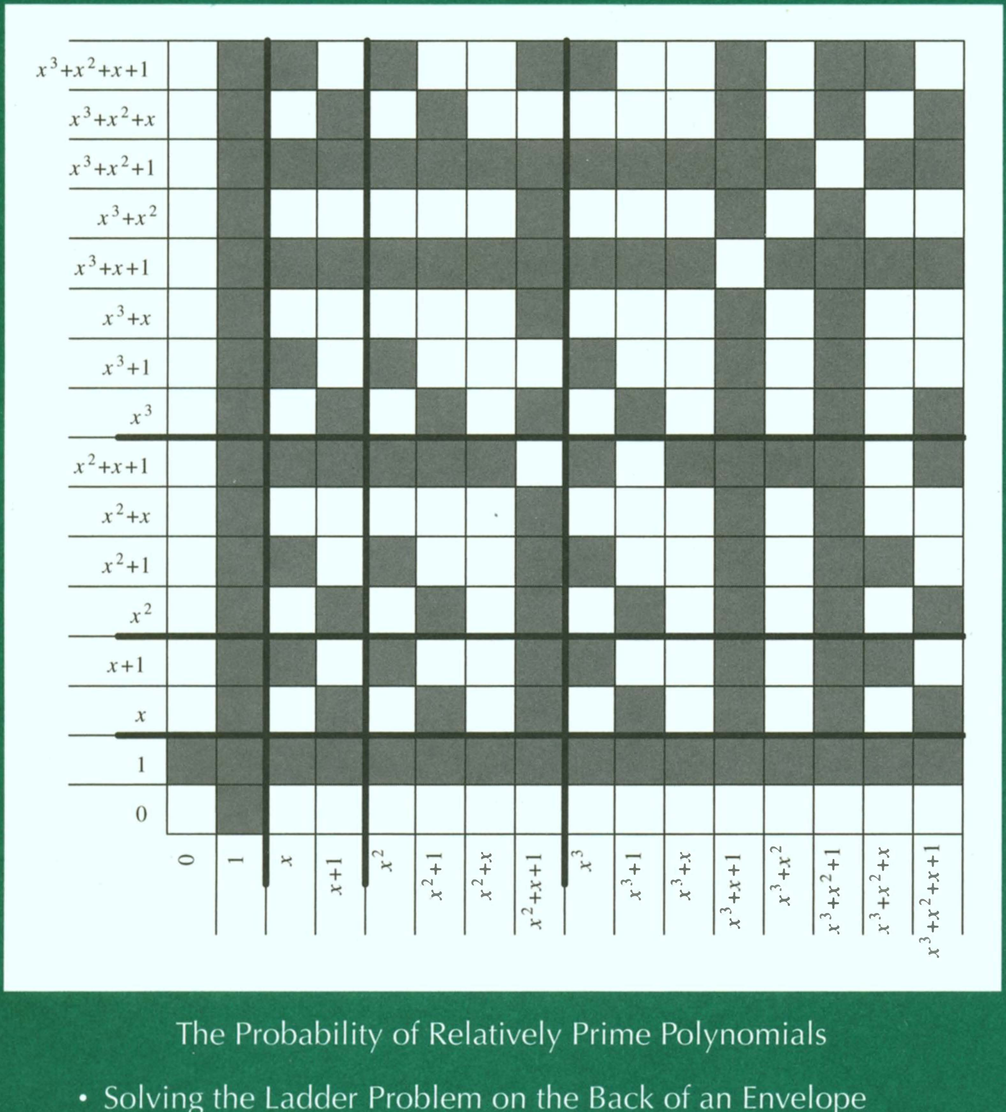
Vol. 80, No. 3, June 2007

MATHEMATICS MAGAZINE



x^3+x^2+x+1		▨	▨		▨			▨	▨			▨		▨	▨	
x^3+x^2+x		▨		▨		▨						▨		▨		▨
x^3+x^2+1		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		▨	▨
x^3+x^2		▨						▨				▨		▨		
x^3+x+1		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		▨	▨	▨	▨
x^3+x		▨						▨				▨		▨		
x^3+1		▨	▨		▨				▨			▨		▨		
x^3		▨		▨		▨		▨		▨		▨		▨		▨
x^2+x+1		▨	▨	▨	▨	▨	▨		▨		▨	▨	▨	▨		▨
x^2+x		▨						▨				▨		▨		
x^2+1		▨	▨		▨			▨	▨			▨		▨	▨	
x^2		▨		▨		▨		▨		▨		▨		▨		▨
$x+1$		▨	▨		▨			▨	▨			▨		▨	▨	
x		▨		▨		▨		▨		▨		▨		▨		▨
1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
0		▨														
	0	1	x	$x+1$	x^2	x^2+1	x^2+x	x^2+x+1	x^3	x^3+1	x^3+x	x^3+x+1	x^3+x^2	x^3+x^2+1	x^3+x^2+x	x^3+x^2+x+1

The Probability of Relatively Prime Polynomials

- Solving the Ladder Problem on the Back of an Envelope
- The Recreational Gambler: Paying the Price for More Time at the Table

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Submit new manuscripts to Allen Schwenk, Editor, Mathematics Magazine, Department of Mathematics, Western Michigan University, Kalamazoo, MI, 49008. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

The **cover image** illustrates that nearly exactly half of the polynomial pairs formed over the finite field $\mathbb{Z}_2[x]$ are relatively prime, but can you find a natural one-to-one correspondence? See the article by Art Benjamin and Curtis Bennett on page 196 for the surprising solution.

AUTHORS

Dan Kalman received his Ph.D. from the University of Wisconsin in 1980 and has been at American University since 1993. Prior to that he had academic appointments (University of Wisconsin, Green Bay; Augustana College, Sioux Falls) and worked for eight years in the aerospace industry in Southern California. Kalman is a past Associate Executive Director of the MAA, author of a book published by the MAA, and frequent contributor to MAA journals. He delights in puns and word play of all kinds, and is an avid fan of Douglas Adams, J. R. R. Tolkien, and Gilbert and Sullivan.

Arthur Benjamin is Professor of Mathematics at Harvey Mudd College, co-editor of Math Horizons, and the MAA Polya Lecturer from 2006 to 2008. His book, Proofs That Really Count: The Art of Combinatorial Proof, received the Beckenbach Book prize in 2006. In 2005, Reader's Digest proclaimed him to be "America's Best Math Whiz." He is a past Associate Editor for Mathematics Magazine, and maintains its searchable database.

Curtis Bennett is a Professor of Mathematics at Loyola Marymount University and Secretary of the Southern California-Nevada Section of the MAA. He works in the areas of groups and geometries, combinatorics, and the scholarship of teaching and learning. He was a CASTL fellow with the Carnegie Foundation for the Advancement of Teaching in 2000 and in 2003.

Joseph Bak has been teaching at City College of New York since 1971. His primary areas of research are approximation theory and complex analysis. His relatively recent interest in probability in general and gambling, in particular, began with a course on probability which he taught at City College. It has been maintained by the feedback from the first article he wrote on the subject, which appeared in Mathematics Magazine in 2001.

Vol. 80, No. 3, June 2007

MATHEMATICS MAGAZINE

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 *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

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

MAA Advertising
c/o Marketing General, Inc.
209 Madison Street Suite 300
Alexandria VA 22201

Phone: 866-821-1221
Fax: 866-821-1221
E-mail: rhall@marketinggeneral.com

Further advertising information can be found online at www.maa.org

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

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

Printed in the United States of America

ARTICLES

Solving the Ladder Problem on the Back of an Envelope

DAN KALMAN
American University
Washington, D.C. 20016
kalman@american.edu

How long a ladder can you carry horizontally around a corner? Or, in the idealized geometry of FIGURE 1, how long a line segment can be maneuvered around the corner in the L-shaped region shown? This familiar problem, which dates to at least 1917, can be found in the max/min sections of many calculus texts and is the subject of numerous web sites. The standard solution begins with a twist, transforming the problem from maximization to minimization. This bit of misdirection no doubt contributes to the appeal of the problem. But it fairly compels the question, *Is there a direct approach?*

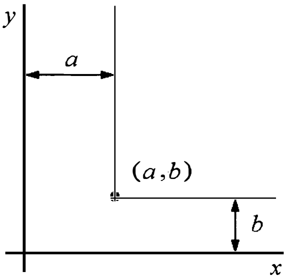


Figure 1 Geometry of ladder problem

In fact there is a beautifully simple direct approach that immediately gives new insights about the problem. It also gives us an excuse to revisit a lovely topic—envelopes of families of curves. This topic was once a standard part of the calculus curriculum, but seems to be largely forgotten in the current generation of texts. A generalized ladder problem considered in [**17**] can also be analyzed using the direct approach.

Ladder problem history

It is not easy to discover when the ladder problem first appeared in calculus texts. Singmaster [**24**] has compiled an extensive chronology of problems in recreational mathematics. There, the earliest appearance of the ladder problem is a 1917 book by Licks [**15**]. As Singmaster notes, this version of the problem concerns a stick to be put up a vertical shaft in a ceiling, rather than a ladder and two hallways, but the two situations are mathematically equivalent. Licks gives what is today the standard solution, finding the maximum length stick that gets stuck in terms of the angle the stick makes with the floor. He concludes *This is a simple way to solve a problem which has proved a stumbling block to many.* Whether this implies an earlier provenance in recreational problem solving, or a more mundane history of people actually putting long sticks up vertical shafts, who can say?

American University is fortunate to possess an extensive collection of mathematical textbooks dating to the 18th century. Haphazardly selecting a sample of eight calculus textbooks published between 1816 and 1902, I searched without success for mention of the ladder problem, or the equivalent, in discussions of maxima and minima. Many of these texts did have quite a number of max/min exercises, including several that our students would recognize. Of particular note is the text by Echols [**7**], published in 1902. Among the 56 max/min exercises in this work, nearly all of today's standard exercises appear, but not the ladder problem. More than half of the books also have a section on envelopes. Coincidentally, in the 1862 work of Haddon [**13**], the envelope we will discuss below appears in an example about a ladder sliding down a wall, but not in connection with any max/min problem.

In modern times, a variation on the ladder problem has been the subject of ongoing research. This is the sofa problem, which seeks the region of greatest area that can go around a corner between halls of given widths. The sofa problem appears in a volume on unsolved problems in geometry [**6**], and gave birth to the *Moving a Sofa Constant* [**9**]. For more on this open problem, see [**27**].

The main focus here is the use of envelopes to solve the ladder problem. The earliest record I have found for this approach is an anecdote of Cooper [**3**], who reports meeting a variant of the ladder problem in 1959 on a physics quiz at Princeton and solving it via envelopes. There is also one reference from Singmaster's compilation in which envelopes are used. Fletcher [**10**] provides five solutions of the ladder problem, with no explicit use of calculus. One of these methods uses envelopes. However, in order to avoid calculus, Fletcher depends on geometric properties of envelopes that are obscure by present day standards. Moreover, the exposition is rather terse, and says almost nothing about what the envelope is, how it arises, or why it provides a solution to the ladder problem.

The direct approach

As mentioned earlier, the standard solution to the ladder problem begins with a restatement: the goal is shifted from finding the longest ladder that will go around the corner to finding the *shortest* ladder that will get stuck. In seeking a direct approach, we consider actually moving a segment around a corner, trying to use as little of the space as possible. Begin with the segment along one of the outer walls, say with the left end at the origin and the right end at the point $(L, 0)$. Slide the left end up the y axis, all the while keeping the right end on the x axis. Intuitively, this maneuver keeps the line segment as far as possible from the corner point (a, b). Surely, if a segment of length L cannot get around the corner using this conservative approach, then it won't go around no matter what we do.

Now as you slide the segment along the walls it sweeps out a region Ω, as illustrated in FIGURE 2. The outer boundary of Ω is part of curve called an *astroid*, with equation

$$x^{2/3} + y^{2/3} = L^{2/3}. \qquad (1)$$

The full astroid is in the shape of a four pointed star (hence the name), but for the ladder problem, we are concerned only with the part of the curve that lies in the first quadrant. Our line segment will successfully turn the corner just when Ω stays within the hallways. And that is true as long as the corner point (a, b) is outside Ω. The extreme case occurs when (a, b) lies on the boundary curve, whereupon

$$a^{2/3} + b^{2/3} = L^{2/3}.$$

This shows that the longest segment that can go around the corner has length $L = (a^{2/3} + b^{2/3})^{3/2}$.

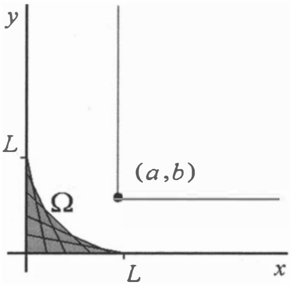


Figure 2 Swept out region Ω

Of course, this solution depends on knowing the boundary curve for Ω. Once you know that, the problem becomes transparent. We can easily visualize the region for a short segment that will go around the corner, and just as easily see what happens if we increase the length of the segment. In contrast with the usual approach to this problem, we are led to a direct understanding of the maximization process. And in the context of the equation for the astroid, we understand *why* the formula for the extreme value of L takes the form that it does.

In fact, in this direct approach, the optimization part of the problem becomes trivial. It is akin to asking "What is the longest segment that can be contained within the unit interval?" This is nominally a max/min problem, but no analysis is needed to solve it. In the same way, the direct approach to the ladder problem renders the solution immediately transparent, once you have found the boundary curve for Ω. But it is not quite fair to claim that this approach eliminates the need for calculus. Rather, the point of application of the calculus is shifted from the optimization question to that of finding the boundary curve.

Envelopes of families of curves

So how is the boundary curve found? The key is to observe that it is the *envelope* of a family of curves. For the current case, notice that each successive position of the line segment can be identified with a linear equation. Let the angle between the segment and the positive x axis be α, as shown in FIGURE 3. Then the x and y intercepts of the line segment are $L\cos\alpha$ and $L\sin\alpha$, respectively, so the line is defined by the equation

$$\frac{x}{\cos\alpha} + \frac{y}{\sin\alpha} = L. \tag{2}$$

This equation defines a family of lines in terms of the parameter α.

The region Ω is the union of all the lines in the family. To be more precise, we restrict α to the interval $[0, \pi/2]$, and intersect all the lines with the first quadrant.

Visually, it seems apparent that the boundary curve is tangent at each point to one of the lines. This observation, which will be proved presently, shows that the boundary curve is an envelope for the family of lines.

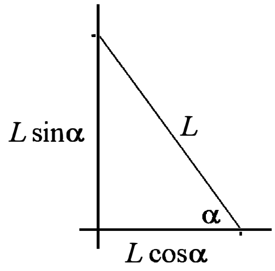


Figure 3 Parameter α

In general, an equation of the form

$$F(x, y, \alpha) = 0 \tag{3}$$

defines a family of plane curves with parameter α if for each value of α the equation defines a plane curve C_α in x and y. An *envelope* for such a family is a curve every point of which is a point of tangency with one of the curves in the family.

There is a standard method for determining the envelope curve: Differentiate (3) with respect to α, and then use the original equation to eliminate the parameter. Technically, (x, y) is a point of the envelope curve only if it satisfies both (3) and

$$\frac{\partial}{\partial\alpha} F(x, y, \alpha) = 0 \tag{4}$$

for some α. Combining equations (3) and (4) to eliminate α produces an equation in x and y. This will be referred to as the *envelope algorithm.*

Obviously, the envelope algorithm depends on certain assumptions about F, requiring at the very least differentiability with respect to α. Also, in the general case, the condition is necessary but not sufficient, so there may exist curves which satisfy (3) and (4), but which are not part of the envelope. For the moment, let us gloss over these issues, and move straight on to applying the algorithm for the ladder problem. A more careful discussion of the technicalities will follow.

The first step is to differentiate (2) with respect to α. That gives

$$\frac{x\sin\alpha}{\cos^2\alpha} - \frac{y\cos\alpha}{\sin^2\alpha} = 0$$

and after rearrangement we obtain

$$x\sin^3\alpha = y\cos^3\alpha. \tag{5}$$

By combining this equation with (2), we wish to eliminate the parameter α. With that in mind, rewrite (5) in the form

$$\tan\alpha = \frac{y^{1/3}}{x^{1/3}}.$$

This leads to

$$\cos\alpha = \frac{x^{1/3}}{\sqrt{x^{2/3}+y^{2/3}}} \quad \text{and} \quad \sin\alpha = \frac{y^{1/3}}{\sqrt{x^{2/3}+y^{2/3}}}.$$

Now we can substitute these expressions in (2), and so derive the following equation in x and y alone.

$$x^{2/3}\sqrt{x^{2/3}+y^{2/3}} + y^{2/3}\sqrt{x^{2/3}+y^{2/3}} = L$$

Simplifying, we have

$$(x^{2/3}+y^{2/3})\sqrt{x^{2/3}+y^{2/3}} = (x^{2/3}+y^{2/3})^{3/2} = L$$

and so we arrive at (1).

We can also derive a parameterization of the envelope. In the equations above for $\cos\alpha$ and $\sin\alpha$, replace $\sqrt{x^{2/3}+y^{2/3}}$ with $L^{1/3}$. Solving for x and y produces

$$\begin{aligned} x(\alpha) &= L\cos^3\alpha \\ y(\alpha) &= L\sin^3\alpha. \end{aligned} \tag{6}$$

In this parameterization, $(x(\alpha), y(\alpha))$ is the point of the envelope that lies on the line corresponding to parameter value α.

Pedagogy

The foregoing computation is an intriguing way to deduce the boundary curve for the region Ω. From that curve we can immediately find the solution to the ladder problem as discussed earlier. As elegant as this solution is, it may be inaccessible to today's calculus students. Interestingly, there is some evidence to suggest that the computation of envelopes via the method above was once a standard topic in calculus. This is certainly the impression left by [**5, 8, 12**], all of which date to the 1940's and 1950's. On the other hand, anecdotal reports by colleagues who were students and teachers of calculus during that time are inconsistent on this point.

In today's calculus texts (or more precisely, in their indices), one finds no mention of envelopes. The topic is covered in older treatments of calculus [**4, 14**] and advanced calculus [**25**] and the expositions in these sources tend to be very similar. My informal survey (as described above) suggests that the topic of envelopes was common in calculus texts throughout the 19th century. Was the topic common enough in the calculus curriculum in the first half of the twentieth century to be considered standard? If so, when and why did this topic fall out of favor? These are interesting historical questions.

If the topic of envelopes has been forgotten in calculus texts, it has not disappeared from the mathematical literature. Indeed, in expository publications like this MAGAZINE, one readily finds recent mention of envelopes and the envelope algorithm. See, for example, [**1, 11, 16, 20, 22, 23**]. There is also an application of envelopes in the field of economics, referred to as the *Envelope Theorem* [**18, 26**]. Nevertheless, I have a feeling that this topic is not as widely known among college mathematics faculty as it should be. Accordingly, a rather detailed discussion of envelopes is presented in the next section.

Outside of calculus courses, where might envelopes by found? The topic appears in works on properties of plane curves (see [**19, 28**]), another subject that seems to have been much more common in an earlier era. To a previous generation of mathematicians who were well acquainted with such terms as *involute*, *evolute*, and *caustic*, the boundary curve (1) would be familiar indeed. It is known not only as an astroid, but more generally as an instance of hypocycloid, the locus of a point on a circle rolling within a larger circle. We obtained it as the envelope of the family of lines (2), identified in **[28]** as the *Trammel ofArchimedes.* The same curve can also be obtained as the envelope of a family of ellipses, the sum of whose axes is equal to L [**28**, p. 2]. See FIGURE 4.

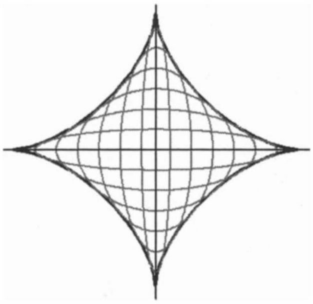


Figure 4 The astroid as envelope of a family of ellipses

The treatment of envelopes in **[28]** implies that this topic is properly a part of the study of differential equations. Perhaps it is in this context that envelopes once were considered a standard calculus topic, although that is certainly not the case in [**4, 14, 25**].

However the historical questions are answered, it is something of a shame that envelopes are not included in modern curricula, even for enrichment. The topic has obvious visual appeal, and the method is an attractive application of differentiation. In addition, the consideration of why and how the method works leads to interesting insights. And of course, if our students knew about envelopes, the solution of the ladder problem would be much simplified. Still, taking everything into consideration, this topic is probably too great a digression for most calculus classes. No doubt, embarking on such a digression just to reach an elegant solution to the ladder problem would be (dare I say it) *pushing the envelope.*

As a compromise, it might be reasonable to guide students through a construction of the boundary curve of Ω, without using the general method of envelopes. Here is one approach. Consider sweeping out the region Ω using a segment of length L. For each value of α, there will be one position of the line segment, given by (2). Now for a fixed value of x_0, consider the points $(x_0, y_0(\alpha)))$ that lie on the various line segments. Evidently, the maximum value $y_0(\alpha)$ defines the point of the boundary curve corresponding to x_0. From (2), we have

$$y_0(\alpha) = L\sin\alpha - x_0\tan\alpha$$

and the maximum value for $0 \leq \alpha \leq \pi/2$ is easily found to be

$$y_0 = (L^{2/3} - x_0^{2/3})^{3/2}.$$

In this way, the boundary curve is obtained. But to solve the ladder problem, we do not really need the entire boundary curve. All we need to know is where the point (a, b) lies relative to the boundary curve. So, in the preceding argument, simply take $x_0 = a$. This provides another approach to the ladder problem.

Technicalities

In deriving the envelope algorithm, one generally assumes that locally the envelope is a curve smoothly parameterized by α. By definition, each point P of the envelope is a point of tangency to some member of the family of curves, and each member of the family is tangent to the envelope at some point P. This suggests that P can be defined as a function of α. However, some caution is necessary. If a curve in the family touches the envelope in multiple points, there will be ambiguity in defining $P(\alpha)$. This is the situation when we generate the astroid as the envelope of a family of ellipses, as in FIGURE 4. In this case there are many functions $P(\alpha)$ that map the parameter domain to the envelope, not all of which are continuous. Of course in this example it is possible to choose $P(\alpha)$ consistently to obtain a smooth parameterization of the envelope. But it is not clear how this can be done in general. Accordingly, we assume the envelope has a smooth parameterization $P(\alpha) = (x(\alpha), y(\alpha))$ such that $P(\alpha)$ is the point of tangency between the envelope and the curve C_α.

With that assumption, observe that the equation $F(x(\alpha), y(\alpha), \alpha) = 0$ holds identically, so the derivative with respect to α is zero. Viewing F as a function of three variables, the chain rule gives

$$\frac{\partial F}{\partial x}\frac{dx}{d\alpha} + \frac{\partial F}{\partial y}\frac{dy}{d\alpha} + \frac{\partial F}{\partial \alpha} = 0. \tag{7}$$

But we can also view F as a function of two variables, thinking of α as a fixed parameter. In this view, the xy gradient of F is normal to the curve $F(x, y, \alpha) = 0$ at each point. Meanwhile, the parameterization of the envelope provides a tangent vector $(\frac{dx}{d\alpha}, \frac{dy}{d\alpha})$ at each point of that curve. At the point $(x(\alpha), y(\alpha))$, the two curves are tangent, so the normal vector $\nabla_{xy}F = (\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y})$ is orthogonal to the velocity vector $(\frac{dx}{d\alpha}, \frac{dy}{d\alpha})$. This shows that the first two terms on the left side of (7) add to 0, and hence $\frac{\partial F}{\partial \alpha} = 0$.

The preceding argument is the basis for the envelope algorithm. It shows that at each point (x, y) of the envelope there is a value of α for which both (3) and (4) hold. This is a necessary condition, and it can be satisfied by points which are not on the envelope. Indeed, we can construct an example of this phenomenon by reparameterizing the family of curves. The general idea behind this construction will be clear from the specific case of the family of lines for the ladder problem.

Let $s(\alpha)$ be any differentiable function from $(0, \pi/2)$ onto itself. Then the equation

$$\frac{x}{\cos s(\alpha)} + \frac{y}{\sin s(\alpha)} = L$$

parameterizes the same family of lines as (2), and so has the same envelope. Applying the envelope algorithm, we compute the partial derivative with respect to α, obtaining

$$\frac{x \sin s(\alpha)\, s'(\alpha)}{\cos^2 s(\alpha)} - \frac{y \cos s(\alpha)\, s'(\alpha)}{\sin^2 s(\alpha)} = 0.$$

Clearly, this equation will be satisfied for any value of α where $s'(\alpha) = 0$. If $\alpha*$ is such a point, then every point of the corresponding line

$$\frac{x}{\cos s(\alpha*)} + \frac{y}{\sin s(\alpha*)} = L$$

satisfies the two conditions of the envelope algorithm. That is, the entire line segment corresponding to $\alpha*$ will be produced by the envelope algorithm. No such line is actually included in the boundary of Ω, nor can any such line be tangent to all the lines in the family. This illustrates how the envelope algorithm can produce extraneous results.

A more complete discussion of these technical points can be found in [**4, 19**]. In particular, conditions that can give rise to extraneous results from the envelope algorithm are characterized. As Courant remarks, once the envelope algorithm produces a curve, "it is still necessary to make a futher investigation in each case, in order to discover whether it is really an envelope, or to what extent it fails to be one." In practice, graphing software can often give a clear picture of the envelope of a family of curves, and so guide our understanding of the results of the envelope algorithm.

A technical point of a slightly different nature concerns the relationship between the envelope of a family of curves, and the boundary of the region that family encompasses. By definition, the envelope is a curve which is tangent at each of its points to some member of the family. This is the definition used to justify the envelope algorithm. But the curve we are interested in for the ladder problem is defined as a boundary curve. How are these two concepts related? Visually, it appears obvious that at each point of the boundary of Ω, the tangent line is a member of the family of lines defining Ω. We substantiate this appearance as follows.

Since each line segment is contained within the region Ω, none of the lines can cross the boundary curve. On the other hand, each point of the boundary must lie on one of the lines. To see this, consider a boundary point P, and a sequence of points P_j in Ω converging to P. Each point P_j is on a line for some parameter value α_j, and these values all lie in the interval $[0, \pi/2]$. So there is a convergent subsequence α_{j_k} with limit $\alpha *$. Now by the continuity of (2), $P = \lim P_{j_k}$ is a point on the line with parameter $\alpha *$. Since this line cannot cross the boundary curve at P, it must be tangent there.

This suggests as a general principle that the boundary of a region swept out by a family of curves lies on the envelope for that family. As in the earlier discussion, some caution is necessary. Here, it is sufficient to assume that the boundary curve is smoothly parameterized by α, the parameter defining the family of curves. On any arc where this is true, the boundary curve will indeed fall along the envelope. On the other hand, consider the following:

$$F(x, y, \alpha) = x^2 + y^2 - \sin^2 \alpha.$$

This describes a family of circles centered at the origin, with radius varying smoothly between 0 and 1. The region swept out by the family of circles is the closed unit disk $x^2 + y^2 \leq 1$, and the unit circle is the boundary curve. But the unit circle is not an envelope for the family of circles. What went wrong? Arguing as above, we can again assign a value of α to each point of the boundary curve. But to do this continuously, we have to take α to be constant, say $\alpha = \pi/2$. Then the entire boundary is one of the curves in the family, but it is not parameterized by α. Notice as well that in this example, the curve for $\alpha = \pi/2$ also satisfies the equation $\frac{\partial F}{\partial \alpha} = 0$. Thus, while the envelope algorithm fails to produce the envelope in this example, it does locate the boundary of the region.

Relating the envelope to the boundary also gives a different insight about why the envelope technique works. View (3) as defining a level surface S of the function F in $xy\alpha$ space. The family of curves is then defined as the set of level curves for this surface. At the same time, the region Ω swept out by the family of curves is the projection of S on the xy plane. Now suppose A is a point on S that projects to a point P on the boundary of Ω. The tangent plane to S at A must be vertical and project to a line in the xy plane. Otherwise, there is an open neighborhood of A on the tangent plane that projects to an open neighborhood of P in the xy plane, and that puts P in the interior of Ω. So the tangent plane is vertical. That implies a horizontal normal vector to S at

A. But that means that the gradient of F, which is normal to the surface at each point, must be horizontal at A. This shows that the partial derivative of F with respect to α vanishes at A, which is the derivative condition of the envelope algorithm.

As a final topic in this section, I cannot resist mentioning one more way to think about the envelope of a family of curves. The idea is to consider a point on the envelope as the intersection point of two *neighboring* members of the family. Each α gives us one member of the family of curves, and the intersection of curves for two *successive* values of α gives a point on the envelope. Of course, that is not literally possible, but we can implement this idea using limits. Just express the intersection of the curves C_α and $C_{\alpha+h}$ as a function of h and α, and take the limit as h goes to 0. It is instructive to carry this procedure out for the example of the ladder problem. It once again yields the envelope as the astroid (1). In the process, one can observe differentiation with respect to α implicitly occuring in the calculation of the limit as h goes to 0. Indeed Courant [**4**] uses this idea to provide a heuristic derivation of the envelope algorithm, before developing a more rigorous justification. In contrast, Rutter [**19**] terms this the *limiting position* definition of *envelope*, one of three closely related but distinct definitions that he considers. He also provides the following interesting example of a family of circles with an envelope, but for which neighboring circles in the family are disjoint.

Begin with the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

and at each point, compute the osculating circle. This is the circle whose curvature matches that of the ellipse at the specified point, and whose center and radius are the center and radius of curvature of the ellipse. The family of osculating circles for all the points of the ellipse is the focus for this example. This situation is illustrated for an ellipse with $a = 8$ and $b = 4$ in FIGURE 5, which shows the ellipse and several members of the family of circles.

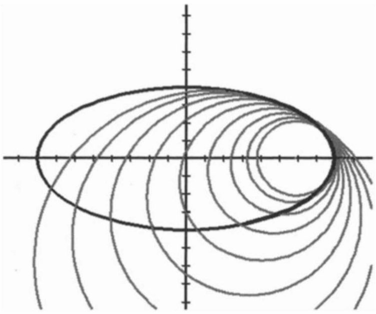


Figure 5 Circles of curvature for an elliptical arc

It is apparent from the construction of this example that the original ellipse is tangent to each of the circles in the family, and so is an envelope for the family. But the ellipse can not be obtained as the limiting points of intersections of neighboring circles. In fact, the neighboring circles are disjoint! This surprising state of affairs is shown in FIGURE 6, with an enlarged view in FIGURE 7.

This same example also exhibits some of the other exceptional behaviors that have been discussed above. One can show that the original ellipse does satisfy the two conditions specified in the envelope algorithm, and so the algorithm would properly identify the envelope for this family of circles. But the entire circles of curvature for each

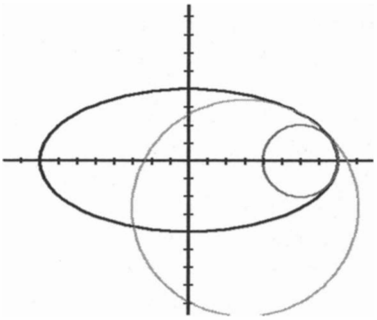


Figure 6 Neighboring circles are disjoint

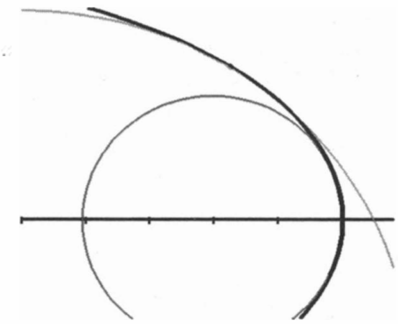


Figure 7 Closeup view of neighboring circles

of the vertices $((\pm a, 0), (0, \pm b))$ also satisfy the conditions of the envelope algorithm, although these circles are *not* part of the envelope. And these circles also contain the boundary of the region swept out by the family of circles. These properties can be observed in the next two figures, defined by two different ellipses. Note also how visually striking these figures are.

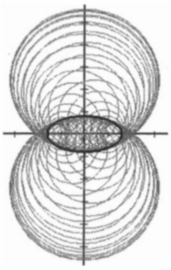


Figure 8 Family of circles for an ellipse

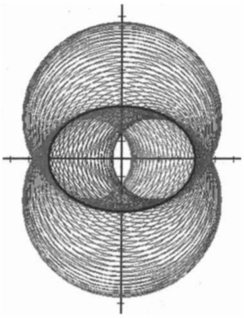


Figure 9 Family of circles for another ellipse

Experimenting with figures of this sort can convey a good deal of understanding of the properties of envelopes discussed above, and I highly recommend it. Modern graphical software is ideally suited to this purpose. For the family of circles in the preceding example, graphical exploration is abetted by the following formulae ([**19**, p. 192]). Identify one point of the ellipse as $(a\cos\alpha, b\sin\alpha)$. Then the radius of curvature at that point is given by

$$\frac{(a^2 \sin^2 \alpha + b^2 \cos^2 \alpha)^{3/2}}{ab}$$

and the center of the circle of curvature is

$$\left(\frac{a^2 - b^2}{a} \cos^3 \alpha, \frac{b^2 - a^2}{b} \sin^3 \alpha\right).$$

Extending the ladder problem

A slight variation on the ladder problem is illustrated in FIGURE 10, with a rectangular alcove in the corner where the two hallways meet. The same configuration might occur if there is some sort of obstruction, say a table or a counter, in one hallway near the corner. As before, the problem is to find how long a line segment will go around this corner.

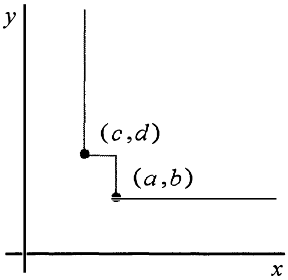


Figure 10 Family of circles for another ellipse

Here, the envelope method again provides an immediate solution. Consider again the region Ω swept out by a family of lines of fixed length L. If this region avoids both points (a, b) and (c, d), then the segment can be moved around the corner. As L increases, the envelope (1) expands out from the origin. The maximal feasible L occurs when the envelope first touches one of the corner points (a, b) and (c, d). This shows that the maximal value of L is given by

$$L_{\max} = \min\left\{(a^{2/3} + b^{2/3})^{3/2}, (c^{2/3} + d^{2/3})^{3/2}\right\}.$$

Going a bit further in this direction, we might replace the inside corner with any sort of curve C (see FIGURE 11). The ladder problem can then be solved by seeking the point (x, y) of C for which $f(x, y) = x^{2/3} + y^{2/3}$ is minimized. Unfortunately, this plan is not so easy to execute. For example, an elliptical arc is a natural choice for the curve C. But even for that simple case the analytic determination of the minimal value of f appears quite formidable, if not impossible. On the other hand, if C is a polygonal path, we need only find the minimum value of f at the vertices.

The couch problem. Extending the problem in a different direction, we can make the situation a bit more faithful to the real world by recognizing that a ladder actually has some positive width. Thus, in the idealized geometry of the problem statement, perhaps we should try to maneuver a rectangle rather than a line segment around the

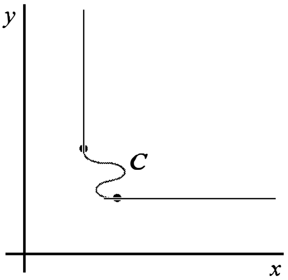


Figure 11 Ladder problem with a curve in the corner

corner. If the width of the rectangle is fixed at w, what is the greatest length L that permits the rectangle to go around the corner?

This version of the problem also provides a reasonable model for moving bulkier objects than ladders. For example, trying to push a desk or a couch around a corner in a corridor is naturally idealized to the problem of moving a rectangle around the corner in FIGURE 1. This is the motivation given by Moretti [**17**] in his analysis of the rectangle version of the ladder problem. In honor of his work, we refer to the rectangular version hereafter as the couch problem. It should not be confused with the *sofa* problem, which concerns the *area* of a figure to be moved around a corner.

Moretti's analysis mimics the standard solution to the ladder problem. Thus, rather than looking for the longest couch that will go around the corner, he seeks the shortest couch that will get stuck. This occurs when the outer corners of the rectangle touch the outer walls of the corridor, and the inner edge touches the inside corner point, as illustrated in FIGURE 12. Using the slope as a parameter, Moretti reduces the problem to finding a particular root of a sixth degree polynomial.

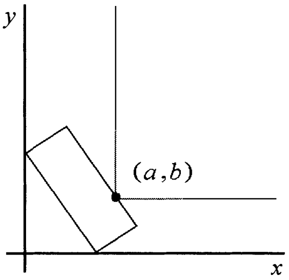


Figure 12 This couch is stuck

For the couch problem, as for the ladder problem, the direct approach using envelopes is illuminating. Indeed, we again make use of the astroid, and one of its *parallel* curves. Here, a *parallel* curve means one whose points are all at a fixed distance from a given curve. From each point P on the given curve C, move a fixed distance w along the normal vector to locate a point Q, being careful to choose the direction of the normal vector consistently. The locus of all such Q is a curve parallel to C at distance w. Parallel curves are discussed in [**19**].

For the couch problem, the envelope we need is parallel to the envelope we found for the ladder problem. This leads to the following appealing geometric interpretation. Let L be the longest rectangle of width w that can be moved around a corner as in the original ladder problem. Then the boundary of Ω (as in FIGURE 2) must be tangent to the circle of radius w centered at the point (a, b). That is, w must be the distance from the corner point (a, b) to the astroid (1). Algebraically, this approach has an appealing simplicity, up to the point of actually finding a solution. Unfortunately, that requires solving a sixth degree equation, which is essentially equivalent to the one considered by Moretti.

On the other hand, the geometric setting of the envelope approach provides a simple method for parameterizing a family of rational solutions to the couch problem. That is, we can specify an infinite set of triples (a, b, w) such that the couch problem has an exact rational solution L. This partially answers one of Moretti's questions. In fairness, though, a similar parameterization can be developed using Moretti's method.

To apply the envelope method to the couch problem, we adopt the same strategy as for the ladder problem. Consider the following process for moving a rectangle around a corner. Initially, the rectangle is aligned with the walls of the corridor, so that the bottom of the rectangle is on the x axis and the left side is on the y axis. Slide the rectangle in such a way that the lower left-hand corner follows the y axis, while keeping the lower right-hand corner on the x axis. Thus, the bottom edge of the rectangle follows the exact trajectory of the segment in the ladder problem, sweeping out the region Ω, as before. But now we want to look at the region swept out by the entire rectangle. The upper boundary of this region is the envelope of the family of lines corresponding to the motion of the *top* edge of the rectangle. For a couch with length L and width w, these lines are characterized as follows. Begin with a line in the family for the original ladder problem, whose intersection with the first quadrant has length L. Construct the parallel line at distance w (and in the direction away from the origin). We seek the envelope of the family of all of these parallel lines.

As before, we parameterize the lines in this new family in terms of the angle α between such a line and the (negatively directed) x axis. The parallel unit vector is given by $\mathbf{m} = (-\cos\alpha, \sin\alpha)$, and the normal unit vector (pointing into the first quadrant) is $\mathbf{n} = (\sin\alpha, \cos\alpha)$. These vectors provide a simple way to define a line at a specified distance d from the origin: begin with the line through the origin parallel to $\mathbf{m}$, and translate by $d\mathbf{n}$. That defines a point on the line as

$$(x, y) = t\mathbf{m} + d\mathbf{n}.$$

Taking the dot product of both sides of this equation with $\mathbf{n}$ thus gives

$$\sin\alpha\, x + \cos\alpha\, y = d.$$

Lines in the original family are described by (2), which we rewrite as

$$\sin\alpha\, x + \cos\alpha\, y = L\ \sin\alpha\ \cos\alpha.$$

This line is at a distance $L\ \sin\alpha\ \cos\alpha$ from the origin. Now we want the parallel line that is w units further away. The equation for that line is evidently

$$\sin\alpha\, x + \cos\alpha\, y = L\ \sin\alpha\ \cos\alpha + w.$$

To make use of the envelope algorithm, let us define the function

$$G(x, y, \alpha) = \sin\alpha\, x + \cos\alpha\, y - L\ \sin\alpha\ \cos\alpha - w.$$

Then, thinking of α as a fixed value, the equation $G(x, y, \alpha) = 0$ defines one line in the family. Similarly, with

$$F(x, y, \alpha) = \sin\alpha\, x + \cos\alpha\, y - L\ \sin\alpha\ \cos\alpha$$

we obtain the lines in the original family by setting $F(x, y, \alpha) = 0$. It will be convenient in what follows to express these functions in the form

$$F(x, y, \alpha) = \mathbf{n} \cdot (x, y) - L \sin\alpha \cos\alpha$$
$$G(x, y, \alpha) = \mathbf{n} \cdot (x, y) - L \sin\alpha \cos\alpha - w$$

Our goal is to find the envelope for the lines defined by G, (hereafter, the envelope for G). According to the envelope algorithm, we should eliminate α from the equations

$$G(x, y, \alpha) = 0$$
$$\frac{\partial}{\partial\alpha} G(x, y, \alpha) = 0$$

But rather than apply this directly, we can use the fact that we know the envelope for F. In fact, since each line in G's family is parallel to a corresponding line in F's family, and at a uniform distance w, it is not surprising that the envelope of G is parallel to the envelope of F, and at the same distance. That is, if (x, y) is on the envelope of F, then the corresponding point of the envelope of G is w units away in the normal direction.

To make this more precise, let us consider a point (x, y) on the envelope of F. There is a corresponding α such that (x, y, α) is a zero of both F and $\frac{\partial F}{\partial\alpha}$. Then (x, y) is on the line with parameter α, which is tangent to the envelope of F at (x, y). Thus, at this point, the line and the envelope share the same normal direction. As observed earlier, the unit normal is given by $\mathbf{n} = (\sin\alpha, \cos\alpha)$. We will now consider a new point $(x', y') = (x, y) + w\mathbf{n}$. We wish to show that (x', y') is on the envelope of G.

To that end, observe that $F(x, y, \alpha) = G(x', y', \alpha)$ and $\frac{\partial F}{\partial\alpha}(x, y, \alpha) = \frac{\partial G}{\partial\alpha}(x', y', \alpha)$. To justify the first of these equations,

$$\begin{aligned} G(x', y', \alpha) &= \mathbf{n} \cdot (x', y') - L \sin\alpha \cos\alpha - w \\ &= \mathbf{n} \cdot [(x, y) + w\mathbf{n}] - L \sin\alpha \cos\alpha - w \\ &= \mathbf{n} \cdot (x, y) + w - L \sin\alpha \cos\alpha - w \\ &= F(x, y, \alpha) \end{aligned}$$

To justify the second equation, we observe first that since F and G differ by a constant, they have the same derivatives. Also, note that $\frac{\partial \mathbf{n}}{\partial\alpha} = -\mathbf{m}$. This gives $\frac{\partial G}{\partial\alpha}(x, y, \alpha) = \frac{\partial F}{\partial\alpha}(x, y, \alpha) = -\mathbf{m} \cdot (x, y) - L(\cos^2\alpha - \sin^2\alpha)$. Now we can write

$$\begin{aligned} \frac{\partial G}{\partial\alpha}(x', y', \alpha) &= -\mathbf{m} \cdot [(x, y) + w\mathbf{n}] - L(\cos^2\alpha - \sin^2\alpha) \\ &= -\mathbf{m} \cdot (x, y) - L(\cos^2\alpha - \sin^2\alpha) \\ &= \frac{\partial F}{\partial\alpha}(x, y, \alpha). \end{aligned}$$

Together, these results show that (x, y) is on the envelope of F if and only if (x', y') is on the envelope of G, and that in each case the points (x, y) and (x', y') correspond to the same value of α. In fact, this result reflects a more general situation: If the family G consists of parallels of the curves in F, all at a fixed distance w, then the envelope for G is the parallel of the envelope of F, at the same distance w. In the context of the couch problem, we can find the needed envelope of G as a parallel to the known envelope of F.

Based on earlier work, we know that the envelope of F is parameterized by the equations

$$x = L\cos^3\alpha$$
$$y = L\sin^3\alpha.$$

That leads immediately to the following parametric description of the envelope of G :

$$x = L\cos^3\alpha + w\sin\alpha$$
$$y = L\sin^3\alpha + w\cos\alpha.$$

For the solution L of the couch problem, the point (a, b) must lie on the envelope of G. Therefore, we can find L (and also find the critical value of α) by solving the system

$$a = L\cos^3\alpha + w\sin\alpha$$
$$b = L\sin^3\alpha + w\cos\alpha.$$

This leads readily enough to an equation in α alone:

$$a\sin^3\alpha - b\cos^3\alpha = w(\sin^2\alpha - \cos^2\alpha). \tag{8}$$

At this point, finding α appears to depend on solving a sixth degree polynomial equation. To derive such an equation, substitute x for $\sin\alpha$ and $\sqrt{1 - x^2}$ for $\cos\alpha$ in (8) to obtain

$$ax^3 - b(1 - x^2)^{3/2} = w(2x^2 - 1).$$

Isolating the term with the fractional exponent and squaring both sides then leads to the equation

$$(a^2 + b^2)x^6 - 4awx^5 + (4w^2 - 3b^2)x^4 + 2awx^3 + (3b^2 - 4w^2)x^2 + w^2 - b^2 = 0.$$

It is is easy to solve this equation numerically (given values for a, b, and w), and very likely impossible to solve it symbolically.

As mentioned, the foregoing analysis leads to a nice geometric interpretation for the solution. If (a, b) is on the envelope of G, then there is a corresponding point (x, y) on the envelope of F. We know that (x, y) is w units away from (a, b), and that the vector between these two points is normal to the envelope of F. This shows that the circle centered at (a, b) of radius w is tangent to the envelope of F at (x, y).

Visually, we can see how to find the maximum value of L. Start with a small enough L so that the astroid (1) stays well clear of the circle about (a, b) of radius w. Now increase L, expanding the astroid out from the origin, until the curve just touches the circle. When that happens, the corresponding value of L is the solution to the couch problem. See FIGURE 13.

The visual image of solving the couch problem in this way is reminiscent of Lagrange Multipliers. Indeed, what we have is the dual of a fairly typical constrained optimization problem: find the point on the curve (1) that is closest to (a, b). The visual image for that problem is to expand circles centered at (a, b) until one just touches the astroid. Our dual problem is to hold the circle fixed and look at level curves for increasing values of the function $f(x, y) = x^{2/3} + y^{2/3}$. We increase the value of f until the corresponding level curve just touches the fixed circle. This geometric conceptualization is associated with the following optimization problem: Find the minimum value of $f(x, y)$ where (x, y) is constrained to lie on the circle of radius w centered at (a, b). We will return to the idea of dual problems in the last section of the paper.

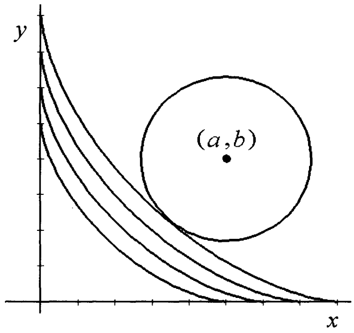


Figure 13 Maximizing L geometrically

Let us examine more closely the Lagrangian-esque version of the couch problem. We wish to find a point of tangency between the following two curves

$$x^{2/3}+y^{2/3}=L^{2/3}$$
$$(x-a)^2+(y-b)^2=w^2.$$

For each curve, we can compute a normal vector as the gradient of the function on the left side of the curve's equation. Insisting that these gradients be parallel leads to the following additional condition

$$x^{1/3}(x-a)=y^{1/3}(y-b).$$

In principle, solving these three equations for x, y, and L would produce the desired solution L to the couch problem. Or, solving the second and third for x and y, and then substituting those values in the first equation, also would lead to the value of L. Unfortunately, every approach appears to lead inevitably to a sixth degree equation.

While the envelope method does not seem to provide a symbolic solution to the couch problem, it does provide a nice procedure for generating solvable examples. Here, we will begin with a value of L and produce a triple (a, b, w) so that L is a solution to the (a, b, w) couch problem. To begin, we generate some *nice* points on the astroid $x^{2/3}+y^{2/3}=L^{2/3}$ using Pythagorean triples. Specifically, if $r^2+s^2=t^2$, we can take $x=r^3$, $y=s^3$ and $L=t^3$ to define a point on an astroid. In particular, we can generate an abundance of rational points on astroids. Notice that the original Pythagorean triple need not be rational. For example, if $(r, s, t)=(3, 4, 5)/\sqrt[3]{5}$, we find $(27/5, 64/5)$ as a rational point on the astroid curve for $L=25$.

Now the equations $x=r^3$, $y=s^3$ are closely related to the parameterization

$$x=L\cos^3\alpha \quad y=L\sin^3\alpha$$

of the astroid. As a result, we can recover α from the equations

$$\cos\alpha=\frac{r}{t} \quad \sin\alpha=\frac{s}{t}.$$

This in turn gives us the normal vector $\mathbf{n}=(\frac{s}{t}, \frac{r}{t})$, and hence, for any value of w, leads to the point (x', y'). Define that point to be (a, b). It necessarily lies on the envelope for G. This shows that the L for the astroid constructed at the outset solves the (a, b, w) couch problem. We formalize these arguments in the following theorem.

THEOREM. *For any positive pythagorean triple* (r, s, t) *and any positive* w *define*

$$a = r^3 + w\frac{s}{t}$$

$$b = s^3 + w\frac{r}{t}$$

$$L = t^3$$

Then L *is the solution to the* (a, b, w) *couch problem.*

For example, with $(r, s, t) = (3, 4, 5)/\sqrt[3]{5}$ and $w = 2$, the equations above give $(a, b) = (7, 14)$ and $L = 25$. So for a rectangle of width 2, 25 is the maximum length that will fit around the corner defined by the point $(7, 14)$. In general, if (r', s', t') is a rational Pythagorean triple, and if u^3 is rational, then taking $(r, s, t) = u(r', s', t')$ and rational w produces rational values of a, b, and L, as well as a rational point (x, y) where the astroid meets the circle centered at (a, b) of radius w.

The preceding example, where $L = 25$ is the solution of the $(7, 14, 2)$ couch problem, was given by Moretti. He mentioned that such examples are relatively rare, and asked for conditions on a, b, and w that make the (a, b, w) couch problem exactly solvable. The theorem above provides a partial answer to Moretti's question, by providing an infinite family of such triples. It would be nice to know whether every rational (a, b, w) with rational solution L to the couch problem arises in this way. If the critical value of α corresponds to a rational point (x, y) on the astroid (1), then a, b, w, and L are related as in the theorem. But there might be rational (a, b, w) for which the solution to the couch problem is also rational, but which does not correspond to a rational point (x, y).

The envelope approach leads in a natural way to the theorem, and provides a nice geometric interpretation of the couch problem solution. But it should be observed that Moretti's approach can also lead to an equivalent method for parameterizing triples (a, b, w) with rational solution L. He formulates the problem in terms of a variable m (corresponding to $\cot\alpha$ in this paper) and derives a sixth degree equation in m with coefficients that depend on a, b, and w. If that equation is solved for w, one can again parameterize solutions in terms of Pythagorean triples. From this standpoint, the envelope method does not seem to hold any advantage over Moretti's earlier analysis.

Duality in the ladder problem

In discussing the tangency condition for an astroid and a circle, the idea of dual optimization problems was briefly mentioned. As a concluding topic, we will look at this idea again.

Segalla and Watson [**21**] discuss what they call the flip side of a constrained optimization problem in the context of Lagrange multipliers. For example, in seeking to maximize the area of a rectangle with a specified perimeter, we have an objective function (the area) and a constraint (the perimeter). At the solution point, the level curve for the extreme value of the objective function is tangent to the given level curve of the constraint function. Here, the roles of the objective and constraint are symmetric, and can be interchanged. Given the maximal area, we can ask what is the minimal perimeter that can enclose a rectangle having this area. The solution corresponds to the same point of tangency between level curves of the objective and constraint functions. Thus, we see that the problem of maximizing area with a fixed perimeter, and minimizing the perimeter with a fixed area are linked.

Maximizing area with fixed perimeter is the famous isoperimetric problem (see Blåsjö [2] for a beautiful discussion), and in that context minimizing the perimeter with fixed area is referred to as the *dual* problem. Duality in this sense corresponds to Segalla and Watson's idea of flip side symmetry. It is also reminiscent of the idea of duality in linear and non-linear programming. There, although the primal and dual optimization problems occur in different spaces, one again finds the idea of linked problems whose solutions somehow coincide.

Duality permits information about one problem to be inferred from information about its dual problem. This property is important in both the isoperimetric problem and in linear programming. As Segalla and Watson point out, a solution to an initial optimization problem immediately leads to a corresponding statement and solution of a dual problem. Thus, discovering that a rectangle with perimeter 40 has maximal area 100, also tells us at once that a rectangle with area 100 has minimal perimeter 40. But there is another way to use duality. If you are unable to solve an optimization problem, try to solve the dual.

Here is how this works for the perimeter-area problem. Imagine that we do not know how to maximize the area of a rectangle with perimeter 40. The dual problem is to minimize the perimeter subject to a given area, but of course we do not know what that fixed area should be. So we solve the general problem for a fixed area of A. That is, we prove that area A occurs with a minimal perimeter of $4\sqrt{A}$. Now relate this to the original problem by insisting on a perimeter of 40. That forces $A = 100$, and tells us this: for area 100, the minimal perimeter is 40. The dual statement now solves the original problem.

Segalla and Watson give several examples of pairs of dual problems. In addition to the area-perimeter example mentioned above, they discuss the *milkmaid problem*: find the minimum distance the milkmaid must walk from her home to fetch water from a river and take it to the barn. In this case the dual problem fixes the length of the milkmaid's hike, and minimally shifts the river to accommodate her. They also give the example of the ladder problem, but do not describe the dual. Indeed, they ask for an interpretation of the dual ladder problem. Let us answer this question using envelopes, and see how the dual version leads to another solution of the ladder problem.

Segalla and Watson use the standard approach to the ladder problem—finding the minimal length segment that will get stuck in the corner. We formulate this as a constrained optimization problem in terms of variables u and v, interpreted as intercepts on the x and y axes of a line segment in the first quadrant. The objective function, $f(u, v) = \sqrt{u^2 + v^2}$, is the distance between the intercepts. The goal is to minimize this distance subject to the constraint that the line must pass through (a, b).

For the dual problem, if we hold f fixed, and look at varying values of the constraint function, what does that mean? An answer will depend, naturally, on how the constraint g is formulated. Here is one approach. A line with intercepts u and v satisfies the equation

$$\frac{x}{u} + \frac{y}{v} = 1.$$

From this equation, the condition that (a, b) lie on the line is

$$\frac{a}{u} + \frac{b}{v} = 1.$$

Accordingly, define $g(u, v) = a/u + b/v$. Now observe that $g(u, v) = t$ means that $(a/t, b/t)$ lies on a line with intercepts u and v. That is, g measures the (reciprocal of the) distance from the origin to the line for u and v along the ray through (a, b). This gives the following meaning to the dual problem: Look at all the lines with intercepts u

and v, where fixing the value of f means that the distance between these intercepts is constant. Among all these lines, find the one whose distance from the origin, measured in the direction of (a, b) is a maximum.

The by-now-familiar astroid appears once again as the envelope of a family of lines. Holding f constant with value L, we are again considering the family of line segments of length L with ends on the positive x and y axes, filling up the region Ω. The point that maximizes g will now be the furthest point you can reach in Ω traveling on the ray from the origin to (a, b). That is, the solution occurs at the intersection of the ray with the envelope (1).

The solution point will be of the form $(a/t, b/t)$ where t is the optimal value of g. Substituting in (1), we find $t = ((a/L)^{2/3} + (b/L)^{2/3})^{3/2}$. This gives us the optimal value of g for the dual problem, in terms of L. To return to the primal problem, we have to choose the value of L that gives us the original constraint value for g, namely $g = 1$. So with $t = ((a/L)^{2/3} + (b/L)^{2/3})^{3/2} = 1$ we again find $L = (a^{2/3} + b^{2/3})^{3/2}$.

It is interesting that the ladder problem has so many formulations. The usual approach is to reverse the original problem, so that we seek a minimal line that cannot go around the corner rather than a maximal line that will go around the corner. The envelope approach presented here deals directly with the problem as stated, finding the maximum line that will fit around the corner. A third approach is to take the dual of the reversed version, viewed as an example of constrained optimization. Although all of these approaches are closely related, each contributes a slightly different understanding of the problem.

REFERENCES

1. Leah Wrenn Berman, Folding Beauties, *College Math. J.* **37** (2006) 176–186.
2. Viktor Blåsjö, The Isoperimetric Problem, *Amer. Math. Monthly* **112** (2005) 526–566.
3. John Cooper, *Ladder Problem Query*, private correspondence, 2006.
4. Richard Courant, *Differential and Integral Calculus, Volume 2,* tranlated by E. J. McShane, Interscience, New York, 1949.
5. L. M. Court, Envelopes of Plane Curves, *Amer. Math. Monthly* **57** (1950) 168–169.
6. H. T. Croft, K. J. Falconer, and R. K. Guy, *Unsolved Problems in Geometry*, Springer-Verlag, New York, 1994.
7. William Holding Echols, *An Elementary Text-Book on the Differential and Integral Calculus,* Holt, New York, 1902.
8. Howard Eves, A Note on Envelopes, *Amer. Math. Monthly* **51** (1944) 344.
9. S. R. Finch, *Moving Sofa Constant,* section 8.12 (pp. 519–523) in Mathematical Constants, Cambridge University Press, Cambridge, England, 2003.
10. T. J. Fletcher, Easy Ways of Going Round the Bend, *Mathematical Gazette* **57** (1973) 16–22.
11. Peter J. Giblin, Zigzags, this MAGAZINE, **74** (2001) 259–271.
12. J. W. Green, On the Envelope of Curves Given in Parametric Form, *Amer. Math. Monthly* **59** (1952) 626–628.
13. James Haddon, *Examples and Solutions of the Differential Calculus,* Virtue, London, 1862.
14. Morris Kline, *Advanced Calculus,* 2nd ed., Wiley, New York, 1972.
15. H. E. Licks, *Recreations in Mathematics,* D. Van Nostrand, New York, 1917, p. 89.
16. Brian J. Loe and Nathanial Beagley, The Coffee Cup Caustic for Calculus Students, *College Math. J.* **28** (1997) 277–284.
17. Christopher Moretti, Moving a Couch Around a Corner, *College Math. J.* **33** (2002) 196–201.
18. The Economics Professor, *Envelope Theorem*, Arts & Sciences Network, `http://www.economyprofessor.com/economictheories/envelope-theorem.php`.
19. John W. Rutter, *Geometry of Curves,* Chapman & Hall/CRC, Boca Raton, 2000
20. Mark Schwartz, The Chair, the Area Rug, and the Astroid, *College Math. J.* **26** (1995) 229–231.
21. Angelo Segalla and Saleem Watson, The Flip-Side of a Lagrange Multiplier Problem, *College Math. J.* **36** (2005) 232–235.
22. Andrew Simoson, An Envelope for a Spirograph, *College Math. J.* **28** (1997) 134–139.
23. Andrew Simoson, The Trochoid as a Tack in a Bungee Cord, this MAGAZINE **73** (2000) 171–184.

24. David Singmaster, *Sources in Recreational Mathematics, an Annotated Bibliography*, article 6.AG. `http://us.share.geocities.com/mathrecsources/`.

25. Angus E. Taylor and W. Robert Mann, *Calculus: an Intuitive and Physical Approach,* 2nd ed., Wiley, New York, 1977.

26. D Thayer Watkins, *The Envelope Theorem and Its Proof*, San Jose State University, `http://www2.sjsu.edu/faculty/watkins/envelopetheo.htm`.

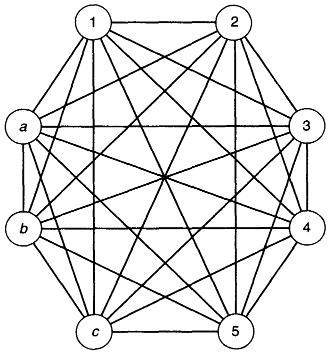
27. Eric W. Weisstein, *Moving Sofa Problem,* From MathWorld—A Wolfram Web Resource, `http://mathworld.wolfram.com/MovingSofaProblem.html`.

28. Robert C. Yates, *Curves and their Properties,* National Council of Teachers of Mathematics, Reston, VA, 1974

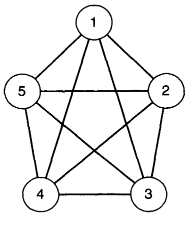
Proof Without Words: A Graph Theoretic Decomposition of Binomial Coefficients

$$\binom{n+m}{2} = \binom{n}{2} + \binom{m}{2} + nm$$

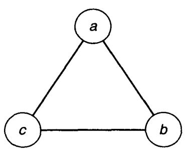
E.g., $n = 5, m = 3$.



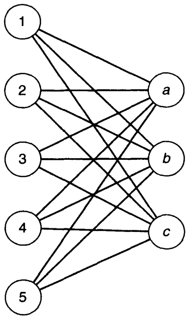
K_8



K_5



K_3



$K_{5,3}$

—JOE DEMAIO
KENNESAW STATE UNIVERSITY

The Recreational Gambler: Paying the Price for More Time at the Table

JOSEPH BAK
City College of New York
New York, NY 10031
jbak@ccny.cuny.edu

Introduction

Suppose a person decides to spend some time in a casino. He sets aside \$100 which he feels he can afford to lose and continually makes even-money bets of \$10, so that his initial "fortune" of \$100 is either increased or decreased by \$10 after each bet. Assume also that his chance of winning each bet is 0.47, roughly what it would be at the roulette wheel. Finally, assume that our gambler has decided to stop betting when his fortune has either increased to \$160 or declined to \$0. What is the probability that the gambler will emerge successfully, leaving the casino with \$160?

In spite of its contemporary trappings, this contest can be viewed as an example of one of the earliest problems in the history of probability. The classic gambler's ruin involves two bettors, one with an initial fortune of A units, the other with an initial fortune of $(B - A)$, who repeatedly bet one unit against each other until one of them is ruined. If we focus on the player with the initial fortune of A units, we could say that he has decided to bet continually until he either loses all of his money (i.e., is *ruined*) or raises his fortune to B. If this player has probability p of winning each individual bet, his probability of reaching B units, which we will denote $P(A, B, p)$, is given by a well-known formula which we will discuss in the next section.

With some very elementary adjustments, our gambler fits the mold of the classic gambler described above. All we have to do is view the money being wagered in units of \$10 each, and think of the casino as the gambler's opponent. Then we can describe the gambling as a contest between two individuals, one with initial fortune of 10 units and the other with an initial fortune of 6, who continually bet one unit until one of them is ruined. Hence the probability that our gambler will reach his goal of \$160 is $P(10, 16, 0.47)$, which is approximately 0.40.

Suppose now that our gambler decides he would rather play more conservatively, betting only \$5 on each game, but maintaining the same stopping values of \$160 or \$0. With this new approach, his initial \$100 represents 20 units, and he is betting one unit at a time until he is ruined or reaches his goal of 32. This "conservative" approach will actually increase his chance of being ruined. His chance of reaching the desired \$160, now equal to $P(20, 32, 0.47)$, is only about 0.22.

The fact that decreasing the size of the bets results in a lower probability of success is hardly surprising. The individual bets are slanted against the gambler, and the larger number of smaller bets increases the likelihood that the "favored" casino will ultimately win the contest. Feller [**4**, p. 346] pointed out that reducing the individual wagers by a factor of k changes the gambler's chance of success from $P(A, B, p)$ to $P(kA, kB, p)$, and he proved that this results in a reduced probability of success, if $k = 2$, as long as $p < \frac{1}{2}$. More recently, it was proved that $P(kA, kB, p) < P(A, B, p)$ for all integers $k > 1$, as long as $p < \frac{1}{2}$ [**6**, p. 405]. Of course, $p < \frac{1}{2}$ if and only if the opposing player's probability of winning each game ($q = 1 - p$) is greater than $\frac{1}{2}$. Thus it follows that, for $p > \frac{1}{2}$, the reverse is true: $P(kA, kB, p) > P(A, B, p)$.

If it is fairly intuitive that smaller bets lead to a decreased chance of success, how do we understand the popularity of such bets? Part of the answer, undoubtedly, is that the typical gambler considers his visit to the casino as a form of entertainment rather than as a good business opportunity. Larger bets may make more sense financially (as would not entering the casino in the first place), but they would have a lower entertainment value. By making smaller bets, the gambler suspects that he will extend his playing time. In fact, reducing the bet size is guaranteed to increase the playing time. Picture our gambler entering the casino with \$100 and making his sequence of \$10 bets, and imagine his more conservative twin brother coming in with the same initial fortune and making the same bets (at the same time!) except that he bets \$5 each time. Since the twin's net gain or loss will always be exactly 1/2 of his bolder brother's, when the bolder brother has finished (gaining \$60 or losing \$100), the conservative twin is still playing.

This leads us to another aspect of the classic gambler's ruin, the *duration of play.* This is defined as the expected number of games until the contest is over, and denoted $D(A, B, p)$. Like $P(A, B, p)$, $D(A, B, p)$ is given by a formula which we will consider in the next section.

As we noted above, reducing the individual bets by a factor of k changes the probability of success from $P(A, B, p)$ to $P(kA, kB, p)$. Similarly, the duration of play changes from $D(A, B, p)$ to $D(kA, kB, p)$. But while the probability of success declines, the duration of play increases. For example, as our gambler switches from \$10 to \$5 bets, his duration of play changes from $D(10, 16, 0.47)$, which is about 60, to $D(20, 32, 0.47)$, which is just over 216. This extended playing time may compensate somewhat for the reduced probability of success.

In Section 2, we consider the precise connection between the reduced probability of success and the increased duration, and we examine the possible values of the ratio of increased duration: $\frac{D(kA,kB,p)}{D(A,B,p)}$. It is fairly easy to see that the ratio is at least as large as k. That is, it should take at least k times as long to lose kA units or win $k(B - A)$ units as it does to lose A units or win $(B - A)$. (If you find this intuitive argument unsatisfying, note that the claim follows from the equations in (7).) Finding the maximum value of the ratio, however, is considerably more difficult. Two recent results in this regard are the following: *There exist values of B, p and k such that*

$$\frac{D(k, kB, p)}{k^2 D(1, B, p)} \text{ is arbitrarily close to 2.} \quad [\mathbf{2}, \text{p. } 186] \tag{1}$$

Thus the ratio of increased duration can be (asymptotically) as large as $2k^2$. This seems to be the maximal ratio of increased duration. Complementing the above result, it was proved that: *For any positive integral values of B and k, and for $\frac{1}{2} < p < 1$,*

$$\frac{D(k, kB, p)}{D(1, B, p)} < 2k^2. \quad [\mathbf{2}, \text{p. } 187] \tag{2}$$

In section 3 we will complete this result, proving the following theorem.

THEOREM. *For any positive integral values of A, B, k with $B > A$, and $0 \leq p \leq 1$,*

$$\frac{D(kA, kB, p)}{D(A, B, p)} < 2k^2. \tag{3}$$

The proof introduces an interesting family of polynomials and a somewhat novel application of Descartes' Rule of Signs.

The probability of success, expected value, bias, and duration of play

The formula for the probability of success [4, p. 345], which is well over 300 years old, is

$$P(A, B, p) = \frac{(q/p)^A - 1}{(q/p)^B - 1} \quad \text{if} \quad p \neq \frac{1}{2}; q = 1 - p;$$
$$P\left(A, B, \frac{1}{2}\right) = A/B. \tag{4}$$

The classic gambler's ruin has only two possible outcomes: a loss of A if he is ruined and a gain of $(B - A)$ if he wins. The probability that the contest will go on forever is 0. (This fact may be intuitive, but in fact its proof is nontrivial [4, pp. 344–345].) Thus if X represents the outcome of the contest, its expected value is

$$E[X] = (B - A)P(A, B, p) - A[1 - P(A, B, p)] = BP(A, B, p) - A. \tag{5}$$

Note that $P(A, B, p)$ is clearly an increasing function of p and, according to (4), the expected value is zero for $p = \frac{1}{2}$. Hence it is positive for $p > \frac{1}{2}$, and negative for $p < \frac{1}{2}$.

The contest is "fair," i.e., has an expected value of 0, when $p = \frac{1}{2}$, in which case $P(A, B, p) = A/B$. For this reason, we will call $|P(A, B, p) - A/B|$ the *bias* in the contest. If we view the outcome once again as a contest between two players, one with an initial fortune of A, and the other with an initial fortune of $(B - A)$, then the total amount of money being contested is B. According to (5), the expected value for *either* player is ± the product of the total amount of money being contested multiplied by the bias. Note also that $P(kA, kB, p)$ is less than or greater than $P(A, B, p)$ if p is less than or greater than $\frac{1}{2}$, respectively, so the bias is increased when the amount wagered on each bet is reduced. That is,

$$|P(kA, kB, p) - A/B| > |P(A, B, p) - A/B|$$

in any unfair contest.

We now consider the duration of the contest. Let X_i be the outcome of the ith bet. Then $E[X_i] = p - q$ and $E[X_i^2] = 1$. The duration, $D(A, B, p)$, is equal to $E[T]$, where T is the smallest integer with $X_1 + X_2 + \cdots + X_T =$ either $-A$ or $B - A$. According to Wald's Identities [5, pp. 266–268],

$$D(A, B, p) = \frac{E[X]}{E[X_i]} = \frac{BP(A, B, p) - A}{p - q}, \quad \text{if} \quad p \neq \frac{1}{2}$$
$$D(A, B, 1/2) = \frac{E[X^2]}{E[X_i^2]} = A(B - A). \tag{6}$$

The ratio of increased duration is

$$\frac{D(kA, kB, p)}{D(A, B, p)} = k\frac{BP(kA, kB, p) - A}{BP(A, B, p) - A} = k\frac{P(kA, kB, p) - A/B}{P(A, B, p) - A/B}, \quad \text{if} \quad p \neq \frac{1}{2}$$
$$\frac{D(kA, kB, 1/2)}{D(A, B, 1/2)} = k^2. \tag{7}$$

Thus, if $p = \frac{1}{2}$, reducing the size of the bets by a factor of k increases the duration by a factor of k^2. In all other cases, reducing the wagers by a factor of k increases the duration by a factor equal to k times the ratio of increase in the bias of the contest.

The minimum ratio of increased duration is k, which is achieved for the limiting values $p = 0$ and $p = 1$ when $P(A, B, p) = P(kA, kB, p) = p$. Finding the maximum ratio of increased duration is considerably more difficult. A natural conjecture would be that for fixed values of A and B, the greatest ratio of increased duration occurs in the case of a fair contest. The conjecture is actually true for a gambler who adopts a double-or-nothing strategy, i.e., when $B = 2A$. (In fact, many aspects of the theory are especially neat in this case, e.g., the proof of formula (4). See [**7**, pp. 101–111].) In all other cases, however, the conjecture is false [**2**, p. 176].

Note that, according to (7), if the contest bias is bounded away from zero, the ratio of increased duration cannot be much larger than k. For example, if $|P(A, B, p) - A/B| \geq \frac{1}{3}$, the ratio of increased duration is at most $3k$. In order to obtain a ratio with an order of magnitude greater than k, it is necessary that

(i) the bias in the original contest, $|P(A, B, p) - A/B|$, is close to 0, and

(ii) the bias in the contest with reduced wagers, $|P(kA, kB, p) - A/B|$, is considerably larger.

The first goal can be accomplished, for any fixed values of A and B, by taking p very close to $\frac{1}{2}$. To accomplish the second goal, suppose that p is slightly greater than $\frac{1}{2}$, A is small and B is large. Since $p > \frac{1}{2}$, $P(A, B, p) > A/B$. But with p close enough to $\frac{1}{2}$, $P(A, B, p)$ will be only slightly greater than A/B. Thus, even though the gambler has a better than 50-50 chance of winning each bet, he still has only a slight chance of not being ruined. Intuitively, this is due to the small value of A, which makes the gambler vulnerable to an early ruin, in spite of his slight advantage on each bet. On the other hand, $P(kA, kB, p)$ might be substantially larger since the larger starting value of kA reduces his vulnerability.

It is not surprising, then, that the large ratio of increased duration in (1) was obtained by taking $A = 1$, a very large value of B, and p just slightly greater than $\frac{1}{2}$.

A universal upper bound for the ratio of increased duration

To prove inequality (3), valid for $0 \leq p \leq 1$, we first consider a (fixed) value of p, $\frac{1}{2} < p < 1$. For integral values of k and B, greater than or equal to 2 and 3, respectively, we will show that the ratio of increased duration is a monotonic decreasing function of A for $1 \leq A \leq B - 1$. Assume, then, that $1 \leq A \leq B - 2$. We want to show that

$$\frac{D(k(A+1), kB, p)}{D(A+1, B, p)} < \frac{D(kA, kB, p)}{D(A, B, p)} \tag{8}$$

or equivalently that

$$\frac{BP(k(A+1), kB, p) - (A+1)}{BP(A+1, B, p) - (A+1)} < \frac{BP(kA, kB, p) - A}{BP(A, B, p) - A}.$$

Applying (4), making the substitution $\frac{q}{p} = x$, and doing the usual algebraic simplifications yields the equivalent formulation:

$$\begin{aligned}&[B(x^{k(A+1)} - 1) - (A+1)(x^{kB} - 1)][B(x^{A} - 1) - A(x^{B} - 1)]\\&\quad < [B(x^{kA} - 1) - A(x^{kB} - 1)][B(x^{A+1} - 1) - (A+1)(x^{B} - 1)].\end{aligned}$$

To further simplify the notation, let $T_N = T_N(x) = x^N - 1$. Then, after canceling the common last term on each side of the inequality and dividing the remaining terms by B, our inequality takes the form:

$$Q(x) < 0, \tag{9}$$

where

$$\begin{aligned} Q(x) = {} & A[T_{kB}T_{A+1} - T_{k(A+1)}T_B] + (A+1)[T_{kA}T_B - T_{kB}T_A] \\ & + B[T_{k(A+1)}T_A - T_{kA}T_{A+1}], \end{aligned} \tag{10}$$

and $x = \frac{q}{p}$ is between 0 and 1, since $\frac{1}{2} < p < 1$.

Before proceeding with the general argument, we consider a particular case with $A = 2$, $B = 4$, and $k = 3$. Then

$$\begin{aligned} Q(x) &= 2[(x^{12}-1)(x^3-1) - (x^9-1)(x^4-1)] \\ &\quad + 3[(x^6-1)(x^4-1) - (x^{12}-1)(x^2-1)] \\ &\quad + 4[(x^9-1)(x^2-1) - (x^6-1)(x^3-1)] \\ &= 2x^{15} - 3x^{14} - 2x^{13} + x^{12} + 4x^{11} + 3x^{10} - 6x^9 + x^6 - x^4 + 2x^3 - x^2 \\ &= (x-1)^5(2x^{10} + 7x^9 + 13x^8 + 16x^7 + 14x^6 + 10x^5 + 6x^4 + 3x^3 + x^2), \end{aligned}$$

so that $Q(x) < 0$ for $0 < x < 1$.

The general argument is remarkably similar. We will show that for *any* choices of A, B and k, the polynomial $Q(x)$ can be factored as $(x-1)^5R(x)$, where $R(x)$ has all positive coefficients. To obtain this result, we will prove two lemmas:

LEMMA 1. *$Q(x)$ has a zero of degree at least five at $x = 1$. That is, $Q(x)$ and its first four derivatives are all zero at* 1.

LEMMA 2. *The coefficients of $\frac{Q(x)}{x-1}$, in its expanded form, have exactly four sign changes.*

Note that while we will be able to show $\frac{Q(x)}{x-1}$ has exactly four sign changes, $Q(x)$ may have more than five. For example, in the case considered above, $Q(x)$ had seven changes of sign, while $\frac{Q(x)}{x-1}$ is equal to

$$2x^{14} - x^{13} - 3x^{12} - 2x^{11} + 2x^{10} + 5x^9 - x^8 - x^7 - x^6 - x^3 + x^2.$$

We will apply the two lemmas along with the following version of Descartes' Rule of Signs.

LEMMA 3. (DESCARTES) *The number of sign changes in $(x-a)P(x)$, where P is any polynomial and a is a positive number, is at least one more than the number of sign changes in $P(x)$.*

We defer the proofs of Lemmas 1 and 2 to the next section, but note how they offer a quick proof of inequality (9): According to Lemma 1, $\frac{Q(x)}{x-1}$ can be factored as $(x-1)^4R(x)$. By Lemma 3, $(x-1)R(x)$ has at least one more sign change than $R(x)$, $(x-1)^2R(x) = (x-1)[(x-1)R(x)]$ has at least two more sign changes than $R(x), \ldots,$ and $(x-1)^4R(x)$ has at least four more sign changes than $R(x)$. However, by Lemma 2, $\frac{Q(x)}{x-1}$ has exactly four sign changes. This implies that $R(x)$ has no sign changes and consequently all its coefficients are positive since its leading coefficient is $A > 0$. Since $Q(x) = (x-1)^5R(x)$, we conclude that $Q(x) < 0$ for $0 < x < 1$.

We can now complete the proof of the Theorem, i.e., inequality (3). As we mentioned previously, inequality (9) implies that the ratio of increased duration, $\frac{D(kA,kB,p)}{D(A,B,p)}$, is a decreasing function of A for $\frac{1}{2} < p < 1$. Since $\frac{D(k,kB,p)}{D(1,B,p)} < 2k^2$, according to inequality (2), we have proved (3) as long as $\frac{1}{2} < p < 1$.

If $0 < p < \frac{1}{2}$, we can consider the duration of the game from the point of view of our gambler's opponent who has the complementary probability, $q = 1 - p$, of winning each game and an initial sum of $B - A$. Thus $D(A, B, p) = D(B - A, B, q)$ and the ratio of increased duration $\frac{D(kA,kB,p)}{D(A,B,p)} = \frac{D(k(B-A),kB,q)}{D(B-A,B,q)}$ has the exact same values as $\frac{D(kA,kB,q)}{D(A,B,q)}$ in reverse order as A varies from 1 to $B - 1$. Since $\frac{1}{2} < q < 1$, inequality (3) is proved for all values of p between 0 and 1, except for $p = 0, \frac{1}{2}$, or 1. For $p = \frac{1}{2}$, however, recall from (7) that the ratio of increased duration is always k^2, and for $p = 0$ or 1, it is obviously equal to k. Hence inequality (3) is completely proved, assuming Lemmas 1 and 2, of course.

Proofs of Lemmas 1 and 2

Proof of Lemma 1. The proof of Lemma 1 is rather straightforward. The facts that Q and Q' are zero at $x = 1$ follow immediately from the fact that $T_N(1) = 0$ for all N. To analyze the higher-order derivatives, we will use Leibniz' Rule for the nth derivative of a product: $(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}$ and the fact that $T_N^{(k)}(1) = N(N-1)\cdots(N-k+1)$ for $k \geq 1$.

To evaluate $Q''(1)$, we need only consider the middle term in Leibniz' Rule for each product, so that

$$\begin{aligned} Q''(1) &= 2A[kB(A+1) - k(A+1)B] + 2(A+1)[kAB - kBA] \\ &\quad + 2B[k(A+1)A - kA(A+1)] = 0. \end{aligned}$$

For $Q'''(1)$, we consider the middle two terms in Leibniz' Rule. Thus

$$\begin{aligned} Q'''(1) &= 3A[kB(kB-1)(A+1) + (A+1)AkB - k(A+1)(kA+k-1)B \\ &\qquad - B(B-1)k(A+1)] \\ &\quad + 3(A+1)[kA(kA-1)B + B(B-1)kA - kB(kB-1)A \\ &\qquad - A(A-1)kB] \\ &\quad + 3B[k(A+1)(kA+k-1)A + A(A-1)k(A+1) \\ &\qquad - kA(kA-1)(A+1) - (A+1)AkA]. \end{aligned}$$

Note then that the six positive terms are exactly matched by the six negative terms, so that $Q'''(1) = 0$.

In evaluating $Q^{(4)}(1)$, we need to consider terms of the form $T_N' T_M'''$ as well as terms of the form $T_N'' T_M''$. The former yield twelve terms and, as was the case in evaluation $Q'''(1)$, the six positive terms are matched exactly by the six negative terms. The middle terms in Leibniz' Rule all have a common factor $kAB(A+1)$, so that $Q^{(4)}(1) = 6kAB(A+1)[(kB-1)A - (kA+k-1)(B-1) + (kA-1)(B-1) - (kB-1)(A-1) + (kA+k-1)(A-1) - A(kA-1)]$. In this case, the positive and negative terms in the brackets are not identical, but expanding the bracketed expressions yields 24 terms in the form of twelve positive terms and their additive inverses. Hence $Q^{(4)}(1) = 0$ and Lemma 1 is complete. ■

Proof of Lemma 2. The proof of Lemma 2 is also elementary, but involves numerous cases, which we will handle in an organized way. First, we revisit formula (10) for $Q(x)$ and find $\frac{Q(x)}{x-1}$, dividing each of the polynomials of the form T_N, where N is *not* expressed as a multiple of k, by $(x-1)$. Since $(x^N-1)/(x-1) = 1+x+\cdots+x^{N-1}$, we obtain

$$\begin{aligned}\frac{Q(x)}{x-1} &= A[(x^{kB}-1)(1+x+\cdots+x^{A}) - (x^{k(A+1)}-1)(1+x+\cdots+x^{B-1})]\\ &\quad + (A+1)[(x^{kA}-1)(1+x+\cdots+x^{B-1})\\ &\qquad\qquad - (x^{kB}-1)(1+x+\cdots+x^{A-1})]\\ &\quad + B[(x^{k(A+1)}-1)(1+x+\cdots+x^{A-1}) - (x^{kA}-1)(1+x+\cdots+x^{A})].\end{aligned}$$

In spite of the fact that the above polynomial could be of arbitrarily large degree, we will be able to describe the sequence of its coefficients. Recall that $B > A+1$, so that $\frac{Q(x)}{x-1}$ is a polynomial of degree $kB+A$. Each of its coefficients is a sum of some subset of the numbers $\pm A$, $\pm(A+1)$ and $\pm B$, and each of those six numbers multiplies two "strings" of powers, one beginning with the constant term, and the other beginning beyond the end of the first string. For example, $-A$ multiplies all monomials from 1 to x^A, and from $x^{k(A+1)}$ to $x^{k(A+1)+B-1}$.

Hence we can picture $\frac{Q(x)}{x-1}$ using the chart in FIGURE 1 below. In this chart, the column headings represent the exponents of all monomials which either begin or end a string with the same multiplier. These exponents are arranged in increasing order. The five exponents whose relative positions have some degree of flexibility are shaded, and we will consider all the various possible permutations. (The indicated positions of the five exponents are the leftmost ones for $B-1$ and B, and the rightmost ones for the other three exponents.) The row labels are the indicated multipliers, and the shaded horizontal bars highlight the monomials which contain that multiplier in their coefficient (lighter shadings for positive multipliers, darker for negative multipliers).

Note that the sequence of nonzero coefficients (from smallest powers to largest) begins with the positive coefficient of x^A and continues with four sign changes, culminating with $-x^{kB+A-1} + Ax^{kB+A}$. This pattern is maintained when we include any possible terms between the indicated powers. The coefficient of any monomial between two adjacent columns is the sum of all multipliers which are present in both of the adjacent columns. The resulting coefficients are summarized in FIGURE 2 below.

The only remaining question is whether this pattern persists under all possible permutations of the five unfixed columns. The fact that it does could be established by considering each of the several dozen possibilities. This is unnecessary, however, since we will show that none of the possible permutations alters the pattern of sign changes in the polynomial.

Note that any change in the coefficients of a column resulting from the permutation of a different column is the result of changes in a horizontal bar which begins or ends in the permuted column. Thus, for example, if $B-1$ is greater than or equal to any of the exponents from $kA-1$ to $kA+A$, the net effect on the overtaken columns would be the addition of -1 to their coefficients. This would result from the addition of A and the loss of $(A+1)$ in their total coefficients. Hence these columns would maintain their negative coefficients (or have their coefficients of zero become -1). Since $B-1 > A$ and $B < kA+B-1$, the only other possibly-altered coefficients would be those related to the exponents $B-1$ and B themselves. As long as $B \leq kA$, however, the coefficients of both $B-1$ and B will be zero or negative, and will remain that way, unaffected by any possible permutations of the other three columns. (The cases where $B > kA$ are the most confusing, so we will consider those last.)

	0	$A-1$	A	$B-1$	B	$kA-1$	kA	$kA+A$	$kA+B-1$	$kA+k$	$kA+k+A-1$	$kA+k+B-1$	kB	$kB+A-1$	$kB+A$
$-A$	■	■	■							■	■	■			
A	▒	▒	▒	▒									▒	▒	▒
$-A-1$	■	■	■	■	■	■							■	■	
$A+1$	▒	▒			▒	▒	▒	▒	▒						
$-B$	■	■					■	■							
B	▒	▒	▒							▒	▒				
Totals	0	0	$B-(A+1)$	-1	0	0	$A+1-B$	$A+1-B$	$A+1$	$B-A$	$B-A$	$-A$	-1	-1	A
			pos	neg			neg	neg	pos	pos	pos	neg	neg	neg	pos

Figure 1

Monomials with exponents between	A and $B-1$	B and $kA-1$	kA and $kA+A$	$kA+A$ and $kA+B-1$	$kA+B-1$ and $kA+k$	$kA+k$ and $kA+k+A-1$	$kA+k+A-1$ and $kA+k+B-1$	$kA+k+B-1$ and kB	kB and $kB+A-1$
Coefficient	-1	0	$A+1-B$	$A+1$	0	$B-A$	$-A$	0	-1

Figure 2

A shift of the kB column to the left will result in a change of -1 in the coefficient of any overtaken column. Since $kB > kA$, this will not result in a change of sign for any of the affected coefficients.

A shift of $kA + k$ to the left will add $B - A$ to either of the two possible overtaken columns. This will increase the positive coefficient of $kA + B - 1$, and will increase the negative coefficient of $kA + A$ to either zero or 1. That is, the original coefficient, $A + 1 - B$, will be increased by $B - A$ if the column is overtaken by $kA + k$. It may also change by -1 if $B \geq kA + A$. The overall pattern of four sign changes, however, is still maintained.

Finally, we consider the possible permutations with $B > kA$. If $B \leq kA + A$, the coefficients of $B - 1$ and B would have the same possible values of those of kA and $kA + A$. Otherwise, $B - 1$ and B occupy adjacent columns as indicated in FIGURES 3–5 below. In all of these cases, we see the same pattern of four sign changes, and it is easy to show once again that the pattern is undisturbed by any additional permutations of the other three columns. ■

Some concluding remarks

To highlight the main result, inequality (3), we return to our gambler at the roulette wheel who continually makes \$10 bets in the hope of leaving the casino with \$160. Rather than assuming that his initial fortune is \$100, we consider all possible initial fortunes from \$10 to \$150. In terms of \$10 units, then, our gambler is willing to stake between 1 and 15 units in the hope of reaching 16. For each of these possible starting values, A, the successive columns in FIGURE 6 below show his probability of success and the duration, as well as the probability of success and the duration if he switches to \$5 bets. The last column shows the ratio of the durations: $\frac{D(2A,32,0.47)}{D(A,16,0.47)}$.

Note that the ratio of durations increases with A, as we saw in the proof of inequality (3), since the probability of success on each bet is below $\frac{1}{2}$. For values of A above 12, the ratio exceeds k^2, which is 4, but in all cases, it is well below $2k^2$. To obtain a ratio close to $2k^2$ would require extremely large values of B and k [**2**, p. 186].

In our analysis, we always assumed that the gambler had a fixed initial fortune as well as a fixed goal, and we focused on his one remaining decision: how much to stake on each bet. We then saw how reducing the bet size yielded a decreased probability of success, an increase in the bias, and a corresponding increase in the duration.

Of course, a gambler must first decide what his initial fortune and goal will be, i.e., how much he is willing to lose and how much he hopes to gain. Once the initial fortune and goal are determined, the size of the individual bets can be based on the associated probabilities of success and durations.

In addition, in most casinos, a gambler will have a choice of a variety of games to play, with varying probabilities of success. If he chooses to play craps, for example, his probability of success on each game will be approximately 0.493. FIGURE 7, below, gives the same information as FIGURE 6, adjusted to assume that the bets take place at the craps table.

The probabilities of success in FIGURE 7 are always greater than the corresponding probabilities in FIGURE 6, and are often significantly higher. The duration is often larger as well. If these were gamblers' sole considerations, the craps table would make almost all other bets extremely unattractive. The fact that people frequent all sorts of casino games shows that gamblers' choices of which games to play (and how much to bet) are affected by a wide range of factors. These may include the particular features of the games as well as peripheral factors such as previous winnings and how many

	0	$A-1$	A	$kA-1$	kA	$kA+A$	$B-1$	B	$kA+B-1$	$kA+k$	$kA+k+A-1$	$kA+k+B-1$	kB	$kB+A-1$	$kB+A$
$-A$	■	■	■							■	■	■			
A	▨	▨	▨	▨	▨	▨	▨						▨	▨	▨
$-A-1$	■	■	■	■									■	■	
$A+1$	▨	▨						▨	▨						
$-B$	■	■			■	■									
B	▨	▨	▨							▨	▨				
Totals	0	0	$B-(A+1)$	-1	$A-B$	$A-B$	A	$A+1$	$A+1$	$B-A$	$B-A$	$-A$	-1	-1	A
			pos	neg	neg	neg	pos	pos	pos	pos	pos	neg	neg	neg	pos

Figure 3

	0	$A-1$	A	$kA-1$	kA	$kA+A$	$kA+k$	$B-1$	B	$kA+k+A-1$	$kA+B-1$	$kA+k+B-1$	kB	$kB+A-1$	$kB+A$
$-A$	■	■	■				■	■	■	■	■	■			
A	▨	▨	▨	▨	▨	▨	▨	▨					▨	▨	▨
$-A-1$	■	■	■	■									■	■	
$A+1$	▨	▨							▨	▨	▨				
$-B$	■	■			■	■									
B	▨	▨	▨				▨	▨	▨	▨					
Totals	0	0	$B-(A+1)$	-1	$A-B$	$A-B$	B	B	$B+1$	$B+1$	1	$-A$	-1	-1	A
			pos	neg	neg	neg	pos	pos	pos	pos	pos	neg	neg	neg	pos

Figure 4

	0	$A-1$	A	$kA-1$	kA	$kA+A$	$kA+k$	$kA+k+A-1$	$B-1$	B	$kA+B-1$	$kA+k+B-1$	kB	$kB+A-1$	$kB+A$
$-A$	■	■	■				■	■	■	■	■	■			
A	▨	▨	▨	▨	▨	▨	▨	▨	▨				▨	▨	▨
$-A-1$	■	■	■	■									■	■	
$A+1$	▨	▨								▨	▨				
$-B$	■	■			■	■									
B	▨	▨	▨				▨	▨							
Totals	0	0	$B-(A+1)$	-1	$A-B$	$A-B$	B	B	0	1	1	$-A$	-1	-1	A
			pos	neg	neg	neg	pos	pos		pos	pos	neg	neg	neg	pos

Figure 5

A	$P(A, 16, .47)$	$D(A, 16, .47)$	$P(2A, 32, .47)$	$D(2A, 32, .47)$	DurRatio
1	0.022	10.834	0.006	30.166	2.784
2	0.047	20.924	0.013	59.472	2.842
3	0.074	30.174	0.023	87.685	2.906
4	0.106	38.477	0.035	114.506	2.976
5	0.141	45.713	0.051	139.558	3.053
6	0.181	51.744	0.071	162.362	3.138
7	0.226	56.418	0.096	182.305	3.231
8	0.277	59.561	0.128	198.611	3.335
9	0.334	60.977	0.168	210.292	3.449
10	0.398	60.447	0.220	216.093	3.575
11	0.471	57.721	0.285	214.414	3.715
12	0.553	52.520	0.369	203.227	3.870
13	0.646	44.526	0.475	179.946	4.041
14	0.750	33.385	0.610	141.288	4.232
15	0.867	18.694	0.782	83.077	4.444

Figure 6

A	$P(A, 16, .493)$	$D(A, 16, .493)$	$P(2A, 32, .493)$	$D(2A, 32, .493)$	DurRatio
1	0.050	13.996	0.039	51.939	3.711
2	0.102	26.366	0.081	98.681	3.743
3	0.155	37.064	0.125	139.924	3.775
4	0.209	46.041	0.172	175.344	3.808
5	0.265	53.248	0.222	204.603	3.842
6	0.323	58.634	0.274	227.340	3.877
7	0.382	62.146	0.330	243.174	3.913
8	0.443	63.730	0.388	251.701	3.949
9	0.506	63.332	0.450	252.494	3.987
10	0.571	60.893	0.516	245.101	4.025
11	0.637	56.355	0.586	229.044	4.064
12	0.706	49.657	0.660	203.817	4.104
13	0.776	40.738	0.738	168.883	4.146
14	0.849	29.533	0.820	123.675	4.188
15	0.923	15.976	0.908	67.593	4.231

Figure 7

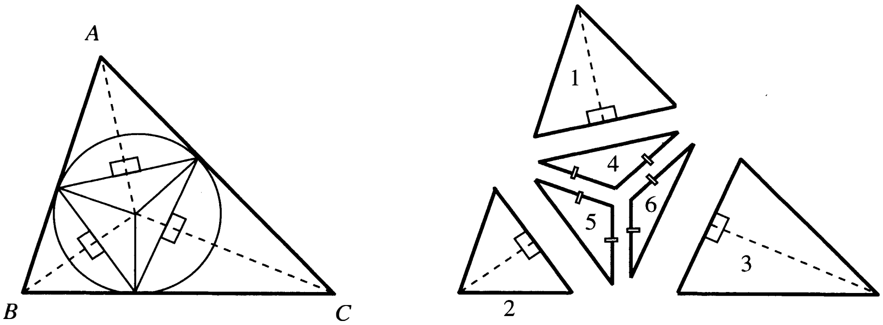
people are watching. Trying to sort out these various factors can be daunting. As one author concluded, “Just what it is that makes one bet more attractive than another is not always clear; what is clear is that the players are trying to do something other than maximize their average winnings.” [**3**, p. 60]

The author would like to thank Ken Ross and the anonymous referees for their many helpful suggestions.

REFERENCES

1. Joseph Bak, The anxious gambler's ruin, this MAGAZINE **74** (2001) 182–193.
2. Joseph Bak, The effect of increased stakes on the duration of play, *Far East J. Math. Sci.* **5** (2002) 173–190.
3. Morton D. Davis, *Game Theory: A Nontechnical Introduction*, Dover Publications, New York, 1983.
4. William Feller, *An Introduction to Probability Theory and Its Applications*, 3rd edition, Wiley, New York, 1968.
5. James D. Harper and Kenneth A. Ross, Stopping Strategies and Gambler's Ruin, this MAGAZINE **78** (2005) 255–268.
6. R. Isaac, Bold play is best: A simple proof, this MAGAZINE **72** (1999) 405–407.
7. Ken Ross, *A Mathematician at the Ballpark: Odds and Probabilities for Baseball Fans*, Pi Press, New York, 2004. Paperback edition, Penguin Plume, New York, 2007.

Proof Without Words: Every Triangle Can Be Subdivided into Six Isosceles Triangles



—ÁNGEL PLAZA
ULPGC, 35017-LAS PALMAS G.C., SPAIN

The Probability of Relatively Prime Polynomials

ARTHUR T. BENJAMIN
Harvey Mudd College
Claremont, CA 91711
benjamin@hmc.edu

CURTIS D. BENNETT
Loyola Marymount University
Los Angeles, CA 90045
cbennett@lmu.edu

Euclid does integers

The Euclidean algorithm for finding greatest common divisors, one of the oldest algorithms in the world, is also one of the most versatile. When applied to integers, Euclid's theorem can be stated as:

$$\text{If } a = qb + r \text{ then } \gcd(a, b) = \gcd(b, r).$$

The one sentence proof is that any number that divides a and b must also divide b and r (since $r = a - qb$) and vice versa; hence, the pairs (a, b) and (b, r) have the exact same set of common divisors. What turns this theorem into an algorithm is that if $b > 0$, then we can find a unique quotient q so that $0 \le r < b$, allowing us to repeat the process with the second coordinate decreasing to zero. That is, if $\gcd(a, b) = c$, then Euclid's algorithm will look like

$$\gcd(a, b) = \gcd(b, r) = \cdots = \gcd(c, 0) = c.$$

For example,

$$\gcd(422, 138) = \gcd(138, 8) = \gcd(8, 2) = \gcd(2, 0) = 2.$$

Better yet, we can keep track of the integer quotients at each step (for example, $q_1 = \lfloor\frac{422}{138}\rfloor = 3$) and remove the gcd label so the above calculation looks like

$$(422, 138) \xrightarrow{q_1=3} (138, 8) \xrightarrow{q_2=17} (8, 2) \xrightarrow{q_3=4} (2, 0) = 2.$$

Now in addition to working from left to right, we can run the algorithm from right to left by holding on to the quotients. That is, given the quotients $q_1 = 3, q_2 = 17, q_3 = 4$, we can start with $(2, 0)$ and (from $q_3 = 4$) derive that it came from $(8, 2)$, which (from $q_2 = 17$) came from $(138, 8)$ which (from $q_1 = 3$) came from $(422, 138)$. In other words, we can run Euclid's algorithm backwards to obtain "dilcuE's algorithm:" $(b, r) \xrightarrow{q} (qb + r, b)$. For example,

$$2 = (2, 0) \xrightarrow{q_3=4} (8, 2) \xrightarrow{q_2=17} (138, 8) \xrightarrow{q_1=3} (422, 138).$$

As a practice problem, let's find the unique pair of relatively prime integers (i.e., whose greatest common divisor is one) for which Euclid's algorithm produces quotients $q_1 = 2, q_2 = 3, q_3 = 5, q_4 = 8$. By dilcuE's algorithm, we have

$$1 = (1, 0) \xrightarrow{q_4=8} (8, 1) \xrightarrow{q_3=5} (41, 8) \xrightarrow{q_2=3} (131, 41) \xrightarrow{q_1=2} (303, 131).$$

Euclid does polynomials

What makes Euclid's algorithm so versatile is that it can also be applied to objects other than integers. For example, given two polynomials $a(x)$ and $b(x)$ with rational coefficients, we define their greatest common divisor $c(x)$ to be the monic polynomial of greatest degree for which $c(x)$ divides $a(x)$ and $b(x)$. Here, Euclid's theorem says

$$\text{If } a(x) = q(x)b(x) + r(x), \text{ then } \gcd(a(x), b(x)) = \gcd(b(x), r(x)).$$

(The proof is exactly as before, except we insert (x) after every term.) To turn this theorem into an algorithm, we note that if the degree of $b(x)$ is at least one, then by the division algorithm for polynomials, we can always find unique quotient polynomial $q(x)$ so that the degree of $r(x)$ is strictly less than the degree of $b(x)$; hence Euclid's algorithm is guaranteed to terminate with an ordered pair $(kc(x), z)$ for some rational numbers $k \neq 0$ and z, and some monic polynomial $c(x)$ of degree at least one. If $z = 0$, then $\gcd(a(x), b(x)) = \gcd(kc(x), 0) = c(x)$; If $z \neq 0$, then $a(x)$ and $b(x)$ are relatively prime. For example,

$$\begin{aligned}(x^3 + 4x^2 + 5x + 2, 2x^2 - 6x - 8) &\xrightarrow{\frac{1}{2}x+\frac{7}{2}} (2x^2 - 6x - 8, 30x + 30)\\ &\xrightarrow{\frac{1}{15}x-\frac{4}{15}} (30x + 30, 0)\end{aligned}$$

where, for example, the first step indicates that

$$x^3 + 4x^2 + 5x + 2 = \left(\frac{1}{2}x + \frac{7}{2}\right)(2x^2 - 6x - 8) + (30x + 30).$$

Since $\gcd(30x + 30, 0) = 30(x + 1)$, it follows that $\gcd(x^3 + 4x^2 + 5x + 2, 2x^2 - 6x - 8) = x + 1$. On the other hand, adding 15 to the constant term of $a(x)$ results in

$$\begin{aligned}(x^3 + 4x^2 + 5x + 17, 2x^2 - 6x - 8) &\xrightarrow{\frac{1}{2}x+\frac{7}{2}} (2x^2 - 6x - 8, 30x + 45)\\ &\xrightarrow{\frac{1}{15}x-\frac{3}{10}} \left(30x + 45, \frac{11}{2}\right),\end{aligned}$$

so the original polynomials are relatively prime, since the constant term, 11/2, is not zero. As with the integers, we can reverse this procedure starting with the final pair of polynomials, and backtracking through the quotients to obtain the original pair. Notice that the Euclidean Algorithm works here, because the coefficients of all of the polynomials, including the quotient polynomials, are allowed to be rational. If all coefficients were restricted to be integers, we could not apply the Euclidean algorithm.

Things become more interesting when we look at the set $\mathbb{Z}_2[x]$ where all of the coefficients come from the set $\{0, 1\}$, and all of the coefficient arithmetic is performed modulo 2. For example, in $\mathbb{Z}_2[x]$, $(x + 1)^3 = x^3 + 3x^2 + 3x + 1 = x^3 + x^2 + x + 1$, and $(x + 1)(x^2 + x + 1) = x^3 + 2x^2 + 2x + 1 = x^3 + 1$. For the exact same reason as in the polynomial case, we can perform the Euclidean algorithm on polynomials from $\mathbb{Z}_2[x]$ too. (In fact, it's easier in $\mathbb{Z}_2[x]$ because all nonzero polynomials are monic, and subtraction is the same as addition.) For instance,

$$\begin{aligned}(x^3 + x^2 + x + 1, x^3 + 1) &\xrightarrow{q_1=1} (x^3 + 1, x^2 + x) \xrightarrow{q_2=x+1} (x^2 + x, x + 1)\\ &\xrightarrow{q_3=x} (x + 1, 0).\end{aligned}$$

Thus $\gcd(x^3 + x^2 + x + 1, x^3 + 1) = x + 1$, which agrees with our earlier calculations. Again, if we hold on to the quotients, we can reverse the process through dilcuE's algorithm.

Euclid does 1-to-1 correspondences

We now are ready ask the main question of this paper. *If we choose two polynomials at random from $\mathbb{Z}_2[x]$, then what is the chance that they are relatively prime?* In FIGURE 1, we have a 16 by 16 matrix representing every pair of polynomials of degree 3 or lower. (Notice that the number of polynomials of degree n is 2^n since the coefficient of n must be one, but every subsequent coefficient can be one or zero. Likewise the number of polynomials of degree less than n is also 2^n.) Every dark square represents an ordered pair of polynomials that are relatively prime. Every light square represents an ordered pair of polynomials that are not relatively prime. We have drawn thick lines separating polynomials of different degrees. Notice that except for the four squares in the lower-left corner representing the ordered pairs of constant polynomials, all other thick rectangles have an equal number of dark and light squares. As the next theorem shows, this is not a coincidence.

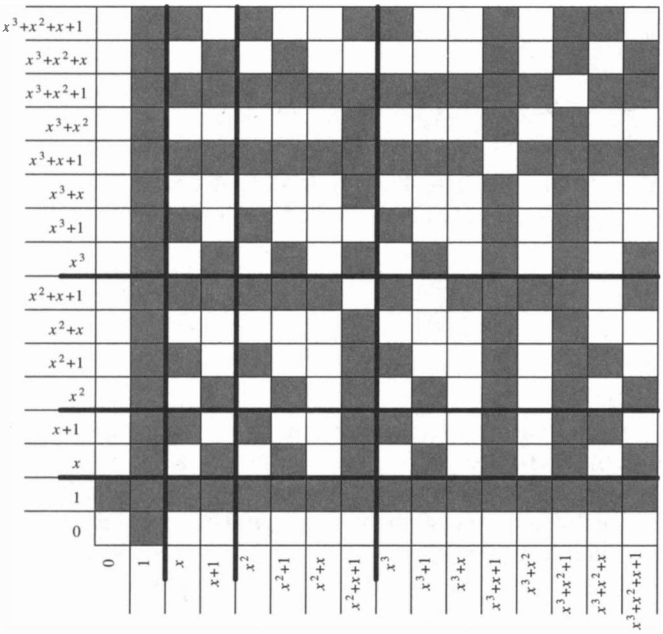


Figure 1 In every solid rectangle, except for the one in the lower left corner, half of the polynomials in $Z_2[x]$ are relatively prime (as represented by the dark squares). But how do you pair up the dark squares with the light squares?

THEOREM 1. *Let $a(x)$ and $b(x)$ be randomly chosen (i.e., uniformly and independently) from the set of polynomials in $\mathbb{Z}_2[x]$ of degree m and n, respectively, where m and n are not both zero. Then the probability that $a(x)$ and $b(x)$ are relatively prime is $1/2$.*

Proof. Without loss of generality, we assume that $m \geq n$. Our goal is to show that every relatively prime pair $(a(x), b(x))$ can be matched up with a non-relatively prime pair $(a_1(x), b_1(x))$, where a_1 and b_1 have the same degree as a and b, respectively.

If $n = 0$, then we match the relatively prime pair $(a(x), 1)$ with the non-relatively prime pair $(a(x), 0)$. Now suppose that $n \geq 1$ and let $(a(x), b(x))$ be a non-relatively prime pair. Then applying Euclid's algorithm gives us a unique sequence

$$(a(x), b(x)) \xrightarrow{q_1} (b(x), r_1(x)) \xrightarrow{q_2} (r_1(x), r_2(x)) \xrightarrow{q_3} \cdots \xrightarrow{q_t} (c(x), 0)$$

where $c(x)$, a polynomial of degree at least one, is the greatest common divisor. Starting with the relatively prime pair $(c(x), 1)$ and using the same quotient polynomials, $q_t, \ldots, q_1$, we can reverse Euclid's algorithm to produce a relatively prime pair $(a_1(x), b_1(x))$, which have the same degrees as $(a(x), b(x))$. ■

For example, when $a(x) = x^3 + x^2 + x + 1$ and $b(x) = x^3 + 1$, the Euclidean algorithm produces the greatest common divisor $c(x) = x + 1$.

$$(x^3 + x^2 + x + 1, x^3 + 1) \xrightarrow{q_1=1} (x^3 + 1, x^2 + x) \xrightarrow{q_2=x+1} (x^2 + x, x + 1) \xrightarrow{q_3=x} (x + 1, 0).$$

Now running the Euclidean algorithm backwards from the relatively prime pair $(x + 1, 1)$, with the same quotients

$$(x + 1, 1) \xrightarrow{q_3=x} (x^2 + x + 1, x + 1) \xrightarrow{q_2=x+1} (x^3 + x, x^2 + x + 1) \xrightarrow{q_1=1} (x^3 + x^2 + 1, x^3 + x)$$

we obtain $(a_1(x), b_1(x)) = (x^3 + x^2 + 1, x^3 + x)$, which is a relatively prime pair since Euclid's algorithm reduces it to $(x + 1, 1)$.

COROLLARY 2. *If $a(x)$ and $b(x)$ are randomly chosen from the set of polynomials in $\mathbb{Z}_2[x]$ of degree less than n, then the probability that they are relatively prime is* $\frac{1}{2} + \frac{1}{4^n}$.

Proof. There are 2^n polynomials of degree less than n and therefore 4^n ordered pairs of polynomials $(a(x), b(x))$. Three of the four constant pairs $(0, 1)$, $(1, 0)$, $(1, 1)$ are relatively prime. From Theorem 1, among the remaining $4^n - 4$ pairs, exactly half of them are relatively prime. Thus the probability of a relatively prime pair is

$$\frac{3 + \frac{1}{2}(4^n - 4)}{4^n} = \frac{1}{2} + \frac{1}{4^n}. \qquad ■$$

In a recent paper [**1**], Corteel, Savage, Wilf, and Zeilberger prove a special case of Theorem 1 (under the assumption that $m = n$) by an elegant generating function argument, but ask for a "nice simple bijection that proves this result." We hope that our Euclidean bijection is nice and simple enough. We note that Reifegerste [**2**] also found a bijection using "resultant matrices" that was essentially the Euclidean algorithm in heavy disguise. As we'll see, the Euclidean bijection leads to interesting generalizations of Theorem 1 (some of which also appear in [**1**]).

Euclid does 1-to-many correspondences

What if the coefficients of our polynomial come from $\mathbb{Z}_3$ instead of $\mathbb{Z}_2$? A picture similar to FIGURE 1 would show that, except for the lower left corner of constant polynomial pairs, in every thick rectangle, precisely two thirds of all polynomial pairs are relatively prime. (The polynomials are listed (from left to right and from bottom to top) in lexicographic order. For example, the first nine columns correspond to the polynomials: 0, 1, 2, $x, x+1, x+2, 2x, 2x+1$, and $2x+2$.)

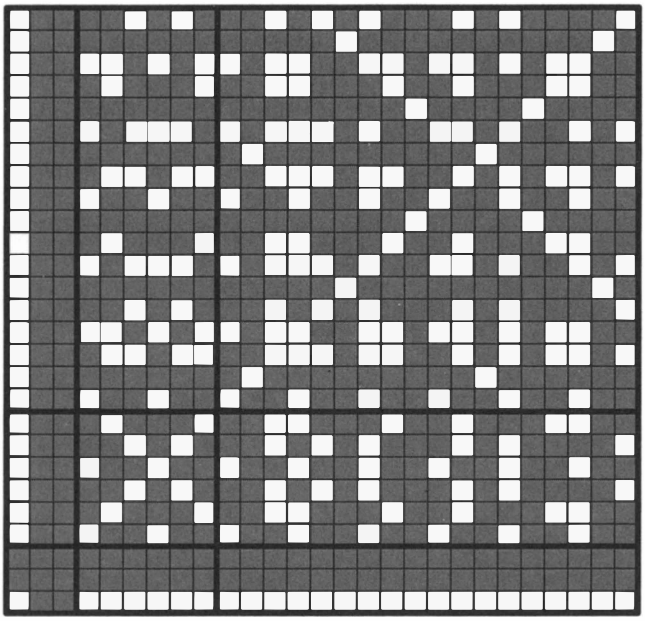


Figure 2 Every solid rectangle, except for the one in the lower left corner, has twice as many dark squares (representing relatively prime polynomials in $Z_3[x]$) as light squares. But how do you assign two dark squares to each light square?

In general, if our coefficients come from a finite field F of q elements (for instance, the set $F = \mathbb{Z}_q$, when q is prime) then we have the following generalization.

THEOREM 3. *Let F be a finite field of q elements, and let $a(x)$ and $b(x)$ be randomly chosen from the set of polynomials in $F[x]$ of degree m and n, respectively, where m and n are not both zero. Then the probability that $a(x)$ and $b(x)$ are relatively prime is* $1 - 1/q$.

Proof. To prove this, we show that for every non-relatively prime pair $(a(x), b(x))$, there are $q-1$ relatively prime pairs; hence the proportion of non-relatively prime pairs is $1/q$. If $n = 0$, then the non-relatively prime pair $(a(x), 0)$ is matched up with the $q-1$ relatively prime pairs $(a(x), z)$ where z is a nonzero element of F. (Note that $\gcd(2x, 2) = 1$, not 2, since 2 divides 1, and we insist that the greatest common divisor be monic.)

When $n \geq 1$, then we can apply Euclid's algorithm to $(a(x), b(x))$, producing a unique set of quotient and remainder polynomials,

$$(a(x), b(x)) \xrightarrow{q_1} (b(x), r_1(x)) \xrightarrow{q_2} (r_1(x), r_2(x)) \xrightarrow{q_3} \cdots \xrightarrow{q_s} (kc(x), z)$$

where $c(x)$ is a monic polynomial, and $k \neq 0$ and z are constants in F. If $z = 0$, then $a(x)$ and $b(x)$ have greatest common divisor $c(x)$; otherwise they are relatively prime.

Now suppose $n \geq 1$, and let $(a(x), b(x))$ be a non-relatively prime pair. Then Euclid's algorithm produces a unique set of quotient and remainder polynomials

$$(a(x), b(x)) \xrightarrow{q_1} (b(x), r_1(x)) \xrightarrow{q_2} (r_1(x), r_2(x)) \xrightarrow{q_3} \cdots \xrightarrow{q_s} (kc(x), 0)$$

where $c(x)$ is a monic polynomial of degree at least one, and k is a nonzero constant in F. Starting with $(kc(x), 0)$ and the quotients $q_s, \ldots, q_1$, we can reverse Euclid's algorithm to reconstruct $(a(x), b(x))$. Likewise, for each nonzero constant z in F, we can start with the relatively prime pair $(kc(x), z)$ and the same quotient polynomials $q_s, \ldots, q_1$ to produce a relatively prime pair $(a_z(x), b_z(x))$, which have the same degree as $a(x)$ and $b(x)$ respectively. Since there are $q - 1$ choices for z we have established the desired 1-to-$(q - 1)$ correspondence. ■

For example, suppose that $q = 3$, $F = \mathbb{Z}_3$, $m = 5$, $n = 3$, and consider the pair $(x^5 + x, x^3 + x + 1)$. By Euclid's algorithm,

$$\begin{aligned}(x^5 + x, x^3 + x + 1) &\xrightarrow{q_1 = x^2+2} (x^3 + x + 1, 2x^2 + 2x + 1)\\ &\xrightarrow{q_2 = 2x+1} (2x^2 + 2x + 1, 0)\end{aligned}$$

and so the pair is not relatively prime. Then starting with the relatively prime pairs $(2x^2 + 2x + 1, 1)$ and $(2x^2 + 2x + 1, 2)$, dilcuE's algorithm gives us two more relatively prime polynomials of degree 5 and 3, namely

$$\begin{aligned}(2x^2 + 2x + 1, 1) &\xrightarrow{q_2 = 2x+1} (x^3 + x + 2, 2x^2 + 2x + 1)\\ &\xrightarrow{q_1 = x^2+2} (x^5 + x^2 + x + 2, x^3 + x + 2)\end{aligned}$$

and

$$\begin{aligned}(2x^2 + 2x + 1, 2) &\xrightarrow{q_2 = 2x+1} (x^3 + x, 2x^2 + 2x + 1)\\ &\xrightarrow{q_1 = x^2+2} (x^5 + 2x^2 + x + 1, x^3 + x).\end{aligned}$$

The number of pairs of polynomials of degree less than n is q^{2n}. Among the q^2 constant pairs, all of them are relatively prime except for $(0, 0)$. (Yes, in $F[x]$, $\gcd(2, 2) = 1$.) Among the others, exactly $1/q$th of them are not relatively prime. Thus the number of nonrelatively prime pairs is $1 + \frac{1}{q}(q^{2n} - q^2)$. Dividing by q^{2n}, the probability of not being relatively prime is $\frac{1}{q} - \frac{q-1}{q^{2n}}$. Consequently, we have

COROLLARY 4. *Let F be a finite field with q elements. If $a(x)$ and $b(x)$ are randomly chosen from the set of polynomials in $F[x]$ of degree less than n, then the probability that they are relatively prime is $1 - \frac{1}{q} + \frac{q-1}{q^{2n}}$.*

Euclid does m-tuples

How about the probability that a random *triple* of polynomials in $F[x]$ is relatively prime? We claim that the probability that three polynomials $a_1(x), a_2(x), a_3(x)$ (not all constant) in $F[x]$ are not relatively prime is $1/q^2$. Without loss of generality, we'll assume that $a_1(x)$, $a_2(x)$, and $a_3(x)$ are chosen randomly from among polynomials of degree $d_1 \geq d_2 \geq d_3 \geq 0$, respectively, where $d_1 \geq 1$. The polynomials will not be relatively prime if and only if $a_1(x)$ and $a_2(x)$ are not relatively prime and

$\gcd(a_1(x), a_2(x))$ and $a_3(x)$ are not relatively prime. By Theorem 3, the probability that $a_1(x)$ and $a_2(x)$ are not relatively prime is $1/q$, and their gcd is a polynomial $c(x)$ with some degree $d \geq 1$. Given that $c(x)$ has degree d, Euclid's algorithm can be used (although we shall skip this detail) to show that it is equally likely to be any of the q^d monic polynomials of degree d. Applying Theorem 3 again, we have the probability that $c(x)$ and $a_3(x)$ are not relatively prime is also $1/q$. (Note that $a_3(x)$ is chosen independently of $c(x)$.) Multiplying the probabilities together, the probability that $a_1(x), a_2(x), a_3(x)$ are not relatively prime is $1/q^2$, and hence the probability that they are relatively prime is $1 - 1/q^2$.

Using induction, this argument can be extended to show

COROLLARY 5. *Let $(d_1, d_2, \ldots, d_m)$ be an ordered m-tuple of nonnegative integers (not all zero) and for $1 \leq i \leq m$, let $a_i(x)$ be a randomly chosen polynomial of degree d_i over $F[x]$, where F is a finite field with q elements. Then the probability that $a_1(x), a_2(x), \ldots, a_m(x)$ are relatively prime is $1 - \frac{1}{q^{m-1}}$.*

Finally, by a counting argument similar to the ones before, our final corollary is obtained.

COROLLARY 6. *If $a_1(x), \ldots, a_m(x)$ are randomly chosen polynomials of degree less than n in $F[x]$, where the field F has q elements, then the probability that they are relatively prime is $1 - 1/q^{m-1} + (q-1)/q^{mn}$.*

Using a similar argument, one can show that the set of pairs of *monic* polynomials of $\mathbb{Z}[x]$ can be partitioned into disjoint infinite sets, such that each set contains at most one pair that is not relatively prime. Thus, if a pair of monic polynomials is chosen at random (in an appropriate sense) from $\mathbb{Z}[x]$, then the probability that they are relatively prime is 1.

Acknowledgment. We thank Ed Scheinerman for bringing this problem to our attention and the referee for many valuable suggestions.

REFERENCES

1. S. Corteel, C. Savage, H. Wilf, D. Zeilberger, A Pentagonal Number Sieve, *Journal of Combinatorial Theory, Series A*, **82** (1998) No. 2, 186–192.
2. A. Reifegerste, On an Involution Concerning Pairs of Polynomials in $\mathbb{F}_2$, *Journal of Combinatorial Theory, Series A*, **90** (2000) 216–220.

NOTES

Fitting One Right Triangle in Another

CHARLES H. JEPSEN
Grinnell College
Grinnell, IA 50112
jepsen@math.grinnell.edu

VALERIA VULPE
Grinnell College
Grinnell, IA 50112

In a survey article in this MAGAZINE [2], Wetzel posed several open questions about fitting one plane figure in another. When does a triangle fit in an equilateral triangle? When does a triangle fit in a rectangle? When does one right triangle fit in another right triangle? In this note, we answer the last question. This result constitutes part of the output of an undergraduate summer research project by the second author under the supervision of the first author.

We find conditions on a, b, c, d so that a right triangle with legs a, b (the initial triangle) fits in a right triangle with legs c, d (the target triangle). We may assume $a \le b$ and $c \le d$. First note some necessary conditions for a fit:

- (diameter condition) $\sqrt{a^2+b^2} \le \sqrt{c^2+d^2}$;
- (thickness condition) $ab/\sqrt{a^2+b^2} \le cd/\sqrt{c^2+d^2}$;
- (area condition) $ab \le cd$.

(The third condition is a consequence of the first and second together.) Easy examples show that these conditions are not sufficient. To find necessary and sufficient conditions, we identify four cases:

(i) $a \le c$ and $b \le d$;
(ii) $a \ge c$ and $b \ge d$, but not both equalities hold;
(iii) $a < c$ and $b > d$;
(iv) $a > c$ and $b < d$.

Case (i). Here the initial triangle fits in the target triangle by aligning their right angles.

Case (ii). Here the area of the initial triangle exceeds that of the target triangle and no fit exists.

The two interesting cases remain. We may restrict our attention to the largest triangle of a given shape that fits in the target (for if a triangle fits in the target, any smaller triangle of the same shape will fit). The following result of Sullivan [1] simplifies the remaining cases:

> Among the polygons inside a triangle T which are similar to a given polygon P, any largest one has two of its vertices along the same edge of T (and some vertex on each edge of T).

Case (iii): $a < c \leq d < b$. First note that for a fit to exist, we must have $b < \sqrt{c^2 + d^2}$. Next note that because b is longer than both c and d, we cannot get a fit by placing an edge of the initial triangle along a leg of the target triangle. Hence to get a fit, we place an edge of the initial triangle along the hypotenuse of the target. Now observe that an initial triangle with a leg placed along the hypotenuse can be reflected so that its hypotenuse lies along the hypotenuse and the reflected triangle remains inside the target. See FIGURE 1.

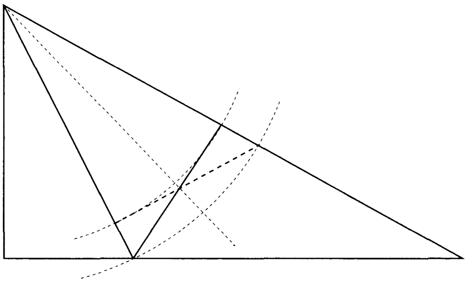


Figure 1

Hence we may assume that the hypotenuses of the triangles are aligned. This can be done in two ways as shown in FIGURE 2.

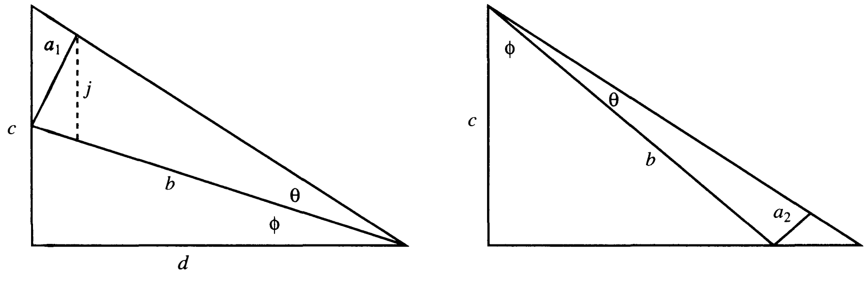


Figure 2

With b, c, d as given, we find the values of a_1 and a_2. Then a fit exists for any initial triangle whose leg a is at most the longer of a_1 and a_2.

In the triangle on the left, from the Law of Tangents,

$$\frac{a_1}{b} = \tan\theta = \tan\bigl((\theta+\phi)-\phi\bigr) = \frac{\tan(\theta+\phi)-\tan\phi}{1+\tan(\theta+\phi)\tan\phi}$$

$$= \frac{\frac{c}{d} - \frac{\sqrt{b^2-d^2}}{d}}{1+\frac{c}{d}\cdot\frac{\sqrt{b^2-d^2}}{d}} = \frac{d(c-\sqrt{b^2-d^2})}{d^2+c\sqrt{b^2-d^2}}.$$

Therefore,

$$a_1 = \frac{bd(c - \sqrt{b^2 - d^2})}{d^2 + c\sqrt{b^2 - d^2}}.$$

A similar computation gives

$$a_2 = \frac{bc(d - \sqrt{b^2 - c^2})}{c^2 + d\sqrt{b^2 - c^2}}.$$

(Or note that the difference between the two triangles is that the roles of c and d are interchanged.)

To compare a_1 and a_2, we compare numerators and denominators separately. To see that the numerator of a_2 is at most the numerator of a_1, observe that $d \geq c$ and $c^2 + d^2 > b^2$ imply $0 \leq (d^2 - c^2)(d^2 + c^2 - b^2) = d^4 - b^2d^2 + b^2c^2 - c^4$. Thus $\left(d\sqrt{b^2 - d^2}\right)^2 \leq \left(c\sqrt{b^2 - c^2}\right)^2$ implies

$$bc(d - \sqrt{b^2 - c^2}) \leq bd(c - \sqrt{b^2 - d^2}).$$

Next compare denominators:

$$\begin{aligned} 0 &\leq \left(d\sqrt{b^2 - d^2} - c\sqrt{b^2 - c^2}\right)^2 \\ &= d^2(b^2 - d^2) - 2cd\sqrt{b^2 - c^2}\sqrt{b^2 - d^2} + c^2(b^2 - c^2). \end{aligned}$$

Add $(d^2 - c^2)^2$ to get

$$\left(d^2 - c^2\right)^2 \leq \left(d\sqrt{b^2 - c^2} - c\sqrt{b^2 - d^2}\right)^2,$$

so

$$d^2 + c\sqrt{b^2 - d^2} \leq c^2 + d\sqrt{b^2 - c^2}.$$

Hence $a_2 \leq a_1$. Thus a fit exists if

$$b < \sqrt{c^2 + d^2} \quad \text{and} \quad a \leq \frac{bd(c - \sqrt{b^2 - d^2})}{d^2 + c\sqrt{b^2 - d^2}}.$$

The set of conditions that allow a fit in this case can be simplified by noting that the second inequality implies the first inequality (a must be positive). Also, the condition $a < c$ is not used in the derivation and is, in fact, a consequence. From the triangle on the left in FIGURE 2, we see $a \leq a_1 < j < c$.

Case (iv): $c < a \leq b < d$. Again we may assume the initial triangle fits in the target triangle so that their hypotenuses are aligned. And again there are two ways to do this as shown in FIGURE 3.

With a, b, c as given, we find d_1 and d_2. Then any right triangle whose leg d is at least the shorter of d_1 and d_2 will allow a fit. In the left triangle, using the Law of Tangents as above, we get

$$\frac{a}{b} = \frac{c(d_1 - \sqrt{b^2 - c^2})}{c^2 + d_1\sqrt{b^2 - c^2}}.$$

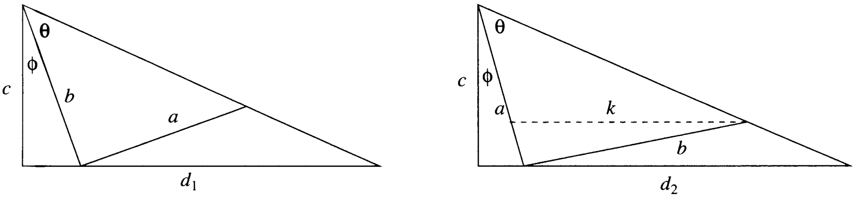


Figure 3

Solve for d_1 to get

$$d_1 = \frac{c(ac + b\sqrt{b^2 - c^2})}{bc - a\sqrt{b^2 - c^2}}.$$

Similarly, by interchanging the roles of a and b,

$$d_2 = \frac{c(bc + a\sqrt{a^2 - c^2})}{ac - b\sqrt{a^2 - c^2}}.$$

To compare d_1 and d_2, we rationalize the denominators:

$$d_1 = \frac{c(ab^3 + c(a^2 + b^2)\sqrt{b^2 - c^2})}{a^2c^2 + b^2c^2 - a^2b^2}, \quad d_2 = \frac{c(a^3b + c(a^2 + b^2)\sqrt{a^2 - c^2})}{a^2c^2 + b^2c^2 - a^2b^2}.$$

Now we need only compare numerators. Observe that $b \geq a$ implies

$$0 \leq ab(b^2 - a^2) + c(a^2 + b^2)\left(\sqrt{b^2 - c^2} - \sqrt{a^2 - c^2}\right).$$

Thus

$$c(a^3b + c(a^2 + b^2)\sqrt{a^2 - c^2}) \leq c(ab^3 + c(a^2 + b^2)\sqrt{b^2 - c^2}).$$

Hence $d_2 \leq d_1$. Thus a fit exists if

$$d \geq \frac{c(bc + a\sqrt{a^2 - c^2})}{ac - b\sqrt{a^2 - c^2}}.$$

If we solve this inequality for b, we get

$$b \leq \frac{ac(d - \sqrt{a^2 - c^2})}{c^2 + d\sqrt{a^2 - c^2}}.$$

As before, the inequality $b < d$, not used in the derivation, is a consequence (from the triangle on the right in FIGURE 3, we have $b < k < d_2 \leq d$).

Putting everything together, we can state our final result (where we choose to reverse the order of the last two cases):

THEOREM. *A right triangle with legs a, b ($a \leq b$) fits in a right triangle with legs c, d ($c \leq d$) if and only if*

(1) *$a \leq c$ and $b \leq d$, or*

(2) *$a > c$ and $b \leq \dfrac{ac(d - \sqrt{a^2 - c^2})}{c^2 + d\sqrt{a^2 - c^2}}$, or*

(3) *$b > d$ and $a \leq \dfrac{bd(c - \sqrt{b^2 - d^2})}{d^2 + c\sqrt{b^2 - d^2}}$*

As an example of (2), we can take $a = 5, b = 6, c = 4, d = 78$; for (3), take $a = 18$, $b = 135$, $c = 106$, $d = 108$. In each case, the second inequality becomes an equality. See FIGURE 4.

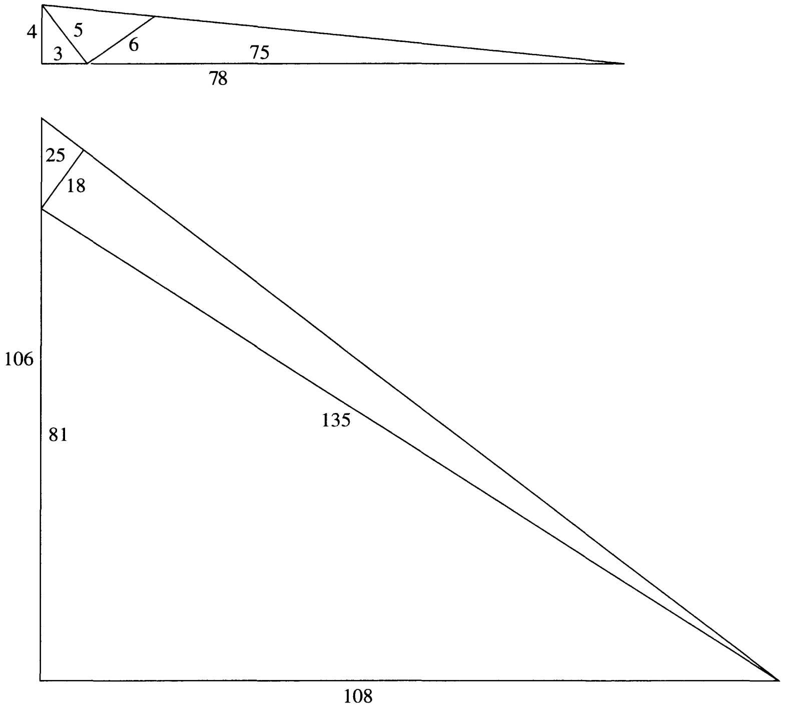


Figure 4

REFERENCES

1. J. M. Sullivan, Polygon in Triangle: Generalizing a Theorem of Post. (Preprint. `http://torus.math.uiuc.edu/jms/Papers/post.pdf`.)
2. J. E. Wetzel, Fits and Covers, this MAGAZINE **76** (2003) 349–363.

Determinants of Matrices over the Integers Modulo *m*

JODY M. LOCKHART
United States Naval Academy
Annapolis, Maryland 21402
jml@usna.edu

WILLIAM P. WARDLAW
United States Naval Academy
Annapolis, Maryland 21402
wpw@usna.edu

As is proven in any elementary course on linear algebra, a square matrix over the real numbers is invertible if and only if its determinant is nonzero. For students who are familiar with $\mathbb{Z}_m$, the ring of integers modulo m, one can easily prove that a square matrix over $\mathbb{Z}_m$ is invertible if and only if its determinant is invertible. In this paper, we discuss determinants of matrices over the ring $\mathbb{Z}_m$ for integers $m \geq 2$. We derive a formula for the number of matrices over $\mathbb{Z}_m$ with a given determinant. Although this formula follows from a result of Richard P. Brent and Brendan D. McKay [1], it was proven independently by the authors and the presentation given below is accessible to advanced undergraduates. Since our proof uses many topics encountered in undergraduate mathematics courses, including modular arithmetic, induction, groups, rings and counting, this paper would be a good basis for a capstone project.

We start with some notation. As usual, let $M_n(\mathbb{Z}_m)$ be the ring of $n \times n$ matrices over $\mathbb{Z}_m$, where $\mathbb{Z}_m$ is the ring of integers modulo m, and let $x \bmod m$ denote the equivalence class representative of x modulo m in $\{0, \dots, m-1\}$. So, for example,

$$\begin{bmatrix} 3 \bmod 5 & 9 \bmod 5 & -1 \bmod 5 \\ 0 \bmod 5 & 4 \bmod 5 & -3 \bmod 5 \\ 6 \bmod 5 & 2 \bmod 5 & 5 \bmod 5 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 4 \\ 0 & 4 & 2 \\ 1 & 2 & 0 \end{bmatrix}$$

is an element of $M_3(\mathbb{Z}_5)$. Also as usual, the matrix $[c_{ij}]$ is the $n \times n$ matrix with c_{ij} in the ith row and jth column. Further, let $D_n(m, k) = \{A \in M_n(\mathbb{Z}_m) \colon \det A \equiv k (\bmod m)\}$ and $d_n(m, k) = |D_n(m, k)|$. For example, the matrix above is an element of $D_3(5, 0)$.

The goal of this paper is to demonstrate a recursive formula for $d_n(m, k)$. Using this formula, we will be able to see, for example, that the number of 5×5 matrices over $\mathbb{Z}_{27}$ of determinant 0 is

$$33{,}193{,}792{,}816{,}322{,}871{,}923{,}020{,}442{,}669{,}672{,}187$$

and the number of 3×3 matrices over $\mathbb{Z}_{100}$ of determinant 35 is

$$7{,}749{,}504{,}000{,}000{,}000.$$

Recall the following facts:

(i) If a and b are relatively prime integers, then the mapping $f : \mathbb{Z}_{ab} \to \mathbb{Z}_a \times \mathbb{Z}_b$ defined by $f(x) = (x \bmod a, x \bmod b)$ is a bijection.

(ii) If m is an integer and if a is relatively prime to m, then a has a multiplicative inverse in $\mathbb{Z}_m$.

(iii) If R and S are rings and if $f : R \to S$ is a surjective ring homomorphism, then $|R| = |\ker f||S|$.

For readers not familiar with these facts, their proofs are relatively easy. (Hints: (i) Since $|\mathbb{Z}_{ab}| = |\mathbb{Z}_a \times \mathbb{Z}_b|$, it suffices to show that f is injective; (ii) use the Euclidean algorithm; (iii) follows immediately from the Fundamental Theorem of Ring Homomorphisms.)

Our first result allows us to consider $d_n(m, k)$ just when m is a power of a prime.

LEMMA 1. (MULTIPLICATIVITY OF d_n) *If a and b are relatively prime integers, then $d_n(ab, k) = d_n(a, k) \cdot d_n(b, k)$.*

Once this lemma is established, we will know, for example, that

$$d_4(72, 6) = d_4(3^2, 6) \cdot d_4(2^3, 6).$$

Proof. Define the map $\gamma : M_n(\mathbb{Z}_{ab}) \to M_n(\mathbb{Z}_a) \times M_n(\mathbb{Z}_b)$ by $\gamma(A) = (\alpha(A), \beta(A))$ where $\alpha([k_{ij}]) = [k_{ij} \bmod a]$ and $\beta([k_{ij}]) = [k_{ij} \bmod b]$. Since a and b are relatively prime, γ is a bijection (using (i) above). We need to show that $|D_n(ab,k)| = |D_n(a,k)||D_n(b,k)|$. Since γ is injective, it is enough to show that γ maps $D_n(ab,k)$ onto $D_n(a,k) \times D_n(b,k)$. Suppose that $M \in D_n(ab,k)$. Since $\det M \equiv k(\operatorname{mod} ab)$ implies $\det(\alpha(M)) \equiv k(\operatorname{mod} a)$ and $\det(\beta(M)) \equiv k(\operatorname{mod} b)$, it follows that $\gamma(M) \in D_n(a,k) \times D_n(b,k)$. On the other hand, suppose $(A,B) \in D_n(a,k) \times D_n(b,k)$. Since $\gamma : M_n(\mathbb{Z}_{ab}) \to M_n(\mathbb{Z}_a) \times M_n(\mathbb{Z}_b)$ is a surjection, there is a $C \in M_n(\mathbb{Z}_{ab})$ such that $\gamma(C) = (A,B)$. Since $A = \alpha(C)$, $B = \beta(C)$, $\det A \equiv k(\operatorname{mod} a)$ and $\det B \equiv k(\operatorname{mod} b)$, it follows that $\det(C) \equiv k(\operatorname{mod} a)$ and $\det(C) \equiv k(\operatorname{mod} b)$. Since a and b are relatively prime, $\det(C) \equiv k(\operatorname{mod} ab)$ and so $C \in D_n(ab,k)$, completing the proof. ■

The next lemma will allow us to compute $d_n(p^r, \ell)$ for p prime and for all $\ell \geq 1$ if we know $d_n(p^r, p^k)$ for $k \geq 0$.

LEMMA 2. *Let p be a prime number and suppose that a is an integer not divisible by p. Let r and s be integers with* $0 \leq s < r$. *Then* $d_n(p^r, p^s) = d_n(p^r, ap^s)$.

For example,

$$d_3(27, 18) = d_3(27, 9) \text{ and } d_3(27, 5) = d_3(27, 1).$$

Proof. Define $f : D_n(p^r, p^s) \to D_n(p^r, ap^s)$ by $f(A) = B$ where $B_{1j} = aA_{1j}$ for $j = 1, \ldots, n$ and $B_{ij} = A_{ij}$ for $i = 2, \ldots, n$ and $j = 1, \ldots, n$. Since a is relatively prime to p, a has a multiplicative inverse a' in $\mathbb{Z}_{p^r}$. The map $g : D_n(p^r, ap^s) \to D_n(p^r, p^s)$ defined by $g(A) = B$ where $B_{1j} = a'A_{1j}$ for $j = 1, \ldots, n$ and $B_{ij} = A_{ij}$ for $i = 2, \ldots, n$ and $j = 1, \ldots, n$ is the inverse of f. Hence, f is a bijection and the result follows. ■

In particular, we know that $d_n(p^r, a) = d_n(p^r, 1)$ for all a relatively prime to p. As usual, let $GL(n, \mathbb{Z}_m)$ denote the group of invertible $n \times n$ matrices over $\mathbb{Z}_m$.

THEOREM 1. *Let p be a prime and let n and r be positive integers. Then*

$$|GL(n, \mathbb{Z}_{p^r})| = p^{rn^2} \prod_{i=1}^{n} (1 - p^{-i})$$

and

$$d_n(p^r, a) = p^{r(n^2-1)} \prod_{i=2}^{n} (1 - p^{-i})$$

for all integers a relatively prime to p.

Proof. The number of matrices in $GL(n, \mathbb{Z}_p)$ is calculated in [**2**, p. 125, 5.7.20]. For the sake of completeness, we include this calculation. We count the number of ways to get n independent rows. There are $p^n - 1$ possible first rows since we need only avoid the zero vector. The second row can be anything but a multiple of the first row and there are p multiples of the first row so there are $p^n - p$ possible second rows. The third row can be anything but a linear combination of the first two rows and there are p^2 such linear combinations so there are $p^n - p^2$ possible third rows. Continuing in this fashion, we get

$$|GL(n, \mathbb{Z}_p)| = \prod_{j=0}^{n-1} (p^n - p^j) = p^{n^2} \prod_{i=1}^{n} (1 - p^{-i}). \tag{1}$$

Define the homomorphism $f : M_n(\mathbb{Z}_{p^{r+1}}) \to M_n(\mathbb{Z}_{p^r})$ by sending matrix $A = [a_{ij}]$ to $[a_{ij} \bmod p^r]$. Then the kernel has p^{n^2} elements since $[a_{ij}]$ is in the kernel exactly when a_{ij} is one of the p multiples of p^r for all $i, j = 1, \ldots, n$. Notice that A is invertible if and only if $\gcd(\det A, p) = 1$ if and only if $\gcd(\det(f(A)), p) = 1$ if and only if $f(A)$ is invertible. Therefore, f maps $GL(n, \mathbb{Z}_{p^{r+1}})$ onto $GL(n, \mathbb{Z}_{p^r})$ and, by (iii),

$$|GL(n, \mathbb{Z}_{p^{r+1}})| = p^{n^2} |GL(n, \mathbb{Z}_{p^r})|. \tag{2}$$

Combining (1) and (2) yields

$$|GL(n, \mathbb{Z}_{p^r})| = p^{rn^2} \prod_{i=1}^{n} (1 - p^{-i}).$$

Since $GL(n, \mathbb{Z}_{p^r})$ is the disjoint union of those $D_n(p^r, a)$ for which a is not divisible by p, since these $D_n(p^r, a)$ all have the same number of elements by Lemma 2, and since the number of such a's not divisible by p is $\phi(p^r) = (p-1)p^{r-1}$ (where ϕ is the Euler phi-function), we have

$$\begin{aligned} d_n(p^r, a) &= \frac{|GL(n, \mathbb{Z}_{p^r})|}{\phi(p^r)} \\ &= \frac{p^{rn^2}}{(p-1)p^{r-1}} \prod_{i=1}^{n} (1 - p^{-i}) \\ &= p^{r(n^2-1)} \prod_{i=2}^{n} (1 - p^{-i}), \end{aligned}$$

the desired result.

It remains for us to find $d_n(p^r, p^k)$ for $k = 1, \ldots, r$. Not surprisingly, the case $k = r$, $d_n(p^r, p^r) = d_n(p^r, 0)$, is different from the others and we consider it first.

Since there are p^{n^2} square matrices of size n over $\mathbb{Z}_p$, we have $d_n(p, 0) = p^{n^2} - |GL(n, \mathbb{Z}_p)|$. Hence,

$$d_n(p, 0) = p^{n^2} - \prod_{k=0}^{n-1} (p^n - p^k) \quad \text{for } n = 1, 2, 3, 4, \ldots. \tag{3}$$

We now investigate $d_{n+1}(p^{r+1}, 0)$ by doing row reduction. We need to determine how many $(n+1) \times (n+1)$ matrices $\bmod p^{r+1}$ have determinant 0. We will first count the number of such matrices A which have at least one element in the first column relatively prime to p. Since there are p^r multiples of p in $\mathbb{Z}_{p^{r+1}}$, there are $p^{r(n+1)}$ first columns with all entries divisible by p.Therefore, there are $p^{(r+1)(n+1)} - p^{r(n+1)}$ choices for the first column. Without loss of generality, assume that A_{11} is relatively prime to p. (If it is not, then switch two rows to make this the case. The new matrix has zero determinant if and only if the original matrix does.) Do elementary row operations and use A_{11} to clear the first column. This can be done since A_{11} has a multiplicative inverse $\bmod p^{r+1}$. The new matrix is

$$\begin{bmatrix} 1 & \frac{A_{1,2}}{A_{1,1}} & \cdots & \frac{A_{1,n+1}}{A_{1,1}} \\ 0 & C_{2,2} & \cdots & C_{2,n+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & C_{(n+1),2} & \cdots & C_{(n+1),(n+1)} \end{bmatrix},$$

where

$$C_{i,j} = A_{i,j} - \frac{A_{1,j}}{A_{1,1}} \cdot A_{i,1}.$$

Let C be the $n \times n$ matrix $[C_{i,j}]$, $i, j = 2, 3, \dots, n+1$. Now, $\det A \equiv 0 (\operatorname{mod} p^{r+1})$ if and only if $\det C \equiv 0 (\operatorname{mod} p^{r+1})$. The number of matrices C with determinant 0 is $d_n(p^{r+1}, 0)$. To complete the choice of matrix A it remains to choose $A_{1,j}$ for $j = 2, \dots, n+1$ since once these are chosen, the relations

$$C_{i,j} = A_{i,j} - \frac{A_{1,j}}{A_{1,1}} \cdot A_{i,1}$$

determine the remaining $A_{i,j}$. The number of choices for $(A_{1,2}, A_{1,3}, \dots, A_{1,n+1})$ is $p^{n(r+1)}$. Therefore, the number of $(n+1) \times (n+1)$ matrices mod p^{r+1} that have determinant 0 and have at least one element in the first column relatively prime to p is

$$(p^{(r+1)(n+1)} - p^{r(n+1)}) p^{n(r+1)} d_n(p^{r+1}, 0).$$

Next we count the number of $(n+1) \times (n+1)$ matrices mod p^{r+1} that have determinant 0 and have no element in the first column relatively prime to p. Let A be such a matrix. The first column of A is divisible by p. Let matrix $B = [B_{i,j}]$, where

$$B_{i,1} = \frac{A_{i,1}}{p}$$

for $i = 1, \dots, n+1$ and $B_{i,j} = A_{i,j}$ for $i = 1, \dots, n+1$ and $j = 2, \dots, n+1$. Finally, let matrix C be B reduced modulo p^r. Then

$$\begin{aligned}
&\det A \equiv 0 (\operatorname{mod} p^{r+1}) \text{ if and only if}\\
&\det B \equiv 0 (\operatorname{mod} p^{r}) \text{ if and only if}\\
&\det C \equiv 0 (\operatorname{mod} p^{r}).
\end{aligned}$$

The number of possible matrices C is $d_{n+1}(p^r, 0)$. For each such C, there are $p^{n(n+1)}$ possible B since for each $i = 1, \dots, n+1$ and $j = 2, \dots, n+1$, there is a $k \in \{0, \dots, p-1\}$ such that $B_{i,j} = C_{i,j} + kp^r$. Matrix A is uniquely determined from B by multiplying the first column by p. Therefore, the number of $(n+1) \times (n+1)$ matrices mod p^{r+1} that have determinant 0 and have no element in the first column relatively prime to p is $p^{n(n+1)} d_{n+1}(p^r, 0)$.

Adding the numbers from the two cases, we get

$$d_{n+1}(p^{r+1}, 0) = (p^{(r+1)(n+1)} - p^{r(n+1)}) p^{(r+1)n} d_n(p^{r+1}, 0) + p^{n(n+1)} d_{n+1}(p^r, 0) \quad (4)$$

for $r = 1, 2, 3, \dots$ and $n = 1, 2, 3, \dots$. ■

THEOREM 2. *If p is a prime number and n and r are positive integers, then*

$$d_n(p^r, 0) = p^{rn^2} \left[1 - \prod_{i=0}^{n-1} (1 - p^{-r-i}) \right]. \quad (5)$$

Proof. Using (3) and (4) above, we induct on n and r.

If $r = 1$, the result follows from (3) above. If $n = 1$, $d_n(p^r, 0) = 1$ since $[0]$ is the only 1×1 matrix with determinant 0; this agrees with (5). Now consider $d_{n+1}(p^{r+1}, 0)$, for $n \geq 1$ and $r \geq 1$. We have

$$\begin{aligned}
d_{n+1}(p^{r+1},0) &= (p^{(r+1)(n+1)} - p^{r(n+1)})p^{(r+1)n}d_n(p^{r+1},0) + p^{n(n+1)}d_{n+1}(p^r,0) \text{ by (4)} \\
&= (p^{(r+1)(n+1)} - p^{r(n+1)})p^{(r+1)n}p^{(r+1)n^2}\left[1-\prod_{i=0}^{n-1}(1-p^{-r-1-i})\right] \\
&\quad + p^{n(n+1)}p^{r(n+1)^2}\left[1-\prod_{i=0}^{n}(1-p^{-r-i})\right] \text{ by induction hypothesis} \\
&= p^{(r+1)(n+1)^2} - (p^{(r+1)(n+1)^2} - p^{r(n+1)^2+n^2+n})\prod_{i=0}^{n-1}(1-p^{-r-1-i}) \\
&\quad - p^{n(n+1)+r(n+1)^2}\prod_{i=o}^{n}(1-p^{-r-i}) \\
&= p^{(r+1)(n+1)^2} - p^{(r+1)(n+1)^2}\prod_{i=0}^{n-1}(1-p^{-r-1-i}) \\
&\quad + p^{rn^2+2rn+n^2+n}\prod_{i=0}^{n-1}(1-p^{-r-1-i}) \\
&= p^{(r+1)(n+1)^2} - p^{(r+1)(n+1)^2}(1-p^{-r-n-1})\prod_{i=0}^{n-1}(1-p^{-r-1-i}) \\
&= p^{(r+1)(n+1)^2} - p^{(r+1)(n+1)^2}\prod_{i=0}^{n}(1-p^{-r-1-i}) \\
&= p^{(r+1)(n+1)^2}\left[1-\prod_{i=0}^{n}(1-p^{-r-1-i})\right],
\end{aligned}$$

completing the induction. ■

As an example, we calculate the number of 2 × 2 matrices over $\mathbb{Z}_6$ with determinant k for $k = 0, 1, 2, 3, 4$, and 5. By Lemma 1, it suffices to calculate $d_2(2,0)$, $d_2(2,1)$, $d_2(3,0)$, $d_2(3,1)$, and $d_2(3,2)$. By Theorem 1,

$$\begin{aligned}
d_2(2,1) &= 6 \\
d_2(3,1) &= d_2(3,2) = 24.
\end{aligned}$$

By Theorem 2,

$$\begin{aligned}
d_2(2,0) &= 10 \\
d_2(3,0) &= 33.
\end{aligned}$$

Therefore, using multiplicativity (Lemma 1),

$$\begin{aligned}
d_2(6,0) &= d_2(2,0)d_2(3,0) = 330 \\
d_2(6,1) &= d_2(2,1)d_2(3,1) = 144 \\
d_2(6,2) &= d_2(2,0)d_2(3,2) = 240 \\
d_2(6,3) &= d_2(2,1)d_2(3,0) = 198 \\
d_2(6,4) &= d_2(2,0)d_2(3,1) = 240 \\
d_2(6,5) &= d_2(2,1)d_2(3,2) = 144.
\end{aligned}$$

Observe that the sum of these values is $1296 = 6^4$, the total number of 2×2 matrices over $\mathbb{Z}_6$.

It remains to consider $d_n(p^r, p^k)$ for $0 < k < r$.

LEMMA 3. *Let r and k be integers with $0 \le k < r$ and let s be a positive integer. Then $d_n(p^{r+s}, p^k) = p^{s(n^2-1)} d_n(p^r, p^k)$.*

Proof. Let $f : M_n(\mathbb{Z}_{p^{r+1}}) \to M_n(\mathbb{Z}_{p^r})$ be the homomorphism defined by sending matrix $[a_{ij}]$ to $[a_{ij} \bmod p^r]$. Then $|\ker(f)| = p^{n^2}$. For $A \in M_n(\mathbb{Z}_{p^{r+1}})$, $\det(f(A)) = p^k$ if and only if $\det(A) = p^k + tp^r$ for some $t = 0, \dots, p-1$. Since, for $0 \le k < r$ and $0 \le t \le p-1$, $p^k + tp^r = p^k(1 + tp^{r-k})$ and since $1 + tp^{r-k}$ is not divisible by p, we have

$$|\{A \in M_n(\mathbb{Z}_{p^{r+1}}) : \det A = p^k + tp^r\}| = |\{A \in M_n(\mathbb{Z}_{p^{r+1}}) : \det A = p^k\}|.$$

Therefore,

$$|\{A \in M_n(\mathbb{Z}_{p^{r+1}}) : \det(f(A)) = p^k\}| = p|\{A \in M_n(\mathbb{Z}_{p^{r+1}}) : \det A = p^k\}|.$$

Since $|\ker(f)| = p^{n^2}$,

$$|\{A \in M_n(\mathbb{Z}_{p^{r+1}}) : \det(f(A)) = p^k\}| = p^{n^2}|\{B \in M_n(\mathbb{Z}_{p^r}) : \det B = p^k\}|.$$

Putting this together, we have

$$p|\{A \in M_n(\mathbb{Z}_{p^{r+1}}) : \det A = p^k\}| = p^{n^2}|\{B \in M_n(\mathbb{Z}_{p^r}) : \det B = p^k\}|.$$

Therefore, $d_n(p^{r+1}, p^k) = p^{n^2-1} d_n(p^r, p^k)$. Using this equation s times, we have $d_n(p^{r+s}, p^k) = p^{s(n^2-1)} d_n(p^r, p^k)$. ■

THEOREM 3. *Let p be a prime number and let n, r, and k be positive integers with $r > k$. Then*

$$d_n(p^r, p^k) = \frac{p^n - 1}{p - 1} p^{rn^2-n-r+1} \prod_{i=1}^{n-1}(1 - p^{-k-i}).$$

Proof. Let $f : M_n(\mathbb{Z}_{p^{k+1}}) \to M_n(\mathbb{Z}_{p^k})$ be the homomorphism defined by sending matrix $[a_{ij}]$ to $[a_{ij} \bmod p^k]$. For $A \in M_n(\mathbb{Z}_{p^{k+1}})$, $\det(f(A)) = 0$ if and only if $\det A = 0$ or $\det A = tp^k$ for some $t = 1, \dots, p-1$. Therefore,

$$\begin{aligned} p^{n^2} d_n(p^k, 0) &= |\ker f| d_n(p^k, 0) \\ &= d_n(p^{k+1}, 0) + (p-1) d_n(p^{k+1}, p^k), \end{aligned}$$

and so

$$d_n(p^{k+1}, p^k) = \frac{1}{p-1}(p^{n^2} d_n(p^k, 0) - d_n(p^{k+1}, 0)).$$

By Lemma 3,

$$\begin{aligned} d_n(p^r, p^k) &= d_n(p^{k+1+(r-k-1)}, p^k) \\ &= p^{(r-k-1)(n^2-1)} d_n(p^{k+1}, p^k). \end{aligned}$$

Combining the last two equations gives the recursive formulation

$$d_n(p^r, p^k) = \frac{p^{(n^2-1)(r-k-1)}}{p-1}(p^{n^2} d_n(p^k, 0) - d_n(p^{k+1}, 0)).$$

Using Theorem 2, we get

$$\begin{aligned}
d_n(p^r, p^k) &= \frac{p^{(n^2-1)(r-k-1)}}{p-1}\left(p^{n^2} p^{kn^2}\left[1 - \prod_{i=0}^{n-1}(1 - p^{-k-i})\right]\right. \\
&\quad \left. - p^{(k+1)n^2}\left[1 - \prod_{i=0}^{n-1}(1 - p^{-k-1-i})\right]\right) \\
&= \frac{p^{n^2 r - r + k + 1}}{p-1}\left(-\prod_{i=0}^{n-1}(1 - p^{-k-i}) + \prod_{i=0}^{n-1}(1 - p^{-k-1-i})\right) \\
&= \frac{p^{n^2 r - r + k + 1}}{p-1}\prod_{i=1}^{n-1}(1 - p^{-k-i})\left[(1 - p^{-k-n}) - (1 - p^{-k})\right] \\
&= \frac{p^{n^2 r - r + 1}}{p-1}(1 - p^{-n})\prod_{i=1}^{n-1}(1 - p^{-k-i}) \\
&= \frac{p^n - 1}{p-1} p^{rn^2 - n - r + 1}\prod_{i=1}^{n-1}(1 - p^{-k-i}),
\end{aligned}$$

completing the proof. ■

For example, if p is a prime number and r and k are positive integers with $r > k$, then $d_2(p^r, p^k) = p^{3r} + p^{3r-1} - p^{3r-k-1} - p^{3r-k-2}$.

The reader now has the tools to verify the assertions that

$$d_5(27, 0) = 33{,}193{,}792{,}816{,}322{,}871{,}923{,}020{,}442{,}669{,}672{,}187$$

and

$$d_3(100, 35) = 7{,}749{,}504{,}000{,}000{,}000.$$

Acknowledgment. The authors would like to thank the referee for his suggestions that improved both the content and the exposition in this paper.

REFERENCES

1. Richard P. Brent and Brendan D. McKay, Determinants and ranks of random matrices over $\mathbb{Z}_m$, *Discrete Mathematics* **66** (1987) 35–49.
2. W. R. Scott, *Group Theory*, Dover, Mineola, N. Y., 1987.

The Classification of Similarities: A New Approach

AAD GODDIJN
Freudenthal Institute, University Utrecht
Utrecht, The Netherlands
A.Goddijn@fi.uu.nl

WIM PIJLS
Erasmus University
Rotterdam, The Netherlands
pijls@few.eur.nl

Two well-known types of geometric transformations are the isometry and the similarity. A similarity with factor k, $k > 0$, is defined as a bijective transformation f of the plane onto the plane such that for every segment XY the distance between $f(X)$ and $f(Y)$ is k times the distance between X and Y. An isometry is a similarity with $k = 1$. For isometries, we have a well-known classification. Any isometry is one of the following transformations: a translation, a rotation, a reflection or a glide reflection. For similarities this classification is extended as follows: any similarity transformation that is not an isometry is either a dilative rotation or a dilative reflection. The former transformation preserves orientations and is called a direct transformation, whereas the latter changes orientations and is called an opposite transformation. The classification of similarities is not fully proved in most of geometry textbooks. To our knowledge, only **[1]** Chapter 5 and **[2]** give a satisfactory proof of the above classification of similarities. In this paper, an alternative proof is presented based upon Apollonius circles.

For completeness, we recall some definitions. A *central dilatation* or *stretch* with center C and factor k, $k \neq 0$, is a bijective transformation f of the plane onto the plane with C as a fixed point and the property $\overrightarrow{Cf(X)} = k \cdot \overrightarrow{CX}$ for any point X. A *dilative rotation* is the composition of a rotation around a center C through an angle α and a stretch with the same center C and a factor k. A *dilative reflection* is the composition of a reflection in line ℓ and a stretch with center C, where C is on ℓ. The length of a segment XY is denoted by $|XY|$.

The proofs of the classification

By definition dilative reflections and rotations have a fixed point, their center C. Similarities, as defined above, with factor $k \neq 1$ also have a fixed point, but this should be proved. In the 1960s A.L. Steger discovered an elegant and interesting proof of the following theorem: any non-isometric similarity has a fixed point. From this theorem one derives readily using the classification of isometries that any non-isometric similarity is a dilative rotation or a dilative reflection; see **[1]** chapter 5.

Our proof consists of a one-step approach. It is common to state, as in **[1]**, that a similarity is determined by its actions on three non-collinear points, but our design starts from a segment with two points. Given two segments A_1B_1 and A_2B_2 of unequal length, a single construction yields two points that are the centers of a dilative reflection and a dilative rotation respectively, both mapping A_1 to A_2 and B_1 to B_2. Consequently, when we take a third point C_1, there are only two possible images for this point.

The proof in the current paper has been found in a classical way, described by Pappus as the method of *analysis*. This means: assume the problem being solved and look for the defining properties of the solution [**4**]. So, given two segments A_1B_1 and A_2B_2 of unequal length and length ratio $1 : k$, we seek all points Z, which are supposed to be the center (fixed point) of a dilatation providing the necessary mapping. Z must have the properties $|ZA_2| = k \cdot |ZA_1|$ and $|ZB_2| = k \cdot |ZB_1|$. Each of these two conditions defines a locus of possible points Z. The loci in question are well-known: they are the so-called Apollonius circles, which we discuss in the next section. The two intersection points of those circles, if they exist, will be taken in consideration as possible centers for the dilatations.

Most of the textbooks lack a satisfactory proof for the fact that any non-isometric opposite similarity is dilative reflection. See among others [**3**, p. 22] and [**5**, p. 43]. As Coxeter stated in [**1**, p. 67]: the direct similarities are treated but opposite similarities seem to have been neglected. Only in [**2**] (after Coxeter made his statement) is a sound proof found for the fact that any opposite similarity is a dilative reflection.

The Apollonius circle

Given two arbitrary points X_1 and X_2 and a constant $k > 0, k \neq 1$, the locus of points Z such that $|ZX_2| = k \cdot |ZX_1|$, is a circle Γ, the so-called Apollonius circle for X_1 and X_2 with factor k. See FIGURE 1. The center of Γ lies on the line through X_1 and X_2. Point P is defined as the interior point of X_1X_2 such that $|PX_2| = k \cdot |PX_1|$, whereas P' is defined as the exterior point of X_1X_2 such that $|P'X_2| = k \cdot |P'X_1|$. Hence, both P and P' lie on Γ and PP' is a diameter of Γ.

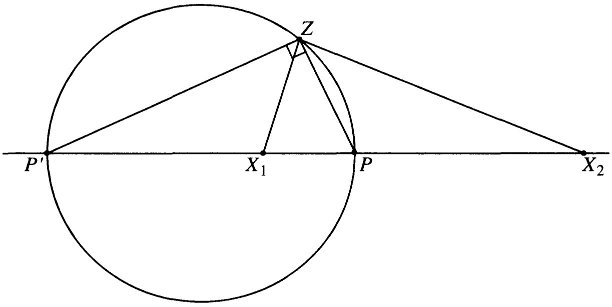


Figure 1 An Apollonius circle

Any point Z on the Apollonius circle Γ satisfies $|ZX_2| = k \cdot |ZX_1|$ and this equality in combination with $|PX_2| = k \cdot |PX_1|$ implies, according to a well-known angle bisector theorem, that ZP is the internal bisector of $\angle X_1ZX_2$. Consequently, for any Z on Γ, there is a dilative reflection with center Z and factor k, which maps X_1 onto X_2. Conversely, if a dilative reflection with center Z and factor k is given that maps X_1 onto X_2, then this center Z lies on Γ. The axis is always ZP.

Likewise, a point Z lies on Γ if and only if it is the center of dilative rotation with factor k transforming X_1 into X_2.

Notice that for any Z on Γ the external bisector of $\angle X_1ZX_2$ is given by ZP', again due to fact that $|P'X_2| = k \cdot |P'X_1|$. The external bisector ZP' and the internal bisector ZP are perpendicular. The aforementioned dilative reflection with center Z, axis ZP, and factor k is identical to the dilative reflection with center Z, axis ZP', and factor $-k$.

Constructing transformations using Apollonius circles

Let two segments A_1B_1 and A_2B_2 be given with $k \neq 1$, where k is defined as $k = |A_2B_2|/|A_1B_1|$. We are looking for a dilative reflection as well as a dilative rotation transforming A_1B_1 into A_2B_2. Let Γ_A and Γ_B denote the Apollonius circles with factor k respectively for the pair A_1, A_2 and the pair B_1, B_2. The possible centers for the dilative reflection and the dilative rotation must lie on the intersection of Γ_A and Γ_B. Unfortunately, we do not know whether these circles intersect. First of all, we shall prove that these circles must intersect.

Let P and P' denote the interior and exterior intersection points of Γ_A with segment A_1A_2. Analogously, the interior and exterior intersection points of Γ_B with segment B_1B_2 are denoted by Q and Q'. The segments PP' and QQ' are diameters of the circles. In FIGURE 2 all these points but not the circles are drawn and two points D and D' are added to produce some similar triangles to be used in the proof. The points D and D' lie on a line parallel to A_1B_1 through B_2 on either side of B_2 at distance $|A_2B_2|$. Since A_1B_1 and A_2B_2 are not parallel, A_2 does not lie on this line. More explicitly we define D and D' by $\overrightarrow{B_2D} = -k \cdot \overrightarrow{B_1A_1}$ and $\overrightarrow{B_2D'} = k \cdot \overrightarrow{B_1A_1}$.

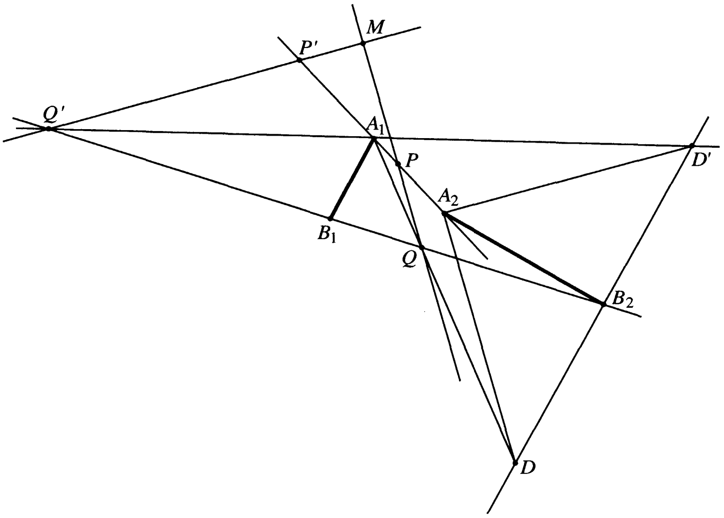


Figure 2 Why are the Apollonius circles intersecting?

Triangle QA_1B_1 is transformed by a stretch with center Q and factor $-k$ into triangle QDB_2 from which we conclude $|QD| = k \cdot |QA_1|$. Now Q is an internal point of segment A_1D, as is P of segment A_1A_2. Remembering $|PA_2| = k \cdot |PA_1|$ we conclude that DA_2 and PQ are parallel. In a similar way using a stretch through Q' with factor k we can prove that $D'A_2$ and $P'Q'$ are parallel.

By the construction of the points D and D' we have the equality $|DB_2| = |A_2B_2| = |D'B_2|$. This implies that $\angle DA_2D' = 90°$. The consequence is that PQ (parallel to DA_2) and $P'Q'$ (parallel to $D'A_2$) are perpendicular, so they surely intersect. The intersection is called M. Since both $\angle PMP'$ and $\angle QMQ'$ are right angles, M lies on Γ_A as well as on Γ_B.

The dilative reflection and the dilative rotation

The dilative reflection defined by factor k, axis MP and center M maps A_1 onto A_2 and the dilative reflection with factor k, axis MQ and center M maps B_1 onto B_2. Since M,

P and Q are collinear, these two transformations are identical. So, the desired dilative reflection has been found. In the next paragraph we focus on the dilative rotation.

In general, Γ_A and Γ_B have besides M another common point, which we call N. A dilative rotation with center N and factor k transforms A_1 onto A_2. Likewise, N is the center of dilative rotation with factor k mapping B_1 onto B_2. These transformations are identical, if $\angle A_1NA_2 = \angle B_1NB_2$. We show that this is indeed the case.

In FIGURE 3, $\angle PMN = \angle PP'N$ and $\angle QMN = \angle QQ'N$, due to the fact that angles inscribed in the same arc of a circle are equal. Since $\angle PMN$ is identical to $\angle QMN$, we conclude $\angle PP'N = \angle QQ'N$. The angles $\angle PNP'$ and $\angle QNQ'$ are right. It follows that the right triangles $PP'N$ and $QQ'N$ are similar. The points A_1 and A_2 divide PP' in the same proportions as B_1 and B_2 divide QQ'. Consequently, figure $PP'NA_1A_2$ is similar to figure $QQ'NB_1B_2$. This implies $\angle A_1NA_2 = \angle B_1NB_2$.

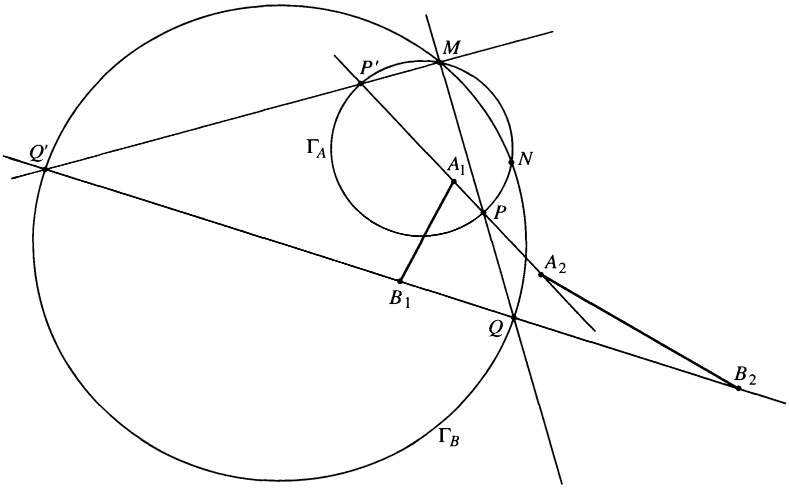


Figure 3 N is the center of the dilative rotation

If Γ_A and Γ_B are tangent in $M = N$, a slightly different derivation applies. We replace in the above derivation $\angle PMN$ and $\angle QMN$ with the angle between PQM and the common tangent line in $M = N$. Similarly to above, we can derive that $\angle PP'N = \angle QQ'N$. As a result of this equality, PP' or A_1A_2 is parallel to QQ' or B_1B_2. The dilative rotation reduces to a central dilatation from $M = N$.

A particular case holds when $P = Q$ or $P' = Q'$. In that case A_1B_1 is parallel to A_2B_2. If $P = Q$, then $P'Q'$, A_1B_1 and A_2B_2 are parallel and we give the line through $P = Q$ perpendicular to $P'Q'$ the role of PQ in the above proof. Then M is again an intersection point of Γ_A and Γ_B. The point given by $P = Q$ plays the role of N. The dilative rotation is a central dilatation.

The case $P' = Q'$ can be handled analogously. Notice that the situation with $P = Q$ and simultaneously $P' = Q'$ cannot happen.

The classification of the similarities

The orientation of a triple (A, B, C) of non-collinear points is either clockwise or counterclockwise according as the traversal A to B to C and back to A is clockwise or not. Let two triangles $A_1B_1C_1$ and $A_2B_2C_2$ be given, such that the lengths of the sides in the latter are k times the lengths in the former, $k \neq 1$. We have shown that there is a unique dilative rotation as well as a unique dilative reflection transforming A_1B_1

into A_2B_2. If the orientations of the triples (A_1, B_1, C_1) and (A_2, B_2, C_2) are the same, the dilative rotation also transforms C_1 into C_2. If the orientations differ, the dilative reflection does so.

As mentioned earlier, any isometry is one of the four transformations: reflection, glide reflection, rotation, or translation. Since a reflection and a rotation are special cases of a dilative reflection and a dilative rotation respectively, we conclude that any similarity is one of the following four transformations: a translation, a dilative rotation, a dilative reflection, or a glide reflection.

REFERENCES

1. H.S.M. Coxeter, *Introduction to Geometry*, second edition, John Wiley, New York, 1969.
2. Clayton W. Dodge, *Euclidean Geometry and Transformations*, reprint in Dover Publications, 2004, first publication in 1972.
3. Roger A. Johnson, *Advanced Euclidean Geometry*, reprint in Dover Publications, 1960, first publication in 1929.
4. J. J. O'Connor and E. F. Robertson, Pappus of Alexandria, at: `http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Pappus.html`
5. I.M. Yaglom, *Geometric Transformations*, vol. II, New Mathematical Library, no. 21, Random House/The Singer Company, 1968.

Perfect Matchings, Catalan Numbers, and Pascal's Triangle

TOMISLAV DOŠLIĆ
University of Zagreb
Kačićeva 26
10000 Zagreb, Croatia
doslic@math.hr

We wish to present a simple combinatorial proof of a determinant formula connecting the Catalan numbers and a matrix derived from Pascal's triangle. We prove the formula by counting perfect matchings in a suitably chosen class of graphs. Although the proof relies on results and techniques from a narrow area, we still believe that it may be interesting also for readers outside this circle, since Catalan numbers are not a very common finding in (or around) the Pascal triangle. We begin with some preliminaries about benzenoid graphs.

A **benzenoid system** is a connected collection of congruent regular hexagons arranged in a plane in such a way that two hexagons are either completely disjoint or have one common edge. To each benzenoid system we can assign a **benzenoid graph**, taking the vertices of hexagons as the vertices of the graph, and the sides of hexagons as the edges of the graph. The resulting graph is simple, planar, 2-connected, bipartite and all its finite faces are hexagons.

A **perfect matching** in a graph G is a collection M of edges of G such that every vertex of G is incident with exactly one edge from M. The number of different perfect matchings in a graph G we denote by $\Phi(G)$.

The motivation for introducing and studying benzenoid graphs came from theoretical chemistry, where they serve as the mathematical model for benzenoid hydrocarbons, a broad and important class of polycyclic carbon and hydrogen compounds in which carbon atoms are arranged in a plane pattern of rings (or cycles) of length six.

It turned out that the stability of a particular benzenoid compound is correlated with the number of perfect matchings in the corresponding benzenoid graph. This fact triggered an avalanche of results concerning the number of perfect matchings (or Kekulé structures, as the chemists call them) in various classes of benzenoid graphs. For an extensive survey, the reader may wish to consult the monograph [**3**]. Also, in recent years, perfect matchings in benzenoid graphs have been intensively studied for their connections with tilings of "Aztec diamonds" and with plane partitions [**11**].

Let B be a benzenoid graph drawn so that some of its edges are vertical. A **peak** of B is a vertex lying above all its neighbors. A **valley** is a vertex lying below all its neighbors. A **monotonic path** in B is a path connecting a peak and a valley, such that after starting at a peak it always goes downwards. An example of a benzenoid graph B with three peaks and three valleys is shown in FIGURE 1. The monotonic path connecting peak 2 with valley 3 is shown in bold lines.

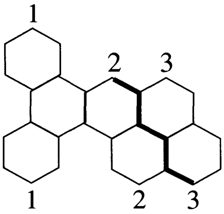


Figure 1 A monotonic path in a benzenoid graph

Obviously, the peaks and the valleys of a given benzenoid graph B belong to different classes of the bipartition of B. Since all other vertices are pairwise connected by vertical edges (and thus their number is balanced), the number of peaks must be equal to the number of valleys if the graph B has a perfect matching.

Let us denote by $w_{i,j}(B)$ the number of monotonic paths that connect the ith peak with the jth valley of B. The matrix whose elements are $w_{i,j}(B)$ we denote by $W(B)$. Since we are interested only in graphs with perfect matchings, we may restrict our attention to the case when $W(B)$ is a square matrix. The matrix $W(B)$ need not be symmetric.

According to John and Sachs [**8**], the number of perfect matchings in a given benzenoid graph B is given by the following formula:

$$\Phi(B) = |\det W(B)|.$$

The same result can be obtained via the Gessel–Viennot theorem that is much better known than the result of John and Sachs (see, e.g., [**1**]). For a given perfect matching in a benzenoid graph we replace its vertical edges by the vertical edges **not** in the matching. This results in a set of disjoint paths from the peaks to the valleys, and Gessel–Viennot applies. For example, the benzenoid graph B from FIGURE 1 has 27 different perfect matchings, since

$$W(B) = \begin{pmatrix} 5 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 3 & 3 \end{pmatrix}$$

and $\det W(B) = 27$.

Let us now apply the the John-Sachs formula to the triangular benzenoid graph T_n consisting of n rows of hexagons, with the number of hexagons in a row decreasing by one from n in the lower-most row to one in the uppermost row, each row shifted one and a half hexagon to the right from the row immediately below it. An example of such a graph is shown in FIGURE 2. First we calculate the matrix elements $w_{i,j}(T_n)$ of the matrix $W(T_n)$.

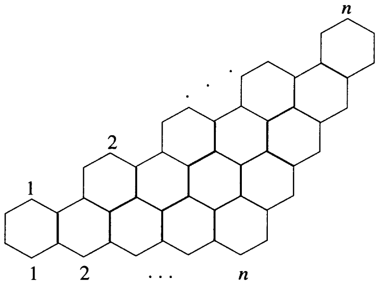


Figure 2 The graph T_n

In the general case, the counting of monotonic paths is a difficult and tedious task, but in our graph T_n we can easily calculate their numbers. To this end, we consider a peak i, $1 \leq i \leq \lceil n/2 \rceil$, and construct an auxiliary graph $R(i)$, induced by all vertices of T_n that can be reached from the peak i by monotonic paths. An example of such a graph is shown in FIGURE 3. (Some authors call such an induced subgraph the **wetting region** of peak i [**7**].)

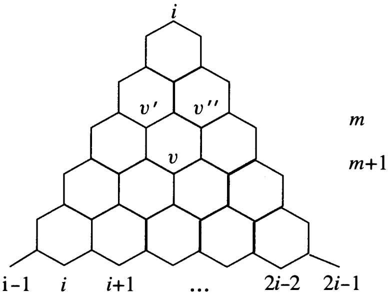


Figure 3 The wetting region of peak i

From FIGURE 3 it is clear that the number of monotonic paths from the peak i to a valley v at level $m+1$ is obtained by summing the numbers of monotonic paths from i to the valleys v' and v'' lying immediately above the considered valley v. Hence, the numbers of monotonic paths obey the same recurrence as the elements of Pascal triangle, with modified initial conditions. From there we obtain

$$w_{i,i+k}(R(i)) = \binom{i+1}{k+1},$$

or, after renaming $j = i + k$,

$$w_{i,j}(R(i)) = \binom{i+1}{j-i+1}.$$

The same reasoning remains valid also in the case when wetting region of the peak i is not triangular, i.e. for $i > \lceil n/2 \rceil$.

Hence, our matrix $W(T_n)$ is given by the matrix elements $w_{i,j}(T_n) = \binom{i+1}{j-i+1}$. In a more explicit form,

$$W(T_n) = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & \cdots \\ 1 & 3 & 3 & 1 & 0 & \cdots \\ 0 & 1 & 4 & 6 & 4 & \cdots \\ 0 & 0 & 1 & 5 & 10 & \cdots \\ 0 & 0 & 0 & 1 & 6 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \end{pmatrix}.$$

We see that, starting from the second row, the ith row of $W(T_n)$ contains the $(i+1)$st row of the Pascal triangle, shifted one place to the right with respect to the row immediately above it. Its determinant could be calculated directly, e.g., by specializing the values of parameters A, B, and L_i in Theorem 26 of reference [**9**], but we proceed here using a bijective approach. By the John-Sachs formula, determinant of $W(T_n)$ is equal to the number of perfect matchings in the graph T_n:

$$\det_{1 \le i,j \le n} \left[\binom{i+1}{j-i+1} \right] = \Phi(T_n). \qquad (1)$$

Let us now determine the quantity $\Phi(T_n)$. In order to find this number, we first consider a class of benzenoid graphs called benzenoid parallelograms. A **benzenoid parallelogram** $B_{m,n}$ consists of $m \times n$ hexagons arranged in m rows, each row containing n hexagons, shifted half a hexagon to the right from the row immediately below it. All such graphs contain perfect matchings, and an example of a perfect matching in $B_{3,4}$ is shown in FIGURE 4.

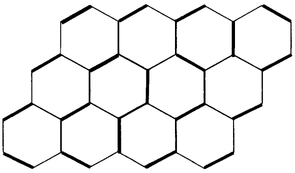


Figure 4 A perfect matching in $B_{3,4}$

It is well-known that the number of perfect matchings in $B_{m,n}$ is equal to $\binom{m+n}{m}$ [**3**]. The bijective correspondence between perfect matchings in $B_{m,n}$ and lattice paths in a rectangular lattice from (0, 0) to (n, m) with steps travelling east and north has been a part of benzenoid folklore at least since early 1950s [**6**]. In fact, it has been so much a part of the folklore that the first proof with all the details fully worked out appeared in print only recently [**4**]. We reproduce here the most important points for the reader's convenience.

We start by observing that the matching shown in FIGURE 4 contains exactly one vertical edge from each row and prove that this is valid for all perfect matchings in $B_{m,n}$.

LEMMA 1. *Every perfect matching in a benzenoid parallelogram $B_{m,n}$ contains precisely one vertical edge of each row.*

Proof. Let us consider a benzenoid parallelogram $B_{m,n}$ and a perfect matching M in $B_{m,n}$. The vertex set of $B_{m,n}$ is partitioned in two sets, W (for white) and B (for black) in such a way that the top vertices of all hexagons are white. Suppose that there is a row, say the ith one, such that no vertical edge from it is contained in M. By removing all vertical edges of this row, we decompose the parallelogram $B_{m,n}$ into the components B^+ and B^-. An example of such procedure is shown in FIGURE 5.

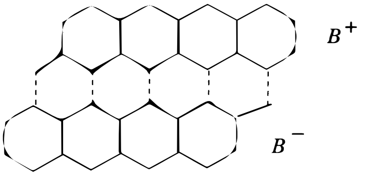


Figure 5 With the proof of Lemma 1

Each of deleted edges connects a black vertex of B^+ with a white one of B^-. Further, in B^+ the number of black vertices exceeds the number of white ones by precisely one (and conversely in B^-). Hence, any perfect matching in $B_{m,n}$ must contain precisely one vertical edge from the specified row. ■

So, every perfect matching M of a benzenoid parallelogram $B_{m,n}$ contains exactly one vertical edge from each row of $B_{m,n}$. There are $n + 1$ vertical edges in every row of $B_{m,n}$. Label them consecutively from the left to the right with integer labels $0, 1, 2, \dots, n$. For a given perfect matching M, let i_p be the label of the vertical edge from the pth row contained in M.

LEMMA 2. *The sequence $(i_1, \dots, i_m)$ is non-decreasing for every perfect matching M of a benzenoid parallelogram $B_{m,n}$.*

Proof. Consider a perfect matching M in $B_{m,n}$ and suppose that there is a $p \in [m-1]$ such that $i_p > i_{p+1}$. Remove all vertical edges from the pth row which are to the left from i_p. (We count the rows from bottom to top.) The remaining graph, $B'_{m,n}$ as shown in FIGURE 6, will have a pendant vertex. Denote this vertex by a. Consider the shortest path connecting the vertex a with x, the lower endpoint of the vertical edge i_{p+1} from M, and denote it by P. No vertex of $P - \{x\}$ is covered by a vertical edge of M, and yet, as no edges from M were removed, all vertices of $p - \{x\}$ must be covered by some edge of M. But the cardinality of $V(P) - \{x\}$ is necessarily odd and we have arrived at a contradiction. ■

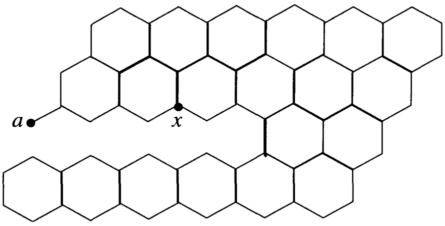


Figure 6 With the proof of Lemma 2

COROLLARY 3. *Let M be a perfect matching in $B_{m,n}$ containing the vertical edge i_p in the row p. Then the part of M lying in the rows $p+1, \dots, m$, left from their respective i_pth vertical edges is uniquely determined. Similarly, the part of M lying in the rows $1, \dots, p-1$, right to their respective i_pth rows is uniquely determined.*

Proof. Let us first consider the part of $B_{m,n}$ above and left from the i_pth vertical edge of the pth row. No vertical edge from this part may be in M, and the conditions on the boundary force both non-vertical edges on the left side of every hexagon in this part of $B_{n,m}$ to be in M. A similar argument holds for the part of $B_{m,n}$ below and right of the considered vertical edge. ■

PROPOSITION 4. *There is a bijection between the set of all perfect matchings in* $B_{m,n}$ *and the set of all non-decreasing sequences of length m with elements from* $\{0, 1, \dots n\}$.

Proof. It follows from Lemma 2 and Corollary 3 that the positions of vertical edges in a perfect matching uniquely define a non-decreasing sequence of length m with elements from $\{0, 1, \dots n\}$. To prove the other part, take a nondecreasing sequence $(i_1, i_2, \dots, i_m)$ with elements from $\{0, 1, \dots, n\}$ and construct a matching in $B_{m,n}$ by taking the vertical edge i_p in the row p. Denote this matching by V and suppose that there are two different perfect matchings, M' and M'', such that $V \subset M'$, $V \subset M''$. Consider their symmetric difference $M' \bigtriangleup M''$. Any edge from $M' \bigtriangleup M''$ must be nonvertical. By Corollary 3, no edge of $M' \bigtriangleup M''$ may lie left and above of any edges of V. Similarly, no such edges can exist right and below the edges from V. The only remaining possibility is that they are on paths connecting the upper end of the vertical edge i_p with the lower end of the edge i_{p+1}. But all such paths must have an even number of inner vertices, and their perfect matchings are unique. So, we have $M' \bigtriangleup M'' = \phi$, and hence $M' = M''$. Therefore, each choice of m vertical edges with nondecreasing labels defines a unique perfect matching M of $B_{m,n}$. ■

Now we describe in more detail the lattice paths we are dealing with. A **lattice path** of length n between the points P_0 and P_n in the (x, y) coordinate plane is any sequence P of n segments $\left(\overline{P_{j-1}P_j}\right)_{j=1}^{n}$ both of whose endpoints have integer coordinates. The segment $\overline{P_{j-1}P_j}$ is called the jth **step** of the path P. By imposing various restrictions on the size and orientation of steps, on the initial and final points, and on the areas of the lattice that must be visited or avoided by the path, we obtain different classes of lattice paths. We consider here lattice paths in a rectangular lattice with integer coordinates from (0, 0) to (n, m) with east and north steps. Denote the set of all such paths with $P_{n,m}$.

PROPOSITION 5. *There is a one-to-one correspondence between the set of all lattice paths from* (0, 0) *to* (n, m) *with east and north steps and the set of all nondecreasing sequences of length m with elements from* $\{0, \dots, n\}$.

Proof. Let us take a lattice path from $P_{n,m}$. It has $m + n$ steps, m of them vertical. By writing down their abscissas, we get a nondecreasing sequence of length m with elements from $\{0, \dots, n\}$, and the correspondence is obviously injective. On the other hand, take a nondecreasing sequence of length m with elements from $\{0, \dots, n\}$ and construct a lattice path starting from $(0, 0)$ with vertical steps connecting the points $(i_j, j - 1)$ and (i_j, j) for $j = 1, \dots, m$. By inserting horizontal steps from (i_j, j) to (i_{j+1}, j), $j = 1, \dots, m - 1$ and the steps from (0, 0) to $(i_1, 0)$ and vertical steps from (i_m, m) to (n, m), if needed, we get a lattice path from (0, 0) to (n, m) with east and north steps and the correspondence is again injective. ■

By combining Proposition 4 and Proposition 5 we obtain the folklore bijective correspondence between perfect matchings and lattice paths. An example is shown in FIGURE 7. (A less formal way to become convinced in the existence of this correspondence is to imagine each right-slanted edge of $B_{m,n}$ gradually contracting to a point.)

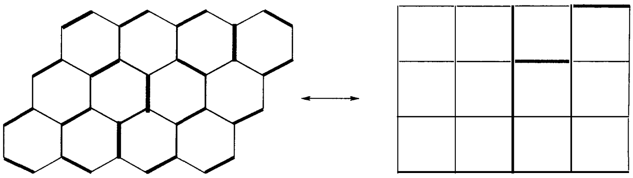


Figure 7 A perfect matching in $B_{3,4}$ and the corresponding lattice path

As a consequence, every perfect matching in a triangular benzenoid graph of the type shown in FIGURE 2 corresponds to a lattice path in a rectangular lattice from $(0, 0)$ to (n, n) with east and north steps that never ventures above the line $y = x$. It is well-known that such lattice paths are enumerated by the nth Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$ ([**12**]). Hence, $\Phi(T_n) = C_{n+1}$. (This result has been obtained earlier ([**13**]) by deriving recurrence formulas for the number of perfect matchings in a wider class of benzenoid graphs and then specializing certain parameters. It was also obtained in the context of rhombus tilings of triangulated regions, by using an approach essentially equivalent to the one presented here ([**2**], [**5**])). Substituting this into formula (1) we obtain our main result.

THEOREM 6. $\det_{1\le i,j\le n}\left[\binom{i+1}{j-i+1}\right] = C_{n+1}.$

As a bonus, by expanding $\det_{1\le i,j\le n}\left[\binom{i+1}{j-i+1}\right]$ by the elements of the last column and shifting the indices, we can establish the following identity.

THEOREM 7. *Let C_i denote the ith Catalan number. Then, for all $n \ge 0$,*

$$\sum_{k\ge 0}(-1)^k\binom{n-k+1}{k}C_{n-k} = \delta_{n,0}.$$

We conclude by noting that matrices similar to our $W(T_n)$ were considered earlier [**10**], and that some of their minors have been expressed in terms of Bernoulli numbers.

Acknowledgments. The author is indebted to the anonymous referee for working out the connection with the Gessel–Viennot theorem. Partial support of the Ministry of Science, Education and Sport of the Republic of Croatia via Grant No. 0037117 is gratefully acknowledged.

REFERENCES

1. A.T. Benjamin, N.T. Cameron, Counting on Determinants, *Amer. Math. Monthly* 112 (2005) 481–492.
2. M. Ciucu, C. Krattenthaler, The number of centered lozenge tilings of a symmetric hexagon, *J. Combin. Theory Ser. A* 86 (1999) 103–126.
3. S.J. Cyvin, I. Gutman, *Kekulé Structures in Benzenoid Hydrocarbons*, Lec. Notes in Chemistry 46, Springer, Heidelberg, 1988.
4. T. Došlić, Perfect matchings in lattice animals and lattice paths with constraints, *Croat. Chem. Acta* 78 (2005) 251–259.
5. T. Eisenkölbl, (−1)-enumeration of plane partitions with complementation symmetry, *Adv. Appl. Math.* 30 (2003) 53–95.
6. M. Gordon, W. H. T. Davison, Theory of Resonance Topology of Fully Aromatic Hydrocarbons, *J. Chem. Phys.* 20 (1952) 428–435.
7. I. Gutman, S.J. Cyvin, A new method for the enumeration of Kekulé structures, *Chem. Phys. Lett.* 136 (1987) 137–140.
8. P. John, H. Sachs, Wegesysteme und Linearfaktoren in hexagonalen und quadratischen Systemen, *Graphen in Forschung und Unterricht*, Franzbecker-Verlag, Kiel, 1985. 85–101.
9. C. Krattenthaler, Advanced determinant calculus, *Sém. Lothar. Comb.* 42 (1999) B42q, 67 pp.

10. A. Lascoux, M. P. Schützenberger, Polynômes de Schubert, *Comptes Rendus* 294 (1982) 447.
11. J. Propp, Enumerations of Matchings: Problems and Progress, *New Perspectives in Geometric Combinatorics*, MSRI Publications, Vol. 37, 1999.
12. R.P. Stanley, *Enumerative Combinatorics*, vol. 2, Cambridge Univ. Press, Cambridge, 1999.
13. R. Tošić, S.J. Cyvin, Enumeration of Kekulé structures in benzenoid hydrocarbons: "flounders" *J. Math. Chemistry* **3** (1989) 393–401.

On Candido's Identity

CLAUDI ALSINA
Universitat Politècnica de Catalunya
08028 Barcelona, Spain
claudio.alsina@upc.edu

ROGER B. NELSEN
Lewis & Clark College
Portland, OR 97219, USA
nelsen@lclark.edu

Giacomo Candido [**1**] (1871–1941) proved the equality

$$[F_n^2 + F_{n+1}^2 + F_{n+2}^2]^2 = 2[F_n^4 + F_{n+1}^4 + F_{n+2}^4],$$

where F_n denotes the nth Fibonacci number, by observing that for all reals x, y one has the curious identity

$$[x^2 + y^2 + (x+y)^2]^2 = 2[x^4 + y^4 + (x+y)^4]. \tag{1}$$

Candido's identity (1) can be easily shown to be true not only in $\mathbb{R}^+ := [0, \infty)$ but also in any commutative ring and admits a clear visual description as presented recently in [**3**]. This identity raises the question: is (1) a characteristic property of the polynomial function $y = x^2$ in $\mathbb{R}^+$? In order to answer this we reformulate (1) as follows. Let f be a function from $\mathbb{R}^+$ into $\mathbb{R}^+$ such that

$$f\,(f(x) + f(y) + f(x+y)) = 2\,[f(f(x)) + f(f(y)) + f(f(x+y))]. \tag{2}$$

In general (2) admits trivial solutions like $f \equiv 0$ as well as many bizarre, highly discontinuous solutions. For example, define f to be any function from $\mathbb{R}^+$ to $\mathbb{R}^+$ with the property that $f(x) = 0$ whenever x is rational and $f(x)$ is rational (but arbitrary!) whenever x is irrational. It is an exercise (try it) to show that every possible combination of rational or irrational values for the inputs x and y reduces (2) to the identity $0 = 0$. But if we require f to be a continuous surjection on $\mathbb{R}^+$ with $f(0) = 0$, then we shall show that f can differ from the squaring function only by a multiplicative constant.

LEMMA. *For any two positive real numbers a and b with* $0 < a < b$, *there are integers m and n such that* $a < 2^m 3^n < b$.

Proof. We consider three cases.

Case 1. If $1 \le a < b$ then $0 \le \log_2(a) < \log_2(b)$ and it follows that $\log_2(a)/3^n < \log_2(b)/3^n < 1$ for a sufficiently large positive integer n. Since $2^p \neq 3^q$ for all integers p, q such that $p, q \neq 0$, we deduce $p \log 2 \neq q \log 3$, i.e., $\log_2(3) = \log 3/\log 2$ is clearly irrational (see, e.g., [**2**]). So it follows from the equidistribution theorem [**4**, Theorem 6.2, p. 72] that the sequence $\log_2(3)$, $2\log_2(3)$, $3\log_2(3), \ldots$ is uniformly distributed modulo 1, i.e., there is some positive integer m such that

$$\log_2(a)/3^n < \log_2(3^m) - \lfloor \log_2(3^m) \rfloor < \log_2(b)/3^n,$$

where $\lfloor x \rfloor$ denotes the greatest integer $k \leq x$. Let $r = \log_2(3^m)$ and let $s = r - \lfloor r \rfloor$. Then since $2^r = 3^m$, it follows that $2^s = 3^m/2^{\lfloor r \rfloor}$. With this notation

$$\log_2(a) < 3^n s < \log_2(b)$$

i.e., $a < 2^{(3^n s)} < b$, whence $a < (3^m/2^{\lfloor r \rfloor})^{3^n} < b$. This shows that there is an integral power of 2 times an integral power of 3 between a and b.

Case 2. If $a < 1 < b$ we can use $n = m = 0$.

Case 3. If $0 < a < b \leq 1$ we will have $1 \leq 1/b < 1/a$ so by case 1 there exist integers m, n such that $1/b < 2^m 3^n < 1/a$ and therefore $a < 2^{-m}3^{-n} < b$. ■

Now we prove the following:

THEOREM. *A continuous surjective function f from $\mathbb{R}^+$ to $\mathbb{R}^+$ such that $f(0) = 0$ satisfies Candido's equation (2) if and only if*

$$f(x) = kx^2, \tag{3}$$

where $k > 0$ is an arbitrary constant.

Proof. From Candido's equality (1), it follows that (3) satisfies (2). Conversely, assume that f is a solution of (2) satisfying the above conditions. Since $f(0) = 0$ the substitution $y = 0$ into (2) yields that for all $x \geq 0$: $f(2f(x)) = 4f(f(x))$. Since f is surjective, $f(x)$ ranges throughout $\mathbb{R}^+$ as x ranges throughout $\mathbb{R}^+$, so that if we let $z = f(x)$, we have $f(2z) = 4f(z)$ for all z in $\mathbb{R}^+$. It follows by induction

$$f(2^n z) = (2^n)^2 f(z), \tag{4}$$

for all integers $n \geq 0$.

Since $f(z) = f(2^n(z/2^n)) = (2^n)^2 f(z/2^n)$ we get

$$f(2^{-n}z) = (2^{-n})^2 f(z) \tag{5}$$

for all integers $n \geq 1$. Thus from (4) and (5) we can conclude

$$f(2^n z) = (2^n)^2 f(z), \tag{6}$$

for all integers n. Next, set $y = x$ in (2) to obtain

$$f(2f(x) + f(2x)) = 4f(f(x)) + 2f(f(2x)),$$

and by virtue of (6), using $f(2x) = 4f(x)$, we get:

$$4f(3f(x)) = f(6f(x)) = 4f(f(x)) + 2 \cdot 4^2 \cdot f(f(x)) = 36f(f(x)),$$

i.e., with $f(x) = z \geq 0$ arbitrary, $f(3z) = 3^2 f(z)$ and by induction $f(3^m z) = (3^m)^2 f(z)$, whenever $m \geq 0$. As above, $f(z) = f(3^m(z/3^m)) = (3^m)^2 f(z/3^m)$ so $f(3^{-m}z) = (3^{-m})^2 f(z)$ and therefore

$$f(3^m z) = (3^m)^2 f(z), \tag{7}$$

for all integers m. By means of (6) and (7), we obtain that for all integers m, n:

$$f(2^n 3^m) = (2^n 3^m)^2 f(1). \tag{8}$$

By our previous lemma any real numbers in $[0, \infty)$ may be approximated by a sequence in the set $\{2^n 3^m | n, m \text{ integers }\}$ so from (8) and the continuity of f we can conclude that for all x in $\mathbb{R}^+$, $f(x) = kx^2$, with $k = f(1) > 0$ an arbitrary constant. ∎

Acknowledgment. The authors thank the referees for their helpful remarks and suggestions which improved the final presentation of this paper.

REFERENCES

1. G. Candido, A Relationship Between the Fourth Powers of the Terms of the Fibonacci Series. *Scripta Mathematica* 17:3–4 (1951) 230.
2. S. Lang, *Introduction to Transcedental Numbers*, Addison-Wesley, Reading, 1966.
3. R. B. Nelsen, *Proof Without Words: Candido's Identity*, this MAGAZINE, **78** No. 2 (2005) 131.
4. I. Niven, *Irrational Numbers*, Carus. Math. Mono. 11, MAA, Wiley, New York, 1956.

Monotonic Convergence to *e* via the Arithmetic-Geometric Mean

JÓZSEF SÁNDOR
Department of Mathematics and Computer Sciences
Babeş-Bolyai University
Str. Kogălniceanu Nr.1
400084 Cluj-Napoca, Romania
jjsandor@hotmail.com

Recently, Hansheng Yang and Heng Yang [**3**], by using only the arithmetic-geometric inequality, have proved the monotonicity of the sequences (x_n), (y_n), related to the number e:

$$x_n = \left(1 + \frac{1}{n}\right)^n, \quad y_n = \left(1 + \frac{1}{n}\right)^{n+1} \quad (n = 1, 2, \dots)$$

Such a method probably is an old one and has been applied e.g. in [**1**], or [**2**].

We want to show that the above monotonicities can be proved much easier than in [**3**].

Recall that the arithmetic-geometric inequality says that for $a_1, \dots, a_k > 0$, and

$$G_k = G_k(a_1, \dots, a_k) = \sqrt[k]{a_1 \dots a_k},$$
$$A_k = A_k(a_1, \dots, a_k) = \frac{a_1 + \dots + a_k}{k},$$

we have

$$G_k \le A_k, \tag{1}$$

with equality only when all a_i are equal.

Let $k = n + 1$, $a_1 = 1$, and $a_2 = a_3 = \cdots = a_{n+1} = 1 + \frac{1}{n}$. Then

$$G_{n+1} = \left(1 + \frac{1}{n}\right)^{n/n+1} \quad \text{and} \quad A_{n+1} = 1 + \frac{1}{n+1},$$

and raising both sides to the $n + 1$ power gives

$$x_n < x_{n+1}. \tag{2}$$

For $k = n + 2$, $a_1 = 1$, and $a_2 = a_3 = \cdots = a_{n+2} = 1 - \frac{1}{n+1}$, we have

$$G_{n+2} = \left(1 - \frac{1}{n+1}\right)^{n+1/n+2} \quad \text{and} \quad A_{n+2} = \frac{n+1}{n+2}.$$

Raising both sides to the $n + 2$ power and taking reciprocals gives

$$y_{n+1} < y_n. \tag{3}$$

Clearly $y_n - x_n = x_n \cdot \frac{1}{n} > 0$, so $x_n < x_{n+1} < y_{n+1} < y_n < \cdots < y_1 = 4$ for $n > 1$, thus the sequences (x_n), (y_n) are convergent, having the same limit (denoted by e). Clearly, $x_n < e < y_n$, so (2) and (3) give

$$x_n < x_{n+1} < e < y_{n+1} < y_n. \tag{4}$$

REFERENCES

1. A. Lupaş, On the Number *e* (Romanian), *Gazeta Mat., Seria B* **23**(7) (1972) 393–396.
2. J. Sándor, On Bernoulli's Inequality, *Octogon Math. Mag.* **3**(1) (1995) 34–35.
3. H. Yang and H. Yang, The Arithmetic-Geometric Mean Inequality and the Constant *e*, this MAGAZINE **74** (2001) 321–323.

PROBLEMS

Proposals

To be considered for publication, solutions should be received by November 1, 2007.

1771. *Proposed by Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.*

Let P be a point inside of triangle ABC, and let AA', BB', and CC' be the cevians through P. Prove that if $A'B' = A'C'$ and $BC' = CB'$, then triangle ABC is isosceles.

1772. *Proposed by Rick Mabry, Louisiana State University at Shreveport, Shreveport, LA.*

Let n be an even positive integer and let a be a real number. Let $T_n(x; a)$ denote the degree n Taylor polynomial at the point $x = a$ for the exponential function e^x.

(a) Prove that for each real a, the polynomial $T_n(x; a)$ assumes its minimum at a unique point.

(b) Let t_a denote the value of x for which $T_n(x; a)$ assumes its minimum. Prove that the planar set

$$\{(t_a,\ T_n(t_a; a))\ :\ a \in \mathbb{R}\}$$

is itself the graph of an exponential function.

1773. *Proposed by H. A. ShahAli, Tehran, Iran.*

Let a, b, c, d be nonnegative real numbers with $a + b = c + d = 1$. Determine the maximum value of

$$(a^2 + c^2)(a^2 + d^2)(b^2 + c^2)(b^2 + d^2),$$

and determine conditions under which the maximum is attained.

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

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames IA 50011, or mailed electronically (ideally as a LaTeX file) to ehjohnst@iastate.edu. All communications, written or electronic, should include the reader's name, full address, and an e-mail address and/or FAX number on every page.

1774. *Proposed by Götz Trenkler, Department of Statistics, University of Dortmund, Dortmund, Germany.*

Let P and Q be idempotent, hermetian matrices of the same dimension and rank. Prove that if $PQP = P$, then $P = Q$.

1775. *Proposed by Christopher J. Hillar, Texas A&M University, College Station, TX.*

Characterize those graphs G that satisfy the following conditions: between each pair of vertices A and B in G,

(a) there exist two vertex disjoint paths.

(b) any set of vertex disjoint paths between A and B has at most two elements.

Quickies

Answers to the Quickies are on page 236.

Q971. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let $\{a_n\}$ and $\{b_n\}$ be two sequences of positive terms. Prove that if $\sum_{k=1}^{\infty} a_n$ and $\sum_{k=1}^{\infty} \frac{a_n}{b_n}$ both converge, then $\sum_{k=1}^{\infty} \frac{a_n}{b_n^r}$ converges for any r with $0 < r < 1$.

Q972. *Proposed by Michel Bataille, Rouen, France.*

Find all rational numbers r_1, r_2, r_3, r_4, r_5 such that $r_1r_5 = 1$, $r_1r_4 + r_2r_5 = 2$, $r_1r_3 + r_2r_4 + r_3r_5 = 3$, $r_1r_2 + r_2r_3 + r_3r_4 + r_4r_5 = 4$, and $r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2 = 5$.

Solutions

Fair game **June 2006**

1746. *Proposed by Stephen J. Herschkorn, Highland Park, NJ.*

Alice and Bob play a game in which they alternately flip a (biased) coin that has probability p of coming up heads when tossed. Alice goes first. With one possible exception, each player flips the coin once per turn. The first player to have cumulatively flipped k heads is the winner. To compensate for Alice's advantage in going first, Bob gets a second flip on his first turn if his first flip turns up tails; this is the exception. (Note that if $k = 1$, Bob may not get to flip at all.)

For each of the cases $k = 1$ and $k = 2$, determine the value of p for which the game is fair, and calculate the expected value and variance of the number of flips in the game when p takes on this value.

Solution by Rob Pratt and Emily Lada, Raleigh, NC.

Starting with Alice's second turn, let state (i, j) denote that the current player needs i more heads to win and the opponent needs j more heads to win. Let P_{ij} denote the probability that the current player wins, starting from state (i, j). Define random variable X_{ij} to be the number of flips needed to terminate, starting from state (i, j). Let $E_{ij} = \mathrm{E}X_{ij}$ and $E_{ij}^2 = \mathrm{E}X_{ij}^2$. By conditioning on whether the next flip is heads or tails, we obtain

$$\begin{aligned} P_{i0} &= 0 && 1 \le i \le k, \\ P_{ij} &= p(1 - P_{j,i-1}) + (1-p)(1 - P_{ji}) && 1 \le i \le k,\ 1 \le j \le k. \end{aligned} \tag{1}$$

Similarly,

$$\begin{aligned} E_{i0} &= 0 && 1 \le i \le k, \\ E_{ij} &= 1 + p\ E_{j,i-1} + (1-p)E_{ji} && 1 \le i \le k,\ 1 \le j \le k. \end{aligned} \tag{2}$$

Also,

$$\begin{aligned} E^2_{i0} &= 0 && 1 \le i \le k, \\ E^2_{ij} &= p\,\mathrm{E}(1 + X_{j,i-1})^2 + (1-p)\mathrm{E}(1 + X_{ji})^2 \\ &= 1 + 2p\ E_{j,i-1} + p\ E^2_{j,i-1} \\ &\quad + 2(1-p)E_{ji} + (1-p)E^2_{ji} && 1 \le i \le k,\ 1 \le j \le k. \end{aligned} \tag{3}$$

We compute the desired quantities for $k = 1$ and $k = 2$ by conditioning on the flips before Alice's second turn. For $k = 1$, these four flip sequences are H, TH, TTH, and TTT. The probability that Alice wins is therefore

$$p \cdot 1 + p(1-p)P_{10} + p(1-p)^2 P_{10} + (1-p)^3 P_{11} = p + (1-p)^3/(2-p),$$

where we have substituted the values $P_{10} = 0$ and $P_{11} = 1/(2-p)$ obtained from solving (1). Setting this probability equal to $1/2$ and solving for p yields $p = 1 - 1/\sqrt{2}$. We could compute the expected value and variance of the number of flips by applying the same conditioning argument and substituting the solutions of (2) and (3). But it is simpler to recognize that, for $k = 1$ only, the number of flips is a geometric random variable with success probability p. So the expected value is $1/p = 1/(1 - 1/\sqrt{2}) = 2 + \sqrt{2}$, and the variance is $(1-p)/p^2 = 4 + 3\sqrt{2}$.

For $k = 2$, the six possible flip sequences before Alice's second turn are HH, HTH, HTT, TH, TTH, and TTT. So Alice wins with probability

$$\begin{aligned} &(p^2 + p^2(1-p))P_{11} + p(1-p)^2 P_{12} + (p(1-p) + p(1-p)^2)P_{21} + (1-p)^3 P_{22} \\ &\quad = (4 - 7p + 6p^2 - 2p^3)/(2-p)^3, \end{aligned}$$

where we have substituted the values $P_{11} = 1/(2-p)$, $P_{12} = (3-2p)/(2-p)^2$, $P_{21} = (1-p)/(2-p)^2$, and $P_{22} = (4 - 5p + 2p^2)/(2-p)^3$ obtained from solving (1). Setting this probability equal to $1/2$ and solving for p yields $p = 1 - 1/\sqrt{3}$. By applying the same conditioning argument and substituting the solution of (2), we compute the expected number of flips as

$$\begin{aligned} &p^2(2 + E_{11}) + p^2(1-p)(3 + E_{11}) + p(1-p)^2(3 + E_{12}) \\ &\quad + p(1-p)(2 + E_{21}) + p(1-p)^2(3 + E_{21}) + (1-p)^3(3 + E_{22}) \\ &\quad = (5 - 2p + p^2 - p^3)/(p(2-p)) \\ &\quad = 7/2 + 5/\sqrt{3}. \end{aligned}$$

Similarly, we use conditioning and the solutions of (2) and (3) to compute the second moment of the number of flips as

$$
\begin{aligned}
&p^2\mathrm{E}(2+X_{11})^2 + p^2(1-p)\mathrm{E}(3+X_{11})^2 + p(1-p)^2\mathrm{E}(3+X_{12})^2\\
&\quad + p(1-p)\mathrm{E}(2+X_{21})^2 + p(1-p)^2\mathrm{E}(3+X_{21})^2 + (1-p)^3\mathrm{E}(3+X_{22})^2\\
&\quad = p^2(4+4E_{11}+E_{11}^2) + p^2(1-p)(9+6E_{11}+E_{11}^2)\\
&\qquad + p(1-p)^2(9+6E_{12}+E_{12}^2) + p(1-p)(4+4E_{21}+E_{21}^2)\\
&\qquad + p(1-p)^2(9+6E_{21}+E_{21}^2) + (1-p)^3(9+6E_{22}+E_{22}^2)\\
&\quad = 3(12-14p+7p^2-3p^4+p^5)/(p^2(2-p)^2)\\
&\quad = 39/2 + 17\sqrt{3}.
\end{aligned}
$$

Hence, the variance is $(39/2 + 17\sqrt{3}) - (7/2 + 5/\sqrt{3})^2 = 16\sqrt{3} - 13/12$.

Also solved by Michael Andreoli, Robert Calcaterra, Chip Curtis, Peter W. Lindstrom, Jyoti Shiwalkar and M. N. Deshpande (India), Nicholas Singer, Michael Vowe (Switzerland), Paul Weisenhorn (Germany), Yongjun Yang, and the proposer. There was one incorrect submission.

A countable Hausdorff space? June 2006

1747. *Proposed by Stephen J. Herschkorn, Highland Park, NJ.*

Does there exist a Hausdorff space with a countably infinite topology?

Solution by Paul Budney, Sunderland MA.

There is no such topology. Suppose that X is a Hausdorf space with a countable infinite topology. If X has an infinite subset A of isolated points, then the set of all subsets of A is an uncountable collection of open sets. Hence X has at most finitely many isolated points. Because the topology on X has infinitely many sets, X cannot be finite. Thus we can find two nonisolated points x_1, $y_1 \in X$. Then we can find disjoint open sets U_1, V_1 with $x_1 \in U_1$ and $y_1 \in V_1$ and so that V_1 contains no isolated points. Now suppose that $U_1, U_2, \ldots, U_n, V_n$ are pairwise disjoint, nonempty open sets and that V_n contains no isolated points. Then we can find distinct points x_{n+1}, y_{n+1} and disjoint open sets U_{n+1}, $V_{n+1} \subset V_n$ with $x_{n+1} \in U_{n+1}$ and $y_{n+1} \in V_{n+1}$. It follows that $U_1, U_2, \ldots, U_{n+1}, V_{n+1}$ are pairwise disjoint open sets and that V_{n+1} contains no isolated points. This process generates an infinite sequence $U_1, U_2, \ldots$ of pairwise disjoint open subsets of X. But then the set $\{\cup_{j \in J} U_j \,:\, J \subset \mathbb{Z}^+\}$ is an uncountable collection of distinct open subsets of X. This is impossible if the topology on X is countably infinite.

Also solved by Alejandro Aguado and George F. Seelinger, Samuel F. Barger, Michael W. Botsko, Robert Calcaterra, John Cobb, Fejéntaláltuka Szeged Problem Solving Group (Hungary), Peter W. Lindstrom, Kim McInturff, Jose H. Nieto (Venezuela), Dave Trautman, and the proposer. There was one solution with no name and one incorrect submission.

Triangular Products June 2006

1748. *Proposed by Anonymous*

Let m and n be positive integers such that mn is a triangular number. Prove that there exists an integer k such that the sequence $\{R_j\}$ generated by

$$R_0 = m, \qquad R_1 = n, \qquad R_j = 6R_{j-1} - R_{j-2} + k, \quad j \geq 2,$$

has the property that $R_j R_{j+1}$ is a triangular number for all integer $j \geq 0$.

Solution by Nicholas C. Singer, Annandale, VA.

An integer T is a triangular number if and only if $8T+1$ is a square. Because mn is triangular, $t=\sqrt{8mn+1}$ is an integer. Let $k=2(m+n\pm t)$ and note that k is an even integer. For positive integer j,

$$\begin{aligned}
8R_jR_{j+1}&+1-\left(R_j+R_{j+1}-\frac{k}{2}\right)^2\\
&=8R_j(6R_j-R_{j-1}+k)+1-\left(7R_j-R_{j-1}+\frac{k}{2}\right)^2\\
&=6R_{j-1}R_j+k(R_{j-1}+R_j)+1-R_{j-1}^2-R_j^2-\frac{k^2}{4}\\
&=8R_{j-1}R_j+1-\left(R_{j-1}+R_j-\frac{k}{2}\right)^2.
\end{aligned}$$

It follows that the first expression in the display is invariant for integer j. Setting $j=0$ we see that the value of the invariant is

$$8mn+1-(m+n-(m+n\pm t))^2=8mn+1-t^2=0.$$

Hence $8R_jR_{j+1}+1=\left(R_j+R_{j+1}-\frac{k}{2}\right)^2$ for all integers $j\geq 0$. It follows that with k chosen as above, R_jR_{j+1} is a triangular number for all integer $j\geq 0$.

Also solved by Jim Delany, Charles R. Diminnie, Northwestern University Math Problem Solving Group, Raúl A. Simon (Chile), Marian Tetiva (Romania), Michael Vowe (Switzerland), and the proposer. There was one incorrect submission.

A triangle inequality June 2006

1749. *Proposed by Mihàly Bencze, Nègygalu, Romania.*

Let r and R be, respectively, the inradius and circumradius of a triangle with sides of length a, b, and c and let n be a positive integer. Prove that

$$\frac{R(R^n-r^n)}{(R-r)r^n}\geq\frac{(a^n-b^n)(a^{n+1}+b^{n+1})}{(a-b)a^nb^n}+2^{n+1}-2(n+1).$$

Solution by the proposer.

Let $x=\frac{b+c-a}{2}$, $y=\frac{c+a-b}{2}$, and $z=\frac{a+b-c}{2}$. Then

$$\frac{R}{r}=\frac{2abc}{(b+c-a)(c+a-b)(a+b-c)}=\frac{(x+y)(y+z)(z+x)}{4xyz}.$$

Because

$$\frac{x+y}{4xyz}=\frac{1}{4zx}+\frac{1}{4yz}\geq\frac{1}{(z+x)^2}+\frac{1}{(y+z)^2},$$

it follows that

$$\frac{R}{r}\geq\frac{y+z}{z+x}+\frac{z+x}{y+z}=\frac{a}{b}+\frac{b}{a}\geq 2.$$

This proves the inequality in the case $n=1$ and generalizes Euler's inequality $R\geq 2r$.

We use induction to prove that for any positive integer k,

$$\left(\frac{R}{r}\right)^k \geq \left(\frac{a}{b}\right)^k + \left(\frac{b}{a}\right)^k + 2^k - 2. \tag{1}$$

If (1) is true for some positive integer k, then

$$\begin{aligned}\left(\frac{R}{r}\right)^{k+1} = \left(\frac{R}{r}\right)^k \frac{R}{r} &\geq \left(\left(\frac{a}{b}\right)^k + \left(\frac{b}{a}\right)^k + 2^k - 2\right)\left(\frac{a}{b} + \frac{b}{a}\right)\\ &= \left(\frac{a}{b}\right)^{k+1} + \left(\frac{b}{a}\right)^{k+1} + 2^{k+1} - 2\\ &\quad + (2^k - 2)\left(\frac{a}{b} + \frac{b}{a} - 2\right) + \left(\left(\frac{a}{b}\right)^{k-1} + \left(\frac{b}{a}\right)^{k-1} - 2\right)\\ &\geq \left(\frac{a}{b}\right)^{k+1} + \left(\frac{b}{a}\right)^{k+1} + 2^{k+1} - 2.\end{aligned}$$

The desired inequality now follows by summing the inequalities (1) for $k = 1, 2, \ldots, n$.

Also solved by Paul Weisenhorn (Germany), and John B. Zacharias. There was one incorrect submission.

Covering 3-progressions June 2006

1750. *Proposed by Christopher J. Hillar, Texas A&M University, College Station, TX.*

Let $p > 3$ be prime. Define a sequence x_1, x_2, x_3 of integers to be a 3-progression if they are in arithmetical progression modulo p. If $A_1, A_2, \ldots, A_n \subseteq \mathbb{Z}/p\mathbb{Z}$ is a collection of sets such that each three progression is contained in at least one of the A_k's, then the collection $\{A_1, A_2, \ldots, A_n\}$ is called a 3-covering of $\mathbb{Z}/p\mathbb{Z}$. Find the minimum over all 3-coverings of the quantity

$$\sum_{i=1}^{n} |A_i|^2.$$

Solution by the proposer.

The answer is p^2 achieved with $k = 1$ and $A_1 = \mathbb{Z}/p\mathbb{Z}$. First notice that any 3-progression mod p is an ordered triple of the form $(x_1, (x_2 + x_1)/2, x_2)$ for some $x_1 \neq x_2$. Moreover, since $p > 3$, a permutation of $(x_1, (x_2 + x_1)/2, x_2)$ that is also 3-progression must either be $(x_1, (x_2 + x_1)/2, x_2)$ or $(x_2, (x_2 + x_1)/2, x_1)$. It follows that the number of subsets of $\mathbb{Z}/p\mathbb{Z}$ that are the elements of a 3-progression is equal to $\binom{p}{2}$.

Let $\|A_i\|$ denote the number of 3-progressions that are contained in A_i. We claim next that $|A_i|^2 \geq 2\|A_i\| + |A_i|$. To prove this simply observe that a subset A_i has at most $\binom{|A_i|}{2}$ 3-progressions by the reasoning above. Finally, to complete the proof, examine the chain of inequalities:

$$\sum_{i=1}^{n} |A_i|^2 \geq 2\sum_{i=1}^{n} \|A_i\| + \sum_{i=1}^{n} |A_i| \geq 2\binom{p}{2} + p = p^2,$$

where the final inequality follows since A_i must cover all 3-progressions mod p, and every element must appear in some A_i.

Note. It is not necessary that the sets under consideration be from a 3-progression, but only that each set is "determined" by a set of 2 elements. The result is also true in any $\mathbb{Z}/n\mathbb{Z}$, with $\gcd(6, n) = 1$.

There were two incorrect submissions.

Answers

Solutions to the Quickies from page 231.

A971. First consider $\sum_{b_n \geq 1} \frac{a_n}{b_n^r}$. For $b_n \geq 1$, we have $\frac{a_n}{b_n^r} \leq a_n$. It follows from the comparison test that $\sum_{b_n \geq 1} \frac{a_n}{b_n^r}$ converges. Next consider $\sum_{b_n<1} \frac{a_n}{b_n^r}$. For $0 < b_n < 1$ and $0 < r < 1$ we have $b_n < b_n^r$. Thus $\frac{a_n}{b_n^r} \leq \frac{a_n}{b_n}$ for $0 < b_n < 1$. Again by the comparison test, $\sum_{b_n<1} \frac{a_n}{b_n^r}$ converges. Combining the two results we conclude that $\sum_{n=1}^{\infty} \frac{a_n}{b_n^r}$ converges.

A972. It is clear that

$$(r_1, r_2, r_3, r_4, r_5) = (1, 1, 1, 1, 1)$$

and

$$(r_1, r_2, r_3, r_4, r_5) = (-1, -1, -1, -1, -1)$$

are both solutions. We prove that there are no other rational solutions. Towards this end, let $(r_1, r_2, r_3, r_4, r_5) \in \mathbb{Q}^5$ be a solution to the system of equations. Then $r_1, r_5 \neq 0$ and the following equality holds in $\mathbb{Q}[x]$:

$$(r_1x^4 + r_2x^3 + r_3x^2 + r_4x + r_5)(r_5x^4 + r_4x^3 + r_3x^2 + r_2x + r_1) = f(x),$$

where

$$f(x) = x^8 + 2x^7 + 3x^6 + 4x^5 + 5x^4 + 4x^3 + 3x^2 + 2x + 1.$$

Now $f(x) = (x^4 + x^3 + x^2 + x + 1)^2$ and $x^4 + x^3 + x^2 + x + 1$ is irreducible in $\mathbb{Q}[x]$ (as is any polynomial $x^{p-1} + \cdots + x + 1$ for prime p.) It follows that

$$r_1x^4 + r_2x^3 + r_3x^2 + r_4x + r_5 = r_1(x^4 + x^3 + x^2 + x + 1),$$

and hence that $r_1 = r_2 = r_3 = r_4 = r_5$. Because $r_1r_5 = 1$ we must have $r_1 = r_2 = r_3 = r_4 = r_5 = 1$ or -1.

Note. A similar result clearly holds for the system of p equations $\sum_{j=1}^{k} r_j r_{p-k+j} = k$, $k = 1, 2, \ldots, p$ for p prime.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles, books, and other materials 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.

Wilford, John Noble, In medieval architecture, signs of advanced math, *New York Times* (27 February 2007) D2, `http://www.nytimes.com/2007/02/27/science/27math.html` . Lu, Peter J., and Paul J. Steinhardt, Decagonal and quasi-crystalline tilings in medieval Islamic architecture, *Science* 315 (23 February 2007) 1106–1110, `http://www.sciencemag.org/cgi/content/full/315/5815/1106` ; supporting online material (8 figures), `http://www.sciencemag.org/cgi/content/full/315/5815/1106/DC1` . Makovicky, Emil, 800-year-old pentagonal tiling from Maragha, Iran, and the new varieties of aperiodic tiling it inspired, in *Fivefold Symmetry*, edited by I. Hargittai, Singapore, World Scientific, 1991, 67–86.

The strip and wallpaper patterns that occur in the tilings at the Alhambra in Spain—and that inspired Maurits Escher—were documented long ago by mathematicians. I visited Iran in 1971 but passed up a chance to visit Isfahan, where perhaps I would have visited the Darb-i Imam shrine (1453) and noticed a curious decagonal tiling there. But that was before Roger Penrose discovered in 1973 two tiles ("kite" and "dart") that can tile the plane nonperiodically with fivefold rotational symmetry. In 1984 Paul Steinhardt named the three-dimensional analogues of such patterns *quasicrystals*, and others showed that quasicrystals really do occur in some metallic alloys. From examination of a few hundred medieval Islamic patterns, physicists Lu and Steinhardt offer a convincing explanation that a vast variety of them were constructed from five tile shapes. In particular, the pattern at the Darb-i Imam shrine is almost perfectly quasicrystalline (11 "errors" out of 3700 tiles). Lu and Steinhardt cite this defect as the kind that "an artist could have made inadvertently." Perhaps, but they should know better: Islamic artworks tend to contain deliberate "errors" because only Allah is perfect. I am glad for the discovery in Isfahan, and for the photos of the beautiful pattern at the shrine there, as I suppose now that I will never get there.

Meyerson, Michael, *Political Numeracy: Mathematical Perspectives on Our Chaotic Constitution*, W.W. Norton, 2002; 287 pp, $24.95. ISBN 0-393-04172-7.

Author Meyerson demonstrates that a wide variety of kinds of "mathematics can illuminate various areas of our Constitution" and its interpretation in Supreme Court decisions. The Constitution can be seen as an axiom system of government or as a dynamical system, burdens of proof bring in Type I and Type II errors of statistical decision-making, and of course there are the paradoxes of voting and of methods of apportionment of representatives. Topology, infinity, chaos, theory of types, and even Gödel's Incompleteness Theorem are brought to bear (Meyerson fails to emphasize that the last applies only to systems incorporating arithmetic). Unfortunately, the author does not dig into the basis of Gödel's famous remark at his citizenship hearing that a dictatorship is indeed possible under the Constitution. (Neither "dictatorship" nor "political question" appears in the index; the latter is discussed but receives scant treatment.)

Lock, Kari, Mixing a night out with probability... and making a fortune, *Math Horizons* (February 2007) 8–9. Caitlin, Donald, Oh, New York, bring back those Big Dippers, `http://catlin.casinocitytimes.com/articles/1226.html`

In statistics classes I show students that the local state lottery is the worst game in town, with a net negative payoff of about 50%. Sometimes, though, lottery marketers have some really "valuable" ideas, and it can "profit" us to learn from their inspiration. That's what happened in 1997 to two alumni of a probability course who noted—and took advantage of, to the tune of $100,000 or more—a "special" offered by the New York State Lottery that doubled prizes on Wednesdays. Said one (with a new house and new car), "It shows that paying attention in math class can, in fact, be useful"—which, when you think about it, could be interpreted as a rather backhanded compliment to our discipline. (Thanks to Steve Conrad and J. Laurie Snell.)

Benson, David J., *Music: A Mathematical Offering*, Cambridge University Press, 2006; xiii + 411 pp, $120, $48 (P). ISBN 978-0-521-85387-3, 978-0-521-61999-8. Updated version with further mathematical appendices free at `http://www.maths.abdn.ac.uk/~bensondj/html/maths-music.html` . Harkleroad, Leon, *The Math Behind the Music*, Cambridge University Press, 2006; xiv + 143 pp + audio CD, $70, $24.99 (P). ISBN 978-0-521-81095-7, 978-0-521-00935-5.

Same year, same publisher, two very different books. Benson's book is based on a mathematics course. The prerequisites are not stated, but Fourier series, Fourier integrals, distributions, and Bessel functions appear in Chapter 2; so it's a good thing that Benson advises not reading the book sequentially but skipping around. The chapters treat waves, Fourier theory, instruments, consonance/dissonance, scales/temperaments, digital music, synthesizers, and symmetry; only in the consonance and scales chapters can the reader avoid postcalculus mathematics. Harkleroad's book grew out of a course with no prerequisites in mathematics or music, hence it is "interdisciplinary" in its current academically-correct meaning (you need no background in anything). Unlike Benson's book, Harkleroad's has no exercises. Its chapters cover pitch, tuning, transpositions/retrogrades/inversions, change-ringing, stochastic music, $1/f$ music, dance, and previous bad attempts to "mix mathematics and music." The mathematics discussed is superposition of sine waves, group theory, and tree diagrams of probabilities. The accompanying CD helps bring the subject alive and would be valuable to anyone teaching any kind of course on mathematics and music. In short, Benson's book is for mathematics majors and Harkleroad's is for liberal arts students and general-interest readers; everyone interested in mathematical aspects of music should listen to Harkleroad's CD.

Applegate, David L., Robert E. Bixby, Vašek Chvátal, and William J. Cook, *The Traveling Salesman Problem: A Computational Study*, Princeton University Press, 2006; xi + 593 pp, $45; free Concorde software with supporting documentation at `http://www.tsp.gatech.edu` . ISBN 978-0-691-12003-8.

The back cover rightfully declares this book to be "the definitive book on the subject" and "an essential resource" in this area. The first two chapters, on the history and applications of the problem, are broad-ranging, nontechnical, and engrossing. The book treats only the symmetric TSP (same cost $x \to y$ as $y \to x$). Subsequent chapters explore the connections with linear programming, pursue in detail the cutting-plane approach, specialize the branch-and-cut algorithm, explore heuristic methods, and relate computational experience. The authors' Concorde software is responsible for the largest solved instance of TSP, with 86,000 cities.

Steiglitz, Ken, *Snipers, Shills, and Sharks: eBay and Human Behavior*, Princeton University Press, 2007; xix + 270 pp, $29.95. ISBN 978-0-691-12713-2. Yang, I., and B. Kahng, Bidding process in online auctions and winning strategy: Rate equation approach, *Physical Review E* 73 (2006) 067101, `http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=PLEEE8000073000006067101000001&idtype=cvips&prog=normal` .

What guidance does mathematics give about online (and other) auctions? These works offer data and empirical observations (e.g., there is much overbidding). Steiglitz's book relegates the mathematics to 75 pp of appendices about optimality in auctions (calculus and mathematical probability needed). The main conclusion of both works: Sniping (bidding at the last moment) is a good (rational and effective), if not optimal, strategy to win an auction.

NEWS AND LETTERS

33rd United States of America Mathematical Olympiad April 18 and 19, 2006

edited by Zuming Feng and Cecil Rousseau

PROBLEMS

1. Let p be a prime number and let s be an integer with $0 < s < p$. Prove that there exist integers m and n with $0 < m < n < p$ and
$$\left\{\frac{sm}{p}\right\} < \left\{\frac{sn}{p}\right\} < \frac{s}{p}$$
if and only if s is not a divisor of $p - 1$.

   (For x a real number, let $\lfloor x \rfloor$ denote the greatest integer less than or equal to x, and let $\{x\} = x - \lfloor x \rfloor$ denote the fractional part of x.)
2. For a given positive integer k find, in terms of k, the minimum value of N for which there is a set of $2k + 1$ distinct positive integers that has sum greater than N but every subset of size k has sum at most $N/2$.
3. For integral m, let $p(m)$ be the greatest prime divisor of m. By convention, we set $p(\pm 1) = 1$ and $p(0) = \infty$. Find all polynomials f with integer coefficients such that the sequence $\{p(f(n^2)) - 2n\}_{n \geq 0}$ is bounded above. (In particular, this requires $f(n^2) \neq 0$ for $n \geq 0$.)
4. Find all positive integers n such that there are $k \geq 2$ positive rational numbers $a_1, a_2, \ldots, a_k$ satisfying $a_1 + a_2 + \cdots + a_k = a_1 \cdot a_2 \cdots a_k = n$.
5. A mathematical frog jumps along the number line. The frog starts at 1, and jumps according to the following rule: if the frog is at integer n, then it can jump either to $n + 1$ or to $n + 2^{m_n + 1}$ where 2^{m_n} is the largest power of 2 that is a factor of n. Show that if $k \geq 2$ is a positive integer and i is a nonnegative integer, then the minimum number of jumps needed to reach $2^i k$ is greater than the minimum number of jumps needed to reach 2^i.
6. Let $ABCD$ be a quadrilateral, and let E and F be points on sides AD and BC, respectively, such that $AE/ED = BF/FC$. Ray FE meets rays BA and CD at S and T, respectively. Prove that the circumcircles of triangles SAE, SBF, TCF, and TDE pass through a common point.

SOLUTIONS

Note: For interested readers, the editors recommend the *USA and International Mathematical Olympiads 2006*. There, many of the problems are presented together with a collection of varied solutions developed by the examination committees, contestants, and experts, during or after the contests.

1. First suppose that s is a divisor of $p - 1$; write $d = (p - 1)/s$. As x varies among $1, 2, \ldots, p - 1$, $\{sx/p\}$ takes the values $1/p, 2/p, \ldots, (p - 1)/p$, once each in some order. The possible values with $\{sx/p\} < s/p$ are precisely $1/p, \ldots,$ $(s - 1)/p$. From the fact that $\{sd/p\} = (p - 1)/p$, we realize that the values

$$\left\{\frac{sx}{p}\right\} = \frac{p-1}{p}, \frac{p-2}{p}, \dots, \frac{p-s+1}{p}$$

occur for $x = d, 2d, \dots, (s-1)d$ (which are all between 0 and p), and so the values

$$\left\{\frac{sx}{p}\right\} = \frac{1}{p}, \frac{2}{p}, \dots, \frac{s-1}{p}$$

occur for $x = p-d, p-2d, \dots, p-(s-1)d$, respectively. From this it is clear that m and n cannot exist as requested.

Conversely, suppose that s is not a divisor of $p-1$. Put $m = \lceil p/s \rceil$; then m is the smallest positive integer such that $\{ms/p\} < s/p$, and in fact $\{ms/p\} = (ms-p)/p$. However, we cannot have $\{ms/p\} = (s-1)/p$ or else $(m-1)s = p-1$, contradicting our hypothesis that s does not divide $p-1$. Hence the unique $n \in \{1, \dots, p-1\}$ for which $\{nx/p\} = (s-1)/p$ has the desired properties (since the fact that $\{nx/p\} < s/p$ forces $n \geq m$, but $m \neq n$).

2. The minimum is $N = 2k^3 + 3k^2 + 3k$. The set $\{k^2+1, k^2+2, \dots, k^2+2k+1\}$ has sum $2k^3 + 3k^2 + 3k + 1 = N + 1$ which exceeds N, but the sum of the k largest elements is only $(2k^3 + 3k^2 + 3k)/2 = N/2$. Thus this N is such a value.

   Suppose $N < 2k^3 + 3k^2 + 3k$ and there are positive integers $a_1 < a_2 < \cdots < a_{2k+1}$ with $a_1 + a_2 + \cdots + a_{2k+1} > N$ and $a_{k+2} + \cdots + a_{2k+1} \leq N/2$. Then

$$\begin{aligned}&(a_{k+1}+1) + (a_{k+1}+2) + \cdots + (a_{k+1}+k)\\ &\leq a_{k+2} + \cdots + a_{2k+1} \leq \frac{N}{2} < \frac{2k^3+3k^2+3k}{2}.\end{aligned}$$

   This rearranges to give $2ka_{k+1} \leq N - k^2 - k$ and $a_{k+1} < k^2 + k + 1$. Hence $a_{k+1} \leq k^2 + k$. Combining these we get $2(k+1)a_{k+1} \leq N + k^2 + k$. We also have

$$(a_{k+1} - k) + \cdots + (a_{k+1} - 1) + a_{k+1} \geq a_1 + \cdots + a_{k+1} > \frac{N}{2}$$

   or $2(k+1)a_{k+1} > N + k^2 + k$. This contradicts the previous inequality, hence no such set exists for $N < 2k^3 + 3k^2 + 3k$ and the stated value is the minimum.

3. The polynomial f has the required properties if and only if

$$f(x) = c(4x - a_1^2)(4x - a_2^2)\cdots(4x - a_k^2), \qquad (1)$$

   where $a_1, a_2, \dots, a_k$ are odd positive integers and c is a nonzero integer. It is straightforward to verify that polynomials given by (1) have the required property. If p is a prime divisor of $f(n^2)$ but not of c, then $p|(2n - a_j)$ or $p|(2n + a_j)$ for some $j \leq k$. Hence $p - 2n \leq \max\{a_1, a_2, \dots, a_k\}$. The prime divisors of c form a finite set and do affect whether or not the given sequence is bounded above. The rest of the proof is devoted to showing that any f for which $\{p(f(n^2)) - 2n\}_{n\geq 0}$ is bounded above is given by (1).

   Let $\mathbb{Z}[x]$ denote the set of all polynomials with integral coefficients. Given $f \in \mathbb{Z}[x]$, let $\mathcal{P}(f)$ denote the set of those primes that divide at least one of the numbers in the sequence $\{f(n)\}_{n\geq 0}$. The solution is based on the following lemma.

   LEMMA. *If $f \in \mathbb{Z}[x]$ is a nonconstant polynomial then $\mathcal{P}(f)$ is infinite.*

   Proof. Repeated use will be made of the following basic fact: if a and b are distinct integers and $f \in \mathbb{Z}[x]$, then $a - b$ divides $f(a) - f(b)$. If $f(0) = 0$, then p divides $f(p)$ for every prime p, so $\mathcal{P}(f)$ is infinite. If $f(0) = 1$, then every prime divisor p of $f(n!)$ satisfies $p > n$. Otherwise p divides $n!$, which in turn divides $f(n!) - f(0) = f(n!) - 1$. This yields $p|1$, which is false. Hence $f(0) = 1$ implies that $\mathcal{P}(f)$ is infinite. To complete the proof, set $g(x) = f(f(0)x)/f(0)$ and observe that $g \in \mathbb{Z}[x]$ and $g(0) = 1$. The preceding argument shows that $\mathcal{P}(g)$ is infinite, and it follows that $\mathcal{P}(f)$ is infinite. ■

Suppose $f \in \mathbb{Z}[x]$ is nonconstant and there exists a number M such that $p(f(n^2)) - 2n \leq M$ for all $n \geq 0$. Application of the lemma to $f(x^2)$ shows that there is an infinite sequence of distinct primes $\{p_j\}$ and a corresponding infinite sequence of nonnegative integers $\{k_j\}$ such that $p_j | f(k_j^2)$ for all $j \geq 1$. Consider the sequence $\{r_j\}$ where

$$r_j = \min\{k_j \pmod{p_j},\ p_j - k_j \pmod{p_j}\}.$$

Then $0 \leq r_j \leq (p_j - 1)/2$ and $p_j | f(r_j^2)$. Hence

$$2r_j + 1 \leq p_j \leq p(f(r_j^2)) \leq M + 2r_j,$$

so $1 \leq p_j - 2r_j \leq M$ for all $j \geq 1$. It follows that there is an integer a_1 such that $1 \leq a_1 \leq M$ and $a_1 = p_j - 2r_j$ for infinitely many j. Let $m = \deg f$. Then

$$p_j | 4^m f((p_j - a_1)/2)^2) \quad \text{and} \quad 4^m f((x - a_1)/2)^2) \in \mathbb{Z}[x].$$

Consequently, $p_j | f((a_1/2)^2)$ for infinitely many j, which shows that $(a_1/2)^2$ is a zero of f. Since $f(n^2) \neq 0$ for $n \geq 0$, a_1 must be odd. Then $f(x) = (4x - a_1^2)g(x)$ where $g \in \mathbb{Z}[x]$. (See the note below.) Observe that $\{p(g(n^2)) - 2n\}_{n \geq 0}$ must be bounded above. If g is constant, we are done. If g is nonconstant, the argument can be repeated to show that f is given by (1).

Note: The step that gives $f(x) = (4x - a_1^2)g(x)$ where $g \in \mathbb{Z}[x]$ follows immediately using a lemma of Gauss. The use of such an advanced result can be avoided by first writing $f(x) = r(4x - a_1^2)g(x)$ where r is rational and $g \in \mathbb{Z}[x]$. Then continuation gives $f(x) = c(4x - a_1^2) \cdots (4x - a_k^2)$ where c is rational and the a_i are odd. Consideration of the leading coefficient shows that the denominator of c is 2^s for some $s \geq 0$ and consideration of the constant term shows that the denominator is odd. Hence c is an integer.

4. The answer is $n = 4$ or $n \geq 6$.

I. First, we prove that each $n \in \{4, 6, 7, 8, 9, \ldots\}$ satisfies the condition.

(1) If $n = 2k \geq 4$ is *even*, setting $(a_1, a_2, \ldots, a_k) = (k, 2, 1, \ldots, 1)$, we obtain

$$a_1 + a_2 + \cdots + a_k = k + 2 + 1 \cdot (k - 2) = 2k = n,$$

and $a_1 \cdot a_2 \cdots a_k = 2k = n$.

(2) If $n = 2k + 3 \geq 9$ is *odd*, setting $(a_1, a_2, \ldots, a_k) = \left(k + \frac{3}{2}, \frac{1}{2}, 4, 1, \ldots, 1\right)$, we obtain

$$a_1 + a_2 + \cdots + a_k = k + \frac{3}{2} + \frac{1}{2} + 4 + (k - 3) = 2k + 3 = n,$$

and $a_1 \cdot a_2 \cdots a_k = \left(k + \frac{3}{2}\right) \cdot \frac{1}{2} \cdot 4 = 2k + 3 = n$.

(3) A very special case is $n = 7$, in which we set $(a_1, a_2, a_3) = (4/3, 7/6, 9/2)$. It is trivial to check that $a_1 + a_2 + a_3 = a_1 a_2 a_3 = 7 = n$.

II. Second, we prove by contradiction that each $n \in \{1, 2, 3, 5\}$ *fails to satisfy* the condition.

Suppose, on the contrary, that there is a set of $k \geq 2$ positive rational numbers whose sum and product are both $n \in \{1, 2, 3, 5\}$. By the **AM-GM Inequality**, we have

$$n^{1/k} = \sqrt[k]{a_1 \cdot a_2 \cdots a_k} \leq \frac{a_1 + a_2 + \cdots + a_k}{k} = \frac{n}{k},$$

which gives $n \geq k^{\frac{k}{k-1}} = k^{1+\frac{1}{k-1}}$. From the above inequality, it is easy to check that $n > 5$ whenever $k = 3, 4$, or $k \geq 5$. This proves that none of the integers 1, 2, 3, or 5 can be represented as the sum and, at the same time, as the product of three or more positive numbers $a_1, a_2, \ldots, a_k$, rational or irrational.

The remaining case $k = 2$ also leads to a contradiction. Indeed, $a_1 + a_2 = a_1 a_2 = n$ implies that $n = a_1^2/(a_1 - 1)$ and thus a_1 satisfies the quadratic

$$a_1^2 - na_1 + n = 0.$$

Since a_1 is supposed to be *rational*, the discriminant $n^2 - 4n$ must be a perfect square, and it is easy to check that this is not the case for $n \in \{1, 2, 3, 5\}$. This completes the proof.

5. Suppose $2^i k$ can be reached in m jumps.

Our approach will be to consider the frog's life as a sequence of leaps of certain lengths. We will prove that by removing the longest leaps from the sequence, we generate a valid sequence of leaps that ends at 2^i. Clearly this sequence will be shorter, since it was obtained by removing leaps. The result will follow.

LEMMA. *If we remove the longest leap in the frog's life (or one of the longest, in case of a tie) the sequence of leaps will still be legitimate.*

Proof. By definition, a leap from n to $n + \nu$ is legitimate if and only if either (a) $\nu = 1$, or (b) $\nu = 2^{m_n+1}$. If all leaps are of length 1, then clearly removing one leap does not make any others illegitimate; suppose the longest leap has length 2^s.

Then we remove this leap and consider the effect on all the other leaps. Take an arbitrary leap starting (originally) at n, with length ν. Then $\nu \leq 2^s$. If $\nu = 1$ the new leap is legitimate no matter where it starts. Say $\nu > 1$. Then $\nu = 2^{m_n+1}$. Now if the leap is before the removed leap, its position is not changed, so $\nu = 2^{m_n+1}$ and it remains legitimate. If it is after the removed leap, its starting point is moved back to $n - 2^s$. Now since $2^{m_n+1} = \nu \leq 2^s$, we have $m_n \leq s - 1$; that is, 2^s does not divide n. Therefore, 2^{m_n} is the highest power of 2 dividing $n - 2^s$, so $\nu = 2^{m_{n-2^s}+1}$ and the leap is still legitimate. This proves the Lemma. ∎

We now remove leaps from the frog's sequence of leaps in decreasing order of length. The frog's path has initial length $2^i k - 1$; we claim that at some point its length is $2^i - 1$.

Let the frog's m leaps have lengths $a_1 \geq a_2 \geq a_3 \geq \cdots \geq a_m$. Define a function f by

$$f(0) = 2^i k$$

$$f(i) = f(i-1) - a_i, 1 \leq i \leq m.$$

Clearly $f(i)$ is the frog's final position if we remove the i longest leaps. Note that $f(m) = 1$—if we remove all leaps, the frog ends up at 1. Let $f(j)$ be the last value of f that is at least 2^i. That is, suppose $f(j) \geq 2^i$, $f(j+1) < 2^i$. Now we have $a_{j+1}|a_k$ for all $k \leq j$ since $\{a_k\}$ is a decreasing sequence of powers of 2. If $a_{j+1} > 2^i$, we have $2^i|a_p$ for $p \leq j$, so $2^i|f(j+1)$. But $0 < f(j+1) < 2^i$, contradiction. Thus $a_{j+1} \leq 2^i$, so, since a_{j+1} is a power of two, $a_{j+1}|2^i$. Since $a_{j+1}|2^i k$ and $a_1, \dots, a_j$, we know that $a_{j+1}|f(j)$, and $a_{j+1}|f(j+1)$. So $f(j+1)$, $f(j)$ are two consecutive multiples of a_{j+1}, and 2^i (another such multiple) satisfies $f(j+1) < 2^i \leq f(j)$. Thus we have $2^i = f(j)$, so by removing j leaps we make a path for the frog that is legitimate by the Lemma, and ends on 2^i.

Now let m be the minimum number of leaps needed to reach $2^i k$. The Lemma and the argument above show that the frog can reach 2^i in only $m - j$ leaps. Since $j > 0$ trivially ($j = 0$ implies $2^i = f(j) = f(0) = 2^i k$) we have $m - j < m$ as desired.

6. Let P be the second intersection of the circumcircles of triangles TCF and TDE. Because the quadrilateral $PEDT$ is cyclic, $\angle PET = \angle PDT$, or

$$\angle PEF = \angle PDC. \tag{2}$$

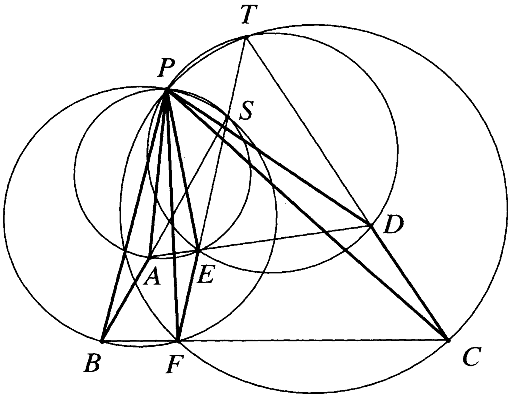
Because the quadrilateral $PFCT$ is cyclic,

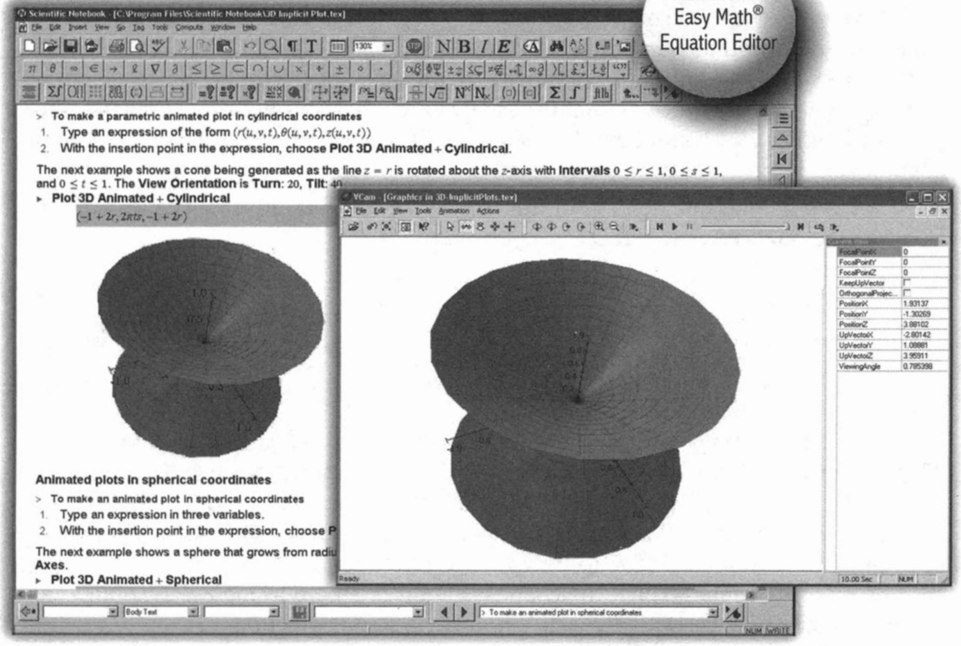
$$\angle PFE = \angle PFT = \angle PCT = \angle PCD. \tag{3}$$

By equations (2) and (3), it follows that triangle PEF is similar to triangle PDC. Hence $\angle FPE = \angle CPD$ and $PF/PE = PC/PD$. Note also that $\angle FPC = \angle FPE + \angle EPC = \angle CPD + \angle EPC = \angle EPD$. Thus, triangle EPD is similar to triangle FPC. Another way to say this is that there is a spiral similarity centered at P that sends triangle PFE to triangle PCD, which implies that there is also a spiral similarity, centered at P, that sends triangle PFC to triangle PED, and vice versa. Because $AE/ED = BF/FC$, points A and B are obtained by extending corresponding segments of two similar triangles PED and PFC, namely, DE and CF, by the identical proportion. We conclude that triangle PDA is similar to triangle PCB, implying that triangle PAE is similar to triangle PBF. Therefore, as shown before, we can establish the similarity between triangles PBA and PFE, implying that

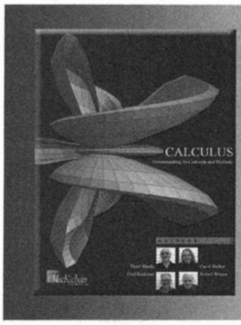
$$\angle PBS = \angle PBA = \angle PFE = \angle PFS \quad \text{and} \quad \angle PAB = \angle PEF.$$

The first equation above shows that $PBFS$ is cyclic. The second equation shows that $\angle PAS = 180° - \angle BAP = 180° - \angle FEP = \angle PES$; that is, $PAES$ is cyclic. We conclude that the circumcircles of triangles SAE, SBF, TCF, and TDE pass through point P.





File Edit Insert View Go Tag Tools Compute Window Help
130%
Body Text
> To make an animated plot in spherical coordinates
NUM WRITE
File Edit View Tools Animation Actions
Ready
10.00 Sec
NUM

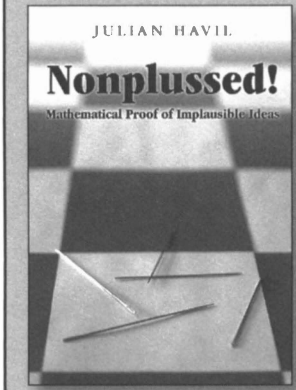


CALCULUS
MacKichan

Nonplussed!

Mathematical Proof of Implausible Ideas

Julian Havil



"*Nonplussed!*, as the title suggests, is a marvelous study of some two dozen choice mathematical problems that boggle the mind. Unlike so many books on recreational math, Havil doesn't hesitate to give crystal-clear proofs and their necessary equations. . . . No one interested in recreational mathematics on an intermediate advanced level should pass up this stimulating, delightful volume."
—Martin Gardner

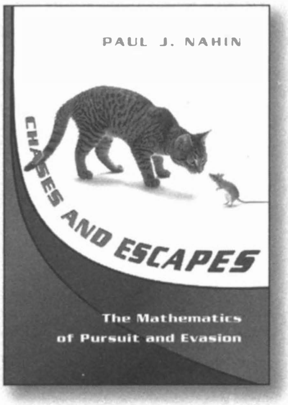
216 pages. 18 halftones. 143 line illus. 6 x 9.

Cloth $24.95 978-0-691-12056-0

Chases and Escapes

The Mathematics of Pursuit and Evasion

Paul J. Nahin



"I know of no better way to grasp the basic concepts of calculus than to study pursuit-and-escape problems. Paul Nahin has made a superb survey of the vast field of such problems. . . . Not only does he make clear the required differential equations, but he traces each problem's colorful history. No book on the topic could be more definitive or a greater pleasure to read."
—Martin Gardner

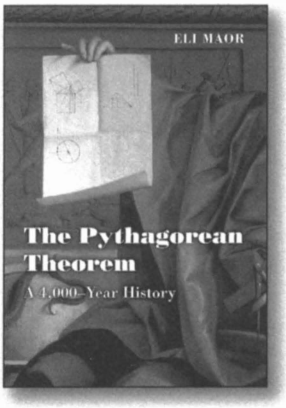
272 pages. 5 halftones. 67 line illus. 6 x 9.

Cloth $24.95 £15.95 978-0-691-12514-5 Due July

The Pythagorean Theorem

A 4,000-Year History

Eli Maor



"There's a lot more to the Pythagorean theorem than $a^2 + b^2 = c^2$, and you'll find it all in Eli Maor's new book. Destined to become a classic, this book is written with Maor's usual high level of skill, scholarship, and attention to detail. He's also got a sense of humor that will please a range of readers. As we used to say in the 1950s, 'Miss it and be square!'"
—Paul J. Nahin, author of *An Imaginary Tale: The Story of the Square Root of -1*

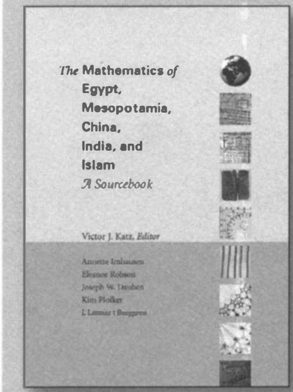
288 pages. 8 color plates. 141 line illus. 2 tables. 6 x 9.

Cloth $24.95 978-0-691-12526-8 Due June

The Mathematics of Egypt, Mesopotamia, China, India, and Islam

A Sourcebook

Edited by Victor J. Katz


The Mathematics of Egypt, Mesopotamia, China, India, and Islam
A Sourcebook
Victor J. Katz, Editor

Addressing a critical gap in the mathematics literature in English, this book is an essential resource for anyone with at least an undergraduate degree in mathematics who wants to learn about non-Western mathematical developments, and how they helped shape and enrich world mathematics.

712 pages. 76 halftones. 210 line illus. 42 tables. 7 x 10.

Cloth $75.00 £48.95 978-0-691-11485-4 Due August

PRINCETON UNIVERSITY PRESS
800.777.4726
math.press.princeton.edu

CONTENTS

ARTICLES

163 Solving the Ladder Problem on the Back of an Envelope, *by Dan Kalman*

182 Proof Without Words: A Graph Theoretic Decomposition of Binomial Coefficients, *by Joe DeMaio*

183 The Recreational Gambler: Paying the Price for More Time at the Table, *by Joseph Bak*

195 Proof Without Words: Every Triangle Can Be Subdivided into Six Isosceles Triangles, *by Ángel Plaza*

196 The Probability of Relatively Prime Polynomials, *by Arthur T. Benjamin and Curtis D. Bennett*

NOTES

203 Fitting One Right Triangle in Another, *by Charles H. Jepsen and Valeria Vulpe*

207 Determinants of Matrices over the Integers Modulo m, *by Jody M. Lockhart and William P. Wardlaw*

215 The Classification of Similarities: A New Approach, *by Aad Goddijn and Wim Pijls*

219 Perfect Matchings, Catalan Numbers, and Pascal's Triangle, *by Tomislav Došlić*

226 On Candido's Identity, *by Claudi Alsina and Roger B. Nelsen*

228 Monotonic Convergence to e via the Arithmetic-Geometric Mean, *by József Sándor*

PROBLEMS

230 Proposals 1771–1775

231 Quickies 971–972

231 Solutions 1746–1750

236 Answers 971–972

REVIEWS

237

NEWS AND LETTERS

239 33rd Annual United States of America Mathematical Olympiad

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036