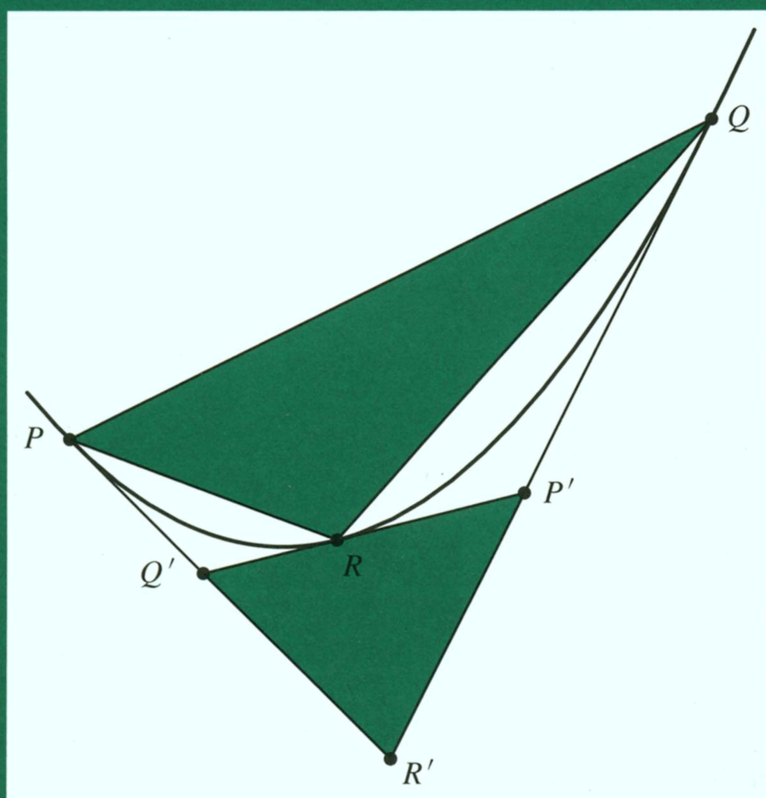
Vol. 81, No. 2, April 2008

MATHEMATICS MAGAZINE



Archimedean Quadrature Redux: The Two Triangle Theorem

- Three Transcendental Numbers
- An Exploration of Pick's Theorem in Space
- Revisiting James Watt's Linkage
- Extra Inning Baseball Games

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*.

Cover: Is it a coincidence that the upper triangle on the cover seems to be about twice the area of the lower triangle? Never abide coincidence as an explanation when there may be a theorem lurking about. Our lead article by Larry Cusick provides an elegant proof of this simple fact, and many more intriguing observations.

AUTHORS

Larry W. Cusick received his PhD from the University of California at San Diego in 1981. He is currently professor of mathematics at the California State University at Fresno, where he has been since 1983. When not teaching, conducting research, or enjoying time with his wife and two sons, he can usually be found playing fiddle or flute at any local Irish music session.

Gregory Dresden received his BS from Stanford, his MA from the University of Wisconsin, and his PhD from the University of Texas. He now finds himself at Washington and Lee University in Virginia, where he teaches algebra and number theory (among other things) and indulges in his fascination with the Fibonacci numbers. In his spare time he ambles along the Appalachian Trail, writes long letters in Spanish, and delights in the company of his lovely wife and charming daughter.

Howard Iseri has been on the faculty of Mansfield University of Pennsylvania since 1991, and he received his PhD from the University of California at Davis in 1992. While Howard is mostly interested in the geometry of polyhedral surfaces, he stumbled across a computer science article on Pick's theorem and got sidetracked for awhile.

Chris Sangwin read mathematics at Brasenose College, Oxford and his PhD was in the mathematics of control systems at the University of Bath. He is currently a lecturer in the School of Mathematics at the University of Birmingham, UK. His research interests include control theory and online mathematics assessment.

Darren Glass teaches mathematics at Gettysburg College. He received his BA from Rice University and his PhD from the University of Pennsylvania. Despite bouncing around the country quite a bit, Darren has never given up his allegiance to the Atlanta Braves, whom he first started following in 1982 when he was eight years old. He has continued to follow them through the highs of many consecutive division titles and the lows of the 1980's. When he is not researching baseball, he likes to work on Galois Theory, Number Theory, and applications to cryptography. He also regularly reviews books for MAA Online.

Phil Lowry teaches at Normandale College. He received his AB and MBA from Harvard University, and will defend his PhD dissertation in 2008 at City University of New York. Before his academic career, he was an infantry captain in the US Army, and has traveled to ninety countries. He is the author of *Green Cathedrals*, the definitive study of all major league baseball parks.

Vol. 81, No. 2, April 2008

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
1529 Eighteenth St. NW
Washington DC 20036

Phone: (866) 821-1221
Fax: (202) 387-1208
E-mail: advertising@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

Archimedean Quadrature Redux

LARRY W. CUSICK
California State University, Fresno
Fresno, CA 93740
larryc@csufresno.edu

Archimedes' use of Eudoxos' method of exhaustion to determine the area bounded by a parabolic arc and a line segment was a crowning achievement in Greek mathematics. The promise of the method, so apparent to us now, seems to have died with Archimedes, only to rise again in different form some 1900 years later with the modern calculus. Archimedes' result though is not just about computing an area. It is about comparing a parabolic area with a related triangular area. That is, there is a geometric content in the comparison that is interesting in its own right. In recognition of this, there have been several generalizations discovered more recently that highlight the geometry using methods of modern analysis ([**1**], [**2**], [**4**], [**6**], [**7**], and [**14**]).

In this short article we would like to make the case that Archimedes' area comparisons deserve more attention, not so much because of his methods, but rather because of the interesting geometric content of the comparisons and the new questions they suggest. We feel that there are more results to be had, and present a few here with some speculation on further research directions.

History

Eudoxos of Cnidos (408–355 BC, in modern day Turkey) is generally credited with the discovery of the so-called method of exhaustion for determining the volumes of a pyramid and cone [**9**]. The ancient Greeks were fond of comparisons between volumes. For example, Eudoxos showed that the volume of a pyramid, respectively a cone, was one-third of the volume of the prism, respectively the cylinder, with like base and height. Although Eudoxos did not have the modern apparatus of limits, his technique amounted to approximating the volume by many simpler figures whose volumes are understood and essentially passing to a limit.

The apex of the method of exhaustion comes with Archimedes of Syracuse (287–212 BC). Archimedes deftly used the method to prove several area and volume (circles and spheres) comparisons ([**5**], [**8**], [**9**], [**13**]). One might argue that parabolic curves are the natural next step. And a lesser mathematician of the time may have passed given the difficulty and apparent lack of obvious applications. Archimedes, however, solved the area problem for parabolas in his two related theorems (1) the quadrature of the parabola and the (2) squaring of the parabola. Both theorems compare a parabolic area to that of related triangle areas and can be found in his *Quadrature of the Parabola* and *The Method*.

The setting for Archimedes' theorems is a region in the plane bounded by a straight line segment and a parabolic arc, meeting at respective points P and Q (FIGURE 1). In his quadrature theorem, Archimedes locates the point R on the parabolic arc that is a maximum distance, measured perpendicularly, from the line segment $\overline{PQ}$ and calls this point the *vertex* of the parabolic arc. (The tangent line to the parabolic arc at the vertex R is parallel to $\overline{PQ}$.) Quadrature states that the area bounded by the parabolic arc and the line segment $\overline{PQ}$ is equal to $\frac{4}{3}$ of the area of $\triangle PQR$.

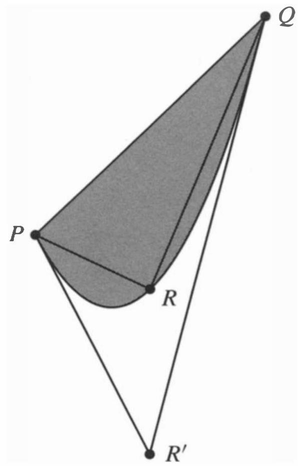


Figure 1 Archimedes' Quadrature and Squaring of the Parabola

For his squaring of the parabola, Archimedes compares the area bounded by the parabolic arc and $\overline{PQ}$ to that of the so-called Archimedes triangle, $\triangle PQR'$ where R' is the intersection point of the two tangent lines to the parabola at P and Q respectively. He then goes on to prove that the parabolic area is $\frac{2}{3}$ of the area enclosed by this triangle.

Archimedes' methods for proving his theorems relied on several properties of the parabola that are not common knowledge today. (For full details of the proofs, see [**5**, pp. 239–242] and [**13**, pp. 51–62].) But we do have the powerful tools of analysis that can allow us to go further.

Archimedean quadrature and squaring for analytic plane curves

The context for our generalization will be *analytic plane curves*. A curve $\mathcal{C}$ will be called *analytic of order n* at a point $R \in \mathcal{C}$ if there is a coordinate system at R with the two respective axes tangent and normal to $\mathcal{C}$ at R so that $\mathcal{C}$ is the graph of an analytic function

$$f(x) = c_n x^n + c_{n+1} x^{n+1} + \cdots,$$

where $c_n \neq 0$. For our purposes, n will always be an even positive number. In the language of [3, p. 17], the curve $\mathcal{C}$ has n-fold contact with its tangent line at R.

Note that a point R on a curve $\mathcal{C}$ is of order 2 precisely when the curvature of $\mathcal{C}$ is non-zero at R (because the curvature function is given by

$$\kappa(x) = f''(x)/(1 + f'(x)^2)^{3/2}).$$

And consequently, every point on a parabolic arc is of order 2.

Let $T_R\mathcal{C}$ denote the tangent line to $\mathcal{C}$ at R. We will consider the family of triangles $\triangle PQR$ for which P and Q are on $\mathcal{C}$, where R lies between P and Q on $\mathcal{C}$, and $\overline{PQ}$ is parallel to $T_R\mathcal{C}$. This situation can be pictured as FIGURE 1. Let A denote the area bounded by the curve and the segment $\overline{PQ}$ (the shaded area in FIGURE 1) and T the area of $\triangle PRQ$. Archimedes' quadrature of the parabola states that $A/T = 4/3$ if C is a parabolic arc. While we would not expect his theorem to be true for other curves, we could ask if his theorem holds "in the limit" for analytic curves. We find that the answer is yes with some additional consideration in the case of zero curvature. The following generalization was proved in the case $n = 1$ in [**6**, Theorem 1].

THEOREM 1. (GENERALIZED ARCHIMEDEAN QUADRATURE) *Suppose $\mathcal{C}$ is an analytic plane curve and $R \in \mathcal{C}$ is a point of order $2n$, with $n \geq 1$, and that A and T are as described in the previous paragraph, then*

$$\lim_{\substack{PQ \to 0 \\ \overline{PQ} \| T_R\mathcal{C}}} \frac{A}{T} = \frac{4n}{2n+1},$$

where the limit is taken over points P, $Q \in \mathcal{C}$ that are on opposite sides of R and $\overline{PQ}$ is parallel to the tangent line $T_R\mathcal{C}$.

Proof. By assumption, $\mathcal{C}$ is the graph of $f(x) = c_{2n}x^{2n} + c_{2n+1}x^{2n+1} + \cdots$ with $c_{2n} \neq 0$. In this coordinate system, $R = (0, 0)$, $P = (a, f(a))$ and $Q = (b, f(b))$, $a < 0 < b$, and $f(a) = f(b)$. (The last point is because the line $\overline{PQ}$ is assumed to be parallel to the tangent line to $\mathcal{C}$ at R which is the horizontal axis in our coordinate system.) We may also assume $f(x) > 0$ for $x \neq 0$. By the inverse function theorem, we may write $b = \gamma(a)$ for some function γ. (See FIGURE 2.)

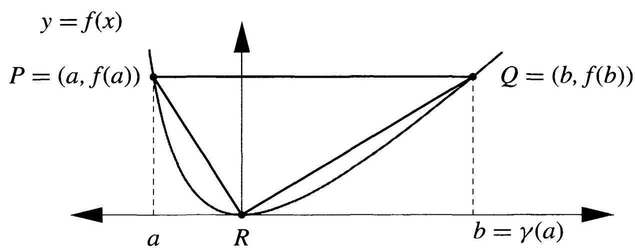


Figure 2 Generalized Archimedean Quadrature

We will need to know how to compute $\gamma'(a)$ later in the proof, and this is also provided by the inverse function theorem:

$$\gamma'(a) = \frac{f'(a)}{f'(\gamma(a))}. \tag{1}$$

The area bounded by $\mathcal{C}$ and $\overline{PQ}$ is equal to

$$A = f(a)(b - a) - \int_a^b f(x)\,dx.$$

The area enclosed by $\triangle PQR$ is

$$T = \frac{1}{2} f(a)(b - a). \tag{2}$$

Thus, the ratio A/T is

$$\frac{A}{T} = 2\frac{f(a)(b-a) - \int_a^b f(x)\,dx}{f(a)(b-a)} = 2 - 2\left(\frac{\int_a^b f(x)\,dx}{f(a)(b-a)}\right). \tag{3}$$

Letting $b = \gamma(a)$, noting that $f(\gamma(a)) = f(a)$ and using L'Hospital's rule along with the Fundamental Theorem of Calculus, we get

$$\lim_{a\to 0^-} \frac{\int_a^b f(x)\,dx}{(b-a)f(a)} = \lim_{a\to 0^-} \frac{\int_a^{\gamma(a)} f(x)\,dx}{f(a)(\gamma(a)-a)} = \lim_{a\to 0^-} \frac{f(a)(\gamma'(a)-1)}{f'(a)(\gamma(a)-a) + f(a)(\gamma'(a)-1)} = \lim_{a\to 0^-} \frac{1}{\frac{f'(a)(\gamma(a)-a)}{f(a)(\gamma'(a)-1)} + 1}. \tag{4}$$

Now focusing on the denominator term, we see that

$$\lim_{a\to 0^-} \frac{f'(a)(\gamma(a)-a)}{f(a)(\gamma'(a)-1)} = \lim_{a\to 0^-} \left(\frac{2nc_{2n}a^{2n-1} + \cdots}{c_{2n}a^{2n} + \cdots}\right)\left(\frac{\gamma(a)-a}{\gamma'(a)-1}\right) = \lim_{a\to 0^-} \left(\frac{2nc_{2n}a^{2n} + \cdots}{c_{2n}a^{2n} + \cdots}\right)\left(\frac{\frac{\gamma(a)}{a} - 1}{\gamma'(a)-1}\right). \tag{5}$$

The first term in the product (5) is easily seen to approach $2n$. As for the second term, we first compute using L'Hospital's rule.

$$\lim_{a\to 0^-} \frac{\gamma(a)}{a} = \lim_{a\to 0^-} \gamma'(a).$$

And by formula (1),

$$\lim_{a\to 0^-} \gamma'(a) = \lim_{a\to 0^-} \frac{f'(a)}{f'(\gamma(a))} \tag{6}$$

$$= \lim_{a\to 0^-} \frac{2nc_{2n}a^{2n-1} + \cdots}{2nc_{2n}\gamma(a)^{2n-1} + \cdots} \tag{7}$$

$$= \lim_{a\to 0^-} \left(\frac{a}{\gamma(a)}\right)^{2n-1} \tag{8}$$

$$= \lim_{a\to 0^-} \left(\frac{1}{\gamma'(a)}\right)^{2n-1}. \tag{9}$$

Upon equating the first and last term in the above string of equalities (6–9), we obtain $(\lim_{a\to 0^-} \gamma'(a))^{2n} = 1$. But $\gamma(a)$ is clearly a decreasing function of a for $a < 0$, so

$$\lim_{a\to 0^-} \frac{\gamma(a)}{a} = \lim_{a\to 0^-} \gamma'(a) = -1,$$

Substituting this into the product (5), gives us

$$\lim_{a\to 0^-} \frac{f'(a)(\gamma(a)-a)}{f(a)(\gamma'(a)-1)} = 2n. \tag{10}$$

This can now be substituted into the fraction (4) to get

$$\lim_{a\to 0^-} \frac{\int_a^b f(x)\,dx}{(b-a)f(a)} = \frac{1}{2n+1}.$$

Finally, we have the last piece to substitute into the ratio (3), completing the proof.

$$\begin{aligned}\lim_{a\to 0^-} \frac{A}{T} &= 2 - \frac{2}{2n+1}\\ &= \frac{4n}{2n+1}.\end{aligned}$$

■

If $2n = 2$, we see that the ratio in Theorem 1 tends to $4/3$. So it would appear that there is not some other curve for which the ratios are a constant value other than $4/3$, except possibly at points of order larger than 2. In fact, this can happen. If $\mathcal{C}$ is the graph of $f(x) = x^{2n}$ and $R = (0, 0)$, then it is easy to check that A/T is constantly $4n/(2n+1)$ for $\overline{PQ}$ parallel to $T_R\mathcal{C}$. Also, the necessity of the condition $\overline{PQ}||T_R\mathcal{C}$ in the limit becomes apparent in the example $f(x) = x^2$. We leave it for the reader to verify that the limit would not exist without this extra condition.

Archimedes' squaring of the parabola can be generalized in a similar way. Using the same hypotheses as in Theorem 1, let R' be the intersection of the two tangents lines to $\mathcal{C}$ at P and Q respectively. (See FIGURE 3.)

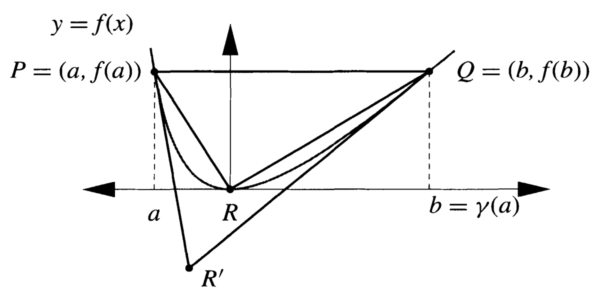


Figure 3 Generalized Archimedean Squaring

Let $\tilde{T}$ be the area enclosed by $\triangle PQR'$. Using the same notation as in the previous proof, we have

$$\frac{A}{\tilde{T}} = \frac{A}{T}\cdot\frac{T}{\tilde{T}}. \tag{11}$$

It is easy to derive

$$\tilde{T} = \frac{f'(a)f'(b)(b-a)^2}{2(f'(a)-f'(b))}.$$

And so using formulas (1) and (10),

$$\begin{aligned}\lim_{a\to 0^-}\frac{T}{\tilde{T}} &= \lim_{a\to 0^-}\frac{f(a)(f'(a)-f'(b))}{f'(a)f'(b)(b-a)}\\ &= \lim_{a\to 0^-}\frac{f(a)(f'(a)-f'(\gamma(a)))}{f'(a)f'(\gamma(a))(\gamma(a)-a)}\\ &= \lim_{a\to 0^-}\frac{f(a)(\gamma'(a)-1)}{f'(a)(\gamma(a)-a)}\\ &= \frac{1}{2n}. \end{aligned} \tag{12}$$

Putting this together with Theorem 1 and (11), we get the generalization to Archimedes' squaring of the parabola.

THEOREM 2. (GENERALIZED ARCIMEDEAN SQUARING) *Given the hypotheses of Theorem 1 and the definition of $\tilde{T}$ as the area enclosed by $\triangle PQR'$,*

$$\lim_{\substack{PQ\to 0\\ \overline{PQ}\|T_R\mathcal{C}}}\frac{A}{\tilde{T}} = \frac{2}{2n+1},$$

where the limit is taken over pairs of points $P, Q \in \mathcal{C}$, on opposite sides of $R \in \mathcal{C}$ and such that $\overline{PQ} \parallel T_R\mathcal{C}$.

It is necessary in the limit that $PQ \parallel T_R\mathcal{C}$ for otherwise the limit may not exist. Also, just as in quadrature, the ratio $A/\tilde{T}$ is constantly $2/(2n+1)$ if $\mathcal{C}$ is the graph of $f(x) = x^{2n}$.

The two triangle theorem

There is another area fact about parabolas that can be gleaned from Archimedes' Theorem. In FIGURE 4, the lines $\overline{PR'}$, $\overline{QR'}$ and $\overline{Q'P'}$ are all tangent to the parabolic arc at P, Q and R respectively. There is no assumption about R other than it lies between P and Q on the arc.

We will now show how Archimedes could have used his squaring of the parabola to prove

$$\frac{\text{Area }\triangle PQR}{\text{Area }\triangle P'Q'R'} = 2. \tag{13}$$

A simplified notation will help in the proof. For the parabolic arc and associated tangent lines pictured in FIGURE 5, (XY) and $[XY]$ will denote the indicated areas.

According to Archimedes' squaring of the parabola, (XY) is two-thirds of the area of $\triangle XYZ$, and so $(XY) = 2[XY]$. Thus

$$\begin{aligned}\frac{\text{Area }\triangle PQR}{\text{Area }\triangle P'Q'R'} &= \frac{(PQ)-(PR)-(QR)}{[PQ]-[PR]-[QR]}\\ &= \frac{2([PQ]-[PR]-[QR])}{[PQ]-[PR]-[QR]}\\ &= 2.\end{aligned}$$

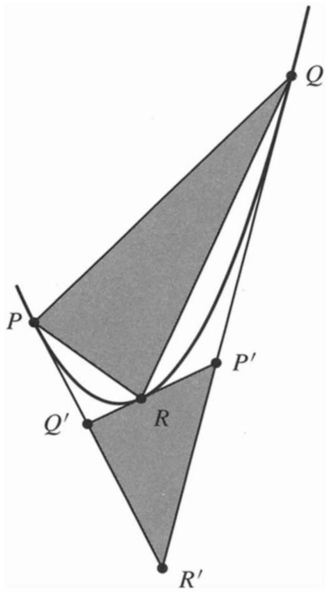


Figure 4 Two Triangle Theorem

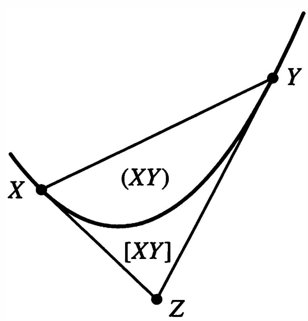


Figure 5 Definitions of (XY) and $[XY]$

So the question is, to what extent does equation 13 transfer to analytic plane curves? The answer is given by the next theorem. The theorem will refer to the same named points P, Q, R, P', Q', R' used above (see FIGURE 4) on an analytic plane curve $\mathcal{C}$. We will also let $T = \text{Area } \triangle PQR$ and $T' = \text{Area } \triangle P'Q'R'$.

THEOREM 3. (TWO TRIANGLE THEOREM) *Referring to the previous paragraph, if $R \in \mathcal{C}$ is a point of order 2, then*

$$\lim_{PQ \to 0} \frac{T}{T'} = 2,$$

where the limit is taken over pairs of points P, $Q \in \mathcal{C}$ on opposite sides of R. If $R \in \mathcal{C}$ is of order $2n$, $n > 1$, then

$$\lim_{\substack{PQ \to 0 \\ PQ \| T_R\mathcal{C}}} \frac{T}{T'} = \frac{2n}{(2n-1)^2},$$

where the limit is taken over points $P, Q \in C$ that are on opposite sides of R and $\overline{PQ}$ is parallel to the tangent line $T_R C$.

Before we set about on the proof of the Two Triangle Theorem, a few remarks are in order. The additional hypotheses in the second part of the theorem, $\overline{PQ} || T_R C$ in the limit, is necessary, for otherwise the limit may not exist. Also, just as before, if C is the graph of $f(x) = x^{2n}$ and $\overline{PQ} || T_R C$, then it is readily verified that $T/T' = 2n/(2n-1)^2$.

Proof of the Two Triangle Theorem. We will prove the second assertion first. Our starting point will be the proof of Theorem 1, and we will freely use the notation and definitions from that proof. It is fairly straightforward to derive the area enclosed by triangle $\triangle P'R'Q'$, assuming $f(a) = f(b)$ (you may use FIGURE 6 with $\overline{PQ}$ tilted so that it is parallel to the horizontal axis),

$$T' = \frac{(f(a)(f'(b) - f'(a)) + f'(a)f'(b)(b-a))^2}{2f'(a)f'(b)\,(f'(a) - f'(b))}. \tag{14}$$

Using this, along with formula (2), we get

$$\frac{T}{T'} = \frac{f(a)(f'(a) - f'(b))f'(a)f'(b)(b-a)}{(f(a)(f'(b) - f'(a)) + f'(a)f'(b)(b-a))^2} \tag{15}$$

If we let $X = f(a)(f'(a) - f'(b))$ and $Y = f'(a)f'(b)(b-a)$, and use the fact $f(a) = f(b)$, then

$$\begin{aligned} \frac{T}{T'} &= \frac{XY}{(-X+Y)^2} \\ &= \frac{\frac{X}{Y}}{\left(-\frac{X}{Y}+1\right)^2}. \end{aligned} \tag{16}$$

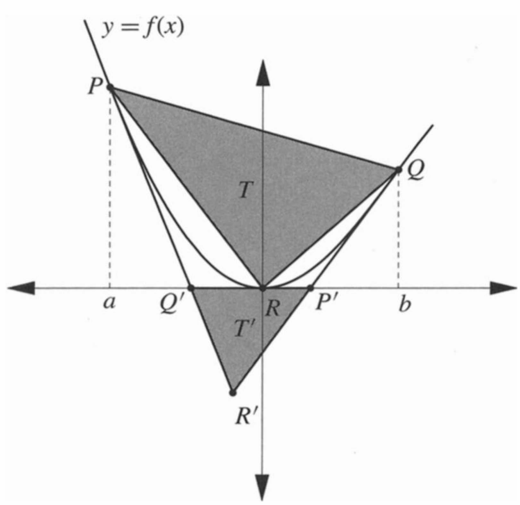


Figure 6 Proof of the Two Triangle Theorem

But

$$\begin{aligned}\frac{X}{Y} &= \frac{f(a)(f'(a)-f'(b))}{f'(a)f'(b)(b-a)}\\ &= \frac{T}{\tilde{T}}\\ &\to \frac{1}{2n}\end{aligned}$$

by formula (12). Substituting this into (16) proves the second part of the theorem,

$$\lim_{a\to 0^-}\frac{T}{T'} = \frac{2n}{(2n-1)^2}.$$

The proof of the first part will require a few preliminaries. We assume $f(x) = c_2x^2 + \cdots$ and $c_2 > 0$. (See FIGURE 6.)

Since we may no longer assume $f(a)$ and $f(b)$ are the same, we must use different formulas for T and T', derived from the standard vector formula for triangle area

$$\text{Area of } \triangle ABC = \frac{1}{2}\|\overrightarrow{AB}\times\overrightarrow{AC}\|.$$

Both formulas are now straightforward calculations,

$$T = \frac{bf(a)-af(b)}{2} \tag{17}$$

and

$$T' = \frac{(f(a)f'(b)-f'(a)f(b)+f'(a)f'(b)(b-a))^2}{2f'(a)f'(b)(f'(a)-f'(b))}.$$

So

$$\frac{T}{T'} = \frac{f'(a)f'(b)(bf(a)-af(b))(f'(a)-f'(b))}{(f(a)f'(b)-f'(a)f(b)+f'(a)f'(b)(b-a))^2}.$$

To complete the proof, we will need four factorizations:

1. $f'(a) = a\varphi_1(a)$ where $\lim_{a\to 0}\varphi_1(a) = 2c_2$.
2. $f'(a) - f'(b) = (a-b)\varphi_2(a,b)$ where $\lim_{a,b\to 0}\varphi_2(a,b) = 2c_2$.
3. $bf(a) - af(b) = ab(a-b)\varphi_3(a,b)$ where $\lim_{a,b\to 0}\varphi_3(a,b) = c_2$.
4. $f'(a)f(b) - f(a)f'(b) = ab(b-a)\varphi_4(a,b)$ where $\lim_{a,b\to 0}\varphi_4(a,b) = 2c_2^2$.

Once we establish these factorizations the proof is complete,

$$\begin{aligned}\lim_{a,b\to 0}\frac{T}{T'} &= \lim_{a,b\to 0}\frac{a\varphi_1(a)b\varphi_1(b)ab(a-b)\varphi_3(a,b)(a-b)\varphi_2(a,b)}{(ab(b-a)\varphi_4(a,b)-a\varphi_1(a)b\varphi_1(b)(b-a))^2}\\ &= \lim_{a,b\to 0}\frac{\varphi_1(a)\varphi_1(b)\varphi_3(a,b)\varphi_2(a,b)}{(\varphi_4(a,b)-\varphi_1(a)\varphi_1(b))^2}\\ &= \frac{8c_2^4}{\left(2c_2^2-4c_2^2\right)^2}\\ &= 2.\end{aligned}$$

We will leave the proofs of the first three factorizations to the reader. They are straightforward infinite series manipulations. The fourth is a little more difficult and is given below. We start with

$$f(x) = \sum_{k=2}^{\infty} c_k x^k.$$

Then,

$$\begin{aligned}
f'(a)f(b) - f(a)f'(b) &= \sum_{k=2}^{\infty}\sum_{m=2}^{\infty} kc_k c_m (a^{k-1}b^m - a^m b^{k-1}) \\
&= ab\sum_{k=2}^{\infty}\sum_{m=2}^{\infty} kc_k c_m (a^{k-2}b^{m-1} - a^{m-1}b^{k-2}) \\
&= ab\Bigg(\sum_{m-1<k-2} kc_k c_m a^{m-1}b^{m-1}(a^{k-m-1} - b^{k-m-1}) \\
&\quad + \sum_{k-2<m-1} kc_k c_m a^{k-2}b^{k-2}(b^{m-k+1} - a^{m-k+1})\Bigg) \\
&= ab(b-a)\Bigg(-\sum_{m-1<k-2} kc_k c_m a^{m-1}b^{m-1}\sum_{j=0}^{k-m-2} b^j a^{k-m-2-j} \\
&\quad + \sum_{k-2<m-1} kc_k c_m a^{k-2}b^{k-2}\sum_{j=0}^{m-k} a^j b^{m-k-j}\Bigg) \\
&= ab(b-a)\varphi_4(a,b),
\end{aligned}$$

where we have defined $\varphi_4(a, b)$ in the previous line. Then,

$$\lim_{a,b\to 0} \varphi_4(a,b) = \varphi_4(0,0)$$

and the only non-zero contribution to $\varphi_4(0, 0)$ is from the second sum when $k = 2$ and $m = 2$, giving

$$\lim_{a,b\to 0} \varphi_4(a,b) = 2c_2^2.$$

This completes the proof of the Two Triangle Theorem. ■

New directions

Archimedes' squaring of the parabola inspired the Two Triangle Theorem, which in turn bears a certain kinship to the so-called *osculating circle* of differential geometry. Let us recall the definition of the osculating circle. For any three points P, Q and R on the curve $\mathcal{C}$, we construct the circumcircle of $\triangle PQR$. The osculating circle, $\mathcal{O}$, to $\mathcal{C}$ at R is the limit circle of these circumcircles as P and Q approach R along $\mathcal{C}$ and on opposite sides of R. (See FIGURE 7.) If the curvature to $\mathcal{C}$ at R, κ, is non-zero then the osculating circle exists and its radius is equal to $1/\kappa$, see [**11**, p. 39].

Now it is perhaps natural to construct the circumcircles of the triangles $\triangle P'Q'R'$. (See FIGURE 8.) As $PQ \to 0$ and P and Q are on $\mathcal{C}$ and on opposite sides of R, it

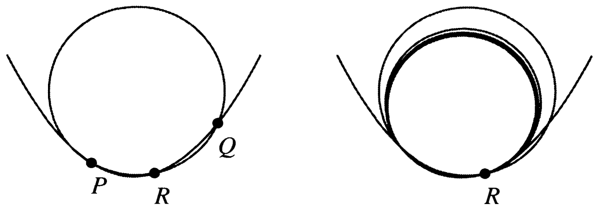


Figure 7 Constructing the Osculating Circle

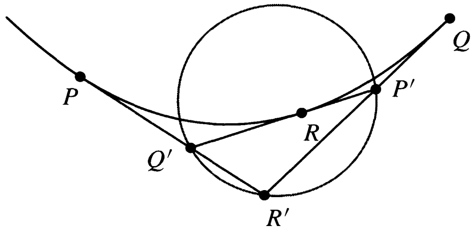


Figure 8 Circumcircle of $\triangle P'Q'R'$

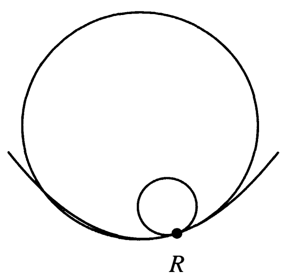


Figure 9 The osculating circle $\mathcal{O}$ (large) and derived osculating circle $\mathcal{O}'$ (small)

appears that we also get a sequence of circles that converges to a circle $\mathcal{O}'$ that we might call the *derived osculating circle*. The two limit circles at R, $\mathcal{O}$ and $\mathcal{O}'$, are pictured in FIGURE 9. The larger circle is the osculating circle $\mathcal{O}$.

If we let r and r' be the respective radii of $\mathcal{O}$ and $\mathcal{O}'$, then computer experiments suggest

$$\lim_{PQ \to 0} \frac{r}{r'} = 4, \tag{18}$$

so long as the point R is of order 2. However, if R is a point of order $2n$ with $n > 1$, then limit 18 may not exist, so we restrict the limit:

$$\lim_{\substack{PQ \to 0 \\ \overline{PQ} || T_R C}} \frac{r}{r'} = \frac{4n^2}{2n-1}, \tag{19}$$

where the limit is taken over pairs of points $P, Q \in \mathcal{C}$ on opposite sides of R and so that $\overline{PQ} || T_R \mathcal{C}$.

There is clearly something of a general nature going on here. We suggest the following setting.

By a *triangle function* we will mean a real valued function $\mathcal{T}$ defined on triangles in the plane so that $\mathcal{T}(\Delta_1) = \mathcal{T}(\Delta_2)$ if Δ_1 is congruent to Δ_2. The area and circumradius are two examples of triangle functions. The general question is this. Using the points P, Q, R, P', Q', R' as we have been doing, R is a point of order 2, and assuming $\mathcal{T}$ is a triangle function, what is the value, if it exists, of

$$L = \lim_{PQ\to 0} \frac{\mathcal{T}(\Delta PQR)}{\mathcal{T}(\Delta P'Q'R')}?$$

And similarly, what about the limit

$$L_{||} = \lim_{\substack{PQ\to 0\\ PQ||T_R C}} \frac{\mathcal{T}(\Delta PQR)}{\mathcal{T}(\Delta P'Q'R')}$$

if the order of R is $2n$ ($n > 1$)?

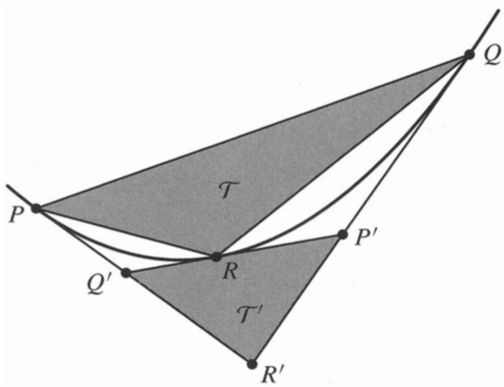


Figure 10 $\mathcal{T} = \mathcal{T}(\Delta PQR)$ and $\mathcal{T}' = \mathcal{T}(\Delta P'Q'R')$

Some more analysis and computer experiments suggest there is no simple general answer to this question. Here are some results to ponder.

1. As we saw earlier, if $\mathcal{T}(\Delta) = \text{area}(\Delta)$, then $L = 2$ and $L_{||} = 2n/(2n-1)^2$.
2. Experimentally, $\mathcal{T}(\Delta) = \text{perimeter}(\Delta)$, then $L = 2$ and $L_{||} = 2n/(2n-1)$.
3. But if $\mathcal{T}(\Delta) = c + \tau(\Delta)$, where c is a fixed non-zero number and τ is either area or perimeter , then $L = L_{||} = 1$.
4. And as we saw in the limits 18 and 19, if $\mathcal{T}(\Delta) = \text{circumradius}(\Delta)$, then $L = 4$ and $L_{||} = 4n^2/(2n-1)$. But if $\mathcal{T}(\Delta) = c + \text{circumradius}(\Delta)$ where c is a constant, then $L = (4\kappa c + 4)/(4\kappa c + 1)$ where κ is the curvature to $\mathcal{C}$ at R. On the other hand, $L_{||} = 4n^2/(2n-1)$ even if $c \neq 0$.
5. If $\mathcal{T}(\Delta) = \text{inradius}(\Delta)$, then computer experiments suggest $L = 1$ and $L_{||} = 1/(2n-1)$.
6. Also experimentally, if $\mathcal{T}(\Delta)$ is the cube root of the product of the three side lengths of Δ, then $L = 2$ and $L_{||} = 2n/(2n-1)$.

Eureka, anyone?

Acknowledgment. I would like to thank the anonymous referees for their insightful and helpful suggestions that greatly improved the exposition of the ideas presented in this article.

REFERENCES

1. A. Benyi, P. Szeptycki & F. Van Vleck, Archimedean Properties and Parabolas, *Amer. Math. Monthly* **107** (2000) 945–949.
2. O. Bottema, Archimedes Revisited, this MAGAZINE **57** (1984) 224–225.
3. J. W. Bruce & P. J. Giblin, *Curves and Singularities*, Cambridge Univ. Press, 1984.
4. D. W. De Temple & J. M. Robertson, Lattice Parabolas, this MAGAZINE **50** (1977) 152–158.
5. H. Dörrie. *100 Great Problems of Elementary Mathematics*, Dover Publications (1965).
6. M. Golomb & H. Haruki, An Inequality for Elliptic and Hyperbolic Segments, this MAGAZINE **46** (1973) 152–155.
7. M. Golomb, Variations on a Theorem of Archimedes, *Amer. Math. Monthly* **81** (1974) 138–145.
8. T. Heath, *The Works of Archimedes*, Dover Publications, 2002.
9. M. Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Press, 1972.
10. J. E. Marsden and M. J. Hoffman, *Elementary Classical Analysis*, 2nd ed., W. H. Freeman, 1974.
11. R. S. Millman and G. D. Parker, *Elements of Differential Geometry*, Prentice-Hall, 1977.
12. A. Rosenthal, The History of Calculus, *Amer. Math. Monthly* **58** (1951) 75–86.
13. S. Stein, *Archimedes: What Did He Do Besides Cry Eureka?* Mathematical Association of America, 1999.
14. G. Swain and T. Dence, Archimedes' Quadrature of the Parabola Revisited, this MAGAZINE **71** (1998) 123–130.

Three Transcendental Numbers from the Last Non-Zero Digits of n^n, F_n, and $n!$

GREGORY P. DRESDEN
Washington & Lee University
Lexington, VA 24450
dresdeng@wlu.edu

In this article, we will construct three infinite decimals from the last nonzero digits of n^n, F_n (the Fibonacci numbers), and $n!$, respectively, and we will show that all three are transcendental. Along the way, we will learn a bit about the history of transcendental numbers, discuss two major theorems in the field, and pose some questions for future research. Let's begin by recalling what it means for a number to be transcendental.

DEFINITION. For α a complex number, we say α is **algebraic** if it is the root of a polynomial with integer coefficients. If no such polynomial exists, we say that α is **transcendental**.

Early mathematicians, of course, were ignorant of this distinction. Indeed, the Pythagoreans of ancient Greece believed that everything in the universe could be measured by whole numbers and their ratios. It must have come as quite a shock when Hippasus (fifth century BC) first demonstrated that certain ratios, such as the ratio between the diagonal and the side of a square, were not the ratios of two whole numbers. Legend has it that angry Pythagoreans threw Hippasus into the sea for his heresy [**9**], but the idea of irrational numbers lived on. Indeed, our word "irrational" dates back to the Greek word $\alpha\rho\rho\eta\tau o\varsigma$ (*arrētos*), meaning "unspeakable", perhaps reflecting the Greeks' disgust of such messy objects as $\sqrt{2}/1$ [**9**], [**16**].

We now skip ahead to seventeenth-century Europe. With the introduction of algebra and modern notation, it finally became possible to ask if there existed numbers that were not roots of polynomials. James Gregory [**3**],[**8**] appears to have been the first to attempt to prove that π and e were not algebraic, an ambitious (and ultimately fruitless) goal given that it was not yet known if they were even irrational (their irrationality was finally shown in the eighteenth century by Lambert and Euler, respectively).

It was not until the nineteenth century that we see the first proofs of the existence of transcendental numbers. Liouville was the first to do so, with the specially-constructed number

$$L = \sum_{n=1}^{\infty} \frac{1}{10^{n!}} = 0.110001000000000000000001\ldots$$

The numbers π and e were shown to be transcendental by the later part of the century by Lindemann and Hermite, respectively. Lindemann's proof finally put to rest the old problem of squaring the circle, first studied by the Greeks over two millenia earlier. Lindemann later reported [**12**, p. 246] that Kronecker said to him (probably in jest, and perhaps alluding to the ancient Greeks' distaste for irrational numbers), "Of what use is your beautiful investigation of π? Why study such problems since irrational numbers do not exist?" Some have seen this as a demonstration that Kronecker believed only in the existence of integers (recall also his famous quote "God created the integers, all else is the work of man"), but it is clear from his work that this is not the case. For a summary of the controversy over Kronecker's words, see the article by Edwards [**5**, Essay 5.5].

We note that even in the twenty-first century there are many open questions about transcendental numbers. It's not known if $\zeta(3) = 1 + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{4^3} + \cdots$ is transcendental, although Apéry did prove that it is irrational [**14**]. As for the Euler constant $\gamma = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i} - \ln n$, we do not yet know if it is even irrational, let alone transcendental. For a further discussion of recent work on transcendental numbers, Ribenboim in [**13**] provides an English summary of a 1983 historical overview by Waldschmidt [**17**]. See also Ribenboim's extensive bibliography in [**13**, Chapter 10].

Two Modern Theorems

It is possible to prove that certain numbers (such as π and e) are transcendental by assuming them to be algebraic and then working towards a contradiction. Such proofs are fairly complicated and involve a lot of auxiliary polynomials [**7**], [**2**]. Fortunately, there exist several different characterizations of transcendental numbers that will prove to be much easier to work with. This first theorem is the result of the work of three mathematicians over the first half of the twentieth century; it is restated slightly for convenience.

THEOREM 1. (THUE, SIEGEL, ROTH) *For α algebraic and $\epsilon > 0$, there exist only finitely many rational numbers p/q such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\epsilon}}.$$

Thus, if we can show that for some fixed $\epsilon > 0$ there are infinitely many such numbers p/q that satisfy the above inequality, then α must be transcendental. In contrast to Thue-Siegel-Roth, consider the following fairly trivial result from Diophantine approximation [**7**], [**11**]: if we have α *rational* and $\epsilon > 0$ then there exist only finitely many rational numbers p/q such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{1+\epsilon}}.$$

Note also that both these results are best possible in terms of the exponents. That is, if we replace the epsilons with zeros, then there would actually be infinitely many p/q's (found by continued fractions) satisfying the first inequality and infinitely many p/q's (for any q, take $p = \lfloor \alpha q \rfloor$) for the second inequality.

Our second theorem comes from an article [**1**] just published in 2004, and it is an extremely useful result. The following presentation is translated from the original French and is simplified for convenience.

THEOREM 2. (ADAMCZEWSKI, BUGEAUD, LUCA) *Let α be irrational, and suppose there exist two sequences $\{U_n\}$, $\{V_n\}$ of finite words on $\{0, 1, \ldots 9\}$ and a real number $x > 1$ such that*

(1) *For every $n \geq 1$, the word $U_n V_n^x$ is a prefix for α,*

(2) *The set $\left(\frac{|U_n|}{|V_n|}\right)_{n\geq 1}$ is bounded,*

(3) *The set $|V_n|$ is strictly increasing.*

Then, α is transcendental.

A few comments and definitions are in order. Recall that a **finite word** W is simply a collection of digits, such as 411 or 314159. In saying that a word W is a **prefix** for a number α we mean that α begins with the digits from W. By $|W|$ we mean the number of digits in W. For x an integer, then W^x is simply the word $WW\cdots W$ (repeated x times), while if x is not integral we define W^x as being $W^{\lfloor x\rfloor}$ followed by the first $\lceil (x-\lfloor x\rfloor)\cdot |W|\rceil$ elements of W. In particular, if W is 411, then $W^{1.5}$ would be 41141. Finally, the original statement and proof of this theorem in [**1**] is slightly more general in that it holds for α expressed in any base.

Both of these theorems give us new ways of identifying transcendental numbers. The Thue-Siegel-Roth theorem requires us to find extremely close rational approximations, while the Adamczewski-Bugeaud-Luca theorem relies on a pattern in the digits. Let us now move on to our three transcendental numbers and see how we can apply these two theorems.

Forming numbers from the digits of n^n

We begin by looking at the pattern formed from the last (i.e. unit) digit of n^n. Since

$$1^1 = \boxed{1}, \qquad 2^2 = \boxed{4}, \qquad 3^3 = 2\boxed{7}, \qquad 4^4 = 25\boxed{6}, \qquad 5^5 = 312\boxed{5}, \qquad \cdots$$

then if we take the last digit of each number and form a decimal, we get

```
0. 1476563690 1636567490
   1476563690 1636567490 ···
```

This looks a lot like a repeating decimal, and indeed it is not hard to prove (see [**6**]) that

$$n^n \equiv (n+20k)^{n+20k} \bmod 10$$

which allows us to conclude that we have a rational number with period 20 equal to (1476563690 1636567490) / (9999999999 9999999999). Since rational numbers are by definition not transcendental, we need to modify our construction to produce a more interesting decimal.

With this in mind, let us now construct a *new* decimal number $A = 0.d_1d_2d_3\ldots d_n\ldots$ such that the nth digit d_n of A is the last *nonzero* digit of n^n; that is,

```
A = 0. 1476563691 1636567496
       1476563699 1636567496 ···
```

The reader will note that this new number A differs from the previous number at the tenth decimal place, the twentieth, the thirtieth, and so on. In a recent paper [**4**], we showed that this A is an irrational number (despite "almost" having a period of twenty digits). We will now show that it is transcendental.

To prove transcendence, we will be using the Thue-Siegel-Roth Theorem. In particular, we will demonstrate the existence of an infinite sequence of rational numbers p_n/q_n such that

$$\left|A - \frac{p_n}{q_n}\right| < \frac{1}{{q_n}^{2.1}}.$$

Let's begin by creating a sequence of irrational numbers $A_0 = A, A_1, A_2, \ldots$ such that each A_n is formed by replacing every digit in A_{n-1} with zeros except for every tenth nonzero digit (which will be left alone). This means that each A_n has nonzero digits only every 10^nth place, at 10^{-10^n}, $10^{-2\cdot 10^n}$, $10^{-3\cdot 10^n}$, and so on. Visually, this sequence looks like the following, where the dots represent zeros.

```
A0 = 0.1476563691 1636567496 1476563699 1636567496 ···
A1 = 0.........1 .........6 .........9 .........6 ···
A2 = 0.......... .......... .......... .......... ···
```

(The number A_2 doesn't have a nonzero digit until the one hundredth decimal place, at 10^{-100}.) If we remove the dots and condense these decimals a bit, we see an interesting pattern develop (recall that the nonzero digits in A_n are actually 10^n decimal places apart; think of the spaces in A_1, A_2, etc. as representing lots and lots of zeros)

```
A0 = 0.1476563691 ···
A1 = 0. 1 6 9 6 5 6 9 6 1 1 ···
A2 = 0.  1  6  1  6  5  6  1  6  1  1 ···
A3 = 0.   1   6   1   6   5   6   1   6   1   1 ···
```

A simple application of [4, Lemma 3] shows that for $n \geq 2$, the sequences of nonzero digits in each A_n are identical: 1, 6, 1, 6, 5, 6, 1, 6, 1, $*$, where $*$ is either 1, 6, or 5 depending on the position. This implies that $R_n = A_n - A_{n+1}$ is rational for $n \geq 2$; the cases $n = 0$ and $n = 1$ follow immediately from Lemma 2 below. Since each R_n is rational, then if we can show that A_n is well approximated by rationals (in the context of the Thue-Siegel-Roth theorem), this might help us to approximate A as well. Let's investigate these A_n's a bit further.

If we write out just the nonzero digits of A_n (with appropriate spacing)

$$A_n = 0.\ 1\ 6\ 1\ 6\ 5\quad 6\ 1\ 6\ 1\ 1\ \cdots \quad (\text{for } n \geq 2)$$

we clearly see that each A_n (for $n \geq 2$) is quite close to the rational number

$$\frac{s_n}{t_n} = 0.\ 1\ 6\ 1\ 6\ 1\quad 6\ 1\ 6\ 1\ 6\ \cdots$$

Here, just as with A_n, these nonzero digits of s_n/t_n are actually 10^n decimal places apart. Thus, s_n/t_n is easily seen to be $\frac{1\cdot 10^{10^n}+6}{10^{2\cdot 10^n}-1}$. Also, it is easy to see that A_n and s_n/t_n differ in the fifth visible position (among other places), which means that they differ by about $\frac{4}{10^{5\cdot 10^n}}$. As a result, we have

$$\left| A_n - \frac{s_n}{t_n} \right| \approx \frac{4}{10^{5\cdot 10^n}} < \frac{1}{10^{4.2\cdot 10^n}} \approx \frac{1}{{t_n}^{2.1}}$$

(The attentive reader will notice that we could have easily replaced the 2.1 with 2.4 or even 2.49. However, 2.1 will work fine for our purposes.)

Let's now relate this back to A. Recall that we have $A_{n+1} = A_n - R_n$, and it is easy to show that the rational number R_n has denominator $10^{10^{n+1}} - 1$ for $n > 0$ and denominator $10^{20} - 1$ at $n = 0$. We define p_n/q_n as

$$\frac{p_n}{q_n} = \left(\sum_{i=0}^{n-1} R_i \right) + \frac{s_n}{t_n}.$$

The denominator q_n is $10^{2\cdot 10^n}$, the same as t_n, because the denominator of each R_i divides evenly into t_n. Thus,

$$\left|A-\frac{p_n}{q_n}\right|=\left|A-\sum_{i=0}^{n-1}R_i-\frac{s_n}{t_n}\right|=\left|A_n-\frac{s_n}{t_n}\right|\approx\frac{4}{10^{5\cdot 10^n}}<\frac{1}{10^{4.2\cdot 10^n}}\approx\frac{1}{{t_n}^{2.1}}=\frac{1}{{q_n}^{2.1}}$$

The conditions of Theorem 1 being satisfied, we can conclude that A is transcendental.

A number from the Fibonacci sequence

Let us look at a new decimal number, this one constructed from the Fibonacci sequence F_n. It's easy to show that $F_{n+60} \equiv F_n \bmod 10$, which means that, as in the case of n^n, taking the last digit of each number would result in a repeating decimal. Instead, we construct a decimal (call it B) from the last *nonzero* digits. Since the Fibonacci numbers are

$$\boxed{1},\quad \boxed{1},\quad \boxed{2},\quad \boxed{3},\quad \boxed{5},\quad \boxed{8},\quad 1\boxed{3},\quad 2\boxed{1},\quad 3\boxed{4},\quad 5\boxed{5},\quad 8\boxed{9},\quad 14\boxed{4},\quad 23\boxed{3},\quad \cdots$$

then we get

$B =$ 0.112358314594371 774156178538194
998752796516737 336954932572912
112358314594375 774156178538192 ...

Our technique in proving that B is transcendental is the same as our proof earlier with A (formed from the last nonzero digit of n^n). We will construct a sequence of decimals $B_0 = B$, B_1, B_2, each formed by selecting a few digits from the previous number. Each B_n will be well approximated by a rational number s_n/t_n, which will be used to create another rational number p_n/q_n. Using Theorem 1 and these rational numbers p_n/q_n, we will be able to conclude that B is transcendental.

Let's begin with B_1, which will be formed from B_0 by keeping every fifteenth digit in B_0, and replacing all the others with zeros.

$B_1 =$ 0..............1 4
..............7 2
..............5 ....2 ...

We condense and continue B_1 to get

$B_1 =$ 0. 1 4 7 2 5 2 3 4 9 2 9 6 3 8 5 8 7 6 1 6
1 4 7 2 5 2 3 4 9 6 9 6 3 8 5 8 7 6 1 2 ...

(digits are 15 decimal places apart, separated by a stream of 0's).

Now define B_2 to be every tenth nonzero digit of B_1, with zeros elsewhere.

$B_2 =$ 0. 2 6 6 2 1 8 4 4 8 8
2 6 6 2 3 8 4 4 8 6 ...

(digits are 150 decimal places apart, separated by zeros).

Define B_3 to be every fifth nonzero digit of B_2.

$B_3 =$ 0. 1 8 3 6 5 4 7 2 9 9
1 8 3 6 5 4 7 2 9 8 ...

(digits are 750 decimal places apart).

Let B_4 be every tenth nonzero digit of B_3, and we will finally see a beautiful pattern appear.

$$B_4 = 0.\ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1\ 9\ \ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1\ 8 \cdots$$

(digits are now 7500 decimal places apart).

If we define B_n (for $n \geq 4$) to be every tenth nonzero digit of B_{n-1}, we will always get (Lemma 4) the pattern

$$B_n = 0.\ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1\ 9\ \ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1\ 8 \cdots \quad (\text{for } n \geq 4)$$

(digits are $75 \cdot 10^{n-2}$ decimal places apart). Furthermore, these B_n's have all been chosen such that they differ from their predecessors by a rational number; that is, if we define $R_n = B_{n+1} - B_n$, then R_n is rational.

Consider the rational number

$$\frac{s_n}{t_n} = \frac{9 \cdot 10^{75 \cdot 10^{n-2}} - 10}{(10^{75 \cdot 10^{n-2}} - 1)^2}.$$

Writing out the first few nonzero digits, we find that this rational number is

$$\frac{s_n}{t_n} = 0.\ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2 \cdots.$$

After that 2, the decimal expansion becomes rather complicated, with most of the zeros replaced by nines, and then some eights start to show up, but that doesn't really concern us. What's important is that s_n/t_n is a good approximation to B_n for quite a few decimal places; in fact, they are the same up to the eighth nonzero digit. Thus, for $n \geq 4$,

$$\left| B_n - \frac{s_n}{t_n} \right| < \frac{1}{10^{8 \cdot (75 \cdot 10^{n-2})}} \approx \frac{1}{t_n{}^4}$$

Let's now relate this back to B_4. If we can show that B_4 is transcendental, then since B_4 and B differ by a rational number, then B itself will also be transcendental.

Recall that we have $B_{n+1} = B_n - R_n$, and it is easy to show that R_i has denominator $10^{75 \cdot 10^{i-1}} - 1$ (for $i \geq 4$). We now define p_n/q_n as

$$\frac{p_n}{q_n} = \left(\sum_{i=4}^{n-1} R_i \right) + \frac{s_n}{t_n}.$$

The denominator q_n is $(10^{75 \cdot 10^{n-2}} - 1)^2$, the same as t_n, because the denominator of each R_i divides evenly into t_n. Thus,

$$\left| B_4 - \frac{p_n}{q_n} \right| = \left| B_4 - \sum_{i=4}^{n-1} R_i - \frac{s_n}{t_n} \right| = \left| B_n - \frac{s_n}{t_n} \right| < \frac{1}{10^{8 \cdot (75 \cdot 10^{n-2})}} \approx \frac{1}{t_n{}^4}$$

Thus, by Theorem 1, B_4 is transcendental, and hence so is B.

A number from the sequence $n!$

Finally, let us look at what we get from the last nonzero digit of $n!$. Starting with

$\boxed{1}$, $\boxed{2}$, $\boxed{6}$, $2\boxed{4}$, $1\boxed{2}0$, $7\boxed{2}0$, $50\boxed{4}0$, $403\boxed{2}0$, $3628\boxed{8}0$, $3628\boxed{8}00$, $\cdots$

we get

C = 0.1264 22428 88682 88682 44846 44846 88682 $\cdots$

It looks more appealing if we replace the first block of 1264 with 66264, and continue out a few thousand digits. We then get the very nice pattern

C' = 0.66264 22428 88682 88682 44846 $\cdots$ (625 digits per line)
66264 22428 88682 88682 44846 $\cdots$
22428 44846 66264 66264 88682 $\cdots$
66264 22428 88682 88682 44846 $\cdots$
44846 88682 22428 22428 66264 $\cdots$

22428 44846 66264 66264 88682 $\cdots$
$\vdots$ $\vdots$ $\vdots$ $\vdots$ $\vdots$

Note that C and C' are related by the linear formula $C = 10C' - 6.5$, so if C' is identified as rational, algebraic, or transcendental, then the same must hold for C itself. The number C' turns out to be much easier to work with (notice the nice repetition in the first two lines, for example), and with this in mind, let us define ℓ to represent the last nonzero digit with a slight variation.

DEFINITION. For $n \geq 0$, define $\ell(n) = \begin{cases} 6 & : \quad n = 0, 1 \\ \text{lnzd}(n!) & : \quad n \geq 2. \end{cases}$

In a previous paper [**4**], we proved that C (and thus C') is irrational. Thanks to the new Adamczewski-Bugeaud-Luca theorem, we can now show that it is transcendental. An important step in this direction is the following formula for the decimal digits in C'. Writing n in base-5 notation as $n = \sum_{i=0}^{N} a_i 5^i$ (for $a_i \in \{0, 1, 2, 3, 4\}$), then $\ell(n)$ is given by the reduction mod 10 of $6 \prod_{i=0}^{N} (a_i!) 2^{i \cdot a_i}$ (see Lemma 7). We can use this to explain why the first two lines in the expression for C' are identical, and in fact we can go a bit further. The following lemma gives the full story.

LEMMA 1. *For r a multiple of four, the first 5^r digits of C' equal the second 5^r digits.*

Proof. Let n be a positive integer less than 5^r. We can write $n = \sum_{i=0}^{r-1} a_i 5^i$, and thus $n + 5^r = \sum_{i=0}^{r-1} \left(a_i 5^i\right) + 1 \cdot 5^r$. So, by our formula for $\ell(n)$, we have $\ell(n + 5^r) \equiv \ell(n) \cdot (1!) 2^{r \cdot 1}$ mod 10. Since r is a multiple of 4, then $2^r \equiv 6$ mod 10, and recalling that 6 acts as an identity on the set $\{2, 4, 6, 8\}$ under multiplication mod 10, we have $\ell(n + 5^r) \equiv \ell(n)$ mod 10, but of course these are both single-digit integers, so we conclude $\ell(n + 5^r) = \ell(n)$. Since this holds for all $n < 5^r$, the lemma follows. ■

We now use this lemma to immediately prove the transcendence of the irrational number C'. In the notation of the Adamczewski-Bugeaud-Luca theorem, let $U_n = \{\}$, let V_n be the first 5^{4n} digits of C', and let $x = 2$. All three criteria of Theorem 2 having been satisfied, we conclude that C', and hence C, is transcendental.

Proofs

The following technical lemmas are used in the paper.

LEMMA 2. *For n not divisible by* 100, *then*

$$lnzd(n^n) = lnzd((n+100)^{n+100}).$$

Proof. We first note that if $100 \nmid n$ then $\text{lnzd}(n) = \text{lnzd}(n+100)$. Also, $\text{lnzd}(a^b) \equiv \text{lnzd}(a)^b \bmod 10$ for all $a, b > 0$. This allows us to state that $\text{lnzd}((n+100)^{n+100}) \equiv \text{lnzd}(n+100)^{n+100} \equiv \text{lnzd}(n)^{n+100} \bmod 10$, so the lemma reduces to proving that $\text{lnzd}(n)^{n+100} \equiv \text{lnzd}(n)^n \bmod 10$. This is clearly true for $\text{lnzd}(n) = 5$. If $\text{lnzd}(n) = 1, 3, 7$, or 9, then since each of these raised to the fourth power gives 1 mod 10, we have $\text{lnzd}(n)^{n+100} \equiv \text{lnzd}(n)^n \cdot 1^{25} = \text{lnzd}(n)^n$. If on the other hand $\text{lnzd}(n) = 2, 4, 6$, or 8, then since each of these gives 6 when raised to the fourth power mod 10, we have $\text{lnzd}(n)^{n+100} \equiv \text{lnzd}(n)^n \cdot 6^{25} \equiv \text{lnzd}(n)^n \cdot 6$. Now, $\text{lnzd}(n)$ is even in this case, and 6 acts as a multiplicative identity on even numbers mod 10, so we end up with $\text{lnzd}(n)^n$, as desired. ■

LEMMA 3. *For* $n = 7500 \cdot 10^k$, $k \geq 0$, *then* $lnzd(F_n) = 9$.

Proof. This is certainly true for $k = 0$, by inspection. For $k > 0$ we use induction and the identity [**19**, Formula 42]

$$F_{10n} = F_n \left[\binom{9}{0} L_n{}^9 - \binom{8}{1} L_n{}^7 + \binom{7}{2} L_n{}^5 - \binom{6}{3} L_n{}^3 + \binom{5}{4} L_n \right].$$

Assume that n is as given and that $\text{lnzd}(F_n) = 9$. By observation, the Lucas numbers mod 100 have period 60; since $60 | 7500$, this implies that $L_n \equiv 2 \bmod 100$. An application of [**15**, Theorems 6 and 7] and [**18**] gives that F_n ends in exactly $k+4$ zeros, and F_{10n} in $k+5$ zeros. With all this in mind, we write the above formula as

$$\begin{aligned}\left(\frac{F_{10n}}{10^{k+4}} \bmod 100\right) &\equiv \left(\frac{F_n}{10^{k+4}} \bmod 100\right) \\ &\quad \cdot \left[2^9 - 8 \cdot 2^7 + 21 \cdot 2^5 - 20 \cdot 2^3 + 5 \cdot 2\right] \bmod 100 \\ &\equiv \left(\frac{F_n}{10^{k+4}} \bmod 100\right) [10] \bmod 100 = 90.\end{aligned}$$

This implies that $\text{lnzd}(F_{10n}) = 9$, as desired. ■

LEMMA 4. *For* $n = 7500 \cdot 10^k$, $k \geq 0$, *and* $1 \leq i \leq 9$, *then* $lnzd(F_{in}) = 10 - i$.

Proof. The case $i = 1$ is covered by Lemma 3. For $i = 2$ we apply the formula $F_{2n} = F_n L_n$ and the fact (mentioned in Lemma 3) that $\text{lnzd}(L_n) = 2$. The other cases proceed by induction and the formula $F_{in} = F_{(i-1)n} L_n - F_{(i-2)n}$ from [**10**, p. 92]. ■

LEMMA 5. *For k not a multiple of* 5 *and for* $b \geq 1$, *then* $\ell(5^b k) \equiv 8^b k \cdot \ell(5^b k - 1)$ *mod* 10.

Proof. $\ell(5^b k) = \text{lnzd}((5^b k)!) = \text{lnzd}((5^b k - 1)! \cdot 5^b k)$. It's easy to show that $2^{b+1} | (5^b k - 1)!$, so we can write this last expression as $\text{lnzd}((5^b k - 1)! 2^{-b} \cdot 10^b k) = \text{lnzd}((5^b k - 1)! 2^{-b} \cdot 6^b \cdot 10^b k)$ because 6 acts as the identity on $\{2, 4, 6, 8\}$ under multiplication mod 10. Replacing 6^b with $2^b 8^b$, cancelling the 2's and moving the 8's outside, we get $8^b \cdot \text{lnzd}((5^b k - 1)! k) \bmod 10$, which gives us $8^b k \cdot \ell(5^b k - 1) \bmod 10$. ■

LEMMA 6. *For $k \geq 1$ not a multiple of 5 and for $c \geq 1$, then*

(1) $\ell(5^b k) \equiv 2^b k \cdot \ell(5^b(k-1)) \bmod 10$.

(2) $\ell(5^b(5c-1)) \equiv 4 \cdot \ell(5^{b+1}(c-1)) \bmod 10$.

Proof. We induct on b. For $b = 0$, the first statement is trivial and the second follows from applying Lemma 5 four times. We now assume that (1) and (2) hold for $b < a$, and attempt to prove them for $b = a$. For the first,

$$\begin{aligned}
\ell(5^a k) &\equiv 8^a k \cdot \ell(5^a k - 1) \mod 10 && \text{(by Lemma 5)}\\
&\equiv 8^a k \cdot 4 \cdot \ell(5 \cdot (5^{a-1}k - 1)) && \text{(by equation (2) with } b = 0, c = 5^{a-1}k)\\
&\equiv 8^a k \cdot 4^2 \cdot \ell(5^2 \cdot (5^{a-2}k - 1))\\
&\vdots\\
&\equiv 8^a k \cdot 4^a \cdot \ell(5^a \cdot (k-1))\\
&\equiv 2^a k \cdot \ell(5^a \cdot (k-1))
\end{aligned}$$

and for the second, $\ell(5^a(5c-1) \equiv 2^a(5c-1) \cdot 2^a(5c-2) \cdot 2^a(5c-3) \cdot 2^a(5c-4) \cdot \ell(5^a(5c-5))$ by applying (1) four times, and after simplifying mod 10 we get $4\ell(5^a(5c-5))$, which is $4 \cdot \ell(5^{b+1}(c-1))$. ■

LEMMA 7. *For $n = \sum_{i=0}^N a_i 5^i$, then $\ell(n) \equiv 6 \prod_{i=0}^N (a_i!) 2^{i \cdot a_i} \bmod 10$.*

Proof. Suppose 5^{i_0} is the largest power of 5 dividing n. We can thus write $n = 5^{i_0}(m + a_{i_0})$ for m some integer multiple of 5. Applying part (1) of Lemma 6 a_{i_0} times, we get $\ell(n) \equiv 2^{i_0 a_{i_0}}(a_{i_0}!)\ell(5^{i_0}m) \bmod 10$. We now repeat the process with n replaced by $5^{i_0}m = n - a_{i_0}5^{i_0}$ (and so on) to eventually arrive at the desired formula. ■

For Further Study

For those interested in the history of mathematics, Morris Kline's book **[9]** is an excellent resource. One also shouldn't ignore Kurt von Fritz's article **[16]**, which gives a delightful discussion of how in the fifth century BC Hippasus might have discovered irrational numbers by using a pentagon, thus suggesting that the first irrational number could have been the golden ratio, $(1 + \sqrt{5})/2$. Of course, the golden ratio is also the limit of the ratio of successive Fibonacci numbers, which brings us back to one of the subjects of this article.

For those interested in creating more transcendental numbers like the three given in this article, a good place to start might be with other famous sequences such as the squares or the triangular numbers. Euler and Sadek **[6]** suggest looking at the last digit of the primes p, or perhaps of p^p. If one wishes more sequences to study, there are well over a hundred thousand of them at N. J. A. Sloane's On-Line Encyclopedia of Integer Sequences (`http://www.research.att.com/~njas/sequences/`).

Finally, J. Siehler suggested looking at algebraic numbers. Is there a way to recognize or to perhaps write down a decimal expansion for an algebraic number, using a theorem similar to the two theorems of this paper? What would such a number look like? It might be easy to prove such a number is algebraic yet hard to actually find its minimal polynomial!

Acknowledgments. The author would like to thank J. Siehler for his assistance on Theorem 2, and classics professors M. Carlisle and K. Crotty for their comments on Greek etymology.

REFERENCES

1. Boris Adamczewski, Yann Bugeaud, and Florian Luca, Sur la complexité des nombres algébriques, *C. R. Math. Acad. Sci. Paris* **339**(1) (2004) 11–14.
2. Edward B. Burger and Robert Tubbs, *Making Transcendence Transparent*, Springer-Verlag, New York, 2004.
3. Max Dehn and E. D. Hellinger, Certain mathematical achievements of James Gregory, *Amer. Math. Monthly* **50** (1943) 149–163.
4. Gregory Dresden, Two irrational numbers from the last non-zero digits of $n!$ and n^n, this MAGAZINE **74** (2001) 316–320.
5. Harold M. Edwards, *Essays in Constructive Mathematics*, Springer-Verlag, New York, 2005.
6. R. Euler and J. Sadek, A number that gives the unit digit of n^n, *J. Rec. Math.* **29**(3) (1998) 203–204.
7. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., The Clarendon Press/Oxford University Press, New York, 1979.
8. E. W. Hobson, H. P. Hudson, A. N. Singh, and A. B. Kempe, *Squaring the Circle and Other Monographs*, Chelsea Publishing Company, New York, 1953.
9. Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, New York, 1972.
10. Thomas Koshy, *Fibonacci and Lucas Numbers with Applications*, Wiley-Interscience, New York, 2001.
11. John McCleary, How not to prove Fermat's last theorem, *Amer. Math. Monthly* **96** (1989) 410–420.
12. Henri Poincaré, *Wissenschaft und Hypothese*, F. Lindemann, ed., B. G. Teubner, Leipzig, 1904.
13. Paulo Ribenboim, *My Numbers, My Friends*, Springer-Verlag, New York, 2000.
14. Alfred van der Poorten, A proof that Euler missed. . .Apéry's proof of the irrationality of $\zeta(3)$, *Math. Intelligencer* **1**(4) (1978/79) 195–203. An informal report.
15. Andrew Vince, The Fibonacci sequence modulo N, *Fibonacci Quart.* **16**(5) (1978) 403–407.
16. Kurt von Fritz, The discovery of incommensurability by Hippasus of Metapontum, *Ann. of Math. (2)* **46** (1945) 242–264.
17. Michel Waldschmidt, Les débuts de la théorie des nombres transcendants (à l'occasion du centenaire de la transcendance de π), In *Proceedings of the Seminar on the History of Mathematics, 4*, pages 93–115, Inst. Henri Poincaré, Paris, 1983.

→ D. D. Wall, Fibonacci series modulo m, *Amer. Math. Monthly* **67** (1960) 525–532.

19. Eric W. Weisstein, Fibonacci number, from MathWorld—A Wolfram Web Resource, `http://mathworld.wolfram.com/FibonacciNumber.html`.

To appear in *The College Mathematics Journal* May 2008

Articles

A Break for Mathematics, *by Deanna Haunsperger*

An Alternate Approach to Alternating Sums: A Method to DIE For, *by Arthur T. Benjamin and Jennifer J. Quinn*

Dinner Tables and Concentric Circles: A Harmony of Mathematics, Music, and Physics, *by Jack Douthett and Richard J. Krantz*

Squaring a Circular Segment, *by Russell A. Gordon*

Dependent Probability Spaces, *by William F. Edwards, Ray C. Shiflett, and Harris S. Shultz*

Classroom Capsules

From Mixed Angles to Infinitesimals, *by Jacques Bair and Valérie Henry*

The Perimeter of a Polyomino and the Surface Area of a Polycube, *by Wiley Williams and Charles Thompson*

The Cross Product as a Polar Decomposition, *by Götz Trenkler*

An Exploration of Pick's Theorem in Space

HOWARD ISERI
Mansfield University
Mansfield, PA 16933
hiseri@mansfield.edu

Georg Pick described the following remarkable theorem involving the integer **lattice points** of the xy-plane, the points in the plane with integer coordinates, and **lattice polygons**, simple polygons with all of their vertices at lattice points.

PICK'S THEOREM. *Given a lattice polygon P, the area of P is given by*

$$A = I + \frac{B}{2} - 1,$$

where A is the area enclosed by the polygon, I is the number of lattice points in the interior of the polygon, and B is the number lattice points on the boundary of the polygon.

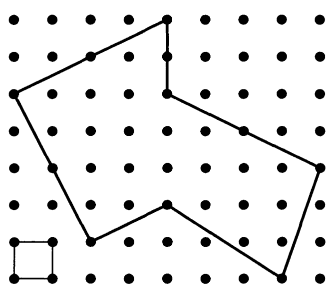


Figure 1 A simple lattice polygon

For example, in FIGURE 1, the polygon shown has $I = 22$ interior lattice points and $B = 11$ boundary lattice points. By Pick's theorem, the area is

$$A = 22 + \frac{11}{2} - 1 = \frac{53}{2}.$$

We will not prove Pick's theorem here, since many proofs are easily found in the literature (see **[3, 4, 7, 8, 10, 12, 13]**, for example). All of these proofs are accessible and interesting, and their variety is also, but for us, Pick's theorem will be a basic assumption.

The conclusion of Pick's theorem, while unexpected, is easy to make sense of, and deeper understandings seem to come in manageable and satisfying chunks. Best of all, like its close relative, Euler's theorem, there always seems to be something more to know. *The Mathematics Monthly* and this MAGAZINE alone contain an extensive series of articles about Pick's theorem **[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]**. In addition to the proofs already mentioned, these contain a variety of explorations, and they discuss, for example, extensions of Pick's theorem to higher dimensions and more general shapes. Here, we will explore how Pick's theorem applies to polygons in a three-dimensional lattice. More accurately, we will consider flat polygonal disks sitting in space, since the area of a non-planar polygon described by its vertices would be difficult to work with, even if we could make sense of it. So keep in mind that when we talk about the area of a polygon in space, the vertices of the polygon must lie in a plane, and the area of the polygon is the area of the region enclosed by the polygon in that plane. We will be expanding on ideas contained in a computer science article [**1**] regarding *inverse Pick's problems* like the following.

> A (flat) quadrilateral (disk) with vertices $(0, 0, 0)$, $(-7, 8, 3)$, $(-10, 10, 6)$, and $(-8, 7, 6)$ lies on the plane $9x + 6y + 5z = 0$. How many integer lattice points are contained in this quadrilateral?

The answer to this question might be of interest to us, if we wanted to display a three-dimensional image of this quadrilateral on a computer screen, or perhaps even in a hologram consisting of a three-dimensional array of pixels, since the number of pixels that naturally lie on the quadrilateral would have some influence on its appearance. For a polygon drawn on a piece of graph paper, it is easy to count the lattice points, and Pick's theorem provides a satisfying way to compute the polygon's area from this information. For a lattice polygon sitting in space, however, finding the number of three-dimensional lattice points that lie on the disk is actually more difficult than finding its area. In this context, therefore, the more satisfying problem tends to be the problem of finding the number of lattice points from the area, that is, in solving the inverse Pick's problem.

The computer science article [**1**] in question, which presents algorithms that solve inverse Pick's problems, is not easy to understand, but the underlying mathematical results are well-worth exploring, and as we work through these ideas using elementary methods, we will see connections between some of the fundamental structures of linear algebra, number theory, and geometry.

1. Variations on Pick's theorem

A few easy variations on Pick's theorem, and some insight into our main explorations come from considering other kinds of lattices.

1.1. Skewed lattices. Pick's theorem is preserved, in some sense, under any linear transformation from the plane to itself, if we use the images of the lattice points to form a new *skewed* lattice. For example, consider the linear transformation that maps $(1, 0) \mapsto (2, 0)$ and $(0, 1) \mapsto (1, 1)$. Specifically,

$$x' = 2x + y$$
$$y' = y,$$

or in matrix form

$$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}.$$

The lattice points and polygon in FIGURE 1 map to the lattice points and polygon in FIGURE 2. Not surprisingly, the number of interior and boundary lattice points stays the same, but the area will increase by a factor of two. The reason for this can be seen by looking at what happens to the unit square shown at the lower left in FIGURE 1. It maps to the parallelogram shown at the lower left of FIGURE 2 with vertices

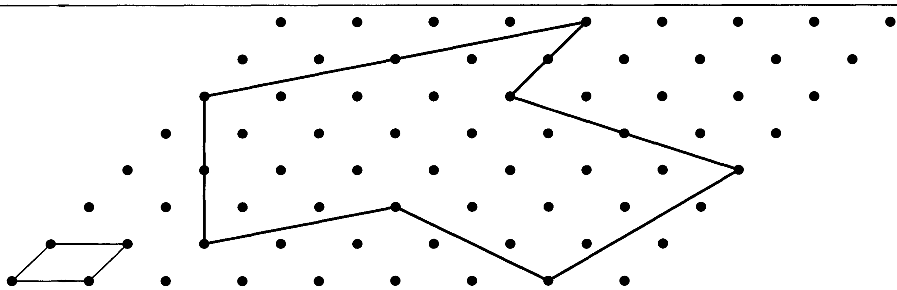


Figure 2 A transformation of the lattice shown in FIGURE 1

(0, 0), (2, 0), (3, 1), and (1, 1). Its area is 2. A basic fact of linear algebra is that the determinant of the matrix of transformation,

$$\begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = (2)(1) - (1)(0) = 2,$$

is the area of the image of a unit square, and in general, a conversion factor for areas. We will call a parallelogram like the small one in FIGURE 2, a *unit parallelogram.*

In Pick's theorem, the area can be interpreted as measuring how many unit squares fit in a given polygon, and for a skewed lattice, like the one in FIGURE 2, Pick's theorem tells us how many unit parallelograms fit in the image of the original polygon. In order to find the new area, therefore, we use Pick's theorem and multiply by the area of the unit parallelogram. In FIGURE 2, therefore, we count the lattice points, put this into Pick's theorem to get $\frac{53}{2}$, and multiply by 2. The area is $A' = 53$. In general, we have the following, if U is the area of a unit parallelogram, then the area formula in Pick's theorem becomes the following.

SKEWED PICK'S THEOREM. *For a lattice polygon in a skewed lattice with unit parallelograms of area U, the area of the polygon is given by*

$$A = U\left(I + \frac{B}{2} - 1\right).$$

1.2. One-dimensional lattices. Before moving on, it will be useful to lay out how the one-dimensional lattice on a line relates to the two-dimensional lattice of the xy-plane.

Consider a line segment from the lattice point (0, 0) to the lattice point (a, b). For any integer n, the point (na, nb) must be a lattice point that lies on the line through (0, 0) and (a, b). It follows that the line segment from (0, 0) to (a, b) has no interior lattice points if and only if a and b are relatively prime (i.e., if $\gcd(a, b) = 1$). In general, given a line segment from (0, 0) to (a, b) with $d = \gcd(a, b)$, then $(\frac{a}{d}, \frac{b}{d}), (\frac{2a}{d}, \frac{2b}{d}), \ldots, (\frac{(d-1)a}{d}, \frac{(d-1)b}{d})$ are the $d - 1$ interior lattice points, and we have the following fact.

SEGMENT PICK'S THEOREM. *Given a lattice line segment in the plane with vector* $[a, b]$, *in addition to the two lattice end points, the segment would contain* $I = d - 1$ *interior lattice points, where* $d = \gcd(a, b)$.

The spacing between the lattice points on the line segment, that is, the length of a unit segment, is

$$U = \frac{\sqrt{a^2 + b^2}}{d},$$

and we have a convoluted length formula

$$L = (I+1)\frac{\sqrt{a^2+b^2}}{d},$$

where $I = d - 1$ is the number of interior lattice points. This is a Pick's theorem for lattice segments lying in the plane. We would not typically use this formula to compute the length of a segment, of course, but given the length of the segment and its slope, we could use the formula to find the number of lattice points. This is the idea we will use in solving our inverse Pick's problems for lattice polygons in space, but for segments in the plane, our segment Pick's theorem is all we will need.

2. Lattice points on a plane in space

A line will pass through some of the lattice points of the xy-plane, and just miss others. The same is true for planes lying in xyz-space. In both cases, the lattice points that lie on the line or plane exhibit regular patterns.

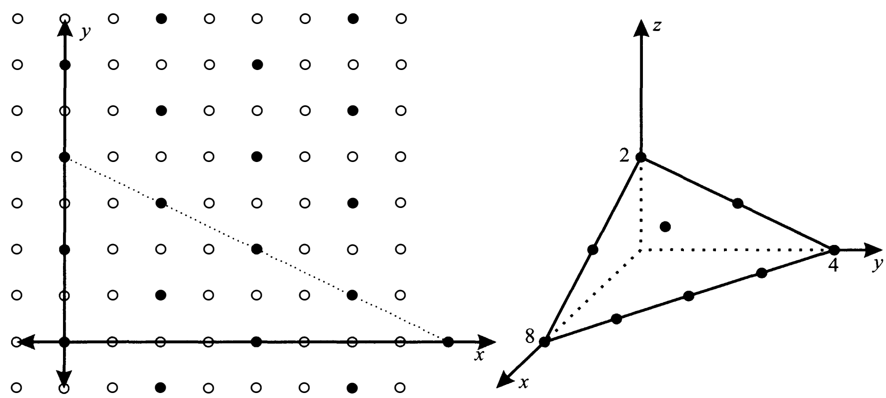


Figure 3 A lattice triangle in space

Consider the plane p with equation $x + 2y + 4z = 8$. The lattice points on a small section of p are shown on the right side of FIGURE 3. All of the three-dimensional lattice points project naturally onto the two-dimensional lattice points of the xy-plane. On the left of FIGURE 3, the lattice points of the xy-plane that correspond to the three-dimensional lattice points on p are shown with solid dots. For example, the lattice point (2, 1, 1) lies on p, and the point (2, 1) is shown with a solid dot. If we were to extend the plane p beyond the triangle, we would see that the three-dimensional lattice points that lie on p form a skewed lattice on p. We can, therefore, apply the skewed Pick's theorem to any lattice polygon lying on p. In order to do that, we would need to find a unit parallelogram, which would be easy while looking at a picture, like FIGURE 3, but drawing the picture would be hard, and we would like to be able to get this information directly from the equation of the plane.

The plane in this example has positive x-, y-, and z-intercepts, and planes like this are easy to graph, but in what follows, it will be more convenient to work with planes passing through the origin. This will not matter, however, since Pick's theorem does not care where the origin is.

2.1. Lattice planes. The version of Pick's theorem that we are after applies to planar polygonal disks, that is, a region of a plane bounded by a polygon. As mentioned earlier, when we refer to a polygon in space, we will do so with the understanding that we are really referring to a planar polygonal disk. With this in mind, a lattice polygon in space will need to lie on something we will call a *lattice plane*. The lattice points from the three-dimensional lattice that lie on this lattice plane will form a skewed lattice, and the skewed Pick's theorem will apply directly. We need, therefore, to better understand these lattice planes.

First we will establish that those planes containing three non-collinear lattice points (including the origin) are precisely those with an equation of the form $ax + by + cz = 0$ with a, b, and c being integers, not all zero. We will call a plane satisfying either of these conditions a **lattice plane**.

Suppose a plane p has three non-collinear lattice points, (0, 0, 0), (x_1, y_1, z_1) and (x_2, y_2, z_2). These points determine the vectors $\mathbf{u_1} = [x_1, y_1, z_1]$ and $\mathbf{u_2} = [x_2, y_2, z_2]$. The cross product $\mathbf{u_1} \times \mathbf{u_2} = [a, b, c]$ is such that a, b, and c are integers, and since the three points are non-collinear, this cross product is non-zero. Therefore, at least one of a, b, and c must be non-zero. Furthermore, $ax + by + cz = 0$ is an equation for the plane p.

Conversely, if $ax + by + cz = 0$ is the equation of a plane p with integer coefficients, then (0, 0, 0), $(b, -a, 0)$, and $(0, c, -b)$ are three distinct, non-collinear lattice points lying on the plane, as long as $b \neq 0$. Similarly, If $a \neq 0$ then the points (0, 0, 0), $(b, -a, 0)$, and $(c, 0, -a)$ are the three distinct, non-collinear lattice points desired, or if $c \neq 0$ we can use the points (0, 0, 0), $(0, c, -b)$, and $(c, 0, -a)$. In any case, as long as at least one of a, b, and c is non-zero, we must have three distinct, non-collinear lattice points on the plane p.

This confirms that our two defining conditions for a lattice plane are equivalent.

2.2. Level lines of lattice points. As we saw in the example, it appears that the three-dimensional lattice points that lie on a lattice plane form regular horizontal rows. These correspond to level curves, or contour lines, for the plane viewed as a surface. Consider a lattice plane p with equation $ax + by + cz = 0$, and for convenience, we assume also that the three coefficients have no common divisors. The situation is very simple, if any of the coefficients are zero, so we will assume that all the coefficients are non-zero. The plane p intersects the xy-plane in the line $ax + by = 0$, and this is a level line, or contour line, for p. All the level lines on p are parallel to this one, and we will see that all the three-dimensional lattice points on p sit nicely above these level lines. Note that if any of our coefficients were zero, our level lines would be parallel to at least one of the coordinate axes or all stacked up on each other. While these are simpler cases, they are probably best dealt with separately.

Remember that the level lines for p, or the level curves for any surface in general, are defined to lie in the xy-plane. We will begin by figuring out which of the level lines actually contain two-dimensional lattice points from the xy-plane, and we will then figure out which of these correspond to three-dimensional lattice points on p.

FIGURE 4 illustrates this for the lattice plane $x + 2y + 4z = 0$. The level lines for this plane take the form $x + 2y = k$, and the level lines for $k = -1, 0, 1, 2, 3, 4$, and 5 are shown. The integer values of k, in this case, partition the lattice points of the xy-plane, and the values of k that are divisible by 4 correspond to integer values of z, and therefore, the three-dimensional lattice points on the plane $x + 2y + 4z = 0$. These are indicated with solid dots. Note that the level line $x + 2y = 0$ is the level line on the intersection between p and the xy-plane, and the other lattice points are organized along the other level lines.

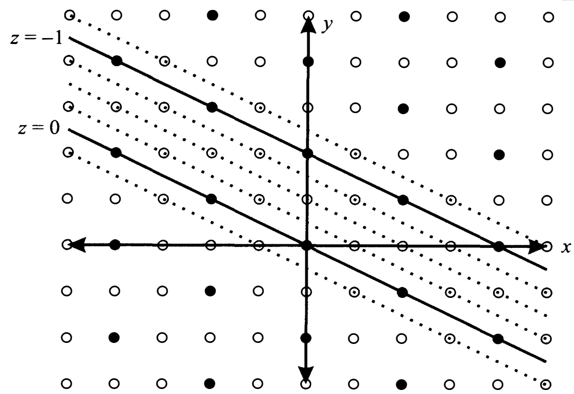


Figure 4 Some lattice level lines

Coming back to the general case, the line $ax + by = 0$ contains the lattice points (0, 0, 0) and $(b, -a, 0)$. From what we have seen before, if $d = \gcd(a, b)$, then the lattice points on this line are the points $(\frac{nb}{d}, -\frac{na}{d})$ for any integer n, and the lattice point $\left(\frac{b}{d}, -\frac{a}{d}\right)$ is adjacent to (0, 0) on this level line. In other words, the vector $[\frac{b}{d}, -\frac{a}{d}]$ determines the spacing for the rest of the lattice points on this level line.

In general, the level lines take the form $ax + by = k$. A basic fact of number theory states that for all the integer linear combinations of a and b, $ax + by$, the smallest positive linear combination is the greatest common divisor $d = \gcd(a, b)$, and furthermore, all the linear combinations are multiples of d. It follows that the lines of lattice points in the xy-plane fall on level lines of the form $ax + by = nd$ for any, and all, integers n.

Each of the lines $ax + by = nd$ are level lines for the plane p, but not all of them correspond to integer values of z. The ones that do are the level lines that contain the three-dimensional lattice points on p. Since z must satisfy $ax + by + cz = 0$, we must have $cz = -nd$, and

$$z = -\frac{nd}{c}.$$

Since a, b, and c are relatively prime, d and c must be also. Therefore, in order for z to be an integer, c must divide n. It follows that the lines of lattice points in the xy-plane that correspond to spatial lattice points on p must satisfy

$$ax + by = cmd$$

for all integers m, and the corresponding z-coordinates are $z = -md$.

The level line $ax + by = d$ is adjacent to $ax + by = 0$ as a level line of lattice points in the xy-plane. If (x_0, y_0) is an integer solution to $ax + by = d$, then integer multiples of the vector $[x_0, y_0]$ will determine a lattice point on each level line of lattice points. The points (cmx_0, cmy_0) for each m, therefore, will be a lattice point on one of the level lines corresponding to three-dimensional lattice points on p. The basic problem, therefore, lies in finding (x_0, y_0), and this can be done by inspection or by using the Euclidean algorithm. We have established the following.

LEMMA. *Let p be a lattice plane with equation $ax + by + cz = 0$, where a, b, and c are non-zero integers that are relatively prime. Then the three-dimensional lattice points on p will be those with coordinates*

$$\left(cmx_0 + \frac{nb}{d}, cmy_0 - \frac{na}{d}, -md\right),$$

where $d = \gcd(a, b)$, (x_0, y_0) *is an integer solution to* $ax + by = d$, *and* m *and* n *run through the integers.*

In vector notation, these lattice points can be expressed as

$$\left[cmx_0 + \frac{nb}{d}, cmy_0 - \frac{na}{d}, -md\right] = m\,[cx_0, cy_0, -d] + n\left[\frac{b}{d}, -\frac{a}{d}, 0\right], \qquad (1)$$

and the vectors $[cx_0, cy_0, -d]$ and $[\frac{b}{d}, -\frac{a}{d}, 0]$ form a *lattice basis* for the skewed lattice on p. This lattice basis determines a unit parallelogram, and so the cross product of these two vectors, specifically $[-a, -b, -c]$, determines the area of a unit parallelogram. This area is $U = \sqrt{a^2 + b^2 + c^2}$. The skewed Pick's theorem now applies to the three-dimensional lattice points on the boundary and in the interior of any lattice polygon lying on p.

SPATIAL PICK'S THEOREM. *Suppose* P *is a lattice polygon in space that lies on a plane* p *with equation* $ax + by + cz = 0$, *and* a, b, *and* c *are relatively prime. If* B *is the number of spatial lattice points on the boundary of* P, *and* I *is the number of spatial lattice points in the interior of* P *(i.e., lying on the planar polygonal disk bounded by* P*), then the area of* P *is given by*

$$A = \sqrt{a^2 + b^2 + c^2}\left(I + \frac{B}{2} - 1\right). \qquad (2)$$

As an example, consider the plane p with equation $x + 2y + 4z = 0$ (p is parallel to the plane in FIGURE 3, so as far as Pick's theorem is concerned, they're the same plane). Since $d = \gcd(1, 2) = 1$, we wish to find an integer solution to the equation $x + 2y = 1$. We could use the Euclidean algorithm to find this, but the solution $(x_0, y_0) = (1, 0)$ is obvious. The corresponding z-coordinate is $z = -1$. Our lattice basis vectors from equation 1 are

$$[(4)(1), (4)(0), -(1)] = [4, 0, -1] \qquad \text{and} \qquad \left[\frac{(2)}{(1)}, -\frac{(1)}{(1)}, 0\right] = [2, -1, 0].$$

The cross product of the basis vectors is $[-1, -2, -4]$, and so the area of the unit parallelogram is the magnitude, $\sqrt{21}$. The area of the triangle shown in FIGURE 3 can now be computed using Pick's theorem. Since the number of interior lattice points is $I = 1$, and the number of boundary lattice points is $B = 8$, the area of the triangle is

$$A = \sqrt{21}\left(1 + \frac{8}{2} - 1\right) = 4\sqrt{21}.$$

As you can see here, finding I is more difficult than finding A, since we can use the cross product to compute the area of the triangle directly. This version of Pick's theorem, therefore, is generally more useful in solving the inverse Pick's problem, as we will see later. If we look a little closer, however, we will see that areas can be computed nicely using Pick's theorem and a little trick.

2.3. A related area formula. The area of a lattice triangle or parallelogram is easy to compute with a cross product, but this becomes cumbersome for more complicated polygons. One nice aspect of Pick's theorem is that more complicated polygons do not make counting lattice points significantly more difficult. As we have already noted, however, counting three-dimensional lattice points is hard. If we were to project a polygon in space onto the xy-plane, however, counting lattice points becomes much easier.

For our lattice plane p with equation $ax + by + cz = 0$ and non-zero, relatively-prime coefficients, we know that the area of the unit parallelogram is $\sqrt{a^2 + b^2 + c^2}$. If we project the lattice basis vectors onto the xy-plane, we get $[cx_0, cy_0, 0]$ and $[\frac{b}{d}, -\frac{a}{d}, 0]$. The cross product of these two vectors is

$$\left[0, 0, -\frac{cax_0}{d} - \frac{cby_0}{d}\right] = -\frac{c}{d}[0, 0, ax_0 + by_0] = -\frac{c}{d}[0, 0, d] = -c[0, 0, 1],$$

and so the area of the parallelogram spanned by these two vectors is $|c|$. It follows that if A is the area of a polygon P lying on p, and A_{xy} is the area of the projection of P onto the xy-plane, then

$$\frac{A}{A_{xy}} = \frac{\sqrt{a^2 + b^2 + c^2}}{|c|}.$$

We can use Pick's theorem to find A_{xy}, and then this formula gives us A. There is nothing inherently special about projecting onto the xy-plane, so it is not surprising that

$$\frac{A}{A_{xz}} = \frac{\sqrt{a^2 + b^2 + c^2}}{|b|},$$

and

$$\frac{A}{A_{yz}} = \frac{\sqrt{a^2 + b^2 + c^2}}{|a|}.$$

Putting these together, we see that

$$A = \sqrt{a^2 + b^2 + c^2} \cdot \frac{A_{xy}}{|c|} = \sqrt{a^2 + b^2 + c^2} \cdot \frac{A_{xz}}{|b|} = \sqrt{a^2 + b^2 + c^2} \cdot \frac{A_{yz}}{|a|}, \tag{3}$$

and twisting these together surprisingly results in

$$\begin{aligned} A &= \sqrt{a^2 + b^2 + c^2} \cdot \frac{A_{xy}}{|c|} \\ &= \sqrt{\frac{a^2 A_{xy}^2}{c^2} + \frac{b^2 A_{xy}^2}{c^2} + \frac{c^2 A_{xy}^2}{c^2}} \\ &= \sqrt{\frac{a^2 A_{yz}^2}{a^2} + \frac{b^2 A_{xz}^2}{b^2} + \frac{c^2 A_{xy}^2}{c^2}} \\ &= \sqrt{A_{yz}^2 + A_{xz}^2 + A_{xy}^2}. \end{aligned} \tag{4}$$

We can find A using Pick's theorem without needing a, b, and c! Even more striking is the resemblance between this formula and the distance formula. Another proof of equation 4 can be found in **[10]**, along with the following sentence, which is almost as nice as the formula.

> The result is undoubtably known, but an informal survey revealed that it is not "well-known." [**10**, p. 255]

In any case, this formula is so nice, everyone should know it.

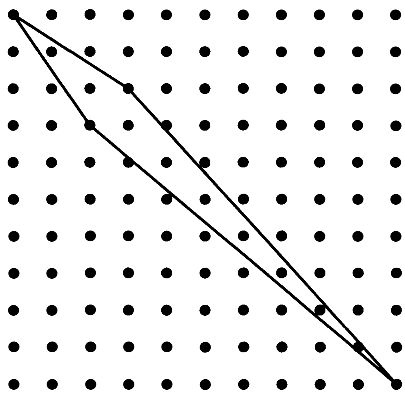


Figure 5 The projection onto the xy-plane with area A_{xy}

We now have a way of computing areas for lattice polygons in space. Consider the quadrilateral mentioned at the beginning of this article. It has vertices (0, 0, 0), (−7, 8, 3), (−10, 10, 6), and (−8, 7, 6), and lies on the plane $9x + 6y + 5z = 0$. Projecting this quadrilateral onto the xy-plane produces a quadrilateral with vertices (0, 0), (−7, 8), (−10, 10), and (−8, 7). This projection is shown in FIGURE 5, and it is easy to see that there are 9 interior lattice points and 4 boundary lattice points. By (the original) Pick's theorem, the area is

$$A_{xy} = (9) + \frac{(4)}{2} - 1 = 10.$$

We also know the equation of the plane, so by equation 3

$$A = \sqrt{(9)^2 + (6)^2 + (5)^2} \cdot \frac{10}{5} = 2\sqrt{142}. \tag{5}$$

With these two values, the areas of the other projections are easily found with equation 3. Since

$$\frac{10}{5} = \frac{A_{xz}}{6} = \frac{A_{yz}}{9},$$

we see that $A_{xz} = 12$ and $A_{yz} = 18$. On the other hand, if we did not have the equation of the plane, the areas A_{xz} and A_{yz} could have been computed with Pick's theorem, and then A could be found with equation 4. The computation of A would then look like

$$A = \sqrt{(10)^2 + (12)^2 + (18)^2} = 2\sqrt{142}.$$

2.4. The inverse Pick's problem. Finally, we can now solve the inverse Pick's problem posed at the beginning of this article. How many interior and boundary (spatial) lattice points lie on the quadrilateral? From equations 2 and 5, we know that

$$2\sqrt{142} = \sqrt{142}\left(I + \frac{B}{2} - 1\right),$$

so

$$3 = I + \frac{B}{2}.$$

The number of boundary lattice points are easy to find. In this case, the edges have vectors [−7, 8, 3], [−3, 2, 3], [2, −3, 0], and [8, −7, −6]. Since each triple of components in each of these vectors are relatively prime, there are no spatial lattice points on the sides other than at the vertices. (Even if these were not relatively prime, the number of non-vertex lattice points on each side would be the greatest common divisor minus one.) Therefore, there are $B = 4$ three-dimensional lattice points on the boundary. It then follows that there is $I = 1$ three-dimensional lattice point in the interior.

In general, therefore, given any lattice polygon P, we can find its area directly or by using Pick's theorem on the projections of P, find the number of boundary lattice points from the side vectors, and finally, find the number of interior lattice points from the spatial version of Pick's theorem.

REFERENCES

1. M. Agfalvi, I. Kadar, and E. Papp, Generalization of Pick's theorem for surface of polyhedra, *ACM SIGAPL APL Quote Quad* **29**(2) (1998) 1–12.
2. D. B. Coleman, Stretch: A geoboard game, this MAGAZINE **51** (1978) 49–54.
3. W. W. Funkenbusch, From Euler's formula to Pick's formula using an edge theorem, *Amer. Math. Monthly* **81** (1974) 647–648.
4. R. W. Gaskell, M. S. Klamkin, and P. Watson, Triangulations and Pick's theorem, this MAGAZINE **49** (1976) 35–37.
5. B. Grünbaum and G.C. Shephard, Pick's theorem, *Amer. Math. Monthly* **100** (1993) 150–161.
6. M. Klamkin and H. E. Chrestenson, Polygon imbedded in a lattice, *Amer. Math. Monthly* **70** (1963) 447–448.
7. A. C. F. Liu, Lattice points and Pick's theorem, this MAGAZINE **52** (1979) 232–235.
8. I. Niven and H. S. Zuckerman, Lattice points and polygonal area, *Amer. Math. Monthly* **74** (1967) 1195–1200.
9. D. Ren, K. Kolodziejczyk, G. Murphy, and J. Reay, A fast Pick-type approximation for areas of H-polygons, *Amer. Math. Monthly* **100** (1993) 669–673.
10. I. Rosenholtz, Calculating surface area from a blueprint, this MAGAZINE **52** (1979) 252–256.
11. P. R. Scott, An inequality for convex polygons, this MAGAZINE **52** (1979) 239–240.
12. P. R. Scott, The fascination of the elementary, *Amer. Math. Monthly* **94** (1987) 759–768.
13. D. E. Varberg, Pick's theorem revisited, *Amer. Math. Monthly* **92** (1985) 584–587.
14. C. S. Weaver, Geoboard triangles with one interior point, this MAGAZINE **50** (1979) 92–94.

Revisiting James Watt's Linkage with Implicit Functions and Modern Techniques

CHRISTOPHER J. SANGWIN
University of Birmingham
Birmingham, B15 2TT, United Kingdom
C.J.Sangwin@bham.ac.uk

The engineer James Watt (1736–1819) was a pioneer of steam power in the United Kingdom. His practical work revolutionised the rather inefficient atmospheric engines of his predecessors such as Newcomen. He vastly improved these engines in a variety of ways so that steam power became the "prime mover" of his age. In doing so he accelerated the Industrial Revolution and helped to usher in the modern industrial era.

In this article I want to re-examine a mechanism he invented to constrain the piston of a steam engine to move in a straight line. It consists of the simple linkage system illustrated in FIGURE 1. This may appear rather trivial to us now but with the rise of the importance of mechanical engineering during the Victorian period this, and related mechanisms, had many important applications. Such linkages still find many contemporary uses and modern research in robotics and flexible structures rely on the geometry which we are going to examine here.

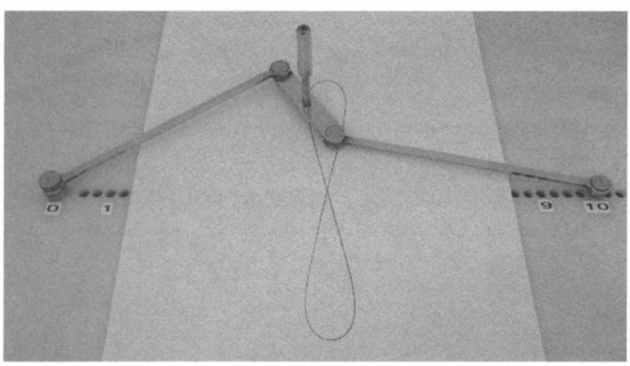


Figure 1 Watt's linkage

Watt published his linkage in a patent dated August 24th, 1784 and it is important to remember that he did not claim it produced a true straight line. He understood its importance and in his old age he wrote to his colleague Matthew Boulton

> Although I am not over anxious after fame, yet I am more proud of the parallel motion than of any other invention I have ever made.

The *parallel motion* is a simple development of the linkage shown in FIGURE 1 and this is the phrase used in historical engineering books.

Mathematical functions

The mathematical description of Watt's linkage will be in terms of *implicit functions* and so we shall consider these first. The usual modern definition of a function f is a rule which takes each element of the *domain* X and assigns a unique element in the *codomain* Y. Sometimes we think of the rule as a mapping or a procedure. Sometimes we write the function $f : X \to Y$ as $y = f(x)$, where $x \in X$ and $y \in Y$.

Some examples, where X and Y are both the set $\mathbb{R}$ of real numbers, are $f(x) = x^3$ and $f(x) = e^x$. It is surprising to learn that this particular definition of function is a relatively recent innovation. I'd like in this article to point out an older notion of function which, because of some rather exciting new techniques for solving systems of polynomial equations, is likely to become more important again. These techniques rely on the pure mathematics of rings and groups, but already have important applications in mechanical engineering and robotics design.

Just over one hundred years ago the English mathematician G. H. Hardy, in his famous book **[5]**, made the following remarks about functions.

> We must point out that the simple examples of functions mentioned above possess three characteristics which are by no means involved in the general idea of a function, viz:
>
> 1. y is determined *for every value of* x;
> 2. to each value of x for which y is given corresponds *one and only one value of* y;
> 3. the relation between x and y is expressed by means of *an analytical formula*.
>
> ... All that is essential is that there should be some relation between x and y such that to some values of x at any rate correspond values of y.

Hardy then goes on to give a number of further examples to illustrate these ideas which can be broadly separated into two groups. Firstly are those which involve a formula, equation or *algebraic* expression in x and y. This might include an infinite sum such as a series. The second are when the relationship between x and y follow from some *geometrical* construction. In this article we shall also look at both these constructions. In particular we shall find an equation which describes the geometric curve shown in FIGURE 1 by algebraic means with the help of a computer algebra system.

To begin, for a clearer separation between algebraic and the more geometric notions of function, we shall go back even further and look at the work of Leonhard Euler. Euler wrote rather a lot of mathematics. For us, the separation between algebraic and geometric notions of function are clearly explained by him in the two volumes **[2]** and **[3]**.

> §4. *A function of a variable quantity is an analytic expression composed in any way whatsoever of the variable quantity and numbers or constant quantities.* Hence every analytic expression, in which all component quantities except the variable z are constants, will be a function of that z; thus $a + 3z$; $az - 4z^2$; $az + \sqrt{a^2 - z^2}$; c^z; etc. are functions of z. **[2]**

Here Euler proposes that a *function* is that which can be expressed using an analytic expression. This notion is not that of an input-output machine, in which the domain and codomain are distinguished. "§*16. If y is any kind of function of z, then likewise, z will be a function of y.*" It is important to realise in connection with this statement that the *same algebraic expression* represents both functions. For example, if we think about $y = x^3$ again, for Euler, this is as much a function of y as it is a function of x. It all depends on how you are thinking about it at any moment.

A result of this is that functions can be multiple valued: "§*10. Finally we make a distinction between single-valued and multiple-valued functions.*" In particular, Euler gives $\sqrt{2z + z^2}$ as an example of a two-valued function.

> Whatever value is assigned to z, the expression $\sqrt{2z + z^2}$ has a twofold significance, either positive or negative.

As another example, the usual way of expressing a circle, for example

$$x^2 + y^2 = 1, \qquad (1)$$

is to Euler a "function". It fails to be a function in the modern sense, even if restricted to the domain $(-1, 1)$, since to each value of x in this range there are two choices for y. Likewise, for each value of $-1 < y < 1$ there are two possible values of x. The modern definition requires only one value for each x in the domain. This is not just nit-picking, but an important restriction on what can be a function.

We compare this with the opening of the next volume [**3**], in which quite a different notion of function is examined. This is of a single-valued function of a real variable, which can be represented by a *graph.*

> Thus any function of x is translated into geometry and determines a line, either straight or curved, whose nature is dependent on the nature of the function. [**3**, §6]

Conversely, explains Euler, *"a curve can define a function"*. It is this notion of function as *curve in space* to which [**3**] is devoted. In particular, the topic of curves generated by an algebraic equation relating x and y is developed in detail in [**3**].

What both these definitions have in common is the notion we would describe as an *implicit function.* Implicit functions do not, I think, have the popularity they deserve. In this article I want to show some situations in which using them leads to tidier mathematical results, and then to explain some applications in which they arise naturally. In particular, the curve shown in FIGURE 1 will be described by an implicit function.

The straight line

The straight line is usually described by the equation $y = mx + c$. The first point to note is that the value of y is given as an *explicit* algebraic expression in x, namely $mx + c$. So, it is clear that to each x is assigned a unique value of y. As a result of this we can draw a graph, and we find that m is the slope, and c the intersection of the line with the y-axis.

Conversely, if we have a line in the plane, then unless the line is vertical, we can write the equation relative to a pair of axes. But such a description cannot capture the case in which the line is parallel to the y-axis. Here, we have only one value of x for which there exists values of y, and indeed every value of y is identified with this value of x. We would need to write this line as $x = a$, say.

However, if we expand our notion of function to include "expressions composed howsoever from the quantities", we may include equations such as the following.

$$ax + by = p. \qquad (2)$$

In this, we can recover $y = mx + c$ by division, provided $b \neq 0$. If $b = 0$ then, $ax = p$, which expresses a vertical line. Initially this equation appears more complex, having three unknowns a, b and p instead of the usual two. But it is more general. I shall now explain why I prefer this form with some more substantial observations.

Let us assume that we have two different points (x_a, y_a) and (x_b, y_b). The task is to find a straight line between them. It turns out that the expression representing a straight line through these two points is given by the standard slope-intercept form by a rather complicated equation

$$y = \frac{(y_a - y_b)x + x_a y_b - x_b y_a}{x_a - x_b}. \tag{3}$$

If $x_a = x_b$ then we would divide by zero, which is forbidden. This corresponds to a vertical line, and as before, we cannot express this in the form $y = mx + c$.

If we re-write (3) in the form (2), then we have

$$(y_b - y_a)x + (x_a - x_b)y = x_a y_b - x_b y_a.$$

This appears to be complex, but there is a symmetry between the x_a, x_b, y_a and y_b which is arguably easier to see, and hence remember, than in (3). Putting the point a on the y-axis as $(0, a)$, and the point b on the x-axis as $(b, 0)$ this reduces to the form

$$ax + by = ab.$$

So an easy way to remember the formula is to look at the two axis intercepts: there is no need to calculate the slope, just to find the equation of the line. If we define $p := \frac{ab}{\sqrt{a^2+b^2}}$ then this can again be re-written in the form

$$\sin(t)x + \cos(t)y = p,$$

where t is the angle of the line to the x-axis and p now represents the perpendicular distance of the line from the origin. In this form we recover an equation in only two unknowns, t and p. In all these forms the symmetry between x and y, and the two interpolated points, is arguably more natural than in the traditional form of the equation for a straight line.

Circles are almost always expressed in an implicit way, as is the ellipse

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

If we use the form (2) for the equation of a line, then it an exercise for you to show that the tangent to this ellipse, through the point (x_0, y_0) is given by

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

Again, this seems to be simpler and more general than the form $y = mx + c$ would permit.

Solving the cubic equation

In this section I want to illustrate how the straight lines (2) can help us solve a cubic equation

$$w^3 + aw^2 + bw + c = 0,$$

by graphical means. Our first observation is that by defining $z = w - \frac{a}{3}$ we have

$$z^3 + pz + q = 0. \tag{4}$$

Notice the z^2 term is missing. The equation (4) is known as the *reduced cubic* and it is the first step in the method of finding the general formula for the roots of the cubic, which [4] develops in full detail.

We shall divide by z^3 and then define $p = x$ and $q = y$. This gives us the equation

$$\frac{x}{z^2} + \frac{y}{z^3} = -1.$$

For each value of z this gives us a straight line. Furthermore, for each point on this straight line the equation (4) holds. If we think of the plane as the (p, q) space of all cubic equations (4), then points on these straight lines are solutions of the equation (4). Hence, to solve a particular equation (4) we look to see which line(s), if any, pass through the point (p, q).

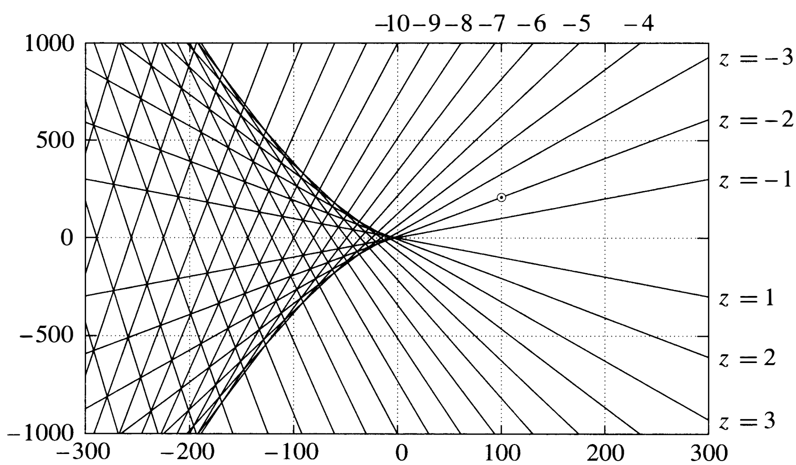


Figure 2 Graphical solution to the cubic equation: real solutions

These lines have been plotted in FIGURE 2. Let us examine the line labeled $z = -2$. This certainly passes through the point $p = 100$, $q = 208$. Hence, we know that the reduced cubic $z^3 + 100z + 208 = 0$ has a real root $z = -2$. Furthermore, from FIGURE 2 it appears that every point in the (p, q) plane has at least one line through it, so every reduced cubic has at least one real root. To the left of the figure it appears that we have a region in which three lines always pass through a given point (p, q). Hence, here we always have three real roots for (p, q) in this region.

The boundary of the region in which the cubic has three real roots is given by the equation

$$4p^3 = -27q^2,$$

which is another example of an implicit function.

Visualizing implicit functions

One clear advantage of the contemporary function-as-function-machine approach is the ease with which such functions can be visualised. You can simply plot the graph, or have a machine approximate this for you. Functions defined implicitly by equations are hard to vizualise, at least initially.

The following is a key observation when trying to sketch the graph of an expression such as $p(x, y) = 0$. We begin by factoring $p(x, y)$ over the real numbers. The expression is satisfied if any of the factors equal zero. Hence, we plot the graphs of each of the factors separately, and then combine them by superposition. The graph of $p(x, y) = 0$ is the superposition of the graphs of its factors. For example, the graph of the expression $x^2 = y^2$, or rather $(x - y)(x + y) = 0$, is the superposition of the two lines $y = x$ and $y = -x$.

Irreducible expressions, such as (1), have their own particular *forms*, something which [3] examines at length. We know this, from familiarity, to be a circle of radius 1, centered at the origin. Familiarity with the other second order curves, that is to say conic sections, is something which comes with regular use. Your might like to discover why adding x^2y^2 to the right hand side of (1) might be described as *"squaring the circle"*.

Computer algebra systems, or other technology, can help here, although you should beware that many CAS's fail to plot simple implicit functions by failing to factor the expressions and act on this simple observation. For example, $(x - 2)^2 = 0$ fails to change sign for any x and y, and as a result an embarrassing number of mainstream CAS's fail to plot this convincingly as the line $x = 2$.

Linkages

To illustrate somewhat more substantially the contemporary importance of implicit functions, we shall examine Watt's linkage, and four-bar linkages in general. Watt's linkage is shown in FIGURE 1. This consists of three movable bars, fixed to a base which constitutes a fourth bar. We are interested in the path of the pen, fixed to the middle bar, over all physically realistic positions of the linkage. You can see this path in the figure.

The original purpose was to constrain the movement of a piston in the cylinder of a steam engines to move in an approximately straight line. Other four-bar linkages which were proposed for this purpose are shown in FIGURE 3. In the linkage to the left, the pen is on an extension of the middle link. In the linkage to the right, the pen is connected to the middle linkage, but is offset from the linkage itself, and this represents

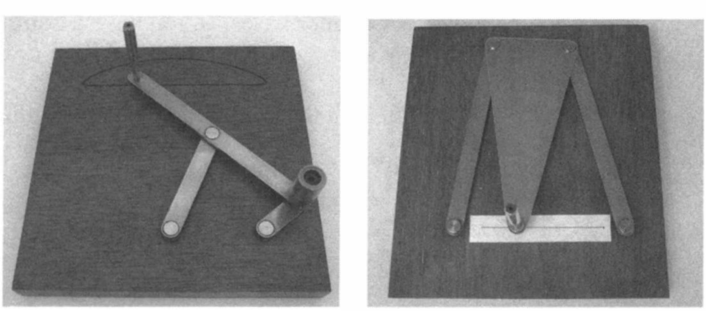


Figure 3 Other variations of the four-bar linkage

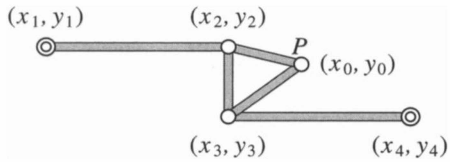


Figure 4 A schematic of general four-bar linkages

the most general situation. For more details of the history of the applications of this problem see [**1**].

From our point of view, all these will be treated in an identical way, as shown in FIGURE 4. From this sketch, we are most interested in the locus of P, for all positions of the linkage. It is very worthwhile making physical models of these linkages, either from commercially available model kits or from materials you have on hand. Alternatively you might like to implement these on a dynamic geometry package, such as GeoGebra (`http://www.geogebra.at/`).

We shall take a more algebraic approach. We first notice that the distance between (x_1, y_1) and (x_2, y_2) is fixed, at r_1 say. Hence by the Pythagorean Theorem we have

$$(x_1 - x_2)^2 + (y_1 - y_2)^2 = r_1^2.$$

Indeed, to describe the whole linkage three more applications of the Pythagorean Theorem provide us with the following equations.

$$(x_2 - x_3)^2 + (y_2 - y_3)^2 = r_2^2,$$
$$(x_3 - x_4)^2 + (y_3 - y_4)^2 = r_3^2,$$
$$(x_4 - x_1)^2 + (y_4 - y_1)^2 = r_4^2.$$

Now, to describe the position of $P = (x_0, y_0)$, relative to (x_2, y_2) and (x_3, y_3) we need two further applications

$$(x_2 - x_0)^2 + (y_2 - y_0)^2 = r_5^2,$$
$$(x_3 - x_0)^2 + (y_3 - y_0)^2 = r_6^2.$$

Since we fix both ends of the linkage, we specify (x_1, y_1) and (x_4, y_4), hence defining r_4 and making one equation redundant. The task is to solve the system consisting of the remaining five equations. By "solve", we mean to specify the lengths of the links $r_1, \dots, r_6$, and then to eliminate (x_2, y_2) and (x_3, y_3) to leave a single equation in only (x_0, y_0) as the solution.

This looks hopeless, but in fact it can be done with computer algebra in a straightforward way using a concept known as Gröbner bases. If you have a computer algebra system, such as Maple, Mathematica, Maxima or some other CAS, you will probably already have the software necessary to do this.

For reference, if you have Maple 9.5, then the commands look something like this.

```
> restart:with(Groebner):with(Ore_algebra):
> P1:=(x1-x2)^2+(y1-y2)^2-r1^2;
> P2:=(x2-x3)^2+(y2-y3)^2-r2^2;
> P3:=(x3-x4)^2+(y3-y4)^2-r3^2;
> P4:=(x4-x1)^2+(y4-y1)^2-r4^2;
> P5:=(x0-x2)^2+(y0-y2)^2-r5^2;
> P6:=(x0-x3)^2+(y0-y3)^2-r6^2;
```

Notice that we are using an expression $(x_1 - x_2)^2 + (y_1 - y_2)^2 - r_1^2$ instead of an equation. This is really only an input syntax issue, and from this point onwards it is implied that such an expression represents an equation with right hand side zero.

Next to examine in more detail a specific example we assign some lengths to these expressions.

```
> y1:=0; y4:=0; x1:=-5; x4:=5;
> r1:=5; r2:=2; r3:=5;
> r5:=1; r6:=1;
> S := [P1,P2,P3,P5,P6];
```

Notice that we have not included the redundant equation P4 in the list S, which is the resulting system of expressions representing our equations. Next the CAS computes the "Gröbner basis" for this system, and then we remove any expressions which have any of the variables x_2, y_2, x_3, or y_3.

```
> LinkageGB:=gbasis(S,lexdeg([x2,y2,x3,y3],[x0,y0])):
> Linkage:=op(remove(has,LinkageGB,{x2,y2,x3,y3})):
> factor(Linkage);
```

The result of this calculation, which took approximately six minutes to complete, is the expression

$$\begin{aligned}(y_0{}^6 + 2\,y_0{}^4 - 99\,y_0{}^2 + 3\,x_0{}^2y_0{}^4 - 96\,x_0{}^2y_0{}^2 \\ + 2401\,x_0{}^2 + 3\,x_0{}^4y_0{}^2 - 98\,x_0{}^4 + x_0{}^6)^2.\end{aligned} \tag{5}$$

In terms of the solution to the original system, this reduces to

$$\begin{aligned}y_0{}^6 + 2\,y_0{}^4 - 99\,y_0{}^2 + 3\,x_0{}^2y_0{}^4 - 96\,x_0{}^2y_0{}^2 + 2401\,x_0{}^2 \\ + 3\,x_0{}^4y_0{}^2 - 98\,x_0{}^4 + x_0{}^6 = 0.\end{aligned} \tag{6}$$

While we have not been able to find y_0 in terms of an explicit expression in x_0, even finding this implicit function is quite an achievement. In fact, it is hopeless to suppose that the figure of eight curve shown in FIGURE 1 could result in a single-valued $y_0 = f(x_0)$.

At this point you may be feeling some disquiet that I am not going to explain exactly what a Gröbner basis *is* or exactly what you have *done* with it. That is your task to investigate using the many available references. A good place to start is the "help" files on your computer algebra system. For example, in Maple type `help(Grobner);`. What I do hope to convince you of is that these (relatively) new computational techniques are particularly useful by applying them to a classical problem which does not appear to be solvable by traditional means, such as those of [**6**]. Indeed, so useful are they for solving apparently hopeless systems, such as that above, that I confidently predict that the implicit function itself will become much more important and widely used.

In fact, we can do rather a lot more with these techniques. In particular, rather than specifying the end points and linkage lengths at the outset, we shall keep these variables in the system of equations. Now we shall solve the same system, and eliminate (x_2, y_2) and (x_3, y_3), finding a single implicit equation for (x_0, y_0), in terms of the end points and linkage lengths.

This appears to be an even more hopeless a task, since we have five nonlinear equations with four variables to eliminate and a further nine parameters which will be left. And yet it can be done. Specifying only that $y_1 = 0$, so that one end of the linkage is effectively anchored on the x-axis, Maple is (eventually) able to find the required equation. The restriction $y_1 = 0$ is not necessary, but it does make the computations finish in a sensible amount of time. Unfortunately even then, this is rather too long to print here, having some 27255 terms in the equation.

Having obtained this equation we can use it to investigate the general behavior of four-bar linkages in which one end is anchored to the x-axis. A first experiment is suggested by FIGURE 1. Notice that the end points of the linkage can be moved to various positions along the x-axis. While these have been labeled from 0 upwards on the diagram, it makes sense for us to have

$$(x_1, y_1) = (-r, 0), \quad (x_4, y_4) = (r, 0)$$

to obtain symmetry, and then as before take

$$r_1 = r_3 = 5, \quad r_2 = 2, \quad r_5 = r_6 = 1.$$

Doing this we obtain, as a polynomial in r, the following.

$$r^4\left(y_0^2 + x_0^2\right) + 2r^2\left(y_0^4 - 26y_0^2 - x_0^4 + 24x_0^2\right) + \left(y_0^2 + x_0^2 - 24\right)^2\left(y_0^2 + x_0^2\right) = 0.$$

Not surprisingly, by substituting $r = 5$ into this equation we recover (6).

FIGURE 5 illustrates the locus of the center point in the middle arm of Watt linkages, for various values of r. When $r = 0$ both long arms are fixed at the origin, and we have the circle $y_0^2 + x_0^2 = 24$. The figure eight shape shown in the model of FIGURE 1 is clearly reproduced for $r = 5$ in FIGURE 5, and the algebraic expression for this curve is given in (6).

Recall that Watt's original intention was to draw an approximate straight line. For $r = 2$, FIGURE 5 appears to show a much longer, approximately straight section in the curve. Perhaps moving the fixed points to $r = 2$, rather than $r = 5$, gives a better straight line? Indeed it does, and this configuration was actually proposed for this purpose by the Russian mathematician Pafnuty Chebyshev (1821–1894), who was fascinated by linkages. We shall refer to this as "Chebyshev's approximate straight line" and you are encouraged to actually make this for yourself, or at the very least sketch the linkage. If you do this you will see how the two disconnected parts of the curve correspond to physical configurations of the linkage.

Notice also that the curve shown to the left of FIGURE 3 looks very similar indeed to that for Watt's linkage with $r = 2$. And yet Watt's linkage has quite a different form than this model, in particular the pen on Watt's linkage lies mid-way along the center link. This suggests another line of inquiry: can we find more than one linkage which generates a particular curve?

To do this let us start with our general expression for the four-bar linkage. Into this we substitute the values for Chebyshev's approximate straight line to obtain the equation

$$\begin{aligned}\big(&{y_0}^6 - 40\,{y_0}^4 + 384\,{y_0}^2 + 3\,{x_0}^2{y_0}^4 - 96\,{x_0}^2{y_0}^2 + 784\,{x_0}^2 \\ &+ 3\,{x_0}^4{y_0}^2 - 56\,{x_0}^4 + {x_0}^6\big)^2 = 0.\end{aligned} \tag{7}$$

This is plotted in FIGURE 5, and labeled $r = 2$. Now we compare the coefficients of (7) with those of the 27255 terms in the general equation. Each comparison provides an equation in the unknown positions of the end points, and also the lengths of the links. Again we have a large number of non-linear equations in nine unknowns. How can we possibly hope to solve these, and hence find alternative four-bar linkages which

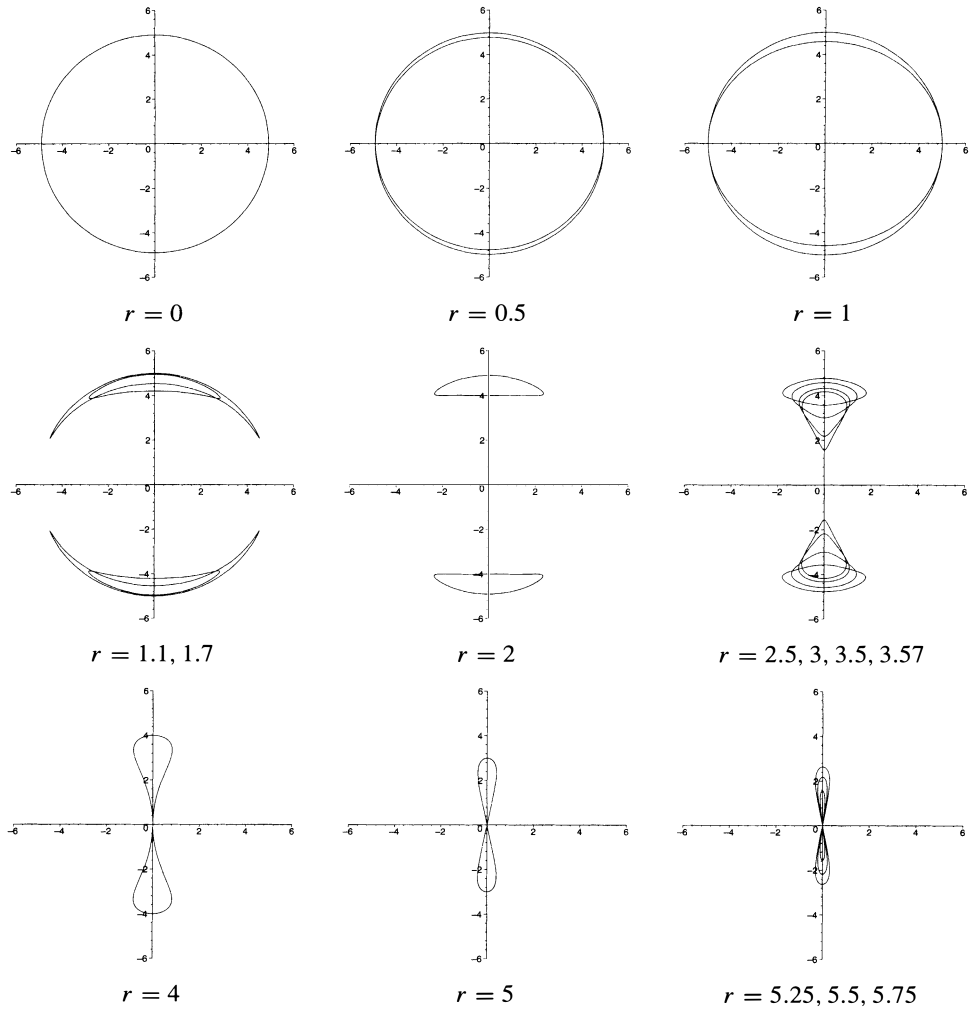


Figure 5 The locus of Watt's linkage for various separations r

produce Chebyshev's approximate straight line? This is simple; we apply the Gröbner basis technique again to solve this system of equations, just as we did before.

One result is

$$(x_1, y_1) = (0, 0), \quad (x_4, y_4) = (2, 0)$$

and

$$r_1 = \frac{5}{2}, \quad r_2 = \frac{5}{2}, \quad r_3 = 1, \quad r_5 = \frac{5}{2}, \quad r_6 = 5.$$

A model of this linkage is shown to the left of FIGURE 3. Having found this result by algebraic techniques, it is relatively straightforward to find a simple and purely geometrical proof that the curves generated are identical by drawing the two alternative linkages on the same diagram. If you use dynamical geometry then the proof "jumps out" as the linkages move in unison.

What about the most general case? If we take a linkage, such as that shown to the right of FIGURE 3, then it is always possible to find exactly *two* others which generate the same curve. This is the famous *triple generation theorem*, and a simple purely geometric proof is given in, for example, [**7**].

What these algebraic expressions, and their associated graphs, lack is the *movement* obtained by the linkages. For example, if the point (x_2, y_2) is rotated at a constant speed, then how does the velocity of P change? These linkages have a satisfying aesthetic quality to them, which can only be experienced by making the linkages. This can be either with a physical model of your own, or virtually in dynamic geometry. For a particularly intriguing example, try $r_1 = r_3 = 5$, $r_2 = 6$, the point P the midpoint of the bar, that is to say $r_5 = r_6 = 3$, and with $(x_1, y_1) = (-2, 0)$, and $(x_1, y_1) = (2, 0)$. Many other configurations are possible.

The four-bar linkage is perhaps the simplest of mechanisms: a three bar linkage forms a triangle and hence is rigid, and any more bars give the potential for greater degrees of freedom. The techniques I have sketched above are becoming widely used for the design of mechanisms in general, and the design of robots in particular. They allow the user to accurately model the movement of complex joints, both in the plane and in three dimensions. They allow a designer to search for alternative configurations of links with the same, or similar, paths. Furthermore, there are many other situations in mathematics which generate systems of polynomial equations. Where these need to be manipulated, the Gröbner basis technique is invaluable. All that can be hoped for as an outcome in general is an implicit function. As a result of this, I predict that implicit functions will become much more important in the near future.

REFERENCES

1. J. Bryant and C. J. Sangwin, *How Round Is Your Circle? Where Engineering and Mathematics Meet*, Princeton University Press, 2008.
2. L. Euler, *Introduction to Analysis of the Infinite*, vol. I, trans. by J. Blanton from the Latin *Introductio in Analysin Infinitorum, 1748*, Springer, 1988.
3. L. Euler, *Introduction to Analysis of the Infinite*, vol. II, trans. by J. Blanton from the Latin *Introductio in Analysin Infinitorum, 1748*, Springer, 1990.
4. L. Euler, *Elements of Algebra*, Tarquin Publications, 2006.
5. G. H. Hardy, *A Course of Pure Mathematics*, Cambridge University Press, 1908.
6. J. A. Hrones and G. L. Nelson, *Analysis of the Four Bar Linkage*, Press of the Massachusetts Institute of Technology, and Wiley, New York, 1951.
7. R.C. Yates, *Geometrical Tools: A Mathematical Sketch and Model Book*, Educational Publishers, Incorporated Saint Louis, 1949.

Quasigeometric Distributions and Extra Inning Baseball Games

DARREN GLASS
Gettysburg College
200 N. Washington St.
Gettysburg, PA 17325
dglass@gettysburg.edu

PHILIP J. LOWRY
City University of New York
New York, NY 10016
plowry1176@aol.com

Each July, the eyes of baseball fans across the country turn to Major League Baseball's All-Star Game, gathering the best and most popular players from baseball's two leagues to play against each other in a single game. In most sports, the All-Star Game is an exhibition played purely for entertainment. Since 2003, the baseball All-Star Game has actually 'counted', because the winning league gets home field advantage in the World Series. Just one year before this rule went into effect, there was no winner in the All-Star Game, as both teams ran out of pitchers in the 11th inning and the game had to be stopped at that point. Under the new rules, the All-Star Game must be played until there is a winner, no matter how long it takes, so the managers need to consider the possibility of a long extra inning game. This should lead the managers to ask themselves what the probability is that the game will last 12 innings. What about 20 innings? Longer?

In this paper, we address these questions and several other questions related to the game of baseball. Our methods use a variation on the well-studied geometric distribution called the quasigeometric distribution. We begin by reviewing some of the literature on applications of mathematics to baseball. In the second section we will define the quasigeometric distribution and examine several of its properties. The final two sections examine the applications of this distribution to models of scoring patterns in baseball games and, more specifically, the length of extra inning games.

1. Sabermetrics

While professional baseball has been played for more than a century, it has only been in the last few decades that people have applied mathematical tools to analyze the game. Bill James coined the term 'Sabermetrics' to describe the analysis of baseball through objective evidence, and in particular the use of baseball statistics. The word Sabermetrics comes from the acronym SABR, which stands for the Society for American Baseball Research [**12**].

Before SABR was ever organized, and before sabermetrics was a word, the influence of statistics over the strategy used by a manager in professional baseball was minimal. No manager would have ever thought of having charts on what each batter had done against each pitcher in the league. Now things are different. Since the publication of Michael Lewis's book *Moneyball* in 2003 [**10**], even most casual baseball fans have become familiar with Sabermetric statistics such as OPS ("on-base plus slugging", which many people feel is a better measure of offensive skill than the traditional statistics such as batting average or RBIs) and Win Shares (a statistic developed by Bill

James in [**8**] which attempts to measure the all-around contributions of any player), and there has been a proliferation of books and websites for the more dedicated fans to pursue these interests.

Sabermetrics has had a profound influence not just in the living room, but also in the clubhouse as it has begun to affect the strategy of the game. In the last decade, Sabermetrics devotees such as Billy Beane, Theo Epstein, Paul DePodesta, and Bill James himself have all worked in the front offices of Major League baseball teams, and these approaches are often given some of the credit for the Red Sox winning the 2004 World Series [**6**].

Sabermetricians attempt to use statistical analysis to answer all sorts of questions about the game of baseball: whether teams should intentionally walk Barry Bonds, whether Derek Jeter deserves his Gold Glove, which players are overpaid (or underpaid), when closing pitchers should be brought into the game, and whether or not batting order matters are just some of the questions that have had many words written about them. For readers interested in these questions, websites such as *Baseball Prospectus* [**2**] and journals such as *By The Numbers* [**5**] are a great place to start reading. Alan Schwarz's book *The Numbers Game* [**13**] provides an excellent historical perspective, and Albert and Bennett's book *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game* [**1**] is a good introduction to some of the mathematical techniques involved.

One recurring theme in the sabermetric literature is the question of how likely certain records are to be broken and how unlikely these records were to begin with. For example, now that Barry Bonds has set the career homerun record, many people are curious whether we should expect to see any player pass Bonds in our lifetime. Several recent articles ([**3**], [**4**]) in *The Baseball Research Journal* have asked the question "How unlikely was Joe DiMaggio's 56 game hitting streak?" and have come to different answers depending on the methods they use to look at the question. This question is of the same flavor as the question we address in Section Four, as we use the mathematical models developed to examine how likely a 20 inning game is to occur, and how unlikely the longest recorded game of 45 innings really was.

2. Distributions

Geometric distributions. To begin, let us recall what we mean by a distribution in the first place.

DEFINITION 2.1. A probability distribution on the natural numbers is a function $f : \mathbb{N}_0 \to [0, 1]$ (where $\mathbb{N}_0$ denotes the nonnegative integers) such that $\sum_{n=0}^{\infty} f(n) = 1$. The mean (or expected value) of a discrete distribution f is given by $\mu = \sum nf(n)$ and the variance is given by $\sigma^2 = \sum (n - \mu)^2 f(n)$.

DEFINITION 2.2. A geometric distribution is a distribution such that for all $n \geq 1$, $f(n) = f(0)\ell^n$ for some fixed $0 < \ell < 1$.

We note that geometric distributions are the discrete version of the exponential decay functions which are found, for example, in half-life problems. In particular, if f is a geometric distribution, then we see that

$$1 = \sum_{n=0}^{\infty} f(n)$$

$$
\begin{aligned}
&= \sum_{n=0}^{\infty} f(0)\ell^n \\
&= f(0)\sum_{n=0}^{\infty} \ell^n \\
&= \frac{f(0)}{1-\ell},
\end{aligned}
$$

therefore $f(0) = 1 - \ell$. Thus, the entire distribution is determined by the value of ℓ. It is a straightforward computation to see that the mean of this distribution is $\frac{\ell}{1-\ell}$ and the variance is $\frac{\ell}{(1-\ell)^2}$.

Quasigeometric distributions. In this paper, we wish to discuss a variation of geometric distributions which can reasonably be referred to as quasigeometric distributions, as they behave very similarly to geometric distributions. These distributions are defined so that they are geometric other than at a starting point. In particular, we want there to be a common ratio between $f(n)$ and $f(n+1)$ for all $n \geq 1$ but not (necessarily) the same ratio between $f(0)$ and $f(1)$. To be explicit, we make the following definition:

DEFINITION 2.3. A quasigeometric distribution is a distribution so that for all $n \geq 2$, $f(n) = f(1)d^{n-1}$ for some $0 < d < 1$. We call d the *depreciation constant* associated to the distribution.

Just as geometric distributions are completely determined by the value of k, a quasigeometric distribution is entirely determined by the values of d and $f(0)$ (which we will often denote by a). In particular, a computation analogous to the one above shows that for $n \geq 1$, $f(n) = (1-a)(1-d)d^{n-1}$. Given this, it is possible to compute the mean and variance of the distribution as follows:

$$
\begin{aligned}
\mu &= \sum_{n=0}^{\infty} nf(n) \\
&= \sum_{n=1}^{\infty} n(1-a)(1-d)d^{n-1} \\
&= (1-a)(1-d)\sum_{n=1}^{\infty} nd^{n-1} \\
&= (1-a)(1-d)(1-d)^{-2} \\
&= \frac{1-a}{1-d}, \qquad (1) \\
\sigma^2 &= \sum_{n=0}^{\infty} n^2 f(n) - \mu^2 \\
&= \sum_{n=1}^{\infty} n^2(1-a)(1-d)d^{n-1} - \mu^2 \\
&= (1-a)(1-d)\sum_{n=1}^{\infty} n^2 d^{n-1} - \mu^2
\end{aligned}
$$

$$= \frac{(1-a)(1+d)}{(1-d)^2} - \frac{(1-a)^2}{(1-d)^2}$$

$$= \frac{(1-a)(a+d)}{(1-d)^2}. \qquad (2)$$

Conversely, we note that a quasigeometric distribution is uniquely determined given μ and σ^2 (although not all pairs (μ, σ^2) determine a quasigeometric distribution). In particular, if $s > |m - m^2|$ and we set

$$a = \frac{m + s - m^2}{m + s + m^2} \quad \text{and} \quad d = \frac{m^2 + s - m}{m^2 + s + m},$$

then the quasigeometric distribution given by a and d will have mean m and variance s. In statistics, this method of describing a distribution is called the method of moments.

3. Baseball scoring patterns

Runs scored per inning. It has been observed by several people (see **[9]**, **[15]**, **[16]**) that the number of runs scored per inning by a given baseball team fits a quasigeometric distribution (although they do not use this language). In TABLE 1, we have provided a table of the probabilities that a given number of runs is scored in an inning based on several different datasets and we see that the same general pattern persists. Woolner's data **[16]** separates teams by their strength, trying to see if teams that score an average of 3.5 runs per game have different scoring patterns than those that score 5.5 runs per game. The data compiled by Jarvis **[9]** separates teams by league to see how scoring patterns are affected by the different rules (designated hitter, etc.) as well as the different cultures in the American League and the National League.

TABLE 1: Probability of scoring a given number of runs in an inning

Dataset	0 runs	1	2	3	4	5
Woolner (all)	0.730	0.148	0.068	0.031	0.014	0.006
Woolner (3.5 RPG)	0.760	0.140	0.059	0.024	0.011	0.004
Woolner (5.5 RPG)	0.679	0.161	0.079	0.042	0.022	0.009
Jarvis (AL)	0.722	0.151	0.070	0.032	0.014	0.006
Jarvis (NL)	0.731	0.150	0.067	0.030	0.013	0.006

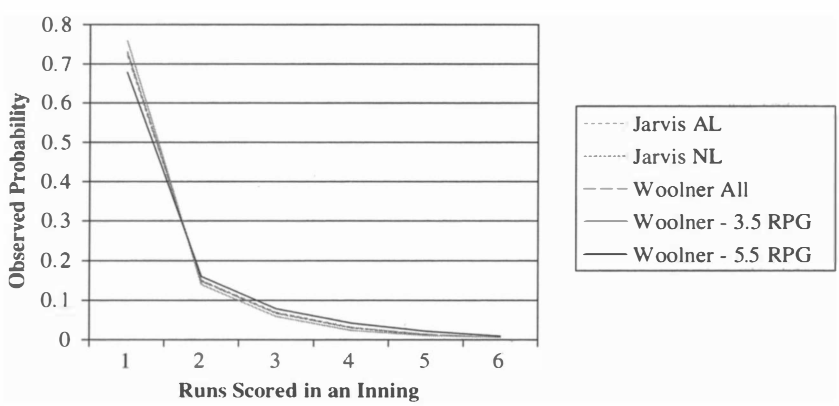


Figure 1 Runs scored per inning

One notes from FIGURE 1 that, after an initial dropoff, the probability of scoring a given number of runs appears to fit a typical exponential curve. This is not a surprising result to baseball fans, because it is what one would intuitiviely expect if the only way that a runner reached base was by hitting a single: the first run would be more difficult to score as it requires multiple hits, but the probability that each additional run will score coincides with the probability that the batter gets a hit.

Over each of these different data sets, one can compute the mean and standard deviation, and in turn compute the associated values of a and d that would define the appropriate quasigeometric distribution from the equations at the end of Section 2.

TABLE 2: Computations of a and d for datasets

Dataset	m	s	a	d
Woolner (all)	0.484	0.999	0.727	0.436
Woolner (3.5 RPG)	0.408	0.896	0.772	0.444
Woolner (5.5 RPG)	0.627	1.173	0.642	0.429
Jarvis (AL)	0.503	1.024	0.715	0.435
Jarvis (NL)	0.478	0.986	0.730	0.434

We see from TABLE 2 that the value of the depreciation constant d does not change very much, even when we look across leagues or across varying strengths of teams. In fact, it does not change significantly even if we compare different eras. This observation will be the key assumption of our model. For the duration of this paper, we will assume that scoring patterns in a given inning fit a quasigeometric distribution with a value of $d = 0.436$ for the depreciation constant as suggested by the full database in [**16**]. The value of a, on the other hand, does change significantly with the strength of a team. One way of interpreting this result is that the difference between the quality of teams is mainly in the probability that they score a single run in a given inning. Then, after the first batter crosses the plate, all teams are more or less equally successful at continuing the scoring drive.

The work of Smith, featured in [**14**], shows that an average major league team scores 0.487 runs per inning. TABLE 3 computes the probability that a team which scores at this rate will score a given number of runs in an inning according to this quasigeometric model, and compares this with the probability observed in Woolner's dataset.

TABLE 3: Runs per inning: Quasigeometric Model vs. Woolner's Data

Number of Runs	Predicted Prob.	Observed Prob.
0	0.725	0.730
1	0.155	0.148
2	0.068	0.068
3	0.029	0.031
4	0.013	0.014
5	0.006	0.006

Runs scored per game. Of course, as any baseball fan who has watched his team squander a lead can tell you, games are not won or lost by the number of runs scored in a given inning but instead by the number of runs scored in the full nine (or more) innings. So one would like a formula to determine the number of runs scored in a nine-inning game. In order to do so, we first make the assumption that all innings are independent of one another. While this assumption is almost certain to be overly strong—teams are likely to face similar pitchers, weather, and park effects in consecutive innings—it greatly simplifies the problem. Furthermore, we will see that it leads to mathematical results that match with actual game data.

We denote the probability that a team scores n runs in nine innings as $F(n)$, and note that

$$F(n) = \sum f(n_1)f(n_2)\dots f(n_8)f(n_9),$$

where the sum ranges over all 9-tuples of nonnegative integers $(n_1, \dots, n_9)$ which sum to n and $f(n_j)$ is the probability that the team scores n_j runs in inning j.

If a team scores n runs in a game, then we know that the team must score in between one and $\min(n, 9)$ different innings. Breaking up by these cases, we can compute

$$F(n) = \sum_{i=1}^{\min(n,9)} \binom{9}{i} f(0)^{9-i}\left(\sum f(n_1)\dots f(n_i)\right),$$

where the interior sum is over all ordered i-tuples of positive integers summing to n. If we now invoke our assumption that the probability of scoring a given number of runs in an inning is quasigeometric (and independent of the inning), and therefore that $f(0) = a$ and $f(n_i) = (1-a)(1-d)d^{n_i-1}$ for all $n_i \geq 1$, we can calculate that

$$F(n) = \sum_{i=1}^{\min(n,9)} \binom{9}{i}\binom{n-1}{i-1} a^{9-i}d^{n-i}(1-a)^i(1-d)^i.$$

In this formula, i represents the number of innings in which the team scores, a represents the probability that a team goes scoreless in a given inning, and d represents the depreciation constant, which we are assuming is equal to 0.436 for all teams. One way to view the $\binom{n-1}{i-1}$ term is that it counts the number of ways to divide n runs among i innings. It will be more useful to us to translate this result in terms of the strength of a given team. To do this, we note that Equation (1) showed that to model a team that scores an average of m runs per inning we should choose $a = 1 - m(1-d)$. Doing so, we compute:

$$F(n) = \sum_{i=1}^{\min(n,9)} \binom{9}{i}\binom{n-1}{i-1} m^i(1-(1-d)m)^{9-i}(1-d)^{2i}d^{n-i},$$

where again d is the depreciation constant 0.436 and m represents the average number of runs per inning that a team scores. TABLE 4 computes $F(n)$ for a team that scores the historical average of 0.487 runs per inning and compares these values with the empirical distribution of runs per game scored by National League teams between 1969 and 2002.

One sees that this quasigeometric model appears to give a good approximation of reality, and therefore we might want to see how this type of model can be used to answer many different types of questions. In the following section, we will look at the question of how often we should expect games to last 20 innings or more, but before moving on to that, we think it would be interesting to note that one could use this model to compute the odds that a team of a given strength would beat another team of a given strength. In particular, we note that the 2003 Atlanta Braves scored an average of 0.618 runs per inning, whereas the 2003 New York Mets scored an average of 0.443 runs per inning. While this is clearly a lopsided matchup, one of the beautiful things

TABLE 4: Number of runs per game predicted by model vs. actual game data

Number of Runs	$F(n)$ = Prob in game	% of NL Games
0	0.055	0.062
1	0.107	0.108
2	0.138	0.139
3	0.145	0.148
4	0.135	0.134
5	0.114	0.113
6	0.091	0.088
7	0.068	0.068
8	0.046	0.049
9	0.034	0.033
10	0.023	0.023

about the game of baseball is that underdogs often win, and one wonders what the probability of the Mets winning a given game against the Braves would have been.

One can use the quasigeometric model in order to approach this question. In particular, we can use the strengths of each team to calculate $F_B(n)$ (resp. $F_M(n)$), the probability that the Braves (resp. the Mets) will score n runs in nine innings. Given these functions and the assumption that their scoring is independent of each other, we can compute that there is roughly a 31% chance that the Mets will be ahead after nine innings, a 60% chance that the Braves will win, and a 9% chance that the game will go into extra innings. If one looks at what actually happened in the games played between the two teams in 2003, we see that the Braves won 11 of the 19 (or 58%) of the games, with none going into extra innings. These results correspond quite closely with the predictions of our model, given the small sample size involved.

4. Extra inning games

One of the things about baseball that its fans love the most, and its detractors like the least, is the fact that it is free of the artificial boundaries of time within which the clock confines other sports. This freedom from time constraints helps to shape the unique charm that is an evening at the ballpark, for fans never know when they may be the first to be enchanted until past sunrise by the first-ever wild ten-hour 46-inning slugfest.

This idea brings us back to the question posed in the introduction: what is the probability that a given baseball game lasts twenty innings or more? Alternatively, there has only been one Major League Baseball game to last twenty-six innings in history, and one could ask if the mathematical models predict more or fewer than have actually occurred.

To answer these questions, one must first consider what the probability is that a game goes into extra innings at all. In particular, this asks whether or not the two teams have scored the same number of runs after nine innings of play. To compute this, we make the assumption that the scoring of the two teams is independent of one another, and thus that T, the probability that the game is tied after 9 innings, can be computed as

$$T = \sum_{i=0}^{\infty} F_A(i) F_B(i)$$

where $F_A(i)$ and $F_B(i)$ are the probabilities that Team A and Team B score i runs in nine innings, the formula for which was given above.

We note that the formula above tells us that if we assume both teams score the major league average of 0.487 runs per inning, then $T = 0.103$, so that we would expect just over 10% of games to go into extra innings. In reality, 9.22%—18,440 of the 199,906 major league games played between 1871 and 2005—have gone into extra innings. The discrepancy between this number and what our model predicts likely arises from two facts. First, our model assumes that the teams are scoring independently of one another. In reality, this assumption is likely to be not quite true, as external factors (humidity, altitude, pitching, etc.) may cause games to be either high or low scoring, and there may be a psychological factor that promotes teams to score more if the other team is a few runs ahead, or to stop trying once they are blowing out the other team.

The other factor that we can think of is trickier to get a handle on. The above calculation assumes that both teams are average, but in most games one team will be better than the other. For an extreme example, we look at the AL East in 2003, where the Detroit Tigers scored an average of 0.405 runs per inning and the Boston Red Sox scored an average of 0.659 runs per inning. This is the largest discrepancy between two teams in the same league in over 25 years. In this case, the formula predicts that only 8.4% of games will go into extra innings. While this specific example is an extreme, it suggests that when two teams have differing abilities to score runs, we should expect fewer extra inning games even if the overall average number of runs scored is held constant. This expectation is confirmed by the data in TABLE 5, where the rows and columns represent the strengths of the two teams playing, and T is the probability that they will be tied after nine innings, according to our model.

Given that a large number of games are played between teams with widely differing abilities to score runs, this would suggest that our model will predict a larger number of extra inning games than actually occur.

After the ninth inning, the game will conclude at the end of the first inning after which the score is not tied. Therefore, if we let k be the probability that the two teams score the same number of runs in a given inning, then the probability that a game is still tied after n innings is Tk^{n-9} and for $n > 9$ the probability that it ends after n innings is $Tk^{n-10}(1-k)$.

We note that we are making several assumptions here. First, we are assuming that there is no effective difference between the tenth inning and any later inning as far as offensive production is concerned. We also assume that, at least as far as extra innings go, if k is the probability that the two teams score the same number of runs in a given inning then the probability that they score the same number of runs in each of

TABLE 5: Probability of a tie game between two teams of various strengths

	0.405	0.437	0.487	0.537	0.617	0.659
0.405	0.1148	0.1119	0.1065	0.1006	0.0903	0.0848
0.437	0.1119	0.1097	0.1056	0.1007	0.0918	0.0869
0.487	0.1065	0.1056	0.1033	0.1000	0.0932	0.0892
0.537	0.1006	0.1007	0.1000	0.0982	0.0936	0.0905
0.617	0.0903	0.0918	0.0932	0.0936	0.0921	0.0904
0.659	0.0848	0.0869	0.0892	0.0905	0.0904	0.0896

n consecutive innings is k^n. We note that our intuition suggests that due to different strategies in the late parts of the game, as well as fatigue amongst the players, that the scoring distribution might be different as a game progresses, but the data seems to suggest that this difference is negligible. For details, see [**14**].

In order to proceed, it will now suffice to figure out what value k should have. Our first attempt to do so was to use an empirical number coming from the data itself, as detailed in [**11**]. In this paper, we will use the quasigeometric model of scoring which we have developed in order to construct a theoretical value of k. In particular, if we let $a = f_A(0)$ and $b = f_B(0)$ be the respective probabilities of each team going scoreless in an inning, we can compute:

$$
\begin{aligned}
k &= \sum_{i=0}^{\infty} f_A(i) f_B(i) \\
&= ab + \sum_{i=1}^{\infty} f_A(i) f_B(i) \\
&= ab + \sum_{i=1}^{\infty} (1-a)(1-d_A)d_A^{i-1}(1-b)(1-d_B)d_B^{i-1} \\
&= ab + \frac{(1-a)(1-b)(1-d_A)(1-d_B)}{d_A d_B (1 - d_A d_B)}.
\end{aligned}
$$

If we continue with our assumption that $d_A = d_B = 0.436$, and we let m_A (resp. m_B) be the average number of runs per inning scored by team A (resp. team B), then this simplifies to give us

$$k = 1 - 0.564 m_A - 0.564 m_B + 0.4423 m_A m_B.$$

We are now ready to see the fruits of our labor. Let us first look at the case where both of our teams score the major league average number of runs, which means $m_A = m_B = 0.487$. Then it follows that $T = 0.103$ and that $k = 0.55588$. In particular, the probability of a game lasting n innings is $(0.103)(0.4442)(0.5558)^{n-10}$ for all $n \geq 10$. The chart below calculates this probability for games of varying lengths. We have also included the actual number of major league ballgames from 1871 through 2005 that have lasted that long, as well as the number of games that our model predicts.

Comparing the model to the past ... and to the future. So how "rare" are extremely long marathon baseball games? The second author has built a database, discussed in detail in [**11**], of baseball games lasting 20 innings or more. Among these are included the Brooklyn at Boston 26-inning major league record game in 1920, the Rochester at Pawtucket 33-inning minor league game in 1981, and the longest known ballgame: a 45-inning amateur game in Mito, Japan in 1983. Our theoretical model predicts the 26-inning major league record game is not as rare as empirical data would indicate, but the 33-inning minor league record game and 45-inning amateur record game are significantly more rare than empirical data would indicate.

In the previous section we saw that there is approximately a 0.00029 probability that any given game lasts 20 or more innings. Assuming that the probability of any two games lasting this long is independent of one another we can compute that the probability that out of any collection of x games at least one of them lasts 20 or more innings is $1 - (1 - 0.00029)^x$. There are 2340 major league games played each year and therefore we should expect a 50% chance to experience a major league game of 20 or more innings in any given season. Similarly, our model predicts that there will

TABLE 6: Number of games of a given length predicted vs. actual number

# Innings	Prob in given game	Actual MLB	Expected MLB
≤ 9	0.8973	181,466	179,349.6
10	0.04574	8106	9142.3
11	0.02542	4561	5080.9
12	0.01413	2549	2824.3
13	0.007857	1413	1570.4
14	0.004367	831	872.8
15	0.002427	426	485.1
16	0.001349	259	269.6
17	0.0007502	140	149.9
18	0.0004170	69	83.3
19	0.0002318	40	45.9
20	0.0001288	20	25.1
21	7.163E-05	10	13.9
22	3.982E-05	8	7.8
23	2.213E-05	2	4.3
24	1.230E-05	3	2.4
25	6.839E-06	2	1.3
26	3.802E-06	1	0.74
27	2.113E-06	0	0.41
28	1.174E-06	0	0.23
29	6.530E-07	0	0.13
30	3.630E-07	0	0.071
Total	1.0	199,906	199,906

be 0.939 major league games that would have lasted 27 or more innings by now. In fact, we have not yet had such a game in 135 years of major league play. These results indicate that the 26-inning game in Boston is not an outlier from what one would expect from our model.

If we assume that the scoring patterns in minor league games are similar to those in major league games (an assumption for which there is some evidence), and in particular that scoring is quasigeometric with the same values of a and d, then we should expect 6.68 minor league games to have gone 27 or more innings. In fact, we have had 6 such games. If we look further we see that the model predicts that we will have had only 0.087 minor league games which lasted 33 innings. In fact, we have had one such game.

Furthermore, there is a 99.3% chance we will have a minor league marathon of 20 or more innings in any given season, a 0.13% chance we will have a minor league game of 34 or more innings in any given season, a 1.32% chance of seeing a minor league game of 34 innings or more in any given decade, and a 9.4% chance of seeing a minor league game of 34 innings or more in a lifetime of 75 years.

Our model allows us to estimate the probability of games lasting a certain number of innings or longer. This is an alternative method, and perhaps a more easily understood way to express how unlikely are marathons of a certain length. We will now use this approach to compare relative probabilities of breaking the current records for major league and minor league games.

Assuming that major league baseball continues to have 30 teams play a 162-game season, there is a 50% chance we will see a major league game go 27 innings or more in the next 60 years. There is a 95% chance we will see a major league game go 27 innings or more in the next 260 seasons. So the 85-year old 26-inning major league record, while rare, is not so rare that we should assume it will stand for another ninety seasons.

As far as minor league games go, if we assume that there continue to be 13,714 minor league games played per year, then there is a 50% chance we will see a minor league game go 34 innings or longer in the next 565 years. There is a 95% chance we will see a minor league game go 34 innings or more in the next 2,445 years. So the 24-year old 33-inning minor league record may be very rare, and although it could be broken at any time, we should not expect to see it broken anytime soon.

It is interesting to note that, despite several assumptions that seem like they are not entirely accurate, this model does a good job of predicting the number of marathon games. This gives us hope that the quasigeometric model of baseball scoring can be used to answer a variety of questions about the game of baseball, and that it will be a useful tool in the growing research in Sabermetrics.

REFERENCES

1. James Albert and Jay Bennett, *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game,* Copernicus, New York, 2001.
2. Baseball Prospectus, `http://www.baseballprospectus.com`
3. Charles Blahous, The DiMaggio Streak: How Big a Deal Was It? *Baseball Research Journal* (1994) 41–43.
4. Bob Brown and Peter Goodrich, Calculating the Odds: DiMaggio's 56-Game Hitting Streak, *Baseball Research Journal* (2003) 35–40.
5. *By The Numbers*, Statistical Analysis Committee of SABR, issues available for download at `http://www.philbirnbaum.com/`
6. *Mind Game: How the Boston Red Sox Got Smart, Won a World Series, and Created a New Blueprint for Winning.* Steven Goldman, ed., Workman Publishing, New York, 2005.
7. Bill James, *1977 Baseball Abstract,* self-published, 1977.
8. Bill James, *Win Shares,* STATS Publishing, New York, 2002.
9. John F. Jarvis, A Collection of Team Season Statistics, `http://www.knology.net/~johnfjarvis/stats.html`
10. Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*, W. W. Norton, New York, 2003.
11. Philip J. Lowry, I Don't Care If I Ever Come Back: Marathons Lasting 20 or More Innings, *Baseball Research Journal* (2004) 8–28.
12. Society for American Baseball Research, `http://www.sabr.org`
13. Alan Schwarz, *The Numbers Game: Baseball's Lifelong Fascination with Statistics*, Thomas Dunne Books, New York, 2004.
14. David W.Smith, Coming from Behind: Patterns of Scoring and Relation to Winning, presentation at SABR Denver Convention, 2003.
15. TangoTiger. Tango Distribution. Tango On Baseball website, `http://www.tangotiger.net`
16. Keith Woolner, An analytic model for per-inning scoring distributions, Baseball Prospectus, March 4, 2000. `http://www.baseballprospectus.com/article.php?articleid=472`

NOTES

A Closer Look at the Crease Length Problem

SEAN ELLERMEYER
Kennesaw State University
Kennesaw, GA 30144-5591
sellerme@kennesaw.edu

An optimization problem that appears as an exercise in most modern calculus textbooks (Larson [**3**] and Stewart [**5**] for example) is the "crease length problem":

> One corner of a rectangular piece of paper with dimensions $a \times b$ (where a and b are given with $0 < a \leq b$) is folded to a point on the long side of the paper (the side of length b) and the fold is then flattened to form a crease. What is the minimum possible length of such a crease and to what point on the long side of the paper must the corner be folded in order to achieve this minimum?

Although not always stated as such, the problem that the textbooks authors actually intend for their readers to solve is a more restricted version of the problem stated above. Upon consulting the solutions manuals that accompany many of the textbooks, we find that the solutions to the crease length problem that are provided only take into account those paper foldings that do not produce a flap that protrudes over one of the edges of the paper. However, as we can convince ourselves by grabbing a piece of paper and doing some folding experiments, some of the possible folds (as described in the problem above) do produce protruding flaps. Specifically, referring to FIGURE 1, we can perform a Case 1 fold which produces a flap that protrudes over the short edge of the paper, a Case 2 fold which has no protrusion, or a Case 3 fold which produces a flap that protrudes over the long edge of the paper. In addition, there are two "critical" folds, illustrated in FIGURE 2, that separate Case 1 from Case 2 and Case 2 from Case 3, and there are also two other critical folds (not illustrated)—folding the lower

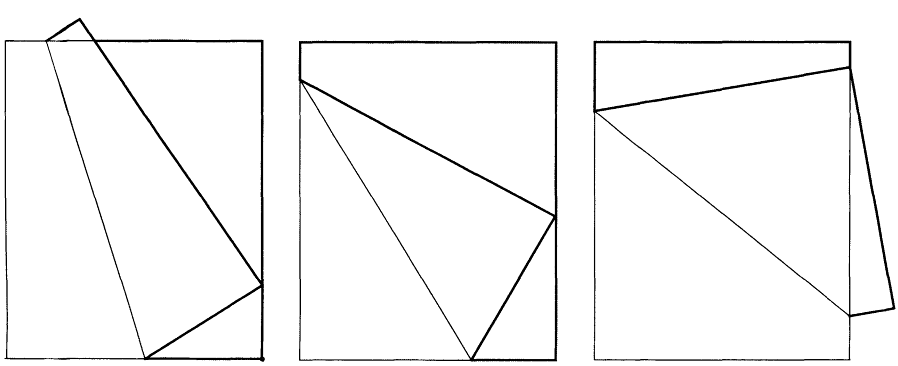


Figure 1 Three possible folds

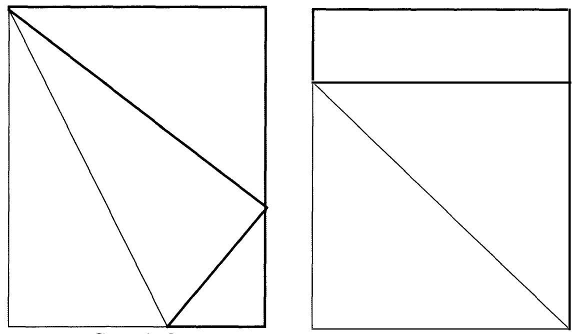


Figure 2 Critical cases

left corner onto the lower right corner (thus folding the paper in half) and folding the lower left corner onto the upper right corner (which can be viewed as an extreme case of Case 3). In [**1**], Haga considers a much wider array of possibilities for making a single fold of a rectangular sheet of paper but the focus is on studying the areas and ratios of side lengths of the polygons that are formed by such folds rather than on determining optimal crease lengths.

In this note, we provide a solution of the general crease length problem in which all possible foldings of a corner to the opposite edge (as described above) are taken into account. One of our findings will be that the minimum crease length is never produced by a Case 2 fold (no matter the dimensions of the paper) and hence that the general crease length problem always yields a different minimum than the constrained problem that is treated in the textbooks. Our more interesting discovery, however, will be a criterion that determines which foldings must be performed in order to achieve the minimum (and maximum) crease lengths. This criterion, which does not manifest itself when only the constrained problem is considered, is a condition relating the paper dimensions to the *Golden Ratio*, which is the number $\phi = \left(1 + \sqrt{5}\right)/2$. This number is one of the "special" constants of mathematics (like π and e) that seems to show up frequently, often when least expected, in investigations of many different phenomena (geometric and otherwise). For those who would like to become better acquainted with the Golden Ratio, we recommend the book of Huntley [**2**] and the article of Markowsky [**4**].

Supposing our paper to have dimensions $a \times b$ where $0 < a \leq b$, we can view the crease length as a function of y, where y is the distance from the lower right corner of the paper to the point on the right edge of the paper to which the lower left corner has been folded. (Refer to FIGURE 1.) Our goal is to determine the absolute minimum and maximum values of the crease length and the values of y at which these extrema occur as y ranges from 0 to b. In order to construct the function that gives the length of this crease, we will find it convenient to first consider a slightly different folding problem in which the paper to be folded has no top or right boundaries.

The crease function for infinite paper

If we begin with an infinite piece of paper or "infinite open rectangle" $R = (0, \infty) \times (0, \infty)$ and fold the point $(0, 0)$ (which is the lower left corner of the rectangle) onto an arbitrary point $(x, y) \in R$, then the fold cannot protrude over any of the boundaries of

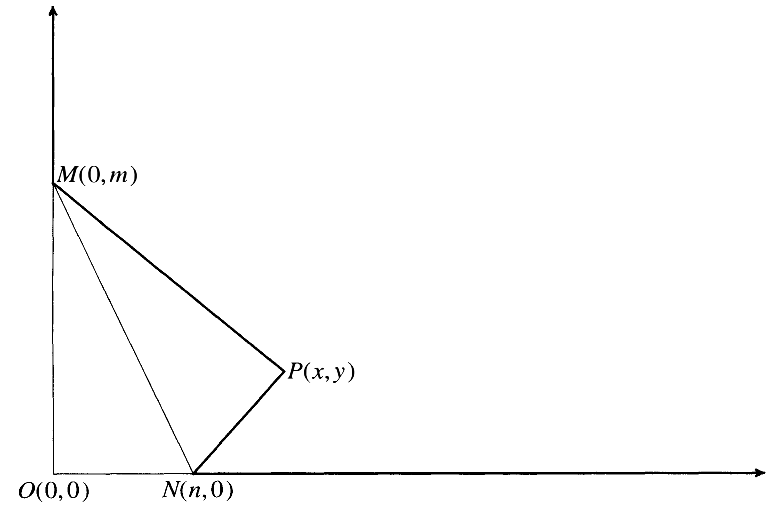


Figure 3 Folding infinite paper

the original rectangle and we obtain a situation as depicted in FIGURE 3. By referring to this figure, the crease length $|MN|$ can be determined as a function of x and y.

Since $|NP| = |ON|$, we see that $(x-n)^2 + y^2 = n^2$ and hence that

$$n = \frac{x^2+y^2}{2x}. \tag{1}$$

Likewise, the equality of $|MP|$ and $|OM|$ implies that

$$m = \frac{x^2+y^2}{2y}. \tag{2}$$

Then, by the Pythagorean Theorem, we find that the square of the crease length (for convenience we will always square the crease length) is given by the function

$$F(x,y) = |MN|^2 = n^2 + m^2 = \frac{(x^2+y^2)^2}{4x^2} + \frac{(x^2+y^2)^2}{4y^2} = \frac{(x^2+y^2)^3}{4x^2y^2}. \tag{3}$$

Before moving on to examine creases of finite paper, we pause to make an observation that might be of interest to students and teachers of multivariable calculus. A major topic in multivariable calculus is the study of indeterminate form limits of functions of two variables. Many of the examples and exercises through which students learn about this topic involve rational functions (ratios of polynomials in x and y). However, examples of these types of limit problems for which physical or geometric intuition can be brought to bear seem to be rare. The function F defined in (3) does provide such an example though. Specifically, let us consider the problem of evaluating $\lim_{(x,y)\to(0,0)} F(x,y)$.

The standard method of evaluating this limit is to let (x, y) approach $(0, 0)$ along various curves (such as $y = x$, $y = x^2$, $y = x^3$) and to observe that each curve of approach yields a different limit $(0, 1/4, \infty)$, thus allowing us to conclude that $\lim_{(x,y)\to(0,0)} F(x,y)$ does not exist. However, our intuitive understanding of why this limit does not exist is greatly enhanced by referring to FIGURE 3 in which $F(x, y)$ is the square of the length of the crease MN. In particular, it is easy to visualize that if we let the point $P(x, y)$ approach $O(0,0)$ along the line $y = x$, then the crease length approaches 0. On the other hand, we observe that within any prescribed distance of O we can find points $P(x, y)$ that yield arbitrarily large crease lengths (obtained by choosing $P(x, y)$ close enough to the boundary of the paper). This visual reasoning provides us with a geometry–based understanding of why the limit in question does not exist and also suggests that polar coordinates should be useful for the purpose of obtaining a more detailed mathematical description of the behavior of F. Indeed, by letting $x = r\cos(\theta)$ and $y = r\sin(\theta)$, we see that the level curve $F(x, y) = K^2$ (corresponding to a given crease length $K > 0$) is $r = K\sin(2\theta)$, $0 < \theta < \pi/2$. The fact that each of these level curves (for each $K > 0$) lies in the open first quadrant and intersects every neighborhood of (0, 0) shows that $\lim_{(x,y)\to(0,0)} F(x, y)$ does not exist and, furthermore, shows that F assumes every positive value in every neighborhood of (0, 0).

The crease function for finite paper

In order to derive the crease function for a finite piece of paper of dimensions $a \times b$ where $0 < a \leq b$, we position the lower left corner of our paper at the point (0, 0) and the lower right corner at the point $(a, 0)$. The function that we want to derive is

$$L(y) = \text{square of the crease length when } O(0,0) \text{ is folded onto } P(a, y)$$

with domain $0 \leq y \leq b$.

Since the finite rectangle $(0, a) \times (0, b)$ is a subset of the infinite rectangle $(0, \infty) \times (0, \infty)$, we will be able to make use of the infinite paper crease function, F, in deriving L. In fact, the work has already been done for a Case 2 fold (see FIGURE 1) since this type of fold produces a crease that has the same length as the crease that would be formed in folding an infinite piece of paper. Thus, for a Case 2 fold, we have by (3) that

$$L(y) = F(a, y) = \frac{(a^2 + y^2)^3}{4a^2y^2}.$$

The derivation of L for Cases 1 and 3 requires a little additional geometry. In these cases, we regard the finite paper as being superimposed on the infinite paper. Although only the finite paper is to be folded, we extend the lines formed by the fold onto the infinite paper as shown in FIGURES 4 and 5. It then follows from (1) and (2) that

$$n = \frac{a^2 + y^2}{2a} \quad \text{and} \quad m = \frac{a^2 + y^2}{2y}.$$

For Case 1 (FIGURE 4), the crease length is $|RN|$. The triangles MBR and MON are similar, so we obtain

$$\frac{|MR| + |RN|}{|OM|} = \frac{|MR|}{|BM|}$$

which gives us

$$|RN| = \frac{|MR|}{|BM|}|OB| = \frac{|MN|}{|OM|}|OB|$$

and thus

$$L(y) = |RN|^2 = \frac{m^2 + n^2}{m^2} b^2 = \left(1 + \frac{n^2}{m^2}\right) b^2 = \left(1 + \frac{y^2}{a^2}\right) b^2 = \frac{b^2}{a^2}(a^2 + y^2).$$

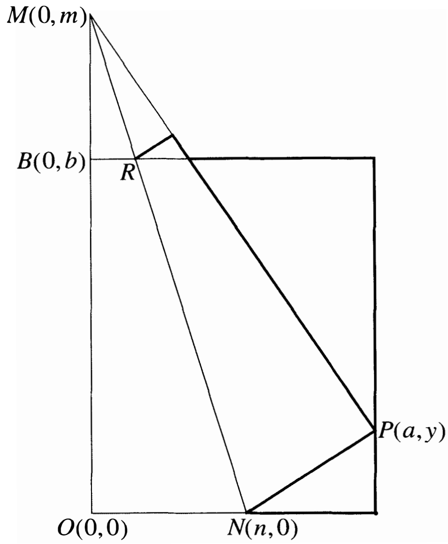


Figure 4 Case 1 fold

For Case 3 (FIGURE 5), the crease length is $|MR|$ and since

$$\frac{|MR| + |RN|}{|ON|} = \frac{|RN|}{|AN|}$$

which implies that

$$|MR| = \frac{|RN|}{|AN|}|OA| = \frac{|MN|}{|ON|}|OA|,$$

we obtain

$$L(y) = |MR|^2 = \frac{m^2 + n^2}{n^2}a^2 = \left(\frac{m^2}{n^2} + 1\right)a^2 = \left(\frac{a^2}{y^2} + 1\right)a^2 = \frac{a^2}{y^2}(a^2 + y^2).$$

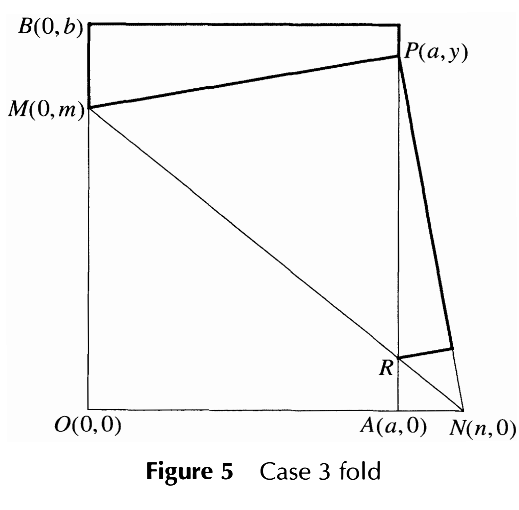


Figure 5 Case 3 fold

By comparing FIGURES 2, 4, and 5, we observe that the condition that corresponds to the critical Case 1–2 is $m = b$ which is equivalent to $y = b - \sqrt{b^2 - a^2}$ and that the condition that corresponds to the critical Case 2–3 is $n = a$ which is equivalent to $y = a$. The crease function for an $a \times b$ piece of paper is thus the piecewise–defined function

$$L(y) = \begin{cases} \dfrac{b^2}{a^2}(a^2 + y^2) & \text{if } 0 \le y \le b - \sqrt{b^2 - a^2} \\ \dfrac{(a^2 + y^2)^3}{4a^2y^2} & \text{if } b - \sqrt{b^2 - a^2} < y < a \\ \dfrac{a^2}{y^2}(a^2 + y^2) & \text{if } a \le y \le b \end{cases}.$$

It can readily be seen that L is increasing on the interval $\left(0, b - \sqrt{b^2 - a^2}\right)$ and decreasing on the interval (a, b). On the middle interval, $\left(b - \sqrt{b^2 - a^2}, a\right)$, since

$$L'(y) = \frac{(a^2 + y^2)^2}{a^2y^3}\left(y + \frac{\sqrt{2}}{2}a\right)\left(y - \frac{\sqrt{2}}{2}a\right),$$

we observe that $y = \sqrt{2}a/2$ is a critical point of L that corresponds to a local minimum value of L if and only if $\sqrt{2}a/2 \in \left(b - \sqrt{b^2 - a^2}, a\right)$. While the relation $\sqrt{2}a/2 < a$ is certainly always true, the relation $b - \sqrt{b^2 - a^2} < \sqrt{2}a/2$ is true (as the reader can check) if and only if $b^2/a^2 > 9/8$. In all of the textbook exercises that we have seen, the paper dimensions are given to be such that $b^2/a^2 > 9/8$ and, since only Case 2 folds are addressed in these exercises, the local minimum value $L\left(\sqrt{2}a/2\right)$ is regarded as the absolute minimum value and hence as the "right answer" for the minimum crease length. However, in what follows we will see that this is in fact never the absolute minimum in the more general problem (no matter the values of a and b).

The golden ratio makes the call

In order to economize on notation, we introduce a new parameter $q = b/a$ (with the assumption that $0 < a \le b$ implying that $q \ge 1$) and we also give names to the critical points of L: $y_0 = 0$, $y_1 = b - \sqrt{b^2 - a^2}$, $y_2 = \sqrt{2}a/2$, $y_3 = a$, and $y_4 = b$. Even though y_2 is not a critical point unless $q^2 > 9/8$, it will not be necessary to treat this case separately in what follows.

We have determined that the candidates for the absolute minimum value of L are

$$L(y_0) = b^2 = q^2a^2$$

$$L(y_2) = \frac{27}{16}a^2$$

$$L(y_4) = \frac{a^2}{b^2}(a^2 + b^2) = \left(\frac{1}{q^2} + 1\right)a^2$$

and that the candidates for the absolute maximum value of L are

$$L(y_1) = \frac{2b^3y_1}{a^2} = 2q^3\left(q - \sqrt{q^2 - 1}\right)a^2$$

$$L(y_3) = 2a^2.$$

To see where the absolute extrema actually occur, we need to compare the values $L(y_0)$, $L(y_2)$ and $L(y_4)$ and also compare the values $L(y_1)$ and $L(y_3)$. It is in doing this that we will see the Golden Ratio,

$$\phi = \frac{1+\sqrt{5}}{2} \approx 1.618,$$

make its appearance. The property of the Golden Ratio that will be used in our comparisons is the property that it is the only positive real number that is exactly one greater than its reciprocal; that is, $\phi^{-1} + 1 = \phi$, or equivalently $\phi^2 - \phi - 1 = 0$. The other fact that will be used is the fact that $\phi < 27/16$. Our results are given in the following:

PROPOSITION 1. *Let $q = b/a$ and let ϕ denote the Golden Ratio.*

1. *If $q^2 < \phi$, then the minimum possible crease length is achieved by folding the paper in half and the maximum is achieved by performing a Case 2–3 fold (*FIGURE *2).*
2. *If $q^2 > \phi$, then the minimum possible crease length is achieved by folding the lower left corner of the paper to the upper right corner of the paper and the maximum is achieved by performing a Case 1–2 fold.*
3. *The minimum possible crease length can be achieved with two distinct foldings if and only if $q^2 = \phi$. (The same is true of the maximum possible crease length.)*

Proof. First we compare the values $L(y_0)$, $L(y_2)$, and $L(y_4)$: If $q^2 < \phi$, then

$$\frac{1}{a^2}(L(y_2) - L(y_0)) = \frac{27}{16} - q^2 > \phi - q^2 > 0$$

and

$$\frac{1}{a^2}(L(y_4) - L(y_0)) = \frac{1}{q^2} + 1 - q^2 > \frac{1}{\phi} + 1 - \phi = 0;$$

whereas if $q^2 > \phi$, then

$$\frac{1}{a^2}(L(y_2) - L(y_4)) = \frac{27}{16} - \frac{1}{q^2} - 1 > \frac{27}{16} - \frac{1}{\phi} - 1 = \frac{27}{16} - \phi > 0$$

and

$$\frac{1}{a^2}(L(y_0) - L(y_4)) = q^2 - \frac{1}{q^2} - 1 > \phi - \frac{1}{\phi} - 1 = 0.$$

The above comparisons show that L achieves its absolute minimum value at y_0 if $q^2 < \phi$ and at y_4 if $q^2 > \phi$. If $q^2 = \phi$, then L achieves its minimum at both y_0 and y_4. In no case is the minimum achieved at y_2.

We now compare the values $L(y_1)$ and $L(y_3)$. Since the quantity

$$\frac{1}{2a^2}(L(y_1) - L(y_3)) = q^4 - q^3\sqrt{q^2-1} - 1$$

is positive if and only if

$$q^3\sqrt{q^2-1} < q^4 - 1$$

which is true if and only if

$$q^6 - 2q^4 + 1 > 0,$$

and since

$$q^6 - 2q^4 + 1 = (q^2 - 1)\left(q^2 + \frac{1}{\phi}\right)(q^2 - \phi),$$

we conclude that $L(y_1) > L(y_3)$ if and only if $q^2 > \phi$. Therefore L achieves its absolute maximum value at y_1 if $q^2 > \phi$ and at y_3 if $q^2 < \phi$. If $q^2 = \phi$, then L achieves its maximum at both y_1 and y_3. ■

In order to illustrate the somewhat surprising nature of our results, let us compare the constructions of extreme crease lengths for 8.5 × 11 paper and 8.7 × 11 paper. Since these paper dimensions are not very different (probably not visible to the naked eye), intuition would lead us to believe that the extrema would be obtained by performing similar folds. However, for $a = 8.5$ and $b = 11$ we have $q^2 \approx 1.675 > \phi$ meaning that the minimum crease length is obtained by folding the lower left corner onto the upper right corner and the maximum crease length is obtained by performing a Case 1–2 fold (FIGURE 2); whereas for $a = 8.7$ and $b = 11$ we have $q^2 \approx 1.599 < \phi$ meaning that the minimum crease length is obtained by folding the paper in half and the maximum crease length is obtained by performing a Case 2–3 fold. A comparison of the crease functions, giving actual crease lengths $\sqrt{L(y)}$, for 8.5 × 11 paper and 8.7 × 11 paper is shown in FIGURE 6.

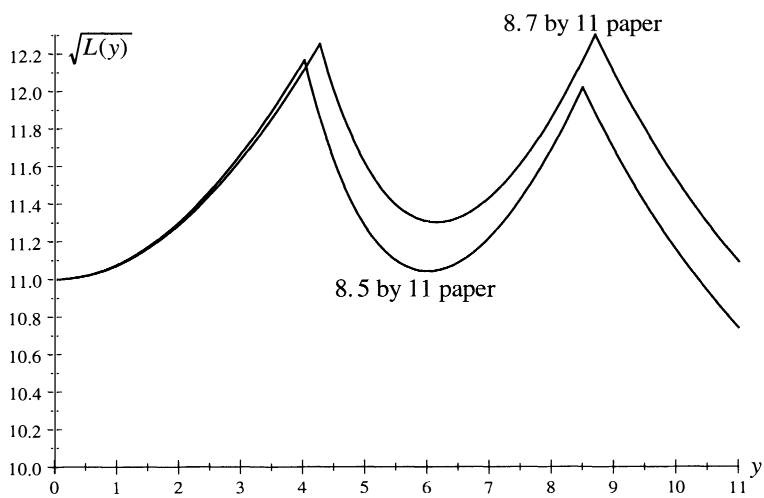


Figure 6 Comparison of crease length functions

REFERENCES

1. Kazuo Haga, Fold paper and enjoy math: origamics, in *Origami³: Third International Meeting of Origami Science, Math, and Education*, T. Hull (ed.), A.K. Peters, Natick, MA, 2002, 307–328.
2. H. E. Huntley, *The Divine Proportion (A Study in Mathematical Beauty)*, Dover, New York, 1970.
3. Ron Larson, Robert P. Hostetler, and Bruce H. Edwards, *Calculus*, 8th ed., Houghton Mifflin, Boston, 2006.
4. George Markowsky, Misconceptions about the golden ratio, *College Math. J.* **23**(1) (1991) 2–19.
5. James Stewart, *Calculus Concepts and Contexts*, 3rd ed., Thomson Brooks/Cole, Belmont, CA, 2005.

An Offer You Can't Refuse

R. HIRSHON
College of Staten Island
Staten Island, NY 10314
rhirshon@aol.com

R. DE SIMONE
394 Fanning Street
Staten Island, NY 10314
rdesimone@si.rr.com

The offer

My friend Bob, who is a gambling man, says to me: "Come on Ron, let's play a game. You can make some easy money. Here are the rules: We will flip this unfair coin that I have that has a probability of falling heads which is .52, just a tiny bit more than $\frac{1}{2}$. Each time it falls heads, you will pay me \$1. If it falls tails, I will pay you \$1. Whoever runs out of money first will pay the other guy \$500. But since I have a slightly better chance of winning on each flip, I have to do something extra for you. Here's what I'm going to do for you. I will start with a meager bankroll of \$10. But Ron, because you are my dear friend, I'm going to let you start with any bankroll you desire. You can start with a thousand dollars, or a million dollars or a billion dollars or if you wish you can start with $T = \$1{,}000{,}000{,}000^{1{,}000{,}000{,}000}$. That's right, that's what I said, Ron, you can start with one billion raised to the one billionth power dollars. That's a big bankroll! I'm not sure how long it would take to write that number out, or how many miles it would stretch if we did. But I think a lot. And Ron, not to confuse the matter, let us say that our bankroll is really composed of worthless chips, so that you really don't have to risk T at all (as if anyone had that amount of money) and the only money that is at stake for either one of us is the original \$500 wager. And since your bankroll may be very large, we will need a really good computer to help us play the game. Fortunately, I have just such a computer that can handle these large bankrolls, and will simulate and play our game in a minute or two." (O.K for the fun and intrigue of the problem, let's make believe such a computer is just at our finger tips!)

Well, Bob has tempted me. I know he is basically honest (not too dishonest), and I trust that he has described the rules accurately. But he is not above "conning" me in an honest way. O.K, to play it safe, since Bob told me that I can start with whatever I want, I am not going to start with T. I am going to play it a lot safer and start with T^T. That should be enough, I believe, to overpower Bob's meager 10 chip bankroll. For even though Bob has a slight edge of winning on each individual flip of the coin, it is only necessary for me to win 10 more flips than Bob does to take the \$500 prize. Would I be wrong to play Bob's game? Before answering Ron's question, reading the next two paragraphs may be in order.

Some historical remarks

The game proposed in the preceding section and the problems engendered by it go by the name of Gambler's Ruin. Let players X and Y play the flipping game, starting with bankrolls of x, y, respectively and let p be the probability that X wins on a single flip and $q = 1 - p$ be the probability that X loses on a single flip. Let $L = L(x, y, p)$ be the event that X goes broke before Y. Then we have:

$$P(L) = \begin{cases} \dfrac{q^x(p^y - q^y)}{p^{x+y} - q^{x+y}} & \text{if } p \neq q \\ \dfrac{y}{x+y} & \text{if } p = q = \frac{1}{2} \end{cases} \tag{1}$$

The above result in slightly different notation can be found in [**6**, p. 92]. (We will find the above notation useful in the sequel.) The proof is a short argument using only elementary probability theory. Other proofs of (1) have used more sophisticated mathematics [**1**], [**4**]. There are interesting articles in [**2**], [**3**], and [**5**] explaining various generalizations of the problem and giving numerous references. The general form of the problem was solved by the mathematician James Bernoulli and published eight years after his death in 1713 [**6**]. If one makes a quick search on the Internet and plugs Gambler's Ruin into a search, a wealth of material comes up including simulations for this problem and problems of this type. Evidently the problem has continued to fascinate for nearly 300 years.

The only mathematical prerequisites for reading this paper are the first six weeks of an undergraduate course in probability and a solid understanding of the limit process discussed in a good undergraduate calculus course.

The answer to Ron's question

Ron should not play the game as described in the introduction. It might be surprising, but no matter how large Ron's bankroll, Ron's probability of losing (going broke before Bob does) is more than .55. And a "law of average" type of argument won't explain a possible paradox. For example, if $p = .52$ were replaced by $p = .5173$ when $x = 10$, then Bob might be in trouble. (See The Turning Probability Theorem in the sequel.) In any case, if Ron played the game a thousand times, with $p = .52$, (don't forget, we have a very speedy computer to help us) he could expect to win 450 times and to lose about 550 times so that he could expect to lose approximately 100(\$500) = \$50,000 or more! Things can even get worse for Ron. Here are some facts:

LEMMA. *If Bob starts with a bankroll of \$$x$ and has probability p of winning on an individual flip of the coin and if $q = 1 - p$, then no matter what Ron's initial bankroll is, Bob's probability of winning a game (that is of Ron going broke, before Bob does) is always greater than $Z = 1 - (q/p)^x$.*

If $p > q$, and Ron's initial bankroll y is kept constant, and p is kept constant, it is obvious that Bob's chance of winning the match increases as $x \to \infty$. But the Lemma above shows even more. Bob's chance of winning approaches 1 uniformly, independently of y as x increases without bound. If $p > q$, the situation where x and p are held constant, and either y is unknown or is allowed to vary and increase without bound will be some of the more interesting cases to study.

We will show how to prove the lemma in a small paragraph in the sequel. But first let us calculate a few values of Z for various x and p to get a feel for the above result. We tabulate the results below. The middle line of the table below is for $x = 10$ and the bottom line is for $x = 20$.

	p	.51	.5175	.52	.53	.54	.55	.56	.57	.58	.59
$x = 10$	Z	.3297	.5036	.5509	.699	.799	.866	.910	.940	.960	.974
$x = 20$	Z	.5507	.7535	.798	.910	.960	.982	.992	.996	.998	.999

We leave it to the reader to decide if Bob is a "con artist." The one good thing we can say about him is that at least he wasn't too greedy and he did not propose to start with a bankroll of 20. But he may have been smart enough to know that a smaller bankroll for him might have (to use the vernacular) sucked Ron right into his trap. He would have even trapped Ron with $x = 10$ and $p = .5175$, and perhaps these choices would have enticed Ron even more. So a new question that Ron might ask is: If x and p are kept constant, can I know in advance, whether "very deep" pockets for me can trump a tiny edge for Bob? This time, the answer is yes!

The other side of the story: The turning points of the game

In the sequel, we replace Bob and Ron with X and Y respectively, with starting bankrolls of x and y and with p being the probability of X winning on a single flip. The following result shows that in certain cases, it pays for player Y to be rich. Recall that we are concerned with the case that X has an edge so that $p > 1/2$, x and p are held constant, and there are no restrictions on y.

When big bucks really count, the turning probability theorem

Given any positive integer x, there exists a corresponding probability $P = P(x)$ such that $1/2 < P < 1$ and such that X's fate in the game can change depending on whether his probability p of winning on a single flip during the game is less than P or bigger than P. Namely, if $1/2 < p < P$, then Y's probability of winning the match will be greater than $1/2$ provided that Y starts with a sufficiently large bankroll. If $P \leq p < 1$, then X's probability of winning the match will always be greater than $1/2$ no matter how much Y starts with. The value of $P(x)$ is obtained easily by solving the equation $1 - Z = 1/2$, or $Z = 1/2$ for P. In particular,

$$P(x) = \frac{1}{(1/2)^{1/x} + 1}, \quad \left(\frac{Q}{P}\right)^x = \frac{1}{2}, \quad Q = 1 - P. \tag{2}$$

We will prove the above theorem in the sequel, but it might be of interest to calculate approximately here a few values of P, which we do in the table below. It seems appropriate to call $P(x)$ the turning probability of the game for the bankroll x. We observe that the turning probability is always irrational except for the case $x = 1$. The graph of P decreases from 2/3 to .5 as x increases and the graph becomes asymptotic to $P = .5$ very quickly.

x	1	3	5	10	15	20
$P = P(x)$	2/3	.5575	.5346	.51732	.5116	.5087

The turning probability may be viewed as keeping x constant in the problem and finding an optimal minimal "winning" probability P for player X for that x. For example, consider the following possible cases:

(a) $x = 10$, $p = .517$,

(b) $x = 10$, $p = .5173$,

(c) $x = 10$, $p = .51734$.

Should we expect a significant difference between these cases? We think most of us wouldn't or shouldn't. But there is a huge difference especially as concerns the repeated play of the game with stakes as in our original game between Ron and Bob. The critical P value which distinguishes the cases for $x = 10$ is approximately .51732. In case (c), X's probability of winning the match (Y going broke before X) is always greater than $1/2$ no matter what Y's initial bankroll is. That is "deep pockets" for Y do not help. However in the other two cases, if Y starts with a sufficiently large bankroll, his probability of winning the match will be greater than $1/2$. That is, "deep pockets" works for Y.

The next result takes a slightly different point of view. It answers the following question: If $p > 1/2$ is fixed and is given, and we have no prior knowledge of y, what is an "optimal minimal" bankroll x for player X for this p? We state below the theorem on the turning bankroll. Again, in the sequel, $L = L(x, y, p)$ is the event that player X goes broke before player Y does when they start with respective bankrolls x, y, and p is the probability of X winning on a single flip.

The turning bankroll theorem

Given any fixed $p > 1/2$, there exists a bankroll x^* such that for any bankroll $x \geq x^*$, if X begins with x, then X's probability of winning the match (Y going broke before X) is always greater than $1/2$ no matter what bankroll y, Y starts with. In addition, if X begins with a bankroll $x < x^*$, then for all sufficiently large bankrolls y for Y, X's probability of winning the match is less than $1/2$. We call x^* the turning bankroll for the probability p. We can state the turning bankroll point explicitly. To do this, we first solve the equation $Z = 1/2$ for x and obtain a solution $\ln 2/\ln(p/q)$, $.5 < p < 1$. Since this solution depends on p, we represent it in functional notation by $x = \alpha(p)$. Then $x^* = \lceil \alpha(p) \rceil$ is the smallest positive integer greater than or equal to $\alpha(p)$. Below is a sample calculation of $\alpha(p)$ and x^* for various p.

p	.51	.514	.516	.5172	.5174	.52	.53	.6666	2/3	.6667
$\alpha(p)$	17.33	12.37	10.8	10.07	9.95	8.7	5.8	1.0004	1	.9997
x^*	18	13	11	11	10	9	6	2	1	1

The solutions $\alpha(p)$ above are approximated sufficiently to obtain x^*. We see that x^* will be 1 for $p \geq 2/3$ and x^* will be greater than 1 otherwise. It seems natural to call $p = 2/3$ a turning point probability for the turning bankroll point $x^* = 1$. (We will deal with this comment in greater detail in the next theorem.) For the values of p which yield $x^* = 1$, the part of the turning bankroll theorem which deals with $x < x^*$ is vacuously true since we treat only positive integral bankrolls. We point out again that from the Big Bucks Theorem both

(a) $x = 10$, $p = .517$ and

(b) $x = 10$, $p = .5172$

are losing strategies for X, but from the present result

(a_1) $x = 11$, $p = .517$ and

(b_1) $x = 11$, $p = .5172$

are both winning strategies for X. The function x^* is a step function and the next result describes the step function behavior of x^* explicitly.

The turning probability theorem for turning bankrolls

Let j be a positive integer. Let, p_j be the solution p to the equation $j = \ln 2/\ln(p/q)$. Namely, $p_j = h_j/(h_j + 1)$, where $h_j = \sqrt[j]{2}$. Then $\alpha(p_j) = j$ and the p_j form a monotonically decreasing sequence of numbers less than one. The sequence decreases to $1/2$. For all p, $1/2 < p < 2/3$, if $p_{j+1} \le p < p_j$, the turning bankroll x^* of p is precisely $j + 1$. If $p \ge 2/3$, the turning bankroll for p is 1. To restate this in another way: The turning bankroll is a step function whose discontinuities are the probabilities p_j.

We give below approximations for the first ten p_j. The table values are approximations of the actual values except for p_1 which is precisely 2/3.

j	1	2	3	4	5	6	7	8	9	10
p_j	2/3	.586	.558	.543	.535	.529	.525	.522	.519	.517

For example, to determine the turning bankroll of .53 from the above table, note that $p_6 < .53 < p_5$, so the turning bankroll x^* of .53 is 6. Note that in the turning bankroll theorem, we "plug" in p to obtain x^*. In the turning probability theorem for turning bankrolls, we find the "probability intervals" on which x^* takes a constant value. We will now prove the Turning Theorems.

Proofs of the assertions. We first deal with the Lemma. In this entire discussion, as well in the one immediately below, x is fixed as X's initial bankroll. We use (1) above to verify the validity of the first assertion. We can assume $p > q$ or else the statement is a triviality. (And besides, we are concerned with the cases when X has an edge.) As y gets larger, X's chance of losing strictly increases. So if x stays constant, $P(L)$ will be strictly less than the limit of the right hand side of (1) as $y \to \infty$. If this is not clear, remember that in any particular case, $P(L)$ is going to be equal to one of the terms in (1) for some fixed y, and that the limit of a strictly monotonic increasing sequence of numbers is strictly greater than any of the individual numbers. Now note that if we divide numerator and denominator of (1) by p^{x+y}, then the expression (1) may be rewritten as $(q/p)^x \cdot [1 - (q/p)^y]/[1 - (q/p)^{x+y}]$. Clearly the second fraction approaches 1 as $y \to \infty$ so that (1) increases monotonically to $(q/p)^x$ and $P(L) < (q/p)^x$ so that $1 - P(L) > 1 - (q/p)^x$, which is our assertion.

The big bucks turning probability assertion. This is merely a matter of interpreting what we already have written above. We just have to note that we have actually shown above that as $y \to \infty$, $P(L)$ monotonically increases and approaches $(q/p)^x$. Now if we think of what all this means in the limiting process, we obtain the story of Big Bucks. But we will write a few details below to help clarify the assertion. Also recall that x is fixed in this discussion. We will write $L = L(y, p)$ to indicate the parameters y and p involved. First note that for fixed y as p gets larger, X's chance of losing strictly decreases. Again, let $P = P(x)$. Note that $L(y, p_1) < L(y, p_2)$ if $p_1 > p_2$. Consequently if $p \ge P$, then for any y, by the use of (2), $L(y, p) \le L(y, P) < (Q/P)^x = 1/2$. Thus for $p \ge P$, the probability that X wins the match is greater than 1/2 for all y.

Now let us consider the situation when $p < P$. Then $L(y, P) < L(y, p)$. Now in this very last inequality, let $y \to \infty$. We then conclude that for some large enough value y_1, $L(y_1, p) \geq (Q/P)^x = 1/2$. But then for all $y_2 > y_1$, $1/2 \leq L(y_1, p) < L(y_2, p)$. Rephrased, this says that for all y_2 sufficiently large and all $p < P$, we have X's chance of losing is larger than $1/2$. Thus for these y and p, Y's chance of winning is larger than $1/2$.

The turning bankroll assertion. Again this is a matter of interpreting the limit process used in obtaining Z values. First suppose that $x \geq x^*$. Then $x \geq \alpha(p)$. Hence $1/2 = (q/p)^{\alpha(p)} \geq (q/p)^x$. This is the same as $1 - (q/p)^x \geq 1/2$. But the probability of X winning the match for any bankroll y is always strictly greater than $1 - (q/p)^x$. This gives the first part of the assertion. Now suppose $x < x^* = \lceil \alpha(p) \rceil$. Since x is an integer, this means $x < \alpha(p)$. Consequently, $1/2 = (q/p)^{\alpha(p)} < (q/p)^x$. Since $P(L) \to (q/p)^x$ as $y \to \infty$, we see $P(L) > 1/2$ for all y sufficiently large. Hence for all y sufficiently large, X's probability of winning the match is less than $1/2$.

The turning probability theorem for turning bankrolls. First of all, one can readily verify that the function α is a strictly decreasing function of p. To verify that x^* takes the value $j + 1$ on the interval $[p_{j+1}, p_j)$, suppose $p_{j+1} \leq p < p_j$. Apply α to this inequality to obtain $\alpha(p_{j+1}) \geq \alpha(p) > \alpha(p_j)$ so that $j + 1 \geq \alpha(p) > j$. But then by definition, the value of x^* at p must be $j + 1$.

Last remarks. In this paragraph we describe a combinatorial problem that we encountered while thinking about Gambler's Ruin. Essentially, the problem arose when we attempted to write a power series expression for $P(L)$. The proofs of the assertions below will be submitted in the future in a separate manuscript. Let L_n be the event that the match comes to an end with X losing in exactly n flips. A major problem is to determine the number of ways in which L_n can happen. That first important observation is that if X loses the match (goes broke) in exactly n flips and loses exactly l of the first n flips, and wins exactly w of the first n flips, then $l + w = n$ and $l - w = x$, so $2l - x = n$. Thus x and n have the same parity. In other words, for any (x, y), the events that have the subscripts n are either all even or all odd. The first subscript on an L symbol is x, since it takes at least x flips for X to lose the match. To illustrate, it might be helpful to look at a simple case, $(x, y) = (1, 3)$. In this situation, the game must end with a loss for X with one of the events L_n with n odd. Let's say we agree that X will win a flip if head comes up on the flip and will lose the flip if tails comes up. We calculate by hand the number of subevents in L_7. In considering any subevent in L_7, we must have exactly 4 tails and exactly 3 heads in the subsequence, and end with a T. So right away we have a simple upper bound $\binom{6}{3} = 20$ for the number of elements in L_7. One can verify that there are exactly 4 subevents in L_7, namely, *HTHTHTT*, *HHTTHTT*, *HTHHTTT*, and *HHTHTTT*. For example, we can not place *HTHTTHT* in this list because in this situation X would already have lost the match on the fifth flip. We emphasize that in checking for the validity of a certain sequence as a possible subevent of L_n we must check the sequence to be certain that X has lost the match exactly on the nth flip and not lost the match before the nth flip and that Y has not lost the match before the nth flip. Without a good way to attack this question, the reader may be able to imagine how tedious such calculations can get. The question concerning the size of L_n is a completely combinatorial problem, depending only on x and y. Using (1), we have developed an algorithm such that for any x and y, we find a rational function $g(u)$ such that the coefficients of its Maclaurin series expansion yield the number of elements in L_n. We call this rational function the loss function of X with parameters

x, y. We refer to the coefficients in the Maclaurin series as the loss sequence. The table below gives some sample results. Taking $(x, y) = (5, 1)$, the table shows that L_5, L_7, and L_9 have 1, 4, and 13 elements, respectively. If we take $(x, y) = (4, 2)$, the table shows that L_4, L_6, and L_8 have 1, 4, and 13 elements, respectively.

x	y	Loss Function $g(u)$	Loss Sequence—first ten terms
1	2	$1/(1-u)$	1, 1, 1, 1, 1, 1, 1, 1, 1, 1
1	3	$(1-u)/(1-2u)$	$1, 1, 2, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8$
1	4	$(1-2u)/(1-3u+u^2)$	1, 1, 2, 5, 13, 34, 89, 233, 610, 1597
5	1	$1/(1-4u+3u^2)$	1, 4, 13, 40, 121, 364, 1093, 3280, 9841, 29524
4	2	$1/(1-4u+3u^2)$	1, 4, 13, 40, 121, 364, 1093, 3280, 9841, 29524
3	3	$1/(1-3u)$	$1, 3, 3^2, 3^3, 3^4, 3^5, 3^6, 3^7, 3^8, 3^9$

We have been able to show that the radii of convergence r of the Maclaurin series of all loss functions obey $1/4 \leq r < 1$. The case $(x, y) = (1, 1)$ is an exceptional case, since in this case, we do not get an infinite sequence since the game then ends in one flip. For $n > 1$, the pairs $(x, y) = (n, 1)$ and $(x, y) = (n-1, 2)$ yield the same loss function. Also, for any particular loss function, there can be only finitely many pairs (x, y) that have this same loss function. The loss sequence for $(x, y) = (1, 4)$ yields the odd indexed Fibonacci numbers (plug Fibonacci numbers or Fibonacci polynomials into an internet search).

Acknowledgements. The authors would like to thank the referees, the editor, and Emeric Deutsch for valuable suggestions.

REFERENCES

1. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., John Wiley, 1968.
2. J. Bak, The Recreational Gambler: Paying the Price for More Time at the Table, this MAGAZINE **80** (2007) 183–194.
3. J. Harper and K. Ross, Stopping Strategies and Gambler's Ruin, this MAGAZINE **78** (2005) 255–269.
4. E. Parzen, *Modern Probability Theory and Its Applications*, Wiley Classics, 1992.
5. R. I. Jewett and K. A. Ross, Random Walks on Z, *College Math. J.* **19** (1988) 330–342.
6. S. Ross, *A First Course in Probability*, 5th Edition, Prentice Hall, 1997.

Lazy Student Integrals

GREGORY GALPERIN
GREGORY RONSSE
Eastern Illinois University
Charleston, IL 61920
ggalperin@eiu.edu
gronsse@eiu.edu

A challenging integral

Let $\alpha \in \mathbb{R}$ and consider the problem of evaluating

$$I(\alpha) = \int_0^\infty \frac{dx}{(1+x^\alpha)(1+x^2)}$$

as a function of α. As a first attempt, we might substitute $x = \tan\theta$ to obtain the integral

$$\int_0^{\pi/2} \frac{\cos^\alpha\theta\, d\theta}{\cos^\alpha\theta + \sin^\alpha\theta}. \tag{1}$$

This seems no better than the original. As a second attempt, split the integral over intervals [0, 1] and $[1, \infty)$. On the second interval substitute $u = 1/x$ to obtain

$$I(\alpha) = \int_0^1 \frac{dx}{(1+x^\alpha)(1+x^2)} + \int_0^1 \frac{u^\alpha du}{(1+u^\alpha)(1+u^2)}.$$

Replacing the dummy variable of integration u with x and combining the two integrals, we obtain

$$I(\alpha) = \int_0^1 \frac{dx}{1+x^2} = \frac{\pi}{4},$$

which is independent of α.

Many people, at first glance, think we should obtain a decreasing function of α and find it surprising that $I(\alpha)$ is constant. It is a trick of the mind. Upon looking at an integral of a bounded function over $[0, \infty)$, we tend to think of the behavior of the function for large values of x and ignore the behavior for $0 \le x \le 1$. The natural symmetry of inversion between [0, 1] and $[1, \infty)$ reveals that the integral over each of these two intervals must sum to a constant.

Now, of course, the integral in equation (1) is the constant $\pi/4$, independent of α. Is there symmetry here?

Tale of the lazy student

In first semester calculus class, students learn to evaluate definite integrals by finding anti derivatives. In order to remind them that an integral is a limit of Riemann sums, it is wise for an instructor to ask them to evaluate an integral similar to

$$\int_{-3}^{3} \frac{\sin x^3 dx}{\sqrt{1+x^4} + \cos(2x)}.$$

The answer is, of course, 0. We are integrating an odd function over an interval which is symmetric about 0. The area above the x-axis is equal to the area below the x-axis.

The lazy student, upon seeing such complicated integrals, has become conditioned to write down 0 immediately and get the right answer. He has noticed that such problems always seem to have positive and negative portions that cancel each other. The instructor must grudgingly admire this valid insight, but he seeks to enforce more careful analysis by altering the problem. So he adds a constant to the integrand, but keeps it mysterious by combining the constant with the fraction to keep the previous denominator but alter the numerator. Our lazy, but perceptive, student now notices a new rule. "Complicated integrals" can be evaluated by evaluating the integrand at 0 and then multiplying by the length of the interval. The exasperated instructor throws in another gimmick by translating the integral along the x-axis by translation to an interval $[a, b]$ in order to disguise the symmetry. The "good students" are completely baffled and angry. However, our lazy, but ingenious, student rises to the occasion. He evaluates the integrand at the interval's midpoint and multiplying by the length of the interval.

The lazy student's method will not, of course, work for all "complicated integrals". Nevertheless, the lazy student would consider the value of the integral in equation (1) as obvious (provided this student was not too lazy in algebra and trig class). Another formula that is obvious to the lazy student is

$$\int_0^{10} \frac{(3 + x^{\sqrt{7}})\,dx}{6 + x^{\sqrt{7}} + (10 - x)^{\sqrt{7}}} = 5. \tag{2}$$

Let us endeavor to make these formulas obvious to the non lazy, by generalization and abstraction.

Lazy student formulas

Let $f\colon [0, a] \longrightarrow \mathbb{R}$ be any continuous function. Substitute $u = a - x$ to obtain

$$\int_0^a f(x)\,dx = \int_0^a f(a - u)\,du.$$

Geometrically, the substitution simply reflects the graph of f about the line $x = a/2$, which obviously leaves the area under the curve invariant.

Now suppose that f satisfies the following **symmetry condition**,

$$f(x) + f(a - x) = 1.$$

Integrating, we obtain the lazy student formula

$$\int_0^a f(x)dx = \frac{a}{2}.$$

This is fine, but how do we obtain functions f that satisfy this restrictive symmetry condition? Set

$$f(x) = \frac{g(x)}{g(x) + g(a - x)},$$

where g is any continuous function on $[0, a]$ such that the denominator of the above does not vanish. It is easily verified that f satisfies the symmetry condition. Furthermore, any function f satisfying the symmetry condition has this form, just take $g = f$.

Equation (2) is now obvious by taking $g(x) = 3 + x^{\sqrt{7}}$ and $a = 10$.

As a special case, note that $\sin(x) = \cos(\pi/2 - x)$, to obtain the lazy student formula:

$$\int_0^{\pi/2} \frac{f(\cos x)dx}{f(\sin x) + f(\cos x)} = \frac{\pi}{4},$$

where f is **any** continuous function defined on [0, 1], such that the denominator in the above integrand does not vanish. This includes the integral in (1) as a special case.

Reference

James Stewart, *Calculus*, 5th ed., Brooks/Cole-Thomson Learning, Belmont, CA, 2003.

PROBLEMS

Proposals

To be considered for publication, solutions should be received by September 1, 2008.

1791. *Proposed by Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.*

Let ABC be a triangle with circumcenter O, perimeter P, and area K. Prove that if

$$\frac{BC}{P} = \frac{[OBC]}{K} = \frac{1}{3},$$

then ABC is equilateral. (Here $[XYZ]$ denotes the area of triangle XYZ.)

1792. *Proposed by H. A. ShahAli, Tehran, Iran.*

Let N be a positive integer. Prove that there is a positive integer n such that $n^2 + 3$ is divisible by at least N distinct primes.

1793. *Proposed by Götz Trenkler, University of Dortmund, Dortmund, Germany*

Let A be an $n \times n$ matrix with complex entries such that $A^2 = A^*$, where A^* denotes the conjugate transpose of A. Show that

a. $\text{rank}(A + A^*) = \text{rank}(A)$

b. $I_n + A$ is nonsingular.

1794. *Proposed by Dorin Marghidanu, Colegiul National "A. I. Cuza", Corabia, Romania*

Let $x_1, x_2, \ldots, x_n \geq e$. Prove that

$$x_1^{\frac{x_1+x_2+\cdots x_n}{x_1}} + x_2^{\frac{x_2+\cdots+x_n}{x_2}} + \cdots + x_{n-1}^{\frac{x_{n-1}+x_n}{x_{n-1}}} + x_n \geq x_1 + 2x_2 + \cdots + (n-1)x_{n-1} + nx_n.$$

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

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

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 **on each page** the reader's name, full address, and an e-mail address and/or FAX number.

1795. *Proposed by Jeff Groah, Montgomery College, Conroe, TX.*

Find a function $f : [0, 1] \to [0, 1]$ such that for each nontrivial interval $I \subseteq [0, 1]$, we have $f(I) = [0, 1]$.

Quickies

Answers to the Quickies are on page 161.

Q979. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

If f is a real-valued differentiable function on $[0, 1]$, then it is well known that f' satisfies the Intermediate Value Property on $[0, 1]$. In addition, it is easy to show that $|f'|$ also satisfies the Intermediate Value Property on $[0, 1]$. Let $\mathbf{f} : [0, 1] \to \mathbb{R}^3$ be a differentiable vector valued function. Must it be true that $\|\mathbf{f}'\|$ satisfies the Intermediate Value Property on $[0, 1]$?

Q980. *Proposed by Ovidiu Furdui, The University of Toledo, Toledo, OH.*

Find all integer solutions to the diophantine equation $x^3 + x^2 + x + 1 = y^3$.

Solutions

Deducing a limit **April 2007**

1766. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let f be differentiable on $(0, \infty)$ and let ω be a positive real number. Prove that if $\lim_{x\to\infty}(f'(x) + \omega f(x)) = A$, then $\lim_{x\to\infty} f(x) = A/\omega$.

I. *Solution by Evangelos Mouroukos, Agrino, Greece.*
Because $\lim_{x\to\infty} e^{\omega x} = \infty$, we may apply L'Hôpital's rule [1] to get

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \frac{f(x)e^{\omega x}}{e^{\omega x}}$$
$$= \lim_{x\to\infty} \frac{f'(x)e^{\omega x} + \omega f(x)e^{\omega x}}{\omega e^{\omega x}} = \frac{1}{\omega} \lim_{x\to\infty} \left(f'(x) + \omega f(x)\right) = \frac{A}{\omega}.$$

1. Rudin, Walter, *Principles of Mathematical Analysis*, 3rd Edition, 1976, pg. 109.

II. *Solution by Henry J. Ricardo, Medgar Evers College, CUNY, Brooklyn, NY.*
Given $\epsilon > 0$, there is a $c > 0$ such that

$$\left|\left(e^{\omega x}\left(f(x) - \frac{A}{\omega}\right)\right)'\right| = \left|e^{\omega x}\left(f'(x) + \omega f(x) - A\right)\right| \le \omega\epsilon e^{\omega x} \quad \text{for} \quad x \ge c.$$

Therefore integrating on an interval $[c, x]$ we find

$$\left|e^{\omega x}\left(f(x) - \frac{A}{\omega}\right) - e^{\omega c}\left(f(c) - \frac{A}{\omega}\right)\right| \le \epsilon\left(e^{\omega x} - e^{\omega c}\right),$$

from which

$$\left|\left(f(x) - \frac{A}{\omega}\right) - e^{\omega(c-x)}\left(f(c) - \frac{A}{\omega}\right)\right| \le \epsilon\left(1 - e^{\omega(c-x)}\right).$$

It follows that

$$\left|f(x) - \frac{A}{\omega}\right| \leq e^{\omega(c-x)} \left|f(c) - \frac{A}{\omega}\right| + \epsilon\left(1 - e^{\omega(c-x)}\right),$$

so $\left|f(x) - \frac{A}{\omega}\right|$ can be made arbitrarily small by taking x sufficiently large.

Also solved by Michel Bataille (France), Gerald E. Bilodeau, Robert Calcaterra, Adam Coffman, Apostolis Demis (Greece), David Doster, Robert L. Doucette, Jayanthi Ganapathy, Peter Gressis, Eugene A. Herman, Chris Hill, Enkel Hysnelaj (Australia), Geoffrey A. Kandall, Victor Y. Kutsenok, Elias Lampakis (Greece), Kee-Wai Lau (China), Jerry Metzger, Northwestern University Math Problem Solving Group, Paolo Perfertti (Italy), Albert Stadler (Switzerland), Marian Tetiva (Romania), Dave Trautman, Michael Vowe (Switzerland), and the proposer. There was one incorrect submission.

An equilateral characterization **April 2007**

1767. *Proposed by Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.*

Let G be the centroid of $\triangle ABC$. Prove that if $\angle BAC = 60°$ and $\angle BGC = 120°$, then the triangle is equilateral

Solution by Minh Can, Irvine Valley College, Irvine, CA.

Let $a = BC$, $b = CA$, $c = AB$, $m = GB$, and $n = GC$. By the Law of Cosines $a^2 = b^2 + c^2 - bc$. Substituting this into Stewart's formula for the length of a median we find

$$m^2 = \frac{4}{9}\left(\frac{a^2}{2} + \frac{c^2}{2} - \frac{b^2}{4}\right) = \frac{1}{9}\left(3c^2 + (c-b)^2\right)$$

and, similarly,

$$n^2 = \frac{1}{9}\left(3b^2 + (b-c)^2\right).$$

From $[BAC] = 3[BGC]$ with $\angle BAC = 60°$ and $\angle BGC = 120°$, we find $bc = 3mn$. Substituting the above expressions for m^2 and n^2 we find

$$9a^2b^2 = 9^2m^2n^2 = \left(3c^2 + (c-b)^2\right)\left(3b^2 + (b-c)^2\right).$$

It follows that $b = c$, so triangle ABC is equilateral.

Also solved by Herb Bailey, Michel Bataille (France), Jany C. Binz (Switzerland), Robert Calcaterra, Hojin Choi (Korea), John Christopher, Adam Coffman, Miguel Amengual Covas (Spain), Chip Curtis, Prithwijit De (Ireland), Jim Delany, Apostolis Demis (Greece), David Doster, Robert L. Doucette, Euler's FOILers, Dmitry Fleischman, Michelle Ghrist, Michael Goldenberg and Mark Kaplan, G.R.A.20 Problem Solving Group (Italy), Peter Gressis, Jeff Groah, Chris Hill, Enkel Hysnelaj (Australia), Geoffrey A. Kandall, Victor Y. Kutsenok, Elias Lampakis (Greece), Kee-Wai Lau (China), Charles McCraken, Kim McInturff, Evangelos Mouroukos (Greece), José H. Nieto (Venezuela), Northwestern University Math Problem Solving Group, Pθlar Bear Problem Solvers, Jawed Sadek, H. A. ShahAli (Iran), Raul Simon (Chile), Seshadri Sivakumar, Skidmore College Problem Group, Earl A. Smith, Albert Stadler (Switzerland), H. T. Tang, Marian Tetiva (Romania), Michael Vowe (Switzerland), Paul Weisenhorn (Germany), Paul Zorn, and the proposer. There were two solutions with no name and one incorrect submission.

Square partitions **April 2007**

1768. *Proposed by G.R.A.20 Problem Solving Group, Rome, Italy.*

For which positive integers n can the set $\{1, 2, \ldots, 2n\}$ be partitioned into n two element subsets so that the sum of the two numbers in each subset is a perfect square?

Solution by John Christopher, California State University Sacramento, Sacramento CA.

Examining the set $S = \{1, 2, \dots, 2n\}$ for each positive integer $n \le 30$, we find that there is at least one partition of S into n two element subsets so that the sum of the two numbers in each subset is a perfect square as long as n is not one of the seven integers 1, 2, 3, 5, 6, 10, 11. We show by induction that the set S will have a partition of the required type for all integers $n \ge 31$.

Assume a partition of the required type exists for $n = 12, 13, \dots, k-1$, where $k \ge 31$, and consider $n = k$. Find the unique positive integer t such that

$$\frac{(2t+1)^2}{2} < 2k < \frac{(2t+3)^2}{2}, \tag{1}$$

and note $t \ge 5$. Let $m = (2t+1)^2 - 2k$. We show that $24 < m < 2k$. From (1) it follows that

$$(2t+1)^2 - \frac{(2t+3)^2}{2} < m < (2t+1)^2 - \frac{(2t+1)^2}{2},$$

which leads to $2t^2 - 2t - 7/2 < m < 2k$. For $t \ge 5$, $24 < 2t^2 - 2t - 7/2$ which implies that $24 < m < 2k$.

Now write $S = \{1, 2, 3, \dots, 2k\} = X \cup Y$, where

$$X = \{1, 2, \dots, m-1\} \qquad \text{and} \qquad Y = \{m, m+1, \dots, 2k\}.$$

Because $m-1$ is even and $12 \le (m-1)/2 < k$, we know from the inductions hypothesis that the set X can be partitioned into two element subsets such that the sum of the elements in each of the subsets is a square. The set Y can be expressed as

$$Y = \{m, 2k\} \cup \{m+1, 2k-1\} \cup \cdots \cup \{2t^2 + 2t, 2t^2 + 2t + 1\},$$

and the sum of the two elements in each of these subsets is $(2t+1)^2$. This completes the induction.

Therefore the set $\{1, 2, \dots, 2m\}$ can be partitioned as desired for positive any integer $m \in \{4, 7, 8, 9\} \cup \{12, 13, 14, \dots\}$.

Also solved by Dmitry Fleischman, Chris Hill, Peter Hohler (Switzerland), Enkel Hysnelaj (Australia), Eugen J. Ionascu and Albert VanCleave, Jerry Metzger and Thomas Richards, José H. Nieto (Venezuela), Paul Weisenhorn (Germany), and the proposers. There were five incorrect submissions.

Expansion coefficient **April 2007**

1769. *Proposed by Michel Bataille, Rouen, France.*

For positive integer n, let

$$P_n(x, y) = \sum_{k=0}^{n} \binom{2n+1}{2k+1} x^{n-k}(x+y)^k.$$

Find a closed form expression for the coefficient of $x^i y^j$ when P_n is expanded.

Solution by Brian Bradie, Christopher Newport University, Newport News, VA.

First note that when P_n is expanded, each term is of degree n; therefore, the coefficient of $x^i y^j$ is zero unless $i + j = n$. Expanding $(x+y)^k$ using the binomial theorem, we see

$$P_n(x,y) = \sum_{k=0}^{n} \binom{2n+1}{2k+1} x^{n-k} \left[\sum_{j=0}^{k} \binom{k}{j} x^{k-j} y^j \right]$$
$$= \sum_{k=0}^{n} \sum_{j=0}^{k} \left[\binom{2n+1}{2k+1} \binom{k}{j} x^{n-j} y^j \right]$$
$$= \sum_{j=0}^{n} \left[\sum_{k=j}^{n} \binom{2n+1}{2k+1} \binom{k}{j} \right] x^{n-j} y^j$$
$$= \sum_{j=0}^{n} 4^{n-j} \binom{2n-j}{j} x^{n-j} y^j.$$

With $i = n - j$, it follows that the coefficient of $x^i y^j$ is

$$4^i \binom{2i+j}{j}.$$

Also solved by Michael S. Becker and Charles K. Cook, Robert Calcaterra, Chip Curtis, G.R.A.20 Problem Solving Group (Italy), Enkel Hysnelaj (Australia), Simone Lamont and Farley Mawyer, Jerry Metzger, Ángel Plaza and Sergio Falcón (Spain), Ossama A. Saleh and Terry J. Walters, Volkhard Schindler (Germany), Edward Schmeichel, Nicholas C. Singer, Albert Stadler (Switzerland), Michael Vowe (Switzerland), Paul Weisenhorn (Germany), and the proposer. There was one incomplete submission.

Limit of a recursive sequence **April 2007**

1770. *Proposed by Scott N. Armstrong, University of California, Berkeley, CA, and Christopher J. Hillar, Texas A&M University, College Station, TX.*

Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be nonnegative real numbers summing to 1, and let $a_1, a_2 \ldots, a_k$ be complex numbers. For $n > k$, define

$$a_n = \lambda_1 a_{n-1} + \lambda_2 a_{n-2} + \cdots + \lambda_k a_{n-k}.$$

Prove that if there is a j, $1 \leq j \leq n-1$, such that λ_j and λ_{j+1} are both nonzero, then $\lim_{n\to\infty} a_n$ exists. In addition, determine the value of this limit.

Solution by Eugene Herman, Grinnell College, Grinnell, IA.

When $k = 1$, we have $a_n = a_{n-1}$ for all $n > 1$, which implies that $a_n = a_1$ for all n and hence that a_1 is the limit. Now assume $k \geq 2$ and for $n \geq k$, let $\mathbf{v}_n$ denote the row vector $\langle a_{n+1-k}, \ldots, a_{n-1}, a_n \rangle$ in $\mathbb{C}^k$. Then the given recurrence relation can be expressed as a recurrence relation on the vectors $\mathbf{v}_n$ with

$$\mathbf{v}_n = \mathbf{v}_{n-1} S \quad \text{for all } n > k, \tag{1}$$

where S is the $k \times k$ stochastic matrix whose last column is $\lambda_k, \ldots, \lambda_1$ and whose jth column, $1 \leq j \leq k-1$ is the $(j+1)$st column of the $k \times k$ identity matrix. Thus, S has 1 as an eigenvalue, and all of its remaining eigenvalues are less than or equal to 1 in absolute value.

We now show that the eigenvalue 1 has algebraic multiplicity 1 and that none of the remaining eigenvalues have absolute value 1. The characteristic polynomial of S is

$$p(x) = x^k - \sum_{j=1}^{k} \lambda_j x^{k-j}$$

Therefore

$$p'(1) = k - \sum_{j=1}^{k-1} \lambda_j (k-j) = k - k\sum_{j=1}^{k-1} \lambda_j + \sum_{j=1}^{k-1} j\lambda_j > k - k\sum_{j=1}^{k-1}\lambda_j \geq 0.$$

Since $p'(1) \neq 0$, the eigenvalue 1 has algebraic multiplicity 1. If z is an eigenvalue of S with $|z| = 1$ and $z \neq 1$, then $p(z) = 0$, and so

$$1 = |z^k| = \left| \sum_{j=1}^{k} \lambda_j z^{k-j} \right| \leq \sum_{j=1}^{k} \lambda_j = 1$$

Moreover, the above inequality is strict, because there are two consecutive terms $\lambda_j z^{k-j}$ and $\lambda_{j+1} z^{k-j-1}$ where both λ_j and λ_{j+1} are nonzero. Since these consecutive terms (regarded as vectors in $\mathbb{R}^2$) are not positive scalar multiples of one another, the length of their sum is strictly less than the sum of their lengths. Therefore z cannot be an eigenvalue of S.

Given these facts about the eigenvalues of S, it is well known that $\lim_{n\to\infty} S^n$ exists. Let U denote this limit. Hence, by (1),

$$\lim_{n\to\infty} \mathbf{v}_n = \lim_{n\to\infty} \mathbf{v}_k S^{n-k} = \mathbf{v}_k \lim_{n\to\infty} S^n = \mathbf{v}_k U.$$

Therefore

$$\lim_{n\to\infty} a_n = \mathbf{v}_k \mathbf{u} \quad \text{where } \mathbf{u} \text{ is column } k \text{ of } U. \tag{2}$$

Furthermore, every column of U is the same vector $\mathbf{u}$, and $\mathbf{u}$ is the unique probability vector that is an eigenvector of S associated with the eigenvalue 1. The equation $S\mathbf{u} = \mathbf{u}$ can be written as the system of equations

$$\lambda_k u_k = u_1, \quad u_1 + \lambda_{k-1} u_k = u_2, \quad u_2 + \lambda_{k-2} u_k = u_3, \quad \ldots, \quad u_{k-1} + \lambda_1 u_k = u_k$$

We can easily solve this system to get u_j, $1 \leq j \leq k$, in terms of u_k:

$$u_j = \left(\sum_{i=k+1-j}^{k} \lambda_i \right) u_k, \qquad j = 1, \ldots, k. \tag{3}$$

Because $\mathbf{u}$ is a probability vector,

$$\sum_{j=1}^{k} \left(\sum_{i=k+1-j}^{k} \lambda_i \right) u_k = 1, \qquad \text{and hence} \quad u_k = \frac{1}{\sum_{j=1}^{k} \sum_{i=k+1-j}^{k} \lambda_i} = \frac{1}{\sum_{j=1}^{k} j\lambda_j}.$$

Therefore, by (2) and (3), we conclude that

$$\lim_{n\to\infty} a_n = \frac{1}{\sum_{j=1}^{k} j\lambda_j} \sum_{j=1}^{k} \left(\sum_{i=k+1-j}^{k} \lambda_i \right) a_j.$$

Note: The above proof is also valid when $a_1, \ldots, a_k$ belong to a normed linear space V over $\mathbb{R}$. In this case, the vectors $\mathbf{v}_n$ belong to V^k in place of $\mathbb{C}^k$.

Editor's Note: Michael Vowe noted that the problem appears in William Feller's *An introduction to Probability Theory and Its Applications*, Vol. 1, third edition, on page 333.

Also solved by Robert Calcaterra, Dmitry Fleischman, Russell Jay Hendel, Enkel Hysnelaj (Australia), Mark Kaplan and Michael Goldenberg, Jerry Metzger, Nicholas C. Singer, Albert Stadler (Switzerland), Paul Weisenhorn (Germany), and the proposer. There was one incorrect submission.

Answers

Solutions to the Quickies from page 156.

A979. The answer is no. As an example, consider the function $\mathbf{f}$ defined by

$$\mathbf{f}(t) = \begin{cases} (0,\ 0,\ 0) & t = 0 \\ \left(t^2 \cos\left(\frac{1}{t}\right),\ t^2 \sin\left(\frac{1}{t}\right),\ t\right) & 0 < t \le 1. \end{cases}$$

Then

$$\mathbf{f}'(t) = \begin{cases} (0,\ 0,\ 1) & t = 0 \\ \left(2t \cos\left(\frac{1}{t}\right) + \sin\left(\frac{1}{t}\right),\ 2t \sin\left(\frac{1}{t}\right) - \cos\left(\frac{1}{t}\right),\ 1\right) & 0 < t \le 1. \end{cases}$$

It follows that

$$\|\mathbf{f}'(t)\| = \begin{cases} 1 & t = 0 \\ \sqrt{4t^2 + 2} & 0 < t \le 1. \end{cases}$$

Because $\|\mathbf{f}'(0)\| = 1$ and $\|\mathbf{f}'(t)\| > \sqrt{2}$ for $0 < t \le 1$, we see that $\|\mathbf{f}'\|$ does not satisfy the Intermediate Value Theorem on $[0,\ 1]$.

A980. The solutions are $(x,\ y) = (0,\ 1)$ and $(x,\ y) = (-1,\ 0)$. To prove this first observe that

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

The left inequality is immediate, while the right inequality is equivalent to $x(x + 1) \ge 0$, which is true for all integers x. Thus, $(x - 1)^3 < y^3 \le (x + 1)^3$, and it follows that $y = x$ or $y = x + 1$. The solutions are then obtained by straight forward calculations.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

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

Brams, Steven J., *The Presidential Election Game*, 2nd ed., A K Peters, 2008; xxiii + 194 pp, \$29 (P). ISBN 978-1-56881-348-6. *Mathematics and Democracy: Designing Better Voting and Fair-Division Procedures*, Princeton University Press, 2008; xvi + 373 pp, \$27.95. ISBN 978–0–691–13321–8.

By the time you read this, most of the presidential caucuses and primaries will be over and the the major parties' candidates may have been determined well in advance of their conventions. In the first book, Brams applies decision theory and game theory to analyze campaigns and elections, devoting the first chapter to primaries, followed by chapters on party conventions, the general election, coalition politics, the White House tapes game of 1974, and approval voting. The chapters are reproduced from the first edition in 1978, so all examples are from an era before current students were born; but a new introduction summarizes more recent developments, and all the ideas are once again very timely to a new generation galvanized by politics (as I write just after "Super Tuesday," Phil Straffin's bandwagon curve on pp. 114–117 seems quite relevant to the race between Clinton and Obama). In the new book, Brams treats the mathematics of voting (he espouses approval voting) and fair division (such as cake-cutting or allocation of cabinet ministries). The heightened consciousness of college students about issues of democracy in the U.S. and elsewhere (e.g., Iraq) makes this a good time to teach a special course on the underlying mathematics, and this book would be a good source for such a course.

Dudley, Underwood (ed.), *Is Mathematics Inevitable? A Miscellany*, MAA, 2008; x + 325 pp, \$56.95 (\$45.50 to MAA members). ISBN 978-0-88385-566-9.

The provocative title is from an included 1958 essay by Nathan Altshiller Court. Don't expect a negative answer from this volume! It is a collection of essays without a real theme except that editor Dudley—known for his good taste and his own provocative writing—feels that these essays deserve preservation and presentation to a new generation. They range from a defense before Parliament of quadratic equations to why students (correctly) perceive classes as larger than average, from how driver's license numbers are assigned to why mathematics is applicable to the physical world. Authors include Jean Dieudonné, Richard Guy, Morris Kline, Paul Halmos, and Lewis Carroll. Of course, there is mathematical humor and a couple of pieces by mathematical cranks. Each piece has a prelude and a postlude by Dudley, together with brief biographical information about the author.

Maor, Eli, *The Pythagorean Theorem: A 4,000-Year History*, Princeton University Press, 2007; xvi + 259 pp, \$24.95. ISBN 978-0-691-12526-8.

Most adults remember little of high school geometry, but they remember the Pythagorean theorem. This book goes beyond the theorem and its proofs to set it beautifully in the context of its time and subsequent history, up through its connection to relativity and its association with the Fermat's Last Theorem. Author Maor writes conversationally but does not shy away from using equations. So this book is not for the average adult; it does, however, belong in your public library and your local high school library (have you checked the mathematics section of either of those lately?).

Weinstein, Lawrence, and John A. Adams, *Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin*, Princeton University Press, 2008; xv + 301 pp, $19.95 (P). ISBN 978-0-691-12949-5.

How many airplane flights do Americans take in one year? How much farmland would the U.S. need to devote to corn to power all its cars on ethanol? How much would the ocean surface rise if the ice caps melted? These and other questions (some serious, some silly) are asked (with hints given) and answered in this book, which is devoted to cultivating the art of "back of the envelope" (or napkin) calculation. An introductory chapter explains scientific notation, how to convert units, and "back of the napkin" rule for significant digits (keep only one!). The questions are divided by topic (general, animals and people, transportation, energy and work, hydrocarbons and carbohydrates, the Earth, energy and the environment, the atmosphere, and risk), and a final chapter has 33 unanswered questions. Brief appendices give useful numbers and formulas, metric prefixes and abbreviations, and sample objects along the scales of length, area, density, and mass. Readers who enjoy this book may want to graduate to John Harte's *Consider a Spherical Cow* (University Science Books, 1988) and *Consider a Cylindrical Cow* (2001) (I am eagerly hoping for a successor *Conical Cow* book).

Schiff, Joel L., *Cellular Automata: A Discrete View of the World*, Wiley-Interscience, 2008; xvi + 252 pp, $105. ISBN 978-0-470-16879-0.

"I expect the children of 50 years from now will learn cellular automata before they learn algebra." Stephen Wolfram (writing in 2006) may yet turn out to be right, but alternatively he may instead have underestimated conservatism in education. After all, 25 years ago some thought that students would long since have been learning discrete mathematics in college before they learned calculus. This book tries to move toward his goal, though it is not for the pre-algebra set. It treats dynamical systems, one- and two-dimensional cellular automata, applications (biology, sociology, game theory, physical phenomena), and complexity (including autonomous agents, such as honey bees and ants). The book is billed as a textbook, "hitting the highlights as the author sees them," and it covers a lot of territory briefly. It has a generous complement of images, but there are no exercises or suggested projects.

Wainer, Howard, The most dangerous equation, *American Scientist* (May–June 2007) 249–256.

We mathematicians are not accustomed to thinking of equations as dangerous. Statistician Wainer cites $E = mc^2$ as an equation that is dangerous if you know it, but his concern here is equations that are dangerous if you *don't* know them. The three examples he cites are Kelley's equation (you estimate best by regressing an observation toward the historic mean), the standard linear regression equation, and what he calls *de Moivre's equation*. Only for the last does he give the equation itself: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, the relation between standard error of the mean of a sample, the standard deviation of the population, and the size of the sample. Finding de Moivre's equation more dangerous than the other two, Wainer cites five situations where its misuse have led to enormous losses and hardship. Those situations are the trial of the Pyx (quality control at the British mint over 600 years), kidney-cancer rates (urban vs. rural), the small-schools movement (does small produce better scores?), the safest cities (hint: they aren't large), and sex differences in academic test performance (males have greater variation). You'll have to read the article to learn the details, but the fundamental root of the misuse is misunderstanding how variation changes with size.

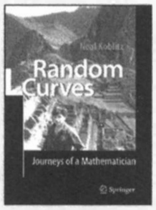
Stigler, Stephen M., Eight centuries of sampling inspection: The trial of the Pyx, *Journal of the American Statistical Association* 72 (1977) 493–500. Isaac Newton as a probabilist, *Statistical Science* 21 (2006) (3) 400–403.

Wainer's article reviewed above criticizes the trial of the Pyx. Stigler's first article tells about it and mentions the involvement of Newton, Master of the Mint for 28 years. Stigler's second article demonstrates that Newton was acquainted with probability: Before his stint at the Mint, he was drawn into correspondence about the very same dice problem as Pascal had been 40 years earlier. Newton had the right intuition and calculations but an incorrect general argument. Unfortunately, his experience at the Mint did not lead him to discover de Moivre's equation.

Springer
the language of science

springer.com

New and Noteworthy from Springer



Random Curves

Journeys of a Mathematician

N. Koblitz, University of Washington, Seattle, WA, USA

These autobiographical memoirs of Neal Koblitz, co-inventor of one of the two most popular forms of encryption and digital signature, cover many topics besides his own personal career in mathematics and cryptography - travels to the Soviet Union, Latin America, Vietnam and elsewhere, political activism, and academic controversies relating to math education, the C. P. Snow two-culture problem, and mistreatment of women in academia.

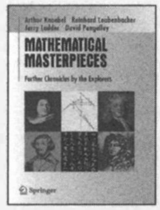
2008. X, 392 p. 28 illus., 18 in color. Hardcover
ISBN 978-3-540-74077-3 ► **$49.95**

An Introduction to Mathematical Cryptography

J. Hoffstein, J. Pipher, J. H. Silverman, Brown University, Providence, RI, USA

An ideal introduction for mathematics and computer science students to the mathematical foundations of modern cryptography, this self-contained text emphasizes the mathematics behind the theory of public key cryptosystems and digital signature schemes. Key topics include ► cryptographic constructions, such as Diffie-Hellmann key exchange, discrete logarithm-based cryptosystems, the RSA cryptosystem, and digital signatures ► mathematical tools for cryptography, including primality testing, factorization algorithms, probability theory, information theory, and collision algorithms ► an in-depth treatment of recent cryptographic innovations, such as elliptic curves, elliptic curve and pairing-based cryptography, lattices, lattice-based cryptography, and the NTRU cryptosystem.

2008. Approx. 550 p. (Undergraduate Texts in Mathematics) Hardcover
ISBN: 978-0-387-77993-5 ► **approx. $49.95**



Mathematical Masterpieces

Further Chronicles by the Explorers

A. Knoebel, Albuquerque, NM, USA; **R. Laubenbacher**, Virginia Bioinformatics Institute, Blacksburg, VA, USA; **J. Lodder, D. Pengelley**, New Mexico State University, Las Cruces, NM, USA

Here are the stories of four mathematical adventures, including the Bernoulli numbers as the passage between discrete and continuous phenomena, the search for numerical solutions to equations throughout time, the discovery of curvature and geometric space, and the quest for patterns in prime numbers. Each story is told through the words of the pioneers of mathematical thought. Particular advantages of the historical approach include providing context to mathematical inquiry, perspective to proposed conceptual solutions, and a glimpse into the direction research has taken.

2007. XII, 336 p. 75 illus., 2 in color. Softcover
ISBN 978-0-387-33061-7 ► **$39.95**
Also available in hardcover
ISBN 978-0-387-33060-0 ► **$79.95**

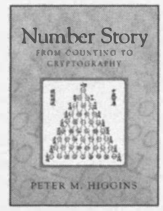
History of Mathematics

A Supplement

C. Smorynski, Westmont, IL, USA

History of Mathematics offers deeper coverage of select topics, providing students with material which encourages more critical thinking. It also includes proofs of important results typically neglected in the modern history of mathematics curriculum. Coverage includes ► A new approach to the historical development of the natural numbers ► Historically interesting (and hard to find) poems and humorous song lyrics with mathematical themes

2008. VI, 274 p. 42 illus. Hardcover
ISBN 978-0-387-75480-2 ► **$49.95**



Number Story

From Counting to Cryptography

P. M. Higgins, University of Essex, UK

Numbers have fascinated people for centuries. They are familiar to everyone, forming a central pillar of our understanding of the world, yet the number system was not presented to us "gift-wrapped" but, rather, was developed over millennia. Peter Higgins distills centuries of work into one delightful narrative that celebrates the mystery of numbers and explains how different kinds of numbers arose and why they are useful. Full of historical snippets and interesting examples, the book ranges from simple number puzzles and magic tricks, to showing how ideas about numbers relate to real-world problems.

2008. XII, 324 p. 33 illus. Hardcover
ISBN 978-1-84800-000-1 ► **$25.00**

The Mathematical Intelligencer

Co-Editor-in-Chief: **M. L. Senechal; C. Davis**

This journal publishes articles about mathematics, mathematicians, and the history and culture of mathematics. It presents expository articles on all kinds of mathematics and interdisciplinary trends, and articles that portray the diversity of mathematical communities and mathematical thought. Not only does The Mathematical Intelligencer inform, it also entertains. Throughout the journal, humor, puzzles, poetry, fiction, and art can be found. The journal also features information new interdisciplinary trends and relations between mathematics and other areas of culture.

ISSN 0343-6993 (print version)

Sign Up for New Book Alerts at springer.com

Easy Ways to Order for the Americas ► **Write:** Springer Order Department, PO Box 2485, Secaucus, NJ 07096-2485, USA ► **Call: (toll free)** 1-800-SPRINGER ► **Fax:** 1-201-348-4505 ► **Email:** orders-ny@springer.com or **for outside the Americas** ► **Write:** Springer Distribution Center GmbH, Haberstrasse 7, 69126 Heidelberg, Germany ► **Call:** +49 (0) 6221-345-4301 ► **Fax :** +49 (0) 6221-345-4229 ► **Email:** SDC-bookorder@springer.com
► Prices are subject to change without notice. All prices are net prices.

013579x

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036