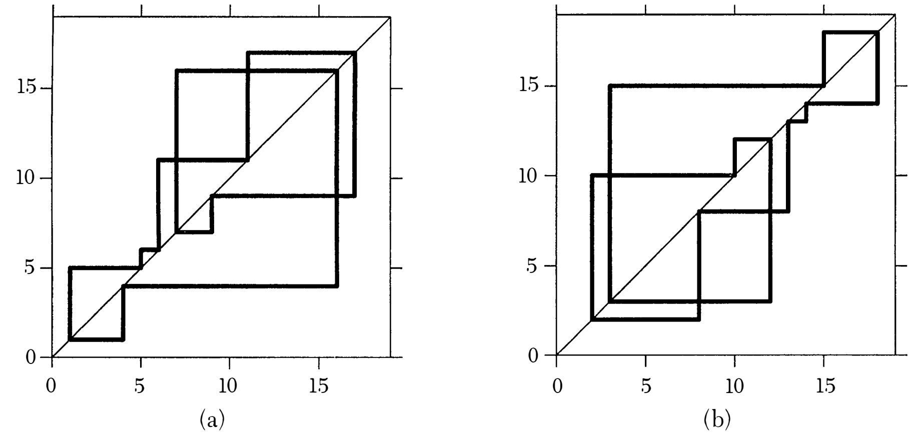


FIGURE 3
Graphical analysis of 17/19 base 5 vs. 2/19 base 5.

Since the graph of a fraction in base b depends entirely on its sequence of remainders, we can show that two fractions are rotational graph pairs simply by showing that "rotating" the sequence of remainders of one fraction about the point $\left(\frac{n}{2}, \frac{n}{2}\right)$ produces precisely the other sequence. In other words, the sequences must be zero in exactly the same places, and whenever the ith remainders of both sequences are nonzero, they must be equidistant from the point $\left(\frac{n}{2}, \frac{n}{2}\right)$. This is true if and only if the remainders in question sum to n. Thus we need only show that adding corresponding nonzero terms in the two sequences of remainders invariably yields n.

THEOREM 1. *In each base b, $\frac{a}{n}$ and $\frac{n-a}{n}$ are rotational graph pairs.*

Proof. First note that the only possible remainders at any stage of the long division of $\frac{a}{n}$ in base b belong to the set $\{0, 1, 2, \ldots, n-1\}$. Now, for any i, $ab^i \pmod{n} + (n-a)b^i \pmod{n} = (ab^i + (n-a)b^i) \pmod{n} = nb^i \pmod{n}$. Since $nb^i \equiv 0 \pmod{n}$ we have that the sum of the ith remainders of each sequence must be either 0 or n. Note that it is impossible for the ith remainder of one sequence to be zero and the ith remainder of the other nonzero: the nonzero remainder would make the sum necessarily greater than zero, and the zero remainder would make the sum necessarily less than n. Hence the sequences are zero in precisely the same places. Finally, if corresponding terms in the two sequences are nonzero, they cannot sum to zero, and so must sum to n. ■

Part of the appeal of Theorem 1 lies in its breadth: it applies to any fraction in any base, regardless of the behavior of the fraction's sequence of remainders. However, in order to have breadth, one often must sacrifice depth. If we consider more restricted classes of fractions, we will be able to prove several stronger, more penetrating results.

We can extend Theorem 1 significantly if we restrict ourselves to fractions and bases that produce purely repeating sequences of remainders—that is, those satisfying $(b, n) = 1$. Since the graphs of these fractions consist of a single repeated figure, beginning with any point in the cycle will yield the same image. Thus if c_1 is a term in the sequence of remainders for $\frac{a}{n}$ base b, then the sequence of remainders of $\frac{c_1}{n}$ base b will go through exactly the same cycle, beginning at c_1 instead of a. Hence the two fractions $\frac{a}{n}$ and $\frac{c_1}{n}$ will produce identical graphs. Similarly, if c_2 is a term in the sequence of remainders of $\frac{n-a}{n}$, then $\frac{c_2}{n}$ and $\frac{n-a}{n}$ will produce identical graphs. This corollary then follows immediately from Theorem 1:

COROLLARY 2. *Suppose b and n are relatively prime. If $ab^i \equiv c_1 \pmod{n}$ for some i and $(n-a)b^j \equiv c_2 \pmod{n}$ for some j, then $\frac{c_1}{n}$ and $\frac{c_2}{n}$ are rotational graph pairs in base b.*

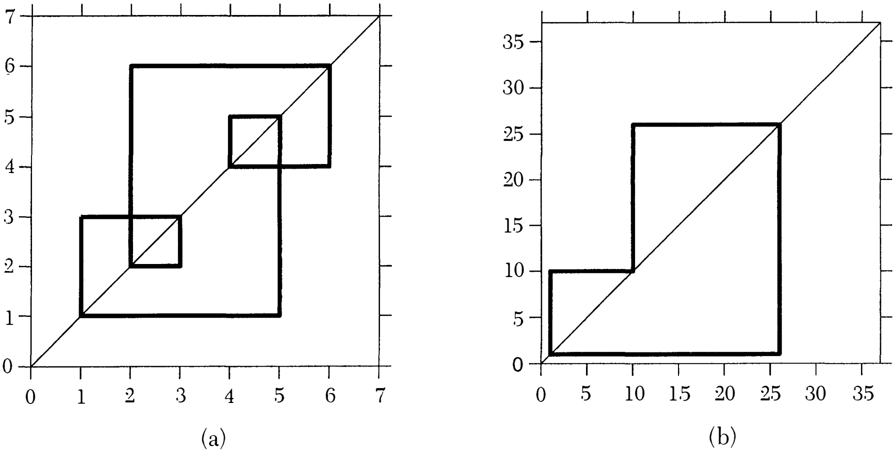
For example, $2 \times 100 \equiv 10 \pmod{19}$ and $17 \times 10 \equiv 18 \pmod{19}$, so $\frac{10}{19}$ base 10 and $\frac{18}{19}$ base 10 are rotational graph pairs.

Although we will return to this limited class of fractions later, in the next section we enlarge our consideration to include all sequences of remainders that do not terminate. The discussion hinges on a different sort of symmetry in the graphical analysis graph of a fraction: a rotational symmetry of a single graph, rather than of one graph to another.

Rotational symmetry

Consider the following two very different images in FIGURE 4:

The lovely rotational symmetry present in the graphical analysis graph of $\frac{1}{7}$ base 10 is strikingly absent in the graph of $\frac{1}{37}$. One might wonder why: after all, both 7 and 37



(a)

(b)

FIGURE 4
Graphical analysis of 1/7 base 10 vs. 1/37 base 10.

are not only relatively prime to 10, but also prime numbers. The following theorem will help to explain this difference.

THEOREM 3. *If $(n, a) = 1$ and n contains at least one prime factor that does not divide b then the following are equivalent:*

A. *$n - a$ appears in the sequence of remainders produced by the long division in base b of $\frac{a}{n}$ (i.e., $r_m = n - a$ for some m).*
B. *There exists an m, $0 < m < n$, such that for each natural number i, we have $r_i + r_{i+m} = n$.*
C. *The graphical analysis graph of the function $F_{b:n}$ beginning with $r_0 = a$ has 180° rotational symmetry about the point $\left(\frac{n}{2}, \frac{n}{2}\right)$.*

Proof. We will show A $\Rightarrow$ B by induction on i. By hypothesis, $r_0 + r_m = a + (n - a) = n$, so induction begins. Assuming that $r_i + r_{i+m} = n$, we must show that $r_{i+1} + r_{i+m+1} = n$. Using the remainder function, we have

$$r_{i+1} + r_{i+m+1} = F_{b:n}(r_i) + F_{b:n}(r_{i+m}) = b \times r_i \pmod{n} + b \times r_{i+m} \pmod{n}$$
$$= b \times (r_i + r_{i+m}) \pmod{n} = b \times n \pmod{n} = 0.$$

Thus $r_{i+1} + r_{i+m+1} \equiv 0 \pmod{n}$. Since n contains at least one prime factor that does not divide b, the sequence of remainders of $\frac{a}{n}$ base b does not terminate, so no remainder can be zero. Therefore $0 < r_{i+1} + r_{i+m+1} < 2n$, implying that $r_{i+1} + r_{i+m+1} = n$.

We now turn to B $\Rightarrow$ C. Condition B guarantees the existence of some positive integer m such that $r_i + r_{m+i} = n$ for each i. Let s be the smallest such integer. Since $r_s + r_{2s} = n = r_s + r_0$, it follows that $r_0 = r_{2s}$, and thus the length of the repeating cycle of the sequence of remainders is $2s$. Furthermore, the cycle is composed of the two halves $r_0, r_1, \ldots, r_{s-1}$ and $r_s, r_{s+1}, \ldots, r_{2s-1}$. Since $r_i + r_{s+i} = n$ for each i, these halves are essentially rotational graph pairs, and thus the whole graph is rotationally symmetric by itself.

Finally we address C $\Rightarrow$ A. Condition C means that our graph is rotationally symmetric about $\left(\frac{n}{2}, \frac{n}{2}\right)$, and since $r_0 = a$, (a, a) must be a point on the graph. Because of the graph's symmetry, $(n - a, n - a)$ must also be a point on the graph, implying that $n - a$ is a term in the sequence of remainders. Thus $r_m = n - a$ for some m. ■

Remarks and observations

In the example given above, 36 is indeed nowhere to be found in the sequence of remainders for $\frac{1}{37}$ base 10, which is 1 → 10 → 26, whereas 6 is the fourth number in the sequence for $\frac{1}{7}$ base 10. The equivalence of parts A and B thus predicts the visual discrepancy. In general, one need not go to the trouble of graphically analyzing a fraction to see if its graph is symmetric: it's enough to compute the sequence of remainders and examine it for a single number, $n - a$.

Interestingly, the symmetry among the remainders mentioned in part B of Theorem 3 is related to a similar symmetry among the digits. Suppose that the condition described in part B holds for a fraction $\frac{a}{n}$ in base b. The long division algorithm tells us that for each i, $b \times r_{i-1} = nd_i + r_i$ where d_i is the ith digit in the decimal expansion of $\frac{a}{n}$ in base b. Thus $nd_i + nd_{i+m} = b(r_{i-1} + r_{i+m-1}) - (r_i + r_{i+m}) = bn - n$. This implies that $d_i + d_{i+m} = b - 1$ for each i, a symmetry that we noted regarding $\frac{1}{7}$ base 10. A similar argument shows that the symmetry of remainders follows from the symmetry of digits, implying that the two are inseparable.

Symmetries in fractions with $(b, n) = 1$

Already the subject of Corollary 2, this class of fractions and its subclass of fractions with prime denominators will prove worthy of close scrutiny. Members of the larger class share one outstanding quality: in a given base b, rotational symmetry depends only on the denominator of the fraction in question (provided, of course, that the fraction is reduced). We make this precise in the next theorem.

THEOREM 4. *Let $\frac{a}{n}$ be a reduced fraction in base b, where $(b, n) = 1$. Then the graphical analysis graph of $\frac{a}{n}$ is rotationally symmetric in base b if and only if the graphical analysis graph of $\frac{1}{n}$ is rotationally symmetric in base b.*

Proof. Suppose that the graphical analysis graph of $\frac{1}{n}$ is rotationally symmetric in base b. The formula b^i (mod n) gives us the ith remainder of the long division of $\frac{1}{n}$ and ab^i (mod n) gives us the ith remainder of the long division of $\frac{a}{n}$. Since $\frac{1}{n}$ is rotationally symmetric, by Theorem 3 we have b^i (mod n) $+ b^{m+i}$ (mod n) $= n$ for each i and for some m satisfying $0 < m < n$. Thus $b^i + b^{m+i} \equiv 0$ (mod n). Multiplying through by a yields $ab^i + ab^{m+i} = an \equiv 0$ (mod n), implying that ab^i (mod n) $+ ab^{m+i}$ (mod n) $= 0$ or n. Since $(b, n) = 1$, the sequence of remainders of $\frac{a}{n}$ base b does not terminate, and thus no remainders can be zero. We therefore conclude that ab^i (mod n) $+ ab^{m+i}$ (mod n) $= n$, proving the rotational symmetry of $\frac{a}{n}$ in base b.

The converse argument is quite similar. Supposing ab^i (mod n) $+ ab^{m+i}$ (mod n) $= n$ for all i and for some m, we clearly have $ab^i + ab^{m+i} \equiv 0$ (mod n). We need only find a positive integer c such that $ca \equiv 1$ (mod n), and we will be able to

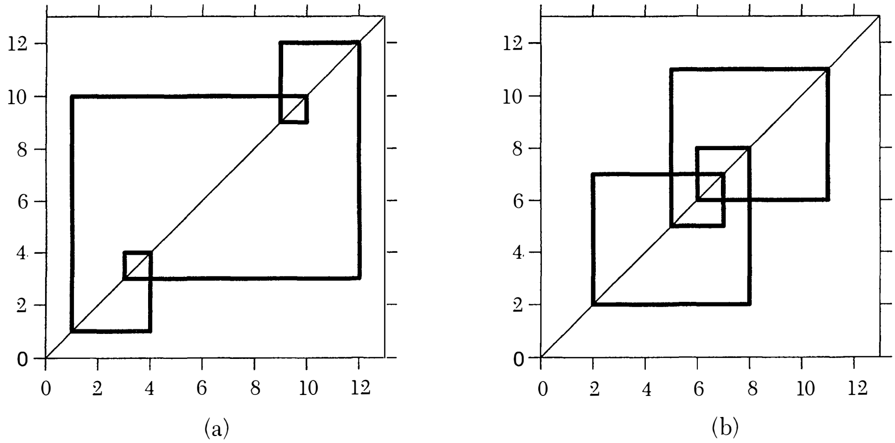


FIGURE 5
Graphical analysis of 1/13 base 10 vs. 5/13 base 10.

multiply through by c and complete the approach used above. Since our fraction is reduced, $(a, n) = 1$, so there exist positive integers c and d such that $ca + dn = 1$. This implies that $ca = 1 - dn \equiv 1 \pmod{n}$, so the desired positive integer does indeed exist. ■

This theorem guarantees that, for our limited class of fractions, if $\frac{1}{n}$ is rotationally symmetric in base b, then $\frac{a}{n}$ will be as well, provided $(a, n) = 1$. It often happens that $\frac{1}{n}$ and $\frac{a}{n}$ in fact produce identical graphs in base b; this is the case for $\frac{1}{7}$ and $\frac{a}{7}$ in base 10, where $(a, 7) = 1$. However, this need not happen, as FIGURE 5 shows.

Theorem 4 allows us to say that every reduced fraction with denominator n is either symmetric or not symmetric in any base b satisfying $(b, n) = 1$, since the value of the numerator plays no role. Thus for short, we will occasionally say simply that n is symmetric or not symmetric in base b.

The Euler totient function

We will be better able to understand symmetries in fractions with prime denominators with the aid of the *Euler totient function*. Denoted $\varphi(n)$, this function takes as input a positive integer n and produces as output the number of positive integers m that are less than or equal to n and satisfy $(m, n) = 1$. Some examples are $\varphi(4) = 2$, $\varphi(6) = 2$, $\varphi(12) = 4$, and, for any prime p, $\varphi(p) = p - 1$. The Euler totient function boasts two convenient properties which allow us to evaluate it easily for any small positive integer: First, if m and n are relatively prime positive integers, then $\varphi(mn) = \varphi(m)\varphi(n)$; and second, if p is prime and j is a positive integer, then $\varphi(p^j) = p^{j-1}(p-1)$ **[3]**. Thus, $\varphi(12) = \varphi(2^2 3) = \varphi(2^2)\varphi(3) = 2(2-1)2 = 4$. One of the better known theorems involving the Euler totient function is as follows:

LEMMA 5. (EULER'S FORMULA). *If b and m are positive integers and $(b, m) = 1$, then $b^{\varphi(m)} \equiv 1 \pmod{m}$.*

In particular, if p is prime and a not a multiple of p, we have $a^{p-1} \equiv 1 \pmod{p}$. Given a fraction with a purely repeating sequence of remainders, it's natural to be curious about the length of the repeating cycle (also known as the sequence's *period*). Euler's formula gives us some information about this period. Suppose $a < m$ and $\frac{a}{m}$ base b has a purely repeating sequence of remainders; we noted earlier that this is the case if and only if $(b, m) = 1$. The first remainder r_0 in the sequence is $ab^0 = a$, so the period of the sequence is the smallest nonzero k such that $r_k = a$. In other words, the period is the smallest nonzero k such that $ab^k \equiv a \pmod{m}$. Since $(b, m) = 1$, Euler's formula tells us that $b^{\varphi(m)} \equiv 1 \pmod{m}$, and thus $ab^{\varphi(m)} \equiv a \pmod{m}$. Because the period is the smallest nonzero k with $ab^k \equiv a \pmod{m}$, and $\varphi(m)$ satisfies this congruence, it follows that the period must divide $\varphi(m)$. In the special case where p is prime and b is not a multiple of p, we have the useful fact that the period of the sequence of remainders of $\frac{a}{p}$ in base b divides $p - 1$.

Fractions with prime denominators

Consider for a moment a reduced fraction with a prime denominator p in a base b that is not a multiple of p. Clearly $(b, p) = 1$, so Theorem 4 applies, showing that the value of the numerator does not affect the symmetry of the fraction's graph. Thus to determine if p is symmetric in base b, it is enough to examine the behavior of $\frac{1}{p}$ in base b. Although this is nice, we can use our restriction to fractions with prime denominators to get something even nicer: a convenient characterization of rotational symmetry.

Any reduced fraction with prime denominator p in a base b satisfying $(b, p) = 1$ must have a purely repeating sequence of remainders. The period of this sequence has everything to do with the rotational symmetry of the fraction: an even period means symmetry, an odd period no symmetry. We enshrine this convenient characterization in the following theorem:

Theorem 6. *Let m be the smallest positive integer such that $b^m \equiv 1 \pmod{p}$, where p is an odd prime and $(b, p) = 1$. Then $\frac{1}{p}$ is rotationally symmetric in base b if and only if m is even.*

Proof. First note that because $(b, p) = 1$ and p is prime, it follows from Euler's formula that $b^{p-1} \equiv 1 \pmod{p}$, so there exists some positive integer satisfying $b^m \equiv 1 \pmod{p}$. Hence it makes sense to discuss the smallest such integer. Now suppose $\frac{1}{p}$ is rotationally symmetric in base b, and let c_1 be a term in the sequence of remainders of $\frac{1}{p}$ base b. Then for some i, $b^i \equiv c_1 \pmod{p}$. Since $0 < c_1 < p$, we have $(c_1, p) = 1$, so, by Theorem 6, $\frac{c_1}{p}$ must be rotationally symmetric in base b. Thus, by Theorem 3, $p - c_1$ must appear in the sequence of remainders of $\frac{c_1}{p}$ base b. Hence for some j, $c_1 b^j \equiv p - c_1 \pmod{p}$. Since $b^i \equiv c_1 \pmod{p}$, we have $b^{i+j} \equiv c_1 b^j \equiv p - c_1 \pmod{p}$, implying that $p - c_1$ is in the sequence of remainders of $\frac{1}{p}$ base b. Thus each remainder r in the repeating cycle of $\frac{1}{p}$ base b occurs together with $p - r$. Since p is odd, we cannot have $r = p - r$, so the elements of the cycle occur in distinct pairs. Hence the cycle length must be even. Given that the first remainder is $b^0 \pmod{p} = 1$, this means that the smallest positive integer satisfying $b^m \equiv 1 \pmod{p}$ is even.

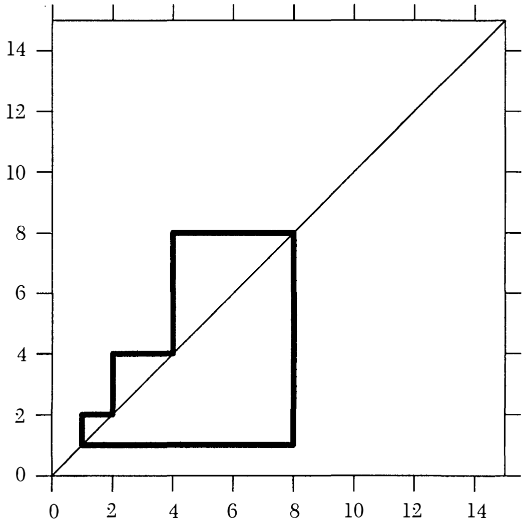


FIGURE 6
Graphical analysis of 1/15 base 2.

To show the converse, let m be the smallest positive integer satisfying $b^m \equiv 1 \pmod{p}$, and suppose m is even. Then $m = 2d$ for some positive integer d. Hence $b^m = b^{2d} = (b^d)^2 \equiv 1 \pmod{p}$. Thus $(b^d + 1)(b^d - 1) \equiv 0 \pmod{p}$ and since p is prime either $b^d \equiv 1 \pmod{p}$ or $b^d \equiv -1 \pmod{p}$. The first case is clearly impossible since $d = \frac{m}{2} < m$, and m was assumed to be the smallest positive integer such that $b^m \equiv 1 \pmod{p}$. Thus we conclude that $b^d \equiv -1 \equiv p - 1 \pmod{p}$. Therefore $p - 1$ is the dth remainder of $\frac{1}{p}$ base b, so by Theorem 3 $\frac{1}{p}$ is rotationally symmetric in base b. ■

Note that p must be prime for the above theorem to hold. Consider FIGURE 6, which shows that $\frac{1}{15}$ base 2 is not symmetric, though $2^m \equiv 1 \pmod{15}$ gives us a smallest m of 4.

Counting bases that produce symmetry

One might be tempted to guess that a prime number is symmetric in some randomly distributed number of bases; delightfully, this is not so. As we noted earlier, to find the ratio of bases in which a prime p is symmetric, we need only consider a single base b from each congruence class mod p. We will be interested here only in the bases in which $\frac{1}{p}$ has a repeating sequence of remainders. Therefore our considered bases will be all bases except those in the 0 congruence class.

For example, the reciprocal of 19 is symmetric in 9 of the 18 bases between 2 and 20, excluding 19; thus it is symmetric in half of the considered bases. The reciprocals of many other prime numbers are also symmetric in $\frac{1}{2}$ of the considered bases; some examples are 3, 7, 11, 23, 31, and 59. Other primes have reciprocals that are symmetric in $\frac{3}{4}$ of the considered bases; the first few are 5, 13, 29, 37, and 61. Still other primes, including 41, 73, and 89, have reciprocals symmetric in $\frac{7}{8}$ of the considered bases. In fact, one can find prime numbers that are symmetric in $\frac{(2^n - 1)}{2^n}$ of the considered bases for many positive integers n. We can explain this separation of the prime numbers into families, but to do so we will need a couple of number-theoretic results. [For details, see for example **[3]**.]

LEMMA 7. *For any $n \geq 1$, we have $n = \sum_{d|n} \varphi(d)$ where the sum is taken over all divisors of n.*

LEMMA 8. *Let p be a prime number and d a positive divisor of $p - 1$. Then there are exactly $\varphi(d)$ numbers b that are incongruent* (mod p) *and have the property that d is the smallest positive integer satisfying $b^d \equiv 1 \pmod{p}$.*

To illustrate Lemma 8, let $p = 7$ and choose $d = 3$. Lemma 8 tells us that there are $\varphi(3) = 2$ possible bases b which are not congruent (mod 7) and have the property that while b^1 and b^2 are not congruent to 1 (mod 7), b^3 is congruent to 1 (mod 7). In other words, were we to compute the sequence of remainders for $\frac{1}{7}$ in one base from each of the seven congruence classes mod 7, we would find that exactly two of them produce a sequence of period 3. If we want to know how many will yield a sequence with period 6, we simply have to calculate $\varphi(6)$, which is 2. The same holds for any other divisor of 6. To find the number of these bases in which the graph of $\frac{1}{7}$ is rotationally symmetric, Theorem 6 tells us we need only determine in how many of them $\frac{1}{7}$ has a sequence of remainders of even period. Since we ignore bases in the zero congruence class, all the bases we consider satisfy $(b, 7) = 1$. Since the period of $\frac{1}{7}$ in any of these bases must divide $\varphi(7) = 6$, the only possible even periods are 6 and 2. Thus our answer is $\varphi(6) + \varphi(2) = 3$, and we see that $\frac{1}{7}$ is symmetric in one half of the considered bases. This sort of analysis underlies the following proof.

THEOREM 9. *Suppose p is an odd prime number and n is the largest integer satisfying $2^n | p - 1$. Then, excluding bases $b \equiv 0 \pmod{p}$, $\frac{1}{p}$ is symmetric in $\frac{2^n - 1}{2^n}$ of the remaining bases.*

Proof. We need only consider a single representative base from each nonzero congruence class mod p. By Theorem 6, it suffices to find the number of bases in which $\frac{1}{p}$ produces a sequence of remainders of even period. The period of $\frac{1}{p}$ in any representative base must divide $\varphi(p) = p - 1$; so we want to find for each even divisor m of $p - 1$ the number of bases in which $\frac{1}{p}$ has a sequence of period m. Lemma 8 tells us that for a divisor q of $p - 1$, $\frac{1}{p}$ will produce a sequence of period q in exactly $\varphi(q)$ of the representative bases. Hence we need only compute $\Sigma\varphi(m)$, where m varies over the even divisors of $p - 1$. We will call the value of this sum k.

Suppose 2 divides $p - 1$ exactly n times. Then the prime factorization of $p - 1$ is $2^n p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N}$, where each p_j is an odd prime. The largest odd divisor of $p - 1$ is thus $D = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N}$. Now every even divisor m of $p - 1$ has the form $2^i t$, where $1 \le i \le n$ and t divides D. So we have

$$k = \sum_{i=1}^{n} \sum_{t|D} \varphi(2^i t).$$

Now since $t|D$ and D is odd, t must be odd. So $(2^i, t) = 1$ for any i, and by the first convenient property of the Euler function, we have $\varphi(2^i t) = \varphi(2^i)\varphi(t)$, so

$$k = \sum_{i=1}^{n} \sum_{t|D} \varphi(2^i)\varphi(t) = \sum_{i=1}^{n} \varphi(2^i) \sum_{t|D} \varphi(t).$$

By Lemma 7, $\Sigma_{t|D}\varphi(t) = D = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N}$, so

$$k = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N} \sum_{i=1}^{n} \varphi(2^i).$$

Now $\Sigma_{i=1}^{n}\varphi(2^i) = \varphi(2) + \varphi(2^2) + \cdots + \varphi(2^n)$. By the second convenient property of the Euler function, the right side is $2^0(1) + 2^1(1) + 2^2(1) + \cdots + 2^{n-1}(1) = 2^n - 1$, so we have $\Sigma_{i=1}^{n}\varphi(2^i) = 2^n - 1$. Finally we arrive at our value for k:

$$k = (2^n - 1) p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N}.$$

Since we are considering only a single base from each of the $p - 1$ nonzero congruence classes mod p, we have that $\frac{1}{p}$ is symmetric in

$$\frac{k}{p-1} = \frac{(2^n - 1) p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N}}{2^n p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_N^{\alpha_N}} = \frac{2^n - 1}{2^n}$$

of the considered bases. ∎

COROLLARY 10. *Suppose p is an odd prime and D is the largest odd divisor of $p-1$. Then, excluding bases $b \equiv 0 \pmod{p}$, $\frac{1}{p}$ fails to be symmetric in $\frac{D}{p-1}$ of the remaining bases.*

Proof. Suppose 2 divides $p-1$ exactly n times. Applying Theorem 9 we get that $\frac{1}{p}$ *fails* to be symmetric in $1-\frac{2^n-1}{2^n}=\frac{1}{2^n}$ of all the considered bases. The prime factorization of $p-1$ is 2^nD. Thus $\frac{1}{p}$ fails to be symmetric in $\frac{1}{2^n}=\frac{D}{2^nD}=\frac{D}{p-1}$ of the possible bases. ■

We noted earlier that in a base of the form $ap+1$, where a is a positive integer, $\frac{1}{p}$ will have a sequence of remainders that is simply an infinite string of ones. This leads to a graph consisting only of the fixed point $(1,1)$. If $p=2$, this graph is in fact symmetric about $(\frac{p}{2},\frac{p}{2})$, but for any odd prime it fails to be symmetric. Thus an odd prime must fail to be symmetric in all bases belonging to the 1 congruence class mod p. However, there exist odd primes that are symmetric in *all* of the other considered bases, and thus are as symmetric as it is possible for an odd prime to be.

Perfectly symmetric numbers and Fermat primes

DEFINITION. *A positive integer $n>1$ is perfectly symmetric if its reciprocal is symmetric in any base b provided $b \not\equiv 0 \pmod{n}$ and $b \not\equiv 1 \pmod{n}$.*

Clearly, 2 is trivially perfectly symmetric. This membership in the set of perfectly symmetric numbers makes 2 a spectacularly rare positive integer, joined only by widely-spaced comrades:

THEOREM 11. *The only perfectly symmetric numbers are 2 and the Fermat primes.*

Proof. Recall that a Fermat prime is a prime of the form $2^{2^m}+1$, where m is a natural number. Suppose n is a perfectly symmetric number, and suppose also that n is composite. Then there is some base b that divides n and satisfies $1<b<n$. Clearly b and n are not relatively prime. So the sequence of remainders of $\frac{1}{n}$ in base b cannot be purely repeating: either it terminates or has an initial string of unrepeated remainders. In the latter case, the string of unrepeated digits creates a tail in the graphical analysis graph of $\frac{1}{n}$ base b, and the tail ruins any symmetry. In the former case, $n-1$ cannot appear in the sequence of remainders, for if it did, we would have $r_k=n-1$ for some nonzero k, implying that $r_{2k}=1$. But the sequence of remainders terminates, so this is not possible. In either case n is not symmetric in base b, contradicting our assumption.

Therefore n must be prime. We have already seen that 2 is perfectly symmetric. If n is an odd prime, then, by Corollary 10, $\frac{1}{n}$ will fail to be symmetric in bases belonging to D of the $n-1$ nonzero congruence classes mod n, where D is the largest odd divisor of $n-1$. Any odd prime must fail to be symmetric in at least one of these base congruence classes, but since n is perfectly symmetric, it cannot fail in any of the others; therefore $D=1$. Thus no odd number greater than one can divide $n-1$, implying that $n-1$ is of the form 2^i for some i. Therefore $n=2^i+1$ and n is prime. If $i=uv$, where u is odd and $u>1$, then $2^v+1|2^i+1$, so 2^i+1 fails to be prime. Thus in this case our i must be of the form 2^k, where $k \in \mathbb{N}$. Therefore our prime n is of the form $2^{2^k}+1$, and is thus a Fermat prime.

Conversely, if n is either 2 or a Férmat prime, then clearly either $n = 2$, and is thus perfectly symmetric, or n is odd. In the latter case, by Corollary 10 we have that $\frac{1}{n}$ fails to be symmetric in bases belonging to D of the $n - 1$ nonzero congruence classes mod n, and must be symmetric in all the rest. Here $n - 1 = 2^i$, so $D = 1$. But the reciprocal of any number m must fail to be symmetric in bases belonging to the 1 congruence class mod m. Hence if $x > 1$ and $b \equiv x \pmod{n}$, $\frac{1}{n}$ is symmetric in base b. Therefore n is perfectly symmetric. ■

Currently there are only five known Fermat primes: 3, 5, 17, 257, and 65537. Thus, only six known perfectly symmetric numbers lurk among all the positive integers greater than one, suggesting that perfect symmetry is among the more unusual properties a number can have. However, precisely how many perfectly symmetric numbers exist remains an open question.

Questions and conclusions

Our discussion of the number of symmetry-producing bases for various fractions raises two questions about certain kinds of prime numbers:

Question 1. *Does there exist, for each positive integer n, a natural number k such that $2^n(2k + 1) + 1$ is prime?*

If so, then for any positive integer n one can find a prime p such that 2 divides $p - 1$ exactly n times. This would mean that for any positive integer n, primes exist that are symmetric in $\frac{2^n - 1}{2^n}$ of the considered bases.

Question 2. *How many Fermat primes are there?*

No one has any idea; we know only that there are at least five. Pierre de Fermat thought that all numbers of the form $2^{2^k} + 1$ were prime, but history has proven otherwise: All the numbers generated using $k = 5, \ldots, 11$ have turned out to be composite, as well as selected others, including the monstrous $2^{2^{23471}} + 1$. There remain, however, infinitely many more as-yet-undetermined possibilities. An answer to this question would also tell how many perfectly symmetric numbers exist.

Thus ends our exploration of fractions and symmetry. Postmodernism has taught us that all ways of looking at a problem are not equivalent: different perspectives highlight different properties. Adopting our society's penchant for images led us to examine more closely the symmetries of certain fractions, and opened our eyes to unexpected visions.

Note on the computer program During the course of this project, we wrote a simple computer program that graphically analyzes any fraction in any base. We found many of the images quite striking and beautiful, and were sorry not to be able to include all of them in this article. For those of you who would like to generate some of these images yourselves, our program is in an electronic supplement at `http://www.maa.org/pubs/mm_supplements/index.html`.

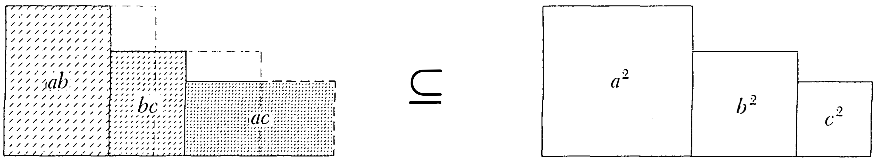
Acknowledgments. We would like to thank Dr. Mark Hanisch for his help with MATLAB, Dr. James Lynch for his interest and encouragement, the Reverend Kent Gilbert for moral support and occasional entertainment, and Dr. Libby Jones for timely sustenance.

REFERENCES

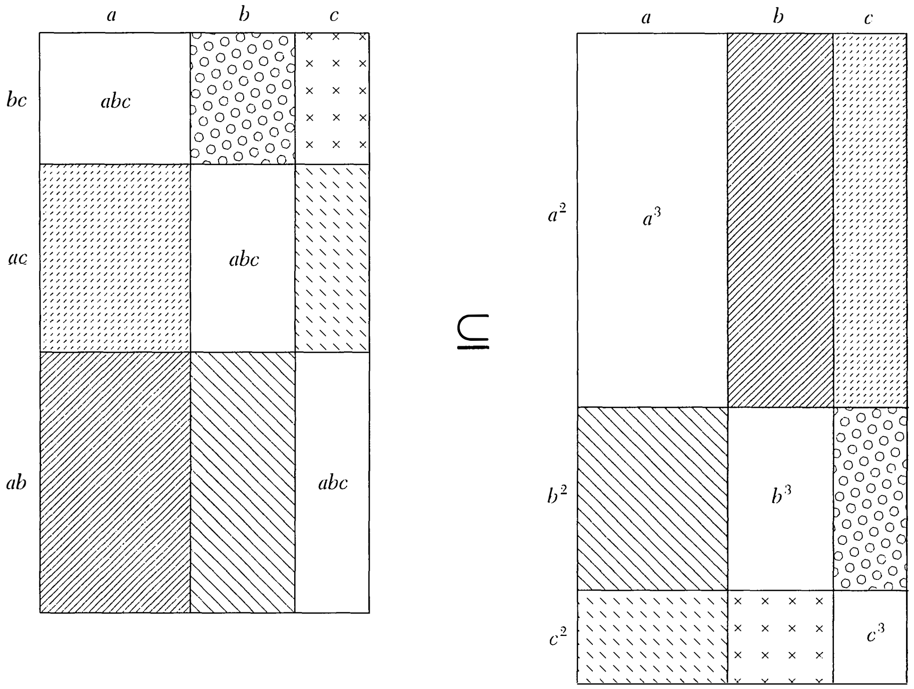
1. K. H. Becker and M. Dörfler, *Dynamical Systems and Fractals*, Cambridge University Press, Cambridge, UK, 1989.
2. R. Devaney, *Chaos, Fractals, and Dynamics*, Addison-Wesley, Reading, MA, 1990.
3. J. Silverman, *A Friendly Introduction to Number Theory*, Prentice-Hall, Upper Saddle River, NJ, 1997.

Proof Without Words: The Arithmetic–Geometric Mean Inequality for Three Positive Numbers

LEMMA 1. $ab + ac + bc \leq a^2 + b^2 + c^2$



THEOREM. $3abc \leq a^3 + b^3 + c^3$



—CLAUDI ALSINA
UNIVERSITAT POLITECNICA CATALUNYA
DIAGONAL 649, 08028 BARCELONA
SPAIN

Counting on Continued Fractions

ARTHUR T. BENJAMIN
FRANCIS EDWARD SU
Harvey Mudd College
Claremont, CA 91711

JENNIFER J. QUINN
Occidental College
Los Angeles, CA 90041

Introduction

You might be surprised to learn that the finite continued fraction

$$3+\cfrac{1}{7+\cfrac{1}{15+\cfrac{1}{1+\cfrac{1}{292}}}} \quad \text{and its reversal} \quad 292+\cfrac{1}{1+\cfrac{1}{15+\cfrac{1}{7+\cfrac{1}{3}}}}$$

have the same numerator. These fractions simplify to $\frac{103993}{33102}$ and $\frac{103993}{355}$ respectively. In this paper, we provide a combinatorial interpretation for the numerators and denominators of continued fractions which makes this reversal phenomenon easy to see. Through the use of counting arguments, we illustrate how this and other important identities involving continued fractions can be easily visualized, derived, and remembered.

We begin by defining some basic terminology. Given an infinite sequence of integers $a_0 \geq 0, a_1 \geq 1, a_2 \geq 1, \ldots,$ let $[a_0, a_1, \ldots, a_n]$ denote the finite continued fraction

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cfrac{1}{\ddots+\cfrac{1}{a_n}}}}$$

The *infinite* continued fraction $[a_0, a_1, a_2, \ldots]$ is the limit of $[a_0, a_1, \ldots, a_n]$ as $n \to \infty$. This limit always exists and is some irrational number α [3]. The rational number $r_n := [a_0, a_1, \ldots, a_n]$ is a fraction p_n/q_n in lowest terms, called the n-th *convergent* of α. It is well-known that p_n and q_n satisfy the recurrences

$$p_n = a_n p_{n-1} + p_{n-2}$$
$$q_n = a_n q_{n-1} + q_{n-2}$$

for $n \geq 2$, with initial conditions $p_0 = a_0$, $p_1 = a_1 a_0 + 1$, $q_0 = 1$, $q_1 = a_1$.

Now let's do some combinatorics. For a given continued fraction $[a_0, a_1, a_2, \ldots]$, consider the following tiling problem. Let P_n count the number of ways to tile a $1 \times (n+1)$ board with dominoes and stackable square tiles. All cells (numbered $0, 1, \ldots, n$) must be covered by a tile. Nothing can be stacked on top of a domino, but cell number i may be covered by a stack of as many as a_i square tiles, $i = 0, \ldots, n$.

FIGURE 1 shows an empty board with the *height conditions* $a_0, a_1, \dots, a_n$ indicated. FIGURE 2 gives an example of a valid tiling for a 1×12 board with height conditions $5, 10, 3, 1, 4, 8, 2, 7, 7, 4, 2, 3$.

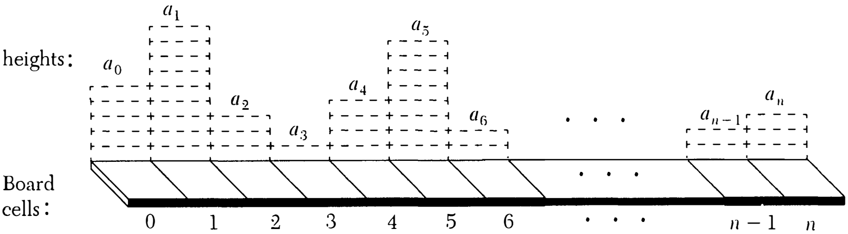


FIGURE 1

An empty $1 \times (n+1)$ board.

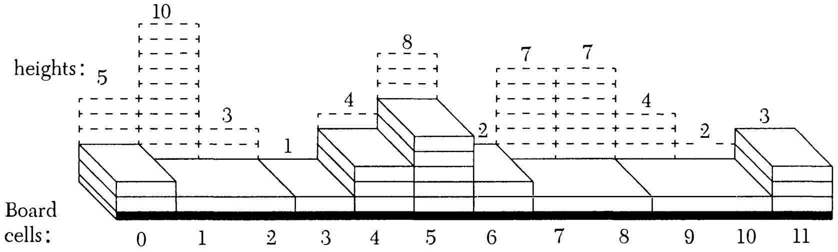


FIGURE 2

A tiling satisfying the height conditions $5, 10, 3, 1, 4, 8, 2, 7, 7, 4, 2, 3$.

For $n \geq 2$, we show

$$P_n = a_n P_{n-1} + P_{n-2}.$$

This follows from the observation that a tiling either ends with a stack of square tiles or a single domino. In the first case, there are a_n choices for the stack size and P_{n-1} ways to tile cells 0 through $n-1$. In the second case, there is only one choice for the last domino, and there are P_{n-2} ways to tile cells 0 through $n-2$. Using FIGURE 3 one can check that $P_0 = a_0$ and $P_1 = a_0 a_1 + 1$. Since P_n and p_n satisfy the same recurrence and initial conditions, we have $P_n = p_n$.

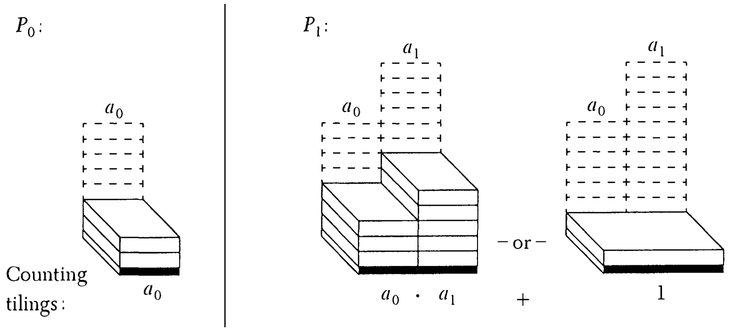


FIGURE 3

Verifying the initial conditions for the recurrence relation $P_n = a_n P_{n-1} + P_{n-2}$.

Removing cell 0 from the previous board, let Q_n count the number of ways to tile the $1 \times n$ board with dominoes and stackable square tiles, where the ith cell may be covered by a stack of as many as a_i square tiles, $i = 1, \dots, n$. By the same reasoning as before, (and letting $Q_0 = 1$ denote the "empty" tiling) we see that $Q_n = q_n$.

To illustrate, consider the continued fraction representation for π, which begins $[3, 7, 15, 1, 292, \dots]$. See Figure 4. If we count the number of ways to tile cells 0, 1, and 2, we get $p_2 = 333$. Counting the number of ways to tile only cells 1 and 2 easily gives us $q_2 = 106$. This produces the π approximation $r_2 = 333/106$. The reader should verify that tiling cells 0 through 3 produces $r_3 = 355/113$.

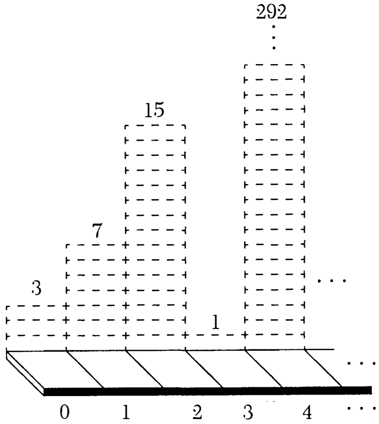


FIGURE 4
The beginning of the π board.

When $a_i = 1$ for all $i \geq 0$, it is well-known that the nth convergent p_n/q_n is the ratio of two consecutive Fibonacci numbers. Specifically, if we define $f_0 = 1$, $f_1 = 1$, and for $n \geq 2$, $f_n = f_{n-1} + f_{n-2}$, then $p_n = f_{n+1}$ and $q_n = f_n$. You may recall that the Fibonacci number f_n counts the number of ways to tile a $1 \times n$ board with 1×1 squares and 1×2 dominoes. So the continued fraction tiling problem generalizes the tiling interpretation of Fibonacci numbers [**1**, **2**].

Identities

Armed with our tiling interpretation, many well-known continued fraction identities can be explained combinatorially. We begin with the reversal identity.

Theorem 1. *If* $[a_0, a_1, \dots, a_{n-1}, a_n] = p_n/q_n$, *then* $[a_n, a_{n-1}, \dots, a_1, a_0] = p_n/p_{n-1}$.

Proof. Although one can easily prove this by induction, the theorem is nearly obvious when viewed combinatorially. To understand the common numerator, we see that the number of ways to tile the board with height conditions $a_n, a_{n-1}, \dots, a_1, a_0$ is the same as the number of ways to tile the board with height conditions $a_0, a_1, \dots, a_{n-1}, a_n$. The denominator of $[a_n, a_{n-1}, \dots, a_1, a_0]$ is the number of ways to tile the board with height conditions $[a_{n-1}, \dots, a_1, a_0]$, which by reversal is p_{n-1}.

The next few identities are useful for measuring the rate of convergence of convergents.

THEOREM 2. *The difference between consecutive convergents of* $[a_0, a_1, a_2, \dots]$ *is:* $r_n - r_{n-1} = (-1)^{n-1}/q_n q_{n-1}$. *Equivalently, after multiplying both sides by* $q_n q_{n-1}$, *we have*

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}.$$

Proof. Given height conditions $a_0, a_1, \dots, a_n$, let $\mathscr{P}_n$ denote the set of all tilings on cells $0, \dots, n$ and let $\mathscr{Q}_n$ denote the set of all tilings on cells $1, \dots, n$. Note that these sets have sizes $|\mathscr{P}_n| = p_n$ and $|\mathscr{Q}_n| = q_n$.

We will construct an *almost one-to-one correspondence* between the sets $\mathscr{P}_n \times \mathscr{Q}_{n-1}$ and $\mathscr{P}_{n-1} \times \mathscr{Q}_n$. Consider $(S, T) \in \mathscr{P}_n \times \mathscr{Q}_{n-1}$. For $i \geq 1$, we say (S, T) has a *fault* at cell i if both S and T have tiles that end at i. We say (S, T) has a fault at cell 0 if S has a square at cell 0. For instance, in FIGURE 5, there are faults at cells 0, 3, 5, and 6.

If (S, T) has a fault, construct (S', T') by swapping the "tails" of S and T after the rightmost fault. See FIGURES 5 and 6. Note that $(S', T') \in \mathscr{P}_{n-1} \times \mathscr{Q}_n$. Since (S', T') has the same rightmost fault as (S, T), this procedure is completely reversible.

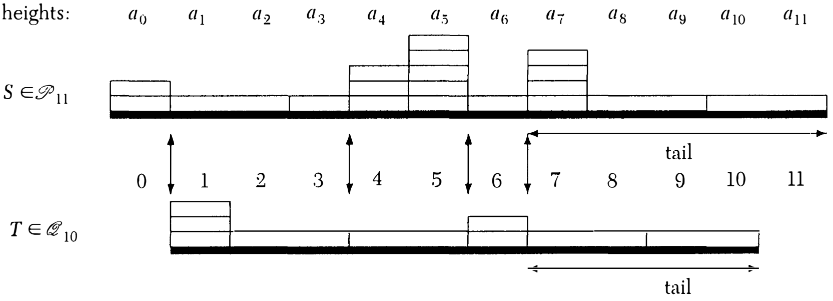


FIGURE 5
A pair of tilings with faults and tails indicated.

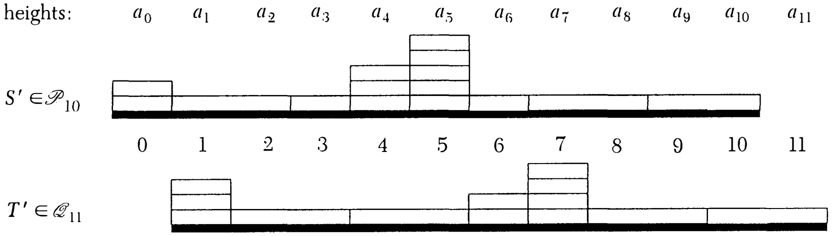

heights:
a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11}
$S' \in \mathscr{P}_{10}$
0 1 2 3 4 5 6 7 8 9 10 11
$T' \in \mathscr{Q}_{11}$

FIGURE 6
Result of swapping tails in FIGURE 5.

Notice when either S or T contains a square, (S, T) must have a fault. Thus the only fault-free pairs occur when S and T consist of all dominoes in *staggered formation* as illustrated in FIGURE 7. When n is odd (i.e., S and T both cover an even number of cells), there is precisely one fault-free element of $\mathscr{P}_n \times \mathscr{Q}_{n-1}$ and no fault-free elements of $\mathscr{P}_{n-1} \times \mathscr{Q}_n$. Therefore when n is odd, we have $|\mathscr{P}_n \times \mathscr{Q}_{n-1}| - |\mathscr{P}_{n-1} \times \mathscr{Q}_n| = 1$.

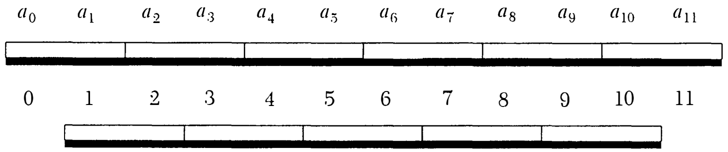

a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11}
0 1 2 3 4 5 6 7 8 9 10 11

FIGURE 7
The fault-free pair consists of staggered dominoes.

Similarly when n is even, there are no fault-free elements of $\mathscr{P}_n \times \mathscr{Q}_{n-1}$ and exactly one fault-free element of $\mathscr{P}_{n-1} \times \mathscr{Q}_n$. Hence when n is even, $|\mathscr{P}_n \times \mathscr{Q}_{n-1}| - |\mathscr{P}_{n-1} \times \mathscr{Q}_n| = -1$. Treating the odd and even cases together, we obtain

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}.$$

The next identity shows that the even convergents are increasing, while the odd convergents are decreasing.

THEOREM 3. $r_n - r_{n-2} = (-1)^n a_n / q_n q_{n-2}$. *Equivalently, after multiplying both sides by* $q_n q_{n-2}$, *we have*

$$p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n.$$

Proof. As in the last proof, we use tail swapping after the last fault to create a one-to-one correspondence between the "faulty" elements of $\mathscr{P}_n \times \mathscr{Q}_{n-2}$ and $\mathscr{P}_{n-2} \times \mathscr{Q}_n$. The proof is essentially given in FIGURES 8, 9, and 10.

The only unmatched elements are those that are fault-free. When n is odd, there are no fault-free elements of $\mathscr{P}_n \times \mathscr{Q}_{n-2}$, but there are precisely a_n fault-free

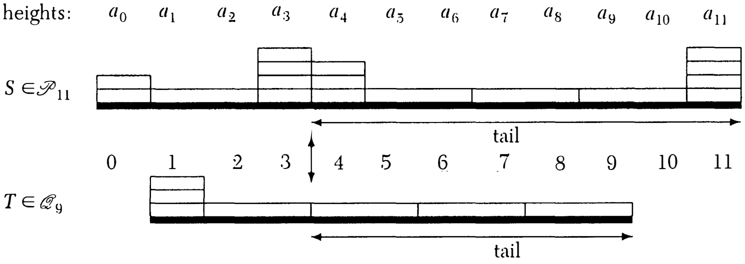

heights: a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11}
$S \in \mathscr{P}_{11}$
tail
0 1 2 3 4 5 6 7 8 9 10 11
$T \in \mathscr{Q}_9$
tail

FIGURE 8
An element of $\mathscr{P}_{11}\mathscr{Q}_9$ with rightmost fault indicated.

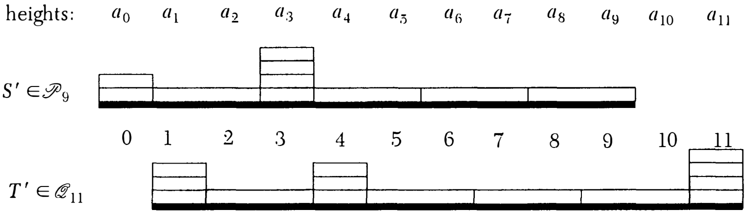

heights: a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11}
$S' \in \mathscr{P}_9$
0 1 2 3 4 5 6 7 8 9 10 11
$T' \in \mathscr{Q}_{11}$

FIGURE 9
The result of swapping tails in Figure 8.

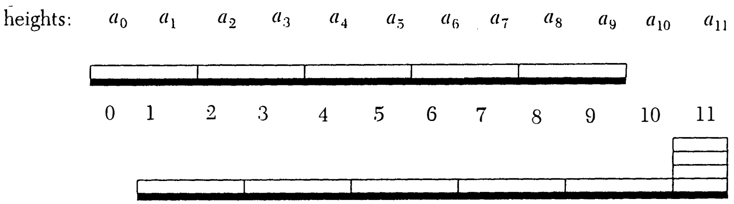

heights: a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11}
0 1 2 3 4 5 6 7 8 9 10 11

FIGURE 10
Problem pairings are fault-free.

elements of $\mathscr{P}_{n-2} \times \mathscr{Q}_n$, consisting of a stack of squares on the nth cell, and dominoes everywhere else (FIGURE 10). Likewise when n is even, there are no fault-free elements of $\mathscr{P}_{n-2} \times \mathscr{Q}_n$, but there are a_n fault-free elements of $\mathscr{P}_n \times \mathscr{Q}_{n-2}$, consisting of a stack of squares on the nth cell, and dominoes everywhere else. Thus we have established $|\mathscr{P}_n \times \mathscr{Q}_{n-2}| - |\mathscr{P}_{n-2} \times \mathscr{Q}_n| = (-1)^n a_n$, as desired.

Using the combinatorially clear fact that $q_n \to \infty$ as $n \to \infty$, the last two identities demonstrate that $(r_0, r_1), (r_2, r_3), (r_4, r_5), \ldots$ is a sequence of nested intervals whose lengths are going to zero. Hence, $\lim_{n \to \infty} r_n$ exists.

Extensions

Next we examine the quantity $K(i,j)$, for $i \leq j$, that counts the number of tilings of the sub-board with cells $i, i+1, \ldots, j$ with height conditions $a_i, a_{i+1}, \ldots, a_j$. For convenience we define $K(i, i-1) = 1$. We see that $K(i,j)$ is the numerator of the continued fraction $[a_i, a_{i+1}, \ldots, a_j]$ and the denominator of the continued fraction $[a_{i-1}, a_i, \ldots, a_j]$. Thus the $K(i,j)$ are identical to the classical *continuants* of Euler [4].

The following theorem, due to Euler, can also be proved by the same tail-swapping technique.

THEOREM 4. *For $i < m < j < n$,*

$$K(i,j)\ K(m,n) - K(i,n)\ K(m,j) = (-1)^{j-m} K(i,m-2) K(j+2,n).$$

This result follows by considering tilings of sub-boards S from cells i to j and T from m to n. Every faulty pair (S,T) corresponds to another faulty pair (S',T') obtained by swapping the tails after the last fault. The term on the right side of Theorems 4 counts the number of fault-free tilings that only occur when the overlapping regions (of S and T, or of S' and T', depending on the parity of $j - m$) consist entirely of dominoes in staggered formation. See FIGURES 11 and 12. Setting $i = 0$ and $m = 1$, Theorem 4 generalizes Theorems 2 and 3 by allowing us to compare arbitrary convergents r_j and r_n.

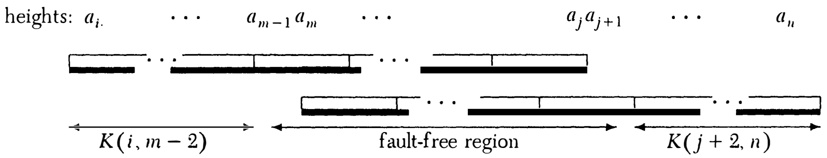

heights: a_i . . . $a_{m-1} a_m$ . . . $a_j a_{j+1}$ . . . a_n
$K(i, m-2)$
fault-free region
$K(j+2, n)$

FIGURE 11
When $j - m$ is even, there are $K(i, m-2)K(j+2, n)$ fault-free tilings (S,T).

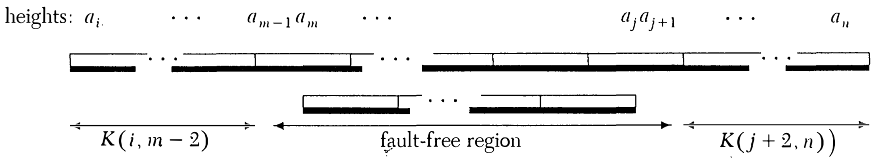


FIGURE 12

When $j - m$ is odd, there are $K(i, m-2)K(j+2, n)$ fault-free tilings (S', T').

Finally, we generalize in a different direction. Suppose we allow dominoes to be stacked as well as squares. Specifically, suppose we impose height conditions $b_1, b_2, \ldots$ so that we may stack as many as b_i dominoes on cells $i-1$ and i. We let $\hat{P}_n$ count the number of ways to tile the board with cells $0, 1, \ldots, n$ and height conditions $a_0, \ldots, a_n$ and $b_1, \ldots, b_n$ for the squares and dominoes respectively. We let $\hat{Q}_n$ count the same problem with cell 0 removed. As before, we see that $\hat{P}_n$ and $\hat{Q}_n$ satisfy

$$\hat{P}_n = a_n\hat{P}_{n-1} + b_n\hat{P}_{n-2}$$

$$\hat{Q}_n = a_n\hat{Q}_{n-1} + b_n\hat{Q}_{n-2}$$

for $n \geq 2$, with initial conditions $\hat{P}_0 = a_0$, $\hat{P}_1 = a_1a_0 + b_1$, $\hat{Q}_0 = 1$, $\hat{Q}_1 = a_1$. But these are precisely the conditions that define the convergents of the expansion

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{\ddots + \cfrac{b_n}{a_n + \ddots}}}}$$

In other words, when the above continued fraction is truncated at the b_n/a_n term, it simplifies to the rational number $\hat{P}_n/\hat{Q}_n$. All of the preceding theorems have generalizations along these lines with similar combinatorial interpretations. We invite the reader to *continue* these investigations.

Acknowledgment. We thank Chris Hanusa for asking an inspiring question.

REFERENCES

1. A. T. Benjamin and J. J. Quinn, Recounting Fibonacci and Lucas identities, *College Math. J.* 30 (1999), 359–366.
2. R. C. Brigham, R. M. Caron, P. Z. Chinn, and R. P. Grimaldi, A tiling scheme for the Fibonacci numbers, *J. Recreational Math.*, Vol. 28, No. 1 (1996–97), 10–16.
3. I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An Introduction to the Theory of Numbers*, John Wiley and Sons, Inc., New York, NY, 1991.
4. O. Perron, *Die Lehre von den Kettenbrüchen*, Chelsea Publishing Co., New York, NY, 1929.
5. S. Vajda, *Fibonacci and Lucas Numbers, and the Golden Section*, John Wiley and Sons, New York, NY, 1989.

Venn Said It Couldn't Be Done

PETER HAMBURGER
RAYMOND E. PIPPERT
Indiana University-Purdue University Fort Wayne
Fort Wayne, IN 46805

Introduction

Margaret E. Baron, in her fascinating essay **[1]** on the historical development of logic diagrams, writes:

> The scope and content of ancient formal logic was determined by Aristotle's *Organon* and, in particular, the Doctrine of the Syllogism which, as it has come to us, contains no diagrams. Nonetheless, so suggestive is the language and manner of presentation of the syllogistic schema, that many logicians have speculated as to the possibility that Aristotle made use of spatial concepts in his actual lectures.... Isolated diagrams using some form of geometric figure to denote a proposition or syllogism occurred in the works of a number of sixteenth-century logicians.... It was, however, Gottfried Wilhelm Leibniz who first devoted serious study to the analysis of logical propositions by means of diagrams.

According to Baron, Leibniz used not only diagrams with three circles or ellipses, wherever it appears more convenient, but three straight lines as well. She also attributes the popularization of circle diagrams to Euler: "Through him, knowledge of the diagrams became widespread and they had some considerable influence in the nineteenth century." For mathematicians like Gergonne and Lambert, claims Baron, the diagrams constituted a vital starting point for an investigation of syllogism. In England, diagrammatic representation was regarded to be of major importance and many books had a section entirely devoted to this topic. Baron writes: "It was, however, John Venn who gave the most detailed consideration to the whole question of diagrams and it is to his credit that he took all possible steps to survey the contributions of his contemporaries ... ."

The inclusion/exclusion arguments of logic and the development of their diagrammatic representations have a long history. They are associated with many famous figures, such as the Greek Aristotle (B.C. 384–322), the Germans Wilhelm Leibniz (1646–1716) and Gottfried Ploucquet (1716–1790), the Swiss Leonhard Euler (1707–1783), the French Joseph Diaz Gergonne (1771–1859) and Johann Heinrich Lambert (1728–1777), the Scottish Sir William Hamilton of Edinburgh (1788–1856), and the British George Boole (1815–1864) and John Venn (1834–1923) **[6, 7]**.

The spatial diagrams are named after Venn, a logicián. Most people have heard of Venn diagrams, and seen drawings with two or three circles, which illustrate the idea in simple cases. The famous three-circle diagram is widely used; its applications range from counting to actuarial science, and from biology to English drama. But fewer people may know that Venn diagrams also have applications in computerized industrial design and automated industrial manufacturing as well.

We will need a few definitions. Formally speaking, a *Venn diagram* consists of n simple closed curves in the plane so that all possible intersections of the interiors and the exteriors of these curves are nonempty and connected. More precisely, an n-*Venn* *diagram* in the plane is a collection of simple closed Jordan curves $\mathcal{F} = \{C_1, C_2, \ldots, C_n\}$ such that each of the 2^n sets $X_1 \cap X_2 \cap \cdots \cap X_n$ is a nonempty and connected region; here, X_i is either the bounded interior or the unbounded exterior of C_i, $i = 1, 2, \ldots, n$. We note that each of the 2^n sets can be described by an n-tuple of zeros and ones where the ith coordinate is 0 if X_i is the unbounded exterior of C_i and 1 otherwise. A Venn diagram is *simple* if at most two curves intersect (transversally) at any point in the plane. Among *non-simple* Venn diagrams, we shall consider only those in which any two curves meet (not necessarily transversally) in points and not in segments of curves. Two Venn diagrams are *isomorphic* if, by continuous transformation of the plane, one of them can be changed into the other or its mirror image. An n-Venn diagram $\mathcal{F}$ is called *irreducible* if each of the n families of $n - 1$ curves, obtained from $\mathcal{F}$ by deleting in turn one of the n curves, fails to be a Venn diagram. Otherwise, it is called *reducible*. The projection of a Venn diagram from the plane to the sphere via stereographic projection yields a *spherical Venn diagram*. Two planar Venn diagrams that can be projected to the same spherical Venn diagram are said to belong to the same *class*.

It is easy to see that there is only one Venn diagram with one curve and only one with two curves. It is also known that there is only one simple Venn diagram on three curves. (Each of these three Venn diagrams can be drawn with a circle or circles.) There are two different planar ones with four curves. But it is also known that there is only one spherical simple Venn diagram with four curves. Furthermore, it easily follows from Euler's theorem ($F + V - E = 2$, where F, V, and E are the numbers of faces, vertices, and edges of a planar graph) that one cannot form simple (or non-simple) Venn diagrams with more than three circles. We show this only for the simple case. Indeed, since two circles can intersect at most twice, we have $V \leq 12$ for four circles. Now each vertex in a simple Venn diagram has degree 4, so the number of edges $E = 2V$, whence $F + V - 2V = 2$ or $F = 2 + V$. Since $F = 16$, we find that $16 \leq 2 + 12$, a contradiction. Similarly, the fact that two ellipses can intersect in at most four points can be used to show that one cannot form Venn diagrams with *more than five ellipses*.

Previous results

The study of the geometrical and topological properties of more complicated planar and spherical Venn diagrams, and their applications in logic set theory, algebra, measure theory, geometry, topology, combinatorics, graph theory, and other areas of mathematics became important in this century, and has produced an impressive journal literature. This study diverged from the diagrammatic representation of syllogistic schema, and has developed into a study of its own. Mathematicians such as the Hungarians Alfréd Rényi (1921–1970), Kató Rényi (1929–1969), and János Surányi, the Poles E. Marczewski and P. Nowicki, and the Americans Branko Grünbaum and Peter Winkler have investigated properties such as convexity of Venn diagrams, the number of k-gons in a Venn diagram on n-curves, and the existence of many different types of Venn diagrams.

One of the most frequently investigated problems of Venn diagrams is the existence of Venn diagrams with five (congruent) ellipses. An erroneous statement of John Venn **[6]** made the problem famous, and this is the subject of this paper. Venn wrote:

> Beyond three terms circles fail us, since we cannot draw a fourth circle which shall intersect three others in the way required. But there is no

theoretic difficulty in carrying out the scheme indefinitely. Of course any closed figure will do as well as a circle, since all that we demand of it, in order that it shall adequately represent the contents of a class, is that it shall have an inside and an outside, so as to indicate what does and what does not belong to the class. There is nothing to prevent us from going on for ever thus drawing successive figures, doubling the consequent number of subdivisions. The only objection is, that since diagrams are primarily meant to assist the eye and the mind by the intuitive nature of their evidence, any excessive complication entirely frustrates their main object.

For four terms the simplest and neatest figure seems to me to be one composed of four equal ellipses thus arranged:

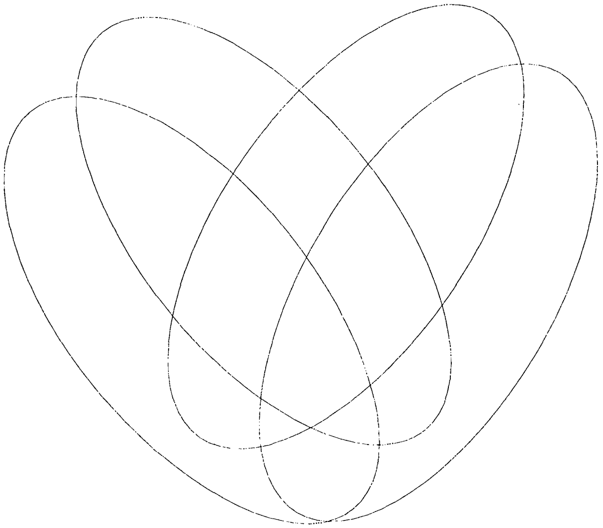


FIGURE 1
Venn's own 4-Venn diagram.

It is obvious that we thus get the sixteen compartments that we want, counting, as usual, the outside of them all as one compartment. The eye can distinguish any one of them in a moment by following the outlines of the various component figures. ... The desired condition that these sixteen alternatives shall be mutually exclusive and collectively exhaustive, so as to represent all the component elements yielded by the four terms taken positively and negatively, is of course secured.

With five terms ellipses fail [emphasis added], at least in the above simple form. It would be quite possible to sketch out figures of a somewhat horse-shoe shape which should answer the purpose—that is, five of which should fulfill the condition of yielding the desired thirty-two distinctive and exhaustive compartments. For all practical purposes, however, any outline which is not very simple and easy to follow with the eye, fails entirely in its main purpose of affording intuitive and sensible illustration. What is wanted is that we should be able to distinguish and identify any assigned compartment in a moment, so as to see how it lies in respect of being inside and outside each of the principal component figures... .

It must be admitted that such a diagram is not quite so simple to draw as one might wish it to be; but then we must remember what are the

> alternatives before any one who wishes to grapple effectively with five terms and all the thirty-two possibilities which they yield. He must either write down or in some way or other have set before him all those thirty-two compounds of which XYZWV is a sample; that is, he must contemplate the array produced by 160 letters. In comparison with most ways of doing that, the sketching out of such a figure is a pleasure, besides being far more expeditious; for, with a very little practice, any of the diagrams here offered might be drawn in but a minute fraction of the time requisite to write down all the letter-compounds. I can only say for myself that, after having for various purposes worked through hundreds of logical examples, I generally resort to diagrams of this description; it not only avoids a deal of unpleasant drudgery, but is also a valuable security against error and oversight. The way in which this last advantage is secured will be best seen presently, when we come to inquire how these diagrams are to be used to represent propositions as distinguished from mere terms or classes.
>
> Beyond five terms it hardly seems as if diagrams offered much substantial help; but then we do not often have occasion to meddle with problems of a purely logical kind which involve such intricacies.

A new result

Branko Grünbaum, one of the most influential geometers of this century and a 1975 recipient of the Lester R. Ford award for an article related to the present topic, wrote as follows [**3**]:

> In [**6**] Venn gave examples of Venn diagrams with four ellipses. However, he mistakenly stated that no five ellipses can form a Venn diagram; indeed, it takes only a little patience to verify that the five congruent ellipses in each part of [Figure 2] form a Venn diagram.
>
> Venn's erroneous assertion was repeated—unchecked and unchallenged—by several authors[1] for almost a century.
>
> The first Venn diagram of five ellipses ... was published only in 1975 [**2**]; a non-simple example ... was found by Schwenk [**5**]. [See Figure 3]
>
> Using Euler's theorem and the fact that two ellipses can intersect in no more than four points, it follows easily (by an argument similar to the one concerning circles) that there can be no Venn diagrams with six or more ellipses. One possible explanation for Venn's error is that he may have believed that all Venn diagrams can be constructed following a sort of "greedy algorithm"...: to get a diagram with n curves first make a diagram with $n - 1$ curves and then add the last one. However, it is easy to verify that none of the Venn diagrams in [Figures 2 and 3] (*nor any other simple Venn diagram of five ellipses) can be obtained by adding a fifth ellipse to a Venn diagram of four ellipses* [emphasis added]. Probably the same is true without the assumption of simplicity.

[1] Among others, in the article *Logic Diagrams* in P. Edwards (editor), 1967, *Encyclopedia of Philosophy*, Macmillan, New York, NY, 1967, *Logic Diagrams*, by M. Gardner, Vol 5. pp. 77–81.

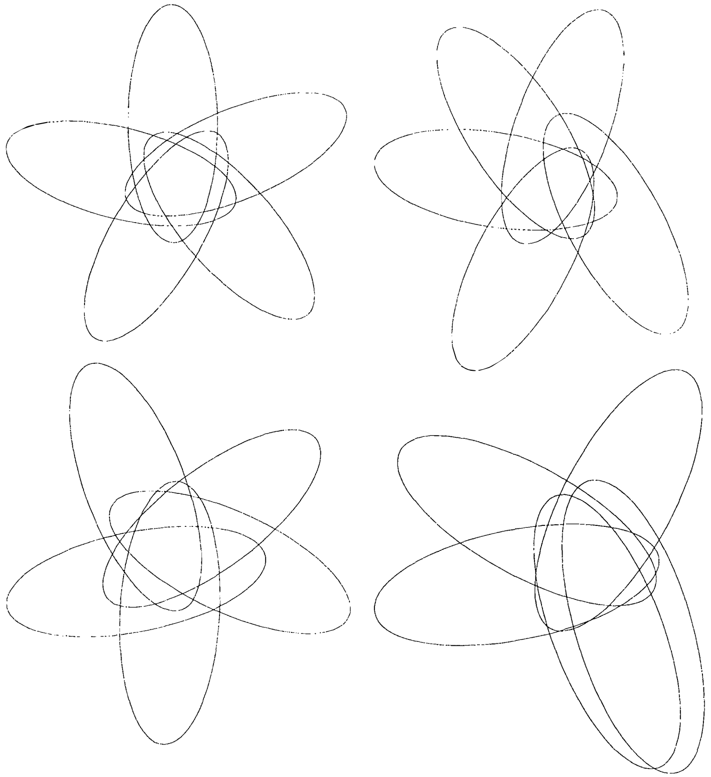


FIGURE 2
Grünbaum's four 5-Venn diagrams.

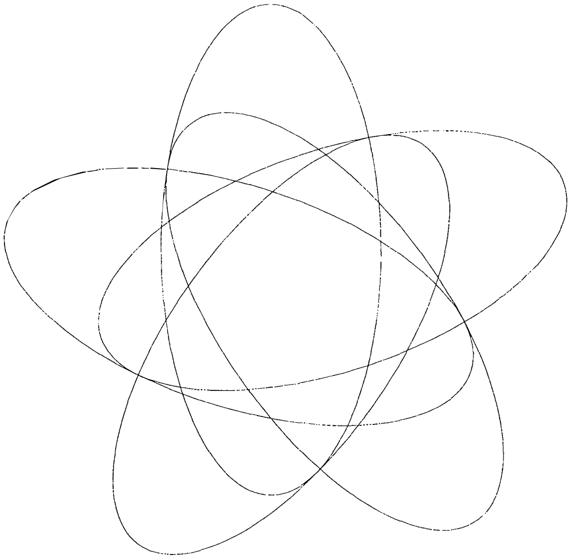


FIGURE 3
Schwenk's non-simple 5-Venn diagram.

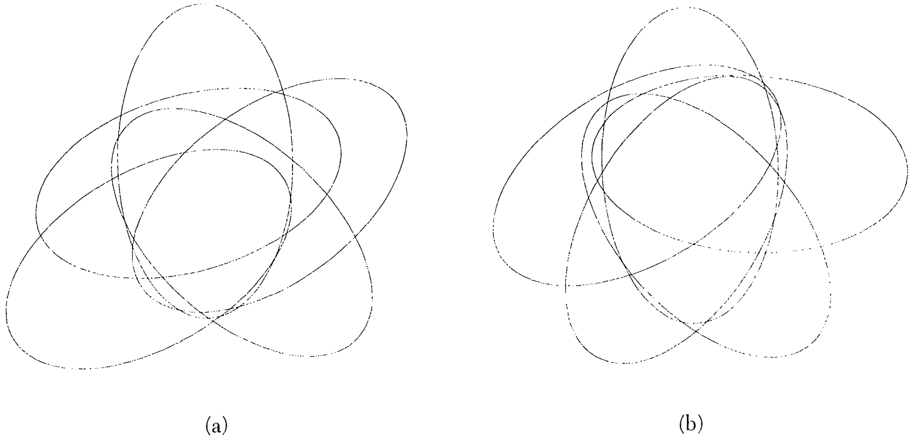


FIGURE 4
Two reducible, simple 5-Venn diagrams with five congruent ellipses. (The deletion of the (unique) vertical ellipse yields a Venn diagram of four ellipses.)

Using graph theory, the authors, together with Kiran B. Chilakamarri, developed methods to analyze and construct new Venn diagrams. One of the results is the objective of this note—the Venn diagram that, more than a century ago, Venn erroneously said couldn't be drawn. The Venn diagrams in FIGURE 4 are *simple, reducible Venn diagrams with five congruent ellipses*. (This is very special in one sense, since each ellipse has the same size. With ellipses of different sizes one can obtain diagrams in which no region is very small.) The two Venn diagrams belong to the same class, and these are the only simple, reducible Venn diagrams with five ellipses. The lengthy and technical proof of this statement appeared in **[4]**.

An important part of the creation of mathematical research is the formulation of good questions which lead to large quantities of related results, which in turn fuel the engine of progress in mathematics. Grünbaum's problems and conjectures (some true and some false) have been the inspiration for all of our work on Venn diagrams. We are very grateful to Professor Grünbaum for his encouragement and helpful comments.

REFERENCES

1. M. E. Baron, A note on the historical development of logic diagrams: Leibniz, Euler and Venn, *Mathematical Gazette* 53 (1969), 113–125.
2. B. Grünbaum, Venn diagrams and independent families of sets, this MAGAZINE 1 (1975), 12–23.
3. B. Grünbaum, Venn diagrams I, *Geombinatorics* 1 (1992), 5–12.
4. P. Hamburger and R. E. Pippert, Simple, reducible Venn diagrams on five curves and Hamiltonian cycles, *Geometriae Dedicata* 67 (1997), 1–20.
5. A. J. Schwenk, Venn diagram for five curves, this MAGAZINE 57 (1984), 297.
6. J. Venn, On the diagrammatic and mechanical representation of propositions and reasonings, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 9 (1880), 1–18.
7. J. Venn, *Symbolic Logic*, Macmillan, London, UK, 1881, second edition, 1894.

Superexponentiation and Fixed Points of Exponential and Logarithmic Functions

STEPHEN R. WASSELL
Sweet Briar College
Sweet Briar, VA 24595

1. Introduction

We shall investigate an application of "superexponentiation," an operation that we denote by $\Uparrow$ (following [5]) and define as follows:

$$b \Uparrow n := b\,^\wedge(b\,^\wedge(\cdots{}^\wedge b)\cdots) \quad (b > 0,\, n = 1, 2, \ldots), \tag{1}$$

where exponentiation occurs n times. Superexponentiation simply continues the pattern of addition, multiplication, and exponentiation. It seems to have first appeared in the literature in [4], for the purpose of exhibiting extremely large (albeit finite) numbers (its *implicit* use in [7], an earlier work than [4], is shown in [5]). Superexponentiation is used in [5] and [6] to examine the logical foundation of mathematical induction. In [2] superexponentiation and its inverse operation, iteration of logarithms, are used to analyze the running time of certain algorithms. A general discussion of superexponentiation is given in [1].

For our purposes, superexponentiation naturally arises from analyzing fixed points of exponential functions (and hence of the corresponding logarithms). We approach this using orbit analysis as in [3], i.e., by iterating the exponential function $F(x) = b^x$, starting with input $x = b$. For $b > 1$ the resulting orbit clearly will be strictly increasing, and for bases such as 2 or 10 (those used in the applications cited above), the orbit will diverge rapidly to infinity. Will this divergence occur for all $b > 1$? What kind of behavior is observed if $0 < b < 1$? Will the orbit converge to a single number for suitable values of b?

In Section 2 we discuss the context under which the author arrived at the problem. In Section 3 we answer the question of convergence, finding the set of b for which the orbit does indeed converge; this result is stated as a theorem at the end of the section. In Section 4 we offer several suggestions for further exploration.

2. Motivation

When introducing exponential and logarithmic functions,

$$F(x) = b^x \quad \text{and} \quad G(x) = \log_b x \quad (b > 0, b \neq 1),$$

the instructor will inevitably display the graphical consequence of the fact that $F(x)$ and $G(x)$ are inverses. FIGURE 1 shows the case $b = \mathrm{e}$, surely the most popular case to present, given the importance of the functions e^x and $\ln x$.

For other bases, however, the graphical situation may not be as familiar. FIGURE 2 shows the case $b = \{1/2\}$. For $0 < b < 1$, b^x exhibits exponential decay rather than exponential growth. While this often may be presented, the corresponding logarithm

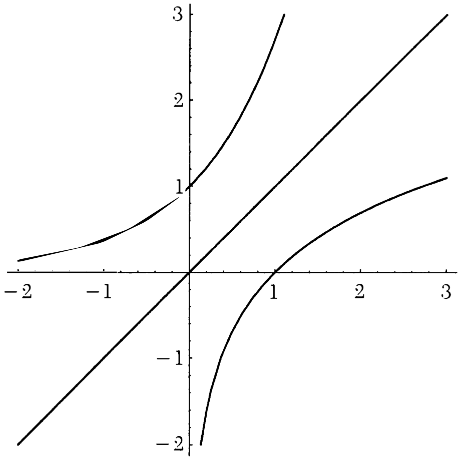


FIGURE 1
The familiar graphs of e^x, $\ln x$, and x.

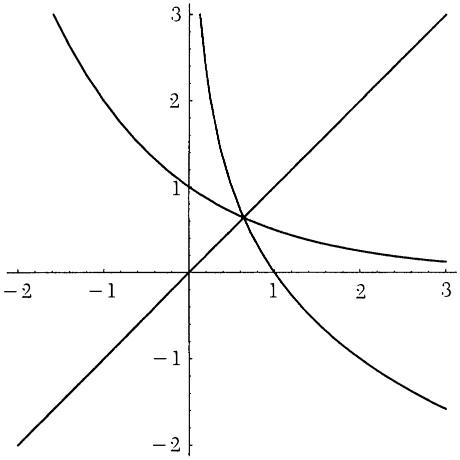


FIGURE 2
The graphs of $(1/2)^x$, $\log_{1/2} x$, and x.

may often be omitted (after all, the three most prevalent logarithmic bases, $b = 2$, $b = e$ and $b = 10$, all satisfy $b > 1$). Consider, however, that in Figure 2, unlike in Figure 1, the graphs of the exponential and logarithmic functions intersect (along the line $y = x$). Thus, there is some real number x_0, clearly between 0 and 1, for which

$$(1/2)^{x_0} = x_0 = \log_{1/2} x_0 .$$

In other words, x_0 is simultaneously a fixed point of the functions $f(x) = (1/2)^x$ and $g(x) = \log_{1/2} x$. How can we find the value of x_0? Trying to solve $(1/2)^x = x$ using the inverse function results in $x = \log_{1/2} x$ (and vice versa, of course), and the equation $(1/2)^x = \log_{1/2} x$ does not seem any more promising.

Let us try a recursive approach. Take the given equation, $x = 0.5^x$, and substitute it into itself:

$$x = 0.5^{(0.5^x)}.$$

(Note that, as indicated by the parentheses, this cannot be simplified using the familiar property $(b^m)^n = b^{mn}$. Unlike addition and multiplication, exponentiation is nonassociative; cf. **[1]**) Repeating this process ad infinitum leads to a formal expression for the solution:

$$x_0 = 0.5^{(0.5^{(\cdot^{\cdot^{\cdot}})})}. \tag{2}$$

It remains to determine whether or not (2) converges. Numerical calculation suggests oscillatory convergence, as evidenced by the orbit of 0.5 under iteration of $f(x) = 0.5^x$ (rounded to three significant figures):

$$0.500, 0.707, 0.613, 0.654, 0.635, 0.643, 0.640, 0.641, \ldots.$$

The oscillation and convergence of this orbit can also be seen graphically in FIGURE 3, which stems from FIGURE 2. At least we can thus approximate the value of the fixed point x_0 to any desired degree of accuracy. In the next section, we shall generalize the above discussion to other bases and explore the behavior of the resulting family of functions.

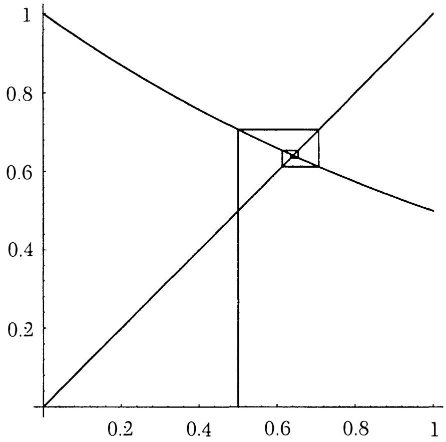


FIGURE 3
The graphs of $(1/2)^x$ and x, as well as the orbit of $1/2$ under iteration of $f(x) = (1/2)^x$, approaching the limit of (2).

3 Generalization

Given a base $b > 0$, we wish to determine whether

$$b \Uparrow \infty := \lim_{n \to \infty} b \Uparrow n \tag{3}$$

exists as a real number, or is infinite, or does not exist. In case the limit exists as a number, which we denote by x, it follows that x is a solution to

$$b^x = x, \text{ or equivalently, } x = \log_b x. \tag{4}$$

We must first ask, for what $b > 0$ is it even possible to solve (4)? The case $b = 1$ is trivial: we can easily solve $1^x = x$ (but note that $\log_b x$ is not defined for $b = 1$). For $b \in (0,1)$ the graphs of b^x, x, and $\log_b x$ will intersect, as in FIGURE 2, and so it is possible to solve (4). What of the case $b > 1$? For sufficiently large b (e.g., $b = \mathrm{e}$) the graphs of b^x, x, and $\log_b x$ do not intersect, and so (4) is impossible. Since, however, b^x does intersect the line $y = x$ when $0 < b \leq 1$, (4) should be possible when $b > 1$ is sufficiently close to 1 as well.

We explore this by focusing on the exponential function (results for the logarithmic function would follow immediately). To this end, consider the one-parameter family of functions

$$F_b(x) = b^x \quad (b > 0), \tag{5}$$

parameterized by the base b. (Unlike with $F(x)$, we now explicitly indicate that the base b is a parameter.) What happens to the graph of $F_b(x)$ as $b > 0$ varies? We show five snapshots of $F_b(x)$ plotted in the same coordinate plane in Figure 4. Note that the family pivots about the point $(0, 1)$ since $b^0 = 1$ for all $b > 0$. For b sufficiently close to 0, $F_b'(0)$ is a negative number of large magnitude. While $F_b'(0)$ stays negative for all $0 < b < 1$, its magnitude lessens as $b \to 1$. When $b = 1$, $F_b'(0) = 0$ (and the graph is simply a horizontal line). When b is just slightly larger than 1, then, $F_b'(0)$ is just slightly positive, and we see that this allows for *two* intersection points of $F_b(x)$ with x, i.e., two solutions of (4).

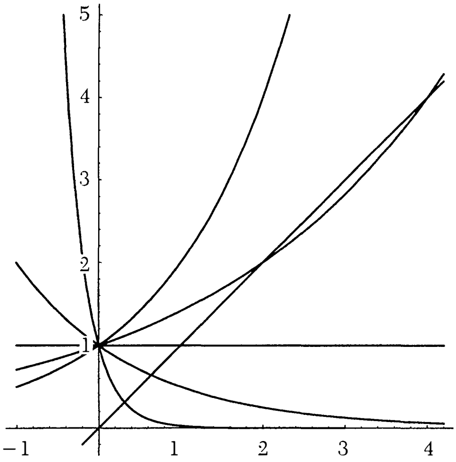


FIGURE 4

The graphs of $F_b(x) = b^x$, for $b = 1/40, 1/2, 1, \sqrt{2}$, and 2, along with the line $y = x$.

As $b > 1$ becomes larger, the value of $F_b'(0)$ becomes larger in magnitude, and eventually the graphs of $F_b(x)$ and x no longer intersect, so that (4) is no longer possible to solve. We can determine the unique value of $b > 1$ for which the graphs of $F_b(x)$ and x intersect at just one point by solving the system of two equations,

$$\{F_b(x) = x,\ F_b'(x) = 1\}, \tag{6}$$

in the two unknowns b and x; see Figure 5. These two equations are simply

$$\{b^x = x,\ b^x \ln b = 1\}, \tag{7}$$

from which it immediately follows that

$$1 = b^x \ln b = x \ln b = \ln(b^x) = \ln x,$$

providing the solution to (6): $x = \mathrm{e}$ and $b = \mathrm{e}^{1/\mathrm{e}}$, the latter of which may be called the e^{th} root of e!

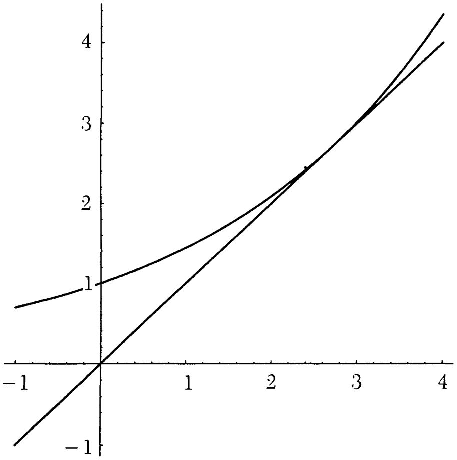


FIGURE 5

The graph of $F_{e^{1/e}}(x) = e^{x/e}$ tangentially intersecting the line $y = x$ at the point (e, e).

Therefore, we seek to solve (4) only for $0 < b < e^{1/e}$. Furthermore, while there may be other ways to solve this equation (see items i. and ii. in Section 4), we wish to use the iterative process that was outlined in Section 2 for the case $b = 1/2$. To do so, we employ the methods given in **[3]** (see especially Chapters 3–6, 12), which introduces the mathematics of chaos using a dynamical systems approach. This involves analyzing the evolution of fixed points of a family of functions as the parameter of the family varies. In fact, our present query is framed within the same context: our family is defined in (5), with parameter b. For a given $b \in (0, e^{1/e})$, we are attempting to find the fixed point of $F_b(x)$ by analyzing the orbit of the initial input (seed), b, under iteration of F_b. (While this seed is a natural choice for us, it is wise to choose seeds more deliberately in general; see Chapter 16 of **[3]**.)

We use the methods of **[3]** in two cases: $1 < b < e^{1/e}$ and $0 < b < 1$. For $b \in (1, e^{1/e})$, the graph of $F_b(x)$, being qualitatively the same as that of $F_{\sqrt{2}}(x)$ (see FIGURE 6), has two fixed points. At the lower fixed point $|F_b'(x)| < 1$, while at the

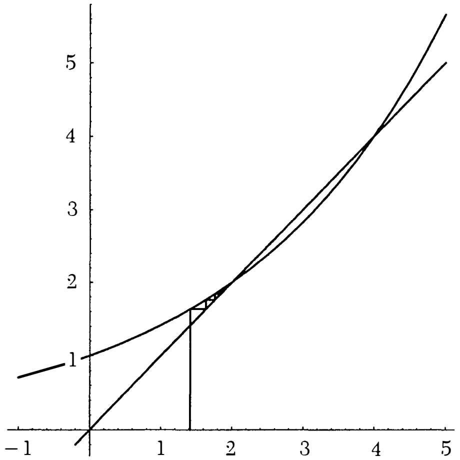


FIGURE 6

The graphs of $F_{\sqrt{2}}(x) = (\sqrt{2})^x$ and x, as well as the orbit of the seed $\sqrt{2}$ under iteration of $F_{\sqrt{2}}$, approaching the lower fixed point of $F_{\sqrt{2}}$, which is $x = 2$.

upper fixed point, $|F_b'(x)| > 1$ (this is clear since the slope of $y = x$ itself is 1). Therefore, the lower fixed point is attracting and the upper fixed point is repelling. Orbit analysis readily reveals that for any seed chosen less than the value of the upper fixed point, its orbit will be attracted to the lower fixed point. For any seed greater than the upper fixed point, however, its orbit will diverge to ∞. Luckily, our seed b is in the basin of attraction of the lower fixed point!

We can conclude, therefore, that for $b \in (1, e^{1/e})$, the limit $b \Uparrow \infty$ exists and yields the value of one of the two solutions to (4), namely the lower fixed point. Before we move on to the second case, consider the endpoints of the first case: $1 \Uparrow \infty = 1$ is trivial, and we have seen $b = e^{1/e}$ in Figure 5. In fact, we know that $F'_{e^{1/e}}(e) = 1$, so that a saddle-node bifurcation should occur at the parameter value $b = e^{1/e}$. Indeed, F_b has no fixed points for $b > e^{1/e}$, one fixed point for $b = e^{1/e}$, and two fixed points (one attracting and one repelling) for $1 < b < e^{1/e}$. We shall revisit this upper endpoint below.

To analyze the case $0 < b < 1$, let us get better acquainted with the family of functions (5) for such values of b; Figure 7 shows five members. Clearly, for $b \in (0, 1)$, the function F_b will have exactly one fixed point, which will necessarily be in the interval $0 < x < 1$. This fixed point will be attracting or repelling depending on whether the derivative F_b', evaluated at the fixed point, will be less than or greater than 1 in magnitude. Moreover, a period-doubling bifurcation will occur at the value of b for which the derivative is equal to -1. We can find this particular parameter value and corresponding fixed point by solving the system (6) again, except that we replace the 1 on the right side of the second equation with -1. Solving as before, it is straightforward to show that the solution is $b = (1/e)^e$ and $x = 1/e$. (The critical parameter values $b = e^{1/e}$ and $b = (1/e)^e$ were found only empirically in **[1]**.)

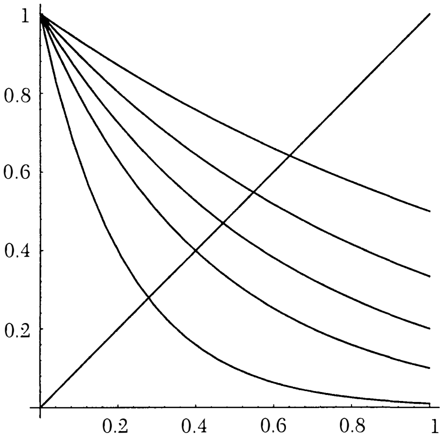
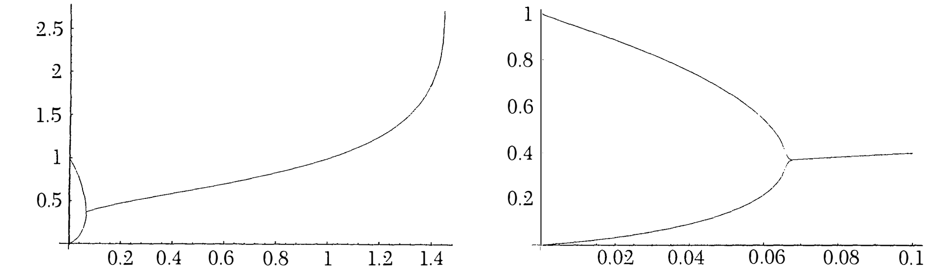


FIGURE 7

The graphs of $F_b(x) = b^x$, for $b = 1/40$, $1/10$, $1/5$, $1/3$, and $1/2$, along with the line $y = x$.

To better understand the saddle-node bifurcation at $b = e^{1/e} \approx 1.44$ and the period-doubling bifurcation at $b = (1/e)^e \approx .0660$, we plot two versions of a bifurcation diagram in Figure 8. Let us start in the upper right corner of the top diagram, at the point $(e^{1/e}, e)$. This point on the bifurcation diagram signifies that for the base value $b = e^{1/e}$, the corresponding function from the family, $F_{e^{1/e}}$, has a fixed point $x = e$. To the right of this point, i.e., for larger values of b, there are no fixed points. When b becomes smaller than $e^{1/e}$ (but larger than 1), we know that there are, in fact, two fixed points of F_b. The upper fixed point is repelling, however, and so the



(a) (b)

FIGURE 8

Two bifurcation diagrams for $F_b(x) = b^x$, the first for $0 < b < e^{1/e}$ and the second for $0 < b < 0.1$.

upper fixed point branch (which would trace a curve upwards and to the left of $(e^{1/e}, e)$) does not appear in this bifurcation diagram. Only the lower (attracting) fixed point branch appears.

Although it is not apparent in the bifurcation diagram, something special occurs at $b = 1$, for there is suddenly only one fixed point of F_b here. This would be seen in the repelling fixed point branch, if it were present. It would increase without bound, asymptotically approaching the line $b = 1$, since as b decreases to 1, the upper (repelling) fixed point increases to ∞. On the other hand, since the attracting fixed points change continuously as b decreases through 1, the attracting fixed point branch is uneventful at $(1, 1)$.

The single fixed point branch of the bifurcation diagram continues down to the point $((1/e)^e, 1/e)$, where the period-doubling bifurcation occurs (see the second graph in FIGURE 8). Indeed, as b decreases through $(1/e)^e$, the attracting fixed point branch bifurcates into an attracting 2-cycle branch. The remaining fixed points are repelling, and hence this fixed point branch for $b \in (0, (1/e)^e)$ does not appear in the diagram. For completeness we show, in FIGURE 9, the orbit analysis for a base less than $(1/e)^e$, namely $b = 1/40$. Of course, in this case the spiral approaches a 2-cycle rather than a fixed point (see FIGURE 3).

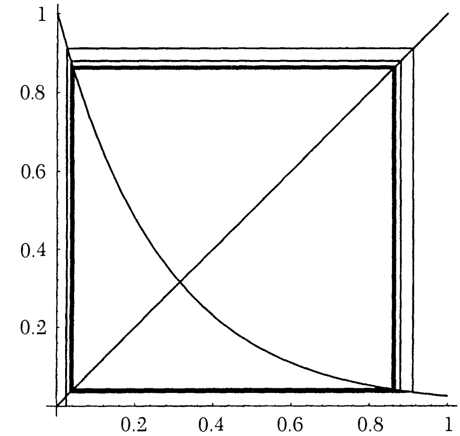


FIGURE 9

The graphs of $F_{1/40}(x) = (1/40)^x$ and x, as well as the orbit of the seed $1/40$ under iteration of $F_{1/40}$, approaching a 2-cycle.

Referring to Chapter 16 of **[3]**, we can show that for any $b < \mathrm{e}^{1/\mathrm{e}}$, b is in the basin of attraction of the attractor of $F_b(x)$, which is simply the lower fixed point branch together with the 2-cycle branch (i.e., the first graph of FIGURE 8). Indeed, for $b < \mathrm{e}^{1/\mathrm{e}}$,

$$F_b'(b) = (\ln b)b^b < (\ln \mathrm{e}^{1/\mathrm{e}})b^b = (1/\mathrm{e})b^b < 1,$$

provided that $b^{\mathrm{b}} < \mathrm{e}$, or equivalently, $b \ln b < 1$. However, this also follows from the fact that $b < \mathrm{e}^{1/\mathrm{e}}$:

$$b \ln b < \mathrm{e}^{1/\mathrm{e}} \ln \mathrm{e}^{1/\mathrm{e}} = \mathrm{e}^{(1/\mathrm{e})-1} < 1.$$

We can now state our main theorem:

THEOREM. *The limit $b \Uparrow \infty$, defined in* (1) *and* (3), *exists, and hence provides a solution to $b^x = x$, if and only if $b \in [(1/\mathrm{e})^{\mathrm{e}}, \mathrm{e}^{1/\mathrm{e}}]$. For $b \in [(1/\mathrm{e})^{\mathrm{e}}, 1]$, moreover, $b \Uparrow \infty$ is the unique solution to $b^x = x$.*

It remains only to verify the convergence of $b \Uparrow \infty$ when $b = (1/\mathrm{e})^{\mathrm{e}}$ and when $b = \mathrm{e}^{1/\mathrm{e}}$. At both of these endpoints, the fixed points are neutral, since $|F_{(1/\mathrm{e})^{\mathrm{e}}}'(1/\mathrm{e})| = |F_{\mathrm{e}^{1/\mathrm{e}}}'(\mathrm{e})| = 1$. Through orbit analysis, it is straightforward to show that the fixed point of $F_{(1/\mathrm{e})^{\mathrm{e}}}(x)$, namely $1/\mathrm{e}$, is weakly attracting. Essentially, this describes the fact that $((1/\mathrm{e})^{\mathrm{e}}) \Uparrow \infty$ exists, but that the convergence of the orbit to $1/\mathrm{e}$ is extremely slow. Similarly, the fixed point of $F_{\mathrm{e}^{1/\mathrm{e}}}(x)$, namely e, weakly attracts seeds less than e but weakly repels greater seeds. Since the seed we use in this case, namely the base $\mathrm{e}^{1/\mathrm{e}}$, is less than the fixed point e (lucky again!), we have $(\mathrm{e}^{1/\mathrm{e}}) \Uparrow \infty = \mathrm{e}$.

4. Exploration

We propose six avenues for further investigation:

i. Equation (4) has at least one solution for all $b \in (0, \mathrm{e}^{1/\mathrm{e}}]$ (see FIGURES 4 and 5). As previously stated the solutions to (4) for $0 < b < (1/\mathrm{e})^{\mathrm{e}}$, as well as the upper fixed points for $1 < b < \mathrm{e}^{1/\mathrm{e}}$, are precisely the repelling fixed points missing from the bifurcation diagrams of FIGURE 8. How can the repelling fixed points be found? What is the behavior of the fixed point as $b \to 0$, i.e., what does the repelling fixed point branch for $0 < b < (1/\mathrm{e})^{\mathrm{e}}$ look like? A nice graphical project is to plot the repelling fixed point branches where they belong in the bifurcation diagram.

ii. Write $b = \mathrm{e}^a$, so that the equation $b^x = x$ becomes $\mathrm{e}^{ax} = x$. Solve this last equation for a as a function of x. Show that this function is strictly increasing for $0 < x < \mathrm{e}$ with image $-\infty < a < 1/\mathrm{e}$, and strictly decreasing for $x > \mathrm{e}$ with image $1/\mathrm{e} > a > 0$. Conclude that for each pair (x, a) the exponential function with base $b = \mathrm{e}^a$ has a fixed point x. Note that for each $a \in (0, 1/\mathrm{e})$, i.e., for each $b \in (1, \mathrm{e}^{1/\mathrm{e}})$, there are two pairs (x_1, a) and (x_2, a), which correspond to the lower and upper fixed points! (This method gives us a closed form for the base given a desired fixed point, but we still do not have a closed form for the fixed point(s) given a desired base.)

iii. Provide a classical proof of convergence for $b \in [(1/\mathrm{e})^{\mathrm{e}}, \mathrm{e}^{1/\mathrm{e}}]$ by considering the sequence $\{a_n\}_{n=1}^{\infty}$ where $a_n = b \Uparrow n$. For $1 \leq b \leq \mathrm{e}^{1/\mathrm{e}}$, the proof is straightforward since the sequence is clearly monotone and easily bounded. For $(1/\mathrm{e})^{\mathrm{e}} \leq b < 1$, however, the situation is a bit more complicated since the sequence is not monotone. Considering $r \leq b < 1$ with $r \geq (1/\mathrm{e})^{\mathrm{e}}$, the difficulty of proof increases as r decreases to $(1/\mathrm{e})^{\mathrm{e}}$!

iv. For $x > 0$, define a sequence of functions by $f_n(x) = x \Uparrow n$, so that $f_1(x) = x$, $f_2(x) = x^x, f_3(x) = x^{(x^x)}$, etc. Using a graphing calculator or a computer, plot $f_n(x)$ for several values of n. How does the $\lim_{x \to 0^+} f_n(x)$ depend on n? Why? Does the sequence of functions $\{f_n(x)\}_{n=1}^{\infty}$ seem to converge pointwise to a limit function $f(x)$, at least for some interval of values for x? How does this relate to FIGURE 8?

v. The reader may have seen a teaser where one *begins* with superexponentiation: for example, solve

$$b^{b^{\cdot^{\cdot^{\cdot}}}} = 2, \tag{8}$$

i.e., $b \Uparrow \infty = 2$. Assuming the convergence of $b \Uparrow \infty$, show that $b = \sqrt{2}$ (cf. FIGURE 6). Try solving $b \Uparrow \infty = 4$ in the same way; what is amiss? How big can the right hand side of (8) be for the assumption of convergence of $b \Uparrow \infty$ to be valid? On the other hand, what information is obtained upon solving (8) if the right hand side is "too big"?!

vi. Consider the implications of the tautology $(x^{1/x})^x = x$ $(x > 0)$, and relate your findings to item ii.

Acknowledgment. The author wishes to thank Dr. Loren Pitt, Dr. Douglas Szajda, and the referees for several helpful suggestions.

REFERENCES

1. Nick Bromer, Superexponentiation, this MAGAZINE 60 (1987), 169–173.
2. Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, *Introduction to Algorithms*, MIT Pr., Cambridge, MA, 1990.
3. Robert L. Devaney, *A First Course in Chaotic Dynamical Systems: Theory and Practice*, Addison-Wesley, Reading, MA, 1992.
4. Donald E. Knuth, Mathematics and computer science: coping with finiteness, *Science* 194 (1976), 1235–1242.
5. Edward Nelson, *Predicative Arithmetic*, Mathematical Notes 32, Princeton Univ. Pr., Princeton, NJ, 1986.
6. Edward Nelson, Taking formalism seriously, *Math. Intelligencer* 15:3 (1993), 8–11.
7. Joseph R. Schoenfield, *Mathematical Logic*, Addison-Wesley, Reading, MA, 1967.

When is the Product of Two Oblong Numbers Another Oblong?

TRYGVE BREITEIG
Agder College
N–4604 Kristiansand
Norway

For centuries, mathematicians have been fascinated by the association of numbers with geometric patterns, such as squares or triangles. A less-studied example of this genre—oblong numbers—is the occasion for a nice student investigation: find pairs of oblong numbers whose product is also an oblong number. This activity has historical resonance, and it involves the gathering of data and the search for patterns, the formulation of general results, and the use of Pell's equation.

Oblong numbers are numbers of the form $a(a+1)$, where a is a positive integer. The first few oblong numbers are 2, 6, 12, 20, and 30. The name refers to the geometric form by which these numbers may be represented.

1. Historical perspectives

Greek mathematicians found the relations between numbers and geometrical form of great interest; they studied polygonal numbers of different shapes. A well-known source is the arithmetic book *Introductio Arithmeticae* by Nicomachus (ca. 100 AD). Number patterns and their geometrical representations in general, and square, oblong, and triangular numbers in particular, play an important role in Nicomachus' work.

This interest extended beyond the Greeks. Anicius Boethius (ca. 475–524) holds a special position among those who passed on the philosophical heritage of the classical age to medieval Europe. The bridge Boethius built from Greek culture and mathematics to the European culture of the middle ages is of historical and philosophical significance. His translations and revisions of Greek books, on numbers, astronomy, music, and logic, are parts of the bridge. So is his textbook on numbers, *De Institutione Arithmetica*—even if it is, mathematically speaking, an uncritical paraphrase (some historians would say a translation) of Nicomachus' *Introductio Arithmeticae* from Greek to Latin. According to Burton [**2**, p. 241] this book held an authoritative position for about a thousand years. It does not contain algebraic symbolism.

Boethius points out many number relationships and relates them to his philosophical view of a pattern-governed universe. The following examples may show patterns in the world of numbers:

(i) An oblong number is the sum of consecutive positive even numbers, starting from 2: $2+4+6+\cdots+2n$. A square is similarly the sum of consecutive odd numbers, starting from 1: $1+3+5+\cdots+(2n-1)$.

(ii) An oblong number is twice a triangular number.

(iii) The sum of two consecutive triangular numbers is a square.

(iv) The sum of two consecutive squares added to the square of the oblong number between them is a new square number.

(v) The sum of two consecutive oblong numbers added to twice the square between them is a square.

(vi) An oblong number added to the next square is a triangular number.
(vii) A square plus the next oblong number is a triangular number.
(viii) A number plus the square of that number is an oblong number.

How would we prove these relations? Most of them can be modeled by geometric diagrams. The relations in (i) may be represented as in FIGURE 1; another approach appears in this MAGAZINE [3] as a Proof without Words.

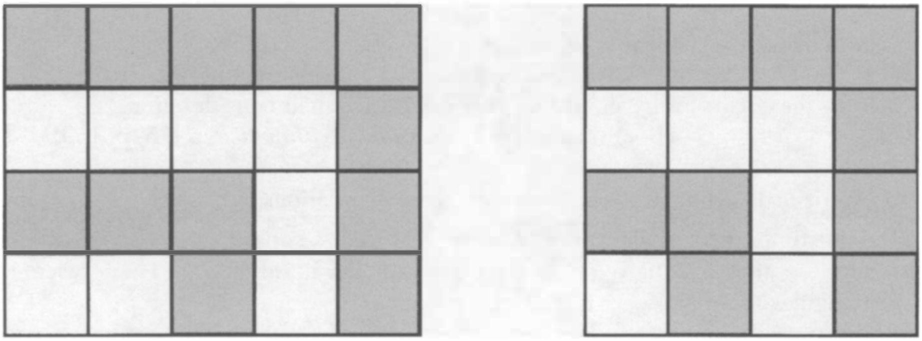


FIGURE 1
Pattern in oblongs and squares. Geometric proof that $\sum_{i=1}^{n} 2i = n(n+1)$ and that $\sum_{i=1}^{n}(2i-1) = n^2$.

Relations (ii) and (iii) are also easily illustrated on well-known, simple figures.

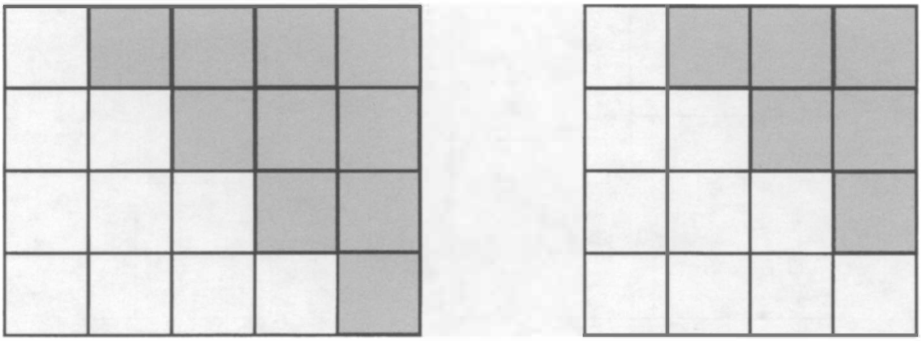


FIGURE 2
Number identities proved geometrically.

All the relationships may be expressed in a symbolic form. The symbolic language of algebra is most useful to express them precisely and prove them. For example, (iv) says

$$n^2 + (n+1)^2 + [n(n+1)]^2 = (n^2+n+1)^2.$$

Properties (v), (vi), (vii), and (viii) also have obvious algebraic representations: Let $T_n = \frac{n(n+1)}{2}$ denote the nth triangular number. Then (v), (vi), (vii), and (viii) can all be written as algebraic identities:

(v) $(n-1)n + 2n^2 + n(n+1) = (2n)^2$ (vii) $n^2 + n(n+1) = T_{2n}$

(vi) $(n-1)n + n^2 = T_{2n-1}$ (viii) $n^2 + n = n(n+1)$

Greek mathematicians did not use our algebraic symbolism. The number relationships they found, like the previous examples, were verified by numerical examples and geometrical proofs.

Numbers play a substantial role in Boethius' philosophy. Squares and oblong numbers are basic:

> ...all things consist of the same nature and then of the nature of another, and ... this can first be seen in numbers. ...
>
> From squares and from figures longer by one side [= oblong numbers] the idea of every form takes its being.
>
> The fact that the entire development of all forms may be seen to arise from these two forms should be noted with no small consideration.
>
> (Boethius, in *De Institutione Arithmetica*; see **5**, p. 159])

In the spirit of Boethius we add two more number relations:

(ix) Eight times a triangular number is one less than a square.

(x) Suppose that a square n^2 is also a triangular number T_m. Then twice the triangular number T_{m-n} is also triangular.

These relations are expressed algebraically as follows:

$$\text{(ix)}\ 8\frac{n(n+1)}{2}+1=(2n+1)^2 \qquad \text{(x)}\ n^2=T_m \Leftrightarrow T_{2n-m-1}=2T_{m-n}$$

FIGURE 3 illustrates (ix) and (x); note the use of a geometric proof.

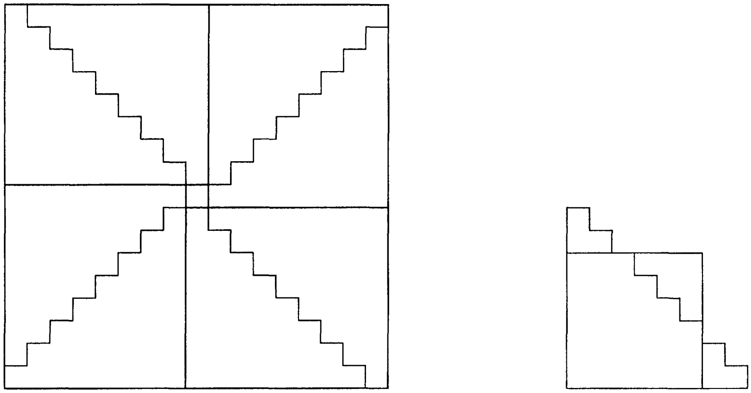


FIGURE 3

Geometric proof of number relations (ix) and (x). The right figure shows the equivalence between $6^2 = T_8$ and $T_3 = 2T_2$.

2. Investigating number patterns

The product of two square numbers is another square. This suggests an analogous question for oblong numbers:

When is the product of two oblong numbers an oblong number?

Obviously, the product of two *consecutive* oblong numbers is another oblong:

$$(x-1)x\cdot x(x+1)=(x^2-1)x^2.$$

To solve the general problem, however, we have to solve the Diophantine equation

$$x(x+1)y(y+1)=z(z+1), \tag{1}$$

a fourth-degree equation in three unknowns. The multiplication seems to make it hard to model by a diagram. Since we aim not to prove a general property of figurate numbers but to find numbers satisfying a given condition, the situation seems to call for a search and some pattern recognition.

Generating numbers We will find numerical solutions, search for patterns, and then move between numerical and algebraic modes of thinking.

A computer program can generate triples (x, y, z) satisfying (1). Since (1) is symmetric in x and y, we may look for triples where $x < y$; we will list results in order of y. (We omit in the program the trivial triples $(0, k, 0)$ and also the solution just mentioned, $(k, k+1, k^2+2k)$.) Following is a simple BASIC program to generate the list:

```
10  FOR Y=2 TO 2000
20   FOR X=1 TO Y-2
30    U=X*(X+1)*Y*(Y+1)
40    Z=INT(SQR(U))
50    IF Z*(Z+1)=U THEN PRINT X,Y,Z
60   NEXT X
70  NEXT Y
```

The numerical results invite closer investigation:

TABLE 1 Integer solutions (x, y, z) of equation (1), where $x < y$

x	y	z	x	y	z
1	2	3	11	574	6600
1	14	20	12	574	7175
2	14	35	12	674	8424
2	34	84	13	674	9099
3	34	119	13	782	10556
3	62	216	14	782	11339
4	62	279	3	870	3015
1	84	119	62	870	54404
14	84	1224	14	898	13020
4	98	440	15	898	13919
5	98	539	15	1022	15840
5	142	780	16	1022	16863
6	142	923	4	1121	5015
2	143	351	62	1121	70091
14	143	2079	16	1154	19040
6	194	1260	17	1154	20195
7	194	1455	17	1294	22644
7	254	1904	18	1294	23939
8	254	2159	2	1420	3479
8	322	2736	143	1420	203840
9	322	3059	18	1442	26676
2	341	836	19	1442	28119
34	341	11780	19	1598	31160
9	398	3780	20	1598	32759
10	398	4179	20	1762	36120
3	480	1664	21	1762	37883
34	480	16575	4	1767	7904
10	482	5060	98	1767	174096
11	482	5543	21	1934	41580
1	492	696	22	1934	43515
84	492	41615			

Looking for patterns, our first question may be: *Are there any recursion relations*? For each value of x there are obviously many solutions. Collecting triples with the same x may reveal a pattern. For example, for $x = 1$, we get the following list, and a pattern starts to appear:

TABLE 2 Solutions where $x = 1$

x	1	1	1	1	1	1	1
y	0	2	14	84	492	2870	16730
z	0	3	20	119	696	4059	23660

Here we observe the recursive pattern

$$x_n = 1; \qquad y_n = 6y_{n-1} - y_{n-2} + 2; \qquad z_n = 6\,z_{n-1} - z_{n-2} + 2\,.$$

The starting triples are $(1,0,0)$ and $(1,2,3)$, and the following triples are recursively defined. Thus we conjecture one infinite string of solutions of the Diophantine equation (1). We will generalize this conjecture and, in Section 3, show that the conjecture is sound.

The case for $x = 1$ appears atypical; here we observe just one recursive string. So let us look at the case $x = 2$. The recursive relationships will become clearer if we split the triples into two strings, extend the list, and adjust the starting numbers to clarify each pattern.

TABLE 3 Solutions with $x = 2$

x	2	2	2	2	2	2
y	0	1	14	143	1420	14061
z	-1	3	35	351	3479	34443

x	2	2	2	2	2	2
y	0	3	34	341	3380	33463
z	0	8	84	836	8280	81968

The starting values are different in each string, but the recursion relation is the same. Again by observation, this is

$$x_n = 2; \qquad y_n = 10y_{n-1} - y_{n-2} + 4; \qquad z_n = 10z_{n-1} - z_{n-2} + 4.$$

What happens for larger x? We look in the same way at the case $x = 3$. These two strings appear:

TABLE 4 Solutions with $x = 3$

x	3	3	3	3	3	3
y	0	2	34	480	6692	93214
z	-1	8	119	1664	23183	332904

x	3	3	3	3	3	3
y	0	4	62	870	12124	168872
z	0	15	216	3015	39201	545805

Both strings are produced by the same recursion relation:

$$x_n = 3; \qquad y_n = 14 y_{n-1} - y_{n-2} + 6; \qquad z_n = 14\, z_{n-1} - z_{n-2} + 6$$

with different initial values.

The case $x = 14$ is especially interesting; it gives *four* strings of solutions:

TABLE 5 Solutions with $x = 14$

x	14	14	14	14
y	0	13	782	45371
z	−1	195	11339	657495

x	14	14	14	14
y	0	15	898	52097
z	0	224	13020	754964

x	14	14	14	14
y	2	1	84	4899
z	−36	20	1224	71000

x	14	14	14	14
y	1	2	143	8320
z	−21	35	2097	120575

all of which obey the same recursion relation:

$$x_n = 14; \qquad y_n = 58 y_{n-1} - y_{n-2} + 28; \qquad z_n = 58 z_{n-1} - z_{n-2} + 28.$$

Now a pattern emerges in the recursion formulae themselves. The following general form includes all the preceding recursion formulae:

$$x_n = k; \quad y_n = (4k+2) y_{n-1} - y_{n-2} + 2k; \quad z_n = (4k+2)\, z_{n-1} - z_{n-2} + 2k. \quad (2)$$

The initial values for each k have to be determined, and each pair will give rise to a chain of solutions. In Section 3 we prove that these chains give *all* solutions of (1).

We have seen a web of recursive strings of solutions. Our next question may be: *Are there any other internal relations between solutions*? From the data in Table 1 we observe that the solutions appear in *related triplets*. This relation can be illustrated by the following examples.

$$\begin{array}{ccccc}
(2,14,35) & — & (2,143,351) & — & (14,143,2079) \\
(2,3,8) & — & (2,34,84) & — & (3,34,119) \\
(2,34,84) & — & (2,341,836) & — & (34,341,11780)
\end{array}$$

This triplets relationship is described by $(a, b, r) — (a, c, s) — (b, c, t)$, where $a < b < c$, in the following way: If (a, b, r) is a solution, then there is another one with the same x, say (a, c, s), and also a third one, related to them by having the form (b, c, t). The solution (a, b, r) "gives birth" to the two others. This relation is illustrated in FIGURE 4.

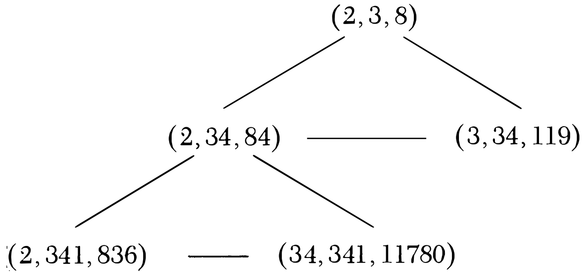


FIGURE 4

The solutions appear as related triplets: A "mother" solution with two "children."

Next we ask: *Can the solutions be expressed by explicit formulae*? To put these observations into a more precise form, we will use algebraic expressions for the solutions. We have already mentioned the solutions $(x, y, z) = (0, k, 0)$ and

$$x = k; \qquad y = k + 1; \qquad z = k^2 + 2k \tag{3}$$

for $k = 1, 2, 3, \ldots$. If we start with the solutions given in (3) as the "mother generation," what triplets are generated? As Table 6 illustrates, each solution generates two more. The upper grid shows the "mother" triples, all of type (3); the lower grids show the corresponding triples in each case.

TABLE 6 Every solution generates two more

x	1	2	3	4	5	6
y	2	3	4	5	6	7
z	3	8	15	24	35	48

x	1	2	3	4	5	6
y	14	34	62	98	142	194
z	20	84	216	440	780	1260

x	2	3	4	5	6	7
y	14	34	62	98	142	194
z	35	119	279	539	923	1455

From these facts we may, by observation, find algebraic expressions for classes of solutions. Since the second difference in y is constant, the formula for y is a polynomial in the variable x of degree 2. Similarly, z is expressed by a polynomial in x of degree 3. By using undetermined coefficients, we find the formulae

$$x = k; \qquad y = 4k^2 + 8k + 2; \qquad z = 4k^3 + 10k^2 + 6k \tag{4}$$

for entries in the middle grid of Table 6, and

$$x = k + 1; \qquad y = 4k^2 + 8k + 2; \qquad z = 4k^3 + 14k^2 + 14k + 3 \tag{5}$$

for entries in the bottom grid.

A simpler description of the latter family of triples is obtained by retarding the parameter k by 1 to obtain

$$x = k; \qquad y = 4k^2 - 2; \qquad z = 4k^3 + 2k^2 - 2k - 1 \tag{6}$$

for $k = 2, 3, 4, \ldots$. All of the triples generated by (4) and (6) are solutions of the Diophantine equation (1), as is easily verified algebraically.

We could continue using the same method and find formulae for the next generation of solutions, based upon the numerical solutions established, until the situation grows too complicated or we run out of numerical data. The solutions we find in this way are only partial, but we see that there are infinitely many solutions, and that explicit expressions exist for classes of solutions.

3. Relations to the Pell equation

Our problem is to find all triples of integers x, y, and z satisfying equation (1):

$$x(x+1)y(y+1) = z(z+1).$$

We observe immediately that x and y appear symmetrically.

Chains of solutions Suppose that (x, y, z) is a solution of (1). Let us fix $x = k$ and look for more solutions with this x. For simplicity, we put $k(k+1) = d$; then (1) gives

$$dy(y+1) = z(z+1),$$

which is equivalent to

$$(2z+1)^2 - d(2y+1)^2 = 1 - d.$$

Substituting $u = 2y + 1$ and $v = 2z + 1$ produces the equation

$$v^2 - du^2 = 1 - d. \tag{7}$$

Equation (7) can be solved using the theory of *Pell's equation*

$$v^2 - du^2 = 1, \tag{8}$$

an interesting Diophantine equation in its own right, well known in number theory (see, e.g., [**1**, Ch. 8], or [**4**, Ch. 8.2]). There is a "fundamental solution" (v_1, u_1) of equation (8), which generates a complete set of solutions in the following manner.

The equation $v_1^2 - du_1^2 = 1$ may be written in the surd form

$$\left(v_1 + u_1\sqrt{d}\right)\left(v_1 - u_1\sqrt{d}\right) = 1.$$

Raising this to the nth power, for any integer n, yields

$$\left(v_n + u_n\sqrt{d}\right)\left(v_n - u_n\sqrt{d}\right) = 1 = v_n^2 - du_n^2,$$

where $v_n \pm u_n\sqrt{d} = (v_1 \pm u_1\sqrt{d})^n$. The solution in integers of $v^2 - du^2 = 1$ are precisely of the form $(v, u) = (\pm v_n, \pm u_n)$. When $d = k(k+1)$, the fundamental solution is $(v, u) = (2k+1, 2)$.

Now suppose we have the more general equation $v^2 - du^2 = c$, where c is an integer. Then, given any particular solution $(v, u) = (p, q)$, we can generate infinitely many additional solutions from the relation

$$v - u\sqrt{d} = \left(p - q\sqrt{d}\right)\left(v_n - u_n\sqrt{d}\right),$$

where $v_n^2 - du_n^2 = 1$.

In the case of equation (7), with $d = k(k+1)$, we may start with the solutions $(v, u) = (1, 1)$ and $(v, u) = (-1, 1)$ and obtain a chain of solutions.

Thus if (v_n, u_n) is a solution of (7), then also

$$\left(v_n + u_n\sqrt{k(k+1)}\right)\left((2k+1) + 2\sqrt{k(k+1)}\right)$$
$$= \left((2k+1)v_n + 2k(k+1)u_n\right) + \left(2v_n + (2k+1)u_n\right)\sqrt{k(k+1)}$$

yields a solution of the same equation (7). From an initial solution we thus get a chain of solutions (v_n, u_n) where

$$v_{n+1} = (2k+1)v_n + 2k(k+1)u_n; \qquad u_{n+1} = 2v_n + (2k+1)u_n. \tag{9}$$

Letting

$$X_n = \begin{bmatrix} v_n \\ u_n \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} 2k+1 & 2k^2+2k \\ 2 & 2k+1 \end{bmatrix},$$

equation (9) can be written in the form $X_{n+1} = MX_n$. The matrix M satisfies its own characteristic equation, which is easily seen to be $((2k+1) - M)^2 - 2 \cdot 2k(k+1) = 0$ or, equivalently, $M^2 - 2(2k+1)M + I = 0$. This implies that

$$v_{n+1} = 2(2k+1)v_n - v_{n-1} \quad \text{and} \quad u_{n+1} = 2(2k+1)u_n - u_{n-1}.$$

Since $u_n = 2y_n + 1$ and $v_n = 2z_n + 1$, we deduce the recursion formula (2)

$$x_n = k; \qquad y_n = (4k+2)\,y_{n-1} - y_{n-2} + 2k; \qquad z_n = (4k+2)\,z_{n-1} - z_{n-2} + 2k,$$

which we conjectured in Section 2.

Thus, for any value $x = k$, we find infinitely many solutions. The starting pairs $(v_0, u_0) = (1, 1)$, and $(v_0, u_0) = (-1, 1)$ give two chains of solutions, which are different except for $k = 1$. For some k, however, we may have more initial values. For $k = 14$, for instance, the pairs $(v_0, u_0) = (41, 3)$ and $(v_0, u_0) = (-41, 3)$ will generate two chains different from the previous ones, as observed in Table 5.

Triplets of solutions Let $f(x) = x(x+1)$. The solutions appears in triplets, of the form

$$(a, b, r) — (a, c, s) — (b, c, t),$$

where it turns out that, since $f(a)f(b) = f(r)$, $f(a)f(c) = f(s)$, and $f(b)f(c) = f(t)$, it follows that $f(r)f(s) = (f(a))^2 f(t)$. By symmetry, $f(s)f(t) = (f(c))^2 f(r)$ and $f(r)f(t) = (f(b))^2 f(s)$. Let us look at this more closely. From the recursion it follows that if (a, b, r) is a solution of (1), then so is (a, c, s), where

$$(2s+1) + (2c+1)\sqrt{a(a+1)}$$
$$= \left((2r+1) + (2b+1)\sqrt{a(a+1)}\right)\left((2a+1) + 2\sqrt{a(a+1)}\right).$$

Comparing the surd part and the integer part on both sides above gives

$$c = 2ab + a + b + 2r + 1; \qquad s = 2a^2b + a^2 + 2ab + 2ar + 2a + r.$$

Thus c is symmetric in a and b. Since (b, a, r) is a solution, so is (b, c, t), where

$$t = 2ab^2 + 2ab + b^2 + 2br + 2b + r.$$

4. Educational perspectives

Working on polygonal numbers can give students rich experiences with investigations, search for patterns, and deductive study of equations, and may also lead into parts of number theory. One nice source of tasks like this is **[6]**. Following are some suggested related problems, all of which lead to Pell's equation.

1. Study numbers of the form $n(n+2)$: When is the product of two numbers in this family another one of the same form? Generalize it to numbers of the form $n(n+k)$ where k is any positive integer.
2. Study other classes of polygonal numbers: When is the product of two triangular numbers again triangular? (Squares are not interesting, because the set is obviously closed under multiplication.) But when is the product of two pentagonal numbers another pentagonal? When is the product of two hexagonals again hexagonal? And so on. Here, the algebraic formulae for polygonal numbers may be useful: triangular: $n^2/2 + n/2$; square: n^2, pentagonal: $3n^2/2 - n/2$; hexagonal: $2n^2 - n$; etc.
3. Study relations among different polygonal numbers: When is a triangular number also a square? For any integer k, what triangular numbers are k times a square?
4. When is a triangular number twice a triangular number? (FIGURE 3b shows the relation between this and the preceding problem.) For any integer k, what triangular numbers are k times a triangular number?

Acknowledgments. Thanks to Bob Burn and to anonymous referees for useful suggestions.

REFERENCES

1. W. W. Adams and L. J. Goldstein, *Introduction to Number Theory*, Prentice Hall, Englewood Cliffs, NJ, 1976.
2. D. M. Burton, *History of Mathematics: An Introduction*, third ed., W. C. Brown Publishers, Boston, MA, 1991.
3. J. Lehel, Proof without words: the sum of odd numbers, this MAGAZINE 64 (1991), p. 103.
4. W. J. leVeque, *Fundamentals of Number Theory*, Addison-Wesley, Reading, MA, 1977.
5. M. Masi, *Boethian Number Theory*, a translation of the *De Institutione Arithmetica*, Rodopi, Amsterdam, Holland, 1983.
6. J. H. Conway and R. K. Guy, *The Book of Numbers*, Copernicus, New York, NY, 1996.

The following problem, proposed by H.T.R. Aude of Colgate University, appeared as Problem No. 130 in the March 1937 issue of the *Magazine*:

> A man (living now) states that he was x years old in the year x^2. He adds, If to the number of my years you add the number of my month, it equals the square of the date (i.e., the day of the month) of my birthday. When was he born?

See page 134 for the answer.

NOTES

Eisenstein's Lemma and Quadratic Reciprocity for Jacobi Symbols

BRETT A. TANGEDAL
College of Charleston
Charleston, SC 29424

Introduction Almost every textbook that offers an elementary proof of the classical law of quadratic reciprocity follows a pattern laid down by Gauss in his third proof of this famous law. They begin with a lemma named in Gauss's honor and after some manipulations with the greatest integer function, they complete the proof by counting lattice points in the X-Y plane (see, for example **[2]**, **[5]**, **[6]**). The lattice point argument that clinches the proof in the end is actually due to Eisenstein **[1]**. Eisenstein's half-forgotten paper contained other important innovations that were entirely lost until quite recently. In a fine piece of historical scholarship, Laubenbacher and Pengelley **[4]** have displayed the gem-like qualities of Eisenstein's *entire* presentation. A true piece of art has now been fully restored and made whole again! What everyone missed until **[4]** was Eisenstein's replacement of Gauss's lemma by a much easier to use result rightly designated in **[4]** as "Eisenstein's lemma." Eisenstein used his lemma to give a remarkably direct and insightful proof of the classical law of quadratic reciprocity. As is amply demonstrated in **[4]**, Eisenstein's proof simplifies and improves upon Gauss's third proof at every step and truly deserves to replace the standard proof in the textbooks. In order to add weight to their argument we show here that Eisenstein's entire presentation along with his lemma generalize nicely to give a direct proof of the quadratic reciprocity law for Jacobi symbols.

Jacobi's reciprocity law Throughout this note, a small Latin letter denotes an integer. The letter n will always denote an odd integer greater than 1. If $n = p$ is prime, and p does not divide k, the Legendre symbol $\left(\frac{k}{p}\right)$ is equal to 1 if there is an integer solution to the congruence $x^2 \equiv k(\text{mod } p)$. We set $\left(\frac{k}{p}\right) = -1$ if there is no such solution. (In general, $\left(\frac{k}{p}\right)$ is set equal to 0 if p divides k, but this case is of no interest to us.) The Jacobi symbol is a natural generalization of the Legendre symbol to the case where n is composite. By the fundamental theorem of arithmetic, we can factor n uniquely as a product of odd primes:

$$n = p_1^{n_1} \cdots p_t^{n_t} \quad \text{where} \quad t \geq 1 \quad \text{and} \quad n_i > 0 \quad \text{for all } i.$$

We define the Jacobi symbol $\left(\frac{k}{n}\right)$ only in the case where k is relatively prime to n, by

$$\left(\frac{k}{n}\right) = \prod_{i=1}^{t} \left(\frac{k}{p_i}\right)^{n_i}$$

which has a value of ± 1 since none of the p_i divides k. The quadratic reciprocity law for Jacobi symbols (or simply "Jacobi's reciprocity law") is a three-part statement:

1. $\left(\frac{-1}{n}\right) = (-1)^{(n-1)/2}$
2. $\left(\frac{2}{n}\right) = (-1)^{(n^2-1)/8}$
3. If $m > 1$ is odd and relatively prime to n, then $\left(\frac{m}{n}\right)\left(\frac{n}{m}\right) = (-1)^{(m-1)(n-1)/4}$.

The classical law of quadratic reciprocity is the above statement with m and n being distinct odd primes.

We use the following notation to introduce Eisenstein's lemma. Let E denote the set of all positive even integers. We set $E_n = \{a \in E | a < n\}$. Let S be any non-empty subset of the integers. For a given m, we write mS for the set $\{ms | s \in S\}$. So, for example, $-3E_7 = \{-6, -12, -18\}$. Also, if $d > 1$, we let $[S]_d$ denote the set of all remainders of elements in S modulo d. For example, $[-3E_7]_9 = \{3, 6, 0\}$. What is referred to in **[4]** as "Eisenstein's lemma" is the following algebraic formula, when p is an odd prime, for the Legendre symbol:

$$\left(\frac{k}{p}\right) = (-1)^{\Sigma r}, \quad \text{where the summation is over all } r \in \left[kE_p\right]_p.$$

Our goal here is to show that this formula is valid for the Jacobi symbol as well:

EISENSTEIN'S LEMMA. $\left(\frac{k}{n}\right) = (-1)^{\Sigma r}$, *with summation over all* $r \in [kE_n]_n$.

Comment. The set E_n has $(n-1)/2$ elements and so does $[kE_n]_n$ because two distinct elements $a_1, a_2 \in E_n$ give $ka_i = q_i n + r_i$ for $i = 1, 2$, and $r_1 \neq r_2$ since k is relatively prime to n.

As one might expect, this more general version of Eisenstein's lemma is somewhat harder to prove than the original and we postpone the proof until the final section. What we gain in the meantime is direct access to Jacobi's reciprocity law using Eisenstein's superior method. Typically, Jacobi's reciprocity law is obtained from the classical reciprocity law only after a tedious computation that offers very little insight **[5]**.

We now show that the three parts of Jacobi's law follow readily from Eisenstein's lemma.

1. With $k = -1$ in Eisenstein's lemma, the remainders are the elements of the set $\{-2+n, \dots, -(n-1)+n\}$. Note that $\Sigma r \equiv (n-1)/2 (\text{mod } 2)$, since each r is odd.

2. Let $k = 2$. First assume that $n \equiv 1 (\text{mod } 4)$. Then the set of remainders is $\left\{4, \dots, 2 \cdot \frac{(n-1)}{2}\right\} \cup \{2 \cdot \frac{(n+3)}{2} - n, \dots, 2(n-1) - n\}$. The elements of the first set are even, while those of the second are odd. Counting the odd elements, we have $\left(\frac{2}{n}\right) = (-1)^{(n-1)/4}$. Finally, note that $(n-1)/4 \equiv (n^2-1)/8 (\text{mod } 2)$ for either $n \equiv 1 (\text{mod } 8)$ or $n \equiv 5 (\text{mod } 8)$. If $n \equiv 3 (\text{mod } 4)$ and $n > 3$ (note that $\left(\frac{2}{3}\right) = -1$), the set of remainders is $\left\{4, \dots, 2 \cdot \frac{(n-3)}{2}\right\} \cup \{2 \cdot \frac{(n+1)}{2} - n, \dots, 2(n-1) - n\}$. Applying the same analysis as above finishes the proof.

3. (Following Eisenstein.) For $a \in E_n$, we have $\frac{ma}{n} = q + \frac{r}{n}$ with $0 \leq \frac{r}{n} < 1$. Thus, $q = \left[\frac{ma}{n}\right]$, the greatest integer less than or equal to $\frac{ma}{n}$. Working (mod 2), and remembering that a is even and n is odd, the equation $ma = qn + r$ becomes $r \equiv \left[\frac{ma}{n}\right] (\text{mod } 2)$. Eisenstein's lemma gives

$$\left(\frac{m}{n}\right) = (-1)^{\Sigma_{a \in E_n}\left[\frac{ma}{n}\right]}.$$

It is now convenient to modify the sum in the exponent. Let c satisfy $0 < c < \frac{n}{2}$. If $mc = q_1 n + r_1$, then $r_1 > 0$ since m and n are relatively prime. Adding this equation to $m(n-c) = q_2 n + r_2$ gives $mn = (q_1 + q_2)n + r_1 + r_2$ with $0 < r_1 + r_2 < 2n$. Since n divides $r_1 + r_2$, we have $n = r_1 + r_2$, and deduce that

$$m - 1 = \left[\frac{mc}{n}\right] + \left[\frac{m(n-c)}{n}\right].$$

Since m is odd by assumption, $\left[\frac{mc}{n}\right] \equiv \left[\frac{m(n-c)}{n}\right] (\text{mod } 2)$. An odd c in the range $0 < c < \frac{n}{2}$ corresponds uniquely to an even $n - c$ in the range $\frac{n}{2} < n - c < n$. Thus, the values of a greater than $\frac{n}{2}$ can be transformed into the odd values less than $\frac{n}{2}$, yielding

$$\sum_{a \in E_n} \left[\frac{ma}{n}\right] \equiv \sum_{i=1}^{(n-1)/2} \left[\frac{mi}{n}\right] (\text{mod } 2).$$

This latter sum has a nice geometric interpretation, illustrated in Figure 1. If $i = 7$, for example, there are $\left[\frac{29 \cdot 7}{45}\right] = 4$ lattice points of the form $(7, y)$ with $y > 0$ below the line $Y = \frac{29}{45}X$. Thus, this sum is simply the total number of lattice points inside the triangle ABC. No lattice points can lie on the diagonal since m and n are relatively prime. By symmetry,

$$\left(\frac{n}{m}\right) = (-1)^{\sum_{j=1}^{(m-1)/2}\left[\frac{nj}{m}\right]}$$

and the sum in the exponent here counts the number of lattice points inside the triangle ACD. The total number of lattice points inside the rectangle $ABCD$ is $(m-1)/2 \cdot (n-1)/2$, from which the third part follows immediately.

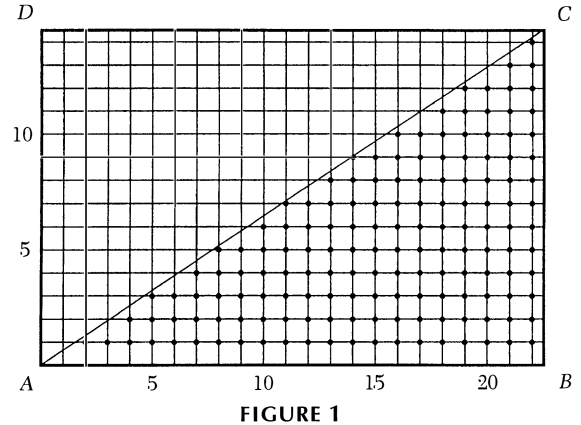


FIGURE 1
Counting lattice points.

Proof of Eisenstein's lemma In this section we require the basic properties of the Euler ϕ-function, and Euler's criterion [6] that $k^{(p-1)/2} \equiv \left(\frac{k}{p}\right) (\text{mod } p)$ for p an odd prime and $(k, p) = 1$. The letter d will denote an odd integer greater than 1 that divides $n = p_1^{n_1} \cdots p_t^{n_t}$, and $E(d) = \{b \in E | b < d \text{ and } (b, d) = 1\}$. The number of elements in $E(d)$ is $\phi(d)/2$ since an odd c with $0<c<d$ and $(c,d)=1$ always corresponds to an even $b=d-c$ with $(b,d)=1$. We can partition the set E_n as follows

$$E_n=\bigcup_{d|n}\frac{n}{d}E(d)$$

since $a\in E_n$ satisfies $(a,n)=\frac{n}{d}$ iff $a\in\frac{n}{d}E(d)$. If we let r denote a generic element of $[kE_n]_n$ and r_d an element of $[k\cdot\frac{n}{d}E(d)]_n$, then, by the comment following the statement of Eisenstein's lemma, we have

$$(-1)^{\Sigma r}=(-1)^{\Sigma_{d|n}(\Sigma r_d)}.$$

Now let $b\in E(d)$ and write $kb=qd+s$ where $0\le s<d$. If we multiply both sides of this equation by $\frac{n}{d}$, the relationship $\frac{n}{d}[kE(d)]_d=[k\cdot\frac{n}{d}E(d)]_n$ becomes apparent. If $s_d\in[kE(d)]_d$, then $\frac{n}{d}\cdot s_d=r_d\in[k\cdot\frac{n}{d}E(d)]_n$ and since $\frac{n}{d}$ is odd, $s_d\equiv r_d \pmod 2$. Thus,

$$(-1)^{\Sigma r}=(-1)^{\Sigma_{d|n}(\Sigma s_d)} \tag{1}$$

and we now turn to a detailed study of the set $[kE(d)]_d$.

LEMMA 1. $k^{\phi(d)/2}\equiv(-1)^{\Sigma s_d} \pmod d$, *with summation over all* $s_d\in[kE(d)]_d$.

Proof. (Following Eisenstein.) Let $b_1,\dots,b_{\phi(d)/2}$ denote the elements of $E(d)$. Write $kb_i=q_id+s_i$ for $i=1,\dots,\phi(d)/2$. Let $S=\{(-1)^{s_i}s_i|i=1,\dots,\phi(d)/2\}$. We claim that $[S]_d=E(d)$. Assuming this is true for the moment, we have

$$\prod_{i=1}^{\phi(d)/2}b_i\equiv\prod_{i=1}^{\phi(d)/2}(-1)^{s_i}s_i\equiv\prod_{i=1}^{\phi(d)/2}(-1)^{s_i}kb_i \pmod d.$$

Since $\prod_{i=1}^{\phi(d)/2}b_i$ is relatively prime to d by construction, we can cancel it from the first and third products above to obtain Lemma 1. To establish the claim, note that every element $a\in[S]_d$ is even and satisfies $0<a<d$ and $(a,d)=1$. We need only show that $(-1)^{s_i}s_i$ and $(-1)^{s_j}s_j$ are distinct mod d for $i\ne j$. The only non-trivial case is where s_i is even and s_j is odd. If $s_i\equiv -s_j \pmod d$, then $d|k(b_i+b_j)$. Let $b_i=2a_i$ where $1\le a_i<\frac{d}{2}$ and, similarly, $b_j=2a_j$. Then $d|2k(a_i+a_j)$ or $d|(a_i+a_j)$ since $(d,2k)=1$. But $2\le a_i+a_j<d$, which gives a contradiction.

LEMMA 2. *We have* $k^{\phi(d)/2}\equiv(\frac{k}{p})(\text{mod } d)$ *if* $d=p^m$, *and* $k^{\phi(d)/2}\equiv 1(\text{mod } d)$ *if* d *is not a prime power.*

Proof. First assume that $d=p^m$ (p is necessarily odd). Euler's criterion gives $k^{(p-1)/2}\equiv\left(\frac{k}{p}\right)(\text{mod } p)$, so we are done if $m=1$. Now assume $m>1$. If $l\ge 1$ and $a\equiv b(\text{mod } p^l)$, then $a^p\equiv b^p(\text{mod } p^{l+1})$ by the binomial theorem. Applying this to Euler's criterion $m-1$ times, we obtain $k^{\frac{1}{2}(p-1)p^{m-1}}\equiv\left(\frac{k}{p}\right)^{p^{m-1}}(\text{mod } p^m)$. The left hand side is just $k^{\phi(d)/2}$ and the right side is equal to $\left(\frac{k}{p}\right)$ since p^{m-1} is odd. Now assume that d is not a prime power. We write $d=p_1^{m_1}p_2^{m_2}\cdots p_t^{m_t}$ and assume without loss of generality that m_1 and m_2 are positive. From above, $k^{\phi(d)/2}=(k^{\frac{1}{2}(p_1-1)p_1^{m_1-1}})^{(p_2-1)p_2^{m_2-1}\cdots}\equiv 1 \;(\text{mod } p_1^{m_1})$ since (p_2-1) is even. Similarly, $k^{\phi(d)/2}\equiv 1$ $(\text{mod } p_i^{m_i})$ for $i=2,\dots,t$, and thus $k^{\phi(d)/2}\equiv 1 \;(\text{mod } d)$.

Combining Lemmas 1 and 2, we have $(-1)^{\Sigma s_d} \equiv \left(\frac{k}{p}\right) \pmod{d}$ when $d = p^m$. Since both sides of this congruence are ± 1 and $d > 2$ this means that $(-1)^{\Sigma s_{p^m}} = \left(\frac{k}{p}\right)$. Thus, $\Sigma s_{p_i} \equiv \Sigma s_{p_i^m} \pmod 2$ for $1 \le m \le n_i$. Similarly, $(-1)^{\Sigma s_d} = 1$ if d is not a prime power, which says that Σs_d is even in this case. We now conclude that

$$\sum_{d|n}\left(\sum s_d\right) \equiv \left(\sum s_{p_1} + \cdots + \sum s_{p_1}^{n_1}\right) + \cdots + \left(\sum s_{p_t} + \cdots + \sum s_{p_t}^{n_t}\right) \pmod 2$$

$$\equiv n_1 \sum s_{p_1} + \cdots + n_t \sum s_{p_t} \pmod 2.$$

Combining with equation (1) finally gives

$$(-1)^{\Sigma r} = \left[(-1)^{\Sigma s_{p_1}}\right]^{n_1} \cdots \left[(-1)^{\Sigma s_{p_t}}\right]^{n_t} = \left(\frac{k}{p_1}\right)^{n_1} \cdots \left(\frac{k}{p_t}\right)^{n_t} = \left(\frac{k}{n}\right).$$

Eisenstein proved Lemma 1 for the special case $d = p$, an odd prime, exactly as above. Since $k^{\phi(d)/2} = k^{(p-1)/2} \equiv \left(\frac{k}{p}\right) \pmod{p}$ by Euler's criterion, his original lemma follows immediately. The proof of the more general "Eisenstein lemma" given here is modeled after a proof of Jenkins [3]. Jenkins was the first to notice that Gauss's lemma could be generalized from the Legendre symbol to the Jacobi symbol.

REFERENCES

1. Gotthold Eisenstein, Geometrischer Beweis des Fundamentaltheorems für die quadratischen Reste, *J. Reine Angew. Math.* 28 (1844), 246–248; also in *Mathematische Werke*, Vol. I, 2nd Ed., Chelsea, New York, NY, 1989, pp. 164–166; English Translation with added commentary by Arthur Cayley in Volume III of Cayley's *Collected Mathematical Papers*, The University Press, Cambridge, UK, 1890, pp. 39–43.
2. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th Ed., Oxford University Press, Oxford, UK, 1979.
3. M. Jenkins, Proof of an Arithmetical Theorem leading, by means of Gauss's Fourth Demonstration of Legendre's Law of Reciprocity, to the extension of that Law, *Proc. London Math. Soc.* 2 (1867), 29–32.
4. Reinhard Laubenbacher and David Pengelley, Gauss, Eisenstein, and the "Third" Proof of the Quadratic Reciprocity Theorem: Ein kleines Schauspiel, *Math. Intelligencer* 16 (1994), 67–72.
5. I. Niven, H. Zuckerman, and H. Montgomery, *An Introduction to the Theory of Numbers*, 5th Ed., John Wiley and Sons, New York, NY, 1991.
6. J. Strayer, *Elementary Number Theory*, PWS Publishing Company, Boston, MA, 1994.

Solution to Problem 130 (see page 129) by Lucille G. Meyer, New Orleans, Louisiana. (The problem appeared in 1937 – Ed.)

A man living now could not have been 43 years old in 1849, that is 43^2. Therefore, the man must have been 44 years old in 1936. From the conditions given,

$$44 + m = d^2 \quad \text{and} \quad 0 < m < 13$$

whence $m = 5$ is the only integral solution, and $d = 7$. The man was born May 7, 1892.

Monty Hall Uses a Mixed Strategy

HERB BAILEY
Rose-Hulman Institute of Technology
Terre Haute, IN 47803

Introduction Marilyn vos Savant posed a problem [12] simulating a popular TV game show hosted by Monty Hall. Her correct solution brought thousands of letters telling her that she was wrong. We quote her statement of the problem: "Suppose you're on a game show, and you are given a choice of three doors. Behind one is a car; behind the others, goats. You pick a door—say, No. 1—and the host, who knows what's behind the doors, opens another door—say, No. 3—which has a goat. He then says to you, 'Do you want to pick door No. 2?' Is it to your advantage to switch your choice?"

Marilyns phrase "say, No. 3" is a little ambiguous. To clarify this, we note that if Doors 2 and 3 hide a goat and a car, then the host opens the door hiding the goat. On the other hand, if Doors 2 and 3 both hide goats then the host flips a coin to choose between these two doors. A number of interesting articles (e.g., [3], [7], and [8]) analyze games in which the host does not make a random choice in the latter case.

In 1959, Martin Gardner [6] posed a problem equivalent to Marilyn's game involving three prisoners with one to be paroled. Gardner describes the game as a "wonderfully confusing little problem." Another equivalent problem, involving three boxes, was posed by S. Selvin [10] in 1975. Ed Barbeau [1] has written a good review of the literature on Marilyn's game and related problems. A recent paper by Fernandez and Piron [4] considers Marilyn's game when the host influences the contestant to switch in certain situations so that the game is less predictable, and thus generates more audience interest.

Two appealing solutions to Marilyn's game are:

S1: After the host shows a goat, the contestant is looking at two closed doors with a car behind one of them. Thus there is a 50-50 chance with either door and there is no advantage in switching,

S2: The probability of picking the car with her first choice was 1/3 and this does not change when the host shows a goat. Since the car is behind one of the two closed doors, the probability is 2/3 that the car is behind the other closed door and she should switch.

In this note, we generalize the game by considering a set of N doors, with a car behind one of the doors and goats behind the rest. The sponsors of the show ask Monty to give away as few cars as possible. We consider three different generalizations, and in the third one, both the host and contestant use mixed strategies. In general, mixed strategy games can only be solved for given numerical values of the parameters involved, however in our game there is an explicit solution.

Generalizations Now imagine a game played with $N = 11$ doors hiding 10 goats and one car. The contestant is asked to select 4 of the doors and leave 7 doors unselected. We call the selected set L with $N_L = 4$ and the unselected set R with $N_R = 7$. Then the host reveals $H = 5$ doors that conceal goats, say by opening $i = 2$ doors from the selected set L and $H - i = 3$ doors from the unselected set R. The contestant then chooses any unopened door. Should she pick from L or R? This example is shown in FIGURE 1.

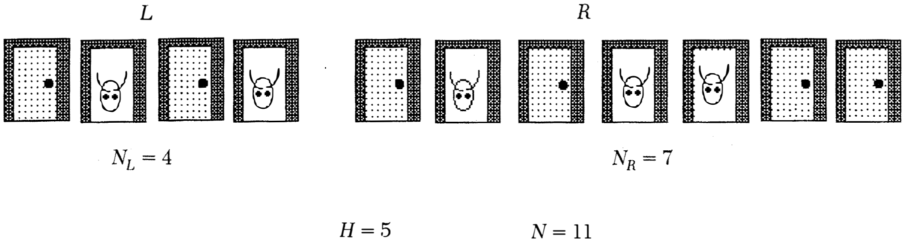


FIGURE 1
Example 1, where the host shows two goats from L and three from R.

In general, we shall assume that H, N_L and N_R are given, with $N_L + N_R = N$. The host opens H of the doors. The contestant's strategy is to pick a door (at random) from the remaining closed doors in L or from those remaining in R in order to maximize her probability of finding the car. If i is the number of doors that the host opens from L, then the host's strategy is to choose i in order to minimize the probability that the contestant picks the car.

The host is not permitted to open all the doors in either L or R, since the possibility must remain that the car can be in either subset. To ensure this, we require $H \leq N_L + N_R - 2$, $i \leq \min(H, N_L - 1)$, and $i \geq \max(0, H - N_R + 1)$. In Example 1, the host has four possible strategies: $i = 0$, 1, 2, or 3. FIGURE 1 corresponds to the choice $i = 2$.

Let P_L be the probability that the contestant will pick the car if she chooses at random from the remaining closed doors in L. Let P_R be defined the same way if she decides to pick from R. The values of P_L are obtained by multiplying the probability N_L/N that the car is in L by the probability $1/(N_L - i)$ that she picks the car from L, given that it is in L. Thus

$$P_L = \frac{N_L}{N}\frac{1}{N_L - i}.$$

Similarly

$$P_R = \frac{N_R}{N}\frac{1}{N_R - (H - i)}.$$

The columns of Table 1 correspond to the four possible strategies ($i = 0, 1, 2, 3$) of the host for Example 1. The rows correspond to the two possible choices of the contestant (L or R). The table entries are the values of P_L and P_R. The last column is the row minimum, and the last row is the column maximum. The *maximin* is the maximum of the row minimums and in this example is 0.127. The *minimax* is the minimum of the column maximums and for this example is 0.182.

TABLE 1. Example 1 with $N_L = 4$, $N_R = 7$, and $H = 5$.

	$i = 0$	$i = 1$	$i = 2$	$i = 3$	Row Min
P_L	$\frac{4}{11}\frac{1}{4} \cong 0.091$	$\frac{4}{11}\frac{1}{3} \cong 0.121$	$\frac{4}{11}\frac{1}{2} \cong 0.182$	$\frac{4}{11}\frac{1}{1} \cong 0.364$	0.091
P_R	$\frac{7}{11}\frac{1}{2} \cong 0.318$	$\frac{7}{11}\frac{1}{3} \cong 0.212$	$\frac{7}{11}\frac{1}{4} \cong 0.159$	$\frac{7}{11}\frac{1}{5} \cong 0.127$	0.127
Column Max	0.318	0.212	0.182	0.364	

Game I In this game, we require the host to declare his strategy, by opening H doors that hide goats, before the contestant makes her choice. The contestant then picks a door at random from the remaining closed doors of L or from those of R. We assume that the host makes the 'conservative' choice to minimize his potential loss: thus in our example, he chooses $i = 2$ corresponding to the minimax. The contestant now compares her probabilities with the host choosing $i = 2$, and picks from the set L with a probability of 0.182 of selecting the car. Note that without the hosts help the contestant must pick from 11 doors with probability of $1/11$ of getting the car.

Marilyn's game is an instance of Game I, since the host must open the door before the contestant chooses the subset. For Marilyn's game $H = 1$, and we let L be the door that the contestant initially selects, and R the remaining two doors. Then $N_L = 1$ and $N_R = 2$. The host has only one choice, namely $i = 0$ with $H - i = 1$, thus $P_L = 1/3$ and $P_R = 2/3$. Hence she should switch her choice to the remaining door in R.

Game II In this game the host and contestant reveal their choices at the same time. For Example 1 (Table 1) the host again makes the conservative choice of $i = 2$ and he can be sure of losing no more than the minimax of 0.182. The contestant, unaware of the host's strategy, also makes the conservative choice of R so that she can be sure of winning at least the maximin of 0.127. Thus in this example the host chooses $i = 2$ and the contestant picks from R giving the contestant the probability 0.159 of getting the car. This is not as high as for Game I but better than without help from the host.

A second example is shown in Table 2, with $N_L = 10$, $N_R = 3$, and $H = 9$.

TABLE 2. Example 2 with $N_L = 10$, $N_R = 3$, and $H = 9$.

	$i = 7$	$i = 8$	$i = 9$	Row Min
P_L	$\frac{10}{39} \cong 0.256$	$\frac{5}{13} \cong 0.385$	$\frac{10}{13} \cong 0.769$	0.256
P_R	$\frac{3}{13} \cong 0.231$	$\frac{3}{26} \cong 0.115$	$\frac{1}{13} \cong 0.077$	0.077
Column Max	0.256	0.385	0.769	

In this example, we note that minimax = maximin = 0.256 and also that $P_L > P_R$ in each of the columns. The contestant in this example has what is called a *dominant* strategy, since choice L is always better than R.

When the minimax $\neq$ maximin, as in Example 1, then the game is said to be *unstable*. In this case the contestant can sometimes do better, over the long run, than in Game II by mixing her strategies, while the host also mixes his strategies. This is Game III as described below. This type game is called two-person, zero-sum, and unstable (e.g., **[2]** and **[13]**).

Game III This game is identical to Game II except that it is played many times, and the contestant uses a mixed strategy, choosing L with probability x, and R with probability $1 - x$. The host also uses a mixed strategy, choosing i with probability p_i, where the sum of the p_i's is equal to 1. Let y be the expected value of the probability that the contestant wins the car. Thus y depends on the strategies of both host and contestant. In Example 1 (Table 1), if the contestant chooses $x = 1/2$ and the host chooses $p_0 = 1/2$, $p_1 = 0$, $p_2 = 0$, and $p_3 = 1/2$, then $y = (0.091 + 0.318 + 0.364 + 0.127)/4 \cong 0.225$.

The contestant seeks a strategy x that maximizes the y that she can be sure of no matter what strategy the host chooses. We will call this maximal y the *contestant's optimal outcome*, and the corresponding x the *contestant's optimal strategy*. The host seeks a strategy p_i to minimize his maximal expected loss no matter what strategy the contestant chooses. These will be called the *host's optimal strategy and outcome*. The *Minimax Theorem* of matrix games (e.g., **[2]** and **[13]**) asserts that the contestant's optimal outcome is equal to the host's optimal outcome.

Mixed strategy games can be solved graphically if one of the players has only two pure choices. This is the case in Game III, since the contestant can choose only L or R in any given play of the game. The appropriate graph for Example 1 is shown in Figure 2, where we define $a_i \equiv P_R(i)$ and $b_i \equiv P_L(i)$. The a_i and b_i are given in the second and third rows of Table 1. The line L_0 is a graph of y as a function of x if the host chooses the pure strategy $i = 0$. Thus on this line, if $x = 0$ then $y = a_0$, and if $x = 1$ then $y = b_0$. Similarly, the lines L_i pass through the points $(0, a_i)$ and $(1, b_i)$.

The heavy line segments in Figure 2 form the lower envelope of these lines. For a given x, the value of y on the lower envelope is the minimum of the y's for the four pure host strategies. The conservative contestant will then choose the x that maximizes the y on the lower envelope. For our example, we solve for the intersection of lines L_1 and L_2 to find the optimal $x = 7/15$ and optimal $y = 28/165 \cong 0.170$. Hence the contestant chooses L with probability 7/15 and R with probability 8/15 and her optimal outcome is 0.170. By the Minimax Theorem we know that the host can also find an optimal strategy (p_0, p_1, p_2, p_3) such that his optimal outcome is 0.170.

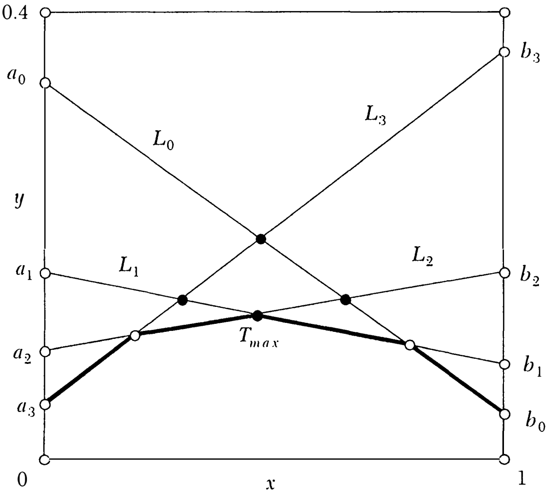


FIGURE 2
Example 1 with mixed strategies.

If the host chooses L_0 or L_3, then he will lose more than 0.170 when the contestant plays her optimal strategy. Thus he must choose $p_0 = p_3 = 0$, and the optimal strategy for the host must include only strategies i_1 and i_2. A similar graphical solution gives the optimal host strategy to be $p_0 = p_3 = 0$, $p_1 = 1/5$ and $p_2 = 4/5$. With this strategy, the host's expected loss will never be more than 0.170, no matter what strategy the contestant uses.

Game III can also be solved using linear programming, but the graphical and the LP solutions are not explicit, since they depend on knowing specific values for N_L, N_R, and H before solving. We now find a solution to Game III resulting in an explicit formula for the optimal x, optimal p_i, and the optimal y in terms of the parameters N_L, N_R, and H.

First note that the optimal solution T_{max} of any Game III must be at the intersection of two of the L_i with slopes of opposite signs; if not, the contestant could increase y by increasing (or decreasing) her choice of x. We call these intersections *viable points* and note that for Example 1 there are four viable points as shown by the small filled-in circles in FIGURE 2. For this example, the optimal solution T_{max} is the viable point with minimal y.

In the general case, let $\bar{S}(\bar{x}, \bar{y})$ be the viable point with minimal y. If $\bar{S}(\bar{x}, \bar{y}) \neq T_{max}$ then $\bar{y}$ will be less than the optimal y at T_{max} and one of the lines intersecting at $\bar{S}$ will pass below T_{max}. This contradicts the assumption that T_{max} is on the lower envelope, and thus $\bar{S}$ and T_{max} are the same point.

We ignored two complications in the above argument, namely that one of the intersecting lines might be horizontal and that more than two lines could pass through an intersection point. The result is essentially the same if these cases are included. Note that if there are no viable points, then the lower envelope of the lines L_i will not have a relative maximum. In this case its maximum will be at $x = 0$ or at $x = 1$ and the contestant has a dominant strategy.

We now find a formula for the coordinates of $\bar{S}$. The line L_i through the points $(0, a_i)$ and $(1, b_i)$ has slope $b_i - a_i$ and y-intercept a_i. Thus its equation is

$$y = a_i + (b_i - a_i)x.$$

Solving the equations for L_i and L_j, gives the intersection point (x_I, y_I), with

$$x_I = (a_j - a_i)/(a_j - b_j + b_i - a_i),$$
$$y_I = (a_j b_i - a_i b_j)/(a_j - b_j + b_i - a_i),$$

where $a_i \equiv P_R(i) = \frac{N_R}{N}\frac{1}{N_R - (H-i)}$ and $b_i \equiv P_L(i) = \frac{N_L}{N}\frac{1}{N_L - i}$. Substituting these equations into y_I gives

$$\hat{y}_I = \frac{(N-H)s(1-s)}{ij - sH(i+j) + [sH^2 + sN^2 + 2s^2NH - s^2N^2 - 2sNH]},$$

where $s = N_L/N < 1$. Since the numerator of the above expression is positive, and the values of y are positive, then the denominator must be positive. For a given game, the numerator and the bracketed terms in the denominator are fixed. Thus to minimize $\hat{y}_I$, we must maximize

$$ij - sH(i+j) = ij - sH(i+j) + s^2H^2 - s^2H^2 = (i - sH)(j - sH) - s^2H^2.$$

Since sH is fixed, we maximize $(i - sH)(j - sH)$.

For the intersection of L_i and L_j to be a viable point, one of these lines, say L_i, must have negative slope, and the other, L_j, must have positive slope. The slope of L_i is negative if $b_i < a_i$ which is equivalent to $i < sH$ (by using the equations for a_i and b_i). Likewise, the slope of L_j is positive provided $j > sH$. Thus $(i - sH)(j - sH) < 0$, and to make this product as large as possible, we choose i to be the largest integer less than sH and j to be the smallest integer greater than sH. Thus the optimal solution T_{max} is at the intersection of lines L_i and L_j, where

$$i = \lfloor sH \rfloor \quad \text{and} \quad j = \lceil sH \rceil.$$

Hence for any Game III, we calculate $sH, i, j, a_i, b_i, a_j, b_j$, and substitute these into the formulas for x_I and y_I. Then x_I is the contestant's optimal strategy, and y_I is the contestant's optimal outcome.

By the Minimax Theorem, the host's optimal outcome is equal to the contestant's optimal outcome. Any strategy other than the i and j found above would result in increasing the expected probability for the contestant when she plays her optimal strategy. Thus the host must use only strategies i and j. We find that $p_j = (b_i - a_i)/(a_j - b_j + b_i - a_i)$, by using calculations similar to the above for x_I. The value of p_i is then $1 - p_j$, since the other p's are zero.

For the parameters $N_L = 4$, $N_R = 7$, and $H = 5$, of Example 1, we have $sH = \frac{N_L H}{N_L + N_R} = 20/11$, and thus $i = 1$ and $j = 2$. Using the corresponding a's and b's gives $x_I = 7/15$, $y_I = 28/165$, $p_1 = 1/5$, and $p_2 = 4/5$, as previously found by the graphical method.

As another example, Example 3, let $N_L = 5$, $N_R = 3$, and $H = 4$. Table 3 summarizes the possible pure strategies.

TABLE 3. Example 3 with $N_L = 5$, $N_R = 3$, and $H = 4$.

	$i = 2$	$i = 3$	$i = 4$	Row Min
P_L	$\frac{5}{24} \cong 0.208$	$\frac{5}{16} \cong 0.313$	$\frac{5}{8} = 0.625$	0.208
P_R	$\frac{3}{8} = 0.375$	$\frac{3}{16} \cong 0.188$	$\frac{1}{8} = 0.125$	0.125
Column Max	0.375	0.313	0.625	

If we play Game III using the parameters of Example 3, then $sH = 20/8$, $i = 2$, and $j = 3$. Substituting the a's and b's gives $x_I = 9/14$, $y_I = 15/56 \cong 0.268$, $p_2 = 3/7$, and $p_3 = 4/7$. Thus, using her optimal mixed strategy, the expected probability that the contestant wins the car is 0.268.

If we play Game II with the parameters of Example 3, the contestant picks L and the host picks $i = 3$. In this case, the contestant has probability 0.313 of winning the car, and she does better if both players use pure strategies rather than mixed strategies.

Closing remarks When the popularity of Marilyn's problem was at its peak, it was used in many classrooms as a Monte Carlo simulation study. Two papers describing these simulations **[5]**, **[11]** found that most solvers initially chose the incorrect solution S_1 (no switch), and most stuck to S_1 even after the simulation indicated that they should switch to the other door. The Monte Carlo simulation of Games I, II, and III of our paper give the expected results and would make a challenging assignment for a computer programming class.

We have tacitly assumed that the contestant prefers the car rather than the goat. This assumption was called into question in a letter from Lore Segal to the editor of *The New York Times* **[9]**, with the comment, "The goat is a delightful animal, although parking might be a problem."

Acknowledgment. My thanks to David Rader, Jack Kinney, and a referee for their helpful suggestions.

REFERENCES

1. E. Barbeau, Fallacies, flaws, and flimflam, *The College Mathematics Journal*, 24 (1993), 149–154 and 26 (1995), 132–134.
2. L. Brickman, *Mathematical Introduction to Linear Programming and Game Theory*, Springer-Verlag, New York, NY, 1989.
3. E. Engel and A. Venetoulias, Monty Hall's probability puzzle, *Chance*, 4 (1991), 6–9.
4. L. Fernandez and R. Piron, Should she switch? A game-theoretic analysis of the Monty Hall problem, *Mathematics Magazine*, 72 (1999), 214–217.
5. D. Friedman, Monty Hall's three doors: construction and deconstruction of a choice anomaly, *The American Economic Review*, 88 (1998), 933–946.
6. M. Gardner, Mathematical Games, *Scientific American*, 201 (1959), 180–182.
7. L. Gillman, The car and the goats, *Amer. Mathematical Monthly*, 99 (1992), 3–7.
8. J. Morgan *et al*, Let's make a deal: The players dilemma, *The American Statistician*, 45 (1991), 284–287.
9. L. Segal, Letters to the editor, *New York Times*, August 16, 1991.
10. S. Selvin, A problem in probability, "Letters to the Editor," *The American Statistician*, 29 (1975), 67 and 134.
11. J. M. Shaughnessy and T. Dick, Monty's dilemma: should you stick or switch?, *Mathematics Teacher* (1991), 252–256.
12. M. vos Savant, Ask Marilyn, *Parade Magazine*, September 9, 1990, December 2, 1990, February 17, 1991.
13. H. Taha, *Operations Research: An Introduction*, Prentice-Hall, New York, NY, 1996.

A Convergence Theorem for the Riemann Integral

RUSSELL A. GORDON
Whitman College
Walla Walla, WA 99362

Let $\{f_n\}$ be a sequence of real-valued functions that converges pointwise to a function f on a closed and bounded interval $[a, b]$ and suppose that each of the functions f_n is Riemann integrable on $[a, b]$. Does it follow that the limit function f is Riemann integrable on $[a, b]$? If f is Riemann integrable on $[a, b]$, is the equation

$$\int_a^b f = \lim_{n\to\infty} \int_a^b f_n$$

valid? Since pointwise convergence is a rather mild condition, it is not difficult to construct examples (see the next paragraph) to show that each of these questions has a negative answer. The goal of this paper is to find further conditions to place on the sequence $\{f_n\}$ in order to obtain positive results.

To show that the pointwise limit of a sequence of Riemann integrable functions may not be Riemann integrable, let $\{r_k\}$ be a listing of the rational numbers in $[0, 1]$ and define ϕ_n and ϕ on $[0, 1]$ by

$$\phi_n(x) = \begin{cases} 1, & \text{if } x = r_1, r_2, \ldots, r_n; \\ 0, & \text{otherwise;} \end{cases} \quad \text{and} \quad \phi(x) = \begin{cases} 1, & \text{if } x \text{ is rational;} \\ 0, & \text{if } x \text{ is irrational.} \end{cases}$$

For each positive integer n, the function ϕ_n is Riemann integrable on $[0, 1]$ since it has only a finite number of discontinuities, but the limit function ϕ is not Riemann integrable on [0, 1]. As a second example, define h_n and h on [0, 1] by

$$h_n(x) = \begin{cases} n, & \text{if } 0 < x < 1/n; \\ 0, & \text{otherwise;} \end{cases} \quad \text{and} \quad h(x) = 0.$$

In this case, the limit function h is Riemann integrable on [0, 1] but

$$0 = \int_0^1 h \neq \lim_{n \to \infty} \int_0^1 h_n = 1.$$

These two examples indicate that a convergence theorem for the Riemann integral will require some condition in addition to pointwise convergence.

The simplest convergence theorem for the Riemann integral involves the notion of uniform convergence. In order for $\{f_n\}$ to converge pointwise to f on $[a, b]$, the sequence $\{f_n(x)\}$ must converge to $f(x)$ for each $x \in [a, b]$. However, the rate of convergence may vary with x (this should be evident for the sequences $\{\phi_n\}$ and $\{h_n\}$) and this variability can create difficulties with limit operations. If the rate of convergence is independent of x, the convergence is said to be uniform. More formally, the sequence $\{f_n\}$ converges uniformly to a function f on the interval $[a, b]$ if for each $\epsilon > 0$ there exists a positive integer N such that $|f_n(x) - f(x)| < \epsilon$ for all $x \in [a, b]$ whenever $n \geq N$. Uniform convergence yields the following theorem; the proof is not difficult and can be found in **[2]** or almost any other introductory textbook in real analysis.

THEOREM 1. *Let $\{f_n\}$ be a sequence of Riemann integrable functions defined on $[a, b]$. If $\{f_n\}$ converges uniformly to f on $[a, b]$, then f is Riemann integrable on $[a, b]$ and$\int_a^b f = \lim_{n \to \infty} \int_a^b f_n$.*

Uniform convergence is a sufficient condition for a convergence theorem for the Riemann integral, but it is by no means a necessary condition. An enlightening example is the following: let $\{c_n\}$ be any sequence of real numbers and for each positive integer n, define ψ_n on [0, 1] by

$$\psi_n(x) = \begin{cases} c_n \sin(n\pi x), & \text{if } 0 \leq x \leq 1/n; \\ 0, & \text{if } x > 1/n. \end{cases}$$

The graph of ψ_n is one hump of a sine wave with amplitude $|c_n|$ and period $2/n$. This sequence converges pointwise to the zero function on [0, 1], and it is not difficult to see that the convergence is uniform if and only if $\{c_n\}$ converges to 0. Since

$$\int_0^1 \psi_n = \int_0^{1/n} c_n \sin(n\pi x)\, dx = \frac{c_n}{n\pi} \int_0^{\pi} \sin\theta\, d\theta = \frac{2c_n}{n\pi},$$

the sequence of integrals can converge to 0 even if $\{c_n\}$ does not converge to 0. In particular, the sequence of integrals converges to 0 if $\{c_n\}$ is bounded. (It is also possible for the sequence $\left\{\int_0^1 \psi_n\right\}$ to converge to 0 when $\{c_n\}$ is unbounded or for the sequence of integrals to be bounded but not convergent.) Although the convergence is not uniform on [0, 1] for some choices of $\{c_n\}$, the convergence is uniform on $[a, 1]$ for each $0 < a < 1$ no matter what sequence is chosen for $\{c_n\}$. In other words, the convergence still exhibits a high degree of uniformity.

Using some elementary properties of the Riemann integral, the uniform convergence result can be extended to sequences that do not converge uniformly but for which the convergence is close to being uniform. Recall the following two results:

1. If f is bounded on $[a, b]$ and Riemann integrable on every closed subinterval of (a, b), then f is Riemann integrable on $[a, b]$.
2. If f is Riemann integrable on the intervals $[a, c]$ and $[c, b]$, then f is Riemann integrable on $[a, b]$ and $\int_a^b f = \int_a^c f + \int_c^b f$.

An additional concept that is required is that of a uniformly bounded sequence of functions. A sequence $\{f_n\}$ is uniformly bounded on $[a, b]$ if there is a number M such that $|f_n(x)| \le M$ for all $x \in [a, b]$ and for all positive integers n. For the record, it is a routine exercise to prove that a uniformly convergent sequence of bounded functions is uniformly bounded.

THEOREM 2. *Let $\{f_n\}$ be a sequence of Riemann integrable functions that converges pointwise to a function f on $[a, b]$. If $\{f_n\}$ converges uniformly to f on each closed subinterval of (a, b) and $\{f_n\}$ is uniformly bounded on $[a, b]$, then f is Riemann integrable on $[a, b]$ and $\int_a^b f = \lim_{n\to\infty} \int_a^b f_n$.*

Proof. Let M be a uniform bound for the sequence $\{f_n\}$ on $[a, b]$ and note that f is bounded by M as well. Since the convergence is uniform on each closed subinterval of (a, b), Theorem 1 shows that the function f is Riemann integrable on each closed subinterval of (a, b). By property (1) listed above, the function f is Riemann integrable on $[a, b]$.

Let $\epsilon > 0$. Choose points $c, d \in (a, b)$ such that $c < d$, $c - a < \epsilon/2M$, and $b - d < \epsilon/2M$. Since $\{f_n\}$ converges uniformly to f on $[c, d]$, by Theorem 1 there exists a positive integer N such that $\left|\int_c^d f_n - \int_c^d f\right| < \epsilon$ for all $n \ge N$. Then

$$\begin{aligned}\left|\int_a^b f_n - \int_a^b f\right| &\le \left|\int_a^c (f_n - f)\right| + \left|\int_c^d f_n - \int_c^d f\right| + \left|\int_d^b (f_n - f)\right| \\ &< 2M(c-a) + \epsilon + 2M(b-d) \\ &< \epsilon + \epsilon + \epsilon = 3\epsilon\end{aligned}$$

for all $n \ge N$ and it follows that $\int_a^b f = \lim_{n\to\infty} \int_a^b f_n$.

THEOREM 3. *Let $\{f_n\}$ be a sequence of Riemann integrable functions that converges pointwise to a function f on $[a, b]$ and let $a = c_0 < c_1 < \cdots < c_{q-1} < c_q = b$ be a partition of $[a, b]$. If $\{f_n\}$ converges uniformly to f on each closed subinterval of (c_{i-1}, c_i) for $1 \le i \le q$ and $\{f_n\}$ is uniformly bounded on $[a, b]$, then f is Riemann integrable on $[a, b]$ and $\int_a^b f = \lim_{n\to\infty} \int_a^b f_n$.*

Proof. The hypotheses of Theorem 2 are satisfied on each of the intervals $[c_{i-1}, c_i]$ so f is Riemann integrable on $[c_{i-1}, c_i]$ and

$$\int_{c_{i-1}}^{c_i} f = \lim_{n\to\infty} \int_{c_{i-1}}^{c_i} f_n$$

for $i = 1, 2, \ldots, q$. By repeated application of property (2) listed above for the Riemann integral, the function f is Riemann integrable on $[a, b]$ and

$$\int_a^b f = \sum_{i=1}^q \int_{c_{i-1}}^{c_i} f = \sum_{i=1}^q \lim_{n\to\infty} \int_{c_{i-1}}^{c_i} f_n = \lim_{n\to\infty} \sum_{i=1}^q \int_{c_{i-1}}^{c_i} f_n = \lim_{n\to\infty} \int_a^b f_n.$$

This completes the proof.

As an application of this convergence theorem, let p be a positive integer and define a sequence $\{g_n\}$ of functions by $g_n(x) = \cos^{2n}(p!\pi x)$. This sequence converges pointwise on [0, 1] to the function g defined by

$$g(x) = \begin{cases} 1, & \text{if } x = i/p! \text{ for } i = 0, 1, 2, \dots, p!; \\ 0, & \text{otherwise.} \end{cases}$$

Each of the functions g_n is Riemann integrable on [0, 1] since it is a continuous function, the convergence is uniform on each closed subinterval of $((i-1)/p!, i/p!)$ for $1 \le i \le p!$, and the sequence $\{g_n\}$ is clearly uniformly bounded on [0, 1]. By Theorem 3, the function g is Riemann integrable on [0, 1] (note also that g is bounded and has only a finite number of discontinuities) and the sequence $\{\int_0^1 g_n\}$ converges to $\int_0^1 g = 0$. Using the periodicity of the cosine function and a standard reduction formula from calculus, we obtain

$$\begin{aligned} \int_0^1 g_n &= \int_0^1 \cos^{2n}(p!\pi x)\,dx = 2p!\int_0^{1/2p!} \cos^{2n}(p!\pi x)\,dx \\ &= \frac{2}{\pi}\int_0^{\pi/2} \cos^{2n}\theta\,d\theta = \frac{(2n)!}{4^n (n!)^2}. \end{aligned}$$

There are other ways (more elementary ways) to prove that this sequence converges to 0, but it is more difficult than it appears at first glance.

Theorem 3 covers almost all of the situations that might occur in applications, but the convergence of $\{f_n\}$ to f can be much more complicated than the examples that have been presented thus far. In fact, it is possible for a sequence of continuous functions to converge pointwise to a continuous function but not converge uniformly to that function on any subinterval. For instance, for each positive integer n, let

$$s_n(x) = \sum_{k=1}^{n} \frac{1}{k!}\sqrt{2e}\, n \sin^2(\pi k! x)\exp\left(-n^2 \sin^4(\pi k! x)\right).$$

Then the sequence $\{s_n\}$ converges pointwise on [0, 1] to the zero function but the convergence is not uniform in any subinterval of [0, 1]. This example is due to Osgood **[5]** and the reader can refer to the original paper for the fairly complicated details. It is enlightening to use a computer algebra system to graph several of the functions s_n.

To prove a convergence theorem that includes sequences whose convergence is far from uniform requires a new approach. The first difficulty lies in the fact that in the absence of uniform convergence, the limit function may not be Riemann integrable (see the sequence $\{\phi_n\}$ presented earlier). Consequently, the Riemann integrability of the limit function must become part of the hypothesis. Other than the fact that the statement of the theorem is then less aesthetically pleasing, this is no great loss as there are easy ways to check that a function is Riemann integrable (for instance, show that it is bounded and has only a finite number of discontinuities). We are thus led to the following convergence theorem which is usually called the Bounded Convergence Theorem since all of the functions have a common bound.

BOUNDED CONVERGENCE THEOREM. *If $\{f_n\}$ is a uniformly bounded sequence of Riemann integrable functions that converges pointwise on $[a, b]$ to a Riemann integrable function f, then $\int_a^b f = \lim_{n\to\infty}\int_a^b f_n$.*

Assuming the hypotheses of the Bounded Convergence Theorem, the sequence $\{|f_n - f|\}$ is a uniformly bounded sequence of nonnegative Riemann integrable functions that converges pointwise on $[a, b]$ to the zero function. If the theorem can be proved for this special case, then the inequality

$$\left|\int_a^b f_n - \int_a^b f\right| \le \int_a^b |f_n - f|$$

indicates that the general result is valid as well. Consequently, it is only necessary to prove this special case of the Bounded Convergence Theorem. The first published proofs of this theorem were given by Arzela in 1885 and independently by Osgood **[6]** in 1897. (For a good summary of the history of the proof of this theorem, see **[4]**.) The first step in these early proofs is to note that the Bounded Convergence Theorem is actually equivalent to a result about sets of points and the concept of a figure plays an important role in this equivalence. A *figure* is a finite union of non-overlapping intervals and the length of a figure is the sum of the lengths of the intervals in the figure. Given a figure V, let $l(V)$ denote its length.

THEOREM 4. *The following are equivalent*:

1. *If* $\{f_n\}$ *is a uniformly bounded sequence of nonnegative Riemann integrable functions that converges pointwise on* $[a,b]$ *to the zero function, then the sequence* $\{\int_a^b f_n\}$ *converges to* 0.
2. *If* $\{V_n\}$ *is a sequence of figures in* $[a,b]$ *such that* $l(V_n) > \delta > 0$ *for all* n, *then there exists a point* z *in* $[a,b]$ *that belongs to infinitely many of the figures* V_n.

Proof. Suppose that (1) holds and let $\{V_n\}$ be a sequence of figures in $[a,b]$ such that $l(V_n) > \delta$ for all n. For each n, let f_n be the characteristic function of the figure V_n. (That is, let $f_n(x) = 1$ if $x \in V_n$ and $f_n(x) = 0$ if $x \notin V_n$.) Each f_n is Riemann integrable on $[a,b]$ since it is a step function and the sequence $\{f_n\}$ is clearly uniformly bounded on $[a,b]$. Suppose there is no point z in $[a,b]$ that belongs to infinitely many of the figures V_n. Then the sequence $\{f_n\}$ converges pointwise on $[a,b]$ to the zero function and (1) implies that the sequence $\{\int_a^b f_n\}$ converges to 0. This contradicts the fact that $\int_a^b f_n = l(V_n) > \delta$ for each n. Therefore, there exists a point z in $[a,b]$ that belongs to infinitely many of the figures V_n.

Now suppose that (2) holds and let $\{f_n\}$ be a uniformly bounded sequence of nonnegative Riemann integrable functions that converges pointwise on $[a,b]$ to the zero function. Suppose that the sequence $\{\int_a^b f_n\}$ does not converge to 0. (This includes the possibility that the sequence does not converge.) By considering a subsequence if necessary, we may assume that there exists a positive number η such that $\int_a^b f_n > 2\eta(b-a)$ for all indices n. Now fix n and choose a partition $a = c_0 < c_1 < \cdots < c_{q-1} < c_q = b$ of $[a,b]$ such that

$$\sum_{i=1}^{q} m(f_n, J_i) l(J_i) > 2\eta(b-a),$$

where $J_i = [c_{i-1}, c_i]$ for $1 \le i \le q$ and $m(f_n, J_i) = \inf\{f_n(x) : x \in J_i\}$. (This represents a set of inscribed rectangles that is close to the "area" under the curve.) Let $S_0 = \{i : m(f_n, J_i) < \eta\}$ and $S_1 = \{i : m(f_n, J_i) \ge \eta\}$ and define a figure by $V_n = \cup_{i \in S_1} J_i$. Note that $f_n(x) \ge \eta$ for all $x \in V_n$. Let M be a uniform bound for the sequence $\{f_n\}$ and compute

$$\begin{aligned} 2\eta(b-a) &< \sum_{i=1}^{q} m(f_n, J_i)\, l(J_i) \\ &= \sum_{i \in S_0} m(f_n, J_i)\, l(J_i) + \sum_{i \in S_1} m(f_n, J_i)\, l(J_i) \\ &\le \sum_{i \in S_0} \eta\, l(J_i) + \sum_{i \in S_1} M\, l(J_i) \\ &\le \eta(b-a) + M\, l(V_n). \end{aligned}$$

It follows that $l(V_n) > \delta$ where $\delta = \eta(b-a)/M$. Since n was an arbitrary positive integer, this process generates a sequence $\{V_n\}$ of figures in $[a,b]$ such that $l(V_n) > \delta$ for all n. Since (2) holds, there exists a point z in $[a,b]$ that belongs to infinitely many of the figures V_n. This implies that $f_n(z) \geq \eta$ for infinitely many n, a contradiction to the fact that $\{f_n(z)\}$ converges to 0. We conclude that the sequence $\left\{\int_a^b f_n\right\}$ converges to 0.

Thus to prove the Bounded Convergence Theorem, a proof of (2) is needed and it is at this point that the proofs of Arzela and Osgood become difficult and tedious. However, an elementary proof of (2) has been given by Lewin **[3]**. Although his proof involves only elementary ideas, a fair number of "obvious but tedious to prove" facts about figures are required. Other methods for proving the Bounded Convergence Theorem have also been discovered; the interested reader should consult **[4]**.

We conclude this paper with two remarks, one applied and one theoretical. In some applications, the key mathematical step is to express a function as a trigonometric series. For example, suppose that f is an odd function on the interval $[-\pi, \pi]$ and a sequence $\{b_k\}$ is required so that the equation $f(x) = \sum_{k=1}^{\infty} b_k \sin(kx)$ is valid for all x in the interval $(-\pi, \pi)$. The standard procedure to find the coefficient b_p is to multiply both sides of the equation by $\sin(px)$ then integrate over the interval $[-\pi, \pi]$:

$$\begin{aligned}
\int_{-\pi}^{\pi} f(x)\sin(px)\,dx &= \int_{-\pi}^{\pi} \left(\sum_{k=1}^{\infty} b_k \sin(kx)\sin(px) \right) dx \\
&= \sum_{k=1}^{\infty} \int_{-\pi}^{\pi} b_k \sin(kx)\sin(px)\,dx \\
&= \int_{-\pi}^{\pi} b_p \sin^2(px)\,dx \\
&= \pi\, b_p .
\end{aligned}$$

The second equality, the interchange of the integral and the infinite sum, is the crucial step and it is at this point that a convergence theorem is required. Since trigonometric series do not necessarily converge uniformly, a more advanced convergence theorem is required. The reader can consult [1] for an application of the bounded convergence theorem that does not involve trigonometric series.

The Bounded Convergence Theorem for the Riemann integral gets very little attention these days because the Riemann integral is no longer the integral of choice in the theory of integration. It has been replaced by the Lebesgue integral since this integral overcomes a number of the deficiencies of the Riemann integral. For instance, the limit of a uniformly bounded sequence of Lebesgue integrable functions is necessarily Lebesgue integrable. The Bounded Convergence Theorem for the Lebesgue integral is quite easy to prove and the corresponding result for the Riemann integral follows as a corollary. However, a fair amount of mathematical sophistication is required to understand the definition of the Lebesgue integral. The first step in the definition is the development of a theory of measure for sets of real numbers and statement (2) in Theorem 4 is a simple consequence of the properties of Lebesgue measure. Nevertheless, a discussion of the Bounded Convergence Theorem for the Riemann integral can enrich an undergraduate course in real analysis. The students will get to see a useful convergence theorem, recognize some of the deficiencies of the Riemann integral, and spend some time thinking about sets of points. The equivalence stated in Theorem 4 indicates the strong relationship between integration and properties of sets of points. All of these ideas will help pave the way for a graduate course in analysis. In addition, working through the details of the proof would make a good project for an undergraduate who is interested in real analysis.

REFERENCES

1. L. R. Bragg and J. W. Grossman, An application of the dominated convergence theorem to mathematical statistics, this MAGAZINE 56 (1983), 41–42.
2. R. A. Gordon, *Real Analysis, A First Course*, Addison-Wesley, Reading, MA, 1997.
3. J. W. Lewin, A truly elementary approach to the bounded convergence theorem, *Amer. Math. Monthly* 93 (1986), 395–397.
4. W. A. J. Luxemburg, Arzela's dominated convergence theorem for the Riemann integral, *Amer. Math. Monthly* 78 (1971), 970–979.
5. W. F. Osgood, A geometrical method for the treatment of uniform convergence and certain double limits, *Bull. Amer. Math. Soc.* 3 (1896), 59–86.
6. W. F. Osgood, Non-uniform convergence and the integration of series term by term, *Amer. J. Math.* 19 (1897), 155–190.

Matrices, Continued Fractions, and Some Early History of Iteration Theory

MICHAEL SORMANI
College of Staten Island, CUNY
Staten Island, NY 10314-6600

Introduction In a 1995 paper in this MAGAZINE, Marafino and McDevitt **[14]** discuss a variety of techniques at the advanced undergraduate level for studying the convergence of the continued fraction

$$\cfrac{1}{1+\cfrac{c}{1+\cfrac{c}{1+\ddots}}} \tag{1}$$

The authors show how ideas from analysis, algebra, number theory, topology, and complex variables can all be used to determine those complex numbers c for which the above continued fraction converges. They prove that (1) converges for all complex numbers c except those on the real line to the left of $-\frac{1}{4}$. In this note, we examine how still another area of undergraduate mathematics—linear algebra—can also be employed to prove this result.

The linear algebra approach to this problem has an interesting history dating back to the very first paper **[8]** on matrix theory, published by Arthur Cayley in 1858. Other historical connections reach backward in time to the early work of Charles Babbage, the designer of the first large scale mechanical computer; and forward in time to the modern theory of dynamical systems. The problem is thus a link in a chain of ideas extending over almost 200 years. Some details of this history appear at the end of this note.

Möbius transformations and associated matrices Following the lead of **[14]**, we consider the Möbius (or linear fractional) transformation

$$f_c(z) = \frac{1}{cz+1};$$

the convergence properties of the infinite continued fraction (1) can be found by studying the sequence of iterates $f_c(0)$, $f_c(f_c(0)) = f_c^2(0)$, $f_c(f_c(f_c(0))) = f_c^3(0), \dots$. It is a straightforward exercise to show that the set of transformations of the form, $f(z) = \frac{az+b}{cz+d}$, where $ad - bc \neq 0$, forms a group under composition. (All coefficients are complex numbers.) If $f(z) = \frac{az+b}{cz+d}$ and $g(z) = \frac{Az+B}{Cz+D}$, then the composite function can be written as

$$f(g(z)) = \frac{(aA+bC)z + (aB+bD)}{(cA+dC)z + (cB+dD)}. \tag{2}$$

The coefficients in (2) suggest that we associate the matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with the Möbius transformation $f(z) = \frac{az+b}{cz+d}$. Note that, if $cz + d \neq 0$, we can write

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} z \\ 1 \end{pmatrix} = \begin{pmatrix} az+b \\ cz+d \end{pmatrix} = (cz+d)\begin{pmatrix} \dfrac{az+b}{cz+d} \\ 1 \end{pmatrix}.$$

Since the composition of Möbius transformations can be accomplished by multiplying their corresponding matrices, we can calculate the iterates of $f(z)$ by finding powers of its associated matrix M, as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^n \begin{pmatrix} z \\ 1 \end{pmatrix} = k_n \begin{pmatrix} f^n(z) \\ 1 \end{pmatrix},$$

where k_n is a normalizing factor that depends on z. Again, we assume that we have not encountered a singular point of $f(z)$ while doing the iterations.

Application of linear algebra We are now ready to apply some elementary theory from linear algebra to solve our original problem. Consider the matrix $M_c = \begin{pmatrix} 0 & 1 \\ c & 1 \end{pmatrix}$, corresponding to the transformation $f_c(z) = \frac{1}{cz+1}$. If $c \neq 0$, then 0 is not an eigenvalue of M_c and $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is not an eigenvector. Hence we can find the eigenvectors and eigenvalues of M_c by examining

$$\begin{pmatrix} 0 & 1 \\ c & 1 \end{pmatrix}\begin{pmatrix} z \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ cz+1 \end{pmatrix} = (cz+1)\begin{pmatrix} \dfrac{1}{cz+1} \\ 1 \end{pmatrix}.$$

We see immediately that $\vec{v} = \begin{pmatrix} z \\ 1 \end{pmatrix}$ is an eigenvector of M_c if and only if $z = f_c(z)$; that is, z is a fixed point of the mapping $f_c(z)$. For such a value of z, the corresponding eigenvalue is $\lambda = cz + 1$.

Solving $z = \frac{1}{cz+1}$ yields $z = \frac{-1 \pm \sqrt{1+4c}}{2c}$ and $\lambda = \frac{1 \pm \sqrt{1+4c}}{2}$. It follows that the two eigenvalues of M_c, λ_1 and λ_2, are of equal magnitude if and only if $\sqrt{1+4c}$ has zero real part, i.e., if and only if c is real and $1 + 4c \leq 0$. There are three cases to consider:

Case 1: Eigenvalues of unequal magnitude. If $1 + 4c$ is either a positive real number or a complex number with nonzero real part, then the two eigenvalues of M_c differ in magnitude. Suppose that $|\lambda_1| > |\lambda_2|$ and that the associated eigenvectors are

$\vec{v}_1$ and $\vec{v}_2$. Following the standard power method, we apply powers of M_c to $\vec{w}_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Since $\vec{w}_0 = c_1\vec{v}_1 + c_2\vec{v}_2$ for some constants c_1 and c_2, we get $M_c^n\vec{w}_0 = k_n\begin{pmatrix} f_c^n(0) \\ 1 \end{pmatrix} = \lambda_1^n(c_1\vec{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^n c_2\vec{v}_2)$. Since $|\lambda_2| < |\lambda_1|$ in this case, it follows that $\left(\frac{1}{\lambda_1^n}\right) M_c^n\vec{w}_0$ converges to $c_1\vec{v}_1$. This, in turn, implies that the iterates $f_c^n(0)$ converge to the fixed point of $f_c(z)$ that maximizes $|\lambda| = |cz + 1|$, the magnitude of the eigenvalue.

Case 2: Equal eigenvalues. If $1 + 4c = 0$, then $\lambda_1 = \lambda_2 = \frac{1}{2}$. There is only one independent eigenvector in this case, namely $\vec{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. We note that the vector $\vec{v}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ obeys $(M_c - \frac{1}{2}I)\vec{v}_2\}\vec{v}_2 = \vec{v}_1$, which implies that

$$M_c^n\vec{v}_2 = \left(\frac{1}{2^n}\right)\vec{v}_2 + \left(\frac{n}{2^{n-1}}\right)\vec{v}_1.$$

Since the vector $\vec{w}_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ can be written as a linear combination of $\vec{v}_1$ and $\vec{v}_2$, it follows in this case also that the iterates $f_c^n(0)$ converge to the fixed point of $f_c(z)$, which is $z = 2$.

Case 3: Unequal eigenvalues of equal magnitude. If c is a real number obeying $1 + 4c < 0$, then we have distinct complex eigenvalues that are complex conjugates of each other. Rewriting the eigenvalues in polar form, we have $\lambda_1 = re^{i\alpha}$ and $\lambda_2 = re^{-i\alpha}$. The vector $\vec{w}_0$ is a linear combination of both eigenvectors; i.e., $\vec{w}_0 = c_1\vec{v}_1 + c_2\vec{v}_2$ for some nonzero constants c_1 and c_2. From

$$M_c^n\vec{w}_0 = k_n\begin{pmatrix} f_c^n(0) \\ 1 \end{pmatrix} = r^n\left(e^{in\alpha}c_1\vec{v}_1 + e^{-in\alpha}c_2\vec{v}_2\right),$$

we can see that the iterates $f_c^n(0)$ remain bounded but do not converge. It should also be noted that if $e^{i\alpha}$ is a kth root of unity, then we will get periodic behavior for the iterates, where $f_c^{n+k}(0) = f_c^n(0)$.

This completes the proof that the complex continued fraction (1) converges for all complex numbers c except for those real c for which $1 + 4c < 0$.

Historical notes The long history of this problem involves the inventor Charles Babbage (1792–1871); the astronomer John Herschel (1792–1871); the logician George Boole (1815–1864); and the mathematician Arthur Cayley (1821–1895), who wrote extensively about matrices and first defined abstract groups.

In 1813, Babbage and Herschel, then mathematics students at Cambridge, edited the first and only issue of the *Memoirs of the Analytical Society* [**4**]. This journal contained three anonymous articles (written by the editors), including one on iteration of Möbius transformations. Although Herschel has been credited [**11**] with the authorship of this article, Babbage was also interested in the subject, and published two papers on functional equations soon afterwards [**2**], [**3**]. These papers deal with questions such as finding functions f for which the nth iterate of f (i.e., the composite function f^n formed by composing f with itself $n - 1$ times) equals itself. Boole, in his 1844 book [**7**] on finite differences, refers repeatedly to Babbage's contributions to the "calculus of functions" (see e.g., [**7**, p. 291) and makes specific reference to his work on iterates of functions. Boole [**7**, p. 298] also examines the specific problem of determining which Möbius functions obey $f^n = f$. By the 1830's, Babbage had abandoned his study of pure mathematics and was deeply involved with his pioneering work on computer design.

In 1858 Arthur Cayley wrote the first paper **[8]** on the theory of matrices. This very readable work contains much of today's standard material in undergraduate linear algebra courses. Among its other items of interest is a formula for the nth power of the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. After developing this formula, Cayley remarks that the "... preceding investigations are intimately connected with the investigations of Babbage and others in relation to the function $\phi x = \frac{ax+b}{cx+d}$." In 1880, Cayley returned specifically to this topic in a paper entitled "On the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and in connection therewith the function $\frac{ax+b}{cx+d}$" **[8]**. Again, he mentions the early work of Babbage and refers both to the 1813 paper **[4]** and to Boole's book.

Cayley's work on iteration theory and the questions he raised, particularly those relating to Newton's method, are usually regarded (see, e.g., **[1]**, **[15]**) as among the primary factors leading to the development by Fatou and Julia of the theory of dynamical systems in the early 20th century. Thus it appears that the study of functional iterations, which has recently become so popular as a result of beautiful computer graphics, can be traced back through Cayley and Boole to Babbage, who designed the first large computer.

Several good references (**[5]**, **[6]**, **[12]**, **[13]**) deal specifically with Möbius transformations. Hans Schwerdtfeger's excellent book **[16]** contains most of the material used above, and much more. Further historical references include **[1]**, **[10]**, **[11]**, and **[15]**.

Acknowledgment. The author is grateful to the referees for their suggestions.

REFERENCES

1. D. Alexander, *A History of Complex Dynamics: From Schroder to Fatou and Julia*, Aspects of Math, Vol. E 24, Friedr. Vieweg & Sohn, Braunschweig, Germany, 1994.
2. J. Babbage, An essay toward the calculus of functions (part I), *Collected Works of Babbage*, Vol. 1. 93–123; originally in *Philosophical Transactions of the Royal Society*, 105 (1815).
3. J. Babbage, An essay toward the calculus of functions (part II), *Collected Works of Babbage*, Vol. 1, 124–193; originally in *Philosophical Transactions of the Royal Society*, 106 (1816).
4. J. Babbage and J. Herschel, *The Memoirs of the Analytical Society*, (1813).
5. A. F. Beardon, *The Geometry of Discrete Groups*, Springer-Verlag, New York, NY, 1983.
6. A. F. Beardon, *Iteration of Rational Functions*, Springer-Verlag, New York, NY, 1991.
7. G. Boole, *The Calculus of Finite Differences*, published 1844; reprinted by Chelsea Publishing Co., 4th ed., New York, NY.
8. A. Cayley, A memoir on the theory of matrices, *Collected Works of Cayley*, Vol. 2, 475–496; originally in *Philosophical Transactions* 148 (1858).
9. A. Cayley, On the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and in connection therewith the function $\frac{ax+b}{cx+d}$, *Collected Works*, Vol. 11, 252–257; originally in *Messenger of Mathematics* IX (1880).
10. J. M. Dubbey, *The Mathematical Work of Charles Babbage*, Cambridge University Press, Cambridge, UK, 1978.
11. P. C. Enros, The Analytical Society (1812–1813): Precursor of the renewal of Cambridge mathematics, *Historia Mathematica* 10 (1983), 24–47.
12. L. Hahn, *Complex Numbers and Geometry*, MAA Spectrum, Mathematical Association of America, Washington, DC, 1994.
13. G. A. Jones and D. Singerman, *Complex Functions: An Algebraic and Geometric Viewpoint*, Cambridge University Press, Cambridge, UK, 1987.
14. J. Marafino and T.J. McDevitt, Convergence of complex continued fractions, this MAGAZINE 68 (1995), 202–208.
15. H. O. Peitgen, D. Saupe, and F. v. Haeseler, Cayley's problem and Julia sets, *Mathematical Intelligencer* 6 (1984), 11–20.
16. H. Schwerdtfeger, *Geometry of Complex Numbers*, Dover, New York, NY, 1979.

From 30 to 60 Is Not Twice as Hard

MICHAEL DALEZMAN
Yeshiva University
New York, NY 10016

Euclid's proof that there are infinitely many primes [3, p. 25] can be modified to yield a proof of a simple inequality: If $p_1, p_2, \dots$ is the sequence of prime numbers and $n \geq 2$ then $p_1 p_2, \dots p_n > p_{n+1}$. In 1907, Bonse [1] gave an elementary proof of a stronger inequality, now called Bonse's Inequality [4]: If $n \geq 4$ then $p_1 p_2 \dots, p_n > p_{n+1}^2$. Bonse then used his inequality to prove that 30 is the largest integer m with the following property:

$$\text{If } 1 < k < m \quad \text{and} \quad (k, m) = 1, \quad \text{then } k \text{ is a prime.}$$

In this note we give an elementary proof of a stronger inequality and use it to prove that 60 is the largest integer m with the following property:

$$\text{If } 1 < k < m \qquad \text{and} \quad (k, m) = 1, \quad \text{then} \quad k \quad \text{is a prime power.}$$

We then indicate how the inequality can be strengthened and how the result can be generalized.

The following notations will be used:

$\omega(k)$: the number of distinct prime divisors of k;
$\Omega(k)$: the number of prime divisors of k, counting multiplicity;
$\lfloor x \rfloor$: the largest integer not greater than x;
$\pi(x)$: the number of primes not exceeding x;
$\phi(k)$: the number of positive integers prime to k and not exceeding k.

THEOREM 1. *If $n \geq 4$, then $p_1 p_2 \cdots p_n > p_{n+1} p_{n+2}$.*

This inequality follows readily from Bertrand's postulate [3, p. 367] but we give here a proof that is self-contained.

Proof. The result can easily be verified for $n < 10$. Let $i = \lceil \frac{n}{2} \rceil$, and suppose

$$p_1 p_2 \cdots p_n \leq p_{n+1} p_{n+2} < p_{n+2}^2.$$

Then

$$(p_1 p_2 \cdots p_i)^2 < p_1 p_2 \cdots p_n < p_{n+2}^2 \quad \text{and} \quad p_1 p_2 \cdots p_i < p_{n+2}.$$

Let us consider the p_i integers $N_t = t p_1 p_2 \cdots p_{i-1} - 1$, $t = 1, 2, \dots, p_i$. For all t, $N_t < p_1 p_2 \cdots p_i < p_{n+2}$ and is prime to $p_1, p_2, \dots, p_{i-1}$. Thus if q_t is the smallest prime dividing N_t, then $p_i \leq q_t < p_{n+2}$. The q_t's are distinct, for if $q_t = q_{t'}$, with $t \neq t'$, then $q_t | N_t - N_{t'} = (t - t')\ p_1 p_2 \cdots p_{i-1}$, so $q_t | t - t'$; this is impossible since $1 \leq t,\ t' \leq p_i$. Hence the number of N_t's must be no greater than the number of primes q such that $p_i \leq q < p_{n+2}$. Therefore $p_i \leq n + 2 - i$. But $i = \lfloor \frac{n}{2} \rfloor$, so $n \leq 2i + 1$ and $p_i \leq i + 3$. The last inequality clearly fails for $i \geq 5$, so it fails for $n \geq 10$. ■

DEFINITION. An integer m satisfies property $\mathscr{P}_s$ if

$$\text{for all } k \text{ such that } 1 < k < m \text{ and } (k, m) = 1, \quad \omega(k) \leq s.$$

LEMMA. *If m satisfies $\mathscr{P}_1$ and $p_n p_{n+1} \leq m$, then $p_1 p_2 \cdots p_n \leq m$.*

Proof. Let m satisfy $\mathscr{P}_1$. Consider all the primes $p_1, p_2, \ldots, p_{n+1}$. If two of these primes, say p_ν and p_μ were both relatively prime to m, we would get $p_\nu p_\mu \leq m$ and $(p_\nu p_\mu, m) = 1$ contradicting the fact that m satisfies $\mathscr{P}_1$. Hence at most one of the primes $p_1, p_2, \ldots, p_{n+1}$ does not divide m, and the lemma follows. ■

MAIN THEOREM. 60 *is the largest integer satisfying $\mathscr{P}_1$.*

Proof. It is easy to verify that 60 satisfies $\mathscr{P}_1$. We need to prove that 60 is the largest such integer. Let $m > 60$ satisfy $\mathscr{P}_1$. If $m \geq 77 = 7 \cdot 11 = p_4 p_5$, we let n be the largest integer such that $p_1 p_2 \cdots p_n \leq m$. By the lemma, $n \geq 4$; by Theorem 1, $p_1 p_2 \cdots p_n > p_{n+1} p_{n+2}$. Hence, $p_{n+1} p_{n+2} < m$ and, by the lemma, $p_1 p_2 \cdots p_{n+1} \leq m$; this contradicts the maximality of n. Thus we must have $60 < m < 77$. Clearly $5 \cdot 7 < m$. By the argument in the proof of the lemma, m must be divisible by all of the primes 2, 3, 5, and 7 with at most one exception. Therefore, m is divisible by 105, by 70, by 42, or 30; the only possibility is 70. But 70 does not satisfy $\mathscr{P}_1$ because 33 is less than 70 and prime to 70, but not a prime power. Hence 60 is the largest integer satisfying $\mathscr{P}_1$. ■

It is noteworthy that the Main Theorem implies Theorem 1. To see why, let $n \geq 4$ and let $a = p_1 p_2 \cdots p_n$. Since $a \geq 210$, a does not satisfy $\mathscr{P}_1$, so there exists an integer b such that $1 \leq b < a$, $(b, a) = 1$ and $\omega(b) \geq 2$. If p and q are 2 distinct primes that divide b, then $p_{n+1} p_{n+2} \leq pq \leq b < a = p_1 p_2 \cdots p_n$.

Generalizations Bonse went further and proved that if $n \geq 5$, then $p_1 p_2 \cdots p_n > p_{n+1}^3$. He used this result to show that 1260 is the largest integer with the property:

$$\text{If } 1 \leq k < m \quad \text{and} \quad (k, m) = 1, \quad \text{then} \quad \Omega(k) \leq 2.$$

Bonse also indicated that similar methods could be used to prove $p_1 p_2 \cdots p_n > p_{n+1}^4$ for sufficiently large n. Using this inequality, he wrote, he had found that 30,030 was the largest integer with the property:

$$\text{If } 1 \leq k < m \quad \text{and} \quad (k, m) = 1 \quad \text{then} \quad \Omega(k) \leq 3.$$

(Actually, Bonse erred; the correct number is 60,060.)

Landau [2] generalized Bonse's results, proving that for every integer $s \geq 1$ there exists an integer n_s such that

$$n \geq n_s \Rightarrow p_1 p_2 \cdots p_n > p_{n+1}^{s+1};$$

he concluded that there exists a largest integer m_s with the property:

$$\text{If} \quad 1 \leq k < m_s \quad \text{and} \quad (k, m_s) = 1, \quad \text{then} \quad \Omega(k) \leq s.$$

Our results too, can be extended and generalized.

THEOREM 2. *For every integer $s \geq 1$, there exists an integer n_s such that*

$$n \geq n_s \Rightarrow p_1 p_2 \cdots p_n > p_{n+1} p_{n+2} \cdots p_{n+s+1}.$$

To prove this, we replace $i = \lfloor \frac{n}{2} \rfloor$ by $i = \lfloor \frac{n}{s+1} \rfloor$ in the proof of Theorem 1; we get

$$p_i \leq si + 2s + 1.$$

The reverse to this inequality will be a consequence of the following 3 lemmas:

LEMMA 1. *For all integers* $a \geq 2$ *and for all* $x > 0$ $\pi(x) \leq \frac{x}{a}\phi(a) + (a-1)$.

This follows from the fact that the interval from ka to $(k+1)a$ (for $k = 1, 2, \dots \lfloor \frac{x}{a} \rfloor$) contains at most $\phi(a)$ primes.

LEMMA 2. *For all* $i \geq 1$ $p_i > i\frac{a}{\phi(a)} - \frac{a^2}{\phi(a)}$.

This is obtained as follows: For any $x > 0$, we let $i = \pi(x) + 1$, then $x < p_i$. We then substitute $i - 1$ for $\pi(x)$ and p_i for x in Lemma 1.

LEMMA 3. $\overline{\lim_{a \to \infty}} \frac{a}{\phi(a)} = +\infty$.

This follows from the fact that

$$\prod_{p \leq x} \frac{p}{p-1} = \prod_{p \leq x} \frac{1}{1 - \frac{1}{p}} = \prod_{p \leq x} \sum_{i=0}^{\infty} \frac{1}{p_i} > \sum_{n \leq x} \frac{1}{n}.$$

Theorem 2 can be used to prove that, for every s, there exists a largest integer m_s satisfying $\mathscr{P}_s$. As in the case $s = 1$, we have the following result:

LEMMA 4. *If* m *satisfies* $\mathscr{P}_s$ *and* $p_n p_{n+1} \cdots p_{n+s} \leq m$, *then* $p_1 p_2 \cdots p_n \leq m$.

To prove the existence of m_s, let m be an integer satisfying $\mathscr{P}_s$, and let us assume that, for some $l \geq n_s$, we have

$$p_1 p_2 \cdots p_{l+1} > m \geq p_1 p_2 \cdots p_l > p_{l+1} p_{l+2} \cdots p_{l+s+1}.$$

By Lemma 4, this implies $p_1 p_2 \cdots p_{l+1} \leq m$; this contradiction proves that $m < p_1 p_2 \cdots p_{n_s}$ and thus establishes the existence of m_s.

Bounds for m_s We will now show that

$$p_1 p_2 \cdots p_{n_s - 1} \leq m_s < p_{n_s} p_{n_s + 1} \cdots p_{n_s + s}.$$

Proof. We saw that $m_s < p_1 p_2 \cdots p_{n_s}$. If $p_{n_s} p_{n_s+1} \cdots p_{n_s+s} \leq m_s$ then Lemma 4 gives $p_1 p_2 \cdots p_{n_s} \leq m_s$. On the other hand, by definition of n_s we have $p_1 p_2 \cdots p_{n_s - 1} < p_{n_s} p_{n_s+1} \cdots p_{n_s+s}$. This shows that $p_1 p_2 \cdots p_{n_s - 1}$ satisfies $\mathscr{P}_s$ and gives the lower bound for m_s. ■

The reader is invited to verify that $n_2 = 6$, $n_3 = 7$, $n_4 = 9$, $m_2 = 2730$, $m_3 = 210{,}210$ and $m_4 = 29{,}099{,}070$.

REFERENCES

1. H. Bonse, Über eine bekannte Eigenschaft der Zahl 30 und ihre Verallgemeinerung, *Archiv der Mathematik und Physik* (3) 12 (1907), 292–295.
2. Edmund Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, vol. I, Chelsea, New York, NY, 1953, pp. 229–234.
3. Ivan Niven, Herbert S. Zuckerman, and Hugh L. Montgomery, *An Introduction to the Theory of Numbers*, 5th Edition, Wiley, New York, NY, 1991.
4. J. V. Uspensky and M. A. Heaslet, *Elementary Number Theory*, McGraw Hill, New York, NY, 1939, p. 87.

Euler's Formula for $\zeta(2k)$, Proved by Induction on k

JI CHUNGANG
CHEN YONGGAO
Nanjing Normal University
Nanjing 210097, Jiangsu
P.R. CHINA

Euler showed in 1735 for any positive integer k that

$$\zeta(2k) = \sum_{n=1}^{\infty} \frac{1}{n^{2k}} = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2(2k)!}, \tag{1}$$

where the Bernoulli numbers B_m $(m = 0, 1, 2, \dots)$ are given by the recurrence relation

$$\sum_{j=0}^{m} \binom{m+1}{j} B_j = 0, \quad m = 1, 2, 3, \dots, \tag{2}$$

with $B_0 = 1$. Many proofs of (1) are known; see, e.g., **[1]**–**[6]**. The purpose of this note is to give a simple proof of (1) using induction on k.

We start with the well-known Fourier expansion

$$\frac{1}{4}t^2 - \frac{\pi}{2}t + \frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{\cos nt}{n^2}, \quad 0 \le t \le 2\pi. \tag{3}$$

Taking $t = 0$ in (3) we obtain (1) in the case $k = 1$. Now suppose that

$$\zeta(2i) = (-1)^{i+1} \frac{(2\pi)^{2i} B_{2i}}{2(2i)!}$$

holds for $1 \le i < k$. From (3) we obtain

$$\int_0^x \int_0^{t_{2k-1}} \cdots \int_0^{t_2} \int_0^{t_1} \left(\frac{1}{4}t^2 - \frac{\pi}{2}t + \frac{\pi^2}{6} \right) dt dt_1 \cdots dt_{2k-1}$$
$$= \int_0^x \int_0^{t_{2k-1}} \cdots \int_0^{t_2} \int_0^{t_1} \sum_{n=1}^{\infty} \frac{\cos nt}{n^2} dt dt_1 \cdots dt_{2k-1},$$

where $0 \le t_i \le 2\pi$ for $i = 1, 2, \dots, 2k-1$ and $0 \le x \le 2\pi$. Carrying out the integration we obtain

$$\frac{x^{2k+2}}{2(2k+2)!} - \frac{\pi x^{2k+1}}{2(2k+1)!} + \frac{\pi^2 x^{2k}}{6(2k)!}$$
$$= (-1)^{k+2} \sum_{n=1}^{\infty} \frac{\cos nx}{n^{2k+2}} + \sum_{i=2}^{k+1} \frac{(-1)^i \zeta(2i)}{(2k+2-2i)!} x^{2k+2-2i}. \tag{4}$$

Putting $x = 2\pi$ in (4) we obtain

$$\zeta(2k) = \frac{(-1)^{k+1}k}{(2k+2)!}(2\pi)^{2k} - \sum_{i=1}^{k-1} \frac{(-1)^{k-i}}{(2k+2-2i)!}2(2\pi)^{2k-2i}\zeta(2i).$$

Appealing to the inductive hypothesis we deduce that

$$\zeta(2k) = \frac{(-1)^{k+1}k}{(2k+2)!}(2\pi)^{2k} - (-1)^{k+1}(2\pi)^{2k}\sum_{i=1}^{k-1} \frac{B_{2i}}{(2k+2-2i)!(2i)!},$$

which can be rewritten as

$$\zeta(2k) = \frac{(-1)^{k+1}(2\pi)^{2k}}{2(2k)!}\frac{1}{(k+1)(2k+1)}\left(k - \sum_{i=1}^{k-1}\binom{2k+2}{2i}B_{2i}\right). \tag{5}$$

As $B_1 = -1/2$ and $B_{2j+1} = 0$ for $j \geq 1$, replacing m by $2k+1$ in (2), we obtain

$$B_{2k} = \frac{1}{(k+1)(2k+1)}\left(k - \sum_{i=1}^{k-1}\binom{2k+2}{2i}B_{2i}\right). \tag{6}$$

The inductive step now follows from equations (5) and (6).

Acknowledgment. This work is supported by the "333 Project" of Jiangsu Province, China. The authors would like to thank the referee for his valuable comments.

REFERENCES

1. T. M. Apostol, Another elementary proof of Euler's formula for $\zeta(2n)$, *Amer. Math. Monthly* 80 (1973), 425–431.
2. T. Estermann, Elementary evaluation of $\zeta(2k)$, *J. London Math. Soc.* 22 (1947), 10–13.
3. K. Knopp, *Theory and Application of Series*, Hafner, New York, NY, 1951.
4. E. C. Titchmarsh, *The Theory of the Riemann Zeta Function*, Oxford U. Pr., Oxford, UK, 1951.
5. G. T. Williams, A new method of evaluating $\zeta(2n)$, *Amer. Math. Monthly* 60 (1953), 19–25.
6. Kenneth S. Williams, On $\sum_{n=1}^{\infty}\frac{1}{n^{2k}}$, this MAGAZINE 44 (1971), 273–276.

From Mathematical World News, a column in *National Mathematics Magazine* (predecessor of *Mathematics Magazine*), April 1937:

> *Mathematische Annalen*, the well-known German journal, was founded in 1868 by Alfred Clebsch and Carl Neumann. It is now edited by David Hilbert ... and is published by the firm of Julius Springer in Berlin.
>
> The Japanese have entered the field of manufacturing slide rules. The Hemmi slide rules, made of bamboo and of laminated construction, are in common use in the United States.
>
> Dr. David Hilbert, professor of mathematics emeritus at the University of Göttingen, celebrated his 75th birthday on January 23, 1937.

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 September 1, 2000.

1594. *Proposed by Kent Holing, Statoil Research Centre, Trondheim, Norway.*

In quadrilateral $ABCD$, $AB + AD = BC + CD$ and $\angle A$ is a right angle. Square $APQR$ has P, Q, and R on segments AB, BD, and AD, respectively, with $AP = BC$.

(a) Find the number of such quadrilaterals (up to congruence) given the lengths BC and BD.

(b) Show how to construct all such triangles in terms of well-known straight-edge-and-compass constructions given the lengths BC and BD.

1595. *Proposed by Wu Wei Chao, Guang Zhou Normal University, Guang Zhou City, Guang Dong Province, China.*

Find all pairs of positive integers a and b such that $ab + a + b$ divides $a^2 + b^2 + 1$.

1596. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

From the vertices $A_0, A_1, \ldots, A_n$ of a simplex S, parallel lines are drawn intersecting the hyperplanes containing the opposite faces in the corresponding points $B_0, B_1, \ldots, B_n$. Determine the ratio of the volume of the simplex determined by $B_0, B_1, \ldots, B_n$ to the volume of S.

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 Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames, IA 50011, or mailed electronically (ideally as a LATEX file) to `johnston@math.iastate.edu`. *Readers who use e-mail should also provide an e-mail address.*

1597. *Proposed by Constantin P. Niculescu, University of Craiova, Craiova, Romania.*

For every $x, y \in (0, \sqrt{\pi/2})$ with $x \neq y$, prove that

$$\ln^2 \frac{1+\sin xy}{1-\sin xy} < \ln \frac{1+\sin x^2}{1-\sin x^2} \cdot \ln \frac{1+\sin y^2}{1-\sin y^2}.$$

1598. *Proposed by Hoe-Teck Wee, student, Massachusetts Institute of Technology, Cambridge, Massachusetts.*

Starting with any n-tuple R_0, $n > 1$, of symbols from A, B, C, we define a sequence $R_0, R_1, R_2, \ldots$ according to the following relation: if $R_j = (x_1, x_2, \ldots, x_n)$, then $R_{j+1} = (y_1, y_2, \ldots, y_n)$, where $y_i = x_i$ if $x_i = x_{i+1}$ (taking $x_{n+1} = x_1$) and y_i is the symbol other than x_i and x_{i+1} if $x_i \neq x_{i+1}$. (For example, if $R_0 = (A, A, B, C)$, then $R_1 = (A, C, A, B)$.)

(a) Find all positive integers $n > 1$ for which there exists some integer $m > 0$ such that $R_m = R_0$ for all R_0.

(b) For $n = 3^k$, $k \geq 1$, find the smallest integer $m > 0$ such that $R_m = R_0$ for every R_0.

Quickies

Answers to the Quickies are on page 162

Q899. *Proposed by Costas Efthimiou, Newman Laboratory of Nuclear Studies, Cornell University, Ithaca, New York, and Florida Southern College, Lakeland, Florida.*

Define the sequence $(a_n)_{n \geq 0}$ by $a_0 = 0$, $-1 < a_1 < 1$, and the recursion

$$a_{n+2} = \frac{a_{n+1} + a_n}{1 + a_n a_{n+1}}.$$

Express a_n in terms of a_1.

Q900. *Proposed by Mircea Radu, Bielefeld University, Bielefeld, Germany, and Institute for Educational Sciences, Bucharest, Romania.*

A given angle measures α degrees. Distinct points $P_0, P_1, \ldots, P_n$ are located on the open rays comprising the angle so that P_k and P_{k+1} are on different rays and

$$P_0P_1 = P_1P_2 = \cdots = P_{n-1}P_n.$$

Find the maximum possible n in terms of α.

Solutions

Triangle with Vertices on Three Concentric Circles April 1999

1569. *Proposed by Ismor Fischer, Department of Biostatistics, University of Wisconsin, Madison, Wisconsin.*

Given three concentric circles in the plane, prove that (up to rotation and reflection) there exists a unique triangle of maximum area having exactly one vertex on each circle, respectively.

Solution by Michael Woltermann, Washington and Jefferson College, Washington, Pennsylvania.

Let $\triangle ABC$ have its vertices A, B, and C on circles with center O and radii a, b, and c, respectively. The triangle of maximal area has orthocenter O lying on or in the interior of $\triangle ABC$.

If the orthocenter of $\triangle ABC$ is not O, some altitude of $\triangle ABC$ fails to pass through O, say the altitude from A. Let A' be on the circle of radius a in the half plane of $\overleftrightarrow{BC}$ containing O (either half plane if O is on $\overleftrightarrow{BC}$) such that $\overleftrightarrow{A'O} \perp \overleftrightarrow{BC}$. Then $d(A', \overleftrightarrow{BC}) > d(A, \overleftrightarrow{BC})$, and the area of $\triangle A'BC$ is greater than the area of $\triangle ABC$. Thus any triangle of maximal area must have orthocenter O. Given A and B, there are at most two choices for C such that $\overleftrightarrow{AO} \perp \overleftrightarrow{BC}$. If the area of $\triangle ABC$ is maximal, then $\overline{AO}$ intersects $\overline{BC}$. We conclude that O is in the interior of or on $\triangle ABC$.

To show uniqueness up to rotation and reflection, we introduce a rectangular coordinate system into the plane so that O is the origin. Without loss of generality assume that $0 < c < b < a$, A is the point $(a, 0)$, and B is in the upper half plane. Then B and C have coordinates $(t, \sqrt{b^2 - t^2})$ and $(t, -\sqrt{c^2 - t^2})$ for some t between $-c$ and 0. The altitude from C to $\overleftrightarrow{AB}$ satisfies

$$y + \sqrt{c^2 - t^2} = \frac{a - t}{\sqrt{b^2 - t^2}}(x - t),$$

and O is on this altitude if and only if $\sqrt{b^2 - t^2}\sqrt{c^2 - t^2} = t(t - a)$. Because $y = \sqrt{b^2 - t^2}\sqrt{c^2 - t^2}$ is increasing from 0 to bc and $y = t(t - a)$ is decreasing from $c(a + c)$ to 0 on $[-c, 0]$, the intermediate value theorem guarantees the existence of a unique t in $[-c, 0]$ satisfying $\sqrt{b^2 - t^2}\sqrt{c^2 - t^2} = t(t - a)$. Therefore, there is a unique such triangle with orthocenter O.

Also solved by Sue Ackermann and Michael Neubauer and Joel Zeitlin, Arthur Berg and Eugene Gutkin, Jean Bogaert (Belgium), Daniele Donini (Italy), Hans Kappus (Switzerland), Victor Y. Kutsenok, Neela Lakshmanan, Laurel and Hardy Problem Group, Stephen Noltie, Rob Pratt and Jesse Frey, Seth Zimmerman, and the proposer. There was one incorrect solution.

A Nonlinear First Order Differential Equation April 1999

1570. *Proposed by Ice B. Risteski, Skopje, Macedonia.*

Solve the differential equation

$$\left(\frac{dy}{dx}\right)^{n+1} + axy^{2n}\frac{dy}{dx} + ay^{2n+1} = 0, \qquad a \neq 0, n \in \mathbb{N}.$$

I. *Solution by Danny Arrigo, University of Central Arkansas, Conway, Arkansas, and Debra P. Otto, student, University of Toledo, Toledo, Ohio.*

The differential equation has solutions

$$y = 0, \qquad y^n = \frac{n}{a}\left(-\frac{n+1}{nx}\right)^{n+1}, \qquad \text{and} \qquad y = \frac{a}{aAx - (-A)^{n+1}}.$$

For $y \neq 0$, the substitution $y = 1/u$ and multiplication by u^{2n+2} gives the Clairaut equation

$$\left(-\frac{du}{dx}\right)^{n+1} - ax\frac{du}{dx} + au = 0.$$

Differentiating with respect to x yields

$$\left[(n+1)\left(-\frac{du}{dx}\right)^{n} + ax\right]\frac{d^2u}{dx^2} = 0,$$

giving rise to two cases. Substituting

$$\left(-\frac{du}{dx}\right)^{n} = -\frac{a}{n+1}x \tag{1}$$

into the Clairaut equation leads to

$$\frac{du}{dx} = \frac{(n+1)u}{nx},$$

which we substitute back into (1) to obtain

$$u^n = -\left(-\frac{nx}{n+1}\right)^n \frac{a}{n+1}x = \frac{a}{n}\left(-\frac{nx}{n+1}\right)^{n+1}.$$

The second case leads to $u = Ax + B$ and substitution into the Clairaut equation implies

$$u = Ax - \frac{(-A)^{n+1}}{a}.$$

Reciprocating to get y yields the claimed solutions.

II. *Solution by Charles K. Cook, University of South Carolina at Sumter, Sumter, South Carolina.*

Let $p = dy/dx$. Then

$$p^{n+1} + axy^{2n}p + ay^{2n+1} = 0 \tag{1}$$

or $ax = -p^n/y^{2n} - ay/p$. Differentiating with respect to y and simplifying yields

$$\left(y\frac{dp}{dy} - 2p\right)\left(np^{n+1} - ay^{2n+1}\right) = 0.$$

If $y\,dp/dy - 2p = 0$, then $p = Cy^2$. Substituting this into (1) yields

$$C^{n+1}y^{2n+2} + aCxy^{2n+2} + ay^{2n+1} = 0,$$

which then yields either $y = 0$ or the general solution

$$y = \frac{-a}{C(C^n + ax)}.$$

If

$$np^{n+1} - ay^{2n+1} = 0, \tag{2}$$

then, substituting this for p^{n+1} in (1), we obtain $y = 0$ or $xp = (-1 - 1/n)y$. Substituting this into (2), we find the singular solution(s)

$$y^n = \frac{n}{a}\left(-\frac{n+1}{nx}\right)^{n+1}.$$

Also solved by Reza Akhlaghi, Jean Bogaert (Belgium), Hans Kappus (Switzerland), Philip Korman, Jerold Lewandowski (student), James Magliano, and the proposer. There was one incomplete solution.

Painting the Digital World April 1999

1571. *Proposed by Michael H. Brill, Sarnoff Corporation, Princeton, New Jersey.*

Let the "real world" be those convex three-dimensional solids whose surfaces are smooth. Let the "digital world" be a three-dimensional tiling of tiny identical cubes, which we call "voxels," analogous to a two-dimensional digital image of square pixels. Each "digital-world" object X' is a maximal subset of these voxels that lies inside the corresponding "real-world" object X.

What is the maximal ratio of the amount of paint needed to cover X' to the amount needed to cover X taken over all convex X whose boundary is a smooth surface and all possible X' as the orientation and size of the voxels vary? In other words, the problem is to find the supremum of the ratio of the exposed surface area of X' to that of X.

Solution by Seth Zimmerman, Oakland, California.

The supremum of the ratio is $\sqrt{3}$.

To see that the ratio, $\sqrt{3}$, can be approached as closely as we wish, consider an octahedron with infinitesimally rounded edges. Assume that the edge lengths of the octahedron are one, so that the "real-world" area is $8 \cdot (\sqrt{3}/4) = 2\sqrt{3}$. Orient a tiling of tiny voxels so that its edges are parallel to the three interior diagonals of the octahedron. As the voxels' size decreases to zero, the "digital-world" area approaches twice the area of the projection of the octahedron onto the three coordinate planes parallel to the voxels faces. The limit of each projection is a unit square so the total digital area is 6. Thus, the limit of the ratio of the digital area to the real world area is $\sqrt{3}$.

We see next that $\sqrt{3}$ is the supremum of the ratio. Considered at the infinitesimal level, a smooth surface can be regarded as locally flat. Thus, we may consider planes passing through $(1,0,0)$, $(0,a,0)$, and $(0,0,b)$, a and b nonnegative and not both zero, and maximize the ratio of sum of the areas of the projections to the three coordinate planes to the area of the triangle. From integration or the vector cross product, we see that the area of the triangle is $\sqrt{a^2b^2 + a^2 + b^2}/2$ and the total area of its projection to the three coordinate planes is $(ab + a + b)/2$. By the Cauchy-Schwarz inequality, $(ab + a + b)/\sqrt{a^2b^2 + a^2 + b^2} \le \sqrt{3}$ with equality if and only if $ab = a = b$, implying $a = b = 1$.

We note that all eight faces of the octahedron were oriented in this way with respect to axes parallel to the edges of the voxels.

Also solved by the proposer.

Limit of a Homogeneous Fractional Recursion April 1999

1572. *Proposed by Western Maryland College Problems Group, Westminster, Maryland.*

Let $b_0 = 1$ and b_1 satisfy $0 < b_1 < 1$. For $n \ge 1$, define b_{n+1} by

$$b_{n+1} = \frac{2b_n b_{n-1} - b_n^2}{3b_{n-1} - 2b_n}.$$

Show that $(b_n)_{n \ge 0}$ converges, and compute its limit in terms of b_1.

Solution by Yan-Loi Wong, The National University of Singapore, Singapore, Republic of Singapore.

We show that $\lim_{n\to\infty} b_n = 0$.

Let $a_1 = 1 - b_1$. First one can prove using induction and the given recursive relation for b_n that for $n \geq 1$,

$$\frac{b_n}{b_{n-1}} = \frac{b_1 + (2n-2)a_1}{b_1 + (2n-1)a_1}.$$

It follows from this that, for $n \geq 1$,

$$b_n = \frac{b_1}{b_1 + a_1} \cdot \frac{b_1 + 2a_1}{b_1 + 3a_1} \cdots \frac{b_1 + (2n-2)a_1}{b_1 + (2n-1)a_1}.$$

This shows that (b_n) is monotone decreasing and bounded below by 0. Hence, it converges. Next we shall prove by induction that for $n \geq 0$, $b_n < 1/\sqrt{b_1 + 2na_1}$. For $n = 0$, the inequality to be proved is $b_0 < 1/\sqrt{b_1}$, which is true because $0 < b_1 < 1$. Suppose that the above inequality is true for $n \geq 0$. Then using the induction hypothesis, we have

$$b_{n+1} = b_n \cdot \frac{b_1 + 2na_1}{b_1 + (2n+1)a_1} < \frac{1}{\sqrt{b_1 + 2na_1}} \cdot \frac{b_1 + 2na_1}{b_1 + (2n+1)a_1} = \frac{\sqrt{b_1 + 2na_1}}{b_1 + (2n+1)a_1}.$$

Direct simplification shows that $(b_1 + 2na_1)(b_1 + (2n+2)a_1) < (b_1 + (2n+1)a_1)^2$. Hence,

$$b_{n+1} < \frac{\sqrt{b_1 + 2na_1}}{b_1 + (2n+1)a_1} < \frac{1}{\sqrt{b_1 + (2n+2)a_1}}.$$

Consequently, $\lim_{n\to\infty} b_n = 0$.

Also solved by Reza Akhlaghi, Tewodros Amdeberhan, Michel Bataille (France), Brian D. Beasley, Jean Bogaert (Belgium), Stan Byrd, David Callan, Jeremy Case, Centre College Problem Solving Group, Knut Dale (Norway), Charles R. Dininnie, Daniele Donini (Italy), Robert L. Doucette, Marty Getz and Dixon Jones, N. H. Guersenzvaig (Argentina), Jim Hartman, E. J. Janowski and G. Ladas, Hans Kappus (Switzerland), Kee-Wai Lau (China), Càn A. Minh (graduate student), Michael Reid, Stew Roberts, C. Ray Rosentrater, Volkhard Schindler (Germany), Heinz-Jürgen Seiffert (Germany), Achilleas Sinefakopoulos (student, Greece), Michael Vowe (Switzerland), Michael Woltermann, Paul J. Zwier, and the proposers. There were two incorrect solutions and one incomplete solution.

Concentric Points in a Triangle April 1999

1573. *Proposed by Jiro Fukuta, Professor Emeritus, Gifu University, Gifu-ken, Japan.*

Given $\triangle ABC$, let AD be a cevian to the side BC, and let E be on segment AD. The circumcircle of $\triangle ACD$ intersects the line BE at points M and N, and the circumcircle of $\triangle ABD$ intersects the line CE at points P and Q. Prove that the points M, N, P, and Q lie on a common circle and its center is on the line perpendicular to the side BC at the point D.

Solution by Michael Reid, Brown University, Providence, Rhode Island.

We have $ME \cdot EN = AE \cdot ED$, because the chords MN and AD intersect at E. Similarly, $PE \cdot EQ = AE \cdot ED$, so $ME \cdot EN = PE \cdot EQ$, from which it follows that M, N,

P, and Q are concyclic. Let O be the center of the circle through M, N, P, and Q, and let r be its radius.

By considering the power of the point B with respect to the circle through M, N, P, and Q, we have $BO^2 - r^2 = BM\cdot BN$. From the circumcircle of $\triangle ADC$, we also have $BM\cdot BN = BD\cdot BC$. Similarly, $CO^2 - r^2 = CP\cdot CQ = CD\cdot CB$. Subtract these two equations to get $BO^2 - CO^2 = BD\cdot BC - CD\cdot CB = BD^2 - CD^2$, or

$$BO^2 - BD^2 = CO^2 - CD^2. \tag{1}$$

The triangle inequality implies $BO + CO \geq BC = BD + CD$, hence both sides of (1) are nonnegative. Let O' be on the line perpendicular to BC at D, on the same side as O and such that $(DO')^2 = BO^2 - BD^2$. Then $BO' = BO$ and, from (1), we have $CO' = CO$. It follows that $O' = O$ and OD is perpendicular to BC as desired.

Also solved by Michel Bataille (France), Daniele Donini (Italy), Marty Getz and Dixon Jones, Victor Y. Kutsenok, Neela Lakshmanan, Volkhard Schindler (Germany), Achilleas Sinefakopoulos (student, Greece), Peter Y. Woo, Robert L. Young, and the proposer.

Answers

Solutions to the Quickies on page 157

A899. Because $-1 < a_1 < 1$, we can write $a_1 = \tanh\theta_1$. Also $a_0 = \tanh\theta_0$, where $\theta_0 = 0$. Inductively,

$$a_{n+2} = \frac{\tanh\theta_n + \tanh\theta_{n+1}}{1 + \tanh\theta_n \tanh\theta_{n+1}} = \tanh(\theta_n + \theta_{n+1}).$$

Therefore $a_{n+2} = \tanh\theta_{n+2}$ with $\theta_{n+2} = \theta_n + \theta_{n+1}$.

The solution to the recursion $\theta_{n+2} = \theta_n + \theta_{n+1}$ has the form

$$\theta_n = b\left(\frac{1+\sqrt{5}}{2}\right)^n + c\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

Using the values $\theta_0 = 0$ and θ_1, we find that the solution in our case is

$$\theta_n = \frac{\theta_1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right],$$

so that

$$a_n = \tanh\left\{\frac{\tanh^{-1}a_1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right]\right\}.$$

A900. The maximum number of lines is $\lceil 90/\alpha \rceil$.

Let O denote the vertex of the angle. There is no loss of generality in assuming P_0 is no further from O than is P_n. Let β denote the measure of $\angle OP_1P_0$. Given $P_0, \dots, P_k$, another point P_{k+1} may be added so long as $\angle OP_{k-1}P_k$ is obtuse. An easy induction yields $\angle OP_{k-1}P_k = 180 - \beta - k\alpha$. Because β may be chosen arbitrarily small, it follows that the maximum value of n satisfies $n\alpha \geq 90 > (n-1)\alpha$, so equals $\lceil 90/\alpha \rceil$.

REVIEWS

PAUL J. CAMPBELL, *editor*
Beloit College

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

Baron-Cohen, Simon, Sally Wheelwright, Valerie Stone, and Melissa Rutherford, A mathematician, a physicist and a computer scientist with Asperger syndrome: Performance on folk psychology and folk physics tests, *Neurocase* 5 (6) (1999) 475–483. Gibbs, W. Wayt, Profile: Monstrous moonshine is true, *Scientific American* (November 1998) 40–41. ASPEN: Asperger Syndrome Education Network, `http:www.AspenNJ.org/` .

Does some of the description in the profile in *Scientific American* of Fields Medalist Richard Borcherds strike a chord? Borcherds was "unnerving [to talk to]," had "palpable unease of his movements," indulged in "frequent far-off stares," was "dress[ed] entirely in wrinkled brown attire," and confessed "I'm not very good at expressing feelings and things like that."

> **nerd** *n.* Slang. 1. A person regarded as stupid, inept, or unattractive. 2. A person who is single-minded or accomplished in scientific pursuits but is felt to be socially inept.—*American Heritage Dictionary*, 3rd ed.

Are you a nerd? Are some of your colleagues or students? Like Borcherds, the mathematician of the *Neurocase* article, some may have *Asperger syndrome* (AS), a relatively new category of developmental disorder, the mildest and highest functioning end of the autism spectrum. Clinical features include paucity of empathy, inappropriate social interaction, pedantic and monotonic speech, poor nonverbal communication, intense absorption in one particular topic, and clumsy movements and odd posture. Males are more likely to be affected, and there may be some inheritance. As in autism, the indicated treatment is support, with little reported effectiveness of specific interventions. One doctor writes: "[S]ome ... individuals with AS represent a unique resource for society, having the single-mindedness and consuming interest to advance our knowledge in various areas of science, math, etc." The study tested Borcherds and two university students vs. 14 controls (what? no Fields Medalists among the controls?) on "folk physics," "folk psychology" (reading mental states from photographs of eyes), and "the most complex test of 'executive function'" (doing the Tower of Hanoi!). The AS individuals did poorly in "folk psychology" but were "at the ceiling" on the other tests. The conclusion is that "theory of mind (folk psychology) is independent of IQ, executive function and reasoning about the physical world." The typology suggests that there may be other syndromes (and corresponding slang words?) that might embrace socially adept and attractive people who are chronically disorganized or incapable of scientific reasoning or rational discussion. (Thanks to Barry Cipra.)

Stewart, Ian, It's a funny old world, *New Scientist* 165 (No. 2224) (5 February 2000) 41–43.

The prolific Ian Stewart here explains the Poincaré conjecture for the general public: "If all loops shrink, is it a 3-sphere?" The exposition is so non-technical that he does not reach the statement of the conjecture until two-thirds of the way through the article; the rest speculates about the likelihood of success for Thurston's geometrization approach and an approach by triangulation, then suggests that perhaps the answer is undecidable [if so, then because of its logical form, it would in fact be true].

Gardner, Martin, *The Last Recreations: Hydras, Eggs, and other Mathematical Mystifications*, Springer-Verlag, 1997; x + 392 pp, $25. ISBN 0-387-94929-1.

Over the period 1956–1986, Martin Gardner—who took no college courses in mathematics—wrote hundreds of columns on "Mathematical Recreations" for *Scientific American.* In the years since 1959, the columns have been collected into books; this is the fifteenth and final collection—"the brilliant capstone to Martin Gardner's unrivalled career as the king of mathematical exposition" (Ron Graham). It features 23 columns plus corrections, addenda, and additional references. The topics range from fun with eggs and checker recreations to the monster group, Bulgarian solitaire, and minimal Steiner trees. (Note to publisher: Excessive leading makes for too little on a page and hence unnecessarily many pages.)

Fernandes, Andrew D., Elliptic-curve cryptography, *Dr. Dobb's Journal* (December 1999) 56–63.

Contemporary cryptographic systems are based on the difficulty of factoring integers (e.g., the RSA cryptosystem) or else on the difficulty of finding *discrete logarithms* in a group: Given a group and elements x and y, find a positive integer k for which $x = y^k$. The Diffie-Hellman key exchange scheme and various digital signature schemes (e.g., the U.S. government Digital Signature Algorithm (DSA)) work with the group Z_p^* over a large prime p. Elliptic-curve cryptography (ECC) replaces this group with one over an elliptic curve. The motivations for ECC are *time* (addition in the elliptic curve group can be much faster than multiplication in Z_p^*), *space* (shorter keys for a security-equivalent elliptic-curve system), and *hope* (that the discrete logarithm problem in ECC is fundamentally harder). Author Fernandes discusses the pros and cons of ECC, experiments with selecting elliptic curves, conducts a comparison benchmarking of DSA and ECC systems, and points the reader to further sources. Code is available at `http://www.ddj.com/ftp/1999/1999_12/ellip.zip` .

The Top Ten Algorithms. Special issue of *Computing in Science & Engineering* 2 (1) (January/February 2000).

Guest editors Jack Dongarra and Francis Sullivan assemble here articles on the 10 algorithms "with the greatest influence on the development and practice of science and engineering in the 20th century." Here they are in chronological order, almost all from the third quarter of the century: Metropolis algorithm for Monte Carlo, simplex method for LP (article by John Nash), Krylov subspace iteration methods, decompositional approach to matrix computation, Fortran optimizing compiler, QR algorithm, quicksort, fast Fourier transform, integer relation detection, and fast multipole method. It would be an interesting exercise to ask the same question about what *theorems* had "the greatest influence on the development and practice of science and engineering in the 20th century."

Bollag, Burton, Notes from academe: Proofs and conundrums for North American students in math-crazy Hungary, *Chronicle of Higher Education*(17 December 1999) B4.

Roughly 15 years after its establishment, the Budapest Semester in Mathematics has at last attracted the attention of the U.S. weekly newspaper of college administrators and professors. Judging from the the activity at the Budapest Semester's booth at the January Mathematics Meetings, the program is healthy; but this article relates worry that Hungary's fame as an "incubator" of mathematicians may not persist in the post-Communist free-market era, when mathematicians are relatively poorly paid: "If a woman marries a mathematician, it's not something she wants to advertise, as if she married a banker," says Dezső Miklós, acting director. Perhaps, given the famous Hungarian mathematical tradition, Hungarian men who marry mathematicians have other sentiments.

NEWS AND LETTERS

New Editor-Elect of *Mathematics Magazine*

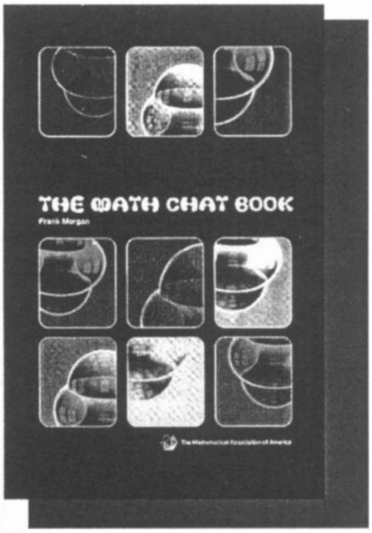
Beginning immediately, please submit new manuscripts to:

Frank Farris
Department of Mathematics and Computer Science
Santa Clara University
500 El Camino Real
Santa Clara, CA 95053-0290

Please read the editorial guidelines posted at `www.maa.org/pubs/mathmag.html`. In addition, we offer the following ideas for potential authors:

- Initial submission continues to be in a physical rather than electronic form. Should your article be accepted, we will ask you to provide a LaTeX file using one of the templates provided at our website. If this is impossible for you, a text file or common word-processor document is acceptable.
- Remember that a good expository article begins with an introduction that grabs the reader's attention and encourages him or her to keep reading.
- If you wish to provide any electronic complement to your article, including such things as color illustrations, Java applets, or animations, supply the URL of your draft site. If your article is accepted, complements will be hosted at `www.maa.org`.
- In the interest of respecting the time of our referees, we recommend a referee's appendix, not for publication, but to guide the referee. Please expand on statements such as, "A simple calculation shows . . . ." It is often appropriate to suppress such things in exposition, but a referee might find the additional information a time-saver.
- We strongly recommend that you search the electronic database of *Mathematics Magazine* and the *College Mathematics Journal* for articles on subjects related to yours. Follow the link to this site from the address above. This should help to fill out your bibliography and avoid any duplication.

The Mathematical Association of America



The Math Chat Book

Frank Morgan

Series: MAA Spectrum

This book shows that mathematics can be fun for everyone. It grew out of Frank Morgan's live, call-in Math Chat TV show and biweekly Math Chat column in *The Christian Science Monitor*. The questions, comments, and even the answers come largely from the callers and readers themselves.

Why does the new year start earlier in Europe? Why is the Fourth of July on a different day of the week each year? How can you be elected President with just 22% of the vote? Can a computer have free will? Didn't some kid find a mistake on the SATs? Do airplanes get lighter as passengers eat lunch? College students make important progress on the still open "Double Bubble Conjecture."

One youngster asks, "If I live for 6000 years, how many days will that be?" His first answer is (6000 years)(365 days/year) = 2,190,000 days. That is not quite right: it overlooks leap years. An older student takes leap years into account, adds 1500 leap year days, and comes up with 2,191,500 days. The answer is still not quite right. Every hundred years we skip a leap year (the year 1900, although divisible by four, was not a leap year), so we subtract 60 days to get 2,191,440. The answer is still not quite right. Every four hundred years we put the leap year back in (2000 will be a leap year), so we add back 15 days to get 2,191,455, the final answer.

This book makes no attempt to fit any mold. Although written by a research mathematician, it goes where the callers and readers take it, over a wide range of topics and levels. Almost anyone paging through it will find something of interest. It is time for everyone to see how much fun mathematics can be.

Catalog Code: MCH/JR 112pp., Paperbound, 2000 ISBN 0-88385-530-5 List: $19.95 MAA Member: $16.00

Name________________ Credit Card No.________________

Address________________ Signature________________Exp. Date____/____

City________________ Qty______ Price $________ Amount $________

State________ Zip ________ Shipping and Handling $________

Phone ________________ Catalog Code: MCH/JR Total $________

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

Order Via:

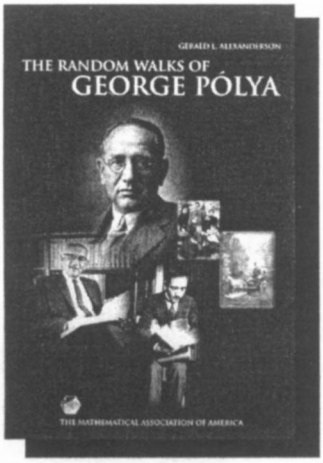
Phone: **1 (800) 331.1622**
Fax: **(301) 206.9789**
Mail: **Mathematical Association of America**
PO Box 91112
Washington, DC 20090-1112
Web: **www.maa.org**

The Mathematical Association of America

The Random Walks of George Pólya

Gerald L. Alexanderson

Series: MAA Spectrum



In the first half of this charming book Gerald Alexanderson presents an insightful portrait of George Pólya, the great teacher and mathematician. In the second half of the book, Alexanderson assembles eight papers that describe Pólya's contribution to various fields.

George Pólya enjoyed the esteem of the mathematical community not only for his deep and influential contributions in a variety of mathematical fields, but also for his groundbreaking work in the teaching of mathematics. His standing in the latter area could rest solely on his having written one of the most widely read books in mathematics, the still-popular *How to Solve It.* In addition to his championing problem-solving, he contributed to mathematics important results in complex and real analysis, inequalitites, mathematical physics, combinatorics, probability theory, number theory, and geometry. He coined the phrases "random walk" and "central limit theorem" and gave to mathematics the Pólya Enumeration Theorem, along with many other ideas used widely today. The present work describes how such versatility came about and, along the way, tells some enlightening stories about mathematics and mathematicians.

The list of articles about Pólya's work include: *Pólya's Work in Probability*, by K.L. Chung, *Pólya's Work in Analysis*, by R.P. Boas, *Comments on Number Theory*, by D.H. Lehmer, *Pólya's Geometry*, by Doris Shattschneider, *Pólya's Enumeration Theorem*, by R.C. Read, *Pólya's Contributions in Mathematical Physics*, by M.M. Schiffer, *George Pólya and Mathematics Education*, by Alan Schoenfeld, and *Pólya's Influence-References to His Work.*

Paperbound Edition:	Casebound Edition:
Catalog Code: RWP	Catalog Code: WAY
320pp., Paperbound, 2000	320pp., Casebound, 2000
ISBN 0-88385-528-3	ISBN 0-88385-531-3
List Price: $29.95 Member Price: $23.95	List Price: $41.95 Member Price: $32.95

Name________________ Credit Card No.________________

Address________________ Signature________________Exp. Date___/___

City________________ Qty______ Price $________ Amount $________

State________ Zip ________ Shipping and Handling $________

Phone ________________ Catalog Code: RWP/JR or WAY/JR Total $________

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

Order Via:

Phone: **1 (800) 331.1622**

Fax: **(301) 206.9789**

Mail: **Mathematical Association of America**
PO Box 91112
Washington, DC 20090-1112

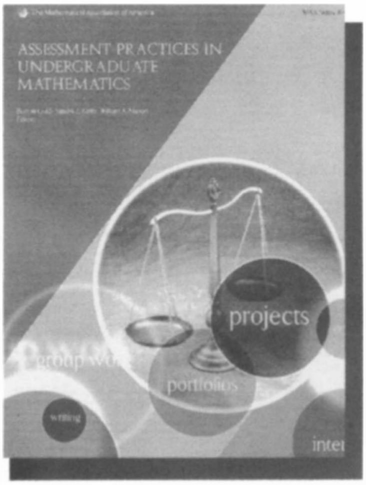
Web: **www.maa.org**

The Mathematical Association of America

Assessment Practices in Undergraduate Mathematics

Bonnie Gold, Sandra Keith, and
William Marion, Editors

Series: MAA Notes



This book, a collection of assessment practices that have been tried by more than 100 contributors in mathematics at a wide variety of schools, attempts to offer the mathematics teacher suggestions from an insider's perspective. The book is not formulaic: no author claims to have "the answer," and many of the projects reported on are still in progress. On the other hand, the articles provide a wealth of suggestions from creative, energetic and concerned individuals who have had the courage to experiment and to critique their own efforts. Without doubt, the reader will find in these pages encouragement to experiment on his or her own, to find assessment methods which are personally meaningful.

Techniques offered in this book range from brief ten-minute classroom exercises and examples of alternative testing, group work and assignments, to examples of how departments may measure the placement of students into courses, the effectiveness of the major, and the quantitative literacy of their graduating students. Teachers beleaguered by formal end-of-term teacher evaluation forms, will find a variety of alternative assessment techniques that provides ways in which the quality of teaching can be better examined.

The book is unique among assessment books in representing the point of view of mathematicians exploring and examining methods of learning in their field.

Catalog Code: NTE-49/JR 350 pp., Paperbound, 1999 ISBN 0-88385-161-x List: $29.95 MAA Member: $23.95

Name________________ Credit Card No.________________

Address________________ Signature________________ Exp. Date____/____

City________________ Qty______ Price $__________ Amount $__________

State__________ Zip __________ Shipping and Handling $__________

Phone ________________ Catalog Code: NTE-49/JR Total $__________

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

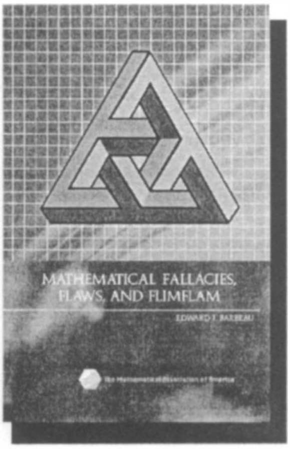
Order Via:

Phone: 1 (800) 331.1622
Fax: (301) 206.9789
Mail: Mathematical Association of
America
PO Box 91112
Washington, DC 20090-1112
Web: www.maa.org

The Mathematical Association of America

Mathematical Fallacies, Flaws, and Flimflam

Edward J. Barbeau


MATHEMATICAL FALLACIES, FLAWS, AND FLIMFLAM
EDWARD J. BARBEAU
The Mathematical Association of America

Series: MAA Spectrum

This book is a collection of mathematical mistakes made by students, teachers, and occasionally seasoned researchers, along with an analysis for most of them. While all the material is for personal enlightenment and amusement, high school and college teachers may use the material to illustrate important and subtle points in mathematics.

Newspapers are responsible for a good number of these mathematical mishaps, particularly in arithmetic (especially percentages) and probability. Quite a number of the "fallacies" come from professional mathematicians. Some are the result of simple oversight, and others are deliberately crafted by the mathematician to drive home an important point to students.

A glimpse at the Table of Contents offers examples from number theory, algebra and trigonometry, geometry, finite mathematics, probability, calculus, linear algebra and advanced undergraduate mathematics.

An example of "mathematical flimflam" from the Calculus Limits and Derivatives section:

> The shortest distance from a point to a parabola
> Problem: Determine the shortest distance from the point (0,5) to a parabola $16y = x2$.
> Solution: We minimize $f(y) = x2 + (y - 5) 2 = 16y + (y - 5) 2$.
> Since $f'(y) = 2y + 6$, the only critical value of f is $y= -3$, which corresponds to an imagainary value of x. Hence the minimum distance does not exist.

Catalog Code: FFL/JR 160 pp., Paperbound, 2000 ISBN 0-88385-529-1 List: $23.95 MAA Member: $19.00

Name__________ Credit Card No.__________

Address__________ Signature__________Exp. Date___/___

City__________ Qty_____ Price $_____ Amount $_____

State_____ Zip _____ Shipping and Handling $_____

Phone __________ Catalog Code: FFL/JR Total $_____

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

Order Via:

Phone: 1 (800) 331.1622
Fax: (301) 206.9789
Mail: Mathematical Association of America
PO Box 91112
Washington, DC 20090-1112
Web: www.maa.org

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

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