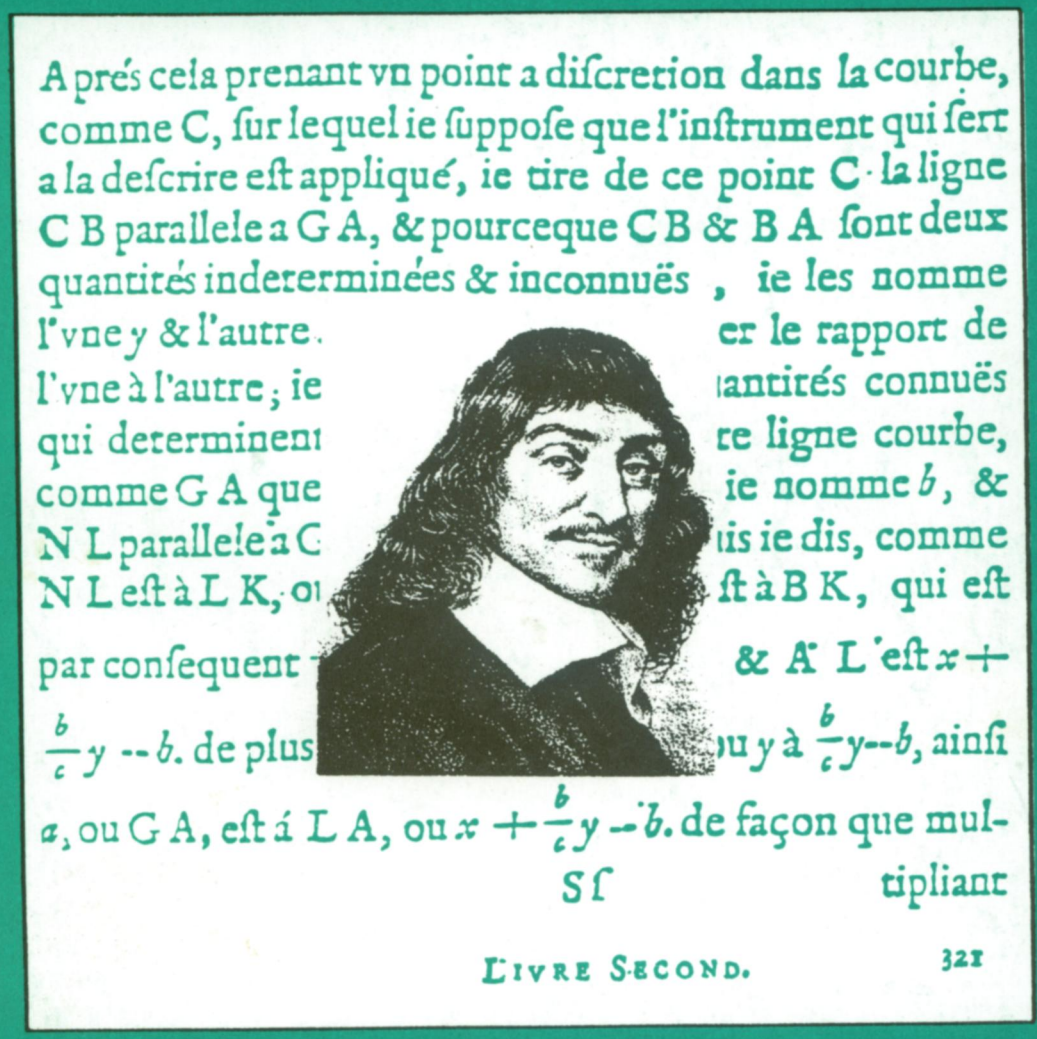
Vol. 68, No. 2 April 1995

MATHEMATICS MAGAZINE



- Descartes and Problem-Solving
- The Ptolemy Inequality and Minkowskian Geometry
- Is There Any Regularity in the Distribution of Prime Numbers at the Beginning of the Sequence of Positive Integers?

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

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

The full statement of editorial policy appears in this *Magazine*, Vol. 64, pp. 71–72, and is available from the Editor. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, nor published by another journal or publisher.

Send new manuscripts to: Paul Zorn, Editor-Elect, Department of Mathematics, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098. Manuscripts should be typewritten and double spaced and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should submit the original and two 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. Illustrations should be carefully prepared on separate sheets in black ink, the original without lettering and two copies with lettering added. Do not use staples.

Cover Illustration: A page from Descartes' *La Géométrie* (1637). Descartes from the David Smith Collection, Columbia University.

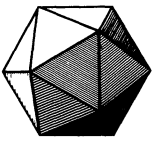
AUTHORS

Judy Grabiner became interested in Descartes' proof of the existence of Descartes ("I think, therefore I am") while earning her B.S. in Mathematics at the University of Chicago. She received her Ph.D. in the History of Science from Harvard. She teaches mathematics and the history of science at Pitzer College in Claremont, CA, including a liberal arts course, "Mathematics, Philosphy, and the 'Real World'," in which the students read Descartes. Her writings include *The Origins of Cauchy's Rigorous Calculus* (MIT, 1981) and "The Centrality of Mathematics in the History of Western Thought," this MAGAZINE, 1988.

John D. Smith first met Ptolemy's theorem at his high school, and it has intrigued him ever since. He graduated in Mathematics from Oxford in 1965 and did a doctorate in relativitistic scattering theory. For 20 years he has taught at Winchester College, a boarding school of mediaeval foundation, where he is Assistant Chaplin and Head of Mathematics. He believes that geometry is a uniquely favorable environment for young students to learn the spirit of pure mathematics and exercise their intuition. His interests include rowing and high-level walking.

Silviu Guiasu graduated in mathematics at the University of Bucharest and received his Ph.D. from the Romanian Academy of Sciences in Bucharest in 1966. In 1982, he was appointed professor at York University in Toronto, teaching Operations Research courses. His research interests are mainly in probabilistic modeling and variational techniques involving global measures of uncertainty for getting order out of chaos and sometimes chaos out of order. Falling under the spell of prime numbers, he does believe that dealing with their mysterious distribution could be addictive, but it is so nice doing it.

Vol. 68, No. 2 April 1995



MATHEMATICS MAGAZINE

The *MATHEMATICS MAGAZINE* (ISSN 0025–570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.

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

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

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

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

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

PRINTED IN THE UNITED STATES OF AMERICA

ARTICLES

Descartes and Problem-Solving

JUDITH GRABINER
Pitzer College
Claremont, CA 91711

Introduction

What does Descartes have to teach us about solving problems? At first glance it seems easy to reply. Descartes says a lot about problem-solving. So we could just quote what he says in the *Discourse on Method* [**12**] and in his *Rules for Direction of the Mind* ([**2**], pp. 9–11). Then we could illustrate these methodological rules from Descartes' major mathematical work, *La Géométrie* [**13**]. After all, Descartes claimed he did his mathematical work by following his "method." And the most influential works in modern mathematics—calculus textbooks—all contain sets of rules for solving word problems, rather like this:

1. Draw a figure.
2. Identify clearly what you are trying to find.
3. Give each quantity, unknown as well as known, a name (e.g., $x, y, \ldots$).
4. Write down all known relations between these quantities symbolically.
5. Apply various techniques to these relations until you have the unknown(s) in equations that you can solve.

The calculus texts generally owe these schemes to George Pólya's *Mathematical Discovery*, especially Chapter 2, "The Cartesian Pattern," and Pólya himself credits them to Descartes' *Rules for Direction of the Mind* ([32], pp. 22–23, 26–28, 55–59, 129ff). So I studied those philosophical works as I began to write about Descartes and problem-solving. But the more I re-read Descartes' *Geometry*, the more convinced I became that it is from this work that his real lessons in problem-solving come. One could claim that, just as the history of Western philosophy has been viewed as a series of footnotes to Plato, so the past 350 years of mathematics can be viewed as a series of footnotes to Descartes' *Geometry*.

Now Descartes said in the *Discourse on Method* that it didn't matter how smart you were; if you didn't go about things in the right way—with the right method—you would not discover anything. Descartes' *Geometry* certainly demonstrates a successful problem-solving method in action. Accordingly, this article will bring what historians of mathematics know about Descartes' *Geometry* to bear on the question, what can Descartes teach the mathematics community about problem-solving? To answer this question, let us look at the major types of problems addressed in the *Geometry* and at the methods Descartes used to solve them.

A First Look at Descartes' *Geometry*

We have all heard that Descartes' *Geometry* contains his invention of analytic geometry. So when we look at the work, we may be quite surprised at what is *not*

there. We do not see Cartesian coordinates. Nor do we see the analytic geometry of the straight line, or of the circle, or of the conic sections. In fact we do not see *any* new curve plotted from its equation. And what curves did Descartes allow? Not, as we might think, any curve that has an equation; that is secondary. He allowed only curves constructible by some mechanical device that draws them according to specified rules. Finally, we do not find the term "analytic geometry," nor the claim that he had invented a new subject—just a new (and revolutionary) method to deal with old problems.

What we *do* see is a work that is problem-driven throughout. Descartes' *Geometry* has a purpose. It is to solve problems. Some are old, some are new; all are hard. For all the lip service in Descartes' *Discourse on Method* to mathematics as logical deduction from self-evident first principles ([**12**], pp, 12–13, 18–19), the *Geometry* is not like that at all. It discovers; it does not present a finished logical structure. The specific purpose of the book is to answer questions like "What is the locus of a point such that a specified condition is satisfied?" And the answer to these questions must be geometric. Not "it is such-and-such a curve," or even "it has this equation," but "it is this curve, it has this equation, and it can be constructed in this way." Everything else in the *Geometry*—and that does include algebra, theory of equations, classifying curves by degree, etc.—are just means to this geometric end. To solve a problem in geometry, one must be able to construct the curve that is its solution.

The Background of Descartes' *Geometry*

To appreciate how much Descartes accomplished, we must first look at some achievements of the ancient Greeks. They solved a range of locus problems, some quite complicated. To find their solutions, they too had "methods." Greek mathematics recognized two especially useful problem-solving strategies: *reduction* and *analysis* ([**25**], pp. 23–24).

First, let us describe the method of reduction [in Greek, *apogōgē*]. Given a problem, we observe that we could solve it if only we could solve a second, simpler problem, and so we attack the second one instead. For instance, consider the famous problem of duplicating the cube. In modern notation, the problem is, given a^3, to find x such that $x^3 = 2a^3$. Hippocrates of Chios showed that this problem could be reduced to the problem of finding two mean proportionals between a and $2a$. That is, again in modern notation, if we can find x and y such that:

$$a/x = x/y = y/2a, \tag{1}$$

then, eliminating y, we obtain $x^3 = 2a^3$ as required ([**25**], p. 23). But more geometric knowledge led to a further reduction ([**25**], p. 61). If we consider just the first two terms of (1),

$$a/x = x/y,$$

we obtain $x^2 = ay$, which represents a parabola. The equation involving the first and third terms in (1) yields

$$a/x = y/2a$$

or $xy = 2a^2$, which represents a hyperbola. Thus the problem of duplicating the cube is reducible to the problem of finding the intersection of a parabola and a hyperbola. This reduction promoted Greek interest in the conic sections.

The other problem-solving strategy is what the Greeks called "analysis"—literally, "solution backwards" (*ànápalin lýsin* [**20**], Vol. ii, p. 400; [**25**], p. 9; cp. pp. 354–360). The Greek "analysis" works like this. Suppose we want to learn how to construct an angle bisector, and suppose that we already know how to bisect a line segment. We proceed by first assuming that we have the problem solved. Then, from the assumed existence of that angle bisector, we work backward until we reach something we do know. In FIGURE 1, take the angle A, and draw AK bisecting it. Then, mark off any length AB on one side of the angle, and an equal length AC on the other side. Connect B and C with the straight line BC, as in FIGURE 2. Now let M be the intersection of the angle bisector with the line BC. Since angle BAM = angle MAC, $AB = AC$, and $AM = AM$, triangle ABM is congruent to triangle ACM. Thus M bisects BC. But wait. Recall that we already know how to bisect a line segment. Thus, we can find such an M. Now we can construct the angle bisector by reversing the process we just went through. That is, suppose we are given an angle A. To construct the angle bisector, construct $AB = AC$, construct the line BC, bisect it at M, and connect the points A and M. AM bisects the angle. This method—assuming that we have the thing we are looking for and working backwards from that assumption until we reach something we do know—was well-named "solution backwards." Pappus of Alexandria, in the early fourth century C.E., compiled a "treasury of analysis" in which he gave the classic definition of "analysis" as "solution backwards"; described 33 works, now mostly lost, by Euclid, Apollonius, Aristaeus, and Eratosthenes, which included substantial problems solvable by the method of analysis; and provided some lemmas that illustrate problem-solving by analysis ([**20**], Vol. ii, pp. 399–427).

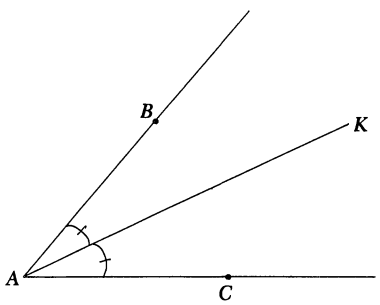


FIGURE 1

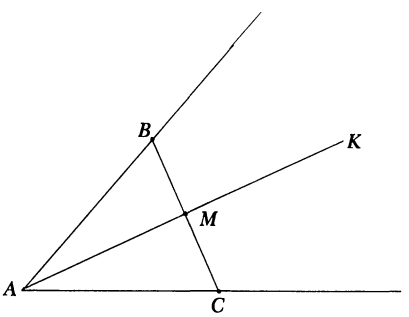


FIGURE 2

In our example of bisecting an angle, the mathematical knowledge needed was minimal. But the Greeks knew all sorts of properties of other geometric figures, notably the conic sections, and so had an extensive set of theorems to draw on in using "analysis" to solve problems in geometry ([**6**], pp. 21–39; [**10**], pp. 43–58; [**20**], passim; [**25**]). (The best and fullest account is that of Knorr [**25**].)

Thus we see that Descartes, though he championed these techniques, clearly did not invent the method of analysis and the method of reduction. Descartes' ideas on problem-solving, moreover, have other antecedents besides the Greek mathematical tradition. First, a preoccupation with finding a universal "method" to find truth appears in the work of earlier philosophers, including the thirteenth-century Raymond Lull, whose method was to list all possible truths and select the right one, the sixteenth-century Petrus Ramus, who saw method as the key to effective teaching and to allowing learners to make their own discoveries ([**29**], pp. 148–9), and the seventeenth-century philosopher of science Francis Bacon, whose method to empirically discover natural laws was one of systematic induction and testing [**1**]. All of these

seekers for method suggested that intellectual progress, unimpressive earlier in history, could be achieved once the right method for finding truth was employed. Descartes shared this view.

A second, more specific antecedent of Descartes' work was the invention of symbolic algebra as a problem-solving tool, a tool that was explicitly recognized as a kind of "analysis" in the Greek sense by its discoverer, Vieta, in 1591 ([6], p. 65; cp pp. 23, 157–173). To say "let $x =$ " the unknown, and then calculate with x—square it, add it to itself, etc., *as if it were known*—is a powerful technique when applied to word problems both in and outside of geometry. Vieta recognized that naming the unknown and then treating it as if it were known was an example of what the Greeks called "analysis," so he called algebra "the analytic art." Incidentally, Vieta's use of this term is the origin of the way we use the word "analysis" in mathematics. In the seventeenth and early eighteenth centuries, the term "analysis" was often used interchangeably with the term "algebra," until by the mid-eighteenth century "analysis" became used for the algebra of infinite processes as opposed to that of finite ones [4].

Descartes was quite impressed with the power of symbolic algebra. But, although he had all these predecessors, Descartes combined, extended, and then exploited these earlier ideas in an unprecedented way. To see how his new method worked, we need to look at a specific problem.

Descartes' Method in Action

We begin with the first important problem Descartes described solving with his new method ([13], pp. 309–314, 324–335). The problem is taken from Pappus, who said in turn that it came from Euclid and Apollonius ([13], p. 304). The problem is illustrated in FIGURE 3 (from [13], p. 309).

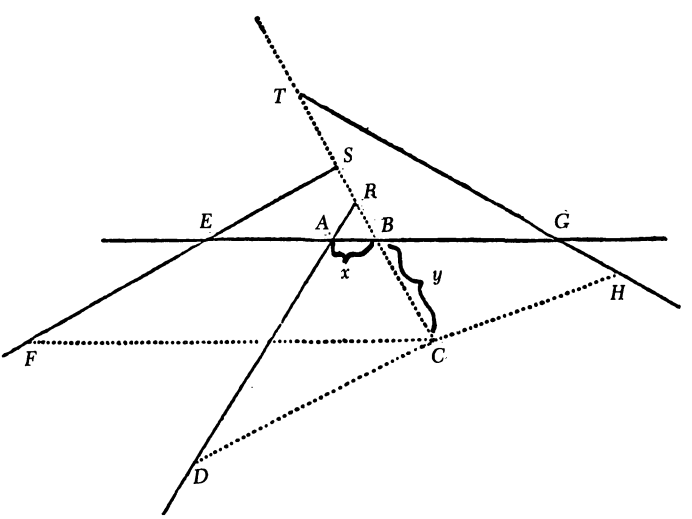


FIGURE 3

Given four lines in a plane, and given four angles. Take an arbitrary point C. Consider now the distances (dotted lines) from C to the various given lines, where the distances are measured along lines making the given angles with the given lines. (For instance, the distance CD makes the given angle CDA with the given line AD.) A further condition on C is that the four distances CD, CF, CB, and CH satisfy

$$(CD \cdot CF)/(CB \cdot CH) = \text{a given constant.} \qquad (2)$$

The problem is to find the locus of all such points C. For Descartes, that means to discover what curve it is, and then to construct that curve. (At this time, any reader who does not already know the answer is encouraged to conjecture what kind of curve it is—or to imagine constructing even *one* such point C.)

Here is how Descartes attacked this problem. First assume, as we must in order to draw FIGURE 3, that we already have one point on the curve. We will then work backwards, by the method of analysis. Draw the point C, and draw the distances. Label the distance from C to the line EG as y, and the line segment between that distance and the given line DA as x. Given these labels x and y, we use them and look for other relationships that can be derived in terms of them. For instance, independent of the choice of C, the angles in the triangle ABR are all known (since angle CBG is one of the given angles in the problem, we have angle ABR by vertical angles; angle RAB is determined by the position of the two given lines that include the segments DR and GE). Thus the shape of triangle ABR is determined, so the side RB is a fixed multiple of x. Descartes therefore called that side $(b/z)\cdot x$, where he took b/z to be a known ratio. Thus $CR = y + (b/z)\cdot x$ ([13], p. 310). Using his knowledge of geometry in this fashion, Descartes found many more such relationships, and was able to express each of the distances CD, CF, CB, and CH as a different linear function of the line segments x and y. For the case where $(CD\cdot CF)/(CB\cdot CH) = 1$, those expressions let him derive an equation between the unknowns x and y and various constants he called m, n, z, o, and p:

$$y = m - (n/z)\cdot x + \sqrt{\{m^2 + ox + (p/m)\cdot x^2\}} \qquad (3)$$

([13], p. 326). Now perhaps the modern reader can guess what type of curve that equation represents. So could Descartes. From his studies of Greek geometry, Descartes knew quite a lot about the conic sections, so he said, though he did not explain, that if the coefficient of the x^2 term is zero, the points C lie on a parabola; if that coefficient is positive, on a hyperbola; if negative, on an ellipse; etc. The positions, diameters, axes, centers, of these curves can be determined also, and he briefly discussed how to do this ([13], p. 329–332).

The reader will have observed that there is no fixed coordinate system here. Descartes labeled as x and y the lengths of line segments that arose in this particular situation. Let us also make a comment about his choice of notation. Vieta had used uppercase vowels for the unknowns, consonants for knowns. Since matters of notation are relatively arbitrary, the fact that we use Descartes' lowercase x and y, rather than Vieta's A and E, testifies to the great influence of Descartes' work on our algebra and geometry. Further, though Descartes himself wrote mm and xx rather than m^2 and x^2 ([13], p. 326), he did use raised numbers, exponents, for integer powers greater than two (e.g. [13], p. 337, p. 344). Today we follow Descartes here too, using exponential notation for all powers.

The Greeks already knew that the Pappus four-line locus was a conic section. Nonetheless, the way Descartes derived this result is impressive. In line with our overall purpose, let us reflect on the method Descartes used. Why is "let x equal the first unknown" so powerful here? Because the technique of "reduction" was used by Descartes to effectively reduce a problem in geometry to a problem in algebra. Once he had done this, he could use the algorithmic power and generality of algebra to solve a formerly difficult problem with relative ease. It is an old problem-solving method, to reduce a problem to a simpler one, but because the simpler one is algebraic, Descartes had something different in kind from what had been done before. Algebra puts muscles on the problem-solving methods of analysis and reduction.

Beyond the Greeks

To fully exploit the power of algebra—to go beyond the Greeks—Descartes had to make a major break with the past. The earlier symbolic algebra of Vieta was based on the theory of geometric magnitudes inherited from the Greeks. Because of this geometric basis, the product of three magnitudes was spoken of as a volume. This created a problem: What might the product of five magnitudes be? Also, Greek geometry presupposed the Archimedean axiom: Quantities cannot be compared unless some multiple of one can exceed the other, so one cannot add a point to a line, or an area to a solid. How then could one write $x^2 + x$ ([6], p. 61, p. 84)? Descartes, like his predecessors, did not envision pure numbers, but only geometrical magnitudes. He too felt constrained to interpret all algebraic operations in geometric terms. But he invented a new geometric interpretation for algebraic equations that freed algebraists from crippling restrictions like being unable to write x^5 or $x^2 + x$. He freed himself, and therefore freed his successors, including us. Here is how he did it.

He took a line that he called "unity," of length one, which could be chosen arbitrarily. This let him interpret the symbol x as the area of a rectangle with one side of length x, the other of length one. He could now write $x^2 + x$ with a clear conscience, since it could be thought of as the sum of two areas. Even more important, he interpreted products as lengths of lines, so that he could interpret any arbitrary power as the length of a line. That is, the product of the line segments a and b for Descartes did not have to be the area ab, but could be another length such that $ab/a = b/1$. And the length ab could be constructed, as in FIGURE 4 ([13], p. 298).

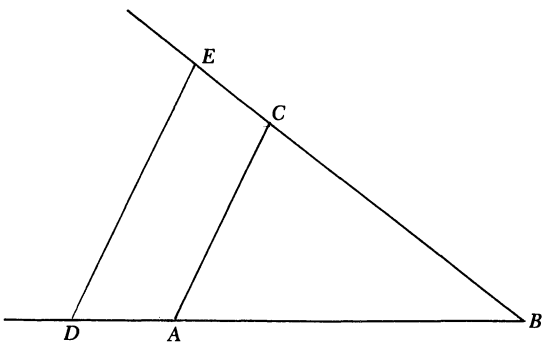


FIGURE 4

In this example, the product of the lines BD and BC is constructed, given a unit line AB. Let the line segments AB and BD be laid off on the same line originating at B, and let the segment BC be laid off on a line intersecting BD. Extending BC and constructing ED parallel to AC yields the proportion $BE/BD = BC/1$, since $AB = 1$. Thus BE is the required product $BD \cdot BC$. Of course this is an easy construction, but he had to give it explicitly. Descartes' philosophy of geometry did not let him merely assert that there was a length equal to the product of the two lines; he had to construct it. Now there was no problem in writing such expressions as x^5. This was just the length such that $x^5/x^3 = x^2/1$.

By showing that all the basic algebraic operations had geometric counterparts, Descartes could use them later at will. Furthermore, he had made a major advance in writing general algebraic expressions. Because of Descartes' innovations, later mathematicians came to consider algebra as a science of numbers, not geometric magnitudes, even though Descartes himself did not explicitly take this step. Descartes took his notational step in the service of solving geometric problems, in order to legitimize

the algebraic manipulations needed to solve these geometric problems. What became a major conceptual breakthrough, then, was in the service of Descartes' problem-solving.

Descartes could now go beyond the Greeks, extending the Pappus four-line problem to five, six, 12, 13, or arbitrarily many lines. With these more elaborate problems, he still followed the same method: Label line segments, work out equations. But when he found the final equation and it was not recognizable as the equation of a conic, what then? To answer this, let me give the simple example he gave, a special case of the five-line problem. He considered four parallel lines separated by a constant distance, with the fifth line perpendicular to the other four ([13], pp. 336–337). (See FIGURE 5.) What, he asked, is the locus of all points C such that

$$CF \cdot CD \cdot CH = CB \cdot CM \cdot AI, \tag{4}$$

where AI is the constant distance between the equally spaced parallel lines and where the distances are all measured at right angles?

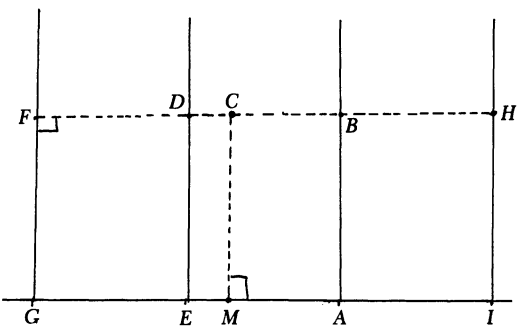


FIGURE 5

Again, Descartes proceeded by analysis. Assuming that he had such a point C, he labelled the appropriate line segments x and y ($x = CM, y = CB$), designated the known distance AI as a, and wrote down algebraic counterparts of all known geometric relationships. For this problem they are simple ones. For instance, $CD = a - y$ and $CF = a + (a - y) = 2a - y$. Thus condition (4) becomes

$$(2a - y)(a - y)(y + a) = y \cdot x \cdot a,$$

which, multiplied out, yields the equation

$$y^3 - 2ay^2 - a^2y + 2a^3 = axy \quad ([13], \text{p. } 337). \tag{5}$$

This is not a conic (it is now often called the cubical parabola of Descartes), so the next question must be, can the curve this represents be constructed? That is, given x, can we find the corresponding value of y and thus construct any point C on the curve? Until these questions are answered affirmatively, Descartes would not consider the five-line problem solved, because, for him, it is a problem in geometry. The algebraic equation was just a means to the end for Descartes; it was not in itself the solution.

So precisely what does "constructible" mean for Descartes? Can the curve represented by that cubic equation be constructed, and, if so, how?

Here another of Descartes' methodological commitments helped him solve this problem: his commitment to generality. The ancients allowed the construction of

straight lines and circles, said Descartes, but classified more complex curves as "mechanical, rather than geometrical" ([13], p. 315). Presumably this was because instruments were needed to construct them. (For instance, Nicomedes had generated the conchoid by the motion of a linkage of rulers ([25], pp. 219–220), and then used the curve in duplicating the cube and trisecting the angle.) But even the ruler and compass are machines, said Descartes, so why should one exclude other instruments ([13], p. 315; tr., p. 43)? Descartes decided to add to Euclid's construction postulates that "two or more lines can be moved, one upon the other, determining by their intersections other curves" ([13], p. 316). The curves must be generated according to a definite rule. And for Descartes, such a rule, at least in principle, was given by the use of a mechanical device that generated a continuous motion. Exactly what this means is complex—for instance, the machine is not allowed to convert an arc length to a straight line—but Bos has provided an enlightening discussion ([3], pp. 304–322, esp. p. 314).

FIGURE 6 reproduces one of Descartes' curve-constructing devices ([13], p. 320). The first curve he generated using it was produced by the intersection of moving straight lines. The straight line KN (extended as necessary) is at a fixed distance KL from a ruler GL. The ruler is attached to the point G, around which it can rotate. The point L can slide along the ruler GL. The segment KL moves up the fixed line AB (extended as needed). As KL moves up, the ruler, which has a "sleeve" attached to L, rotates about G. Note that KL, KN, and the angle between them are all fixed. Then the point at which the ruler GL intersects the straight line KN extended, namely C, will be a point on the curve generated by this device.

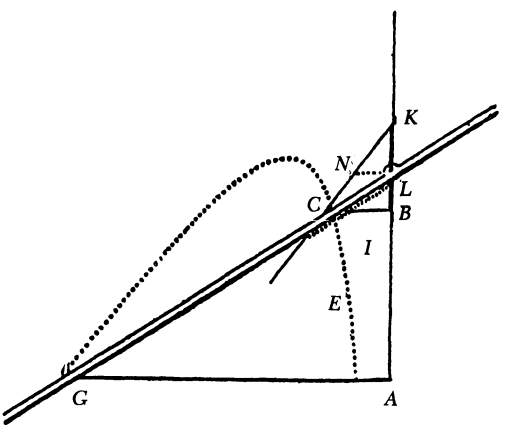


FIGURE 6

To help the reader understand the operation of this device, I show, in FIGURE 7, the construction of a second point C' by this device. KL has moved up; KN thus has a new position; the ruler has rotated to a new position. Where the ruler and KN extended now intersect is another point C' on the curve. If one continues moving KL up and down, the points C, C', etc., trace a new curve.

But what kind of curve is it? Descartes solved this problem in his usual way. He labelled the key line segments (he let the unknowns $y = CB$ and $x = BA$, and the knowns $a = AG$, $b = KL$, and $c = NL$), and algebraically represented the geometric relationships between them. He then showed that if KNC is, as it is in our diagram, a straight line, the new curve generated by the points C, C', etc., is a hyperbola ([13], p. 322). (In fact AB is one of the hyperbola's asymptotes, and the other asymptote is parallel to KN, as was shown by Jan van Schooten in his Latin edition of Descartes'

Geometry ([13], p. 55n).) If instead of the straight line *KNC*, one uses a parabola whose axis is the straight line *KB*, the new curve constructed by the device can again be identified once its equation is found. In this case, Descartes showed by his usual method that the curve produced was precisely the cubic curve of (5) that he got for the simple five-line problem! ([13], p. 322.)

This coincidence must have suggested to Descartes that his construction method could obtain any desired curve. Also, using algebra, Descartes showed that his device would produce curves of successively higher degrees ([13], p. 321–323). For instance, when *KN* was a straight line, it produced a curve represented by a quadratic; when *KN* was a parabola, it produced a third-degree curve. Descartes, struck by the generality of these results, said that any algebraic curve could be defined as a Pappus *n*-line locus ([13], p. 324), but here he went too far. (For a proof that this is incorrect, see [3], pp. 332–338; incidentally, Newton was the first to try to prove that Descartes was in error on this point ([3], p. 338).) Descartes also seems to have believed that any curve with an algebraic equation could be constructed by one of his devices. And here he was right, as was shown in the nineteenth century by A. B. Kempe ([22], cited in [3], p. 324). Thus Descartes' methods really did yield results of the generality he sought. We can now understand and appreciate the claim with which Descartes' *Geometry* begins: "Every problem in geometry can easily [!] be reduced to such terms that a knowledge of the lengths of certain straight lines is sufficient for its construction." (See [13], p. 297.)

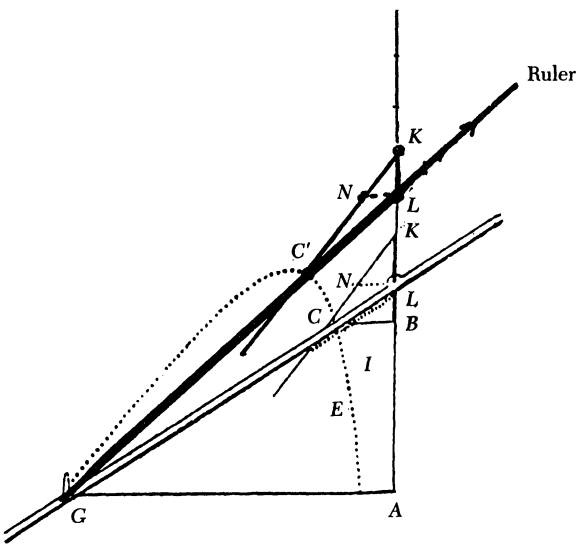


FIGURE 7

The Power of Descartes' Methods: Tangents and Equations

Descartes held that curves were admissible in geometry only if they could be constructed, but of course he also had equations for them. Thus the study of the curves, and of many of their properties, could be advanced by the study of the corresponding equations. Let us briefly consider one example where Descartes did this.

All properties of geometric curves he had not yet discussed, he said, depend on the angles curves make with other curves ([13], pp. 341–342). This problem could be completely solved, he continued, if the normal to a curve at a given point could be found. The reader will recognize that this is an example of the reduction of one problem to another. And how does one find the normal to a curve? Again, by a reduction. It is easy to find the normal to a circle, so we can find the normal to a curve at a point by finding the normal to the circle tangent to the curve at the same point. Thus we must find such a tangent circle. And how did Descartes begin his search for that circle? By yet another reduction, this time to algebra: He sought an algebraic equation for the circle tangent to the given curve at the given point.

He did this by starting with a circle that hit the curve at two points, and then letting the two points get closer and closer together. This required, first, writing an algebraic equation for a circle that hit the curve twice. The equation for the points of intersection of that circle and the original curve would have two solutions. But "the nearer together the points ... are taken, the less difference there is between the roots; and when the points coincide, the roots are equal" —that is, the equation has only one solution when the points coincide, and thus has only one solution when the intersecting circle becomes the tangent circle ([13], pp. 346–7). To find when the two solutions of the algebraic equation became one, Descartes in effect set the discriminant equal to zero, providing another demonstration of the power of algebraic methods to solve geometric problems. Thus, the algebraic equation let him find the tangent circle. Finally, the normal to that circle at the point of tangency gave him the normal to the curve ([6], pp. 94–95). Quite a triumph for the method of reduction!

Descartes applied this technique to find normals to several curves. For instance, he did it for the so-called ovals of Descartes ([13], pp. 360–2), whose properties, including normals, he used in optics. He also discussed finding the normal to the cubical parabola whose equation is (5) ([13], pp. 343–4). Descartes' method was the first treatment of a tangent as the limiting position of a secant to appear in print ([6], p. 95). Thus his method of normals was a step in the direction of the calculus, as was Fermat's contemporary, independent, simpler, and more elegant method of tangents ([6], pp. 80, 94–5; [30], pp. 165–169; [5], pp. 166–169, 157–8).

There is one more important class of problems taken up in Descartes' *Geometry*, the solution of algebraic equations. As we have mentioned, classical problems like duplicating the cube required solving equations. So did constructing arbitrary points on the curve that solved a locus problem. Descartes said in fact that "all geometric problems reduce to a single type, namely the question of finding the roots of an equation." (See [13], p. 401.) Since this process was so important, if one were given an equation, it would be good to learn as much about the solutions as possible before trying to construct them geometrically.

In the last section of the *Geometry*, Descartes tried to do just this, by developing a great deal of what is now called the theory of equations. One example will suffice to illustrate his approach:

$$(x-2)(x-3)(x-4)(x+5)=0. \tag{6}$$

Using this numerical example and multiplying it out, he obtained

$$x^4-4x^3-19x^2+106x-120=0. \tag{7}$$

Descartes pointed out that one can see from the way the polynomial in (7) is generated from (6) that it has three positive roots and one negative one, and that the number of positive roots is given by the number of changes of sign of the coefficients

(this is the principal case of what is now called Descartes' Rule of Signs). Also, a polynomial with several roots is divisible by x minus any root, and it can have as many distinct roots as its degree ([13], pp. 372–4). Descartes was not the first to have pointed out these things, but his presentation was systematic and influential, and the context made clear the importance of the results. The algebra was not an end in itself; it was all done to solve geometric problems.

The last major class of problems addressed in the *Geometry* was constructing the roots of equations of degree higher than two. Going beyond the Greek example of a cubic solved by intersecting conics, Descartes solved fifth- and sixth-degree equations. Why? They come up, he said, in geometry, if one tries to divide an angle into five equal parts ([13], pp. 412), or if one tries to solve the Pappus 12-line problem ([13], p. 324). To illustrate his solution method, he solved a sixth-degree equation with six positive roots by using intersecting cubic curves. The curve he used was not $y = x^3$, which we might think of as simple, but the cubics he had defined as the intersections of moving conic sections and lines. In FIGURE 8, the diagram for one such solution is shown ([13], p. 404). The cubic curve, a portion of which is shown as *NCQ*, intersects the circle *QNC* at the points that solve the sixth-degree equation. The cubic curve involved in this construction, generated by the motion of the parabola *CDF*, is the cubic curve (5) once again.

Descartes said that he could construct the solution to every problem in geometry. We can now see why he thought that!

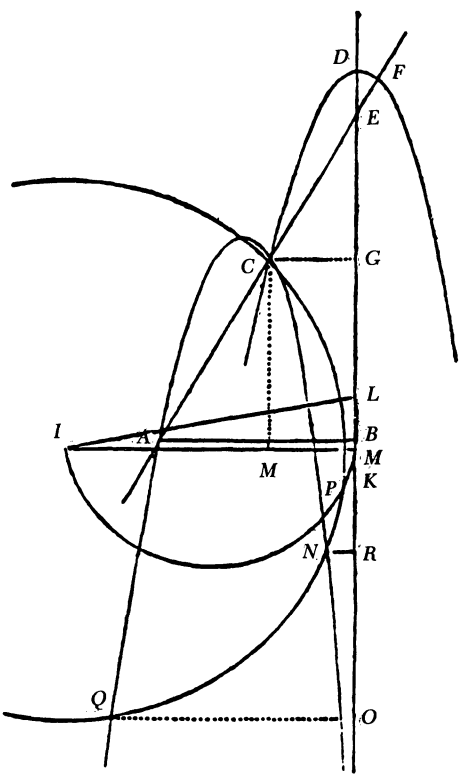


FIGURE 8

Conclusion

Now that we have seen Descartes in action, let us assess his influence on problem-solving. First, consider the mathematics that we now call "analysis." Descartes' *Geometry* solved hard problems by novel methods. There was, as an additional aid for

his successors, the simultaneous and analogous work of Fermat; though Fermat's work on analytic geometry, tangents, and areas was not printed until the 1670s, it was circulated among mathematicians in the 1630s and 1640s and exerted great influence. *Geometry* itself attracted many followers. Continental mathematicians, especially Frans van Schooten and Florimond Debeaune, wrote commentaries and added explanations for Descartes' often cryptic statements. They also extended Descartes' methods to construct other loci. The second edition of Schooten's commentary on Descartes' *Geometry* (with a Latin translation) was published in 1659–1661 together with several other influential works based on Descartes. One was Jan de Witt's *Elements of Curves*, which systematized analytic geometry, including a discussion of constructing conic sections from their equations ([**6**], p. 115–116); another was Hendrik van Heuraet's work on finding arc lengths. Schooten's collection helped inspire both John Wallis and Isaac Newton. Wallis "seized upon the methods and aims of Cartesian geometry" ([**6**], p. 109) and went even further in replacing geometric concepts by algebraic or arithmetic ones. Many mid-seventeenth-century mathematicians, including Wallis, James Gregory, and Christopher Wren, influenced both by Descartes and by Fermat, used algebraic methods to make further progress on the problem of tangents, and—as Descartes had suggested, but did not do—to find areas. Men like van Heuraet, William Neil, and Wren also found arc lengths for some curves this way ([**5**], p. 162), which Descartes, who couldn't do it, had said couldn't be done ([**13**], p. 340). Wallis also extended the algebraic approach of Descartes to infinitesimals. In the 1660s, Isaac Newton carefully studied Schooten's edition of Descartes, using it (together with the work of men like Barrow, Wallis, and Gregory) as a key starting point in his invention of the calculus ([**35**], pp. 106–111, 128–130). In 1674, less than two years before his own invention of the calculus, Gottfried Wilhelm Leibniz worked his way through Descartes' *Geometry*; he was especially interested in the algebraic ideas ([**21**], p. 143). He later even examined some of Descartes' unpublished manuscripts ([**21**], p. 182–183).

Some scholars have credited Descartes with bringing about a revolution in analysis ([**7**], pp. 157–159, 506; [**3**], p. 304; for dissenting views, see [**31**], pp. 110–111; [**21**], pp. 202–210; [**18**], p. 55; [**19**], p. 164). But at the very least we may say of the *Geometry* what Thomas Kuhn once said about Copernicus' *On the Revolution of the Celestial Orbs* ([**26**], p. 134); though it may not have been revolutionary, it was "a revolution-making text." The problem-solving methods introduced in Descartes' *Geometry* and developed in the commentaries on it were clearly seminal throughout the seventeenth century, influencing both Newton and Leibniz, whether or not Descartes was the first inventor of these techniques. And such influence continued through the eighteenth century and beyond ([**17**], pp. 156–158, 505–507).

Incidentally, the systematic approach to analytic geometry we all learned in school is not in either Descartes or Fermat (though Fermat, unlike Descartes, did plot elementary curves from their equations), but dates from various eighteenth-century textbooks, especially those from the hands of Euler, Monge, Lagrange, and Lacroix ([**16**], pp. 192–224). Descartes, though, was not a textbook writer, but a problem-solver. The essence of his influence was in his new approach and his self-consciousness about method. These highlight his achievement as a problem-solver.

Second, then, let us look at his influence on problem-solving in general. The problem-solving methods we teach our students are the direct descendants of Descartes' methods. This is not because he passed them down to us in a set of rules (although he did). Nor is it because his methods work for the problems in elementary textbooks (although they do). It is because his methods solved many outstanding problems of his day. Descartes saw himself as a problem-solver because he had a

method. He saw himself also as a teacher of problem-solving. One can see this even in the way he left hard questions as exercises to the reader, as he put it at the end of the *Geometry*, "to leave for others the pleasures of discovery." (See [13], p. 413.) His *Geometry* teaches us how to solve problems because it contains a set of solved problems whose successful solutions validate his methods. We may not care about the Pappus four-line problem, but we certainly prize the problem-solving power of a generalized algebra. Descartes' methods have come to us indirectly—who reads the *Geometry* nowadays?—but they have come to us because they are embedded in the work of his successors: In algebraic notation and equation theory, in analytic geometry, in calculus, in Lagrange's view that algebra is the study of general systems of operations, and in the more abstract and general subjects built upon these achievements. Because of his influence on later mathematicians, Descartes' methods are embedded also in the way we teach mathematics, in the standard collections of problems and solutions. In fact, for routine problems, the task of applying Descartes' analytic methods is, as he intended, fairly mechanical. Some of the *Rules for Direction of the Mind* explicitly parallel the method of the *Geometry*, ([2], pp. 177–178) and Pólya is thus right to have made such rules explicit for modern students. Inventing new mathematical methods—say, like analytic geometry— is, however, not a routine task. Even here, for Descartes, "method" is crucial.

Third, then, for those of us who want to invent great and new things like analytic geometry, to teachers and students of mathematics, Descartes has something else he wants us to learn, and that is his emphasis on method in general. Here he, together with his great contemporary Sir Francis Bacon, have inspired many. For instance, Leibniz saw his differential calculus as a problem-solving method, explicitly comparing it with analytic geometry, saying "From [my differential calculus] flow all the admirable theorems and problems of this kind with such ease that there is no more need to teach and retain them than for him who knows our present algebra to memorize many theorems of ordinary geometry" ([27], excerpted in [34], p. 281). Or, in our century, there is Pólya's sophisticated emphasis on teaching about method. Let me put Descartes' lesson this way: Raise problem-solving techniques to consciousness. Reflect on the methods that are successful and on their strengths and weaknesses. Then apply them systematically in attacking new problems. That is how Descartes himself invented analytic geometry, as he said in the *Discourse on Method*: "I took the best traits of geometrical analysis and algebra, and corrected the faults of one by the other." (See [12], p. 13, 20.)

Fourth and last, let us briefly consider a key point in Descartes' philosophy: that the methods of mathematics could solve the problems of science. Here, Descartes the philosopher learned from Descartes the mathematician that method was important, that the right method could solve previously intractable problems. He used the ideas of reduction and analysis in his philosophy of science. For instance, he argued that all macroscopic phenomena could be explained by analyzing nature into its component parts, bits of matter in motion. (See [14], pp. 409–414) and ([36], pp. 32–38). Descartes came to believe that the most powerful methods were both general and mathematical. His *Principles of Philosophy* (1644) attempted to deduce all the laws of nature from self-evident first principles; his principles XXXVII and XXXIX are equivalent to Newton's First Law of Motion (1687) ([8], pp. 182–183).) In fact, Descartes went so far as to state that everything that could be known could be found by a method modelled on that of mathematics. He wrote,

> Those long chains of reasoning, so simple and easy, which enabled the geometers to reach the most difficult demonstrations, had made me

> wonder whether all things knowable to man might not fall into a similar logical sequence. If so, we need only refrain from accepting as true that which is not true, and carefully follow the order necessary to deduce each one from the others, and there cannot be any propositions so abstruse that we cannot prove them, not so recondite that we cannot discover them ([**12**], pp. 12–13; 19).

Descartes' vision is clearly echoed by what Leibniz wrote in 1677 about his own search for a general symbolic method of finding truth: "If we could find characters or signs appropriate for expressing all our thoughts as definitely and as exactly as arithmetic expresses numbers or geometric analysis expresses lines, we could in all subjects in so far as they are amenable to reasoning accomplish what is done in Arithmetic and Geometry." (See [**28**], p. 15.) Again, consider the prediction of the great prophet of progress of the Enlightenment, the Marquis de Condorcet, that Descartes' methods could solve all problems. Although the "method" of algebra "is by itself only an instrument pertaining to the science of quantities," Condorcet wrote, "it contains within it the principles of a universal instrument, applicable to all combinations of ideas." This could make the progress of "every subject embraced by human intelligence... as sure as that of mathematics." (See [**9**], pp. 238, 278–279; quoted in [**17**], p. 222.)

Descartes has been attacked as a methodological imperialist and a reductionist, and lauded as an intellectual liberator and one of the founders of modern thought (e.g., [**11**], [**18**], [**24**], [**33**]). For good or ill, the power of Descartes' vision has shaped Western thought since the seventeenth century, and his mathematical work helped inspire his philosophy. But whatever our assessment of Descartes the philosopher may be, his importance for the mathematician is clear. The history of the past 350 years of mathematics can fruitfully be viewed as the story of the triumph of Descartes' methods of problem-solving.

Acknowledgement. I thank Professor Tatiana Deretsky for suggesting this topic to me, and two MATHEMATICS MAGAZINE referees for helpful suggestions.

REFERENCES

1. Bacon, Francis, *Novum Organum* [1620]. Often reprinted, e.g., in E. A. Burtt, ed., *The English Philosophers from Bacon to Mill*, Modern Library, New York, 1913, pp. 24–123.
2. Beck, L. J., *The Method of Descartes: A Study of the Regulae*, Clarendon, Oxford, 1952. [Note: Descartes' *Rules for Direction of the Mind* (*Regulae*) were written about 1628, and published posthumously in 1701].
3. Bos, H. J. M., "On the representation of curves in Descartes' *Géométrie*," *Arch. Hist. Ex. Sci.* 1981, 295–338.
4. Boyer, Carl B., "Analysis: Notes on the evolution of a subject and a name," *Math Teacher* XLVII (1954), 450–62.
5. _____, *The Concepts of the Calculus*, Columbia, New York, 1939.
6. _____, *History of Analytic Geometry*, Scripta Mathematica, New York, 1956.
7. Cohen, I. B., *Revolution in Science*, Harvard, Cambridge, MA, 1985.
8. _____, *The Newtonian Revolution, with Illustrations of the Transformation of Scientific Ideas*, Cambridge University Press, Cambridge, MA, 1980.
9. Condorcet, Marquis de, *Sketch for a Historical Picture of the Progress of the Human Mind*, 1793, tr. J. Barraclough, in Keith Baker, ed., *Condorcet: Selected Writings*, Bobbs-Merrill, New York, 1976.
10. Coolidge, J. L., *A History of Geometrical Methods*, Oxford University Press, Oxford, 1940.
11. Davis, P. J., and Hersh, Reuben, *Descartes' Dream*, Harcourt, Brace, Jovanovich, San Diego and New York, 1986.

12. Descartes, René, *Discourse on the Method of Rightly Conducting the Reason to Seek the Truth in the Sciences*, 1637, tr. L. J. Lafleur, Bobbs-Merrill, New York, 1956. The first set of page numbers in each citation in the present paper are from this translation; the second set are from the edition of Ch. Adam et P. Tannery, eds., *Oeuvres de Descartes*, Paris, 1879–1913, Vol. VI.
13. Descartes, René, *The Geometry*, tr. from the French and Latin by D. E. Smith and M. L. Latham, Dover Reprint, New York, 1954. Contains a facsimile reprint of the original 1637 French edition. In the present paper, page references from Decartes, which appear in the Dover reprint, are given from the French edition, while citations from the Smith-Latham commentary are identified by the page numbers from the Dover reprint itself.
14. Dijksterhuis, E. J., *The Mechanisation of the World-Picture*, Tr. C. Dikshoorn, Oxford University Press, Oxford, 1961.
15. Gaukroger, Stephen, ed., *Descartes: Philosophy, Mathematics, and Physics*, Barnes and Noble, New York, 1980.
16. Gillies, Donald, ed., *Revolutions in Mathematics*, Oxford University Press, Oxford, 1992.
17. Grabiner, Judith V., "The centrality of mathematics in the history of Western thought," this MAGAZINE 61 (1988), pp. 220–230.
18. Grosholz, Emily, *Cartesian Method and the Problem of Reduction*, Clarendon Press, Oxford, 1991.
19. _____, "Descartes' Unification of Algebra and Geometry," in [**15**, pp. 156–168].
20. Heath, Thomas L., *Greek Mathematics*, 2 vols., Clarendon Press, Oxford, 1921.
21. Hofmann, J. E., *Leibniz in Paris*, 1672–1676: *His Growth to Mathematical Maturity*, Tr. A. Prag and D. T. Whiteside, Cambridge University Press, Cambridge, 1974.
22. Kempe, A. B., "On a general method of describing plane curves of the nth degree by linkwork," *Proc. Lond. Math. Soc.* 7 (1876), 213–16.
23. Klein, Jacob, *Greek Mathematical Thought and the Origin of Algebra*, The MIT Press, Cambridge, MA, 1968.
24. Kline, Morris, *Mathematics in Western Culture*, Oxford University Press, Oxford, 1964.
25. Knorr, Wilbur, *The Ancient Tradition of Geometric Problems*, Birkhäuser, Boston, 1986.
26. Kuhn, Thomas S., *The Copernican Revolution*, Harvard, Cambridge, MA, 1957.
27. Leibniz, G. W., "De geometria recondita et analysi indivisibilium atque infinitorum," *Acta Eruditorum* 5 (1686). Excerpted in [**34**, pp. 281–282].
28. _____, "Preface to the General Science," in [**37**, pp. 12–17].
29. Mahoney, Michael S., "The Beginnings of Algebraic Thought in the Seventeenth Century," in [**15**, pp. 141–155].
30. _____, *The Mathematical Career of Pierre de Fermat, 1601–65*, Princeton University Press, Princeton, NJ, 1973.
31. Mancosu, Paolo, "Descartes' *Géométrie* and revolutions in mathematics," [**16**, pp. 83–116].
32. Polya, George, *Mathematical Discovery: On Understanding, Learning, and Teaching Problem Solving*, John Wiley & Sons, Inc., New York, 1981.
33. Russell, Bertrand, *A History of Western Philosophy*, Simon and Schuster, New York, 1945.
34. Struik, Dirk J., *A Source Book in Mathematics, 1200–1800*, Harvard, Cambridge, MA, 1969.
35. Westfall, Richard S., *Never At Rest: A Biography of Isaac Newton*, Cambridge University Press, Cambridge, 1980.
36. _____, *The Construction of Modern Science*, John Wiley&Sons, Inc., New York, 1971.
37. Wiener, P., ed. *Leibniz: Selections*, Scribner's New York, 1951.

Math Bite: $\Sigma a_k b_k \leqslant (\Sigma a_k^2)^{1/2}(\Sigma b_k^2)^{1/2}$.

$$\frac{\Sigma a_k b_k}{(\Sigma a_k^2)^{1/2}(\Sigma b_k^2)^{1/2}} = 1 - \frac{1}{2}\Sigma\left(\frac{a_k}{(\Sigma a_k^2)^{1/2}} - \frac{b_k}{(\Sigma b_k^2)^{1/2}}\right)^2.$$

—PETER SZÜSZ
SUNY AT STONY BROOK
STONY BROOK, NY 11794

12. Descartes, René, *Discourse on the Method of Rightly Conducting the Reason to Seek the Truth in the Sciences*, 1637, tr. L. J. Lafleur, Bobbs-Merrill, New York, 1956. The first set of page numbers in each citation in the present paper are from this translation; the second set are from the edition of Ch. Adam et P. Tannery, eds., *Oeuvres de Descartes*, Paris, 1879–1913, Vol. VI.
13. Descartes, René, *The Geometry*, tr. from the French and Latin by D. E. Smith and M. L. Latham, Dover Reprint, New York, 1954. Contains a facsimile reprint of the original 1637 French edition. In the present paper, page references from Decartes, which appear in the Dover reprint, are given from the French edition, while citations from the Smith-Latham commentary are identified by the page numbers from the Dover reprint itself.
14. Dijksterhuis, E. J., *The Mechanisation of the World-Picture*, Tr. C. Dikshoorn, Oxford University Press, Oxford, 1961.
15. Gaukroger, Stephen, ed., *Descartes: Philosophy, Mathematics, and Physics*, Barnes and Noble, New York, 1980.
16. Gillies, Donald, ed., *Revolutions in Mathematics*, Oxford University Press, Oxford, 1992.
17. Grabiner, Judith V., "The centrality of mathematics in the history of Western thought," this MAGAZINE 61 (1988), pp. 220–230.
18. Grosholz, Emily, *Cartesian Method and the Problem of Reduction*, Clarendon Press, Oxford, 1991.
19. _____, "Descartes' Unification of Algebra and Geometry," in [**15**, pp. 156–168].
20. Heath, Thomas L., *Greek Mathematics*, 2 vols., Clarendon Press, Oxford, 1921.
21. Hofmann, J. E., *Leibniz in Paris*, 1672–1676: *His Growth to Mathematical Maturity*, Tr. A. Prag and D. T. Whiteside, Cambridge University Press, Cambridge, 1974.
22. Kempe, A. B., "On a general method of describing plane curves of the nth degree by linkwork," *Proc. Lond. Math. Soc.* 7 (1876), 213–16.
23. Klein, Jacob, *Greek Mathematical Thought and the Origin of Algebra*, The MIT Press, Cambridge, MA, 1968.
24. Kline, Morris, *Mathematics in Western Culture*, Oxford University Press, Oxford, 1964.
25. Knorr, Wilbur, *The Ancient Tradition of Geometric Problems*, Birkhäuser, Boston, 1986.
26. Kuhn, Thomas S., *The Copernican Revolution*, Harvard, Cambridge, MA, 1957.
27. Leibniz, G. W., "De geometria recondita et analysi indivisibilium atque infinitorum," *Acta Eruditorum* 5 (1686). Excerpted in [**34**, pp. 281–282].
28. _____, "Preface to the General Science," in [**37**, pp. 12–17].
29. Mahoney, Michael S., "The Beginnings of Algebraic Thought in the Seventeenth Century," in [**15**, pp. 141–155].
30. _____, *The Mathematical Career of Pierre de Fermat, 1601–65*, Princeton University Press, Princeton, NJ, 1973.
31. Mancosu, Paolo, "Descartes' *Géométrie* and revolutions in mathematics," [**16**, pp. 83–116].
32. Polya, George, *Mathematical Discovery: On Understanding, Learning, and Teaching Problem Solving*, John Wiley & Sons, Inc., New York, 1981.
33. Russell, Bertrand, *A History of Western Philosophy*, Simon and Schuster, New York, 1945.
34. Struik, Dirk J., *A Source Book in Mathematics, 1200–1800*, Harvard, Cambridge, MA, 1969.
35. Westfall, Richard S., *Never At Rest: A Biography of Isaac Newton*, Cambridge University Press, Cambridge, 1980.
36. _____, *The Construction of Modern Science*, John Wiley&Sons, Inc., New York, 1971.
37. Wiener, P., ed. *Leibniz: Selections*, Scribner's New York, 1951.

Math Bite: $\Sigma a_k b_k \leqslant (\Sigma a_k^2)^{1/2}(\Sigma b_k^2)^{1/2}$.

$$\frac{\Sigma a_k b_k}{(\Sigma a_k^2)^{1/2}(\Sigma b_k^2)^{1/2}} = 1 - \frac{1}{2}\Sigma\left(\frac{a_k}{(\Sigma a_k^2)^{1/2}} - \frac{b_k}{(\Sigma b_k^2)^{1/2}}\right)^2.$$

—PETER SZÜSZ
SUNY AT STONY BROOK
STONY BROOK, NY 11794

The Ptolemy Inequality and Minkowskian Geometry

JOHN D. SMITH
Winchester College,
Winchester, Hampshire, S023 9NA, England

We give a novel interpretation of the Euclidean Ptolemy inequality, which we then extend further; the reversed triangle inequality in Minkowskian geometry is the unexpected key to the extension. This connection between the two geometries shows up again when we consider Ptolemy's theorem with a Minkowskian metric.

1. Introduction

In his *Almagest*, the celebrated astronomical treatise written about A.D. 150, Ptolemy of Alexandria constructs his trigonometric tables with the aid of a beautiful geometrical identity (FIGURE 1):

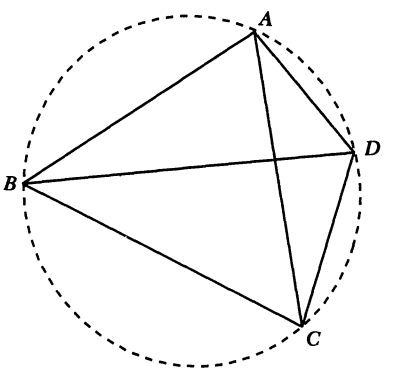


FIGURE 1

If ABCD is a quadrilateral inscribed in a circle,

$$|AB| \cdot |CD| + |BC| \cdot |DA| = |AC| \cdot |BD|, \tag{1.1}$$

where $|AB|$ *denotes Euclidean length.*

'Ptolemy's theorem' (we do not know whether he actually discovered it) proves to be so useful because it is equivalent to the trigonometric addition formulae. If BD is a diameter of length $2R$ and angle $ADB = \xi$, angle $BDC = \eta$ then $\sin\xi\cos\eta + \cos\xi\sin\eta = \sin(\xi + \eta)$ follows easily. Ptolemy's methods are described in [1] and [3]; his tables were used in astronomical calculations for a thousand years.

We call a quadrilateral 'cyclic' if its vertices lie in the given order on a circle. When the quadrilateral $ABCD$ is *not* cyclic the sides and diagonals satisfy a strict *in*equality

$$|AB| \cdot |CD| + |BC| \cdot |DA| > |AC| \cdot |BD|, \tag{1.2}$$

which holds even when the points are not coplanar. We now include this inequality in the definition of Ptolemy's theorem. The identity and inequality are conveniently proved together by the method of inversion, as we show in Section 2 below.

Suppose that the points are all coplanar and A, B, C lie on the unit circle with polar angles α, β, γ. If D is renamed P and has polar coordinates $[p, \phi]$ with respect to the origin O (FIGURE 2), from the cosine formula

$$|PA| = \left[p^2 + 1 - 2p\cos(\alpha - \phi)\right]^{1/2} = (2p)^{1/2}\left[q - \cos(\alpha - \phi)\right]^{1/2},$$

where $q = (p^2 + 1)/2p = \frac{1}{2}(p + p^{-1})$. Hence (1.1) and (1.2) are equivalent to

$$|BC| \cdot \left[q - \cos(\alpha - \phi)\right]^{1/2} - |CA| \cdot \left[q - \cos(\beta - \phi)\right]^{1/2} + |AB| \cdot \left[q - \cos(\gamma - \phi)\right]^{1/2} \geq 0. \tag{1.3}$$

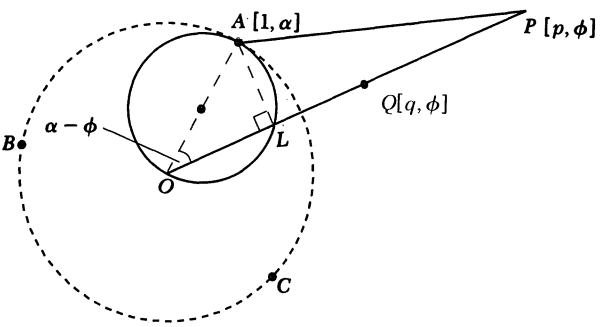


FIGURE 2

We define Q as the point on OP with polar coordinates $[q, \phi]$. The minimum value of q is 1, which is attained when $p = 1$; therefore Q lies on or outside the circumcircle. Conversely, when Q lies outside the circumcircle there are two possible positions of P (which are inverses of each other) and Q is their midpoint. If L is the orthogonal projection of A onto OQ then $q - \cos(\alpha - \phi) = |QL|$. Now L is the point (other than O) where OQ meets the circle on OA as diameter, and since $|PA| = (2p)^{1/2}|QL|^{1/2}$, we can reformulate Ptolemy's theorem as follows (FIGURE 3(a)):

(*) *ABC is a triangle inscribed in a circle with center O. A general point Q lies in the same plane on or outside the circumcircle, and L, M, N are the points (other than O) where OQ meets the circles on OA, OB, OC as diameters. Define*

$$F(Q) = |BC| \cdot |QL|^{1/2} - |CA| \cdot |QM|^{1/2} + |AB| \cdot |QN|^{1/2}. \tag{1.4}$$

Then $F(Q) \geq 0$, with equality if, and only if, Q lies on the arc CA of the circumcircle.

The expression in (1.3) has no natural boundary at $q = 1$, and defines a smooth function when $q > \cos\alpha, \cos\beta, \cos\gamma$. Hence $F(Q)$ extends to a smooth function of Q at all points outside the circles on OA, OB, OC as diameters; these circles are square-root singularities of the function. We found the following inequality by calculation:

(**) *If Q lies inside the circumcircle and on or outside all three circles of singularity, $F(Q)$ is negative in the sector OCA and positive in the other two sectors.*

The graph in FIGURE 3(a) (not drawn strictly to scale) shows values of the function on the x-axis when the polar angles of A, B, C are 30°, 160°, 280°.

Inside the circles of singularity there are up to six 'Boolean' regions defined by the circular arcs, in each of which $F(Q)$ is nonsingular. If there is a region common to the three circles, as in FIGURE 3(b) where the polar angles are 30°, 50°, 70°, the function satisfies a further inequality:

(* * *) *At points (other than O) that lie on or inside all three circles of singularity, F(Q) is negative if this common intersection is opposite sector OCA and positive if it is opposite one of the other two sectors.*

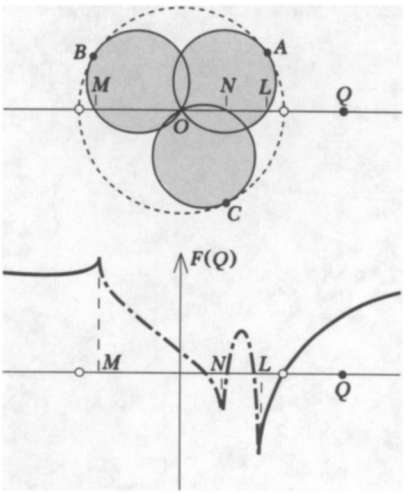


FIGURE 3a

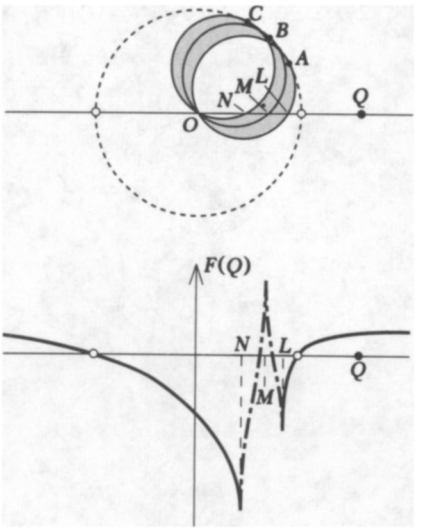


FIGURE 3b

Calculations show that general inequalities like this do not exist for the other Boolean regions.

There seemed little chance of proving the new inequalities by inversion because P is not a real point when Q lies inside the circumcircle. I was therefore overjoyed when I inverted the figure from the complex point and found that the circumcircle becomes a rectangular hyperbola, and the *reversed* inequality $F(Q) < 0$ is a consequence of the *reversed* triangle inequality in Minkowskian geometry! Ptolemy's theorem is thus a bridge between classical Euclidean geometry and the Minkowskian geometry of special relativity. A natural generalization of Ptolemy's theorem [7] can also be extended by this procedure.

Pais's definitive history [5] and Clark's biography of Einstein [2] describe Minkowski's seminal contribution to relativity in recognizing that a four-dimensional geometry binds space and time inseparably together. The text by Yaglom [7] contains a useful section on Minkowskian geometry; we summarize a few basic notions in Sections 3 and 6 below.

2. The Method of Inversion

The proof of Ptolemy's theorem by inversion can be found in [6], for example. We outline the method here, and proceed to complex inversion in Section 4. An inversion with center P and unit radius of inversion maps any point A to the point A' on the ray PA, where $|PA'| = 1/|PA|$ (FIGURE 4). If a second point B is mapped to B', it follows from the defining relations that triangles PAB and $PB'A'$ are similar with ratio of similarity $1/(|PA| \cdot |PB|)$, and hence

$$|A'B'| = \frac{|AB|}{|PA| \cdot |PB|}. \tag{2.1}$$

In FIGURE 4, the vertices A, B, C of a cyclic quadrilateral $ABCP$ are mapped to A', B', C'. By similar triangles, angles $A'B'P$ and $C'B'P$ are respectively equal to angles BAP and BCP; the latter are opposite angles of a cyclic quadrilateral and therefore supplementary. Thus $A'B'C'$ is a straight line. More generally, a circle is inverted into a straight line from a point of the circumference and into a proper circle from a point not on the circumference. The relationship

$$|A'B'| + |B'C'| = |A'C'| \tag{2.2}$$

for the three collinear points now translates into identity (1.1) after using (2.1) and replacing P by D.

An advantage of this proof is that it also gives the corresponding inequality for points that are not concyclic. If $ABCP$ is not a cyclic quadrilateral, the points A', B', C' are *not* now collinear and (1.2) follows from the triangle *in*equality corresponding to (2.2).

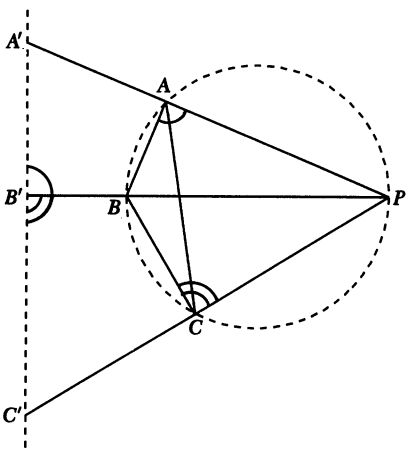


FIGURE 4

3. Minkowskian Geometry

Euclidean geometry in two dimensions can be characterized by a positive definite quadratic form $x^2 + y^2$, the Pythagorean distance function $|P_1P_2|^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$, and the transformations that leave distances invariant. These are combinations of translations, rotations

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

and reflections. The rotations and reflections map each circle $x^2 + y^2 = r^2$ onto itself. If we exclude the reflections we obtain the connected subgroup of 'oriented' Euclidean transformations.

A Euclidean transformation can be understood either as a motion of the plane (the 'alibi' interpretation) or as a transformation between two coordinate systems (the 'alias' interpretation). According to Klein in his influential *Erlanger Programm* of 1872, a 'geometrical property' of a figure is an invariant under such a group of

transformations (see, for example, [4]); thus distance and angle are properties in Euclidean geometry, but orientation (left- and right-handedness) is a property in oriented Euclidean geometry only.

In 2-dimensional *Minkowskian* geometry the interest centers on an *indefinite* quadratic form $x^2 - y^2$ and the expression $(P_1P_2)_h^2 = (x_2 - x_1)^2 - (y_2 - y_1)^2$; this may be positive, zero or negative. The Minkowskian 'h-distance' $|P_1P_2|_h$ is defined as $|(P_1P_2)_h^2|^{1/2}$.

The Minkowskian transformations, or *h-transformations*, leave $(P_1P_2)_h^2$ invariant. For simplicity we shall only consider the connected subgroup of such transformations; they are combinations of translations and *h-rotations*

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cosh \theta_h & \sinh \theta_h \\ \sinh \theta_h & \cosh \theta_h \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3.1}$$

(The full group involves h-reflections as well.) Each branch of a rectangular hyperbola $x^2 - y^2 = \pm r^2$ is transformed onto itself by an h-rotation. These hyperbolae and their translations correspond to circles in Euclidean geometry, and are called *h-circles*.

If a line-segment satisfies $(P_1P_2)_h^2 > 0$, so that its slope is less than 1 in absolute value, we call the line-segment *x-like*; if $(P_1P_2)_h^2 < 0$ and the slope is greater than 1 in absolute value we call it *y-like*; and if $(P_1P_2)_h^2 = 0$ we call the line-segment *null*. This distinction defines the *type* of a segment. An h-rotation maps every null-line of slope ± 1 onto a parallel null-line; in fact an h-rotation is simply a two-way stretch parallel to the null-directions that preserves area. Null-circles, which are translations of $x^2 - y^2 = 0$, consist of pairs of null-lines.

If the slopes of two lines are reciprocals of each other, the lines are *h-orthogonal*; this is a Minkowskian invariant. Each null-line is h-orthogonal to itself!

By an analogue of the perpendicular bisector method, for any triangle ABC the equations $(PA)_h^2 = (PB)_h^2 = (PC)_h^2$ determine the center of the *h-circumcircle* (the h-circle degenerates into two lines if $|AB|_h = 0$, for example). It is clear from FIGURE 5 that if A and B lie on the same branch of $x^2 - y^2 = r^2$, the line-segment AB is y-like, and if A and B lie on different branches then AB is x-like; the opposite holds for points on $x^2 - y^2 = -r^2$. We call a triangle *homogeneous* if the three sides are non-null and of the same type. Thus a triangle inscribed in one branch of an h-circle is homogeneous; conversely, a homogeneous triangle can be inscribed in a single branch of an h-circle.

The h-distance between points $(r \cosh a, r \sinh a)$ and $(r \cosh b, r \sinh b)$ on the right-hand branch of $x^2 - y^2 = r^2$ is $2r \sinh(\frac{1}{2}|b - a|)$. Take three points with parameters a, $b = a + 2v$ and $c = b + 2w$, where $v, w > 0$. Now $\sinh(v + w) = \sinh v \cosh w + \cosh v \sinh w > \sinh v + \sinh w$, since $\cosh v$ and $\cosh w$ exceed 1.

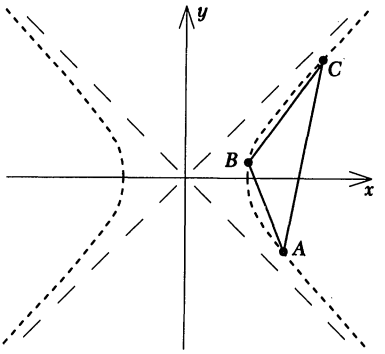


FIGURE 5

This proves the first part of the h-triangle inequality, and the second part is an easy consequence:

A homogeneous h-triangle ABC satisfies $|AB|_h + |BC|_h < |AC|_h$ *if the points lie in the given order on a branch of an h-circle, and* $|AB|_h + |BC|_h > |AC|_h$ *if the points lie in a different order.*

The usual triangle inequality is therefore *reversed* when the vertices lie in the given order.

Three-dimensional Minkowskian geometry is defined in a similar way with respect to a quadratic form $x^2 + y^2 - z^2$. The null-lines through any point lie on a 'null-cone' which is a translation of $x^2 + y^2 - z^2 = 0$, and the 'spheres' are translations of hyperboloids $x^2 + y^2 - z^2 = \pm r^2$, etc.

4. Complex Inversion and the Extended Ptolemy Inequality

We now return to the investigation of $F(Q)$ in (1.4) when Q lies inside the circumcircle; here the value of p is necessarily complex. Choose axes so that Q has Cartesian coordinates $(q, 0)$ with $0 \le q < 1$. If $p = e^{i\rho}$ then $q = \frac{1}{2}(p + p^{-1}) = \cos\rho$. We therefore invert from the complex center $P(e^{i\rho}, 0)$ and map the unit circle in $\mathbb{R}^2$ into a curve in the complex space $\mathbb{C}^2$. Almost miraculously (it seems) there is a natural interpretation of the image as an h-circle and the inequalities we seek follow from the h-triangle inequality.

The formula for inversion from the center $P(x_0, y_0)$ is

$$x' - x_0 = \frac{x - x_0}{(x - x_0)^2 + (y - y_0)^2}, \qquad y' - y_0 = \frac{y - y_0}{(x - x_0)^2 + (y - y_0)^2}.$$

When $x_0 = e^{i\rho}$, $y_0 = 0$, the image of a general point $(\cos\phi, \sin\phi)$ on the unit circle satisfies

$$x' - e^{i\rho} = \frac{\cos\phi - e^{i\rho}}{(\cos\phi - e^{i\rho})^2 + \sin^2\phi} = \frac{\cos\phi - e^{i\rho}}{e^{2i\rho} + 1 - 2\cos\phi \cdot e^{i\rho}}$$

$$= \frac{\cos\phi - e^{i\rho}}{2e^{i\rho}(\cos\rho - \cos\phi)}.$$

Similarly,

$$y' = \frac{\sin\phi}{2e^{i\rho}(\cos\rho - \cos\phi)}.$$

If $x_1 = 2e^{i\rho}(x' - e^{i\rho})$, $y_1 = 2e^{i\rho}y'$ then

$$x_1 = \frac{\cos\phi - e^{i\rho}}{\cos\rho - \cos\phi} = \frac{\cos\phi - \cos\rho - i\sin\rho}{\cos\rho - \cos\phi},$$

and hence

$$x_1 = -1 - \frac{i\sin\rho}{\cos\rho - \cos\phi}, \qquad y_1 = \frac{\sin\phi}{\cos\rho - \cos\phi}.$$

The average of the values of x_1 at $\phi = 0$ and $\phi = \pi$ is $c_1 = -1 + i\cot\rho$. We now "center" the image and define $x_h = i\sin\rho(x_1 - c_1)$, $y_h = \sin\rho\, y_1$, so that

$$x_h = \frac{1 - \cos\rho\cos\phi}{\cos\rho - \cos\phi}, \qquad y_h = \frac{\sin\rho\sin\phi}{\cos\rho - \cos\phi}. \tag{4.1}$$

It is easy to check that the *real* point (x_h, y_h) lies on the h-circle $x_h^2 - y_h^2 = 1$.

We are interested in a triangle ABC with polar angles α, β, γ. If Q lies *outside* all the circles on diameters OA, OB, OC then $q > \cos\alpha, \cos\beta, \cos\gamma$, and if Q lies *inside* all these circles then $q < \cos\alpha, \cos\beta, \cos\gamma$; in both cases there is an arc of the circle containing A, B, and C that is mapped continuously by (4.1) to a single branch of the h-circle.

The mapping is equivalent to the following: Invert in the complex center $(e^{i\rho}, 0)$, perform a complex translation, then a complex enlargement with scale-factor $2e^{i\rho}\sin\rho$, and finally multiply the x-coordinate by i. (The transformation is singular if $\sin\rho = 0$, when the point of inversion is $(\pm 1, 0)$; these are the cases where the inverse-image of the circle is a straight line, and may be regarded as *both* a circle *and* an h-circle!) If A', B', C' are the images under the complex inversion, (2.1) continues to hold:

$$|A'B'|^2 = \frac{|AB|^2}{|PA|^2 \cdot |PB|^2}.$$

From the sequence of transformations above,

$$|A'B'|^2 = -(2e^{i\rho}\sin\rho)^{-2}(A_hB_h)_h^2.$$

Also

$$\begin{aligned} |PA|^2 &= p^2 + 1 - 2p\cos\alpha = (2p)(q - \cos\alpha) \\ &= (2e^{i\rho})(q - \cos\alpha). \end{aligned}$$

Hence

$$\frac{|A_hB_h|_h}{|\sin\rho|} = \frac{|AB|}{|q - \cos\alpha|^{1/2} \cdot |q - \cos\beta|^{1/2}}.$$

Now if Q lies in the sector OCA, the points A_h, B_h, C_h lie in order on the h-circle and $|A_hB_h|_h + |B_hC_h|_h < |A_hC_h|_h$; the result $F(Q) < 0$ follows from the h-triangle inequality for $A_hB_hC_h$ just as in Section 2, except that now the inequality is reversed! When Q lies in one of the other two sectors, A_h, B_h, C_h do *not* lie in that order on the h-circle and the non-reversed h-triangle inequality gives $F(Q) > 0$. By a similar argument, if Q lies in the common intersection of the circles, $F(Q) < 0$ if the intersection is opposite sector OCA and $F(Q) > 0$ if the intersection is opposite one of the other sectors.

$F(Q)$ is continuous except at the origin, and it is continuous at the origin when restricted to any particular diameter. If Q lies *on* one (or more) of the three circles there are two (or fewer) terms left in $F(Q)$. We can still use inversion and show that the inequalities are *strictly* satisfied. This completes the proof of extensions $(**)$ and $(***)$ in Section 1.

5. Increasing the Dimension of the Domain

Suppose that the point P in Section 1 lies in space (FIGURE 6), with spherical polar coordinates $[p, \theta, \phi]$; Q is defined by $[q, \theta, \phi]$, where $q = \frac{1}{2}(p + p^{-1})$. Then

(i) $$\begin{aligned} |PA| &= [p^2 + 1 - 2p\sin\theta\cos(\alpha - \phi)]^{1/2} \\ &= (2p\sin\theta)^{1/2}[q/\sin\theta - \cos(\alpha - \phi)]^{1/2}. \end{aligned}$$

(ii) If $q/\sin\theta = 1$ then Q lies on the surface of a 'doughnut without a hole'; to generate the surface, draw a vertical circle on OA as diameter (for example) and rotate it about the z-axis. The vertical sections through OA, OB and OC divide the surface and its interior into three sectors.

(iii) If $q/\sin\theta = \cos(\alpha - \phi)$ then Q lies on a sphere with diameter OA.

A comparison with (1.3) shows that inequalities $(*)$, $(**)$, and $(***)$ still hold if Q is a point in space and we replace the circumcircle by the surface of the doughnut, and the circles of singularity by three spheres. (The same technique works if ABC is embedded in a space of any dimension.)

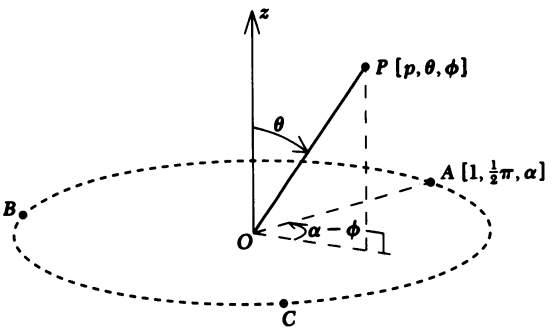


FIGURE 6

6. Ptolemy's Theorem in the Minkowskian Plane

We shall now use 'h-inversion' to derive an analogue of Ptolemy's theorem in the Minkowskian plane. A different coordinate system, with axes along the lines $y = -x$ and $y = x$, makes many of the formulae easier. If $X = x - y$ and $Y = x + y$ then $x^2 - y^2 = 1$ becomes $XY = 1$ (FIGURE 7). The null-lines are parallel to the new coordinate axes and an h-rotation has the simple form $(X, Y) \to (kX, k^{-1}Y)$, as a two-way stretch that preserves area.

If the center of an h-inversion is (x_0, y_0), then

$$x' - x_0 = \frac{x - x_0}{(x - x_0)^2 - (y - y_0)^2}, \qquad y' - y_0 = \frac{y - y_0}{(x - x_0)^2 - (y - y_0)^2},$$

or

$$X' - X_0 = \frac{1}{Y - Y_0}, \qquad Y' - Y_0 = \frac{1}{X - X_0}.$$

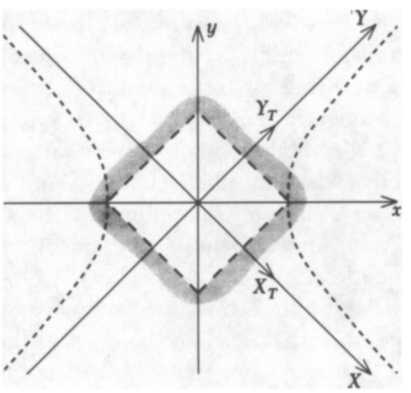


FIGURE 7

Inversion maps one set of parallel null-lines onto the *other* set of parallel null-lines; lines close to (X_0, Y_0) become far-away and vice versa. It is straightforward to verify that an h-circle S is mapped to a straight line if the center of inversion lies on S, and is mapped to another h-circle S' otherwise.

We gain a better understanding of h-inversion by 'completing' the plane. The transformation $(X, Y) \to (X_T, Y_T)$ where

$$(X_T, Y_T) = \left(\frac{X}{|X| + 1}, \frac{Y}{|Y| + 1} \right)$$

maps the whole plane onto the *open* square $|X| < 1$, $|Y| < 1$, and every null-line onto an open interval (FIGURE 7). The two branches of $x^2 - y^2 = 1$ are mapped to the line segments $X_T + Y_T = \pm 1$, but in general an h-circle is mapped to a curve, as shown in FIGURE 8.

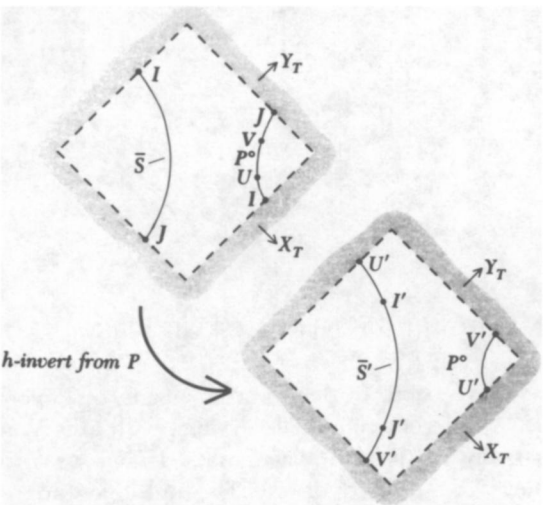


FIGURE 8

For large values of $|X|$ or $|Y|$ the point (X_T, Y_T) approaches the boundary of the square, and if we identify the opposite edges of the square and make a torus, the null-lines are the lines of latitude and longitude. The identified edges of the closed square are thus 'infinite' null-lines of the completed plane. A special doubly infinite point Ω is represented by all four corners of the square. As we might expect, the branches of an h-circle S meet at the two infinite points on its asymptotes, and with these points S becomes a closed curve $\bar{S}$. In the first diagram of FIGURE 8 the points I and J are the infinite points of $\bar{S}$.

The second diagram of FIGURE 8 shows how S is transformed to S' by h-inversion from the center P. The transformation is singular on the null-lines through P; they are mapped to the infinite null-lines at the boundary of the square. The finite points U, V where the null-lines meet S are mapped to the infinite points U', V' of the completed h-circle $\bar{S}'$, and the infinite points I, J of $\bar{S}$ are mapped to finite points I', J' of $\bar{S}'$. The two arcs of $\bar{S}$ with end-points U and V are thus transformed into the branches of S', and there is a continuous mapping of $\bar{S}$ to $\bar{S}'$. The center of inversion P becomes Ω, which has a similar role to the north pole in the stereographic compactification of the Euclidean plane.

Suppose that ABC is a triangle inscribed in S (not necessarily on one branch), and we invert from a general point P that does not lie on S. If the images A', B', C' form a homogeneous triangle, the h-triangle inequality translates into an inequality for the expression $|BC|_h \cdot |PA|_h - |CA|_h \cdot |PB|_h + |AB|_h \cdot |PC|_h$.

Now $A'B'C'$ is homogeneous if, and only if, A, B, C all lie on one of the arcs of S defined by U and V; equivalently, both null-lines from P meet S on one of the arcs AB, BC, or CA. In FIGURE 9, with A and B on one branch of S and C on the other, P can lie anywhere in the three unshaded rectangles on the torus; there is a similar figure when A, B, C lie on a single branch.

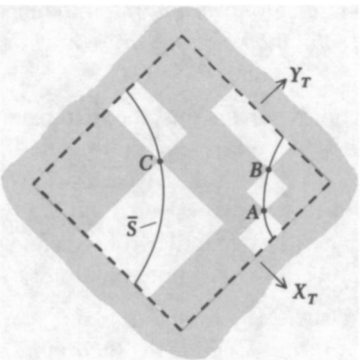


FIGURE 9

We now obtain a simple condition for $A'B'C'$ to be homogeneous. If P is fixed and Q moves along the arc AB, the (x-like or y-like) type of PQ changes when Q crosses a null-line from P or when Q passes through an infinite point of $\bar{S}$. The null lines from P therefore meet the arc AB in 0 or 2 points if, and only if, (i) when A and B lie on the same branch of S, PA and PB have the same type; (ii) when A and B lie on different branches of S, PA and PB have different types. Hence,

(a) If A, B, C lie on the same branch of S, triangle $A'B'C'$ is homogeneous if, and only if, PA, PB, PC have the same type. In this case BC, CA, AB have the same type.

(b) Suppose (for example) that A and B lie on one branch of S and C lies on the other, as in FIGURE 9. Then $A'B'C'$ is homogeneous if, and only if, PA and PB have opposite type to PC. In this case BC and CA have opposite type to AB.

Thus for triangle $A'B'C'$ to be homogeneous, the six edges of the quadrangle $ABCP$ are non-null, and either (i) *any two opposite edges have the same type or* (ii) *any two opposite edges have different types.*

We define a *homogeneous quadrangle* as one with this property. If P lies on S, the h-circle is inverted into a straight line. With P renamed as D, the h-triangle inequality gives the Minkowskian plane Ptolemy theorem:

$ABCD$ is a homogeneous quadrangle in the Minkowskian plane and S is the h-circumcircle of triangle ABC. Define

$$\Pi_h = |BC|_h \cdot |DA|_h - |CA|_h \cdot |DB|_h + |AB|_h \cdot |DC|_h.$$

Then

(i) $\Pi_h \leq 0$ *if the null-lines from D intersect the arc CA of S, with equality if, and only if, D lies on this arc;*

(ii) $\Pi_h > 0$ *otherwise.*

The homogeneous condition has an interesting algebraic interpretation. Because $(AB)_h^2$ may be negative, we define Minkowskian length as $|(AB)_h^2|^{1/2}$; with the definition $[(AB)_h^2]^{1/2}$, however, the homogeneous condition makes the three terms of Π_h either all real or all pure imaginary.

7. Further Results and Connections

(a) Ptolemy's theorem in three-dimensional Minkowskian space. *ABCD is a homogeneous skew quadrangle and 'm' denotes the Minkowskian metric; define*

$$\Pi_m = |BC|_m \cdot |DA|_m - |CA|_m \cdot |DB|_m| + |AB|_m \cdot |DC|_m.$$

Then

(i) $\Pi_m \leq 0$ *if the null-cone at D meets the m-circumcircle of ABC on the arc CA, with equality if, and only if, the intersections are coincident;*

(ii) $\Pi_m > 0$ *otherwise.*

(b) Instead of using complex inversion, the extension of the plane Euclidean theorem in Section 1 can be achieved by embedding A, B, C, and Q in Minkowskian space with metric form $x^2 + y^2 - z^2$. Take $A(\cos\alpha, \sin\alpha, 0)$, etc., and $Q(\cos\rho, 0, 0)$; the extension is equivalent to the Minkowskian Ptolemy theorem (above) for A, B, C, and $D(\cos\rho, 0, \sin\rho)$. The z-axis has a similar role to the imaginary x-axis in complex inversion (though the two procedures are not completely identical).

(c) Ptolemy's theorem in the Minkowskian plane can be reformulated and extended in a similar way to the Euclidean theorem, and there are interesting correspondences between the two:

(i) In the Minkowskian plane, if $(OP_h)_h^2 > 0$ define Q_h on the ray OP_h by $|OQ_h|_h = \frac{1}{2}(p_h + p_h^{-1})$, where $p_h = |OP_h|_h$. Clearly $|OQ_h|_h \geq 1$; the extension to points in $|OQ_h|_h < 1$ is achieved by h-inverting the hyperbola to a *circle*. An 'intertwining' between the extended Minkowskian and Euclidean theorems is suggested by the fact that $x^2 + y^2 = 1$ implies $(1/x)^2 - (y/x)^2 = 1$. The involution $(x, y) \to (1/x, y/x)$ maps the unit circle to the unit h-circle $x^2 - y^2 = 1$ and interchanges 'interiors' and 'exteriors'. Now take A_h, B_h, C_h on the h-circle and h-rotate so that $Q_h(q_h, 0)$ lies on the x-axis; A, B, C and Q are the corresponding points in the Euclidean plane. The extended Minkowskian theorem for A_h, B_h, C_h and Q_h is, when stated algebraically, equivalent to the extended Euclidean theorem for A, B, C and Q.

(ii) If $(OP_h)_h^2 < 0$, define Q_h by $|OQ_h|_h = \frac{1}{2}(p_h - p_h^{-1})$ and h-rotate so that $Q_h(0, q_h)$ lies on the y-axis. Under the same correspondence as above, the Minkowskian theorem for A_h, B_h, C_h and Q_h is equivalent to the (reformulated) Euclidean theorem for A, B, C and the origin O, with projections onto the diameter $x + q_h y = 0$.

8. Conclusion

It is strange indeed that the solution of a Euclidean problem has led us to Minkowskian geometry. The complex method in Section 4 'works' because the unit circle is transformed to a curve where the real part of the x-coordinate is constant and the y-coordinate is real. From general properties of inversion the unit circle becomes a complex circle, and since x_1 is pure imaginary and y_1 is real it is in fact a rectangular hyperbola; the proof is perhaps not so miraculous after all!

The intertwining between the Euclidean and Minkowskian theorems (part (c) of the previous section) gives one more way to deduce the extensions and demonstrates a direct connection between the theorems, as opposed to an indirect one via the triangle inequalities. There is, for example, a relationship between the three unshaded regions inside the circumcircle in FIGURE 3(a) and the three unshaded rectangles on the torus in FIGURE 9. In FIGURE 3(b) there are four unshaded regions

inside the circumcircle (two are hardly visible); the two which are connected at the origin correspond to a single rectangle on the torus.

Acknowledgements. I am very grateful to the referee who made many helpful suggestions and encouraged me to develop the ideas further. Iain Gordon-Ingram kindly drew the figures.

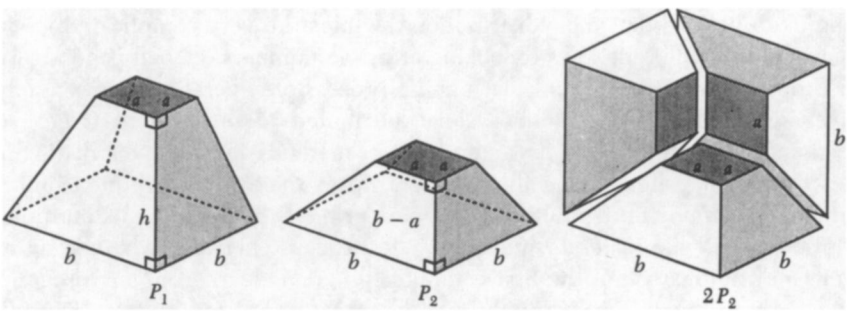
REFERENCES

1. Carl B. Boyer (revised by Uta C. Merzbach), *A History of Mathematics*, 2nd edition, John Wiley & Sons, Inc., New York, 1989, pp. 185–191.
2. Ronald W. Clark, *Einstein: the Life and Times*, Hodder and Stoughton, London, 1973, pp. 127–130.
3. Howard Eves, *Great Moments in Mathematics (Before 1650)*, MAA, Washington, DC, 1983, pp. 96–109.
4. Marvin J. Greenberg, *Euclidean and Non-Euclidean Geometries*, 2nd edition, W. H. Freeman, San Francisco, 1980, pp. 253–257.
5. Abraham Pais, *Subtle is the Lord...*, Oxford University Press, New York, NY, 1982, pp. 151–152.
6. Daniel Pedoe, *A Course of Geometry*, Cambridge University Press, Cambridge, 1970, and Dover Publications, Mineola, NY, 1988, pp. 75–91.
7. J. D. Smith, Generalization of the triangle and Ptolemy inequalities, *Geometriae Dedicata* (in press).
8. I. M. Yaglom, *A Simple Non-Euclidean Geometry and its Physical Basis*, Springer-Verlag New York, 1979, pp. 174–201.

Proof without Words: Volume of a Frustum of a Square Pyramid

[Problem 14, *The Moscow Papyrus*, circa 1850 B.C.]

$$V = \frac{h}{3}(a^2 + ab + b^2)$$



$$V(P_1) = \frac{h}{b-a}V(P_2) = \frac{h}{b-a} \cdot \frac{1}{3}(b^3 - a^3) = \frac{h}{3}(a^2 + ab + b^2)$$

REFERENCES

1. C. B. Boyer, *A History of Mathematics*, John Wiley & Sons, Inc., New York, 1968, pp. 20–22.
2. R. J. Gillings, *Mathematics in the Time of the Pharaohs*, The MIT Press, Cambridge, 1972, pp. 187–193.

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

inside the circumcircle (two are hardly visible); the two which are connected at the origin correspond to a single rectangle on the torus.

Acknowledgements. I am very grateful to the referee who made many helpful suggestions and encouraged me to develop the ideas further. Iain Gordon-Ingram kindly drew the figures.

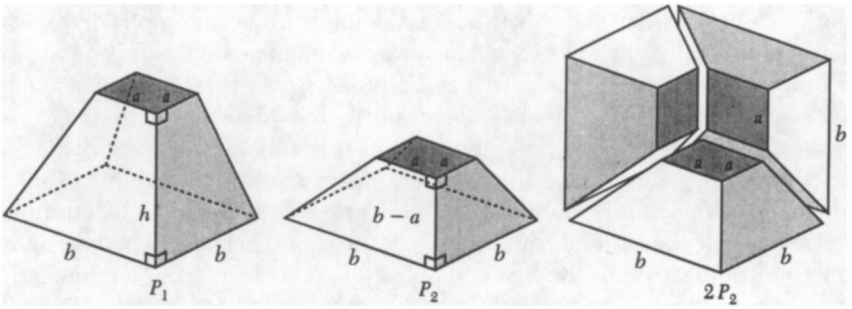
REFERENCES

1. Carl B. Boyer (revised by Uta C. Merzbach), *A History of Mathematics*, 2nd edition, John Wiley & Sons, Inc., New York, 1989, pp. 185–191.
2. Ronald W. Clark, *Einstein: the Life and Times*, Hodder and Stoughton, London, 1973, pp. 127–130.
3. Howard Eves, *Great Moments in Mathematics (Before 1650)*, MAA, Washington, DC, 1983, pp. 96–109.
4. Marvin J. Greenberg, *Euclidean and Non-Euclidean Geometries*, 2nd edition, W. H. Freeman, San Francisco, 1980, pp. 253–257.
5. Abraham Pais, *Subtle is the Lord...*, Oxford University Press, New York, NY, 1982, pp. 151–152.
6. Daniel Pedoe, *A Course of Geometry*, Cambridge University Press, Cambridge, 1970, and Dover Publications, Mineola, NY, 1988, pp. 75–91.
7. J. D. Smith, Generalization of the triangle and Ptolemy inequalities, *Geometriae Dedicata* (in press).
8. I. M. Yaglom, *A Simple Non-Euclidean Geometry and its Physical Basis*, Springer-Verlag New York, 1979, pp. 174–201.

Proof without Words: Volume of a Frustum of a Square Pyramid

[Problem 14, *The Moscow Papyrus*, circa 1850 B.C.]

$$V = \frac{h}{3}(a^2 + ab + b^2)$$



$$V(P_1) = \frac{h}{b-a}V(P_2) = \frac{h}{b-a} \cdot \frac{1}{3}(b^3 - a^3) = \frac{h}{3}(a^2 + ab + b^2)$$

REFERENCES

1. C. B. Boyer, *A History of Mathematics*, John Wiley & Sons, Inc., New York, 1968, pp. 20–22.
2. R. J. Gillings, *Mathematics in the Time of the Pharaohs*, The MIT Press, Cambridge, 1972, pp. 187–193.

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

Is There Any Regularity in the Distribution of Prime Numbers at the Beginning of the Sequence of Positive Integers?

SILVIU GUIASU
York University
North York, Ontario, Canada M3J 1P3

Prime numbers have always fascinated mathematicians. They appear among the integers seemingly at random, and yet not quite: There seems to be some order or pattern, just a little below the surface, just a little out of reach.—Underwood Dudley [3]

Prime numbers are the multiplicative building bricks of the number system. According to the fundamental theorem of arithmetic, every integer number larger than 1 is either a prime or the product of a unique set of primes. In what follows, by an integer we will understand a positive integer. In multiplicative number theory each integer is a word, more exactly a commutative juxtaposition of primes. In this coding process each prime is employed according to a rigid rule (the gap between the consecutive multiples of a prime p is just p) and the set of prime numbers is like an alphabet that is self-generating in order to make the resulting code nondegenerate. But how are the prime numbers themselves generated? Contemplating successive gaps between consecutive primes or the number of prime factors of consecutive integers, we can only notice an apparently chaotic behavior of the prime numbers leading us to believe that their distribution law must be very complicated.

There are two different ways of looking at prime numbers: globally and algorithmically. From an algorithmic point of view the process of generating prime numbers is relatively clear. The prime-number sieve, attributed to the ancient Greek scholar Eratosthenes, was one of the first step-by-step methods invented for distinguishing primes from composites among the numbers up to some predetermined limit: Take the number 2, eliminate its multiples; the next prime is 3, eliminate its multiples; the next prime is 5, eliminate its multiples, etc. Today, checking whether or not an integer is a prime is one of the first computer programs learned in any programming language. Eratosthenes' sieve simply tells us what to do, step-by-step, for selecting the primes in a given set of consecutive integers without revealing any regularity in the distribution of primes.

Those unhappy with an algorithmic approach have tried several ways to approach a global understanding of the behavior of primes. Many papers have dealt with the asymptotic behavior of different functions depending on primes. There is a rich literature on the subject (see for instance [15], [17], [3], [2]) using very subtle mathematical techniques. To give only one example, let $\pi(x)$ denote the number of primes not exceeding the positive real number x. According to the celebrated *prime number theorem* (PNT), we have $\pi(x) \simeq x/\ln x$, $(x \to \infty)$, which means that the ratio of the two functions, namely $\pi(x)/(x/\ln x)$, converges to 1 as x grows without bound, proved independently by J. Hadamard [9] and C. J. de La Vallée Poussin [14] using tools involving functions of complex variables. PNT is a superb example of extracting asymptotic order from chaos.

The Global Approach: probabilistic asymptotic regularity The apparently random occurrence of prime numbers among all integer numbers blocked all attempts at obtaining simple precise formulas for counting primes. Toward the end of the 18th century, A. M. Legendre and C. F. Gauss recast the prime distribution question in a statistical form: About how many of the first n positive integers are primes? A combination of heuristic and theoretical considerations led them to conjecture that the answer was, in some sense, $n/\ln n$. We will say more about their conjectures in the last sections of the paper. That was the beginning of a statistical way of looking at the distribution of primes.

Let $\omega(n)$ be the number of distinct prime factors of the integer n. Thus, $\omega(252) = \omega(2^2 \cdot 3^2 \cdot 7) = 3$. In 1917, Hardy and Ramanujan [**10**] proved the following result: If Φ is a function of x that tends steadily to infinity with x, then $\omega(n)$ satisfies the inequalities

$$\ln\ln x - \Phi(x)(\ln\ln x)^{1/2} < \omega(n) < \ln\ln x + \Phi(x)(\ln\ln x)^{1/2}$$

for almost all numbers n less than x. "Almost all" here is understood in the following way: The frequency of those integers n in the interval $1 \le n \le x$ for which

$$|\omega(n) - \ln\ln x| > \Phi(x)(\ln\ln x)^{1/2}$$

approaches zero as $x \to \infty$. Using only words, the above result was formulated by Hardy and Ramanujan as saying that the normal order of the number of different prime factors of a number n is $\ln\ln n$. In 1939, a fortunate meeting between the probabilist Mark Kac and the number theorist Paul Erdös led to the following deeper result (see [**11**]), known as the Erdös-Kac Theorem (EKT): Asymptotically, when $N \to \infty$, the proportion of integers n from the interval $1 \le n \le N$ for which $\omega(n)$ satisfies the inequalities

$$\ln\ln N - a(\ln\ln N)^{1/2} < \omega(n) < \ln\ln N + b(\ln\ln N)^{1/2}$$

is accurately approximated by the area under the bell-shaped probability density curve

$$(2\pi)^{-1/2} e^{-x^2/2}$$

of the standard normal distribution $N(0, 1)$ between $x = a$ and $x = b$.

The paper containing the complete proof of EKT was published in 1940. In his autobiography [**12**], Mark Kac writes about it: "I wish I could report that when it appeared, our paper catapulted us to heights of fame. It didn't. In fact, it was hardly noticed. Perhaps the war had something to do with it. Perhaps it was because the paper was not very well written—my fault since I did the writing. Most likely it was because the subject was a bit esoteric at the time. It did not take off until the fifties." Other results and bibliographic references about the probabilistic asymptotic regularity in the distribution of primes may be found in Elliott [**4**]. According to him: "The probabilistic theory of numbers may be viewed as a twentieth-century sport."

But what happens at the beginning of the sequence of positive integers? EKT shows that the normal distribution $N(0, 1)$ (with mean 0 and variance 1) is the probabilistic law obeyed by the proportion of those integers n in the interval $1 \le n \le N$ for which

$$|\omega(n) - \ln\ln N| \le (\ln\ln N)^{1/2}$$

but only when $N \to \infty$. It is an asymptotic result and we can jokingly remember what the economist John Maynard Keynes ([13], p. 88) said about probabilistic asymptotic results: "...Now 'in the long run' this is probably true...But this long run is a misleading guide to current affairs. In the long run we are all dead." Therefore, is there any regularity in the distribution of primes in the interval $[1, N]$ when N is not necessarily a large number but something like 100, 1000, 100000, or 10^{16}? Unfortunately, PNT and EKT do not give accurate results when they are applied to the set of primes from a short interval $[1, N]$. Thus, the simple function $n/\ln n$, already mentioned in the formulation of PNT as giving the asymptotic trend followed by the number of primes, is not a very good approximation of $\pi(n)$ for the values of n we are normally dealing with in our computations. Also, the normal probability distribution $N(\ln\ln N, \ln\ln N)$, with the mean $\ln\ln N$ and variance $\ln\ln N$, predicted by EKT for large values of N, cannot be considered a very good approximation of the empirical probability distribution of $\{\omega(n);\ 1 \le n \le N\}$, or even of its trend, for small values of N. As the asymptotic results (PNT and EKT, for instance) fail to describe accurately what happens with the prime numbers at the beginning of the sequence of integers, there is no other solution left but to listen to what John von Neumann once said: "Truth is much too complicated to allow anything but approximations." Therefore, let us try to build up probabilistic models that give the trend followed by the gap between consecutive primes and the behavior of $\pi(n)$ at the beginning of the sequence of positive integers.

Probabilistic models based on entropy maximization More than a century ago, Ludwig Boltzmann [1] introduced a famous H-function, called *statistical entropy*, for measuring the degree of molecular disorder in the evolution of a gas. In 1948, in a celebrated paper [20], Claude Shannon showed that Boltzmann's H-function has the basic properties we can expect from a general measure of the amount of uncertainty contained by a probability distribution and proved the uniqueness of its analytical expression.

Let $\mathbf{q} = (q_0, q_1, \ldots, q_n, \ldots)$ be a discrete probability distribution, i.e. a sequence of nonnegative numbers whose sum is equal to 1. The amount of uncertainty contained by $\mathbf{q}$ is measured by the *entropy*

$$H(\mathbf{q}) = -\sum_{k=0}^{\infty} q_k \ln q_k, \tag{1}$$

where $0 \ln 0 = 0$. Let $f(x)$ be a probability density on the interval $[0, \infty)$. The corresponding entropy is

$$H(f) = -\int_0^{\infty} f(x) \ln f(x)\, dx. \tag{2}$$

Let X be a random variable whose possible values are $\{k;\ k = 0, 1, \ldots, n, \ldots\}$ with the corresponding probabilities $\mathbf{q} = (q_0, q_1, \ldots, q_n, \ldots)$. Then, its mean value, namely

$$E(X) = \sum_{k=0}^{\infty} k q_k, \tag{3}$$

is a number, uniquely determined, showing the main tendency, provided that the series is convergent. Similarly, if Y is a random variable whose possible values are the elements of the set $\{x;\ 0 \le x < \infty\}$ with the probability density $f(x)$, $(0 \le x < \infty)$, then its mean value is

$$E(Y) = \int_0^\infty xf(x)\,dx, \tag{4}$$

provided that the integral does exist. The converse problem is much more difficult. If the mean value $E(X)$ of the random variable X with the possible values (range) $\{k;\ k = 0, 1, \ldots, n, \ldots\}$ is given, then there are infinitely many probability distributions $\mathbf{q} = (q_0, q_1, \ldots, q_n, \ldots)$ compatible with (3). Similarly, if the mean value $E(Y)$ of the random variable Y with the possible values (range) $\{x;\ 0 \le x < \infty\}$ is given, then there are infinitely many probability densities $f(x)$, $(0 \le x < \infty)$ compatible with (4). How do we select the best one and what does "the best" mean? A relatively recent but widely used point of view is based on *entropy maximization*: From the set of all probability distributions compatible with a mean value choose the one that maximizes entropy. Such a probability distribution is the most unbiased one; it will ignore no possibility, being the most uncertain one, subject to the given mean value taken as a constraint of this optimization problem. Theoretically, in order to get the true probability distribution of the random variable X we have to know the sequence of all its moments $\{E(X^k),\ k = 1, 2, \ldots\}$. The probability distribution of maximum entropy compatible with only the first moment $E(X)$ is only a first approximation of the true, but unknown, probability distribution of X. The approximation is accurate to the extent to which the true probability distribution of X gives indeed the mean value $E(X)$, assumed to be known, and does maximize the amount of uncertainty on the possible values of X as measured by entropy. Details about the principle of maximum entropy can be found in [8]. In our context, a straightforward application of the standard Lagrange multipliers method from variational calculus gives the following two results:

PROPOSITION 1. *The probability distribution* $\mathbf{q} = (q_0, q_1, \ldots, q_n, \ldots)$ *that maximizes the discrete entropy* (1) *subject to the mean value* (3) *is*

$$q_k = [E(X)]^k / [1 + E(X)]^{k+1}, \quad (k = 0, 1, 2, \ldots). \tag{5}$$

PROPOSITION 2. *The probability density* $f(x)$, $(0 \le x < \infty)$, *that maximizes the continuous entropy* (2) *subject to the mean value* (4) *is*

$$f(x) = [1/E(Y)]\, e^{-[1/E(Y)]x}, \quad (x \ge 0). \tag{6}$$

Let us mention that (5) is the geometric distribution and (6) its continuous analogue, the exponential distribution. We are going to use (5) in the next section.

The gap between consecutive primes Let us denote by $p_2 = 3 < p_3 < \cdots < p_{\pi(n)}$ the consecutive primes strictly larger than $p_1 = 2$. Let $D(n)$ be the gap between the consecutive elements of this ordered finite set. For each integer n, $D(n)$ is a discrete random variable. Thus, for $n = 100$, we have $\pi(100) = 25$, the above set of primes is $3 < 5 < 7 < 11 < 13 < 17 < 19 < 23 < 29 < 31 < 37 < 41 < 43 < 47 < 53 < 59 < 61 < 67 < 71 < 73 < 79 < 83 < 89 < 97$, the corresponding sequence of gaps between consecutive primes is 2, 2, 4, 2, 4, 2, 4, 6, 2, 6, 4, 2, 4, 6, 6, 2, 6, 4, 2, 6, 4, 6, 8, and $D(100)$ is the random variable with the set of possible values $\{2, 4, 6, 8, 10, 12, \ldots\}$ with the corresponding probabilities $q = (8/23, 7/23, 7/23, 1/23, 0, 0, \ldots)$.

Interpreting the consecutive integers as being instants of time, let us look at the primes as being customers randomly arriving in a shop. Only one customer is allowed to enter the shop at a time. Then, $D(n)$ is the random interarrival time, i.e. the time elapsed between two consecutive arrivals during the time interval $[3, n]$. The gaps between consecutive primes do not follow a regular pattern, as may be seen from the

above example. As the primes arrive randomly, for any n, there is no obvious regularity in the way the possible values of the gap $D(n)$ are distributed. Nevertheless, is there a hidden regular trend followed by the frequencies of the possible gaps between consecutive primes? In order to get this hidden regularity, it is preferable to change the scale of possible values of $D(n)$, by taking $X(n) = -1 + D(n)/2$, in which case $X(n)$ has the same probability distribution as $D(n)$ but its range is $\{k;\ k = 0, 1, 2, \ldots\}$ instead of $\{2k;\ k = 1, 2, \ldots\}$. As there are $\pi(n)$ primes not exceeding the integer n, the mean gap between consecutive primes is $n/\pi(n)$. Thus, without knowing the probability distribution of $D(n)$, we have an estimate for its mean value, namely $E[D(n)] = n/\pi(n)$. This is equivalent to $E[X(n)] = -1 + n/[2\pi(n)]$. According to Proposition 1, the most unbiased probability distribution subject to only this constraint is (5), i.e.

$$q_k = \left(\frac{n}{2\pi(n)} - 1\right)^k \left(\frac{n}{2\pi(n)}\right)^{-(k+1)}, \quad (k = 0, 1, 2, \ldots). \tag{7}$$

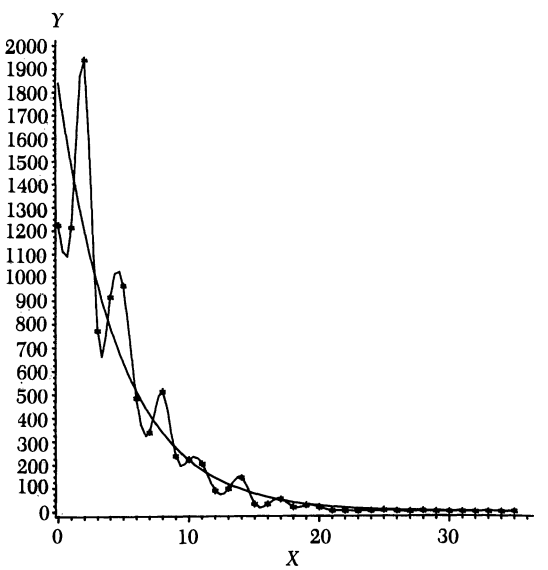


FIGURE 1

TABLE I

k	O_k	E_k
0	1224	1839.58
1	1215	1486.71
2	1940	1201.52
3	773	971.04
4	916	784.77
5	964	634.24
6	484	512.57
7	339	414.25
8	514	334.79
9	238	270.57
10	223	218.67
11	206	176.72
12	88	142.82
13	98	115.43
14	146	93.28
15	32	75.39
16	33	60.93
17	54	49.24
18	19	39.80
19	28	32.16
20	19	25.99
21	5	21.01
22	4	16.98
23	3	13.72
24	5	11.09
25	7	8.96
26	4	7.24
27	1	5.85
28	4	4.73
29	1	3.82
30	1	3.09
31	1	2.50
32	0	2.02
33	0	1.63
34	0	1.32
35	1	1.07

Let O_k be the *observed* absolute frequency of the value k of the random variable $X(n)$ and

$$E_k = q_k\Big(\sum_{k\geq 0} O_k\Big), \quad (k = 0, 1, 2, \ldots),$$

the *expected* absolute frequencies induced by the geometric probability distribution (7). The values of O_k and E_k for $n = 100{,}000$ are given in the second and third columns of Table I, respectively. The moving average smoothing technique and the goodness-of-fit χ^2-test allow us to conclude that the average trend followed by the absolute frequency of the gaps between the consecutive primes smaller than 100,000 is very well approximated by the geometric distribution (7). FIGURE 1 contains the SAS computer graph of the true absolute frequencies $\{O_k;\ k = 0, 1, 2, \ldots, 35\}$ and (the median curve) of the geometric absolute frequencies $\{E_k;\ k = 0, 1, 2, \ldots, 35\}$ for $n = 100{,}000$. The same type of analysis, performed for $n = 200{,}000$, $n = 300{,}000$, and $n = 400{,}000$, gave the same conclusion and similar graphs.

Approximating $\pi(x)$ by special functions Finding a formula for $\pi(x)$ has been a main objective in number theory. Let us look again at primes as being customers arriving in a queueing system while the positive real axis is the time scale. We are interested in the arrival process and at most one arrival is allowed at a time instant. Let $p_1, \ldots, p_{\pi(x)}$ be the consecutive primes not exceeding the real number x; they are the customers arriving during the time interval $[1, x]$. Let λ_x be the mean arrival rate, i.e. the expected number of customers arriving per unit of time, in a neighborhood of the number x. It gives the density of primes in the neighborhood of x. If λ_x is integrable as a function of x, then a continuous approximation of the number $\pi(x)$ of customers arriving during the time interval $[2, x]$ is

$$\pi(x) \cong \int_2^x \lambda_t\, dt.$$

In 1849, in a personal letter to the astronomer Encke [5] (see the English translation of the letter in [6]), after heuristically examining the distribution of the primes between 2,000,000 and 3,000,000, Carl Friedrich Gauss conjectured that $\lambda_x = 1/\ln x$, in which case $\pi(x)$ might be approximated by the logarithmic integral

$$\mathrm{Li}(x) = \int_2^x \frac{dt}{\ln t}.$$

In 1859, Bernhard Riemann [19] assumed that Gauss' density $1/\ln x$ rather describes the mean arrival rate in the neighborhood of x if the customers are thought to be not only the primes but also a half of the prime squares, a third of the prime cubes, a quarter of the prime fourth powers, etc. Since the number of prime squares less than x is obviously equal to the number of primes less than $x^{1/2}$, that is, equal to $\pi(x^{1/2})$, and since in the same way the number of prime nth powers p^n less than x is $\pi(x^{1/n})$, it follows that, according to Riemann's assumption, we have

$$\pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots + \frac{1}{n}\pi(x^{1/n}) + \cdots \cong \mathrm{Li}(x). \tag{8}$$

The series in this formula is finite for any given x because $x^{1/n} < 2$ for n sufficiently large, which implies $\pi(x^{1/n}) = 0$. If we successively replace x in (8) by $x^{1/2}, x^{1/3}, \ldots, x^{1/n}, \ldots$ and multiply the obtained formulas by $-1/2$, $-1/3$, $-1/5$, $+1/6$, $-1/7, \ldots, \mu(n)/n, \ldots$, respectively, where Möbius' function $\mu(n)$ is 0 if n is

divisible by a prime square, 1 if n is a product of an even number of distinct primes, and -1 if n is a product of an odd number of distinct primes, then, summing up (8) and all the formulas obtained, we get the inverse formula $\pi(x) \cong \mathrm{R}(x)$, where $\mathrm{R}(x)$ is Riemann's function

$$\mathrm{R}(x) = \mathrm{Li}(x) + \sum_{n=2}^{\infty} \frac{\mu(n)}{n} \mathrm{Li}(x^{1/n}). \tag{9}$$

At the beginning of the sequence of positive integers, namely up to 4×10^{16}, $\mathrm{Li}(x)$ is a very good approximation and $\mathrm{R}(x)$ an excellent approximation of $\pi(x)$, as can be seen in the comprehensive paper [21].

Approximating $\pi(x)$ by elementary functions Neither Gauss' function $\mathrm{Li}(x)$ nor Riemann's function $\mathrm{R}(x)$ are elementary functions. They are special functions and we need tables or computer subroutines in order to approximate their values. To the extent to which we are interested in discovering some kind of regularity in the apparently chaotic behavior of $\pi(x)$, we should like to have an elementary function, or a finite combination of elementary functions that could well approximate the behavior of $\pi(x)$ at the beginning of the sequence of positive integers. The elementary function $\mathrm{G}(x) = x/\ln x$ suggested by the prime number theorem is not a very good approximation of $\pi(x)$ in the interval $[0, 4 \times 10^{16}]$. Adrien-Marie Legendre [16], on the other hand, conjectured that $\pi(x)$ may be well approximated by the elementary function

$$\mathrm{Le}(x) = \frac{x}{\ln x - 1.08366},$$

which indeed is much better than $\mathrm{G}(x)$. Is it possible to find an even better approximation of $\pi(x)$ in $[0, 4 \times 10^{16}]$ using elementary functions? Regression analysis deals with the problem of finding a linear combination of chosen elementary functions (called predictors) that is the best approximation of the trend followed by a given data set. Here 'the best' refers to an optimization technique proposed independently by the same Legendre and Gauss, namely the least squares method. According to this method, the coefficients of the linear combination of chosen elementary predictors are determined by minimizing the sum of square deviations from the given data set. In our context, as the elementary function $\mathrm{G}(x)$ mentioned in PNT is not a very good approximation of $\pi(x)$ for $x \leq 4 \times 10^{16}$, we would like to compare its values with its deviations from the values of $\pi(x)$. The values of $\mathrm{H}(x) = \{\mathrm{G}(x)/[\pi(x) - \mathrm{G}(x)]\}^4$ for $x = 10^k$, $(k = 1, 2, \ldots, 16)$, are taken as our given data set and we choose the first 12 integer powers of the elementary function, $\ln\ln x$, as predictors. The least squares method implemented by the procedure STEPWISE of the well-known computer statistical package SAS allows us to estimate $\mathrm{H}(x)$ as a polynomial of degree 12 in $\ln\ln x$, and taking into account the definition of $\mathrm{H}(x)$ given above we obtain the approximation of $\pi(x)$ by the elementary function

$$\mathrm{S}(x) = \frac{x}{\ln x}\left[1 + \left\{\sum_{k=0}^{12} a_k (\ln\ln x)^k\right\}^{-1/4}\right] \tag{10}$$

where:

$$a_0 = 229168.50747390; \quad a_1 = -429449.7206839; \quad a_2 = 199330.41355048;$$
$$a_3 = 28226.22049280; \quad a_4 = 0; \quad a_5 = 0; \quad a_6 = -34712.81875914; \quad a_7 = 0;$$
$$a_8 = 33820.10886195; \quad a_9 = -25379.82656589; \quad a_{10} = 8386.14942934;$$
$$a_{11} = -1360.44512548; \quad a_{12} = 89.14545378.$$

As it happens in regression analysis, formula (10) cannot be reliably extrapolated beyond the range used for estimating its coefficients, namely beyond 4×10^{16}; but inside the interval $[0, 4 \times 10^{16}]$ it gives a very good approximation of the values of $\pi(x)$, as can be seen from Table II. Being an elementary function, the values of $S(x)$ may be calculated by using a simple pocket calculator equipped with the keys $+$, $-$, $\times$, $/$, $\sqrt{}$, and ln. Table III contains the differences $\pi(x) - G(x)$, $\pi(x) - Le(x)$, $\pi(x) - Li(x)$, and $\pi(x) - S(x)$ for $x = 11$, 10^2, $k \times 10^3$ $(k = 1, \ldots, 9)$, $k \times 10^4$ $(k = 1, \ldots, 9)$, 10^5, 10^6, 10^7, $k \times 10^8$ $(k = 1, \ldots, 9)$, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 2×10^{16}, 4×10^{16}. The values of the special function $Li(x)$ have been approximated running the IMSL subroutine DLI on an IBM $-$ 4381 computer for $x < 10^9$ and taken from [18] for $x \geq 10^9$. FIGURE 2 contains the SAS graphs of the functions $\pi(x)$, $G(x)$, $Le(x)$, $Li(x)$, and $S(x)$ for $x = 1, 2, \ldots, 1000$.

TABLE II

x	$\pi(x)$	$S(x)$
11	5	5
100	25	25
1000	168	168
2000	303	304
3000	430	431
4000	550	553
5000	669	671
6000	783	787
7000	900	900
8000	1007	1012
9000	1117	1122
10000	1229	1230
20000	2262	2268
30000	3245	3252
40000	4203	4205
50000	5133	5135
60000	6057	6049
70000	6935	6949
80000	7837	7838
90000	8713	8717
100000	9592	9587
1000009	78498	78507
10000000	664579	664596
100000000	5761455	5761426
200000000	11078937	11078977
300000000	16252325	16252266
400000000	21336326	21336122
500000000	26355867	26355484
600000000	31324703	31324617
700000000	36252931	36252745
800000000	41146179	41146303
900000000	46009215	46010033
1000000000	50847534	50847567
10000000000	455052511	455052452
100000000000	4118054813	4118058877
1000000000000	37607912018	37607914982
10000000000000	346065536839	346065467670
100000000000000	3204941750802	3204941555966
1000000000000000	29844570422669	29844571515535
10000000000000000	279238341033925	279238342581326
20000000000000000	547863431950008	547863429443055
40000000000000000	1075292778753150	1075292779267763

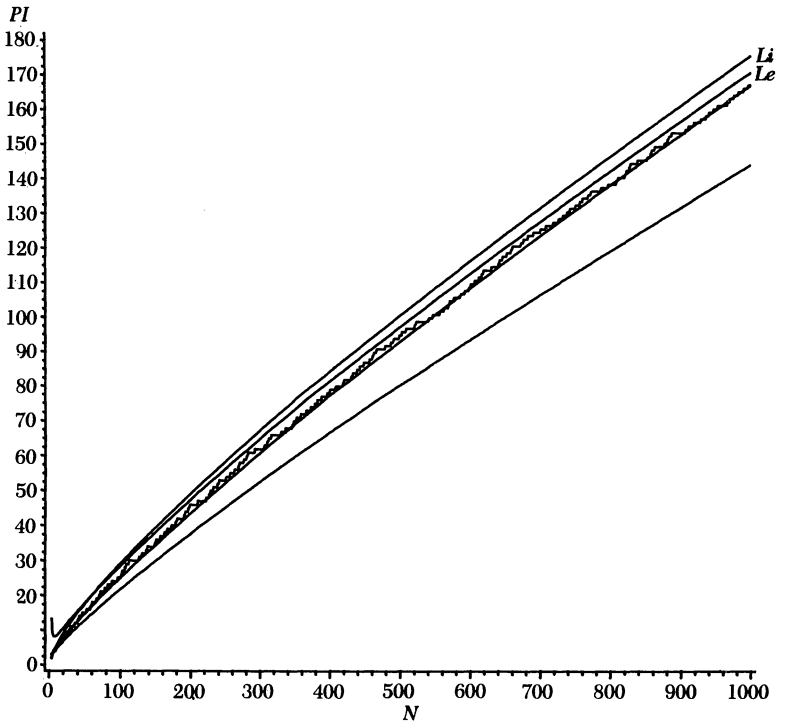


FIGURE 2

Recent results Let $\|a\|$ be the largest integer not exceeding the real number a, μ the Möbius function, and $\Omega(n)$ the total number of prime factors of the integer n (i.e., $\Omega(252) = 5$, for instance). Then, the following exact formula holds ([7]):

$$\pi(x) = -\sum_{k=1}^{\|\log_2 x\|} \mu(k) \sum_{n=2}^{\|x^{1/k}\|} \mu(n)\Omega(n) \left\| \frac{x^{1/k}}{n} \right\|.$$

Also, the following very-easy-to-plot piecewise logarithmic density ([7]):

$$\lambda_u = \begin{cases} \dfrac{1}{\|\log_2 x\|} \dfrac{1}{\ln u}, & \text{if } 2 \le u \le x^{1/\|\log_2 x\|}, \\ \dfrac{1}{k} \dfrac{1}{\ln u}, & \text{if } x^{1/(k+1)} < u \le x^{1/k}, \quad k = 1, \ldots, \|\log_2 x\| - 1, \end{cases}$$

gives an excellent approximation of $\pi(x)$ in the interval $[2, x]$, whose accuracy is comparable to Riemann's by using

$$\pi(x) \cong \mathrm{T}(x) = \int_2^x \lambda_u \, du.$$

Table IV contains the differences $\mathrm{R}(x) - \pi(x)$ and $\mathrm{T}(x) - \pi(x)$ for values of x up to 10^{17}.

TABLE III

x	$\pi(x) - G(x)$	$\pi(x) - Le(x)$	$\pi(x) - Li(x)$	$\pi(x) - S(x)$
11	0	−3	−1	0
100	3	−3	−4	0
1000	23	−4	−9	0
2000	40	−4	−11	−1
3000	55	−3	−12	−1
4000	68	−5	−14	−3
5000	82	−4	−14	−2
6000	93	−5	−16	−4
7000	109	−1	−13	0
8000	117	−5	−18	−5
9000	129	−5	−19	−5
10000	143	−2	−16	−1
20000	243	−6	−26	−6
30000	335	−7	−31	−7
40000	428	−2	−29	−2
50000	512	−3	−32	−2
60000	603	8	−25	8
70000	660	−15	−49	−14
80000	751	−1	−38	−1
90000	823	−5	−43	−4
100000	906	4	−37	5
1000000	6116	−45	−128	−9
10000000	44158	−561	−338	−17
100000000	332774	−6549	−753	29
200000000	615308	−13583	−1037	−40
300000000	882916	−20509	−1086	59
400000000	1141420	−27412	−1034	204
500000000	1393459	−34289	−949	383
600000000	1640014	−41646	−1329	86
700000000	1882918	−48604	−1293	186
800000000	2122022	−55971	−1661	−124
900000000	2357836	−63719	−2417	−818
1000000000	2592592	−69985	−1701	−33
10000000000	20758029	−690493	−3104	59
100000000000	169923159	−6545056	−11588	−4064
1000000000000	1416705193	−60615397	−38263	−2964
10000000000000	11992858452	−555560046	−108971	69169
100000000000000	102838308636	−5070271362	−314890	194836
1000000000000000	891604962452	−46223804313	−1052619	−1092866
10000000000000000	7804289844393	−421692578206	−3214632	−1547401
20000000000000000	15020437343198	−820637985048	−3776488	2506953
40000000000000000	28929900579949	−1597283441125	−5538861	−514613

TABLE IV

x	$\pi(x)$	$R(x)-\pi(x)$	$T(x)-\pi(x)$
10^1	4	1	1
10^2	25	1	1
10^3	168	0	1
10^4	1,229	−2	−1
10^5	9,592	−5	−2
10^6	78,498	29	34
10^7	664,579	88	96
10^8	5,761,455	97	111
10^9	50,847,534	−79	−52
10^{10}	455,052,511	−1,828	−1,776
10^{11}	4,118,054,813	−2,318	−2,218
10^{12}	37,607,912,018	−1,476	−1,279
10^{13}	346,065,536,839	−5,773	−5,383
10^{14}	3,204,941,750,802	−19,200	−18,423
10^{15}	29,844,570,422,669	73,218	74,775
10^{16}	279,238,341,033,925	327,052	330,189
10^{17}	2,623,557,157,654,232	−598,254	−591,910

Conclusion There is a contrast between the apparent randomness of the distribution of primes in the sequence of positive integers and the relative asymptotic regularity shown by the prime number theorem and related limiting results. It has been mentioned that "the primes play a game of chance" [11] and there is a vast literature on different applications of probability theory for describing the behavior of the prime numbers. These results, as the Erdös-Kac theorem for instance, are essentially asymptotic. The question is whether there is any regularity in the distribution of primes at the beginning of the segment of positive integers. The geometric distribution showing the trend followed by the gap between consecutive primes and the good approximation of $\pi(x)$ using an elementary, easy-to-calculate function $S(x)$ obtained by applying regression analysis seem to reveal that hidden order or pattern "just a little below" the apparently random distribution of primes, so nicely described in Dudley's sentence cited at the beginning of the paper.

Acknowledgement. The author is grateful to the referees for valuable and constructive remarks and comments.

REFERENCES

1. L. Boltzmann, *Vorlesungen über Gastheorie*, J. A. Barth, Leipzig, 1896.
2. H. G. Diamond, Elementary methods in the study of the distribution of prime numbers, *Bull. Amer. Math. Soc.* 7 (1982), 553–589.
3. U. Dudley, *Elementary Number Theory*, W. H. Freeman, San Francisco, 1978, p. 163.
4. P. D. T. A. Elliott, *Probabilistic Number Theory*, Springer-Verlag New York, (Vol. 1) 1979, (Vol. 2) 1980.
5. C. F. Gauss, Letter to Encke (24 Dec. 1849), Published in C. F. Gauss, *Werke*, Vol. 2, Königliche Gesellschaft der Wissenschaften, Göttingen, 1863, pp. 444–447.
6. L. J. Goldstein, A history of the prime number theorem, *Amer. Math. Monthly* 80 (1973), 599–615.
7. S. Guiasu, An exact formula for the number of primes and an approximation of their density. *AMS Abstracts* 15 (1994), 333.
8. S. Guiasu and A. Shenitzer, The principle of maximum entropy, *Math. Intelligencer* 7 (1985), 42–48.
9. J. Hadamard, Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques, *Bull. Soc. Math. France* 26 (1896), 199–220.

10. G. H. Hardy and S. Ramanujan, The normal number of prime factors of a number n, *Quart. J. Pure Appl. Math.* 48 (1917), 76–92.
11. M. Kac, *Statistical Independence in Probability, Analysis, and Number Theory*, John Wiley & Sons, Inc., Rahway, NJ, 1959.
12. M. Kac, *Enigmas of Chance*, University of California Press, Berkeley, 1987.
13. J. M. Keynes, *Monetary Reform*, Harcourt, Brace and Company; New York, 1924.
14. C. J. de La Vallée Poussin, Recherches analytiques sur la théorie des nombres premiers, *Ann. Soc. Sci. Bruxelles* 20 (1896), 183–256.
15. E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, Vol. 1, Teubner, Leipzig, 1909.
16. A.-M. Legendre, *Essai sur la Théorie des Nombres*, Duprat, Paris, 1st edition 1798, 2nd edition 1808.
17. W. J. LeVeque, *Fundamentals of Number Theory*, Addison-Wesley, Reading, MA, 1977.
18. P. Ribenboim, *The Book of Prime Number Records*, 2nd edition, Springer-Verlag, New York, 1989.
19. B. Riemann, Über die Anzahl der Primzahlen unter einer gegebenen Grösse, *Monatsber. Königl. Preuss. Akad. Wiss. Berlin* 1859, 671–680; also in *Gesammelte Mathematische Werke*, Teubner, Leipzig, 1892, pp. 145–155.
20. C. E. Shannon, A mathematical theory of communication, *Bell System Tech. J.* 27 (1948), 379–423.
21. D. Zagier, *Die ersten 50 Millionen Primzahlen*, Beihefte zu Elements der Mathematik No. 15, Birhäuser-Verlag, Basel, 1977.

Conjectures in Ramanujan's Notebooks

Srinivasa Ramanujan
Seemed to see, with Shiva's eye,
Peculiar gemstones in the realms
Beyond where daylight's limits lie.

Sapphires in sepulchral caverns
Innocent of Apollo's rays,
Through the intervening blackness
Bluely glimmered in his gaze.

In his absence, none can see them,
So they slumber in the dark—
Rare, unchanging, age by age,
Awaiting some spelunker's spark.

—J. D. MEMORY
GENERAL ADMINISTRATION
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27515-2688

Note for line 2: A third eye, located in the middle of the forehead, sometimes appears in representations of the Hindu god Shiva; it can be taken to represent divine insight.

10. G. H. Hardy and S. Ramanujan, The normal number of prime factors of a number n, *Quart. J. Pure Appl. Math.* 48 (1917), 76–92.
11. M. Kac, *Statistical Independence in Probability, Analysis, and Number Theory*, John Wiley & Sons, Inc., Rahway, NJ, 1959.
12. M. Kac, *Enigmas of Chance*, University of California Press, Berkeley, 1987.
13. J. M. Keynes, *Monetary Reform*, Harcourt, Brace and Company; New York, 1924.
14. C. J. de La Vallée Poussin, Recherches analytiques sur la théorie des nombres premiers, *Ann. Soc. Sci. Bruxelles* 20 (1896), 183–256.
15. E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, Vol. 1, Teubner, Leipzig, 1909.
16. A.-M. Legendre, *Essai sur la Théorie des Nombres*, Duprat, Paris, 1st edition 1798, 2nd edition 1808.
17. W. J. LeVeque, *Fundamentals of Number Theory*, Addison-Wesley, Reading, MA, 1977.
18. P. Ribenboim, *The Book of Prime Number Records*, 2nd edition, Springer-Verlag, New York, 1989.
19. B. Riemann, Über die Anzahl der Primzahlen unter einer gegebenen Grösse, *Monatsber. Königl. Preuss. Akad. Wiss. Berlin* 1859, 671–680; also in *Gesammelte Mathematische Werke*, Teubner, Leipzig, 1892, pp. 145–155.
20. C. E. Shannon, A mathematical theory of communication, *Bell System Tech. J.* 27 (1948), 379–423.
21. D. Zagier, *Die ersten 50 Millionen Primzahlen*, Beihefte zu Elements der Mathematik No. 15, Birhäuser-Verlag, Basel, 1977.

Conjectures in Ramanujan's Notebooks

Srinivasa Ramanujan
Seemed to see, with Shiva's eye,
Peculiar gemstones in the realms
Beyond where daylight's limits lie.

Sapphires in sepulchral caverns
Innocent of Apollo's rays,
Through the intervening blackness
Bluely glimmered in his gaze.

In his absence, none can see them,
So they slumber in the dark—
Rare, unchanging, age by age,
Awaiting some spelunker's spark.

—J. D. Memory
General Administration
University of North Carolina
Chapel Hill, NC 27515-2688

Note for line 2: A third eye, located in the middle of the forehead, sometimes appears in representations of the Hindu god Shiva; it can be taken to represent divine insight.

NOTES

Minimum and Characteristic Polynomials of Low-Rank Matrices

WILLIAM P. WARDLAW
U. S. Naval Academy
Annapolis, MD 21402*

Introduction Given an $n \times n$ matrix A, it can be useful to have a monic polynomial $p(x)$ of low degree that annihilates A; that is, for which $p(A) = 0$. For example, higher degree polynomials in A can be evaluated by reducing them modulo $p(x)$. A low degree annihilator of A can be used in algorithms to calculate the matrix e^{At}, as in [1, Thm. 6.10, p. 284]. Moreover, knowledge of such an annihilator can simplify the determination of the minimum polynomial or the null ideal of A.

A simple dimension argument shows that the set $\{A^k: 0 \le k \le n^2\}$ is dependent, and hence that A satisfies a polynomial equation

$$c_0A^0 + c_1A^1 + \cdots + c_{n^2}A^{n^2} = 0$$

of degree $\le n^2$. The Cayley-Hamilton Theorem greatly improves this result by showing that A is annihilated by its characteristic polynomial $f_A(x) = \det(xI - A)$, which is monic of degree n. And still more improvement is available when the rank of A is less than $n - 1$.

In an interesting note [5] that appeared recently in *The American Mathematical Monthly*, J. Segercrantz used a formula for the coefficients of the characteristic polynomial of a linear transformation on a vector space to show that a transformation of rank r is annihilated by a certain polynomial of degree $r + 1$; this polynomial turns out to be the same as what we shall call the *proper polynomial* of the transformation. This note recounts a more elementary derivation of results used earlier by the author for [6], and then discusses a simple method of computing the proper polynomial as well as the characteristic polynomial. The results are proved for matrices over an arbitrary commutative ring with identity, with some refinements for finding the minimum polynomial when the ring is a unique factorization domain.

The proper polynomial of A Throughout this note we let R be a commutative ring with identity and let A, B, C, D be matrices over R satisfying

$$A = BC \quad \text{and} \quad D = CB. \tag{1}$$

(A and D are necessarily square.) Since $A^{k+1} = BC(BC)^k = B(CB)^kC = BD^kC$ for any nonnegative integer k, it follows easily from the distributive law and the fact that matrices commute with scalars, that if $p(x)$ is any polynomial over R, then

$$Ap(A) = Bp(D)C. \tag{2}$$

*Research supported in part by Naval Research Laboratory, Radar Division, Identification Systems Branch.

Hence we have

$$p(D) = 0 \quad \text{implies} \quad Ap(A) = 0 \tag{3}$$

for any polynomial $p(x)$ over R. In particular, if $f_D(x)$ is the characteristic polynomial of D, the Cayley-Hamilton Theorem gives us

$$Af_D(A) = 0. \tag{4}$$

So far, the relationship between A and D has been symmetric. To exploit (4), let us define the *spanning rank*, denoted by $\mathrm{sr}(M)$, of an $m \times n$ matrix M over R: If $M = 0$, then $\mathrm{sr}(M) = 0$; otherwise, $\mathrm{sr}(M)$ is the smallest positive integer r such that there is an $m \times r$ matrix P and an $r \times n$ matrix Q over R satisfying $M = PQ$. Such a factorization is called a *spanning rank factorization* of M. Note that $\mathrm{sr}(M)$ is just the cardinality of a minimal spanning set of the row space, or of the column space, of M. If R is a field, $\mathrm{sr}(M)$ is the ordinary rank of M. However, if R has divisors of zero, $\mathrm{sr}(M)$ may be larger than the rank of M defined by McCoy in [3, p. 159].

Now the first part of the following theorem is clear.

THEOREM. *Let R be a commutative ring with identity and let A be an $n \times n$ matrix over R with spanning rank r. Then A satisfies a polynomial equation*

$$Af_D(A) = 0 \tag{4}$$

of degree $r + 1$, where D is any $r \times r$ matrix over R such that $D = CB$ and $BC = A$. Moreover, the characteristic polynomial of A is

$$f_A(x) = x^{n-r} f_D(x). \tag{5}$$

To prove (5) we use the following proof from [4]: Let

$$E = \begin{bmatrix} xI_n & B \\ C & I_r \end{bmatrix} \quad \text{and} \quad F = \begin{bmatrix} I_n & 0 \\ -C & xI_r \end{bmatrix}.$$

Multiplication gives

$$EF = \begin{bmatrix} xI_n - A & xB \\ 0 & xI_r \end{bmatrix} \quad \text{and} \quad FE = \begin{bmatrix} xI_n & B \\ 0 & xI_r - D \end{bmatrix},$$

so $x^r f_A(x) = \det(EF) = (\det E)(\det F) = (\det F)(\det E) = \det(FE) = x^n f_D(x)$ follows from the multiplicative properties of the determinants of square matrices over a commutative ring.(See [3, (51), p. 157] or [7, Thm. 1, p. 31]. A more complicated proof of (5) is also given in [7, Thm. 6, p. 32].) Now (5) is given by dividing $x^r f_A(x) = x^n f_D(x)$ by x^r.

The *proper polynomial* of an $n \times n$ matrix A over a commutative ring with $\mathrm{sr}(A) = r$ is the polynomial $p_A(x)$ defined by $p_A(x) = f_A(x)$ if $r = n$ and $p_A(x) = x^{r+1-n} f_A(x)$ if $r < n$. Notice that (5) ensures that $p_A(x)$ is actually a polynomial, and that if $r < n$ it can be computed in a well-defined way as $p_A(x) = xf_D(x)$ for *any* $r \times r$ matrix D satisfying (1). The proper polynomial $p_A(x)$ is a monic annihilator of A with degree the minimum of n and $r + 1$.

If R is a field, then it is clear from (3) that the minimum polynomial $m_A(x)$ of A divides $xm_D(x)$, where $m_D(x)$ is the minimum polynomial of D, and both $m_A(x)$ and $xm_D(x)$ divide $xf_D(x)$. If R is any integral domain, the same is true when the minimum polynomials are taken over the quotient field of R, but the coefficients of these polynomials could fail to be elements of R. However, if R is a unique

factorization domain (UFD), then the minimum polynomials are in $R[x]$. (Since the characteristic polynomial $f_A(x) = \det(xI - A)$ is monic in $R[x]$ and the minimum polynomial $m_A(x)$ of A over the quotient field of R is monic and divides $f_A(x)$, it follows from Gauss' Lemma [2, p. 124] that $m_A(x)$ is in $R[x]$.) Thus we have the following.

COROLLARY. *If R is a unique factorization domain and A and D are matrices over R satisfying* (1), *then $m_A(x)$ divides $xm_D(x)$ and $m_D(x)$ divides $f_D(x)$; that is,*

$$m_A(x) \mid xm_D(x) \mid xf_D(x). \tag{6}$$

Computations and examples Ideally, we would like to find the minimum polynomial $m_A(x)$ of A, but $m_A(x)$ may not be defined over certain rings R, and could be difficult to find even when R is a field. But one can always compute the characteristic polynomial $f_A(x)$ of A and divide out x^{n-r-1} to obtain the proper polynomial $p_A(x)$, as suggested in [5].

However, if the spanning rank r is smaller than the size n of A, and we can find an $r \times r$ matrix D satisfying (1), it is usually easier to calculate $f_D(x)$ from D. And if minimum polynomials exist, it is probably easier to look for $m_D(x)$ and use (6) than to look for $m_A(x)$ directly. We outline a procedure for finding D, which is effective for many rings R and is efficient when R is a field.

First, row reduce A to obtain a row-reduced form $A' = EA$, where E is a product of $n \times n$ elementary matrices and A' has r nonzero rows followed by $n - r$ zero rows. (One can write $A = FA'$, where F is the inverse of E.) Second, let C be the $r \times n$ matrix formed from the first r rows of A' and let B be the $n \times r$ matrix formed from the first r columns of F. Then $A = BC$, and $D = CB$ is the desired $r \times r$ matrix.

Usually, B can be easily found without calculating F. If the left $r \times r$ submatrix of C is the identity matrix, B is formed from the first r columns of A. More generally, if any entry in C is the only nonzero entry in its column, then the corresponding column of B is a quotient by this entry of the corresponding column of A. As we will see, ad hoc methods are often easier than the algorithm of the preceding paragraph.

We conclude with some examples.

EXAMPLE 1. Let $R = Z$ be the ring of integers and consider the matrix M from [5]. We find B and C ad hoc by forming C from the first two rows of M (which span the row space of M) and choosing B so that $M = BC$.

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 3 & 2 & 1 & 0 \\ 4 & 4 & 4 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -2 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} = BC,$$

$$D = CB = \begin{bmatrix} -9 & 9 \\ -17 & 21 \end{bmatrix}, \qquad f_D(x) = \begin{vmatrix} x+9 & -9 \\ 17 & x-21 \end{vmatrix} = x^2 - 12x - 36.$$

Thus, $f_M(x) = x^4 - 12x^3 - 36x^2$ by (5) and $Mf_D(M) = M^3 - 12M^2 - 36M = 0$ by (4). Also, $m_D(x) = f_D(x)$ since $m_D(x) | f_D(x)$ but D clearly does not satisfy any polynomial $x - c$ of degree one. Since M is singular, $m_M(x)$ must have x as a factor, so $m_M(x) = xm_D(x) = x^3 - 12x^2 - 36x$ by (6).

EXAMPLE 2. Again, let $R = Z$ be the ring of integers and consider

$$A = \begin{bmatrix} 3 & 2 & 8 & -2 & 7 \\ -1 & -1 & -3 & 2 & 0 \\ -2 & 2 & -2 & 1 & -2 \\ 0 & 4 & 4 & -1 & 2 \\ 3 & -2 & 4 & -3 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 2 & -2 \\ -1 & -1 & 2 \\ -2 & 2 & 1 \\ 0 & 4 & -1 \\ 3 & -2 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 & 0 & 3 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix} = BC.$$

The rows of the 3×5 matrix C are the three nonzero rows of the row-reduced form of A; the 5×3 matrix B is made up of columns 1, 2, and 4 of A corresponding to the columns in which the initial 1's in the rows of C occur. Thus we have

$$D = CB = \begin{bmatrix} 8 & 0 & -9 \\ 0 & -1 & 0 \\ 6 & 0 & -7 \end{bmatrix},$$

$$f_D(x) = \begin{vmatrix} x-8 & 0 & 9 \\ 0 & x+1 & 0 \\ -6 & 0 & x+7 \end{vmatrix} = x^3 - 3x - 2 = (x+1)^2(x-2),$$

and

$$(x+1)(x-2) \mid m_D(x) \mid f_D(x),$$

since every irreducible factor of $f_D(x)$ is a factor of $m_D(x)$. Trial shows that $(D+I)(D-2I) = 0$, so the minimum polynomial of D is $m_D(x) = (x+1)(x-2) = x^2 - x - 2$. A is singular, so $m_A(x)$ must have x as a factor; thus $m_A(x) = xm_D(x) = x^3 - x^2 - 2x$ follows from (6). The characteristic polynomial of A is $f_A(x) = x^2 f_D(x) = x^5 - 3x^3 - 2x^2$, by (5).

When divisors of zero are present, the minimum polynomial may fail to exist, as we see in our third example.

EXAMPLE 3. Let $R = Z_{12}$ be the ring of integers modulo 12, and let

$$A = \begin{bmatrix} 1 & 1 & 5 & 6 \\ 4 & 0 & 0 & 8 \\ 2 & 4 & 2 & 2 \\ 4 & 1 & 2 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 4 & 0 \\ 2 & 4 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 3 & 5 \\ 0 & 1 & 2 & 1 \end{bmatrix} = BC.$$

The 2×4 matrix C is formed from the two nonzero rows of the row reduced matrix obtained from A. Since the left two columns of C form the identity matrix, the 4×2 matrix B is formed from the first two columns of A. Then

$$D = CB = \begin{bmatrix} 3 & 6 \\ 0 & 9 \end{bmatrix}, \qquad f_D(x) = \begin{vmatrix} x-3 & 6 \\ 0 & x+3 \end{vmatrix} = x^2 + 3.$$

We will determine the *null ideal* $N_D = \{p(x) \in R[x]: p(D) = 0\}$ of D. Each element of N_D can be reduced modulo $f_D(x)$ to obtain an element $g(x) = ax + b$ of N_D. Then $g(D) = aD + bI = 0$ implies $3a + b = 6a = 9a + b = 0$, so a must be even and $g(x)$ is $2x + 6$, $4x$, $6x + 6$, $8x$, $10x + 6$, or 0. Since each of these is a multiple of $2x + 6$, we see that the null ideal of D is the ideal $N_D = (x^2 + 3, 2x + 6)$ generated by $x^2 + 3$ and $2x + 6$. Clearly, D has no minimum polynomial, since N_D is not a principal ideal.

Now A is a zero of $xf_D(x) = x^3 + 3x$, so we can reduce any element of the null ideal N_A modulo $x^3 + 3x$ to obtain an element $h(x) = ax^2 + bx + c$ of N_A. The (1, 2),

(2,2), (3,1), (4,2) entries of $h(A) = aA^2 + bA + cI = 0$ give $3a + b = c = 6a + 2b = 9a + b = 0$, so $6a = 2b = c = 0$. Thus $h(x)$ is a multiple of $2x^2 + 6x$ and $N_A = (x^3 + 3x, 2x^2 + 6x) = xN_D$ is the null ideal of A. Of course, $f_A(x) = x^2 f_D(x) = x^4 + 3x^2$, by (5). McCoy discusses the null ideal of a matrix more fully in [3, sec. 38, pp. 162–168]. Our example arose from his on the bottom of page 164.

Our last example shows that the presence of zero divisors does not preclude the existence of a minimum polynomial of A.

EXAMPLE 4. Let $R = Z_6$ be the ring of integers modulo 6 and let

$$A = \begin{bmatrix} 1 & 2 & 4 & 4 & 3 \\ 0 & 5 & 2 & 2 & 3 \\ 3 & 3 & 4 & 5 & 5 \\ 4 & 1 & 4 & 2 & 5 \\ 2 & 2 & 2 & 4 & 4 \end{bmatrix} = BC,$$

where

$$B = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 5 & 4 \\ 3 & 3 & 2 \\ 4 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 2 & 1 \\ 0 & 0 & 2 & 1 & 1 \end{bmatrix}.$$

As usual, C is formed from the three nonzero rows of the row-reduced matrix obtained from A. The first two columns of B are the first two columns of A, and the third column of B is one of the 32 possible quotients of the third column of A by the initial coefficient 2 in the third row of C. The entries in the third column of B are uniquely determined by the fact that the product of B with the fourth and fifth columns of C must give the fourth and fifth columns of A, respectively. Thus

$$D = CB = \begin{bmatrix} 3 & 1 & 3 \\ 4 & 3 & 0 \\ 0 & 3 & 4 \end{bmatrix},$$

and

$$f_D(x) = \begin{vmatrix} x-3 & 5 & 3 \\ 2 & x-3 & 0 \\ 0 & 3 & x-4 \end{vmatrix} = \left[(x-3)^2 + 2\right](x-4)$$

$$= (x^2 - 1)(x-4) = (x-1)(x-5)(x-4).$$

At first glance, $f_D(x)$ seems to have only simple roots, until we see that

$$m(x) = (x-5)(x-4) = x^2 + 3x + 2 = (x-1)(x-2)$$

in $Z_6[x]$. This prompts us to note that $m(D) = (D - I)(D - 2I) = 0$. Since $aD + bI = 0$ implies $a = b = 0$, it follows that the null ideal of D is $N_D = (m(x))$. Since N_D is a principal ideal, it follows that D has a minimum polynomial $m_D(x) = m(x) = x^2 + 3x + 2$. A similar analysis shows that A has null ideal $N_A = (x^3 + 3x^2 + 2x) = xN_D$ and minimum polynomial $m_A(x) = xm_D(x) = x^3 + 3x^2 + 2x$.

Remark. If the spanning rank r is less than the size of the square matrix A, calculation of the characteristic polynomial, the proper polynomial, and the minimum polynomial or null ideal of A is facilitated by row reducing A to obtain a spanning

rank factorization $A = BC$, and then using the $r \times r$ matrix $D = CB$ to determine its characteristic polynomial and its minimum polynomial or null ideal. Moreover, the polynomial identity (2) provides an easy elementary derivation of the lower degree "characteristic" equation (4) suitable for a first course in linear algebra.

Acknowledgement. The author would like to thank the referees for a number of helpful suggestions.

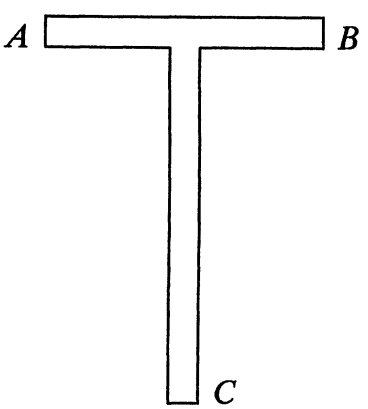
REFERENCES

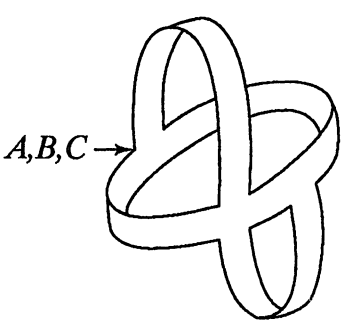
1. C. Cullen, *Linear Algebra & Differential Equations*, Prindle, Weber & Schmidt, Boston, 1979.
2. I. Herstein, *Topics in Algebra*, Xerox College Publ., Lexington, 1964.
3. Neal McCoy, *Rings and Ideals*, Carus Monograph No. 8, MAA, Washington, DC, 1948.
4. Josef Schmid, A remark on characteristic polynomials, *Amer. Math. Monthly*, 77, 9 (November 1970) 998–999.
5. J. Segercrantz, Improving the Cayley-Hamilton equation for low-rank transformations, *Amer. Math. Monthly*, 99, 1 42–44.
6. W. P. Wardlaw, Problem 1179, this MAGAZINE 56, 5 p. 326, and Solution 1179, this MAGAZINE 57, 5 303.
7. W. P. Wardlaw, A transfer device for matrix theorems, this MAGAZINE 59, 1 30–33.

Möbius Shorts

I was once criticized for over indulgence in "indiscriminate reading." In the *Dictionnarie des mathématiques* by A. Bouvier and M. George (Paris, 1979), I found on page 477, an entry for "slip de Möbius," credited to Gourmalin. According to authoritative *Petit Larousse*, the French word "slip" means "shorts." This object is a one-sided surface, which seems not to be well known in American mathematical circles.

To make a paper model, start with a T-shaped piece of paper with a long stem. Bend the top of the T to make a ring (not twisted) and glue A to B. Pass C upward through the ring, turn C down (without twisting), and glue C to the outside of the ring at AB.





Verify that you have produced a one-sided surface. Now guess what results if you cut both the ring and what was originally the stem along their midlines. Then see if you can put the result back together again.

Personal communication, the late Ralph P. Boas, Jr. (1992)

rank factorization $A = BC$, and then using the $r \times r$ matrix $D = CB$ to determine its characteristic polynomial and its minimum polynomial or null ideal. Moreover, the polynomial identity (2) provides an easy elementary derivation of the lower degree "characteristic" equation (4) suitable for a first course in linear algebra.

Acknowledgement. The author would like to thank the referees for a number of helpful suggestions.

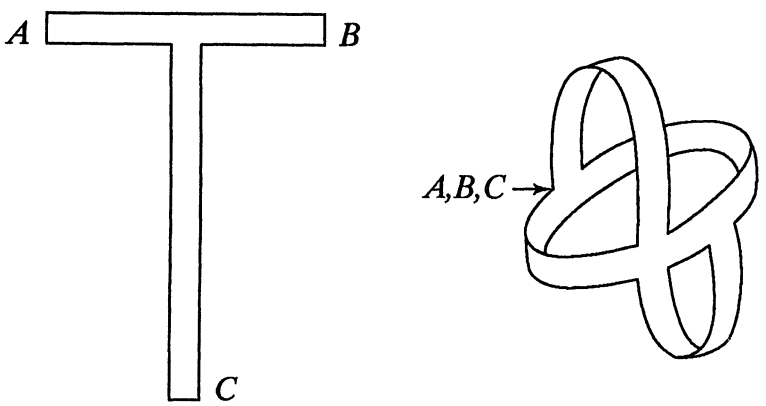
REFERENCES

1. C. Cullen, *Linear Algebra & Differential Equations*, Prindle, Weber & Schmidt, Boston, 1979.
2. I. Herstein, *Topics in Algebra*, Xerox College Publ., Lexington, 1964.
3. Neal McCoy, *Rings and Ideals*, Carus Monograph No. 8, MAA, Washington, DC, 1948.
4. Josef Schmid, A remark on characteristic polynomials, *Amer. Math. Monthly*, 77, 9 (November 1970) 998–999.
5. J. Segercrantz, Improving the Cayley-Hamilton equation for low-rank transformations, *Amer. Math. Monthly*, 99, 1 42–44.
6. W. P. Wardlaw, Problem 1179, this MAGAZINE 56, 5 p. 326, and Solution 1179, this MAGAZINE 57, 5 303.
7. W. P. Wardlaw, A transfer device for matrix theorems, this MAGAZINE 59, 1 30–33.

Möbius Shorts

I was once criticized for over indulgence in "indiscriminate reading." In the *Dictionnarie des mathématiques* by A. Bouvier and M. George (Paris, 1979), I found on page 477, an entry for "slip de Möbius," credited to Gourmalin. According to authoritative *Petit Larousse*, the French word "slip" means "shorts." This object is a one-sided surface, which seems not to be well known in American mathematical circles.

To make a paper model, start with a T-shaped piece of paper with a long stem. Bend the top of the T to make a ring (not twisted) and glue A to B. Pass C upward through the ring, turn C down (without twisting), and glue C to the outside of the ring at AB.



Verify that you have produced a one-sided surface. Now guess what results if you cut both the ring and what was originally the stem along their midlines. Then see if you can put the result back together again.

Personal communication, the late Ralph P. Boas, Jr. (1992)

The Kissing Number of the Square

M. S. KLAMKIN
T. LEWIS
A. LIU
University of Alberta
Edmonton, Canada T6G-2G1

1. Introduction Given a unit square S, what is the maximum number of unit squares that can be arranged so that each touches S, but no two of the squares overlap? Everyone will say 8, and point to the familiar tic-tac-toe board (see FIGURE 1) as the only way to achieve this. It turns out that the result, while true, is not easy to prove. In this paper, we give the history of the problem, and present two elementary solutions. One of them will establish the uniqueness of the optimal configuration.

Two plane figures are said to kiss each other if they do not overlap but their boundaries have non-empty intersection. For a plane figure F, a kissing configuration of order n consists of $n+1$ congruent non-overlapping copies of F such that one copy kisses each of the remaining n. The *kissing number* of F, denoted by $k(F)$, is the largest integer n such that a kissing configuration of order n exists for F. A configuration of order $k(F)$ is said to be optimal. It may or may not be unique.

The kissing number of every plane figure is at least 6. In fact, there is always a kissing configuration in which all 7 copies are translates of one another (see [3]). The circle shows that the universal lower bound of 6 cannot be increased. On the other hand, there is no universal upper bound, as a kissing configuration of order $2n+6$ exists for the 1 by n rectangle, n an arbitrary positive integer.

FIGURE 1 is a kissing configuration of order 8 for the square S. Hence $k(S) \geq 8$. Youngs [5] seems to have been the first to establish that $k(S) = 8$, using multi-variable calculus. However, his proof that FIGURE 1 is optimal does not address the issue of its uniqueness.

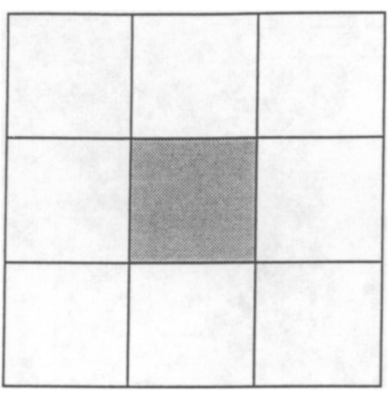


FIGURE 1

The determination of $k(S)$ was later posed as a Putnam competition problem. Youngs' proof was reproduced in the solution book [2]. The book also suggested an alternative argument, and stated that it would establish the uniqueness of the optimal configuration.

Still later, the question appeared in a problem anthology [4], but without reference to [2]. Again, Youngs' proof was given, plus a proof using single-variable calculus.

In Section 2, we give a simple and elementary proof that $k(S) = 8$, along the line of Youngs' argument. In Section 3, we present an alternative proof, simpler than the one suggested in [2]. In Section 4, we show that FIGURE 1 is the unique optimal configuration. In Section 5, we discuss related problems, some of which are as yet unsolved.

2. The angle argument Let O be the center of the given unit square. Let A and B be the centers of any two non-overlapping unit squares that kiss it. Youngs' strategy is to show that angle AOB must exceed 40°. Since $9 \cdot 40° = 360°$, 9 non-overlapping unit squares cannot kiss the given one.

Since the minimum distance from the center of a unit square to its boundary is $1/2$ and the maximum $\sqrt{2}/2$, each of OA, OB and AB is at least 1 and at most $\sqrt{2}$. We now show that, without loss of generality, we may take $OA = OB = \sqrt{2}$ and $AB = 1$.

We may assume that $OA \geq OB$.. If $OA < \sqrt{2}$, replace OAB by a similar triangle $OA'B'$ with $OA' = \sqrt{2}$. If $A'B' > 1$, replace $OA'B'$ by $OA'B''$ with $OB'' = OB'$ and $A'B'' = 1$. Note that angle $A'OB'' <$ angle $A'OB'$ = angle AOB. Since we are trying to show that angle $AOB > 40°$, we may replace AOB by $A'OB''$. See FIGURE 2.1. It follows that we can take $OA = \sqrt{2}$ and $AB = 1$.

In FIGURE 2.2, $OA = OB_2 = \sqrt{2}$ and $OB_1 = AB_1 = AB_2 = 1$. The point B must lie on the circular arc between B_1 and B_2 inclusive, the center of the arc being A. Clearly, angle AOB is minimal if, and only if, B is at B_2. It follows that we can take $OB = \sqrt{2}$.

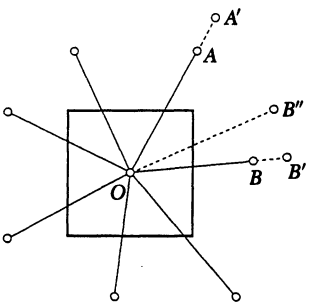


FIGURE 2.1

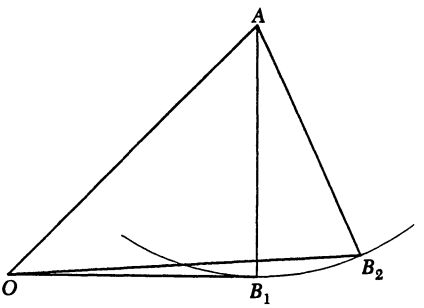


FIGURE 2.2

We now show that this minimum angle $AOB > 40°$. Let angle $AOB = 2\theta$. Then $\sin\theta = (AB/OA)/2 = \sqrt{2}/4$ and $\cos\theta = \sqrt{14}/4$. Hence

$$\cos 3\theta = 4\cos^3\theta - 3\cos\theta = \sqrt{14}/8 < \sqrt{16}/8 = \cos 60°.$$

It follows that $3\theta > 60°$ and $\theta > 20°$.

More directly but less aesthetically, one can use the Law of Cosines and look up the value of $\cos 40°$. It is possible to use a purely geometric argument, although we have not found one that is simple and elegant.

For completeness, we sketch Youngs' original proof. Let $OA = x$, $OB = y$ and $AB = t$, where $1 \leq x, y, t \leq \sqrt{2}$. Then

$$\cos AOB = (x^2 + y^2 - t^2)/2xy \leq F(x, y),$$

where $F(x, y) = (x^2 + y^2 - 1)/2xy$. Youngs used partial derivatives to show that the maximum of F occurs at the endpoint $F(\sqrt{2}, \sqrt{2}) = 3/4 < \cos 40°$.

3. The post argument In this section, we give another proof that $k(S) = 8$. The idea suggested in [2] is to erect a fence of length nl around a plane figure F, in such a way that a copy that kisses F itself must intersect the fence in a section of length at least l. It will then follow that $k(F) \leq n$. This fence argument requires a lot of detailed analysis, especially if the uniqueness of the optimal configuration is also considered. To cut down the number of cases, we introduce a set of evenly spaced posts along the fence.

FIGURE 3 shows a fence of length 8 for the unit square S. It is obtained by translating the sides of S outwards by a distance $1/2$. We put a post at each point of intersection of the fence with the extensions of the sides of S. They divide the fence into 8 sections of equal length.

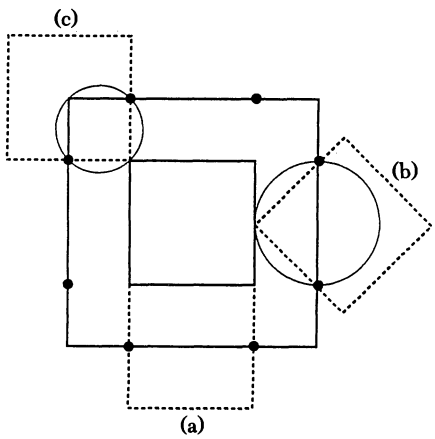


FIGURE 3

We set up a combinatorial "scoring" system as follows. A unit square that kisses S scores 2 points for each post in its interior, and 1 point for each post on its boundary.

Consider a collection C of non-overlapping unit squares kissing S. Clearly, if a post is in the interior of one of them, then it cannot be in the interior or on the boundary of another. Such a post contributes exactly 2 points to the total score. If a post is on the boundary of a square in C, it can be on the boundary of at most one other square in C. Hence it contributes at most 2 points to the total score.

Since there are 8 posts, the total score is at most 16. We now prove that each unit square T kissing S scores at least 2 points. It will then follow that the size of C is at most 8.

We may assume that T does not contain a post in its interior, as otherwise it scores 2 points at the very least. The boundary of T intersects the fence at two points X and Y. Note that $XY \leq 1$ since the distance between two adjacent posts is at most 1.

If X and Y lie on opposite sides of T, then $XY = 1$ and XY is parallel to a side of T. Hence T can only kiss S along a side of T, as shown in FIGURE 3(a). Both X and Y must be posts, and T scores exactly 2 points.

Suppose X and Y lie on adjacent sides of T. Then T must kiss S at a vertex V of T. Now angle $XVY = 90°$. Hence V lies on the circle with diameter XY. This circle cannot touch S unless both X and Y are posts, as shown in FIGURE 3(b) and FIGURE 3(c). Again T scores exactly 2 points, and the proof that $k(S) = 8$ is completed.

4. Uniqueness Let C be a collection of 8 non-overlapping unit squares kissing S. Then each square in C scores exactly 2 points while each post contributes exactly 2 points to the total score.

Suppose at least one square contains a post on its boundary. Then this square must have two posts on its boundary, each contributing 1 point to it. Each of these two posts must contribute 1 point to another square. It follows that every square in C contains two posts on its boundary.

A square T in C cannot kiss S as in FIGURE 3(b), since it will overlap any adjacent square in C. Hence C has four "corner" squares and four "side" squares (see FIGURE 1). This is the only optimal configuration that can arise here.

What happens if none of the posts lies on the boundary of any square in C? We shall shift one post along the fence until it does so, and shift the others so that they divide the fence into 8 equal sections. This may result in FIGURE 4.1 or FIGURE 4.2, the latter representing the general case.

As before, each post contributes at most 2 points to the total score. We claim that each square T in C must score at least 2 points with respect to this new set of posts. Again, we may assume that T does not contain any post in its interior.

If the section of the fence inside T is straight, the argument is the same as that for FIGURE 3(a) or FIGURE 3(b). Hence it is only necessary to deal with the case in FIGURE 4.2(c). We enlarge it to FIGURE 4.3.

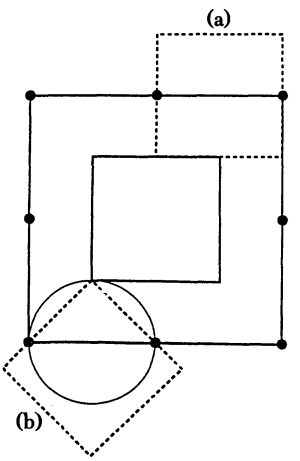


FIGURE 4.1

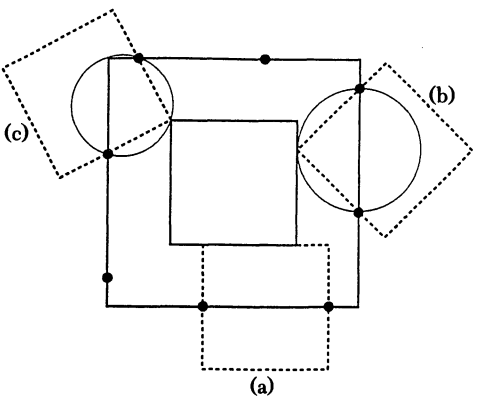


FIGURE 4.2

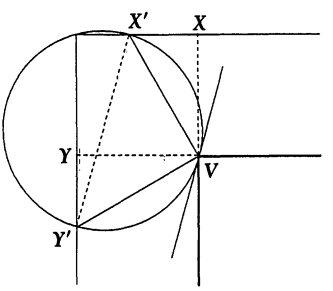


FIGURE 4.3

Here, X and Y are the original positions of the two posts, while X' and Y' are their new ones. Hence $XX' = YY'$. Let V be the corner of S closest to X and Y. Then $VX = 1/2 = VY$ and angle $VXX' = 90° =$ angle VYY'. It follows that triangles VXX' and VYY' are congruent, so that $VX' = VY'$ and angle $XVX' =$ angle YVY'.

Now angle $X'VY' = 90°$ since angle $XVY = 90°$. Hence the circle with diameter $X'Y'$ will pass through V. Moreover, the tangent to this circle at V is parallel to $X'Y'$, so that the circle touches S only at V. The rest of the argument is the same as that for FIGURE 3(c).

It follows that in each of FIGURE 4.1 and FIGURE 4.2, every square in C contains two posts on its boundary. We now complete the uniqueness argument.

We first consider FIGURE 4.1. A square T in C cannot kiss S as in FIGURE 4.1(a), as otherwise the corner post on T cannot contribute to any other square in C. If the squares in C all kiss S as in FIGURE 4.1(b), clearly two adjacent ones will overlap. Hence an optimal configuration cannot arise here.

Finally, consider FIGURE 4.2. The distance between the two posts nearest to a vertex of S is strictly less than 1. The square T of C containing these two posts on its boundary can kiss S only as in FIGURE 4.2(c). T must overlap an adjacent square of C, whether it kisses S as in FIGURE 4.2(a) or FIGURE 4.2(b). Hence an optimal configuration cannot arise here either.

5. Related problems The only other plane-tiling regular polygons are the equilateral triangle T and the regular hexagon H. The tilings suggest that $k(T) = 12$ and $k(H) = 6$, which are indeed correct. However, there are infinitely many optimal configurations for H, two of which are shown in FIGURE 5.1.

The angle argument does not yield any upper bound for $k(T)$, and only $k(H) \leq 7$. The fence argument yields sharp bounds for T, but only $k(H) \leq 12/\sqrt{3}$. Since this is less than 7, we do get $k(H) \leq 6$. The post argument establishes the uniqueness of the optimal configuration for T. We omit the details, which are essentially variations of those for the square. Note that the fence for T is not an equilateral triangle, but has the shape shown in FIGURE 5.2.

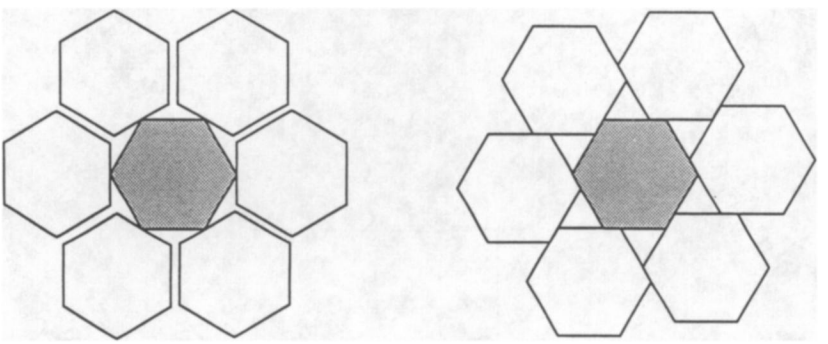


FIGURE 5.1

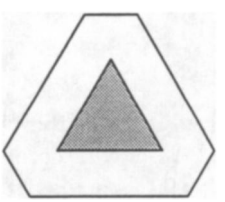


FIGURE 5.2

We can also generalize the problem into higher dimensions. For instance, the kissing number of the unit cube is at least 26, as suggested by the standard space-tiling. While this lower bound seems certainly to be sharp, no proof has yet been found.

Note that only $8 - 2 = 6$ non-overlapping unit squares can kiss a given one along segments with positive lengths. It is surprising that as many as $26 - 2 = 24$ non-overlapping unit cubes can kiss a given one along surfaces with positive areas (see [1]). It has not been proved that the lower bound of 24 cubes cannot be increased.

Acknowledgement. The authors thank the referees for detailed comments and suggestions that led to a substantial revision of the paper.

REFERENCES

1. Martin Gardner, *Wheels, Life and Other Mathematical Amusements*, W. H. Freeman, New York, 1983, pp. 88–93.
2. A. M. Gleason, R. E. Greenwood and L. M. Kelly, *The William Lowell Putnam Mathematical Competition–Problems and Solutions*: 1938–1964, MAA, Washington, DC, 1980, pp. 461–463.
3. C. J. A. Halberg, S. Levine and E. Strauss, On contiguous sets in Euclidean space, *Proc. Amer. Math. Soc.* 10 (1959) 335–344.
4. R. Honsberger, *Mathematical Morsels*, MAA, Washington, DC, 1978, pp. 82–89.
5. J. W. T. Youngs, A lemma on squares, *Amer. Math. Monthly* 46 (1939) 20–22.

Lament of a Professor at the End of the Spring Semester

I took an extra step to bridge the gap
between us, blind to your matching backward step.
We've moved in tandem until I'm angry
at you, and at me—I thought you needed
lenience, but my obstinence instead
would have changed the direction of our cadence
and given you a chance to lead the dance.

—JoAnne Growney
Bloomsburg University
Bloomsburg, PA 17815
(reprinted from *Intersections*,
with permission of the author)

We can also generalize the problem into higher dimensions. For instance, the kissing number of the unit cube is at least 26, as suggested by the standard space-tiling. While this lower bound seems certainly to be sharp, no proof has yet been found.

Note that only $8 - 2 = 6$ non-overlapping unit squares can kiss a given one along segments with positive lengths. It is surprising that as many as $26 - 2 = 24$ non-overlapping unit cubes can kiss a given one along surfaces with positive areas (see [1]). It has not been proved that the lower bound of 24 cubes cannot be increased.

Acknowledgement. The authors thank the referees for detailed comments and suggestions that led to a substantial revision of the paper.

REFERENCES

1. Martin Gardner, *Wheels, Life and Other Mathematical Amusements*, W. H. Freeman, New York, 1983, pp. 88–93.
2. A. M. Gleason, R. E. Greenwood and L. M. Kelly, *The William Lowell Putnam Mathematical Competition–Problems and Solutions*: 1938–1964, MAA, Washington, DC, 1980, pp. 461–463.
3. C. J. A. Halberg, S. Levine and E. Strauss, On contiguous sets in Euclidean space, *Proc. Amer. Math. Soc.* 10 (1959) 335–344.
4. R. Honsberger, *Mathematical Morsels*, MAA, Washington, DC, 1978, pp. 82–89.
5. J. W. T. Youngs, A lemma on squares, *Amer. Math. Monthly* 46 (1939) 20–22.

Lament of a Professor at the End of the Spring Semester

I took an extra step to bridge the gap
between us, blind to your matching backward step.
We've moved in tandem until I'm angry
at you, and at me—I thought you needed
lenience, but my obstinence instead
would have changed the direction of our cadence
and given you a chance to lead the dance.

—JoAnne Growney
Bloomsburg University
Bloomsburg, PA 17815
(reprinted from *Intersections*,
with permission of the author)

An Integral Polynomial

B. SURY
Tata Institute of Fundamental Research
Bombay 400005, India

On many occasions, we find ourselves surprised when some general formula being used for some specific purpose suddenly seems to yield something totally different. For instance, we discover while looking at the so-called Weyl dimension formula for the compact group $SU(n)$ that, for any n integers $a_1 < a_2 < \cdots a_n$ the fraction

$$\prod_{i>j} \frac{a_i - a_j}{i-j}$$

occurs as the dimension of some representation of this group. As a result, we see the unexpected fact that the fraction given above is always an integer. We would like to derive this fact by an elementary method. Surprisingly enough, this does not seem to be very easy to prove. For one thing, an induction argument invariably fails. Even more surprising is the fact that we can give an elementary proof of the following more general fact (whereas the proof itself cannot be applied directly to show the weaker result that the above fraction is an integer!).

THEOREM. *For any integers* $a_1 < a_2 < \cdots < a_n$

$$P(X) := \prod_{n \geq i > j \geq 1} \frac{X^{a_i - a_j} - 1}{X^{i-j} - 1} \in \mathbf{Z}[X].$$

Of course, by L'Hôpital's rule then

$$P(1) = \prod_{i>j} \frac{a_i - a_j}{i-j} \in \mathbf{Z},$$

which was our original assertion. As we will see, we can deduce more from the proof of the theorem.

Proof. Writing $X^r - 1 = \prod_{d/r} \Phi_d(X)$, where Φ_d is the dth cyclotomic polynomial ([1], Theorem 3.4) we have

$$P(X) = \prod_{i>j} \frac{\prod_{d/(a_i - a_j)} \Phi_d(X)}{\prod_{d/(i-j)} \Phi_d(X)}.$$

Fix any positive integer d. Since Φ_d is irreducible ([1], Theorem 3.7), we need only show that the power of $\Phi_d(X)$ occurring in the denominator is at the most the power occurring in the numerator. For $0 \leq i \leq d-1$, we let r_i denote the number of a's that are in the residue class i modulo d. Similarly, we denote by s_i the corresponding numbers when $\{a_1, \ldots, a_n\}$ is replaced by $\{1, \ldots, n\}$. Then $\Sigma r_i = \Sigma s_i = n$. Moreover, if we write $n = qd + r$, $0 \leq r < d$, then it is clear that $s_0 = q = s_i$ for $r < i < d$ and $s_j = q + 1$ for $0 < j \leq r$. The power of Φ_d dividing $\prod_{i>j}(X^{a_i - a_j} - 1)$ equals

$$\frac{1}{2} \sum_{i=0}^{d-1} r_i(r_i - 1) = \frac{1}{2} \sum_{i=0}^{d-1} r_i^2 - \frac{n}{2}.$$

It is reasonable to guess that $\Sigma_{i=0}^{d-1} r_i^2$ is minimum when the r_i are almost equal. To see that this is indeed true, we write $r_i = s_i + t_i$, with $t_i \in \mathbf{Z}$. Then, $\Sigma t_i = 0$. Now, if

$r = 0$ i.e. if d/n, then $s_i = q$ for all $0 \le i < d$ and $\Sigma r_i^2 = \Sigma(q + t_i)^2 = dq^2 + \Sigma t_i^2 \ge dq^2 = \Sigma s_i^2$. If $r > 0$, then

$$\begin{aligned}\sum r_i^2 &= (q + t_0)^2 + \sum_{i=1}^{r} (q + 1 + t_i)^2 + \sum_{i=r+1}^{d-1} (q + t_i)^2 \\ &= (d - r)q^2 + r(q + 1)^2 + \sum_{i=0}^{d-1} t_i^2 + 2\sum_{i=1}^{r} t_i \\ &\ge (d - r)q^2 + r(q + 1)^2 = \sum s_i^2,\end{aligned}$$

provided $\Sigma_{i=0}^{d-1} t_i^2 + 2\Sigma_{i=1}^{t} t_i \ge 0$. But, if $I = \{i : 1 \le i \le r,\ t_i = -1\}$, then

$$\begin{aligned}\sum_{i=0}^{d-1} t_i^2 + 2\sum_{i=1}^{r} t_i &\ge \sum_{0 \le i \le d-1,\, t > 0} t_i^2 + \sum_{1 \le i \le r,\, t_i < 0} (t_i^2 + 2t_i) \\ &\ge \sum_{0 \le i \le d-1,\, t_i > 0} t_i^2 + \sum_{1 \le i \le r,\, t_i = -1} t_i(t_i + 2) \\ &= \sum_{0 \le i \le d-1,\, t_i > 0} t_i^2 - |I| \\ &\ge \sum_{0 \le i \le d-1,\, t_i > 0} (t_i^2 - t_i) \ge 0\end{aligned}$$

since $\Sigma_{0 \le i \le d-1,\, t_i > 0} t_i \ge |I|$ by the equality $\Sigma_{i=0}^{d-1} t_i = 0$ where $|I|$ is the cardinality of I. This completes the proof of the theorem.

If we look at the proof carefully, we can guess at the following result.

Bonus result Let k be any natural number and let $a_1, \ldots, a_n$ be integers such that the number of a_i's in each residue class modulo k is the number of i's in that class. Then

$$\prod_{i \ne j} \frac{\prod_{a_i \equiv a_j(k)} (a_i - a_j)}{\prod_{i \equiv j(k)} (i - j)} \in \mathbf{Z}.$$

In particular, if $k = 1$, we need no restriction on the a_i's.

We notice that the above expression equals $P(e^{2\pi i/k})$. Consequently, $P(e^{2\pi i/k})$ is an algebraic integer as well as a rational number, which forces it to be a rational integer.

Remark We notice that in the proof of the theorem, the cyclotomic polynomials $\Phi_d(X)$ could be replaced by any irreducible polynomials $T_d(X)$ with integer coefficients. Then our argument goes through without change to show that

$$T(X) = \prod_{i > j} \frac{\prod_{d/(a_i - a_j)} T_d(X)}{\prod_{d/(i-j)} T_d(X)} \in \mathbf{Z}[X].$$

For instance, for each d, if we choose $T_d(X)$ to be the constant polynomial 2, we would get

$$\sum_{i > j} \tau(a_i - a_j) \ge \sum_{i > j} \tau(i - j),$$

where $\tau(n)$ is the number of divisors of n.

Acknowledgements. I would like to thank the referee for his/her detailed comments that helped in improving the exposition. I am also indebted to Amit Roy and Rajan for their assistance.

REFERENCE

1. Niven, *Irrational Numbers*, The Carus Mathematical Monographs, MAA, Washington, DC, 1956.

A Certain Property of an Isosceles Trapezoid and Its Application to Chain Circle Problems

FUKUZO SUZUKI
Gunma College of Technology
Maebashi Gunma 371, Japan

Introduction Various studies have already been made on the relations among the radii of osculating circles. In particular, some interesting findings were observed in votive tablets offered by old Japanese mathematicians (see [4]). In old times, only problems and results were indicated on votive tablets offered to shrines. It seems that problems were solved by an elementary method.

In this paper, invariant relations among the radii of osculating circles will be established by clarifying a certain property of the isosceles trapezoid and applying the inversion formula. Also, by applying this property to chain circle problems, invariant relations concerning the radii of chain circle sequences will be obtained, and it will be shown that beautiful results easily can be obtained concerning the radii of chain circle sequences that are observed in several votive tablets.

A certain property of an isosceles trapezoid As the first step, we prove the following lemma for isosceles trapezoids.

LEMMA 1. *If* $\square ABCD$ *is an isosceles trapezoid with* $AB = CD$ *and* S *is an arbitrary point, then*

$$\frac{SA^2 - SD^2}{AD} = \frac{SB^2 - SC^2}{BC}. \tag{1}$$

In particular, if $\square ABCD$ *is a rectangle, we have*

$$SA^2 - SD^2 = SB^2 - SC^2. \tag{2}$$

Proof. Let the feet of perpendiculars from S to the straight lines AD and BC be H and H', respectively. Let the mid-points of AD and BC be M and M', respectively (see FIGURE 1). Then, we have

$$SA^2 = SH^2 + AH^2,$$
$$SD^2 = SH^2 + DH^2, \quad \text{and}$$
$$|SA^2 - SD^2| = |AH^2 - DH^2| = 2MH \cdot AD.$$

Similarly

$$|SB^2 - SC^2| = 2M'H' \cdot BC.$$

Because $\square ABCD$ is an isosceles trapezoid, $MH = M'H'$. Furthermore, $SA^2 - SD^2$ and $SB^2 - SC^2$ are the same sign. Thus we have

$$\frac{SA^2 - SD^2}{AD} = \frac{SB^2 - SC^2}{BC}.$$

In particular, if $\square ABCD$ is a rectangle, $AD = BC$, and we obtain

$$SA^2 - SD^2 = SB^2 - SC^2.$$

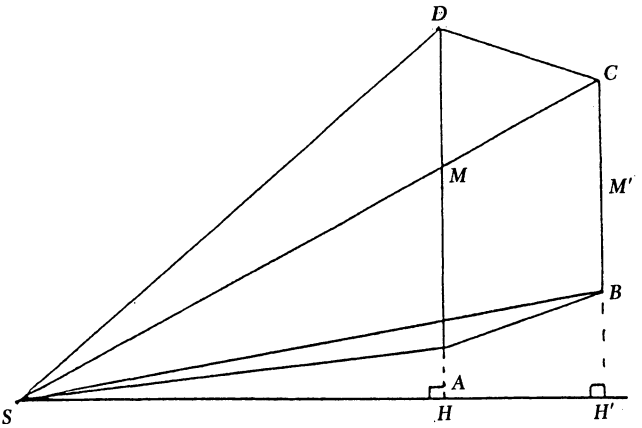


FIGURE 1

Characteristics of the inversion method and main theorems The circle whose center is C and whose radius is a is denoted by $C(a)$. Circle $O(r)$ is sometimes denoted simply by Circle O.

We have a given circle, called the *circle of inversion*, $S(\rho)$, where the point S is called the center of inversion and ρ the constant of inversion, and use this circle to effect a transformation of any given figure.

If P is a point of a given figure F, we find a point P' on SP such that

$$SP \cdot SP' = \rho^2.$$

Then P' is called the inverse of P and the set of points P', where $P \in F$ is called the inverse of F and is denoted by F'. Clearly, P is the inverse of P', and the points P on the circle of inversion $S(\rho)$ map into themselves.

THEOREM 1. *The inverse O' of a circle $O(r)$ that does not pass through S is a circle that does not pass through S and its radius r' satisfies*

$$\frac{1}{r'} = \frac{|d^2 - r^2|}{\rho^2 r}, \tag{3}$$

where $d = SO$.

Proof. From the above definitions, if S is outside Circle O,

$$ST \cdot ST' = \rho^2.$$

On the other hand, since $\triangle SOT$ is similar to $\triangle SO'T'$,

$$\frac{r}{r'} = \frac{ST}{ST'}.$$

Thus we have

$$\frac{1}{r'} = \frac{ST^2}{\rho^2} \cdot \frac{1}{r} = \frac{SO^2 - r^2}{\rho^2 r} = \frac{d^2 - r^2}{\rho^2 r}.$$

Similarly, if S is inside Circle O, we have

$$\frac{1}{r'} = \frac{r^2 - d^2}{\rho^2 r}$$

(see FIGURE 2). Thus we have (3).

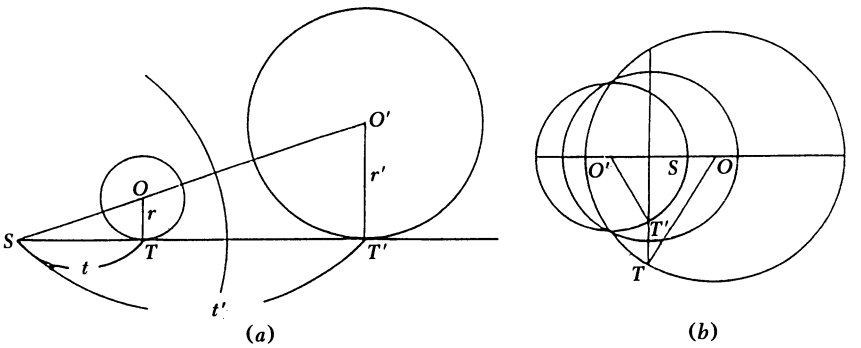


FIGURE 2

The circle whose center is O_j and whose radius is r_j is denoted by Circle O_j ($j = 1, 2, \ldots$) and SO_j by d_j. The length of the common external tangent of Circles O_j and O_l is denoted by t_{jl}, and O_jO_l is denoted by d_{jl}. For O'_j, an inverse circle of Circle O_j ($j = 1, 2, \ldots$), r_j, d'_j, t'_{jl}, etc. are defined similarly.

Then we have

COROLLARY 1. *Let the quadrangle $O_1O_2O_3O_4$ formed by the centers of four disjoint circles of the same radius r, called iso-circles, constitute a rectangle. If S is outside four circles, then, under inversion with respect to Circle $S(\rho)$, the image of a "rectangle of circles" satisfies*

$$\frac{1}{r'_1} + \frac{1}{r'_3} = \frac{1}{r'_2} + \frac{1}{r'_4}. \tag{4}$$

Proof. By Theorem 1 (3), we have

$$\frac{1}{r'_j} = \frac{d_j^2 - r^2}{\rho^2 r} \qquad (j = 1, 2, 3, 4).$$

Therefore,

$$\frac{1}{r'_1} - \frac{1}{r'_4} = \frac{d_1^2 - d_4^2}{\rho^2 r},$$

$$\frac{1}{r'_2} - \frac{1}{r'_3} = \frac{d_2^2 - d_3^2}{\rho^2 r}.$$

From Lemma 1 (2), we have

$$d_1^2 - d_4^2 = d_2^2 - d_3^2.$$

Thus, we obtain

$$\frac{1}{r'_1} + \frac{1}{r'_3} = \frac{1}{r'_2} + \frac{1}{r'_4}.$$

The next properties, which are derived from the definition of inversion, are well known (see [1], [4], and [5]).

(i) There exists an inversion mapping two non-intersecting circles to concentric circles.

(ii) There exists an inversion mapping any two circles to two iso-circles (circles with equal radii).
(iii) The inverse figure of a circle through S is a line not through S.
(iv) The angle of intersection of two curves is invariant under inversion.

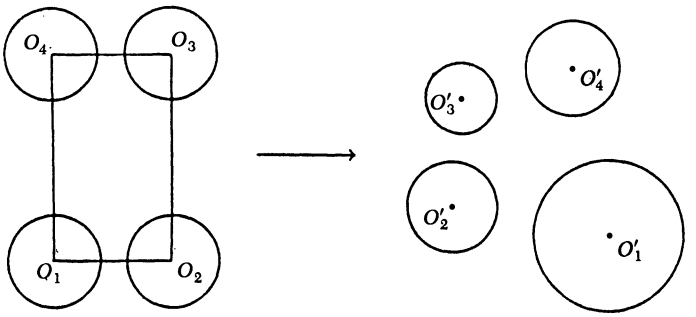


FIGURE 3

The locus of the points with equal powers of the length of the tangent to the two circles is a line and this line is called the *radical axis* of the two circles. The circles with common radical axis is called the *coaxial circles*. In (ii) let the center of inversion be a point on the coaxial circle of the two circles.

Moreover, the following proposition proposed by Casey gives an important invariant by inversion (for details, see [1] or [5]).

PROPOSITION. *Let t_{12} be the common external tangent of Circles $O_1(r_1)$ and $O_2(r_2)$ (see* FIGURE 4*). Then, the quantity*

$$\frac{t_{12}^2}{r_1 r_2} \tag{5}$$

is invariant under inversion.

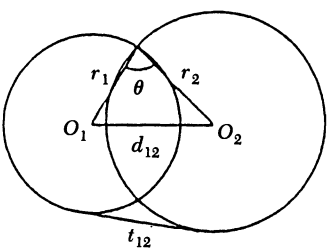


FIGURE 4

Proof. Because the angle of intersection θ of the two circles is invariant by inversion,

$$\cos\theta = \frac{r_1^2 + r_2^2 - d_{12}^2}{2r_1 r_2}$$

$$= \frac{(r_1 - r_2)^2 + 2r_1 r_2 - d_{12}^2}{2r_1 r_2}$$

$$= \frac{(r_1 - r_2)^2 - d_{12}^2}{2r_1 r_2} + 1$$

is invariant, too. But, because $t_{12}^2 = d_{12}^2 - (r_1 - r_2)^2$, relation (5) is invariant as well. If the two circles do not intersect each other, θ is regarded as an imaginary number.

THEOREM 2. *Let the centers of four circles O_1, O_2, O_3, and O_4 constitute an isosceles trapezoid with $O_1O_4 // O_2O_3$ and let $r_1 = r_4 = r$ and $r_2 = r_3 = R$. Then, under inversion in Circle $S(\rho)$, the image of a "trapezium of circles" satisfies*

$$\left(\frac{1}{r'_1} - \frac{1}{r'_4}\right)\frac{\sqrt{r'_1 r'_4}}{t'_{14}} = \left(\frac{1}{r'_2} - \frac{1}{r'_3}\right)\frac{\sqrt{r'_2 r'_3}}{t'_{23}}. \tag{6}$$

Proof. By Theorem 1 (3) and Proposition (5), we have (FIGURE 5)

$$\left(\frac{1}{r'_1} - \frac{1}{r'_4}\right)\frac{\sqrt{r'_1 r'_4}}{t'_{14}} = \frac{d_1^2 - d_4^2}{\rho^2 r} \cdot \frac{r}{t_{14}} = \frac{d_1^2 - d_4^2}{\rho^2 d_{14}}.$$

Similarly,

$$\left(\frac{1}{r'_2} - \frac{1}{r'_3}\right)\frac{\sqrt{r'_2 r'_3}}{t'_{23}} = \frac{d_2^2 - d_3^2}{\rho^2 d_{23}}.$$

On the other hand, because the quadrangle $O_1O_2O_3O_4$ is the isosceles trapezoid, the following equation is obtained from Lemma 1:

$$\frac{d_1^2 - d_4^2}{d_{14}} = \frac{d_2^2 - d_3^2}{d_{23}}.$$

Thus we obtain

$$\left(\frac{1}{r'_1} - \frac{1}{r'_4}\right)\frac{\sqrt{r'_1 r'_4}}{t'_{14}} = \left(\frac{1}{r'_2} - \frac{1}{r'_3}\right)\frac{\sqrt{r'_2 r'_3}}{t'_{23}}.$$

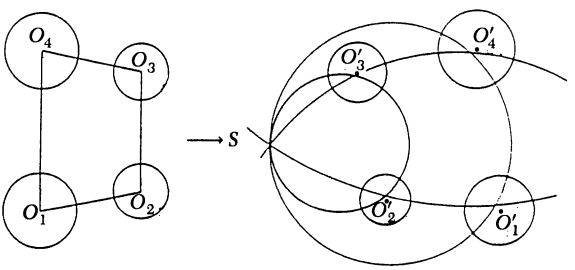


FIGURE 5

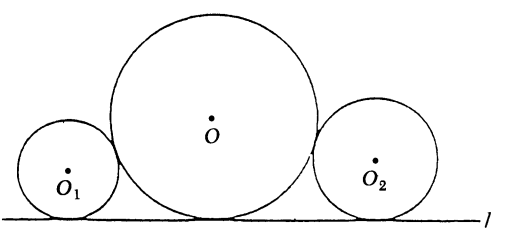


FIGURE 6

COROLLARY 2. *Let four circles O_1, O_2, O_3, and O_4, satisfy the conditions of Theorem 2. If O_1 is externally tangent to O_4 and O_2 to O_3, then we have*

$$\frac{1}{r'_1} + \frac{1}{r'_3} = \frac{1}{r'_2} + \frac{1}{r'_4}. \tag{7}$$

Proof. Because O_1 is externally tangent to O_4, the pairs-invariant is given by

$$\frac{t'_{14}}{\sqrt{r'_1 r'_4}} = \frac{t_{14}}{\sqrt{r_1 r_4}} = 2.$$

Similarly,

$$\frac{t'_{23}}{\sqrt{r'_2 r'_3}} = 2.$$

Thus we obtain (7) from (6).

THEOREM 3. *A circle O has a tangent line l. If two circles O_1 and O_2 both touch O externally and also the line l, the pairs-invariant is given by*

$$\frac{t_{12}}{\sqrt{r_1 r_2}} = 2\sqrt{r}\left(\frac{1}{\sqrt{r_1}} + \frac{1}{\sqrt{r_2}}\right). \tag{8}$$

Proof. Since three circles O, O_1, and O_2 come into contact with the line l, the length t_{12} of the tangent line of two circles O_1 and O_2 satisfies

$$t_{12} = 2\sqrt{r_1 r} + 2\sqrt{r_2 r}.$$

Thus, we have (8).

COROLLARY 3. *Let four circles O_1, O_2, O_3, and O_4 circumscribing a circle O satisfy the conditions of Corollary 2.1. If three image circles O', O'_1, and O'_4 and three image circles O', O'_2, and O'_3 come into contact with the respective lines, then we have*

$$\frac{1}{\sqrt{r'_1}} + \frac{1}{\sqrt{r'_3}} = \frac{1}{\sqrt{r'_2}} + \frac{1}{\sqrt{r'_4}}. \tag{9}$$

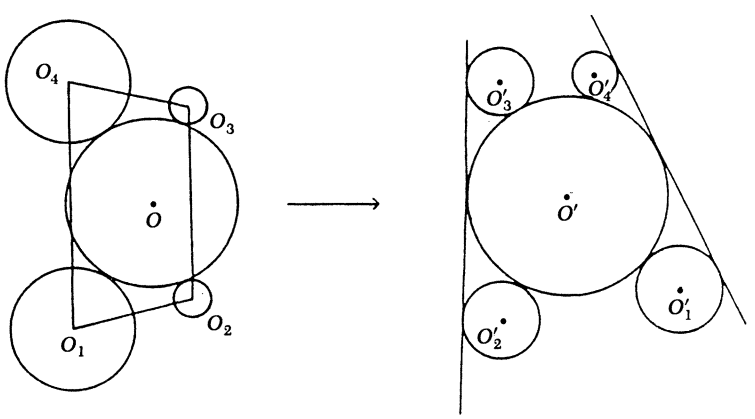


FIGURE 7

Proof. Because the conditions of Corollary 2.1 are satisfied, the following equation holds:

$$\left(\frac{1}{r'_1}-\frac{1}{r'_4}\right)\frac{\sqrt{r'_1r'_4}}{t'_{14}}=\left(\frac{1}{r'_2}-\frac{1}{r'_3}\right)\frac{\sqrt{r'_2r'_3}}{t'_{23}}. \tag{10}$$

On the other hand, as three circles O', O'_1, and O'_4 come into contact with one and the same line, from Theorem 3 (8), we have

$$\frac{t'_{14}}{\sqrt{r'_1r'_4}}=2\sqrt{r'}\left(\frac{1}{\sqrt{r'_1}}+\frac{1}{\sqrt{r'_4}}\right).$$

Similarly,

$$\frac{t'_{23}}{\sqrt{r'_2r'_3}}=2\sqrt{r'}\left(\frac{1}{\sqrt{r'_2}}+\frac{1}{\sqrt{r'_3}}\right).$$

When this is substituted into equation (10), we have (9).

Application to chain circle problems If O_0 and O are two given circles, and if $O_1, O_2, \ldots, O_n$ is a sequence of circles such that (a) O_i is tangent to both O_0 and O, $i = 1, 2, \ldots, n$, and (b) O_i and O_{i+1} are tangent, $i = 1, 2, \ldots, m-1$, then the sequence is called a *Steiner chain of length n relative to* O_0 *and* O. Note that the two circles O_0 and O may be disjoint, tangent, or intersecting.

THEOREM 4. *Let* $O_1, O_2, \ldots, O_m$ *and* $O_{m+1}, O_{m+2}, \ldots, O_{m+n}$ *be two Steiner chains relative to Circles* O_0 *and* O *having a length of m and n, respectively. Then*

$$\left(\frac{1}{r_i}-\frac{1}{r_l}\right)\frac{\sqrt{r_ir_l}}{t_{il}}=\left(\frac{1}{r_j}-\frac{1}{r_k}\right)\frac{\sqrt{r_jr_k}}{t_{jk}}, \tag{11}$$

where $j-i=l-k$ *and* $1 \le i < j \le m < k < l \le m+n$.

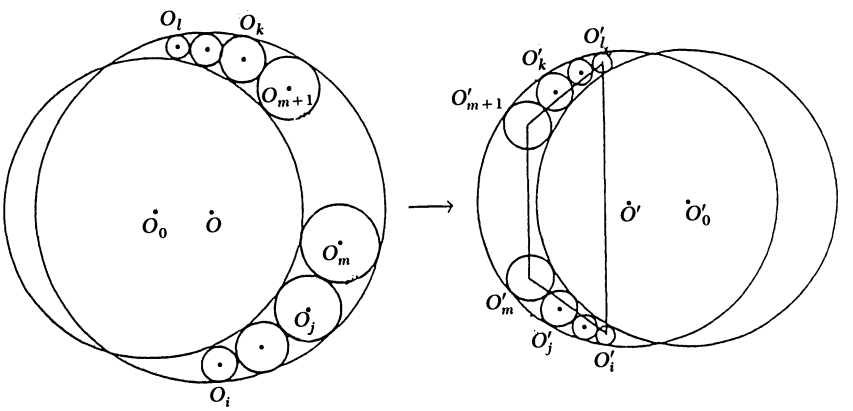


FIGURE 8

Proof. From Proposition 1 (ii), there exists an inversion that transforms the two circles O_i and O_l and two circles O_j and O_k to iso-circles. In the inverse figure, therefore, the quadrangle $O'_iO'_jO'_kO'_l$ formed by the centers of the circles constitutes an isosceles trapezoid with $O'_iO'_l // O'_jO'_k$, and also $r'_i = r'_l$ and $r'_j = r'_k$. Therefore, because the quadrangle of circles O_1, O_2, O_3 and O_4 is the inverse image of an "isosceles trapezium of circles" that satisfies the conditions of Corollary 2.1, Equation (11) follows.

Remark 1. This theorem is obtained in [5] by another elementary method applicable to the case in which Circle O_0 is included in Circle O. In our method, this theorem holds true independent of the positions of the two circles O_0 and O.

Some applications of Theorems 1–4 and Corollary 1–3 to chain circle problems are shown below.

In FIGURES 9 through 12, the following relation holds for the radii a, b, c, and d of four circles A, B, C, and D:

$$\frac{1}{\sqrt{a}} + \frac{1}{\sqrt{c}} = \frac{1}{\sqrt{b}} + \frac{1}{\sqrt{d}}. \tag{12}$$

In FIGURES 9 and 10, there exists an inversion mapping two circles O_1 and O_2 to concentric circles. In FIGURES 11 and 12, there exists an inversion mapping two pairs of two circles O_1 and O_2 and two circles A and D to two pairs of iso-circles. Thus we can prove (12) by using Corollary 3.

Remark 2. The relation (12) for the radii of four circles in FIGURE 9 is proven also by Michiwaki et al ([5], Theorem 2).

In FIGURES 13 through 18, the following relation holds for the radii a, b, c, and d of four circles A, B, C, and D:

$$\frac{1}{a} + \frac{1}{c} = \frac{1}{b} + \frac{1}{d}. \tag{13}$$

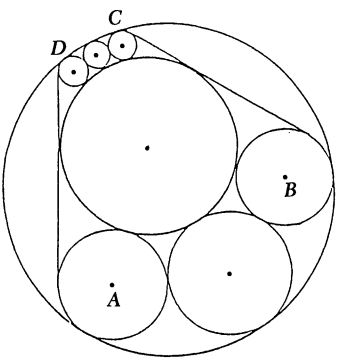


FIGURE 9

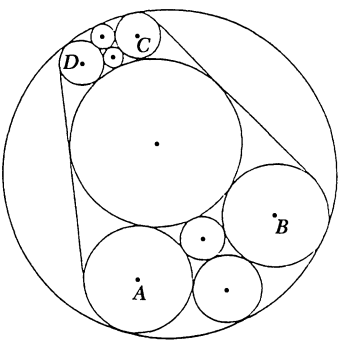


FIGURE 10

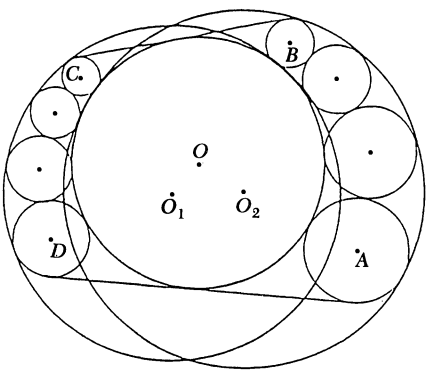


FIGURE 11

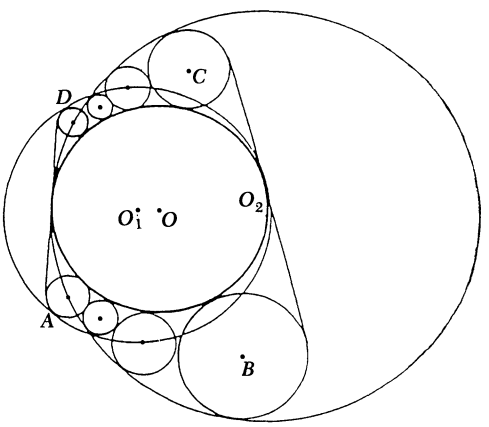


FIGURE 12

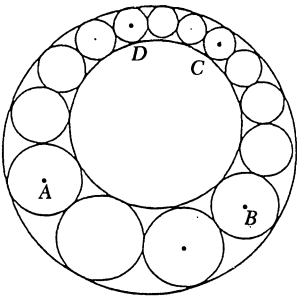


FIGURE 13

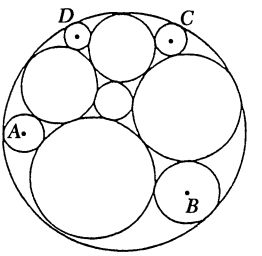


FIGURE 14

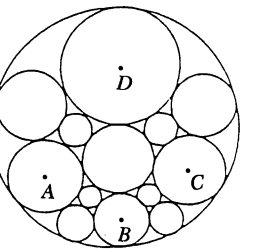


FIGURE 15

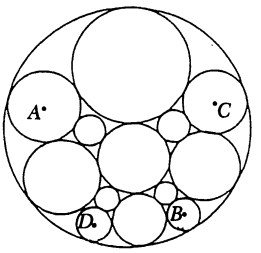


FIGURE 16

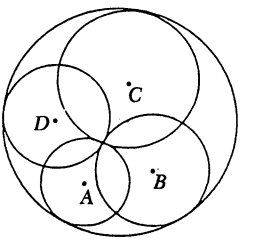


FIGURE 17

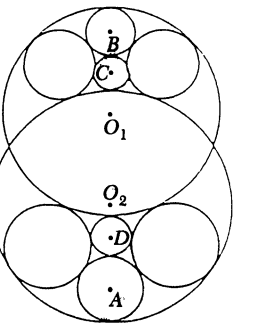


FIGURE 18

This property of FIGURES 13 through 17 can be easily shown from Corollary 1. In the case of FIGURE 18, the inverse circles appear as shown in FIGURE 19 when the circles are inverted with the crosspoint between two circles O_1 and O_2 as the center of inversion. Thus, as FIGURE 18 is the inverse image of FIGURE 19, relation (13) can be obtained from Corollary 2.2.

In FIGURES 20 and 21, the following relation holds for the radii of four circles A, B, C, and D:

$$\frac{1}{a} + \frac{1}{c} = \frac{1}{b} + \frac{1}{d}.$$

This can be shown readily from Corollary 1.

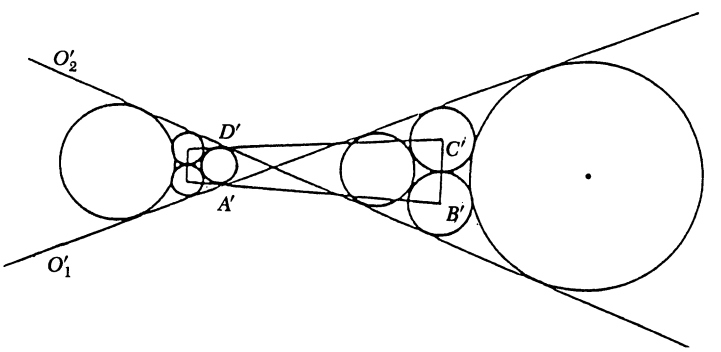


FIGURE 19

Remark 4. FIGURES 9 through 21 except for FIGURES 10, 11, and 12, are observable in votive tablets offered by old Japanese mathematicians. Specific historical and geographical references would be really interesting [4].

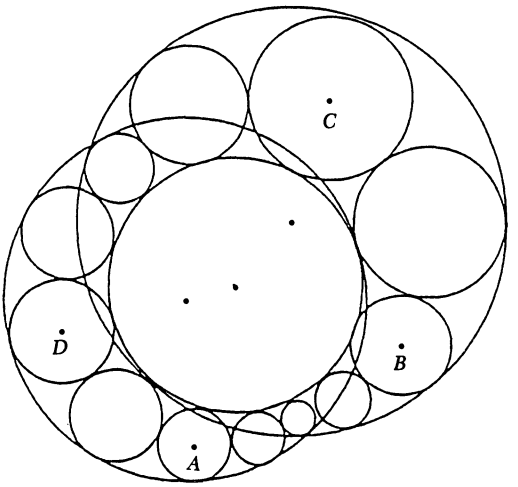


FIGURE 20

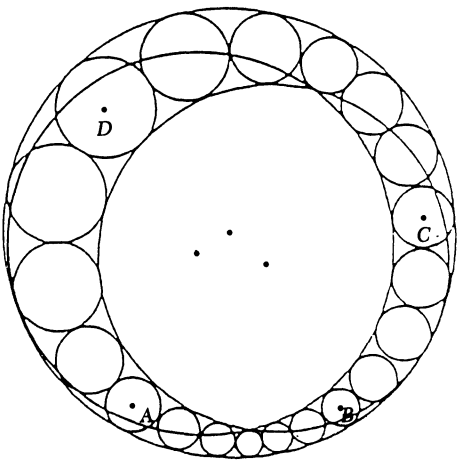


FIGURE 21

Acknowledgments. The author would like to thank the referees and the editor for their helpful suggestions in preparing this paper.

REFERENCES

1. H. S. M. Coxeter, *An Introduction to Geometry*, John Wiley & Sons, Inc., New York, 1961.
2. H. S. M. Coxeter, Mid-circles and loxodromes, *Math. Gazette* 52 (1968) 1–8.
3. H. Eves, *A Survey of Geometry*, Allyn & Bacon, Inc., Boston, 1972.
4. H. Fukagawa and D. Pedoe, *Japanese Temple Geometry Problems*-Sangaku Charles Babbage Research Center, Winnipeg, Canada, 1989.
5. R. A. Johnson, *Modern Geometry*, Houghton Mifflin, Boston, 1929.
6. Y. Michiwaki, M. Ōyama and T. Hamada, An invariant relation in chains of tangent circles, this MAGAZINE 48, 2 (1975), 80–87.
7. J. B. Wilker, Four proofs of a generalization of the Descartes Circle Theorem, *Amer. Math. Monthly* 76, 3 (1969), 278–282.

PROBLEMS

LOREN C. LARSON, *editor*
St. Olaf College

GEORGE GILBERT, *associate editor*
Texas Christian University

Proposals

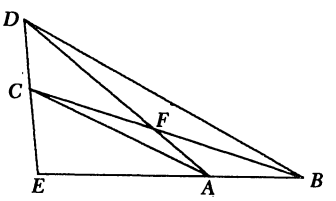
To be considered for publication, solutions should be received by September 1, 1995.

1469. *Proposed by Roger Izard, Dallas, Texas.*

In triangle EDB, shown below, A and C lie on EB and ED, respectively; CB and DA intersect at F. Also,

$$\frac{\text{Area}(\triangle EDB)}{\text{Area}(\triangle ECA)} = 6, \quad DC \cdot AB = 4, \quad \text{and} \quad \text{Area}(\triangle CFA) + \text{Area}(\triangle DFB) = 14/5.$$

Prove that DEB is a right triangle.



1470. *Proposed by John A. Hendrickson, Jr., Academy of Natural Sciences, Philadelphia, Pennsylvania.*

Consider the set of $n \times n$ matrices A, $n \geq 3$, with all entries equal to 0 or 1 and the sum of the entries of each row and of each column equal to 2. Find the maximum and minimum number of quadruples (i, i', j, j'), $i < i'$, for which $A_{ij} = A_{i'j'} = 1$ and $A_{ij'} = A_{i'j} = 0$ as a function of n. For example, there are 12 such quadruples in the

---

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

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

Solutions and new proposals should be mailed in duplicate to Loren Larson, Department of Mathematics, St. Olaf College, 1520 St. Olaf Ave., Northfield, MN 55057-1098 or mailed electronically via fax: (507) 663-3549 or e-mail: larson@stolaf.edu.

matrix

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

1471. *Proposed by Bill Correll, Jr. (student), Denison University, Granville, Ohio.*

For positive integers n, define $f(n)$ to be the smallest positive integer j such that

$$\left\lfloor \frac{n^2}{j} \right\rfloor = \left\lfloor \frac{n^2}{j+1} \right\rfloor,$$

where $\lfloor x \rfloor$ denotes the floor function. Let

$$c(n) = \left\lfloor \frac{n^2}{f(n)} \right\rfloor.$$

Prove that

(i) $\{f(n)\}_{n=1}^{\infty}$ consists of all integers except the perfect squares, and

(ii) $f(n) + c(n) = 2n$.

1472. *Proposed Erhan Gürel, Middle East Technical University, Ankara, Turkey.*

Let Q denote an arbitrary convex quadrilateral inscribed in a fixed circle, and let $\mathcal{F}(Q)$ be the set of inscribed convex quadrilaterals whose sides are parallel to those of Q. Prove that the quadrilateral in $\mathcal{F}(Q)$ of maximum area is the one whose diagonals are perpendicular to one another.

1473. *Proposed Gerald A. Heuer, Concordia College, Moorhead, Minnesota.*

Let $B_n(z)$ denote the determinant of the $(2n+1) \times (2n+1)$ matrix whose entries are given by $b(1, j) = 1$ for all j, $b(j, j) = 2$ for $j = 2, 3, \ldots, 2n+1$,

$$b(i+1, n+i+1) = b(n+i+1, i) = -z \quad \text{for } i = 1, 2, \ldots, n,$$

and all other $b(i, j) = 0$. For example,

$$B_1(z) = \det \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & -z \\ -z & 0 & 2 \end{pmatrix} \qquad B_2(z) = \det \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & -z & 0 \\ 0 & 0 & 2 & 0 & -z \\ -z & 0 & 0 & 2 & 0 \\ 0 & -z & 0 & 0 & 2 \end{pmatrix}.$$

Find all (complex) roots of $B_n(z)$.

Quickies

Answers to the Quickies are on page 152.

Q832. *Proposed by Michael Golomb, Purdue University, West Lafayette, Indiana.*

Show that for every positive integer n,

$$\int_{-\infty}^{\infty} \left(\frac{\sin x}{x} \right)^n dx \Big/ \int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx$$

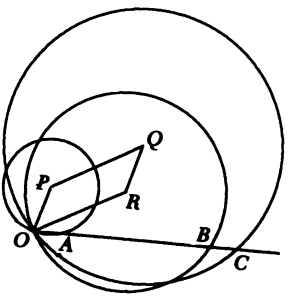
is a rational number.

Q833. *Proposed by David Callan, University of Wisconsin, Madison, Wisconsin.*
Show that for any nonzero integer n,

$$\sum_i \binom{n}{i}\binom{-n}{i} = 0.$$

(For $n = 3$, this reads $1 \cdot 1 + 3 \cdot (-3) + 3 \cdot 6 + 1 \cdot (-10) = 0$.)

Q834. *Proposed by Ismor Fischer, University of Wisconsin, Oshkosh, Wisconsin.*
Consider a parallelogram $OPQR$ in the plane, and three circles centered at P, Q, and R, having radii $|\overline{OP}|$, $|\overline{OQ}|$, and $|\overline{OR}|$, respectively (see FIGURE). For any ray originating from point O that intersects the circles in distinct points A, B, and C (in that order), prove that $|\overline{OA}| = |\overline{BC}|$.



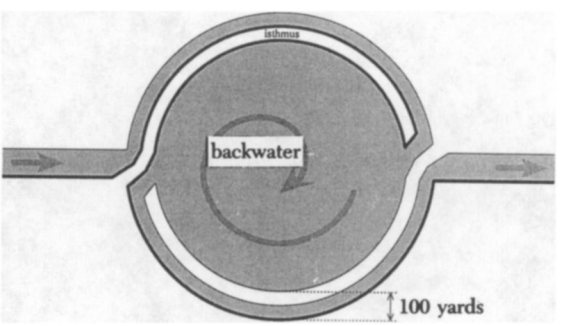
Solutions

Down the river **April 1994**

1443. *Proposed by Allen J. Schwenk, Western Michigan University, Kalamazoo, Michigan.*
A river flows from Town A to Town B and has the property that any point on either of its banks is no farther than 100 yards from some point on the other bank.

a. A boat sails down the river, trying to stay always within a distance d from both banks. For what values of d can we guarantee that such a trip is possible?

b. How far can a swimmer in the river be from the nearest bank?



Solution by Hans Georg Killingbergtrø, Horten, Norway.

a. Clearly $d \geq 100$, because, in this case, the boat could travel along either one of the banks and be within d of the opposite bank. The map shows that d cannot be less than 100.

b. The map also shows that the swimmer can be arbitrarily far from both banks, because the backwater lagoon can be made arbitrarily large.

Also solved by David K. Cahoon, Davide P. Cervone, L. R. King, O. P. Lossers (The Netherlands), Stephen Noltie, Billy D. Read, and the proposer.

Trapezoid **April 1994**

1444. *Proposed by Cristian Turcu, London, England*

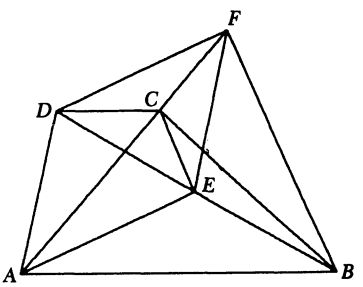
In the following figure, $ABCD$ is a trapezoid, with AB parallel to CD, and the length of AB is the sum of the lengths of AC and CD. E is the midpoint of BD, and F is a point on AC such that BF is parallel to CE.

Prove that

a. AE and DF are perpendicular to BF;

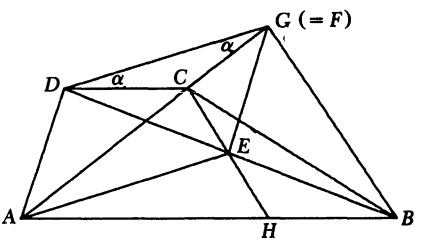
b. C is the incenter of $\triangle DEF$ if, and only if, AD is perpendicular to AB;

c. EF is parallel to AD if, and only if, the length of AB is 3 times the length of CD.



Solution by Helen M. Marston, Princeton, New Jersey.

Extend ray AC to a point G so that $CG = CD$, let $\alpha = \angle CGD$, and let H denote the intersection of CE and AB.



a. We have $\triangle CDE \cong \triangle HBE$, and therefore $AC = AH$. Also, E is the midpoint of the base of isosceles triangle CAH, and therefore AE is perpendicular to CH. The equal sides of isosceles triangle CAH and the equal sides of isosceles triangle GAB are parallel, and therefore CH and BG are also parallel. It follows that $G = F$. $\angle DCA = \angle CAB = 2\alpha$, so $\angle AFB = (\pi - 2\alpha)/2 = \pi/2 - \alpha$; thus, $\angle DFB = \pi/2$. It follows that DF is also perpendicular to BF.

b. If C is the incenter of $\triangle DEF$, then $\angle CDE = \alpha$, and therefore $\angle DBA = \alpha$. It follows that $BE = EA = ED$ and therefore $\angle DAB = 90°$. Conversely, if $\angle DAB = 90°$, then $DE = EB = EA$ and therefore $\angle ABD = \alpha$ and $\angle BDC = \alpha$. Furthermore, $\angle EFC = \alpha$ since $\angle ABF = \angle AFB$ and $\angle EBF = \angle EFB$. Therefore C is the incenter of $\triangle DEF$.

c. Let $CD = r$. If $AB = 3r$, then (since $HB = r$) $AH = 2r$ and $AE = 2r \cos \alpha = DF$. Since AE and DF are parallel and equal, $AEFD$ is a parallelogram, and therefore EF is parallel to AD. Conversely, if EF is parallel to AD, then (since DF is parallel to AE) $DF = AE = 2r \cos \alpha$, so $AH = 2r$ AND $AB = 3r$.

Also solved by Duane M. Broline, Himadri Choudhury (student), Bill Correll, Jr. (student), David Doster, Robert L. Doucette, Ragnar Dybvik (Norway), Frank Eccles, Milton P. Eisner, Francis M. Henderson, John Henle, John G. Heuver (Canada), Dixon Jones, John W. Krussel, Kee-Wai Lau (Hong Kong), Henry S. Lieberman, Nick Lord (England), O. P. Lossers (The Netherlands), Can Anh Minh (student), Robert Patenaude, Billy D. Read, Tetsuo Shinkawa, Jose Ines Escalante Vazquez and Abel Hernandez-Castillo (Mexico), Harry Weingarten, Wai Ling Yee (student, Canada), Sammy Yu (student) and Jimmy Yu (student), David Zhu, and the proposer.

Orthonormal triplets **April 1994**

1445. *Proposed by Ilya V. Burkov, St. Petersburg Technical University, St. Petersburg, Russia.*

Let $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ and $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ be orthonormal right-oriented triplets of vectors in $\mathbf{R}^3$, and k_1, k_2, k_3 be nonzero real numbers with different absolute values such that

$$k_1(\mathbf{r}_1 \times \mathbf{s}_1) + k_2(\mathbf{r}_2 \times \mathbf{s}_2) + k_3(\mathbf{r}_3 \times \mathbf{s}_3) = \mathbf{0}.$$

Prove that $\mathbf{r}_i$ is parallel to $\mathbf{s}_i$, $i = 1, 2, 3$.

Solution by Noah Rosenberg, student, Rice University, Houston, Texas.

Taking the inner product of the given equation with $\mathbf{r}_1$ we obtain

$$k_1(\mathbf{r}_1 \cdot (\mathbf{r}_1 \times \mathbf{s}_1)) + k_2(\mathbf{r}_1 \cdot (\mathbf{r}_2 \times \mathbf{s}_2)) + k_3(\mathbf{r}_1 \cdot (\mathbf{r}_3 \times \mathbf{s}_3)) = 0.$$

Using the scalar triple product identity $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$, and simplifying using the anti-commutativity of the cross product and the hypothesis that $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ is an orthonormal right-handed triplet, we get

$$k_2(\mathbf{r}_3 \cdot \mathbf{s}_2) - k_3(\mathbf{r}_2 \cdot \mathbf{s}_3) = 0.$$

Similarly, taking the inner product of the given equation with $\mathbf{s}_1$, we obtain

$$-k_2(\mathbf{r}_2 \cdot \mathbf{s}_3) + k_3(\mathbf{r}_3 \cdot \mathbf{s}_2) = 0.$$

Solving this linear system in $\mathbf{r}_2 \cdot \mathbf{s}_3$ and $\mathbf{r}_3 \cdot \mathbf{s}_2$ we find that either (i) $k_2^2 = k_3^2$, or (ii) $\mathbf{r}_2 \cdot \mathbf{s}_3 = \mathbf{r}_3 \cdot \mathbf{s}_2 = 0$. Since k_1, k_2, and k_3 differ in absolute value, (i) cannot hold. Therefore, $\mathbf{r}_2 \cdot \mathbf{s}_3 = \mathbf{r}_3 \cdot \mathbf{s}_2 = 0$.

In the same manner we find that $\mathbf{r}_1 \cdot \mathbf{s}_3 = \mathbf{r}_3 \cdot \mathbf{s}_1 = 0$ and $\mathbf{r}_2 \cdot \mathbf{s}_1 = \mathbf{r}_1 \cdot \mathbf{s}_2 = 0$.

Hence we see that $\mathbf{r}_1$ is perpendicular to $\mathbf{s}_2$ and $\mathbf{s}_3$. Thus, $\mathbf{r}_1$ must be parallel to $\mathbf{s}_2 \times \mathbf{s}_3 = \mathbf{s}_1$. Similarly, $\mathbf{r}_2$ is parallel to $\mathbf{s}_2$ and $\mathbf{r}_3$ is parallel to $\mathbf{s}_3$. This completes the proof.

Also solved by Michael H. Andreoli, Duane M. Broline, John Christopher, Robert L. Doucette, Robert Gardner, Thomas Jager, Hans Kappus (Switzerland), Nick Lord (England), O. P. Lossers (The Netherlands), Can Anh Minh (student), Richard Pfiefer, Nora S. Thornber, and the proposer.

Graphing **April 1994**

1446. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Canada.*

Determine the least number of times the graph of

$$y = \frac{a^2}{x^2 - 1} + \frac{b^2}{x^2 - 4} + \frac{c^2}{x^2 - 9} - 1$$

intersects the x-axis (a, b, c are nonzero real constants).

Solution by Jerrold W. Grossman, Oakland University, Rochester, Michigan.

The graph always intersects the x-axis six times. By summetry, it suffices to consider $x \geq 0$. It is clear that $y < 0$ when $x = 0$, that $\lim_{x \to n^-} y = -\infty$ for $n = 1, 2, 3$, that $\lim_{x \to n^+} y = \infty$ for $n = 1, 2, 3$, and that $\lim_{x \to \infty} y = -1$. Furthermore,

$$\frac{dy}{dx} = -2x\left(\frac{a^2}{(x^2-1)^2} + \frac{b^2}{(x^2-4)^2} + \frac{c^2}{(x^2-9)^2}\right),$$

so y is decreasing on $(0,1)$, $(1,2)$, $(2,3)$, and $(3,\infty)$. Since y is also continuous on these intervals, it follows from all of these statements that there is precisely one x-intercept in each of $(1,2)$, $(2,3)$, and $(3,\infty)$ and no x-intercept in $(0,1)$.

Also solved by Casey Abell, Carl Axness, Duane M. Broline, John Christopher, David K. Cohoon, Charles K. Cook, Bill Correll, Jr. (student), Daniel J. Curtin, Thomas Vanden Eynden, Peter Flanagan-Hyde, Arthur H. Foss, David Hankin, Robert M. Hashway, Richard Heeg, Thomas Jager, Hans Kappus (Switzerland), Nick Lord (England), O. P. Lossers (The Netherlands), Patrick Dale McCray, Richard F. Ryan, Heinz-Jürgen Seiffert (Germany), Nora S. Thornber, Robert J. Wagner, and the proposer.

Tetrahedron — April 1994

1447. *Proposed by Florin S. Pîrvănescu, Slatina, Romania.*

Let M denote an arbitrary point inside or on a tetrahedron $A_1A_2A_3A_4$, and let B_i be a point on the face F_i opposite vertex A_i, $i = 1, 2, 3, 4$. For each i, let M_i be the point where the line through M parallel to A_iB_i intersects F_i. Show that

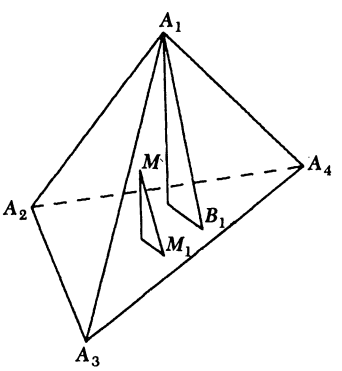
$$\min_{1 \leq i \leq 4} A_iB_i \leq \sum_{i=1}^{4} MM_i \leq \max_{1 \leq i \leq 4} A_iB_i.$$

Solution by O. P. Lossers, Technical University Eindhoven, Eindhoven, The Netherlands.

Both inequalities follow from multiplying the identity

$$\sum_{i=1}^{4} \frac{MM_i}{A_iB_i} = 1,$$

by $\min A_iB_i$ and $\max A_iB_i$, respectively. The identity follows from the fact that MM_i/A_iB_i is the fraction of the volume of the tetrahedron in the cone of M over F_i (see FIGURE).



Also solved by Robert Doucette, Can Anh Minh (student), and the proposer.

Answers

Solutions to the Quickies on page 147.

A832. Repeated integration by parts gives

$$\int_{-\infty}^{\infty} x^{-n} \sin^n x\,dx = \frac{1}{n!}\int_{-\infty}^{\infty} x^{-1} D^{n-1} \sin^n x\,dx.$$

But

$$\sin^n x = \begin{cases} \displaystyle\sum_{k=1}^{n/2} a_k \cos 2kx & \text{if } n \text{ is even,} \\ \displaystyle\sum_{k=1}^{(n+1)/2} b_k \sin(2k-1)x & \text{if } n \text{ is odd,} \end{cases}$$

where a_k, b_k are rational numbers. In either case,

$$D^{n-1} \sin^n x = \sum_{k=1}^{n} c_k \sin kx,\ c_k \text{ rational.}$$

But $\int_{-\infty}^{\infty} x^{-1} \sin kx\,dx = \int_{-\infty}^{\infty} x^{-1} \sin x\,dx$, which gives the result.

A833. By symmetry, assume $n > 0$ and look at the coefficient of x^n in $(1+x)^n(1+x)^{-n}$ (using $\binom{n}{i} = \binom{n}{n-i}$ for $0 \le i \le n$).

A834. With point O as origin, let points P, Q, and R be denoted by (h_i, k_i), $i = 1,2,3$ respectively. Note that $h_1 + h_3 = h_2$ and $k_1 + k_3 = k_2$. The equations of the circles in polar coordinates are, $r_i = 2h_i \cos\theta + 2k_i \sin\theta$, $i = 1,2,3$. Hence, for any suitable angle θ, we have $r_1 + r_3 = r_2$, or $|\overline{OA}| + |\overline{OB}| = |\overline{OC}|$. Subtracting $|\overline{OB}|$ from both sides gives the result.

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.

Kolata, Gina, How a gap in the Fermat proof was bridged, *New York Times* (National Edition) (31 January 1995) B1, B8. Cipra, Barry, Further Fermat ferment, *SIAM News* 27 (10) (December 1994) 2. Leutwyler, Kristin, Finessing Fermat, again, *Scientific American* (February 1995) 16, 18.

The final element of the proof of Fermat's Last Theorem, contained in a paper by Andrew Wiles (Princeton University) and his former student Richard Taylor (Cambridge University), has been read by experts, who say that it is correct. The authors refer to the result as "the theorem of Wiles, completed by Taylor and Wiles." The inspiring news for mathematics students is that the broad nature of the gap, and the technique for filling it, involve techniques familiar to mathematics majors. Wiles needed to glue together an infinite collection of Hecke rings, and his initial attempt used an inductive definition. When that didn't work, he turned to the pigeonhole principle to show that a structure of the desired kind had to exist. "The complete argument involves creating an infinite sequence of sets of pigeonholes and then showing that there must be objects that show up in every set of pigeonholes." Now, on to the full Taniyama conjecture, and the other steps in the Langlands program—but not for Wiles: "It is nice to finish it off, but, for me, the mystery is gone. The challenge now is to go on to problems where no one has any idea where to start."

Godbole, Anant P., Probability theory: A program of undergraduate research, *Council on Undergraduate Research Quarterly* (December 1994) 96–102.

Author Godbole describes his experience as the advisor in an NSF Research Experiences for Undergraduates site at Michigan Technological University for the past four summers. He tells how and why he started his program, what he regards as the keys to its successes, and what his students have worked on (and published). He solicits for the next few issues of the *Council on Undergraduate Research Quarterly* further articles on "the vast spectrum of experiences and possibilities" for undergraduate research in mathematics and computer science. (Thanks to Joe Gallian, University of Minnesota—Duluth.)

Gardner, Martin, Notes of a fringe-watcher: The cult of the golden ratio, *Skeptical Inquirer* 18 (Spring 1994) 243–247.

Despite the fondness of teachers of mathematics for the "golden ratio" $\phi = (1 + \sqrt{5})/2$, the limit of successive terms of the Fibonacci sequence, virtually all of the claims about its application in aesthetics, architecture, painting, sculpture, nature, poetry, and music are nonsense. There is *no* evidence that the golden rectangle is the most pleasing shape for a rectangle, nor that da Vinci or any other Renaissance artist or sculptor used the ratio in his work (some twentieth-century artists have), nor that a person's navel divides their height in the golden ratio, nor that Virgil used ratio in the *Æneid*. Gardner cites "Misconceptions about the golden ratio," by George Markowsky, in the *College Mathematics Journal* 23 (1992) 2–19, and other sources, as well as a 30-year-old article of his own "where I myself fell for some misconceptions."

Brassard, Gilles, Cryptology column—Quantum computing: The end of classical cryptography?, *ACM SIGACT News* 25 (4) (December 19994) 15–21. Shor, Peter W., Algorithms for quantum computation: Discrete logarithms and factoring, to appear in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*; preprint available at WWW location `http://vesta.physics.ucla.edu:7777/` . Brown, Julian, A quantum revolution for computing, *New Scientist* (24 September 1994) 21–24.

Is there a quantum computer in your future? If so, NP-complete problems, like factoring large integers, will be easy. Peter W. Shor (AT&T Bell Labs) has shown how to factor arbitrarily large integers in polynomial time on a quantum computer. That is, any efficient simulation of quantum physics on a classical computer would yield efficient classical algorithms for factoring and for extracting discrete logarithms. Because the difficulty of these problems is the basis for the security of much of current cryptography, current cryptographic methods would be replaced by quantum cryptography. But what is a quantum computer? Its "operations are not restricted by the laws of classical physics," such as taking only one of the potential paths in a computation; "all potential paths are taken simultaneously, in accord with the superposition principle of quantum mechanics"—and they *interfere*, some constructively and some destructively. Even if the probability of going from state a to state b through an intermediary state is nonzero for each of two intermediary states, the probability of going from a to b may be zero. (Welcome to the world of quantum logic!) Of course, if computer scientists can prove that $\mathcal{P} \neq \mathcal{NP}$, there can't be quantum computers, and you can forget about getting one.

Cohen, Daniel E., The mathematician who had little wisdom: A story and some mathematics, in *Combinatorial and Geometric Group Theory: Edinburgh 1993*, edited by Andrew J. Duncan, N.D. Gilbert, and James Howie, Cambridge University Press, 1995, pp. 56–62.

I won't spoil the story by summarizing it here, but I will tell you that you will learn some mathematics from it. I will also tell you the moral: A mathematician should also read computer science journals (including back issues), as much mathematics appears there, too. One important reason is to avoid trying to prove what computer scientists already have established is impossible.

Ayers-Nachamkin, Beverly, A Feminist approach to the introductory statistics course, *Women's Studies Quarterly* (1992) 20 (1 & 2) 86–94.

This pedagogical article by a social psychologist tries to justify to her feminist colleagues her teaching of statistics, "one of the premiere tools of the patriarchal trade." An admitted math avoider (no math in high school), Ayers-Nachamkin realized that the "math-anxious approach" that she was using in teaching was actually an approach urged by feminist philosophy/ideology. The main ingredients are de-emphasizing the authority role of the professor ("I am not an expert in statistics") and increasing cooperative learning. The techniques include relaxation techniques and eliminating graded homework, timed tests (eminently feasible under an honor code, as she points out), and memorization of formulas. She asserts that "my lectures and tests could and should be used to convey Feminist principles." The topics for problems that she uses (at a liberal arts college for women) seem to appeal to stereotyped values of women, such as testing the effect of different flours on the texture of angel-food cake; and she avoids "traditional statistic settings" such as sports records or results from "sterile laboratory domains," the last being "unacknowledged conveyors of patriarchal values." Regardless of your opinion of all this, I urge you to look up the article for its valuable appendix, an edited version of a "Math Anxiety Bill of Rights (and Responsibilities)," by Sandra L. Davis (no source cited).

Stevenson, D.E., Science, computational science, and computer science: At a crossroads, *Communications of the Association for Computing Machinery* 37 (12) (December 1994) 85–96.

For some scientists, science now has three branches: theory, experiment, and computation. "Our purpose is to ask whether today's computer scientists are able to take up the challenge of computational science." Author Stevenson answers in the negative and goes on to explain computational science and advertise the program at Clemson University. He sees computational science as interdisciplinary, standing on a tripod of, first and foremost, applications to science and engineering, then algorithms and architectures. "Computer science is generally not participating in science and engineering applications, nor is it preparing students to do so in the future." Watch for lots of angry letters in the March and April issues.

Adleman, Leonard M., Molecular computation of solutions to combinatorial problems, *Science* 266 (11 November 1994) 1021–1024. Peterson, I., Molecular computing in a DNA soup, *Science News* 146 (12 November 1994) 308. Gifford, David K., On the path to computation with DNA, *Science* 266 (11 November 1994) 993–994. Kolata, Gina, Scientist at work: Leonard Adleman: Hitting the high spots of computer theory, *New York Times* (National Edition) (13 December 1994) B5, B8.

Leonard Adleman (University of Southern California) is famous among mathematicians as discoverer of probabilistic proofs for primality, and among computer scientists as coinventor of the RSA encryption scheme and inventor of the term "computer virus." His latest endeavor is to compute directly with molecules. He conducted a laboratory experiment in which he solved an instance of the directed hamiltonian path problem, an NP-complete problem. Vertices were represented as short DNA sequences and (directed) edges as oligonucleotides (strands of DNA), and the solution emerged as the presence of a particular DNA molecule. "In effect, Adleman has used the enormous parallelism of solution-phase chemistry to solve a hard computational problem" (Gifford). Adleman spent a year learning the necessary laboratory techniques of molecular biology.

Glanz, James, Mathematical logic flushes out the bugs in chip designs, *Science* 267 (20 January 1995) 332–333. Cipra, Barry, How number theory got the best of the Pentium chip, *Science* 267 (13 January 1995) 175.

Could the Pentium bug have been found in advance if Intel had not just simulated its performance but used formal methods to try to prove that its logic was correct? According to Glanz, maybe, but maybe not. Formal verification has now come to the forefront, in terms of both theorem-proving (for command-level operations, like division) and model-checking (e.g., using temporal logic to verify chip behaviors). Says one observer, "It was simply a matter of time before one of these consciousness-raising bugs hit. This won't be the last." Cipra's article gives the mathematical details of what Thomas Nicely (Lynchburg College) was trying to do when he discovered the Pentium error. Viggo Bruns proved in 1919 that the sum of the reciprocals of twin primes converges, to a sum now estimated to be 1.90216054, based on the sum for all twin primes up to 100 billion. Nicely was trying to refine the estimate by considering all twin primes into the trillions, and the Pentium gave reciprocals for 824,633,702,441 and 824,633,702,443 that were incorrect from the tenth significant digit on.

55th ANNUAL WILLIAM LOWELL PUTNAM MATHEMATICAL COMPETITION

These solutions have been compiled and prepared by Loren Larson, St. Olaf College.

A-1. Suppose that a sequence $a_1, a_2, a_3, \ldots$ satisfies $0 < a_n \le a_{2n} + a_{2n+1}$ for all $n \ge 1$. Prove that the series $\sum_{n=1}^{\infty} a_n$ diverges.

Solution. Assume the series converges. Then

$$\sum_{n=1}^{\infty} a_n \le \sum_{n=1}^{\infty} (a_{2n}+a_{2n+1}) = \sum_{n=2}^{\infty} a_n < \sum_{n=1}^{\infty} a_n,$$

a contradiction.

A-2. Let A be the area of the region in the first quadrant bounded by the line $y = \frac{1}{2}x$, the x-axis, and the ellipse $\frac{1}{9}x^2 + y^2 = 1$. Find the positive number m such that A is equal to the area of the region in the first quadrant bounded by the line $y = mx$, the y-axis, and the ellipse $\frac{1}{9}x^2 + y^2 = 1$.

Solution 1. The linear transformation given by $x_1 = \frac{1}{3}x, y_1 = y$ transforms the region R bounded by $y = \frac{1}{2}x$, the x-axis, and the ellipse $\frac{1}{9}x^2 + y^2 = 1$ into the region R' bounded by $y_1 = \frac{3}{2}x_1$, the x_1-axis, and the circle $x_1^2 + y_1^2 = 1$; it also transforms the region S bounded by $y = mx$, the y-axis, and $\frac{1}{9}x^2 + y^2 = 1$ into the region S' bounded by $y_1 = 3mx_1$, the y_1-axis, and the circle. Since all areas are multiplied by the same (nonzero) factor under the transformation, R and S have the same area if and only if R' and S' have the same area. However, we can see by symmetry about the line $y_1 = x_1$ that this happens if and only if $3m = \frac{2}{3}$, that is, $m = \frac{2}{9}$.

Solution 2. (Sketch of a calculus solution.) Setting up integrals for the areas, together with routine techniques of integration, yields

$$\int_0^{3/\sqrt{13}} \left(3\sqrt{1-y^2} - 2y\right) dy$$
$$= \frac{3}{2}\arcsin y + \frac{3}{2}y\sqrt{1-y^2} - y^2\Big]_0^{3/\sqrt{13}}$$
$$= \frac{3}{2}\arcsin\frac{3}{\sqrt{13}},$$

and

$$\int_0^{3/\sqrt{1+9m^2}} \left(\sqrt{1-\frac{x^2}{9}} - mx\right) dx =$$
$$\frac{3}{2}\arcsin\frac{x}{3} + \frac{1}{6}x\sqrt{9-x^2} - \frac{1}{2}mx^2\Big]_0^{3/\sqrt{1+9m^2}}$$
$$= \frac{3}{2}\arcsin\frac{1}{\sqrt{1+9m^2}}.$$

We want $\sqrt{1+9m^2} = \sqrt{13}/3$, or $m = 2/9$.

Variation. Without doing any actual integration, we can see that the first integral above transforms into the second one under the substitution $y = cx$, provided

$$\frac{3}{\sqrt{13}} = c \cdot \frac{3}{\sqrt{1+m^2}}, \quad 3c = 1,\ 2c^2 = m.$$

All these conditions are satisfied for $c = \frac{1}{3}$ and $m = \frac{2}{9}$.

A-3. Show that if the points of an isosceles right triangle of side length 1 are each colored with one of four colors, then there must be two points of the same color which are at least a distance $2 - \sqrt{2}$ apart.

Solution. Suppose the vertices of the isosceles right triangle are $(0,0), (1,0), (0,1)$. Suppose the points of the triangle can be colored in four colors such that points of the same color are always less that a distance $2 - \sqrt{2}$ apart. Then the four points $(0,1)$, $(0, \sqrt{2}-1)$, $(\sqrt{2}-1, 0)$, $(1,0)$ must have different colors, say colors A, B, C, D respectively. The point $(0,0)$ must be of color B or C. Without loss of generality, say $(0,0)$ is of color B. Then the point $(\sqrt{2}-1, 2-\sqrt{2})$

is of distance at least $2-\sqrt{2}$ to points of each of the four colors, and this is impossible.

A-4. Let A and B be 2×2 matrices with integer entries such that $A, A+B, A+2B$, $A+3B$, and $A+4B$ are all innvertible matrices whose inverses have integer entries. Show that $A+5B$ is invertible and that its inverse has integer entries.

Solution. Note that a matrix C with integer entries has an inverse with integer entries if and only if $\det C = \pm 1$. ("Only if": If $D = C^{-1}$, then $\det D \cdot \det C = 1$, and $\det D$, $\det C$ are integers, so $\det C = \pm 1$. "If" follows from the explicit form of the inverse of a matrix.) Therefore, if we consider the function f defined by $f(x) = \det(A + xB)$, we know that the five values $f(0), f(1), f(2), f(3)$, and $f(4)$ must all be 1 or -1, so f takes on at least one of those values three or more times. However, $f(x)$ is a polynomial of degree ≤ 2 in x, and so f can only take on a value more than twice if f is constant. Thus $f(x)$ is one of the constants 1 and -1; in particular, $\det(A + 5B) = \pm 1$, so $A + 5B$ has an inverse with integer entries.

A-5. Let $(r_n)_{n \geq 0}$ be a sequence of positive real numbers such that $\lim_{n \to \infty} r_n = 0$. Let S be the set of numbers representable as a sum

$$r_{i_1} + r_{i_2} + \cdots + r_{i_{1994}},$$

with $i_1 < i_2 < \cdots < i_{1994}$. Show that every nonempty interval (a, b) contains a nonempty subinterval (c, d) that does not intersect S.

Solution. We show, for any k, that if S_k is the set of real numbers representable as a sum

$$r_{i_1} + r_{i_2} + \cdot + r_{i_k}$$

then any nonempty open interval has a nonempty subinterval that does not intersect S_k. This is clearly true for $k = 0$. Suppose it holds for $k = m - 1$.

Let (a, b) be a nonempty interval. We need to show that some nonempty subinterval does not intersect S_m. By the induction hypothesis, we may assume (a, b) does not intersect S_{m-1}. Initially, let c be the midpoint of (a, b) and let $d = b$. Let $\varepsilon = d - c = c - a$.

Choose N so that $r_n \leq \varepsilon$ for $n > N$. For $i = 0, 1, \ldots, N$, in succession, find (c', d') so that the nonempty subinterval $(c' - r_i, d' - r_i)$ of $(c - r_i, d - r_i)$ does not intersect S_{m-1}, replacing (c, d) by (c', d') at each stage. As a result, the final (c, d) does not intersect $S_{m-1} + r_i$ for $0 \leq i \leq N$. On the other hand, for $i > N$, $S_{m-1} + r_i$ cannot intersect (c, d) since $(c - r_i, d - r_i)$ is contained in (a, b). It follows that (c, d) does not intersect S_m.

A-6. Let $f_1, f_2, \ldots, f_{10}$ be bijections of the set of integers such that for each integer n, there is some composition $f_{i_1} \circ f_{i_2} \circ \cdots \circ f_{i_m}$ of these functions (allowing repetitions) which maps 0 to n. Consider the set of 1024 functions

$$\mathcal{F} = \{f_1^{e_1} \circ f_2^{e_2} \circ \cdots \circ f_{10}^{e_{10}}\},$$

$e_i = 0$ or 1 for $1 \leq i \leq 10$. (f_i^0 is the identity function and $f_i^1 = f_i$.) Show that if A is any nonempty finite set of integers, then at most 512 of the functions in $\mathcal{F}$ map A to itself.

Solution. Let A be a nonempty finite subset of the integers $\mathbf{Z}$. By the Pigeonhole Principle, any bijection of $\mathbf{Z}$ which maps A to itself must be a bijection when restricted to A; in particular, its inverse also maps A to itself. Note that not all the bijections $f_1, f_2, \ldots, f_{10}$ can map A to itself, for otherwise if $0 \in A$ we could not map 0 to any $n \notin A$ by a composition $f_{i_1} \circ f_{i_2} \circ \cdots \circ f_{i_m}$, while if $0 \notin A$, we could not map 0 to any $n \in A$ by such a composition.

Let k be the smallest integer such that f_k does *not* map A to itself, and suppose that more than 512 of the functions $\mathcal{F}$ map A to itself. We can write $\mathcal{F}$ as a disjoint union of unordered pairs of functions such that two compositions $f_1^{e_1} \circ f_2^{e_2} \circ \cdots \circ f_{10}^{e_{10}}$ and $f_1^{d_1} \circ f_2^{d_2} \circ \cdots \circ f_{10}^{d_{10}}$ are in the same pair when they differ only in the k-th exponent; that is, when $e_i = d_i$ for $i \neq k$. By the Pigeonhole

Principle, there is then at least one on these 512 pairs in which both functions map A to itself. Since all f_l with $l > k$ also map A to itself, we can use composition with the inverses of f_l, as needed, to conclude that for some $e_1, \ldots, e_{k-1}$, $F_1 = f_1^{e_1} \circ f_2^{e_2} \circ \cdots \circ f_{k-1}^{e_{k-1}}$ and $F_2 = f_1^{e_1} \circ f_2^{e_2} \circ \cdots \circ f_{k-1}^{e_{k-1}} \circ f_k$ both map A to itself. But then $F_1^{-1} \circ F_2 = f_k$ also maps A to itself, a contradiction.

B-1. Find all positive integers that are within 250 of exactly 15 perfect squares.

Solution 1. Answer: $\{N \mid 315 \leq N \leq 325$ or $332 \leq N \leq 350\}$.

Assume $N > 0$ is with 250 of the 15 squares $m^2, (m+1)^2, \ldots, (m+14)^2$, where we can take $m \geq 0$. In fact, m will then be positive, otherwise N would be within 250 of the additional square 225. We have the necessary and sufficient conditions

$$(m+14)^2 \leq N + 250 \leq (m+15)^2 - 1,$$
$$(m-1)^2 + 1 \leq N - 250 \leq m^2.$$

Subtracting (reversing inequalities in the second line), we get

$$28m + 196 \leq 500 \leq 32m + 222,$$

which implies $m = 9$ or 10.

If $m = 9$,

$$23^2 \leq N + 250 \leq 24^2 - 1,$$
$$8^2 + 1 \leq N - 250 \leq 9^2,$$

or $315 \leq N \leq 325$.

If $m = 10$,

$$24^2 \leq N + 250 \leq 25^2 - 1,$$
$$9^2 + 1 \leq N - 250 \leq 10^2,$$

or $332 \leq N \leq 350$.

Solution 2 (This solution was provided by *Daniel J. Bernstein*, whose solutions to all the problems appear each year on the **USENET** newsgroup **sci.math.**)

Let $S(n)$ be the set of perfect squares s such that n is within 250 of s. For example $S(1) = \{0, 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225\}$.

Let $f(n)$ be the size of $S(n)$; e.g., $f(1) = 16$. The difference between $S(n-1)$ and $S(n)$ is that we lose $n - 251$ if $n - 251$ is a square, i.e., at $n \in \{251, 252, 255, 260, 267, 276, 287, 300, 315, 332, 351, \ldots\}$ and we gain $n + 250$ if $n + 250$ is a square, i.e., at $n \in \{6, 39, 74, 111, 150, 191, 234, 279, 326, 375, \ldots\}$. If neither $n - 251$ nor $n + 250$ is a square then $S(n) = S(n-1)$ and $f(n) = f(n-1)$.

We thus calculate successively

$$f(6) = 17,\ f(39) = 18,\ f(74) = 19,$$
$$f(111) = 20,\ f(150) = 21,\ f(191) = 22,$$
$$f(234) = 23,\ f(251) = 22,\ f(252) = 21,$$
$$f(255) = 20,\ f(260) = 19,\ f(267) = 18,$$
$$f(276) = 17,\ f(279) = 18,\ f(287) = 17,$$
$$f(300) = 16,\ f(315) = 15,\ f(326) = 16,$$
$$f(332) = 15, \text{and } f(251) = 14.$$

Hence the integers from 315 through 325 and the integers from 332 through 350 are the only integers up through 351 which are within 250 of exactly 15 perfect squares.

There are no such integers past 350. For, if n is within 250 of m^2, $(m+1)^2, \ldots, (m+14)^2$ then $(m+14)^2 - m^2 \leq 500$ so $28m + 196 \leq 500$, $m \leq 304/28 < 11$, $m \leq 10$. Hence $n \leq m^2 + 250 \leq 100 + 250 = 350$.

B-2. For which real numbers c is there a straight line that intersects the curve

$$y = x^4 + 9x^3 + cx^2 + 9x + 4$$

in four distinct points?

Solution. Answer: For the real numbers c with $c < 243/8$.

Let $y = q(x) = \alpha x + \beta$ denote a line, and set $p(x) = x^4 + 9x^3 + cx^2 + 9x + 4$. The intersections of the line and the curve correspond to solutions of $p(x) = q(x)$. If the line intersects the curve at four distinct points, $r(x) = p(x) - q(x) = 0$ has four distinct real roots. For a fixed α, $r(x)$ has four distinct

real roots for some choice of β if and only if $r(x)$ has three critical points. Similarly, $r'(x)$ has three distinct real roots for some choice of α if and only if it has two distinct critical points. Finally, $r''(x) = 12x^2 + 54x + 2c$ has two distinct real roots if and only if $c < 243/8$.

B-3. Find the set of all real numbers k with the following property:

> For any positive, differentiable function f that satisfies $f'(x) > f(x)$ for all x, there is some number N such that $f(x) > e^{kx}$ for all $x > N$.

Solution. The desired set is $(-\infty, 1)$.

To show this, first note that if $k > 1$ were in the set, then $k = 1$ would also be in the set. However, if f is any function of the form $f(x) = g(x)e^x$, where g is a positive, increasing, differentiable function bounded by 0 and 1 (for example, $g(x) = \frac{1}{\pi}\arctan x + \frac{1}{2}$), we have $f'(x) = e^x(g'(x) + g(x)) > f(x)$ and $f(x) < e^x$ for all x, so $k = 1$ is not in the set.

On the other hand, if $f'(x) > f(x)$ for all x, then (since f is positive) we have

$$\frac{f'(x)}{f(x)} > 1 \quad \text{for all } x,$$

$$\int_0^x \frac{f'(t)}{f(t)}dt > \int_0^x 1\,dt \quad \text{for all } x \geq 0,$$

$$\log(f(x)) > x + \log(f(0)) \quad \text{for all } x \geq 0,$$

$$f(x) > f(0)e^x \quad \text{for all } x \geq 0.$$

If k is any number less than 1, then for large enough x we will have $f(0)e^x > e^{kx}$ (since $f(0)$ is positive), which shows that k is in the set.

B-4. For $n \geq 1$, let d_n be the greatest common divisor of the entries of $A^n - I$, where $A = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}$ and $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Show that $\lim_{n\to\infty} d_n = \infty$.

Solution 1. From experimentation (and then an easy induction on n) we see that A^n has the form

$$A^n = \begin{pmatrix} a_n & b_n \\ 2b_n & a_n \end{pmatrix}$$

with a_n odd, and, since $\det A^n = 1$, we have $a_n^2 - 1 = 2b_n^2$. Thus $a_n - 1$ divides $2b_n^2$, so that $d_n = \gcd(a_n - 1, b_n) \geq \sqrt{(a_n - 1)/2}$. Since $\lim_{n\to\infty} a_n = \infty$ (e.g., $a_n > 3a_{n-1}$), the result follows.

Solution 2. (This solution, due to *Robin J. Chapman*, appeared among those prepared by *Daniel J. Bernstein* (see **B-1**).) Observe that the matrices of the form $\begin{pmatrix} a & b \\ 2b & a \end{pmatrix}$ are closed under multiplication by $\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$. In particular, $\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}^n$ must be of the form $\begin{pmatrix} a & b \\ 2b & a \end{pmatrix}$. Its determinant is $(-1)^n$, so $a^2 - 2b^2 = (-1)^n$. But $\begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}^2$, so

$$\begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}^n - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b \\ 2b & a \end{pmatrix}^2 - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a^2 + 2b^2 - 1 & 2ab \\ 4ab & a^2 + 2b^2 - 1 \end{pmatrix}.$$

If n is odd then $a^2 - 2b^2 = -1$, so $a^2 + 2b^2 - 1 = 2a^2$ and all entries are divisible by a. If n is even then $a^2 - 2b^2 = 1$, so $a^2 + 2b^2 - 1 = 4b^2$ and all entries are divisible by b. Both a and b increase as $n \to \infty$.

B-5. For any real number α, define the function $f_\alpha(x) = \lfloor \alpha x \rfloor$. Let n be a positive integer. Show that there exists an α such that for $1 \leq k \leq n$,

$$f_\alpha^k(n^2) = n^2 - k = f_{\alpha^k}(n^2).$$

Solution. We first show that it suffices for α to satisfy

$$\frac{n^2 - 1}{n^2} < \alpha < \left(\frac{n^2 - n + 1}{n^2}\right)^{1/n}. \quad (1)$$

By the first inequality together with $\alpha < 1$, for any m, $1 \leq m \leq n^2$, we have $f_\alpha(m) = m - 1$, so $f_\alpha^k(n^2) = n^2 - k$ for $1 \leq k \leq n^2$. By the second inequality, $f_{\alpha^n}(n^2) < n^2 - n + 1$.

Since $f_\alpha(f_\beta(m)) \le f_{\alpha\beta}(m)$ for positive β, m, we know $n^2-k = f_\alpha^k(n^2) \le f_{\alpha^k}(n^2)$ for $k \ge 1$. Thus, $f_{\alpha^n}(n^2) = n^2 - n$ and, by a reverse induction on k, $f_{\alpha^k}(n^2) = n^2 - k$ for $1 \le k \le n$ (since $f_{\alpha^k}(n^2) \ge n^2 - k + 1$ would imply $f_{\alpha^{k+1}}(n^2) \ge f_\alpha(f_{\alpha^k}(n^2)) \ge n^2 - k$).

To show that α exists satisfying (1), it suffices to show, for $0 < x < 1$, that $1 - x^2 < (1-x+x^2)^x$. Multiplying through by $(1+x)^x$, this is equivalent to showing

$$(1 - x^2)(1 + x)^x < (1 + x^3)^x.$$

But the right-hand side is greater than 1, while the left-hand side is less than $(e^{-x^2})(e^x)^x = 1$.

B-6. For any integer a, set

$$n_a = 101a - 100 \cdot 2^a.$$

Show that for $0 \le a, b, c, d \le 99$, $n_a + n_b \equiv n_c + n_d \pmod{10100}$ implies $\{a, b\} = \{c, d\}$.

Solution. Observe that $n_a \equiv a \pmod{100}$ and $n_a \equiv 2^a \pmod{101}$.

Suppose $n_a + n_b \equiv n_c + n_d \pmod{10100}$. Then $n_a + n_b \equiv n_c + n_d \pmod{101}$, so

$$2^a + 2^b \equiv 2^c + 2^d \pmod{101}. \quad (1)$$

Also, $n_a + n_b \equiv n_c + n_d \pmod{100}$, so $a + b \equiv c + d \pmod{100}$, and therefore, by Fermat's Theorem (since 101 is prime), $2^{a+b} \equiv 2^{c+d} \pmod{101}$. That is,

$$2^a \cdot 2^b \equiv 2^c \cdot 2^d \pmod{101}. \quad (2)$$

From (1) and (2), we see that $\{2^a, 2^b\}$ and $\{2^c, 2^d\}$ are the same set modulo 101, namely, the set of roots of the quadratic polynomial $(x - 2^a)(x - 2^b) = x^2 - (2^a + 2^b)x + 2^a 2^b = (x - 2^c)(x - 2^d)$ in the field $\mathbf{Z}_{101}$. To see that $\{a, b\} = \{c, d\}$, it suffices to show that the numbers 2^a for $a \in \{0, 1, \ldots, 99\}$ are distinct modulo 101. That is, we need to show that the order of 2 modulo 101 is precisely 100. For this, it suffices to show that $2^{20} \not\equiv 1 \pmod{101}$ and $2^{50} \not\equiv 1 \pmod{101}$. We have $2^{10} = 1024 \equiv 14 \pmod{101}$, so that $2^{20} \equiv 14^2 \equiv -6 \pmod{101}$, from which $2^{50} \equiv 2^{20}2^{20}2^{10} \equiv 36 \cdot 14 \equiv -1 \pmod{101}$.

LETTERS TO THE EDITOR

Dear Editor:

Anne M. Burns has perhaps not wandered our rock-strewn, treeless tundra in her pursuit of arts and mathematics, but she seems to have been inspired by this harsh, remote setting ("Convolutions and Computer Graphics," October 1994, pp. 258-67). Her "jagged lines of random lengths ... topped with disks of random radii" (FIGURE 7) are immediately identifiable as Arctic cotton grass (*Eriphorum callitrix*), whose fluffy heads briefly adorn damp spots on local tundra before early September snows.

Dr. Harold Don Allen
Eastern Arctic Teacher
Education Program
Iqaluit, Baffin Island
Northwest Territories,
Canada X0A 0H0

Editor's Note. For those wishing to contact Bernard Beauzamy (co-author of "Quantitative Estimates for Polynomials in One or Several Variables," October 1994, pp. 243-257), you may reach him at the following address:

Institut de Calcul Mathématique
37, rue Tournefort
75005, Paris France

"What *DERIVE*® says, you can believe.

"Ask it anything. Any number of decimal places, no problem. On a 486, on a Pentium™, no problem.

"As far as I know, *DERIVE* is the only off-the-shelf math utility that you can run (on any system that can put up a DOS window) and use to check any other calculation on the screen, Pentium or no Pentium."

—*Peter Coffee, PC Week, January 1995*

"A fabulous piece of software!

"*DERIVE* is the best buy among the symbolic mathematics software packages. It has less stringent hardware requirements, a superior user interface, better performance, and a lower price than its competitors.

"*DERIVE* also appears to consist of more efficient, better debugged code than its competitors. . . It is even immune to the Pentium flaw!"

—*Dr. Thomas R. Nicely, Professor of Mathematics, Lynchburg College, Virginia, who tracked down the flaw in the Pentium chip with the help of DERIVE.*

NEW VERSION 3

DERIVE®

A Mathematical Assistant

Soft Warehouse Inc
HONOLULU • HAWAII

Soft Warehouse, Inc. • 3660 Waialae Avenue
Suite 304 • Honolulu, Hawaii, USA 96816-3259
Telephone: (808) 734-5801 after 10:00 a.m. PST
Fax: (808) 735-1105 • Email: swh@aloha.com.

CONTENTS

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