Correction

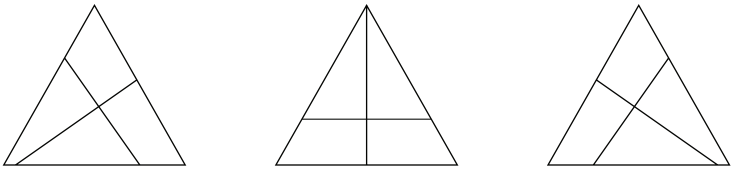
Article title: Zindler Points of Triangles

Authors: Allan Berele & Stefan Catoiu

Journal: *Mathematics Magazine*

DOI: http://dx.doi.org/10.1080/0025570X.2022.2127301

When the above article first published online, Figure 2 appeared incorrectly. The correct figure image is below.



The article has been corrected.

Math. Mag. **95** (2022) i. doi:10.1080/0025570X.2022.2146384 © Mathematical Association of America

Acknowledgements

In addition to our Associate Editors, the following referees have assisted the MAGAZINE *during the past year (May 2021 to May 2022). We thank them for their time and care.*

Claudio Alsina *Universitat Politecnica de Catalunya*
Dave Auckly *Kansas State University*
Jathan Austin *Salisbury University*
Edward Barbeau, Jr. *University of Toronto*
Jennifer Beineke *Western New England University*
Lowell Beineke *Purdue University Fort Wayne*
Arthur Benjamin *Harvey Mudd College*
Ethan Berkove *Lafayette College*
Leah Berman *University of Alaska Fairbanks*
Jeffrey Blanchard *Grinnell College*
Harold Boas *Texas A&M University*
David Bressoud *Macalester College*
David Clark *Randolph Macon College*
Matt Davis *Muskingum University*
Michelle DeDeo *University of North Florida*
Gregory Dresden *Washington and Lee University*
Joshua Ducey *James Madison University*
David Duncan *James Madison University*
Tom Edgar *Pacific Lutheran University*
Rebecca Garcia *Sam Houston State University*
Skip Garibaldi *Center for Communications Research*
Russell Gordon *Whitman College*
Gary Gordon *Lafayette College*
Pamela Gorkin *Bucknell University*
Daniel Heath *Pacific Lutheran University*
Brian Hopkins *Saint Peter's University*
Danrun Huang *Saint Cloud State University*
Stanley Huddy *Fairleigh Dickinson University*
Warren Johnson *Connecticut College*
Edward Keppelmann *Univeristy of Nevada Reno*
Reginald Koo *University of South Carolina Aiken*
Elias Lampakis *Kiparissia, Greece*
Dominic Lanphier *Western Kentucky University*
Martin Lukarevski *Goce Delchev University*
Richard Mabry *Louisiana State University in Shreveport*
Kent Morrison *American Institute of Mathematics*
Roger Nelsen *Lewis and Clark College*
Michael Orrison *Harvey Mudd College*
T. Petersen *DePaul University*
Charles Pruett *James Madison University*
David Richeson *Dickinson College*
Jason Rosenhouse *James Madison University*
Bonita Saunders *National Institute of Standards and Technology*
Brittany Shelton *Albright College*
Deirdre Smeltzer *Mathematical Association of America*
Francis Su *Harvey Mudd College*
Ricardo Teixeira *University of Houston Victoria*
David Treeby *Monash University, Australia*
Anil Venkatesh *Adelphi University*
Kenneth Williams *Carleton University*
Peter Winkler *Darmouth College*
William Wood *University of Northern Iowa*

Math. Mag. **95** (2022) ii. doi:10.1080/0025570X.2022.2142406 © Mathematical Association of America

LETTER FROM THE EDITOR

Welcome to our second straight double-sized issue! Such is the volume of first-rate mathematical exposition the magazine receives that we have amassed a substantial backlog of accepted articles. As a result, some authors have had to wait entirely too long to see their papers in print. To alleviate this problem, Taylor and Francis has allowed us to dramatically increase our page count for a while to work through the backlog more quickly. In addition to thanking them for this largesse, let me also thank Amanda Gedney, Bonnie Ponce, and Annie Petitt for their invaluable assistance and endless patience in putting together these mammoth double issues.

So let's see what have on tap!

Ezra Brown and Adrian Rice get us started with a marvelous exposition of Hurwitz's classical result on sums of squares. At first blush, this theorem arises from a trivial algebraic observation: If each of two positive integers is the sum of two squares, then their product is also the sum of two squares. A similar claim could be made for sums of four and eight squares (or one square, for that matter, though that case seems a bit trivial). But why is such a claim false for other numbers, such as five or six squares? Answering that question quickly takes you into the deep waters of normed division algebras, and from there to the quaternions and octonions. It's all very deep, but Brown and Rice's flawless exposition makes it perfectly comprehensible.

Franklin Gould also takes his inspiration from the quaternions and octonions. Whereas Brown and Rice used a number-theoretic curiosity as their entry point, Gould takes a more geometrical and linear algebraic approach. It is fascinating to see similar questions explored from such different perspectives, and it is a pleasure to publish both of these wonderfully lucid articles.

The knot theorists in the audience will want to check out the article by Louis Kauffman, Devika Prasad, and Claudia Zhu. After introducing readers to the world of knots and their colorings, they provide an exceedingly clever proof of a result on Brunnian links (with the classic Borromean rings being the most famous example of such a link).

Julius Barbanel focuses on a classic geometrical construction task: doubling a cube. This is famously impossible using the standard Euclidean straightedge and compass. But if you relax the rules just a little, then this task becomes possible after all. Barbanel's article is a fascinating mix of history and geometry. He provides an education on an unjustly forgotten chapter in ancient Greek geometry and ruminates on extensions of Greek methods to higher dimensions.

This issue will certainly provide the number theorists with plenty of food for thought. Ethan Berkove and Michael Brilleslyper find the Fibonacci numbers lurking in a problem of polynomial long division. The ensuing investigation leads them to the Tribonacci numbers, generating functions, and novel proofs of familiar summation formulas. Daeyeol Jeon and Heonkyu Lee investigate the endlessly fascinating figurate numbers. Specifically, they determine, among other things, which numbers are expressible as the difference of two polygonal numbers. Fang Chen rounds out our number theory offerings. She uses a classic brainteaser as a gateway for explicating the nature of mathematical research. In addition to the considerable mathematical interest of her article, she also has my personal favorite title from my time as Editor.

Math. Mag. **95** (2022) 420–421. doi:10.1080/0025570X.2022.2136437 © Mathematical Association of America

Geometry is well-represented in our table of contents. Atol Sasane generalizes a result due to Archimedes. Allan Berele and Stefan Catoiu study certain problems from convex geometry and the theory of equipartitions. And Greg Markowsky, Dylan Phung, and David Treeby generalize the familiar HM-GM-AM inequality by realizing each such mean as the centroid of a region in the Cartesian plane.

Mathematics Magazine publishes expository articles, and we include strong undergraduate students among our intended readers. Since analysis tends to be somewhat dense and technical, it does not always lend itself to the sort of conversational tone we like in the articles we publish. With that in mind, it is my great pleasure to be able to publish not one, but two such articles in this issue. Daniel Daners unifies three classic analysis results by showing how all can be proved using the same building block. In their article, Ehssan Khanmohammadi and Omid Khanmohamadi serve up an excursion into functional analysis. A consideration of basic questions regarding convergence and divergence of sequences leads them to a dual of the uniform boundedness principle.

Japanese pencil puzzles are a bottomless pit of interesting mathematical problems. Jacob Boswell, Jacob Clark, and Chip Curtis explore Nurikabe, which is not as well-known as, say, Sudoku or KenKen. For those of a mathematical cast of mind, which I assume includes all readers of this magazine, Nurikabe leads to some very natural combinatorial questions. Boswell, Clark, and Curtis offer some fascinating insights in this area.

Clifford Johnston addresses a very practical problem: how can we ensure that we receive truthful answers to sensitive survey questions? Johnston explores how some elementary probability theory provides an answer to this question, and he shows how these insights can be translated into classroom exercises. As the Editor of the magazine, I wish I had more opportunities to publish pieces of this sort. It shows that an article can be mathematically interesting even without a mountain of dense jargon and notation.

As always, we round out the issue with proofs without words, original problems, and short reviews of newsworthy articles. A fine way to finish off 2022. See you next year!

Jason Rosenhouse, Editor

An Accessible Proof of Hurwitz's Sums of Squares Theorem

EZRA BROWN
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061-0123
ezbrown@math.vt.edu

ADRIAN RICE
Randolph-Macon College
Ashland, VA 23005-5505
arice4@rmc.edu

The purpose of this paper is to give a simple proof, intelligible to undergraduates, that a particular multiplicative formula for sums of n squares can only occur when $n = 1, 2, 4$, or 8, a result originally proved by Hurwitz in 1898. We begin with a brief survey of the history of sums of squares, leading to a discussion of the related topic of normed division algebras over the real numbers.* This story culminates with a crucial paper by Dickson in 1919 that not only contained an exposition of Hurwitz's 1898 proof, but which also outlined a new process for producing division algebras over the reals. That process, now called the Cayley-Dickson construction, is intimately connected with the product formula for sums of squares and the dimensions necessary for its existence. For this reason, we present an introduction to the Cayley-Dickson construction for beginners, together with a proof of Hurwitz's theorem accessible to anyone with a basic knowledge of undergraduate algebra.

Historical background

The question in which we are interested is: for which values of n does it hold that

$$\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)\left(y_1^2 + y_2^2 + \cdots + y_n^2\right) = z_1^2 + z_2^2 + \cdots + z_n^2, \tag{1}$$

where $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, z_1, z_2, \ldots, z_n \in \mathbb{Z}$ and $n \in \mathbb{N}$? Now, since

$$x_1^2 y_1^2 = (x_1 y_1)^2,$$

it is trivial that equation (1) holds when $n = 1$, and as early as the third century, Diophantus was aware that it is also true when $n = 2$. He observed that the number 65 could be written as two different sums of integer squares, namely $16 + 49$ and $64 + 1$, since it is itself the product of two sums of two squares, namely 13×5 or $(3^2 + 2^2)(2^2 + 1^2)$. The formula that he implicitly used would today be written as

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) = (x_1 y_1 \mp x_2 y_2)^2 + (x_2 y_1 \pm x_1 y_2)^2,$$

which is equation (1) when $n = 2$.

Math. Mag. **95** (2022) 422–436. doi:10.1080/0025570X.2022.2125254 © Mathematical Association of America
MSC: Primary 11E25, Secondary 11R52

*A more detailed discussion is contained in a recent paper by the authors [11].

By the 17th century, it was realized that no extension of equation (1) for the $n = 3$ case would be possible. For example, as the French mathematician Albert Girard noticed in 1625, $3 = 1^2 + 1^2 + 1^2$ and $13 = 3^2 + 2^2 + 0^2$ both have three-square sum representations, but their product 39 does not.

The next major step came with Euler, who in 1748 announced that he had derived an expression for equation (1) when $n = 4$, namely:

$$\begin{aligned}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2\right)&\left(y_1^2 + y_2^2 + y_3^2 + y_4^2\right)\\ &= (x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4)^2\\ &+ (x_2y_1 + x_1y_2 - x_4y_3 + x_3y_4)^2\\ &+ (x_3y_1 + x_4y_2 + x_1y_3 - x_2y_4)^2\\ &+ (x_4y_1 - x_3y_2 + x_2y_3 + x_1y_4)^2,\end{aligned} \tag{2}$$

By now it was at least intuitively clear that the dimensions of any further extensions were likely to be of the form $n = 2^m$. Corroboration came seventy years after Euler's result, when a relatively unknown Danish mathematician, Carl Ferdinand Degen, managed to extend it still further, proving the $n = 8$ case. The product

$$\left(x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 + x_7^2 + x_8^2\right) \times \\ \left(y_1^2 + y_2^2 + y_3^2 + y_4^2 + y_5^2 + y_6^2 + y_7^2 + y_8^2\right)$$

is equal to

$$\begin{aligned}&(x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4 - x_5y_5 - x_6y_6 - x_7y_7 - x_8y_8)^2\\ +&(x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3 + x_5y_6 - x_6y_5 - x_7y_8 + x_8y_7)^2\\ +&(x_1y_3 - x_2y_4 + x_3y_1 + x_4y_2 + x_5y_7 + x_6y_8 - x_7y_5 - x_8y_6)^2\\ +&(x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1 + x_5y_8 - x_6y_7 + x_7y_6 - x_8y_5)^2\\ +&(x_1y_5 - x_2y_6 - x_3y_7 - x_4y_8 + x_5y_1 + x_6y_2 + x_7y_3 + x_8y_4)^2\\ +&(x_1y_6 + x_2y_5 - x_3y_8 + x_4y_7 - x_5y_2 + x_6y_1 - x_7y_4 + x_8y_3)^2\\ +&(x_1y_7 + x_2y_8 + x_3y_5 - x_4y_6 - x_5y_3 + x_6y_4 + x_7y_1 - x_8y_2)^2\\ +&(x_1y_8 - x_2y_7 - x_3y_6 + x_4y_5 - x_5y_4 - x_6y_3 + x_7y_2 + x_8y_1)^2.\end{aligned} \tag{3}$$

This eight-squares formula was rediscovered independently a quarter of a century later—in a completely different mathematical context—when in 1843 an Irish mathematician by the name of John Thomas Graves created a new system of hypercomplex numbers. He had been inspired by the recent work of his friend, William Rowan Hamilton, who earlier that year had created the four-dimensional extension of complex numbers [**9**, pp. 106–110], known as the *quaternions*:*

$$\mathbb{H} = \{x_1 + x_2i + x_3j + x_4k : x_1, x_2, x_3, x_4 \in \mathbb{R}, i^2 = j^2 = k^2 = ijk = -1\}.$$

Because of the fundamental equation connecting the three imaginary quantities i, j, k, this new algebra turned out to be noncommutative with respect to multiplication since, for example, $ij = k$ but $ji = -k$.

Furthermore, by analogy with the algebra of complex numbers, letting the conjugate of a quaternion

$$z_1 = x_1 + x_2i + x_3j + x_4k$$

*We explain precisely why Hamilton was unable to come up with a three-dimensional extension of $\mathbb{C}$ in a different article [**12**].

be simply

$$\overline{z_1} = x_1 - x_2 i - x_3 j - x_4 k,$$

and defining the *norm* function $N(z) = z \cdot \overline{z}$, Hamilton found that

$$N(z_1) = x_1^2 + x_2^2 + x_3^2 + x_4^2,$$

and that, for all $z_1, z_2 \in \mathbb{H}$, we have

$$N(z_1)N(z_2) = N(z_1 z_2). \tag{4}$$

In other words, he had discovered that $\mathbb{H}$ was a new kind of structure called a *normed division algebra* over the real numbers, adding to the two that were previously known, namely $\mathbb{R}$ and $\mathbb{C}$. Now, you may be familiar with real normed algebras, but it does not hurt to review some definitions. So here we go.

An *n-dimensional real algebra* $\mathcal{A}$ is an n-dimensional vector space over the real numbers $\mathbb{R}$ equipped with a multiplication that is left- and right-distributive over vector addition, satisfying $(av)(bw) = (ab)(vw)$ for all real numbers a and b and all vectors $v, w \in \mathcal{A}$. We call $\mathcal{A}$ a *division algebra* if every nonzero $v \in \mathcal{A}$ has both a left- and a right-multiplicative inverse. Finally, $\mathcal{A}$ is a *real normed algebra* if there exists a mapping N from $\mathcal{A}$ to the nonnegative real numbers such that $N(v) = 0$ if and only if v is the zero vector and, most importantly, for all $v, w \in \mathcal{A}$,

$$N(v)N(w) = N(vw).$$

These criteria were all satisfied by Hamilton's quaternions.

Spurred on by Hamilton's work, Graves came up with an eight-dimensional extension of $\mathbb{R}$, known today as the *octonions* [1]. This new real algebra $\mathbb{O}$ contained numbers of the form

$$z = x_1 + x_2 i_1 + x_3 i_2 + x_4 i_3 + x_5 i_4 + x_6 i_5 + x_7 i_6 + x_8 i_7,$$

where $x_1, x_2, \ldots x_8 \in \mathbb{R}$ and the seven basic imaginary components $i_1, i_2, \ldots, i_7$ were governed by the following rules:

$$\begin{aligned}
&i_1^2 = \cdots = i_7^2 = -1,\\
&i_\alpha i_\beta = -i_\beta i_\alpha,\\
&i_\alpha i_\beta = i_\gamma \implies i_{\alpha+1} i_{\beta+1} = i_{\gamma+1},\\
&i_\alpha i_\beta = i_\gamma \implies i_{2\alpha} i_{2\beta} = i_{2\gamma},
\end{aligned}$$

with all subscripts belonging to $\{1, 2, 3, 4, 5, 6, 7\}$.

Again, due to the second of the above equations, multiplication in this new algebra is noncommutative. But it also turned out to be nonassociative, since in general

$$i_\alpha(i_\beta i_\gamma) \neq (i_\alpha i_\beta) i_\gamma.$$

Furthermore, letting the conjugate of an octonion z_1 be

$$\overline{z_1} = x_1 - x_2 i_1 - x_3 i_2 - x_4 i_3 - x_5 i_4 - x_6 i_5 - x_7 i_6 - x_8 i_7$$

resulted in the norm function

$$N(z_1) = z_1 \cdot \overline{z_1} = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 + x_7^2 + x_8^2.$$

Now, for $\mathbb{O}$ to be a normed division algebra, equation (4) needed to hold for all $z_i \in \mathbb{O}$. In fact, not only was this the case, but if

$$z_1 = x_1 + x_2 i_1 + x_3 i_2 + x_4 i_3 + x_5 i_4 + x_6 i_5 + x_7 i_6 + x_8 i_7$$

and

$$z_2 = y_1 + y_2 i_1 + y_3 i_2 + y_4 i_3 + y_5 i_4 + y_6 i_5 + y_7 i_6 + y_8 i_7,$$

then equation (4) produced Degen's eight-squares formula (3), which is exactly how Graves was led to this result in the first place.

But Graves was not the only mathematician inspired by Hamilton's discovery of quaternion algebra. Around the same time, and completely independently, the English mathematician Arthur Cayley created an identical system of eight-dimensional algebra which he published in 1845 [2]. Graves had also written up his work for publication. The problem was that he had entrusted his manuscript to the care of his friend, Hamilton, whose memory and organization were not, perhaps, as good as they could have been. The consequence was that, although Graves' work was eventually published [9, pp. 648–656], Cayley's octonions appeared first, winning him much of the credit for their discovery. Indeed, for many years octonions were better known as *Cayley numbers*.

It was quickly realized that the existence of the Cayley-Graves eight-dimensional normed algebra $\mathbb{O}$ is a necessary and sufficient condition for Degen's eight-squares formula, that Hamilton's four-dimensional quaternions $\mathbb{H}$ had the same relationship to Euler's four-squares formula, and that the complex numbers $\mathbb{C}$ and the two-squares formula were similarly codependent. Therefore, the product formula (1) for sums of n squares would hold if and only if a corresponding n-dimensional normed algebra over the real numbers could be found. But did any more of these algebras exist? As early as the 1840s, Cayley and others began to believe that the answer was no. But it was not until 1898 that the German mathematician Adolf Hurwitz managed to prove it [10]. Our explanation of his proof will come later, but first let us convince ourselves that the only possible normed algebras over the reals are indeed $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$, and $\mathbb{O}$.

The Cayley-Dickson construction

In 1919, the American mathematician Leonard Eugene Dickson published a noteworthy paper entitled "On Quaternions and Their Generalization and the History of the Eight Square Theorem" [7]. In it, he explained an ingenious method he had devised[*] that could not only construct the normed algebra $\mathbb{C}$ from $\mathbb{R}$, but could also build $\mathbb{H}$ from $\mathbb{C}$, and $\mathbb{O}$ from $\mathbb{H}$, all using exactly the same procedure.

That procedure was based on an idea first published by Hamilton in 1835 [8]. Consider two real numbers, x and y. Let the ordered pair (x, y) denote the complex number $z = x + yi$, and define its conjugate $\overline{z}$ to be $x - yi$, or $(x, -y)$. Hamilton defined multiplication in $\mathbb{C}$ to be equivalent to

$$(x_1, y_1)(x_2, y_2) = (x_1 x_2 - y_2 y_1,\ y_2 x_1 + y_1 x_2).$$

Since addition was defined merely as the addition of corresponding components, and the multiplicative inverse (or division) operation was easily proved to be

$$z^{-1} = \frac{\overline{z}}{N(z)},$$

[*] Dickson's idea had first appeared in a paper on linear algebras in 1912 [5, pp. 72–73] and was explained in more detail in a book on the subject in 1914 [6, pp. 15–16].

it was clear that Hamilton had devised a new method of obtaining the two-dimensional algebra of $\mathbb{C}$ from the one-dimensional algebra of $\mathbb{R}$.

What Dickson now did was to modify Hamilton's definition of multiplication slightly, so that it became:

$$(x_1, y_1)(x_2, y_2) = (x_1x_2 - \overline{y_2}y_1, y_2x_1 + y_1\overline{x_2}). \tag{5}$$

This clearly had no effect on the construction of $\mathbb{C}$ from $\mathbb{R}$ since for all $x \in \mathbb{R}$, we have $\overline{x} = x$. However, it did result in a remarkable generalization of Hamilton's method which, given an n-dimensional algebra $\mathcal{A}$, enabled the immediate construction of an algebraic extension of $\mathcal{A}$ with dimension $2n$. In effect this amounted to defining the set $\mathbb{C}$ to be equal to $\mathbb{R} + \mathbb{R}i$, where $i^2 = -1$. Similarly, for quaternions, $\mathbb{H}$ could be defined as $\mathbb{C} + \mathbb{C}j$, where

$$i^2 = j^2 = k^2 = ijk = -1,$$

since it can be easily shown that, if $x_1, x_2, y_1, y_2 \in \mathbb{R}$, then we have

$$(x_1 + x_2i) + (y_1 + y_2i)j = x_1 + x_2i + y_1j + y_2k \in \mathbb{H},$$

and if $x_1, x_2, y_1, y_2 \in \mathbb{C}$, then

$$(x_1 + y_1j)(x_2 + y_2j) = (x_1x_2 - \overline{y_2}y_1) + (y_2x_1 + y_1\overline{x_2})j \in \mathbb{H}.$$

Today, this process of successively building algebras of dimension 2^n is known as the *Cayley-Dickson construction* [1, 4]. But, given that it is a generalization of a method due to Hamilton, why is the name of *Cayley* attached to it?

One possible reason lies in the groundbreaking paper of 1858 that introduced the algebra of matrices to the world [3]. In this paper, Cayley put forward the idea of representing quaternions in terms of linear combinations of 2×2 matrices

$$\mathbf{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

In addition to satisfying the fundamental equations for the base elements in $\mathbb{H}$, such as

$$\mathbf{I}^2 = \mathbf{J}^2 = \mathbf{K}^2 = \mathbf{IJK} = -\mathbf{1}$$

these matrices could be used to represent any quaternion $x_1 + x_2i + x_3j + x_4k$ as

$$\begin{aligned} x_1\mathbf{1} + x_2\mathbf{I} + x_3\mathbf{J} + x_4\mathbf{K} &= x_1\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + x_2\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} + x_3\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + x_4\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} \\ &= \begin{bmatrix} x_1 + ix_2 & x_3 + ix_4 \\ -x_3 + ix_4 & x_1 - ix_2 \end{bmatrix} \\ &= \begin{bmatrix} z_1 & w_1 \\ -\overline{w_1} & \overline{z_1} \end{bmatrix}, \end{aligned}$$

where $z_1 = x_1 + ix_2$ and $w_1 = x_3 + ix_4 \in \mathbb{C}$. Neatly, but not at all coincidentally, we have

$$\det\begin{bmatrix} x_1 + ix_2 & x_3 + ix_4 \\ -x_3 + ix_4 & x_1 - ix_2 \end{bmatrix} = x_1^2 + x_2^2 + x_3^2 + x_4^2,$$

showing that the determinant of this matrix representation of a quaternion is simply equal to its norm.

Standard results from linear algebra could then be used to derive crucial properties of $\mathbb{H}$. For example, the associative, distributive, and noncommutative properties of quaternion multiplication followed immediately from the corresponding attributes of matrices. Also, the fact that for any 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$,

$$A^{-1} = \frac{1}{\det A} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

thus meant that any quaternion z with a nonzero norm would have the inverse

$$z^{-1} = \frac{\overline{z}}{N(z)}.$$

And importantly, since for any $n \times n$ matrices A and B, $\det AB = \det A \det B$, the fundamental equation (4) immediately followed.

Most crucially, as Dickson would have realized, if $x_1, x_2, y_1, y_2 \in \mathbb{C}$, then multiplication in $\mathbb{H}$ is defined by the matrix product

$$\begin{bmatrix} x_1 & y_1 \\ -\overline{y_1} & \overline{x_1} \end{bmatrix} \begin{bmatrix} x_2 & y_2 \\ -\overline{y_2} & \overline{x_2} \end{bmatrix} = \begin{bmatrix} x_1x_2 - \overline{y_2}y_1 & y_2x_1 + y_1\overline{x_2} \\ -\overline{y_2x_1 + y_1\overline{x_2}} & \overline{x_1x_2 + \overline{y_2}y_1} \end{bmatrix},$$

which is exactly his equation (5). So perhaps this applicability of matrices to the generation of $\mathbb{H}$ from $\mathbb{C}$ is one reason for the association of Cayley's name with this procedure.

On the other hand, as Dickson himself noted, the Cayley-Dickson construction arose in direct response to Cayley's foundational work on octonions from 1845. Dickson realized that by taking $\mathbb{H}$, re-labeling $i = i_1$, $j = i_2$, $k = i_3$, and defining i_4 to be a new square-root of -1 such that

$$i_1i_4 = i_5, i_2i_4 = i_6 \quad \text{and} \quad i_3i_4 = i_7,$$

he could define the octonions $\mathbb{O}$ as $\mathbb{H} + \mathbb{H}i_4$. Thus, if $x_1, x_2, y_1, y_2 \in \mathbb{H}$, then the product

$$(x_1 + y_1i_4)(x_2 + y_2i_4) = (x_1x_2 - \overline{y_2}y_1) + (y_2x_1 + y_1\overline{x_2})i_4$$

would define all multiplication in $\mathbb{O}$.

There is no need to stop there, however, and the Cayley-Dickson construction can be used to build further composition algebras of increasing dimension: 16, 32, 64, and so on *ad infinitum*. But there is a problem. As the dimension doubles, key algebraic properties are successively lost, the first of which is trivial conjugation, which clearly holds in $\mathbb{R}$, but not in $\mathbb{C}$. This has a knock-on effect, resulting in the loss of commutativity of multiplication when moving from $\mathbb{C}$ to $\mathbb{H}$, which likewise results in the nonassociativity of $\mathbb{O}$. But things get worse, because at dimension 16 the next property to be lost is the nonexistence of zero divisors. In other words, from this point on, for any algebra $\mathcal{A}$ produced by the Cayley-Dickson construction, any element $x \in \mathcal{A}$ could have a nonzero partner $y \in \mathcal{A}$ such that $xy = 0$. This naturally has the consequence that the crucial equation (4) no longer holds in general, meaning that no composition algebras over the reals with dimension 2^n will be normed division algebras if $n > 3$.

This gives us an intuitive idea why the only possible normed algebras over the reals are indeed $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$, and $\mathbb{O}$. But it does not prove it. Nor does it say anything about the existence (or nonexistence) of normed algebras with dimensions other than powers of two. To resolve the matter conclusively, we need to return to sums of squares and look at Hurwitz's 1898 proof. Our exposition will be based on an expanded version given by Dickson in his paper of 1919, where he explained the nature of, and the rationale for, his presentation of Hurwitz's theorem [7, p. 159]:

> Since experience shows that graduate students fail to follow various steps merely outlined by Hurwitz, we shall here give the proof in detailed, amplified form.

But Dickson's exposition is also lacking in certain respects: it gives quite lengthy explanations of some relatively easy portions, while skipping over the more sophisticated details of others. And it would certainly not be fully intelligible to today's undergraduates. We therefore give our own step-by-step, and hopefully intelligible, elucidation of Dickson's expanded proof of Hurwitz's theorem.

Our expanded proof of Dickson's expanded proof of Hurwitz's theorem

Theorem 1 (Hurwitz's Sums of Squares Theorem)**.** *Let n be a positive integer for which there exists an identity of the form* (1)

$$(x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2) = z_1^2 + \cdots + z_n^2,$$

where

$$z_k = \sum_{i,j=1}^{n} A_{ijk} x_i y_j,$$

and the A_{ijk} are constants independent of the values of the x_i and the y_j. Then $n =$ 1, 2, 4, or 8 and no other values.

Proof. Our proof is in six steps:

- By looking at quadratic forms and their associated matrices, we establish that our identity (1) is dependent on the existence of an $n \times n$ matrix A such that

$$A^T A = (x_1^2 + x_2^2 + \cdots + x_n^2) I_n.$$

- This matrix A can be written as a linear combination of matrices $A_1, A_2, \ldots A_n$, in which the entries are 0, 1, or -1, and for which $A_i^T A_i = I_n$ for $i = 1, \ldots, n$. Defining new matrices $B_i = A_n^T A_i$ for $i = 1, \ldots, n-1$, we produce an equivalent expression for $A^T A$ in terms of the B_is, which enables us to derive the following relations:

$$\text{(a) } B_i^T B_i = I_n,$$

$$\text{(b) } B_i^T B_j + B_j^T B_i = 0 \text{ for } i \neq j,$$

$$\text{(c) } B_i + B_i^T = 0.$$

- From these equations, we deduce that the B_is are skew-symmetric matrices. This property leads to the important fact that in formula (1), if $n > 1$, n cannot be odd.
- We next prove that at least half of the matrices in the set of 2^{n-1} products of the various B_is are linearly independent.

- This results in the inequality $2^{n-2} \leq n^2$, which we then prove can only hold if $n = 1, 2, 4, 6$, or 8.
- Finally, we eliminate the $n = 6$ case.

Quadratic forms and their matrices An *n-ary quadratic form* is a homogeneous polynomial of degree two in n variables. Quadratic forms with 2, 3, and 4 variables are called *binary, ternary* and *quaternary* quadratic forms, respectively. Thus, $4z_1^2 + z_1 z_2 + 6z_2^2$ is a binary quadratic form and $z_1^2 + z_2^2 + z_3^2 + z_4^2$ is a quaternary quadratic form.

Consider the quadratic form

$$\begin{aligned} F(z) = \sum_{i,j=1}^{n} b_{ij} z_i z_j &= b_{11} z_1^2 + b_{12} z_1 z_2 + \cdots + b_{1n} z_1 z_n \\ &+ b_{21} z_2 z_1 + b_{22} z_2^2 + \cdots + b_{2n} z_2 z_n \\ &+ \cdots \\ &+ b_{n1} z_n z_1 + b_{n2} z_n z_2 + \cdots + b_{nn} z_n^2. \end{aligned} \tag{6}$$

The matrix of this quadratic form is $B = [b_{ij}]$, and in the special case when $B = I_n$, we have that

$$F(z) = z_1^2 + z_2^2 + \cdots + z_n^2.$$

If we now substitute

$$z_i = a_{i1} y_1 + a_{i2} y_2 + \cdots + a_{in} y_n,$$

where the a_{ij} are all scalars, then equation (6) becomes

$$\begin{aligned} G(y) = \sum_{i,j=1}^{n} m_{ij} y_i y_j &= m_{11} y_1^2 + m_{12} y_1 y_2 + \cdots + m_{1n} y_1 y_n \\ &+ m_{21} y_2 y_1 + m_{22} y_2^2 + \cdots + m_{2n} y_2 y_n \\ &+ \cdots \\ &+ m_{n1} y_n y_1 + m_{n2} y_n y_2 + \cdots + m_{nn} y_n^2, \end{aligned} \tag{7}$$

where the m_{ij} are various linear combinations of the a_{ij} and b_{ij}. The matrix of this quadratic form $G(y)$ is $A^T BA$, where $A = [a_{ij}]$. Hence, if $B = I_n$, then the matrix of the resulting quadratic form is equal to $A^T I_n A = A^T A$.

Now let

$$M = [m_{ij}] = A^T A.$$

Once again, if $M = I_n$, then equation (7) becomes

$$G(y) = y_1^2 + y_2^2 + \cdots + y_n^2.$$

Thus, if we want

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} z_i^2,$$

this will hold provided $m_{ij} = 1$ if $i = j$ and 0 otherwise, i.e., if $M = I_n$. Now suppose we want

$$x_1^2 \sum_{i=1}^n y_i^2 = \sum_{i=1}^n z_i^2,$$

where x_1^2 is a scalar. This will hold provided $m_{ij} = x_1^2$ if $i = j$ and 0 otherwise, i.e., if $M = x_1^2 \cdot I_n$.

In a similar way, let x_k^2 be a scalar for $k = 1, 2, \ldots, n$. Then

$$x_1^2 + x_2^2 + \cdots + x_n^2$$

is also a scalar and it follows that

$$(x_1^2 + x_2^2 + \cdots + x_n^2) \sum_{i=1}^n y_i^2 = \sum_{i=1}^n z_i^2$$

provided

$$m_{ij} = x_1^2 + x_2^2 + \cdots + x_n^2$$

if $i = j$ and 0 otherwise, that is, if

$$M = A^T A = (x_1^2 + x_2^2 + \cdots + x_n^2) I_n.$$

Hence, the existence of the identity

$$\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i^2 = \sum_{i=1}^n z_i^2,$$

i.e., equation (1), is equivalent to the existence of the equality

$$A^T A = (x_1^2 + x_2^2 + \cdots + x_n^2) I_n. \tag{8}$$

For the case $n = 4$, we have

$$(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) = z_1^2 + z_2^2 + z_3^2 + z_4^2,$$

where

$$\begin{aligned} z_1 &= x_1 y_1 - x_2 y_2 - x_3 y_3 - x_4 y_4, \\ z_2 &= x_2 y_1 + x_1 y_2 - x_4 y_3 + x_3 y_4, \\ z_3 &= x_3 y_1 + x_4 y_2 + x_1 y_3 - x_2 y_4, \\ z_4 &= x_4 y_1 - x_3 y_2 + x_2 y_3 + x_1 y_4, \end{aligned} \tag{9}$$

and the relevant matrix A is given by*

$$A = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 \\ x_2 & x_1 & -x_4 & x_3 \\ x_3 & x_4 & x_1 & -x_2 \\ x_4 & -x_3 & x_2 & x_1 \end{bmatrix}.$$

*The equivalence of formulae (2) and (9) is not coincidental!

The matrices A_i and relations among them Notice that the matrix A can be written as the linear combination

$$A = x_1 A_1 + \cdots + x_n A_n, \tag{10}$$

where the entries of the A_i are 0, 1, or -1. In our example above, for instance, we may write A as

$$x_1 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + x_2 \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

where the A_i's are the matrices with real coefficients x_i.

Substituting the linear combinations of type (10) for A and A^T into equation (8) leads to

$$(x_1 A_1^T + \cdots + x_n A_n^T)(x_1 A_1 + \cdots + x_n A_n) = (x_1^2 + x_2^2 + \cdots + x_n^2) I_n \tag{11}$$

Expanding the left side of the above equation gives us

$$\sum_{i,j=1,\ldots,n} x_i x_j A_i^T A_j = \left(\sum_{i=1}^{n} x_i^2 \right) I_n.$$

The coefficient of $A_i^T A_j$ is $x_i x_j$; since $x_i x_j = x_j x_i$, it follows that

$$A_i^T A_i = I_n \text{ for } i = 1, \ldots, n, \text{ and } A_i^T A_j + A_j^T A_i = 0.$$

But what about the individual A_i's? Hurwitz used the equalities

$$A_i^T A_i = I_n = A_i A_i^T$$

to make a change of variables that replaces the terms $x_n A_n$ and $x_n A_n^T$ by the term $x_n I_n$. He did this by defining new matrices $B_i = A_n^T A_i$ for $i = 1, \ldots, n-1$.

Then for $i < n$, $A_i A_i^T = I_n$ implies

$$A_i = A_n A_n^T A_i = A_n B_i,$$

and so

$$A_i^T = B_i^T A_n^T.$$

When we make these substitutions in equation (11) and un-distribute the factors A_n^T and A_n, the result is as follows:

$$\begin{aligned}(x_1^2+\cdots+x_n^2)I_n &= \left(\sum_{i=1}^{n-1} x_iA_i^T + x_nA_n^T\right)\left(\sum_{i=1}^{n-1} x_iA_i + x_nA_n\right)\\ &= \left(\sum_{i=1}^{n-1} x_iB_i^T + x_nI_n\right)A_n^TA_n\left(\sum_{i=1}^{n-1} x_iB_i + x_nI_n\right) \qquad (12)\\ &= \left(\sum_{i=1}^{n-1} x_iB_i^T + x_nI_n\right)\left(\sum_{i=1}^{n-1} x_iB_i + x_nI_n\right).\end{aligned}$$

This time, the coefficients of x_ix_j are I_n if $i = j$ and 0 otherwise, and the coefficients of x_ix_n are 0 for $i < n$. Hence, for $i, j = 1, \ldots, n-1$, we have that

(a) $B_i^TB_i = I_n$,

(b) $B_i^TB_j + B_j^TB_i = 0$ for $i \neq j$,

(c) $B_i + B_i^T = 0$.

From (a) and (c) we deduce that

$$B_i^2 = -B_i^TB_i = -I_n.$$

Using (b) and (c), we deduce that

$$B_i^TB_j + B_j^TB_i = -B_iB_j - B_jB_i = 0,$$

and hence $B_iB_j = -B_jB_i$.[*]

Symmetric and skew-symmetric matrices We now need two definitions. A square matrix M is *symmetric* provided that $M^T = M$, and *skew-symmetric* provided $M^T = -M$. Thus, $B_i^T = -B_i$ and

$$(B_iB_j)^T = B_j^TB_i^T = (-1)^2B_jB_i = B_jB_i = -B_iB_j.$$

Hence, the B_i and the B_iB_j are skew-symmetric. As for a triple product,

$$\begin{aligned}(B_iB_jB_k)^T &= B_k^TB_j^TB_i^T = (-1)^3(B_kB_jB_i)\\ &= (-1)^3(-1^3)B_iB_jB_k = B_iB_jB_k\end{aligned}$$

because the three substitutions $B_r^T = -B_r$ contribute a factor of $(-1)^3$ to the product. In addition, it takes three pairwise adjacent swaps to reverse $B_kB_jB_i$. Each such swap contributes a factor of -1, and so the "swapping" contributes $(-1)^3$ to the product. Hence,

$$(B_iB_jB_k)^T = (-1)^{3+3}B_iB_jB_k = B_iB_jB_k,$$

and so a product of three distinct B_i's is symmetric.

More generally, it takes $\binom{r}{2}$ such swaps to reverse a product of r distinct factors, and there is a straightforward inductive proof of this fact. Thus, if

$$1 \le i_1 < i_2 < \cdots < i_r \le n-1,$$

[*] Given any distinct B_i, B_j, B_k, the fact that $B_i^2 = B_j^2 = B_k^2 = B_iB_jB_k = -I_n$ is not a coincidence either!

then

$$(B_{i_1}B_{i_2}\cdots B_{i_r})^T = B_{i_r}^T\cdots B_2^T B_1^T$$
$$= (-1)^r B_{i_r}\cdots B_2 B_1 = (-1)^{r+\binom{r}{2}} B_{i_1}B_{i_2}\cdots B_{i_r}.$$

Now,

$$r + \binom{r}{2} = \binom{r+1}{2} = \frac{(r+1)r}{2},$$

and this number is even if $r \equiv 0, 3 \mod 4$, and odd if $r \equiv 1, 2 \mod 4$. Thus, a product of $1, 2, 5, 6, \ldots$ B_i's is skew-symmetric, and a product of $3, 4, 7, 8, \ldots$ B_i's is symmetric—and so is the identity matrix I_n.

The skew-symmetry of the individual B_i's also reveals a highly significant fact. Since the B_i's are $n \times n$ matrices such that $B_i^T = -B_i$, this means that, for $n > 1$, $\det B_i = (-1)^n \det B_i$, meaning that if n is odd, then $\det B_i = 0$. But since $B_i^2 = -I_n$, we know that $\det B_i \neq 0$. Therefore, apart from the trivial case when $n = 1$, n cannot be odd. And in what follows, that fact is going to be crucial.

We are now ready to state and prove a key result.

Linear independence of a set of 2^{n-2} $n \times n$ matrices.

Proposition 1. *At least half of the 2^{n-1} matrices in*

$$S = \{I, B_{i_1}, B_{i_1}B_{i_2}, \ldots B_{i_1}B_{i_2}\ldots B_{i_r} | 1 \le i_1 < i_2 < \cdots < i_r \le n-1\}$$

form a linearly independent set.

Proof. First note that $|S| = 2^{n-1}$ since any element of S either does or does not contain $B_1, B_2, \ldots, B_{n-1}$. Let R be any linear combination of the matrices in S involving constants $a_1, a_2, \ldots$—not all zero—such that $R = 0$. Clearly, this means that the matrices in R are linearly dependent. We call this linear dependency *irreducible* if R cannot be written as $R_1 + R_2$, where $R_1 = 0$, $R_2 = 0$, and no element of S belongs to both R_1 and R_2.

In particular, an irreducible linear combination R cannot contain both symmetric and skew-symmetric matrices—and this is the key to proving this proposition. To see this, suppose the contrary. This would mean that $R = 0$ could be rewritten as $M = K$, where M and K are linear combinations of only symmetric and skew-symmetric matrices, respectively. Thus,

$$M = M^T = K^T = -K = -M,$$

implying that $M = -M = 0$, and therefore that $K = 0$.

Next, we show that multiplication by any number of matrices B_i permutes the 2^{n-1} members of

$$S = \{I, B_{i_1}, B_{i_1}B_{i_2}, \ldots, B_{i_1}B_{i_2}\ldots B_{i_r} | i_1 < i_2 < \cdots < i_r \le n-1\}$$

if we ignore sign changes and orderings. The reason is that $B_i^2 = -I_n$ so that B_i is an involution (up to sign) on the set S. Multiplication by B_i permutes the products in S by mapping those products in S that contain the factor B_i to the products in S that do not have the factor B_i and vice-versa. By induction (and ignoring signs and orderings), multiplication by a product of matrices in S is a permutation of the elements of S.

We have already proved that for $n > 1$, n cannot be odd, so we now assume that n is even. An irreducible linear dependency

$$R = \sum_{i=1}^{r} a_i S_i = 0,$$

where S_i is a product of the B_js and the a_i are nonzero, can be rewritten as

$$I_n = \sum c_i T_i$$

by multiplying $R = 0$ through by the inverse of one term and re-arranging the equation.* I_n is symmetric, and so the T_i must also be symmetric. Thus $T_i = B_{i_1} \dots B_{i_j}$ is a product of j factors, with $j \equiv 0$ or $3 \mod 4$.

If T_i has $4k$ factors, multiplying $I_n = \sum c_i T_i$ by B_{i_1} yields the skew-symmetric B_{i_1} on the left-hand side and a symmetric term on the right. Thus, no T_i can be a product of $4k$ terms.

If T_i has $4k + 3$ factors and this number is less than $n - 1$, then multiplying $I_n = \sum c_i T_i$ by a nonfactor B_j yields the skew-symmetric B_j on the left-hand side and a symmetric term on the right. Thus, no T_i can be a product of $4k + 3$ terms with $4k + 3 < n - 1$.

So the only possible linear dependency must be of the form $I_n = cB_1 \dots B_{n-1}$ for some constant c, where $n - 1 \equiv 3 \mod 4$. Hence, $n \equiv 0 \mod 4$. As for c, we know that

$$I_n^2 = c^2 (B_1 \dots B_{n-1})^2.$$

As we have seen, it takes $\binom{r}{2}$ adjacent swaps to reverse a product of r distinct factors. Hence, since $n \equiv 0 \mod 4$, we see that

$$\begin{aligned}(B_1 \dots B_{n-1})^2 &= (-1)^{\binom{n-1}{2}} (B_1 \dots B_{n-1})(B_{n-1} \dots B_1)\\ &= (-1)^{\binom{n-1}{2}+n-1} I_n^{n-1}\\ &= (-1)^{\binom{n}{2}} I_n = I_n.\end{aligned}$$

Therefore, $I_n = c^2 \cdot I_n$, and so $c = \pm 1$. We thus see that the only possible linear dependencies are of the form $I_n = (\pm 1)(B_1 \dots B_{n-1})$—where $n \equiv 0 \mod 4$—and those obtained by multiplying this equation by matrices from the set S. These dependencies take the form of equalities between two products, one of which contains more than half of the B_i's and the other, fewer than half. This means that if $n \equiv 0 \mod 4$, then half of the 2^{n-1} matrices in the set S will be linearly independent.

Finally, we note that if $n \equiv 2 \mod 4$, then there can be no linear dependencies, implying that all 2^{n-1} of the products in S will be linearly independent.

Thus, the set S contains at least 2^{n-2} linearly independent $n \times n$ matrices. ■

We are almost finished ...

The (1, 2, 4, 6, 8) restriction.

Proposition 2. *If n satisfies the conditions of Proposition 1, then $n = 1, 2, 4, 6,$ or 8.*

*Notice that this new relation $I_n = \sum c_i T_i$ is also irreducible since we can multiply it through by a suitable constant c_i and matrix B_j to return us to our original irreducible $R = 0$.

Proof. The n^2 distinct $n \times n$ matrices whose entries are all zeros except for a single 1 are clearly linearly independent. The 2^{n-2} linearly independent matrices from the set S are a subset of these n^2 matrices. Hence $2^{n-2} \leq n^2$. For $n \leq 8$, this inequality is true by inspection, and we show by induction that the inequality is false if $n > 8$.

For $n = 9$, we have that

$$2^{n-2} = 2^7 = 128 > 81 = 9^2 = n^2.$$

Thus, for $n = 9$ the inequality is false.

Next, assume that $n \geq 9$ and that $2^{n-2} > n^2$. Then

$$2^{n-1} = 2 \cdot 2^{n-2} > 2 \cdot n^2 = n^2 + n^2 > n^2 + 2n + 1 = (n+1)^2,$$

because $n^2 > 2n + 1$ whenever $n \geq 3$. It follows that $2^{n-2} > n^2$ for all $n \geq 9$.

Since we know that for $n > 1$, n cannot be odd, this leaves only the possibilities 1, 2, 4, 6, and 8. ■

We have now proved that the existence of the sums-of-n-squares identity (1) is equivalent to the existence of the identity (8), which itself is equivalent to equation (12). This last equation holds if and only if there exist skew-symmetric matrices B_i such that 2^{n-2} of their 2^{n-1} possible products are linearly independent. This is equivalent to saying that $2^{n-2} \leq n^2$, which is only true if $n = 1, 2, 4, 6$, or 8. There is thus only one thing now left to do.

Elimination of the $n = 6$ case. We finally eliminate $n = 6$. To begin with, $6 \equiv 2$ mod 4, so that if a sums-of-six-squares identity exists, then all $32 = 2^5 = 2^{6-1}$ of the relevant matrices form a linearly independent set. Sixteen of these matrices—namely, the five B_i, the ten B_iB_j, and the product $B_1B_2B_3B_4B_5$—are skew-symmetric. Call these matrices $M_1, M_2, \dots, M_{16}$, and let a_{ij}^k be the (i, j)th entry of M_k.

Suppose there exist constants $c_1, c_2, \dots, c_{16}$—not all zero—such that

$$c_1M_1 + c_2M_2 + \cdots + c_{16}M_{16} = 0.$$

Then for each of the 36 pairs (i, j) and each k,

$$c_1a_{ij}^1 + c_2a_{ij}^2 + \cdots + c_{16}a_{ij}^{16} = 0.$$

However, the M_k are skew-symmetric. This has two consequences:

1. For all i and k,

$$a_{ii}^k = -a_{ii}^k = 0,$$

so the six linear combinations when $i = j$ are identically zero.

2. For all i, j, k with $i = j$,

$$a_{ij}^k = -a_{ji}^k,$$

so the 15 linear combinations for $i < j$ are the negatives of the 15 linear combinations for $i > j$, and so they contribute nothing new.

Therefore, there are only 15 distinct linear equations relating these 16 matrices, and a system of 15 homogeneous linear equations in 16 unknowns—namely, $c_1, c_2, \dots, c_{16}$—has more unknowns than equations ... and so it has nontrivial solutions. In short, the 16 skew-symmetric matrices are linearly dependent, contrary to the assumption that all 32 relevant matrices are linearly independent. Hence, there is no sums-of-six-squares identity.

Thus, the sums-of-n-squares identity (1) exists for $n = 1, 2, 4$, and 8 and for no other positive integers n. This completes the proof of Hurwitz's theorem. ■

REFERENCES

[1] Baez, J. C. (2002). The octonions, *Bull. Amer. Math. Soc.* 39(2): 145–205. doi.org/10.1090/S0273-0979-01-00934-x.

[2] Cayley, A. (1845). On Jacobi's elliptic functions . . . and on quaternions. *Philos. Mag.* 26: 208–211. *Collected Math. Papers* 1: 127. doi.org/10.1017/cbo9780511703676.022.

[3] Cayley, A. (1858). A memoir on the theory of matrices. *Philos. Trans. Royal Soc. London* 148: 17–37. *Collected Math. Papers* 2: 475–496. doi.org/10.1017/cbo9780511703683.053.

[4] Conway, J. H., Smith, D. A. (2003). *On Quaternions and Octonions: Their Geometry, Arithmetic, and Symmetry*. New York: A K Peters/CRC Press.

[5] Dickson, L. E. (1912). Linear algebras. *Trans. Amer. Math. Soc.* 13: 59–73. 10.1090/s0002-9947-1912-1500905-3.

[6] Dickson, L. E. (1914). *Linear Algebras*. Cambridge: Cambridge Univ. Press.

[7] Dickson, L. E. (1919). On quaternions and their generalization and the history of the eight square theorem. *Ann. Math.* 2nd ser. 20(3): 155–171. doi.org/10.2307/1967865.

[8] Hamilton, W. R. (1835). Theory of conjugate functions, or algebraic couples; with a preliminary and elementary essay on algebra as the science of pure time. *Trans. Royal. Irish Acad.* 17: 293–422; *Mathematical Papers* 3: 76–96.

[9] Hamilton, W. R. (1967). *The Mathematical Papers of Sir William Rowan Hamilton*, vol. 3. (Halberstam, H., Ingram, R. E., editors). Cambridge: Cambridge Univ. Press.

[10] Hurwitz, A. (1898). Über die Composition der quadratischen Formen von beliebig vielen Variablen, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. 309–316. *Mathematische Werke* 2: 565–571. doi.org/10.1007/978-3-0348-4160-3_39.

[11] Rice, A., Brown, E. (2016). Commutativity and collinearity: A historical case study of the interconnection of mathematical ideas. Part I. *BSHM Bull. J. Br. Soc. Hist. Math.* 31(1): 1–14. doi.org/10.1080/17498430.2015.1046037.

[12] Rice, A., Brown, E. (2021). Why Hamilton couldn't multiply triples. *Coll. Math. J.*. 52(3): 185–192. doi.org/10.1080/07468342.2021.1897418

Summary. We give a simple proof, intelligible to undergraduates, that a particular multiplicative formula for sums of n squares can only occur when $n = 1, 2, 4$, or 8, a result originally proved by Hurwitz in 1898. We begin with a brief survey of the history of sums of squares, leading to a discussion of the related topic of normed division algebras over the real numbers. This story culminates with a crucial paper by Dickson in 1919 that not only contained an exposition of Hurwitz's 1898 proof, but which also outlined a new process for producing division algebras over the reals. That process, now called the Cayley-Dickson construction, is intimately connected with the product formula for sums of squares and the dimensions necessary for its existence. For this reason, we present an introduction to the Cayley-Dickson construction for beginners, together with a proof of Hurwitz's theorem accessible to anyone with a basic knowledge of undergraduate algebra.

EZRA (BUD) BROWN (MR Author ID: 222489) grew up in New Orleans, has degrees from Rice and LSU, taught at Virginia Tech for 48 years, and retired in 2017 as Alumni Distinguished Professor Emeritus of Mathematics. He has done research in number theory, combinatorics, and expository mathematics—but one of his favorite papers is one he wrote with a sociologist. He and the late Richard Guy are the authors of the Carus Monograph, *The Unity of Combinatorics*, published by the AMS in May 2020.

ADRIAN RICE (MR Author ID: 601492) is the Dorothy and Muscoe Garnett Professor of Mathematics at Randolph-Macon College in Ashland, Virginia, where his research focuses on nineteenth-century and early twentieth-century mathematics. In addition to papers on various aspects of the history of mathematics, his books include *Mathematics Unbound: The Evolution of an International Mathematical Research Community, 1800–1945* (with Karen Hunger Parshall), *Mathematics in Victorian Britain* (with Raymond Flood and Robin Wilson), and most recently *Ada Lovelace: The Making of a Computer Scientist* (with Christopher Hollings and Ursula Martin). He is a five-time recipient of awards for outstanding expository writing from the MAA. In his spare time, he enjoys music, travel, and spending time with his wife and son.

Uncolorable Brunnian Links are Linked

LOUIS H. KAUFFMAN
University of Illinois at Chicago
Chicago, Illinois 60607-7045 and
Novosibirsk State University
Novosibirsk, Russia
kauffman@uic.edu

DEVIKA PRASAD
University of Illinois
Urbana-Champaign
Champaign, IL 61820
devika.prasad10@gmail.com

CLAUDIA J. ZHU
University of Pennsylvania
Philadelphia, PA 19104
jiyunzhu@gmail.com

Knot theory studies how ropes and other one-dimensional objects can be entangled in three-dimensional space. The key problem in studying knots is to understand if the knot can be undone. But if you tie a knot on an open length of rope, you can always undo it by slipping the knot off the end of the rope! For this reason, topologists define a knot to be an entangled closed loop of rope. In other words, a knot is an embedding of a circle in three-dimensional space.

We represent knots by diagrams in the plane. The diagram can be seen as a shadow of the three-dimensional curve projected to the plane. The diagram is equipped with extra information, allowing the reconstruction of a (topologically equivalent) curve in three-space. See Figure 1 for an illustration of a knot in three space and its corresponding diagram.

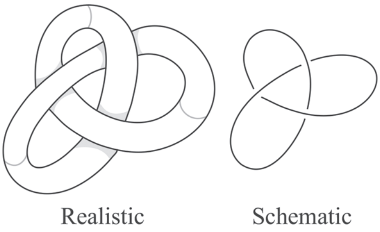


Figure 1 A shaded, three-dimensional representation of a knot (marked as realistic), and a schematic diagram representing the over and under crossings (marked as schematic).

A link is an embedding of a number of disjoint circles and is correspondingly represented by a diagram. This paper focuses on both knots and links, but the first section of this paper will focus on knots. We recommend the reader refer to the books by Kauffman [5, 6], Adams [1], and Crowell and Fox [3] for basic information on knots.

We will generalize a method of Naynes [7] to prove that an infinite collection of Brunnian links are each linked. We begin with an introduction to knots, links, colorings, and quandles.

Math. Mag. **95** (2022) 437–451. doi:10.1080/0025570X.2022.2136462 © Mathematical Association of America
MSC: 57M25

Knot crossings

A *crossing* is where two curve segments in the diagram intersect one another. When one curve crosses another it can weave over or under the second curve. See Figure 2 for the two choices of weaving that are available at each crossing. For a given knot or link, every crossing has a chosen type of weaving.

In the language of graph theory, one can say that the diagram of a knot or link is a 4-regular planar graph with extra structure indicating the weaving at the crossings. Each node of the graph is a crossing with the weaving indicated as in Figure 2.

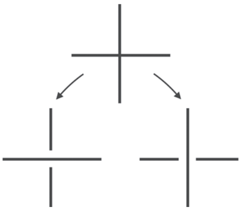


Figure 2 A regular crossing and the over/under choices that are associated with it. The left crossing shows the vertical edge going under the vertex and vice versa for the right crossing.

Any given knot (or link) diagram is a set of instructions for weaving a rope by specifying the crossings and connections. The topological problem about knots and links is to determine whether there is a continuous deformation through embeddings from one knot to another. Such a deformation is called an *ambient isotopy*. In this paper, we use a diagramamtic method to hand the isotopies.

Isotopies and manipulations

Reidemeister and other knot theorists in the early twentieth century were concerned with the ambient isotopy of knots and links [**9**]. Alexander and Briggs [**2**] and Reidemeister [**8**] proved the remarkable theorem that any two knots or links in space are ambient isotopic if and only if their corresponding projection diagrams are isotopic in the sense of the Reidemeister moves (see Figure 4). In this way, Alexander, Briggs, and Reidemeister translated the complex topological problem of classifying knots and links in three-dimensional space to a combinatorial problem of classifying the diagrams of knots and links as they are represented in the plane. We will concentrate on explaining some of the ways to use the Reidemeister moves to show that knots and links are the same or different up to isotopy.

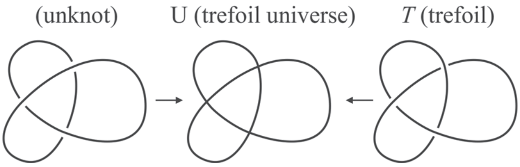


Figure 3 The universe of a knot diagram is the underlying planar graph for that diagram with four edges locally incident to each vertex. The knot diagram includes the over and under crossing information that specifies its weaving.

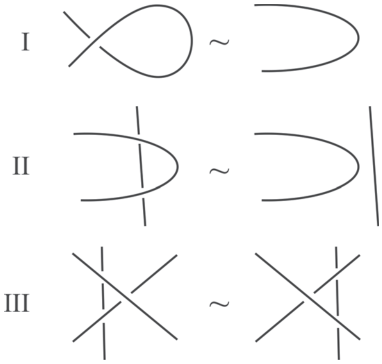


Figure 4 The three Reidemeister moves denoted by the three roman numerals.

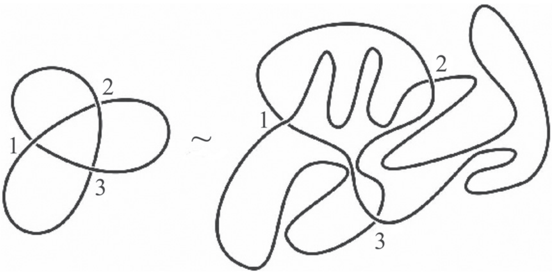


Figure 5 Two planar isotopic versions of the trefoil.

We say that two diagrams K, K' are *planar isotopic* if there is an invertible continuous mapping of the plane that takes one diagram to the other. Such mappings are assumed not to disturb the crossing structure of the diagram, but they can otherwise change the diagram considerably.

The Reidemeister moves are illustrated in Figure 4. The changes caused by these moves occur locally, as shown in the figure. This means that we can only transform the diagram one Reidemeister move at a time. We say two knots are equivalent if we can transform one to the other through some combination of the three moves. This is shown with Reidemeister moves in Figure 10, where we transform the trefoil into a more complicated diagram. It is shown again in Figure 5, where we create a complex diagram using only planar isotopy.

Note that the Reidemeister moves, as illustrated in Figure 4, are representative move types. The single crossing in the illustration of the first move can be switched. The two crossings in the illustration of the second move can also be switched. The crossings shown in the third move can be changed just so long as one arc either overcrosses or undercrosses the other two arcs. The Reidemeister moves apply to links as well.

Knot and link diagrams can be oriented by choosing a direction along each component. See Figure 6 for an illustration of the two orientations possible at a crossing, and Figures 7 and 8 for examples of oriented links. Note that in Figure 6 we have associated a sign of $+1$ or -1 to each oriented crossing. If you place your right hand over a positive oriented crossing so that your fingers point in the direction of the overcrossing arc, then your thumb will point in the direction of the undercrossing arc. This is, of course, called the right-hand rule.

We now define the *linking number* of two oriented curves. However, we need to choose an orientation for each link component to make our definition. See Figure 6.

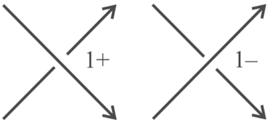


Figure 6 The two orientations possible at a crossing, with a demonstration of how the linking number is calculated for a crossing.

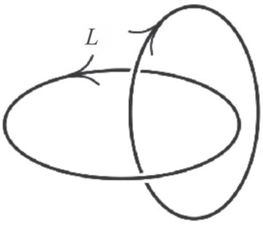


Figure 7 A simple link, L.

For links of two components, the linking number is defined by the formula

$$lk(A, B) = \frac{1}{2}\,(\text{the sum of the signs of the crossings between } A \text{ and } B)\,.$$

In other words, the linking number of A and B is one half the sum of the signs of the crossings that occur between A and B. It is an exercise in using the Reidemeister moves to show that the linking number is unchanged by each move and is hence an invariant of diagram isotopy. For example, the simple link L shown in Figure 7 has linking number $+1$. If a link of two components is designated by a single letter L, then we shall write $lk(L) = lk(A, B)$, where A and B are the two components of L. The fact that the linking number of this simple link is $+1$ means that there is no sequence of Reidemeister moves that will transform it into two disjoint circles. Note also that if you were to switch the orientation of one of the components of L, but not the other, and call this new link L', then $lk(L') = -1$. From this we conclude that there is no way to deform L into L', and that there is no way to deform L' into two disjoint circles.

The link W shown in Figure 8 has linking number zero. However, this Whitehead link W is, indeed, linked. We will use a coloring technique (to be described below) to prove that no isotopy can pull apart the two components of W. For now we will concentrate directly on knots, but the reader interested in the linking number can read the book by Adams [1].

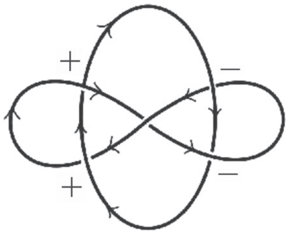


Figure 8 The Whitehead link, W. We can see that the linking number is $lk(W) = \frac{1}{2}(1 + 1 - 1 - 1) = 0$. We found the ±1 by examining the orientation of each of the crossings.

Detecting the trefoil knot by coloring diagrams with three colors

Since we now know that we can transform any knot into an equivalent knot through the Reidemeister moves, we can transform the unknot, whose diagram is a planar circle, into any untangled diagram. We now explain a method for showing that the trefoil (and other examples as well) is knotted. This method involves three different labels $\{r, g, b\}$. We call these labels *colors* (red, green, blue). Each arc in the diagram is assigned a color according to the following rules:

1. Each arc receives a color.
2. Each crossing is either incident to all three colors or to only one color.

We have colored a trefoil in Figure 9. We now demonstrate that these coloring properties are preserved when we do Reidemeister moves on the trefoil. This is shown in Figure 10.

Each time we do a move, we can color the new diagram uniquely in relation to keeping as many colors constant as possible in the original diagram. Note that for a move of type I, we either create a new arc or we join two distinct arcs and no new color is required. For a move of type II, an arc is either removed or an arc appears. The coloring is inherited from one diagram to the other. The same is true for a move

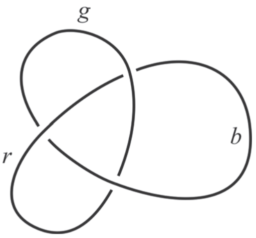


Figure 9 A colored trefoil.

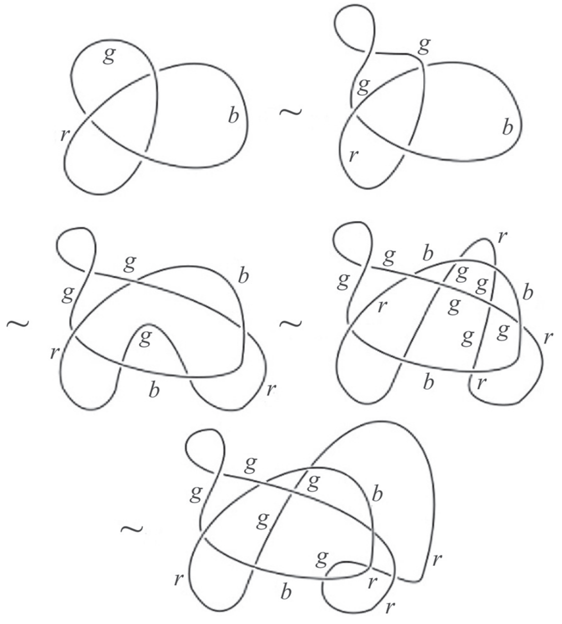


Figure 10 Reidemeister moves and color preservations on the trefoil.

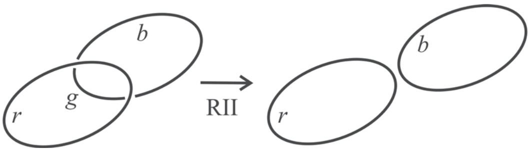


Figure 11 Colors can be lost through Reidemeister moves on a link diagram.

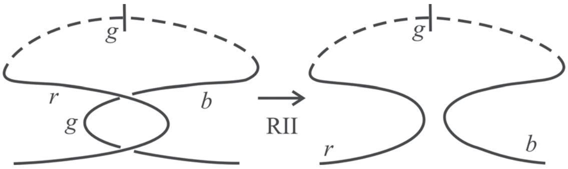


Figure 12 Colors can be lost through Reidemeister moves.

of type III, and we encourage the reader to work through some cases of this before reading further.

We have shown that *every diagram obtained by Reidemeister moves from a three-colored knot diagram inherits a unique three-coloring*. To complete the proof of this statement, there is one point that must be examined. In Figure 11, we show how in a simplifying move of type II, a color can be locally lost. And we illustrate how it can actually be lost in coloring a link of two components. However, in Figure 12, we show that there will be no final loss of this color in a knot diagram. The two arcs that interact in the type II move have different colors. The lost color is the one different from either of these. But in the knot diagram, after the simplifying move, there is a path (the diagram has one component) that starts in one color and ends in the other color. We conclude that the third color must be present to transform one color to the other.

The unknot diagram U is a circle in the plane with no crossings. Thus, by the coloring rules, it is colorable, but only with one color. This means that no sequence of Reidemeister moves can ever create more colors in the diagram other than the starting color. In other words, suppose that there were a sequence of Reidemeister moves starting with the trefoil T and ending with the unknot U. Then, since T has three colors and each diagram obtained from T inherits a unique three-coloring, we would conclude that the unknot U could be three-colored. This is a contradiction. Therefore, because the unknot can only be one-colorable, while the trefoil is three-colorable, we have proved that the trefoil knot is knotted.

Figure 13 is another example of coloring. The diagram is equivalent to a trefoil. This diagram needs a crossing with only one color. It is also a good exercise to see

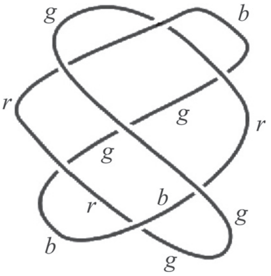


Figure 13 An alternate diagram and coloring of the trefoil.

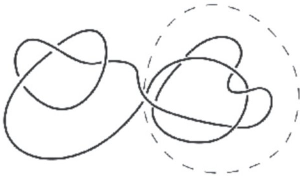


Figure 14 A nugatory diagram. Note that it is possible to surround part of the curve as indicated by the dotted line.

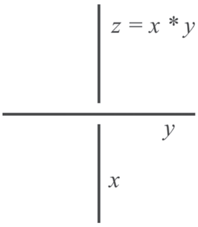


Figure 15 An operation including x, y, z so that it is oriented on the crossing as shown.

how this coloring will reduce to the usual coloring on the trefoil diagram after carrying out some simplifying moves. An interesting unsolved problem is to determine, given a three-colorable knot diagram, the smallest equivalent diagram that demands a special non-nugatory crossing with one color. A crossing is said to be *nugatory* if it is possible to draw a curve in the plane from one side of the crossing to the other without the curve intersecting the knot diagram. See Figure 14.

We have defined the color set $S = \{r, g, b\}$, as well as the binary operation on S, denoted by $*$, such that $x * y$ obeys the following rules:

1. $x * y = x$ if $y = x$.
2. $x * y \neq x$ or y if $y \neq x$.

In Figure 15, we illustrate how one of the undercrossing arcs at a crossing is labeled as the binary product $z = x * y$ of the other undercrossing arc (x) and the overcrossing arc (y). This assignment satisfies the rules listed above.

This system can be used to describe our three or one-coloring rule, which we have developed throughout the paper, by labeling the edges of a crossing as x, y, z (shown in Figure 15) and assigning colors to two of the edges. Note that we can make the assumption of assigning colors to at least two of the arcs because any two colors in the system are arbitrary, and the third color is determined by the other two. Assuming two colors either forces the graph to be colored in one or three colors. It implies the rules shown in Table 1. As the reader can see, the product of any two distinct colors is the third distinct color . The product of a color with itself is itself.

*	r	g	b
r	r	g	b
b	g	b	r
g	b	r	g

TABLE 1: Color Multiplication Table for the $*$ operator

Note that this color product is not associative since $(r * g) * b = b * b = b$, while $r * (g * b) = r * r = r$. The operation is self-distributive in the sense that

$$(x * y) * z = (x * z) * (y * z).$$

For example:

$$(r * g) * b = b * b = b$$

and

$$(r * b) * (g * b) = g * r = b.$$

We will soon see that this self-distributive law corresponds to the third Reidemeister move.

Invariance under the Reidemeister moves

We are now going to show that a three-coloring is invariant under the three Reidemeister moves. Let x, y, and z be arbitrary colors. The first Reidemeister move is shown in Figure 16. The second Reidemeister move and corresponding coloring algebra is shown in Figure 17. Figure 18 shows the algebra for the third Reidemeister move.

1. $x * x = x$.
2. $(x * y) * y = x$.

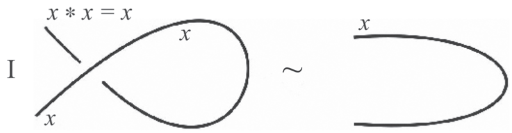


Figure 16 The quandle (see below) we have described applied to the first Reidemeister move. It is easy to verify that $x * x = x$ for any color x.

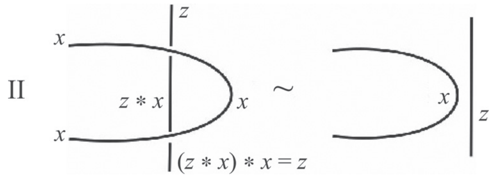


Figure 17 The quandle we have described applied to the second Reidemeister move. It is easy to verify that $(z * x) * x = z$ for any colors x and z.

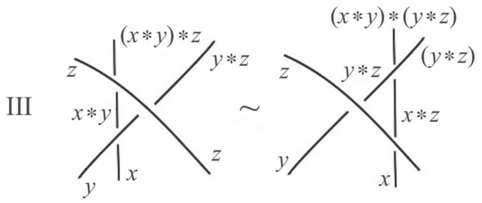


Figure 18 The quandle we have described applied to the third Reidemeister move, which is equivalent to the self-distributive law, $(x * y) * z = (x * z) * (y * z)$.

Via these figures, the Reidemeister moves correspond to the three algebraic rules:

1. $x * x = x$.
2. $(x * y) * y = x$.
3. $(x * y) * z = (x * z) * (y * z)$.

These algebraic rules are the *quandle* axioms [4]. A quandle is an algebraic structure that satisfies these rules. The three-coloring system of r, g, b we have previously defined satisfies the quandle axioms. Note that we can uniquely re-color any diagram after a Reidemeister move has been performed when the coloring system is a quandle.

Coloring links

Many knots can be tricolored, such as the trefoil knot, as we have seen before. And we have seen that a tricolored knot is necessarily knotted. However, it is possible to have a tricolored link that is unlinked! See Figure 19. The coloring for this link is inherited from a coloring of two unlinked circles, as shown in Figure 20.

Because the Reidemeister moves preserve tricolorability, we know that these unlinked links can be tricolored with more than one color, which means that they can be nontrivially tricolored. However, this may not be the case with linked links. In Figure 20 we show the Hopf link, which cannot be tricolored with more than one color.

Theorem 1. *If a link cannot be nontrivially tricolored (meaning with more than just one color), then it must be a linked link (not equivalent to a disjoint collection of unlinked components).*

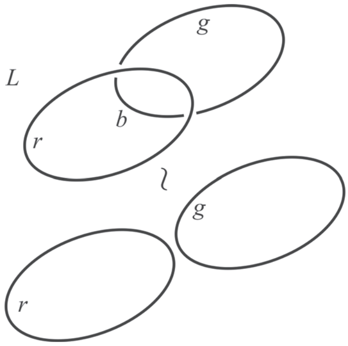


Figure 19 A link, L, that is actually unlinked as shown by the isotopy in the lower figure. L is non-trivially tri-colored.

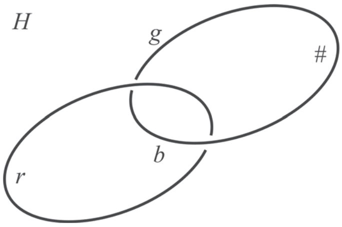


Figure 20 A diagram of the Hopf Link, H which is indeed linked (meaning you cannot perform a series of Reidemeister moves that will make the links unlinked.) The Hopf Link cannot be non-trivially tri-colored.

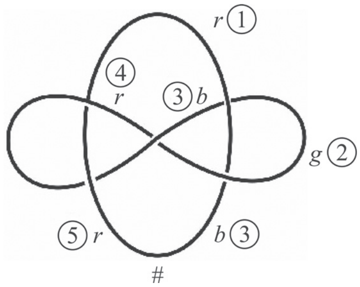


Figure 21 There is no nontrivial coloring of the Whitehead Link.

Proof. If a link L as above is unlinked, then L is equivalent to a collection of disjoint, possibly knotted, components. These components can each be colored one constant color from r, g, b. Since there is more than one such component, we can choose the total coloring of all the components to use at least two colors. Thus, the collection of disjoint components is non-trivially colored. Therefore, L (through Reidemeister moves inducing colorings of diagrams) can be colored non-trivially from r, g, b. Therefore, if L cannot be non-trivially colored, L must be linked. This completes the proof. ■

For example, the Whitehead link W, shown in Figure 21, is uncolorable. To see this, try to color W with three colors. In the figure, we indicate one choice that leads to a contradiction. We first color the arc labeled 1 with r and the arc labeled 2 with g. From this and the coloring rules, it follows that the two arcs labeled 3 are colored b. From that, we conclude that the arc labeled 4 is colored r, from which it follows that the arc labeled 5 must be colored r. But the arc labeled 5 is the same as the arc labeled 3 that we had already colored b. Since blue is not equal to red, it follows that this attempt to color W has failed. Case checking will convince you that any attempt to color the Whitehead link with three colors will fail. The Whitehead link is not three-colorable. This shows that the Whitehead link is linked. Recall that we previously showed that W has linking number zero. Thus, in this case, the coloring method sees more than the linking number.

Remark. A more general quandle structure can be defined by the equation $a * b = 2b - a$ where a and b belong either to the integers or to a modular number system Z/pZ for some natural number p. It is a nice exercise to verify that the quandle axioms are satisfied by this rule. The case where $p = 3$ is equivalent to our three-coloring pattern. Many knots cannot be colored with three colors, but can be colored with a Z/pZ quandle for an appropriate value of p. Such colorings are called *Fox colorings* and can be found in the book by Crowell and Fox [**3**].

The Borromean rings are linked

The Borromean rings, shown in Figure 22, are a link of three components. They have the interesting characteristic that if you remove any one of the three components, the other two components are unlinked. All linking numbers in the Borromean rings vanish. One must go beyond the linking number to verify that they are linked. We will use the method of Nanyes [**7**] to show the linkedness of these rings and then generalize this method to the Brunnian links indicated in Figures 22–25. This is an infinite family of links. We shall refer to it as the *Brunnian Family*. Each link in the family has the property that all pairwise linking numbers are zero, and if any single component in the

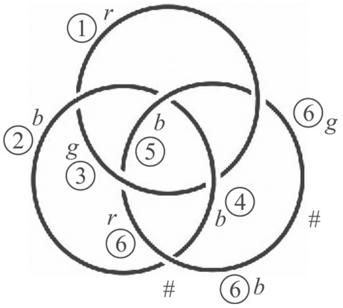


Figure 22 The Borromean rings. Note that if you remove any one of the rings, the remaining rings are no longer linked. We have also shown the steps (in circles) and the colorings for the arcs of the Borromean rings. By showing that there is no nontrivial tricoloring (indicated by the contradiction in step 6) we show that the Borromean rings cannot be tricolored and hence are linked.

link is removed, then it falls apart; the remaining components are entirely unlinked. We shall prove the following theorem:

Theorem 2. *Each link in the Brunnian Family is linked.*

We first demonstrate that the Borromean rings cannot be three-colored. By our previous work, this implies that they are linked (since it is possible to color three unlinked rings with three colors, and a path of Reidemeister moves from three unlinked rings to the Borromean rings would give a three-coloring of the Borromean rings). The goal of this paper is to demonstrate that a certain infinite collection of Brunnian links (generalizing the Borromean rings) are each linked. We accomplish this by showing by induction that each Brunnian link is not tricolorable. First we start with the Borromean rings, as shown in Figure 22.

We complete the proof in the order of the steps shown in the diagram in Figure 22. In the figure, the arcs are indicated by circled numbers as in ①. We will indicate the number assignments using this notation. We being by letting ① = r and ② = b. This implies that ③ = g, which then implies that ④ = b and then ⑤ = b, and ⑥ = r in one place but, on the same arc ⑥ = g and ⑥ = b. This is a triple contradiction! We conclude that the Borrromean rings cannot be non-trivally tricolored.

Now that we have established that the Borromean rings are not tricolorable, we want to do the same for an infinite class of Brunnian rings, a set of links that share the same characteristic as Borromean rings where removing one link unlinks all the other links as well. We call these links the Brunnian links.

Brunnian link proof We now prove that each member of the Brunnian link family indicated in Figures 23–26 is linked. These links are constructed by using as many

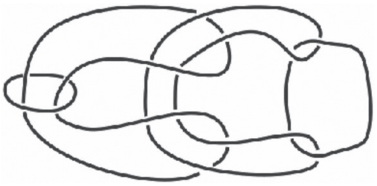


Figure 23 Another generalization of the Borromean rings, but with four rings instead of three. This is known as a Brunnian link, more specifically, a 4-Brunnian link since it has four rings.

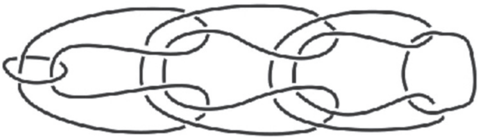


Figure 24 A Brunnian link of five components.

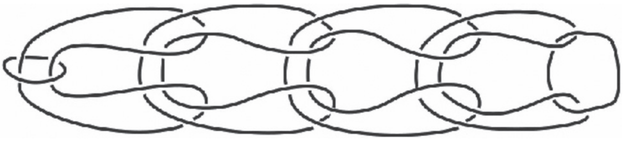


Figure 25 A Brunnian link with six components.

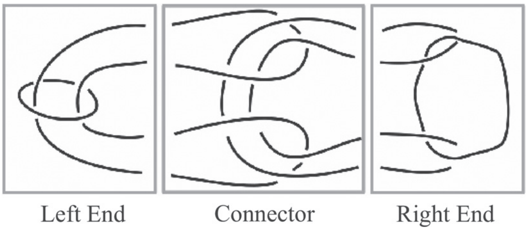


Figure 26 A modular construction for the Brunnian links. We can construct Brunnian links for any n components such that there are n-2 connector (C) parts to the diagram. Adding additional C pieces to the diagram leads to Brunnian links with greater number of components.

center connector pieces as one wants. In this way, we can make Brunnian links with any number of link components greater than or equal to three, using the structure shown in Figure 26. You should verify that the link with no connector piece (a left end connected directly to a right end) is topologically equivalent to the Borommenan rings!

Let R, G, B be three arbitrary colors as previously stated. Note that in Figure 26, we have broken up the Brunnian Links into three different chunks: left ends, connectors, and right ends. We will first focus on the left end, as shown in Figure 27.

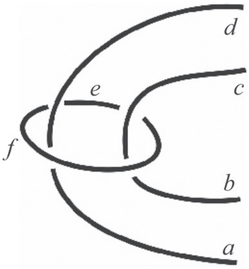


Figure 27 The left end piece and arbitrary letter labels for the edges for clarification purposes.

Left end piece CASE 1. Based on the labeling in Figure 27, we make the following assumption (we assume that all three colors on a, b, and$f$ are different): $a = R$, $b = G$, $f = B$.

We do not assume that all three colors are the same since this would result in a single color labeling of the left end. Based on the above assumption (that all three colors for a, b, and f are distinct), we get that

$$d = a * f = G, \quad c = b * f = R, \quad \text{and} \quad e = f * c = G.$$

But then e must equal R as well because

$$e = f * d = B * G = R \quad \text{and} \quad f = e * c = R * R = R.$$

This would imply $f = R$, but we originally assumed that $f = B$, which is a contradiction. This violates the original assumption about a, b, and f. This implies that two of a, b, f must be the same.

CASE 2. From Case 1, we know that not all of a, b, f can be distinct colors. This means that either $a = f \neq b$ or $a = b \neq f$. In this case, we will find out which two of a, b, f must be the same. First let us suppose that a and f are the same. Let $a = f = R$ and $b = G$. Then $d = R$ and $c = B$. Since d, e, f form a crossing, that means that $e = R$. However, since c, e, f also form a crossing, e must also be G. But e cannot be two different colors. Therefore, there is a contradiction. Thus, the only possibility is when $a = b$, and f is different from them.

For example, let $a = b = R$ and $f = G$. Then we get $e = R$ and $c = d = B$. There is no contradiction, and therefore we see that the only way to obtain a non-trivial coloring of the left connector is to have $a = b$, $c = d$ a different color, and f a third color different from both of them. *Thus, in coloring the left end, no matter what colors we choose, we must have $a = b$, $c = d$, and $a \neq c$.* We move on to the connector piece.

Connector piece For this connector, shown in Figure 28, we will show that no matter what color is assumed for v, since $d = c$ and $b = a$, (as solved from the left end), j and i are always equal and h and g are always equal.

Since we know that $a = b \neq c = d$, from the proof for the left connector piece, let us assume that $a = b = R$, $c = d = B$, and $v = X$, (where X is an arbitrary color).

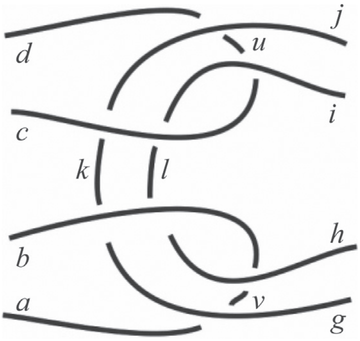


Figure 28 The center connector piece and arbitrary labels for the edges for clarification purposes.

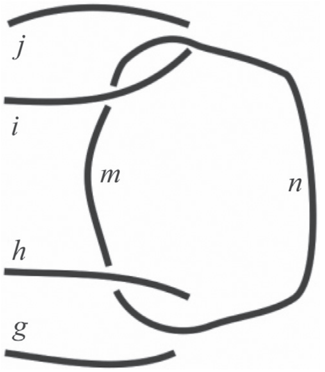


Figure 29 The right end piece and arbitrary labels for the edges for clarification purposes.

Then $h = R * X$ and $g = R * X$ since h touches $b = R$. Also, $v = X$, and g touches $a = R$, and $v = X$. Thus, h and g must be the same color. It follows that

$$k = l = (R * X) * R.$$

Then

$$i = j = ((R * X) * R) * B.$$

We also note that $i = j$ cannot equal $g = h$. For example, if $c = d = B$, then $i = (h * R) * B$. This implies that $i \neq h$.

Since i and j are always equal, and g and h are always equal, we have that regardless of the number of connecting pieces and of whatever colors R, G, and B we chose, we can have an arbitrary number of connector pieces. Had we started with some other combination of two arbitrary colors, we would still end up having i and j be the same and g and h be the same. Thus, if we form a chain of repeating connector pieces, continuing the pattern of the Brunnian links for any number of links, we still have both the inputs (a and b, c and d in this case) and the outputs (i and j, g and h) will be the same color. The colors will alternate, but this does not matter. Given that g and h, and i and j, are the same color, we move on to the final piece of the Brunnian link.

Right end piece Based on Figure 29, since we established that $i = j$ and $g = h$ and $i \neq g$, let us assume that

$$i = j = R \quad \text{and} \quad g = h = G.$$

Then if we solve for n based on the assumptions for i, j, g, and h, we find that $i * n = j$ implies that $R * n = R4$, which implies $n = R$. However, if we solve for n starting from g and h, then we find that $g * n = h$ implies that $G * n = G$, which implies $n = G$. This means that n must be equal to both R and G, which is a contradiction since we stated that R, G, and B are distinct colors. Therefore, no link in the infinite class of Brunnian rings can be can be non-trivially three-colored. Hence, each of them is linked.

This completes our proof that every Brunnian link in our infinite collection of them is linked.

Acknowledgments First and foremost, Claudia and Devika would like to thank their advisor, Dr. Louis Kauffman for giving us the opportunity to study with him, for guiding us through the world of knot theory, and for introducing us to its mysteries and complexities. We would also like to thank Jonathan Schneider for being there when we first proved that the infinite class of Brunnian rings could not be tricolored. We want to thank Ethan Bian for helping us develop a model of the modular Brunnian rings and for patiently studying knot theory with us. We want to thank the Illinois Mathematics and Science Academy, specifically the Student Inquiry and Research department for providing transportation to our research.

Louis H. Kauffman is supported by the Laboratory of Topology and Dynamics, Novosibirsk State University (contract no. 14.Y26.31.0025 with the Ministry of Education and Science of the Russian Federation).

REFERENCES

[1] Adams, C. (1994). *The Knot Book*. New York: W. H. Freeman and Company.

[2] Alexander, J. W., Briggs, G. B. (1926/27). On types of knotted curves. *Ann. Math.* 28(1–4)): 562–586.

[3] Crowwell, R. H., Fox, R. H. (1977). *Introduction to Knot Theory*. New York: Springer.

[4] Joyce, D. (1982). A classifying invariant of knots, the knot quandle. *J. Pure Appl. Algebra* 23(1): 37–65. doi.org/10.106/0022-4049(82)9077-9

[5] Kauffman, L. H. (1987). *On Knots*. Princeton: Princeton University Press.

[6] Kauffman, L. H. (2012) *Knots and Physics*. 4th ed. Singapore: World Scientific.

[7] Nanyes, O. (1993). An elementary proof that the Borromean rings are non-splittable. *American Mathematical Monthly*, 100(8): 786–789. doi.org/10/2307/2324788

[8] Reidemeister, K. (1926). Elementare begründung der knotentheorie. *Abh. Math. Sem. Univ. Hamburg*. 5: 24–32.

[9] Reidemeister, K. (1932). *Knotentheorie*. Berlin: Julius Springer.

Summary. The topology of knots and links can be studied by examining colorings of their diagrams. We explain how to detect knots and links using the method of Fox tricoloring, and we give a new and elementary proof that an infinite family of Brunnian links are each linked. Our proof is based on the remarkable fact (which we prove) that if a link diagram cannot be tricolored then it must be linked. Our paper introduces readers to the Fox coloring generalization of tricoloring and the further algebraic generalization, called a quandle by David Joyce.

LOUIS H. KAUFFMAN is emeritus professor of Mathematics at the University of Illinois at Chicago and is presently visiting Novosibirsk State University in Russia. He is a knot theorist by trade and the Editor in Chief of *Journal of Knot Theory and Its Ramifications*.

DEVIKA PRASAD is a senior at the University of Illinois, Urbana-Champaign, where she is majoring in computer science.

CLAUDIA J. ZHU is a recent graduate of the University of Pennsylvania, where she majored in computer science.

Obtaining Answers to Sensitive Survey Questions Using Venn Diagrams

CLIFFORD JOHNSTON
West Chester University
of Pennsylvania
West Chester, PA 19383
cjohnston@wcupa.edu

When collecting survey data for a sensitive question, we are concerned that the respondents will be untruthful if they believe the survey administrator will know their response. For example, suppose we wished to learn how many students cheated on a university placement test. A student who did cheat but is fearful of being disciplined for answering "yes" to the question, "Did you cheat on the placement test?" will be inclined to lie and give a false answer. If we wish to collect meaningful data on this question, we need to remove the incentive to provide a false answer.

Randomized response methods for sensitive question surveys

A well-established method for obtaining truthful responses to a sensitive question is the use of a randomization technique first proposed by Warner [2]. In this method, a respondent is provided with a sensitive question for which the answer is either "yes" or "no." The respondent is then asked to carry out a random experiment, such as to spin a spinner or roll a die, which also has two outcomes equated with "yes" or "no." The result of the random experiment is not known to the person recording the survey data. The respondent then completes the survey by indicating if the result of the random experiment agrees with his or her answer to the sensitive question. A "yes" answer to the survey question does not reveal the respondent's answer to the sensitive question since the outcome of the random experiment is unknown to the survey administrator. In this way, the respondent does not reveal an answer to the sensitive question. However, since the distribution of the random experiment is known, we can compute the expected number of "yes" responses to the sensitive question. In particular, if q is the proportion of "yes" responses to the sensitive question, then Warner showed that an unbiased, maximum likelihood estimator for q is given by

$$\hat{q} = \frac{1-p}{1-2p} - \frac{f}{1-2p}$$

where p is the probability of a "yes" response in the random experiment and f is the relative frequency of respondents who indicated that their answers to the sensitive question and the random experiment agreed.

For example, suppose the random experiment has a Bernoulli distribution where the likelihood of a "yes" outcome is 25%. Suppose also that in a large survey, 60% of the respondents stated that their answer to the sensitive question agreed with the random outcome. Then, we can estimate the proportion of "yes" responses, q, to the sensitive question as

$$\hat{q} = \frac{1-0.25}{1-2(0.25)} - \frac{0.60}{1-2(0.25)} = 0.30.$$

Math. Mag. **95** (2022) 452–457. doi:10.1080/0025570X.2022.2126676 © Mathematical Association of America
MSC: Primary 03E75, Secondary 97U60

This method does not determine the exact number of “yes” responses in the sample, but rather estimates the proportion of “yes” responses from the population based on the responses to the survey. The accuracy of the result depends on the number of respondents participating in the survey as well as the ratio of “yes” responses among the individuals taking the survey.

Venn diagram application to sensitive question surveys

Based on the randomized response models, we first consider a model consisting of two “yes” or “no” questions—one sensitive and one based on a random experiment—and ask the respondent to indicate in the survey if they would answer both questions the same. For the random outcome, we use a hypergeometric experiment, such as picking a card from a deck of cards without replacement, where we can determine the number of “yes” responses by examining the nonselected outcomes after the survey is completed. Using a two-circle Venn diagram, we then attempt to find the number of “yes” responses to the sensitive question.

In the two-circle Venn diagram model pictured in Figure 1, let N represent the set of respondents answering “yes” to the nonsensitive, randomized question and S represent the set of respondents answering “yes” to the sensitive question. Further, let a, b, c, and d denote the number of respondents in the respective regions of the Venn diagram as shown in the figure. Then, for example, the number of respondents who answered the survey by saying that their answers to both questions agreed would be the sum $b + d$. As noted above, we also know how many people answer “yes” to the nonsensitive question and how many people complete the survey. With the survey complete, we would then have the sums $a + b$, $b + d$, and $a + b + c + d$. Unfortunately, this is not enough information to determine $b + c$, which is the number of respondents that answered “yes” to the sensitive question. For example, the vectors $(6, 4, 4, 6)$ and $(5, 5, 5, 5)$ for (a, b, c, d) are both solutions for $a + b = 10$, $b + d = 10$, and $a + b + c + d = 20$.

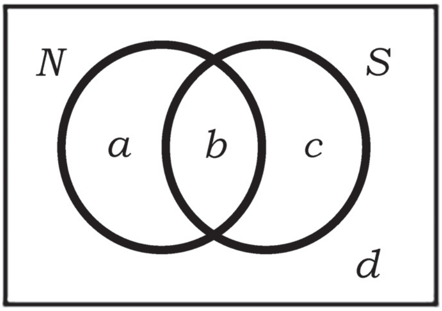
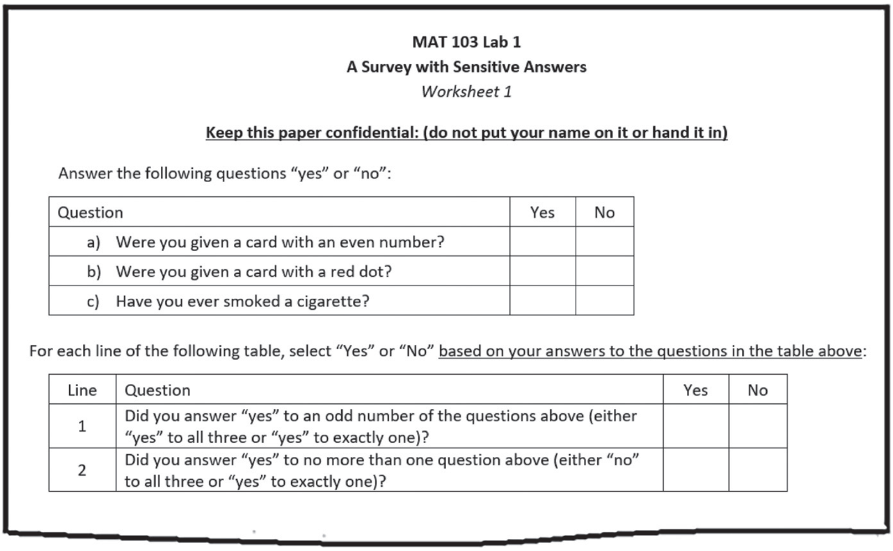


Figure 1 A two circle Venn diagram.

To provide more information, we might attempt to add a question to the survey such as, “Did you answer ‘yes’ to either question?” However, a “no” response to this question identifies the individual as someone who answered “no” to the sensitive question. We think of such a query as a *revealing question* since we can determine the respondent’s answer to the sensitive question for some respondents. Continuing this analysis, we see that any non-revealing question must correspond to a region represented by the sum of two of the variables. A quick application of counting techniques shows there are only six such sums. Also, the sums occur in “complement pairs” so that knowing both provides no substantive information since we know the total number of respondents.



MAT 103 Lab 1

A Survey with Sensitive Answers

Worksheet 1

Keep this paper confidential: (do not put your name on it or hand it in)

Answer the following questions "yes" or "no":

Question	Yes	No
a) Were you given a card with an even number?		
b) Were you given a card with a red dot?		
c) Have you ever smoked a cigarette?		

For each line of the following table, select "Yes" or "No" based on your answers to the questions in the table above:

Line	Question	Yes	No
1	Did you answer "yes" to an odd number of the questions above (either "yes" to all three or "yes" to exactly one)?		
2	Did you answer "yes" to no more than one question above (either "no" to all three or "yes" to exactly one)?		

Figure 2 Lab worksheet.

Therefore, there are only three substantive sums that we could include in the survey. One of those sums is $b + c$, the region we wish to find, and the other two sums, $a + b$ and $b + d$, are already included in the survey. As a result, we cannot "fix" the two-circle Venn diagram method by collecting more information based on the data available in this model. Hence, a two-circle Venn diagram approach results in either too little information or the ability of an administrator to determine at least some respondents' individual answers to the sensitive question.

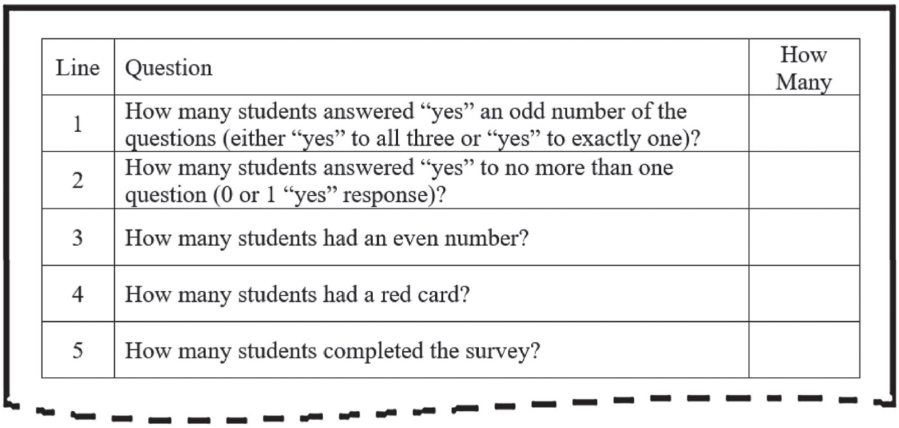
The three-circle model

By moving from a two-circle model to a three-circle model, we gain flexibility. In the two-circle model, each set consisted of two distinct regions in the Venn diagram, whereas in a three-circle model, each set consists of four distinct regions. This increase in regions may allow us to construct questions for the survey that are both non-revealing and provide enough information in total to determine the *exact* number of respondents who answered "yes" to the sensitive question.

The cost is that we need to add another nonsensitive question to the model so that we are now asking the respondents to consider one sensitive question and the outcomes of two random experiments. We also need more information from the survey. In general, a Venn diagram with three sets has eight distinct regions. Hence, to find the cardinality of each region, we will need at least eight pieces of information to generate eight equations. For this problem, though, we do not need to find the cardinality of each region, but the cardinality of the set representing the "yes" responses to the sensitive question.

The survey in practice

We demonstrate our solution with an example. In our adaptation of this technique, we ask each respondent to select a card from a stack of cards numbered from 1 to n in either red or black ink, where n is larger than the number of participants in the survey. The cards are well shuffled before the survey begins and the backs of the cards are indistinguishable.



Line	Question	How Many
1	How many students answered “yes” an odd number of the questions (either “yes” to all three or “yes” to exactly one)?	
2	How many students answered “yes” to no more than one question (0 or 1 “yes” response)?	
3	How many students had an even number?	
4	How many students had a red card?	
5	How many students completed the survey?	

Figure 3 Survey summary table.

The sensitive question for our survey is, “Have you ever smoked a cigarette?” After a survey respondent selects a card, we ask him or her to secretly answer the following questions without revealing his or her answers.

a) Did you pick an even number?

b) Did you pick a red card?

c) Have you ever smoked a cigarette?

Now, on the survey, we ask the respondent to answer the following questions:

(i) Did you answer “yes” to an odd number of questions (exactly one or all three)?

(ii) Did you answer “yes” to no more than one of the three questions?

We then appoint two independent administrators, one for each survey question, and ask the respondents to report to each administrator.

Once all of the respondents have reported to each administrator, the administrators give a tally of the responses to their assigned question and the results are recorded in a table such as Figure 3.

Once the aggregate data is collected, we can determine the number of respondents who answered “yes” to the sensitive question using a three-circle Venn diagram. Three other pieces of information are known besides the collected data. By counting the cards that were not selected, we can determine how many respondents picked an even number and how many respondents picked a red card. We also know the total number of respondents.

Using Figure 4 as a reference, let X be the set of respondents with even numbers, Y be the set of respondents with a red card, and Z be the set of respondents that answered “yes” to the sensitive question. Next, let the variables a through h represent the number of respondents in each corresponding region of the Venn diagram as shown in the figure. Using these identifications, we complete Table 1 indicating the sums of the region variables corresponding to the lines in Figure 3. For example, line 1 corresponds to the sum $a + e + f + g$. We summarize the results in Table 1.

Setting these sums equal to the corresponding results for each line from the survey, we have a system of five equations with eight unknowns. Since we are not looking for a complete solution to this system of equations, the five equations may provide sufficient information. The number of respondents that answered “yes” to the sensitive question is equivalent to the sum $a + c + d + g$. In this particular case, we are able to find the solution using the steps in Table 2.

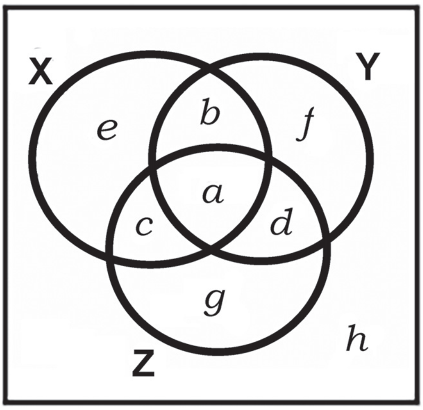


Figure 4 A three-circle Venn diagram.

Line	Sum of the regions
1	$a+e+f+g$
2	$e+f+g+h$
3	$a+b+c+e$
4	$a+b+d+f$
5	$a+b+c+d+e+f+g+h$

TABLE 1: Region sums for collected data.

Mathematical solution

Let x_1 and x_2 be the total number of "yes" responses to survey questions (i) and (ii), respectively; let x_3 and x_4 be the number of respondents who selected an even number and a red card, respectively; and let x_5 be the total number of respondents. Then the number of respondents who answered "yes" to the sensitive question, z, is given by the formula $z = x_1 + 2x_5 - 2x_2 - x_3 - x_4$. The calculations in Table 2 are based on this equation.

This example shows that we can solve this problem using the three-circle Venn diagram model with two nonsensitive questions and a survey with as few as two complex questions. While we have not specifically proven that we need at least two survey questions, our analysis of the two-circle Venn diagram approach strongly suggests that we will need at least two questions. While we do not have another explicit example, we strongly suspect that other formulations of the two question survey will also provide the necessary information to solve the problem.

The lab

The initial goal in formulating this application was to provide students in a general-education mathematics class a substantive application of the Venn diagram techniques used to solve survey problems [1, section 2.4]. As a classroom exercise, we outlined the survey process in a two-part lab paper that was distributed to each student. Two students were appointed as survey administrators. The remaining students each selected a card and completed the first worksheet shown in Figure 2. The students then reported their answers to the survey administrators. Once all the students had reported to the

Action	Result
Multiply line 5 of Table 1 by 2	a. $2a+2b+2c+2d+2e+2f+2g+2h$
Add a. to line 1 of Table 1	b. $3a+2b+2c+2d+3e+3f+3g+2h$
Multiply line 2 of Table 1 by 2	c. $2e+2f+2g+2h$
Add lines 3 and 4 from Table 1	d. $2a+2b+c+d+e+f$
Add c. and d. above	e. $2a+2b+c+d+3e+3f+2g+2h$
Subtract e. from b. above	f. $a+c+d+g=n(Z)$

TABLE 2: Outline of algebraic solution.

survey administrators, the administrators shared the survey tallies with the class. The tallies were recorded on the lab paper in the table shown in Figure 3. The students worked in groups of two or three to complete a "worksheet" that mirrored Table 2 and then answered several questions about the survey process.

Conclusion and contemplations

Student answers to the lab exercise showed that they understood the significance of the problem and the connection to the Venn diagram survey problem techniques. It is less clear that the students grasped the algebra in finding the solution. This method also provides a different approach to a long-studied problem of collecting truthful responses to sensitive questions. There are several possibilities for continuing investigation.

- From a pedagogical standpoint, we wish to assess the extent to which the lab meets the goal of reinforcing the application and usefulness of Venn diagram techniques for solving problems of set cardinality. Also, to what extent can students adapt this process to other applications?
- A complete analysis of the three-set problem should be done similar to the analysis done on the two-set problem. That analysis may provide a "better" set of questions for the survey.
- Since the survey consists of multiple questions, there is a concern that respondents will unknowingly provide incorrect responses as a result of misreading the survey questions or general confusion. An analysis of possible respondent errors may allow us to modify the survey to eliminate or minimize incorrect responses.

REFERENCES

[1] Heeren, C., Heeren, V. E., Hornsby, J., Miller, C. D. (2016). *Mathematical Ideas*, 13th ed. Boston: Pearson.

[2] Warner, S. L. (1965). A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309): 63–69.

Summary. We discuss the application of techniques used to solve Venn diagram survey problems to the problem of collecting truthful responses to sensitive questions. We also develop a method for collecting data for a sensitive question that protects the confidentiality of the respondent's answer. The method uses techniques that students are taught to solve Venn diagram survey problems, and we discuss a classroom exercise based on the solution.

CLIFFORD JOHNSTON (MR Author ID: 351997) is an associate professor of mathematics at West Chester University of Pennsylvania. His professional research interests include PDEs, stochastic processes, mathematical foundations, and math on the web.

Crux's Crux's Crux

AMOL SASANE
London School of Economics
London, United Kingdom WC2A 2AE
A.J.Sasane@lse.ac.uk

The following problem (proposed by Stanley Rabinowitz), appeared as problem 1325 in *Crux*[*] *Mathematicorum*. We call this the *Crux*[†] *Problem* since the accompanying diagram contains a shaded 'cross'=crux.

Crux problem. *Let P be any point inside a unit circle with center C. Perpendicular chords are drawn through P. Rotation of these chords counterclockwise about P through an angle θ sweep out the shaded area shown in Figure 1. Show that this shaded area only depends on θ, but not on P (and hence is easily seen to be 2θ by taking $P = C$).*

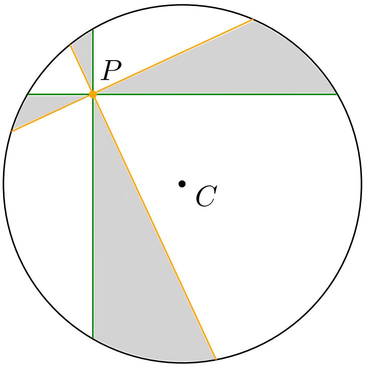


Figure 1 Illustrating the Crux problem.

Two solutions were subsequently published [1, pp. 120-122]:

(I) Jörg Häterich presented a solution using calculus and Archimedes' theorem,

(II) Shiko Iwata presented a non-calculus solution based on trigonometry.

The accompanying editor's note mentioned that Murray Klamkin generalised the problem to n chords through P with equal angles of π/n between successive chords, with the area swept out, when these chords are rotated through an angle of θ about P, then being $n\theta$. The editor's note ended with the following parenthetical remark: "This can also be proved using the solution II. Can it be proved as in solution I?"

In this note, we present a calculus-based solution, based on a special case of a generalization of "Archimedes' theorem," which is proved by employing vectors. We believe that this solution captures in some sense the crux[‡] of the matter.

We begin with a calculus-based proof along lines similar to the first solution given in [1].

[*] First 'Crux' in the title.

[†] Second 'crux' in the title.

[‡] Third 'crux' in the title.

Math. Mag. **95** (2022) 458–464. doi:10.1080/0025570X.2022.2126264 © Mathematical Association of America
MSC: Primary 51M04, Secondary 00A08

A calculus-based proof of the Crux problem

We will use the following result. We call it Archimedes' theorem as it is Proposition 11 in Archimedes' work *The Book of Lemmas* [2, p. 312].

Theorem 1. *(Archimedes) If two mutually perpendicular chords A_1B_1 and A_2B_2 in a unit circle with center C meet at P (see Figure 2), then*

$$PA_1^2 + PB_1^2 + PA_2^2 + PB_2^2 = 4.$$

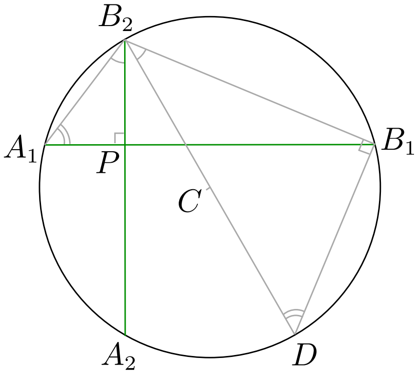


Figure 2 Archimedes' theorem.

Proof. $\triangle B_2PA_1$ is similar to $\triangle B_2B_1D$ since we have

$$\angle B_2A_1B_1 = \angle B_2DB_1 \qquad \text{and} \qquad \angle B_2PA_1 = 90^\circ = \angle B_2B_1D.$$

So

$$\angle PB_2A_1 = \angle B_1B_2D.$$

This implies that

$$\angle A_1CA_2 = \angle B_1CD,$$

and so we obtain $A_1A_2 = B_1D$. By Pythagoras' theorem, we have

$$PB_2^2 + PB_1^2 = B_2B_1^2 \qquad \text{and} \qquad PA_1^2 + PA_2^2 = A_1A_2^2.$$

Adding these, we obtain that

$$\begin{aligned} PA_1^2+PB_1^2+PA_2^2+PB_2^2 &= B_2B_1^2+A_1A_2^2 \\ &= B_1B_2^2+B_1D^2 \\ &= B_2D^2 = 2^2 = 4. \end{aligned}$$

■

Now we give a calculus argument as follows: Rotating A_1B_1 and A_2B_2 about P through an infinitesimal angle $d\theta$, we obtain four sectors, with areas given by

$$\frac{1}{2}PA_k^2\,d\theta, \qquad \frac{1}{2}PB_k^2\,d\theta, \qquad k=1,2.$$

By adding and using Archimedes' theorem, we obtain the rate of change of area

$$\frac{dA}{d\theta}=\frac{1}{2}(PA_1^2+PB_1^2+PA_2^2+PB_2^2)=\frac{1}{2}4=2,$$

and so the total area, if the chords are rotated through an angle θ, is given by

$$\mathrm{A}=\int_0^\theta \frac{dA}{d\theta}\,d\theta=\int_0^\theta 2\,d\theta=2\theta.$$

A vector calculus proof

We will first show the following:

Proposition 1. *Let P be any point inside a unit circle, and, through P, let there be n chords $A_1B_1,\ldots,A_nB_n$ such that there are equal angles of π/n between successive chords. Suppose moreover that A_1B_1 is a diameter. (See Figure 3.) If each chord is rotated counterclockwise through an angle θ, then the total area formed by the resulting sectors is $n\theta$.*

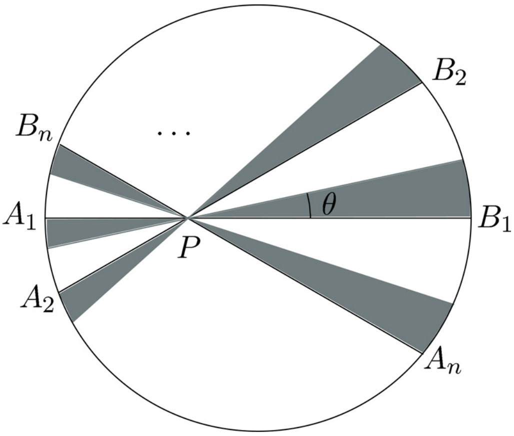


Figure 3 Multiple chords sweeping out equal angles.

This will be shown to yield the generalization (given in Theorem 2) of the Crux problem, where as opposed to the situation above, one of the chords need not be a diameter.

In order to prove Proposition 1, we will first prove a special case of a generalization of Archimedes' theorem (Theorem 3 in the next section, asserting that the sum of the squared distances from a point inside a unit circle to the vertices of n equally angularly spaced chords passing through that point is $2n$), when one of the chords A_1B_1 is the diameter.

Lemma 1 (Generalised Archimedes' theorem, special case)**.** *Let P be any point inside a unit circle, and let there be n chords $A_1B_1, \dots, A_nB_n$ through P such that there are equal angles of π/n between successive chords. Suppose, moreover, that A_1B_1 is a diameter. Then $PA_1^2 + PB_1^2 + \cdots + PA_n^2 + PB_n^2 = 2n$.*

Proof. Let $C_1, \dots, C_n$ be the centers of $A_1B_1, \dots, A_nB_n$. As A_1B_1 is the diameter, C_1 is the center of the circle. See Figure 4. We know that for all $1 \le k \le n$,

$$\langle \overrightarrow{PA}_k - \overrightarrow{PC}_1, \overrightarrow{PA}_k - \overrightarrow{PC}_1 \rangle = \| \overrightarrow{PA}_k - \overrightarrow{PC}_1 \|_2^2 = 1,$$

$$\langle \overrightarrow{PB}_k - \overrightarrow{PC}_1, \overrightarrow{PB}_k - \overrightarrow{PC}_1 \rangle = \| \overrightarrow{PB}_k - \overrightarrow{PC}_1 \|_2^2 = 1.$$

By expanding, adding, and rearranging, we obtain

$$\sum_{k=1}^{n} (PA_k^2 + PB_k^2) = 2\Big\langle \overrightarrow{PC}_1, \sum_{k=1}^{n} (\overrightarrow{PA}_k + \overrightarrow{PB}_k) \Big\rangle - 2nPC_1^2 + 2n. \tag{1}$$

We need to determine the inner product on the right-hand side. We have

$$\overrightarrow{PA}_k = \overrightarrow{PC}_k + \overrightarrow{C_kA}_k \qquad \text{and} \qquad \overrightarrow{PB}_k = \overrightarrow{PC}_k + \overrightarrow{C_kB}_k.$$

But since $\overrightarrow{C_kA}_k + \overrightarrow{C_kB}_k = \mathbf{0}$, we obtain $\overrightarrow{PA}_k + \overrightarrow{PB}_k = 2\overrightarrow{PC}_k$. Hence

$$\sum_{k=1}^{n} (\overrightarrow{PA}_k + \overrightarrow{PB}_k) = 2 \sum_{k=1}^{n} \overrightarrow{PC}_k$$

$$= \sum_{k=1}^{n} \overrightarrow{PC}_k + \sum_{k=1}^{n} \overrightarrow{PC}_{n-k} = \sum_{k=1}^{n} (\overrightarrow{PC}_k + \overrightarrow{PC}_{n-k}).$$

By referring to Figure 4, we see that for all $1 \le k \le n$,

$$\overrightarrow{PC}_k + \overrightarrow{PC}_{n-k} = 2PC_k \left(\cos \frac{k\pi}{n} \right) \frac{\overrightarrow{PC}_1}{PC_1}$$

$$= 2 \left(\cos \frac{k\pi}{n} \right) PC_1 \left(\cos \frac{k\pi}{n} \right) \frac{\overrightarrow{PC}_1}{PC_1} = 2 \left(\cos \frac{k\pi}{n} \right)^2 \overrightarrow{PC}_1.$$

So

$$\sum_{k=1}^{n} (\overrightarrow{PA}_k + \overrightarrow{PB}_k) = \sum_{k=1}^{n} (\overrightarrow{PC}_k + \overrightarrow{PC}_{n-k}) = \sum_{k=1}^{n} 2 \left(\cos \frac{k\pi}{n} \right)^2 \overrightarrow{PC}_1. \tag{2}$$

Now

$$\sum_{k=1}^{n} 2 \left(\cos \frac{k\pi}{n} \right)^2 = \sum_{k=1}^{n} \left(1 + \left(\cos k \frac{2\pi}{n} \right) \right) = n + 0 = n, \tag{3}$$

where we have used

$$\sum_{k=1}^{n} \cos \left(k \frac{2\pi}{n} \right) = 0. \tag{4}$$

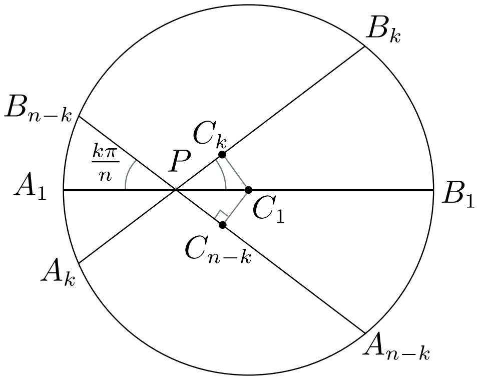


Figure 4 Illustration for the proof of Lemma 1.

To see equation (4), we first note that this sum is the horizontal component of the sum $\overrightarrow{S}$ of n vectors whose tails lie at the center of the unit circle and whose tips lie on the vertices of a regular n-gon inscribed in the circle. To see that $\overrightarrow{S}$ is zero, imagine rotating each vector counterclockwise through an angle of $\frac{2\pi}{n}$, and let the sum of the rotated vectors be $\overrightarrow{S'}$. On grounds of symmetry of the regular polygon, $\overrightarrow{S} = \overrightarrow{S'}$. On the other hand $\overrightarrow{S'}$ ought to be a rotated version of $\overrightarrow{S}$ through an angle of $\frac{2\pi}{n}$. This can only happen if $\overrightarrow{S} = \mathbf{0}$. (Alternative justifications of equation (4) can be given by first summing the geometric series

$$\sum_{k=1}^{n} e^{i\frac{2\pi}{n}k} = e^{i\frac{2\pi}{n}} \frac{1 - e^{i2\pi}}{1 - e^{i\frac{2\pi}{n}}} = 0$$

and taking real parts, or by noticing the sum of the nth roots of unity must add up to 0 since the coefficient of z^1 in $z^n - 1$ is 0, and again taking real parts.)

Consequently, using equations (1), (2), and (3), we obtain

$$\begin{aligned}\sum_{k=1}^{n}(PA_k^2 + PB_k^2) &= 2\left\langle \overrightarrow{PC}_1, \sum_{k=1}^{n}\left(\overrightarrow{PA}_k + \overrightarrow{PB}_k\right)\right\rangle - 2nPC_1^2 + 2n \\ &= 2\left\langle \overrightarrow{PC}_1, n\overrightarrow{PC}_1\right\rangle - 2nPC_1^2 + 2n \\ &= \cancel{2nPC_1^2} - \cancel{2nPC_1^2} + 2n = 2n.\end{aligned}$$

■

We are now ready to prove Proposition 1.

Proof of Proposition 1. Rotating $A_1B_1, \ldots, A_nB_n$ anticlockwise about P through an infinitesimal angle $d\theta$, we obtain $2n$ sectors, with areas given by

$$\frac{1}{2}PA_k^2\,d\theta, \qquad \frac{1}{2}PB_k^2\,d\theta, \qquad k = 1, \ldots, n.$$

By adding and using Lemma 1, the rate of change of the total area is seen to be

$$\frac{dA}{d\theta} = \frac{1}{2}\sum_{k=1}^{n}(PA_k^2 + PB_k^2) = \frac{1}{2}\,2n = n,$$

and so the total area, if the chords are rotated through an angle θ, is given by

$$A = \int_0^\theta \frac{dA}{d\theta}\, d\theta = \int_0^\theta n\, d\theta = n\theta.$$

■

Theorem 2. *Let P be any point inside a unit circle, and let there be n chords through P such that there are equal angles of π/n between successive chords. If each chord is rotated counterclockwise through an angle θ, then the total area formed by the resulting sectors is $n\theta$.*

Proof. To see how this follows from Proposition 1, we first construct the diameter A_1B_1 through P, and consider successive anticlockwise rotations of this diameter through angles of π/n, resulting in the chords $A_2B_2, \ldots, A_nB_n$. Let the given chords from the theorem statement be labeled as $A'_1B'_1, \ldots, A'_nB'_n$, and let their rotated versions (though an angle θ) be labeled as $A''_1B''_1, \ldots, A''_nB''_n$. See Figure 5.

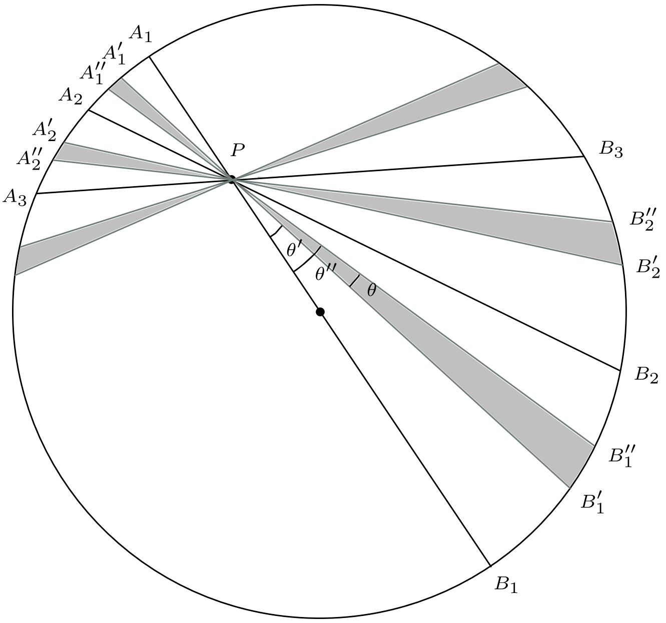


Figure 5 Illustration for the proof of Theorem 2.

Let the angle between A_1B_1 and $A'_1B'_1$ be θ', and that between A_1B_1 and $A''_1B''_1$ be θ''. Then for all $1 \le k \le n$, we use the notation $\widehat{B'_kB''_k}$ for the sector formed by the corresponding arc with P, and denote the area of the sector by $\mathrm{A}(\widehat{B'_kB''_k})$. Then:

$$\sum_{k=1}^{n}[\mathrm{A}(\widehat{B'_kB''_k}) + \mathrm{A}(\widehat{A'_kA''_k})]$$

$$= \sum_{k=1}^{n}[\mathrm{A}(\widehat{B_kB''_k}) - \mathrm{A}(\widehat{B_kB'_k}) + \mathrm{A}(\widehat{A_kA''_k}) - \mathrm{A}(\widehat{A_kA'_k})]$$

$$\begin{aligned}
&= \sum_{k=1}^{n}[\mathrm{A}(\wideparen{B_k B_k''}) + \mathrm{A}(\wideparen{A_k A_k''})] - \sum_{k=1}^{n}[(\mathrm{A}(\wideparen{B_k B_k'}) + \mathrm{A}(\wideparen{A_k A_k'}))] \\
&= n\theta'' - n\theta' = n(\theta'' - \theta') = n\theta.
\end{aligned}$$

■

Archimedes′ theorem

A consequence of Theorem 2 is the following generalization of Archimedes' theorem from the $n = 2$ chord case considered earlier.

Theorem 3 (Generalised Archimedes' theorem)**.** *Let P be any point inside a unit circle, and let there be n chords $A_1B_1, \ldots, A_nB_n$ through P such that there are equal angles of π/n between successive chords. Then*

$$PA_1^2 + PB_1^2 + \cdots + PA_n^2 + PB_n^2 = 2n.$$

Proof. By Theorem 2, we know that if the chords are rotated through an infinitesimal angle $d\theta$, then the sum of the areas of the resulting sectors is $n\, d\theta$. But this area is also equal to

$$\frac{1}{2}(PA_1^2 + PB_1^2 + \cdots + PA_n^2 + PB_n^2)d\theta.$$

Consequently, we obtain that

$$PA_1^2 + PB_1^2 + \cdots + PA_n^2 + PB_n^2 = 2n.$$ ■

Acknowledgment The author is grateful to the reviewer and the Editor for useful suggestions on improving the exposition.

REFERENCES

[1] Solution to problem 1325. (1989). *Crux Mathematicorum*. 15(4): 120–121.

[2] *The works of Archimedes*. (2002). Reprint of the 1897 edition and the 1912 supplement, edited by T. L. Heath. New York: Dover.

Summary. A result of Archimedes states that for perpendicular chords passing through a point P in the interior of the unit circle, the sum of the squares of the lengths of the chord segments from P to the circle is equal to 4. A generalization of this result to $n \geq 2$ chords is given. This is done in the backdrop of revisiting Problem 1325 from *Crux Mathematicorum*, for which a new solution is presented.

AMOL SASANE earned a Bachelors Degree in Electrical Engineering from the Indian Institute of Technology, Mumbai, a PhD in Mathematics from the University of Groningen, the Netherlands, and a Masters in Theoretical Physics from Lund University, Sweden. He is now a professor of mathematics at the London School of Economics. His research interests lie in the field of applicable analysis.

Doubling the Cube and Constructability in Higher Dimensions

JULIUS BARBANEL
Union College
Schenectady, NY 12308
barbanej@union.edu

The ancient Greeks introduced the three so-called "classical geometric construction problems":

1. trisecting an arbitrary angle,
2. squaring the circle, and
3. doubling the cube.

The meaning of problem one is clear. Problem two is the problem of constructing a square with the same area as a given circle. Problem three is the problem of constructing a second cube having twice the volume of a given cube. We shall focus our attention on problem three.

The obvious first question to ask is: what do we man by "construction"? Euclid, in his *Elements* [1], did not offer a direct answer, but he listed postulates that came to be interpreted as the familiar straightedge and compass approach to constructions in the plane. Indeed, the straightedge and compass are often referred to as the "Euclidean construction tools," or simply the "Euclidean tools."

It is now known that none of these three problems can be solved using the Euclidean tools. However, ancient Greek mathematicians solved these three problems using other means. We study and then generalize two such solutions to problem three.

In the first section, we discuss the doubling-the-cube problem and show how it can be recast as the problem of finding mean proportionals. We then introduce Euclid's postulates for two-dimensional constructions and discuss the impossibility of solving the three classical geometric construction problems using only the Euclidean tools. The following section presents a "mechanical" solution to the doubling-the-cube problem. We then present a solution to problem three due to Archytas, which involves a construction in three dimensions, and we generalize the mechanical method presented earlier.

Finally, we close by considering higher-dimensional analogs of the three-dimensional constructability discussed previously. We also ponder what appears to be an omission in Euclid's *Elements*: Euclid performs geometric constructions in both two and three dimensions, but he only gives postulates for constructions in two dimensions.

We shall often write $2D$, $3D$, and nD for "two dimensions" or "two-dimensional," and so forth.

We assume the reader is familiar with the field $\mathbb{Q}$ of rational numbers and the field $\mathbb{R}$ of real numbers. For any positive integer n, we denote by $\mathbb{Q}[\sqrt[n]{}]$ the smallest field that contains $\mathbb{Q}$ and is closed under nth roots. We extend this notation by letting $\mathbb{Q}[\sqrt[n_1]{}, \sqrt[n_2]{}, \dots, \sqrt[n_k]{}]$ denote the smallest field that contains $\mathbb{Q}$ and is closed under n_1 roots, n_2 roots, ..., n_k roots. Each such field is countably infinite.

Math. Mag. **95** (2022) 465–481. doi:10.1080/0025570X.2022.2127300 © Mathematical Association of America
MSC: Primary 51M15, Secondary 51M04

The problem of doubling the cube

It would be easy for a modern reader to wonder what all the fuss is about. If a side of the original cube has length k, then its volume is k^3. It follows that a cube having twice its volume would have volume $2k^3$, and therefore would have sides of length $k\sqrt[3]{2}$. However, no one among the ancient Greeks would accept this as a solution. Ancient Greek geometry was *pure* geometry. There were no coordinate systems, no equations corresponding to curves or surfaces, and, most importantly for us, no numerical measurements of areas or volumes. The modern term for this is *synthetic geometry*.

Since a cube is completely determined by its side, the doubling-the-cube problem is: Given an arbitrary cube, construct a line segment such that the cube having that segment as a side has volume twice that of the original cube. We shall sometimes use the modern perspective of associating real numbers with line segments as an explanatory tool, but the reader should keep in mind that this tool was not available to the ancient Greeks.

The doubling-the-cube problem is also known as the Delian problem. According to one legend, the problem originated in the fourth century BCE, when the people of the Greek island of Delos asked the oracle at Delphi to help them with some difficulty, which is sometimes said to be a plague sent by the god Apollo and sometimes said to be internal political strife [3]. Eratosthenes, in his *Platonicus*, tells us that:

> When the god proclaimed to the Delians by the oracle that, if they would get rid of the plague, they should construct an altar double the existing one, their craftsman fell into great perplexity in their efforts to discover how a solid could be made double of a (similar) solid; they therefore went to ask Plato about it, and he replied that the oracle meant, not that the god wanted an altar of double the size, but that he wished, in setting them to the task, to shame the ancient Greeks for their neglect of mathematics and their contempt for geometry.

We know from other sources that the doubling-the-cube problem actually originated earlier than the time of Plato. In particular, an important breakthrough occurred in the fifth century BCE. Hippocrates of Chios reduced the doubling-the-cube problem to the problem of finding two mean proportionals between two line segments. If p and q are two line segments, then line segments x and y are mean proportionals for p and q if $p : x = x : y = y : q$. Of course, we naturally view these equations as equating numbers, but the ancient Greeks did not. They had a sophisticated and fascinating nonnumerical theory of ratios of geometric objects such as line segments, and they believed that mean proportionals had special significance, both mathematically and esthetically. The theory was developed by Eudoxus [1, Book V], [3, Vol. 1, pp. 325–327].

How does finding two mean proportionals solve the doubling-the-cube problem? Suppose that the cube we wish to double has side k, and let x and y be two line segments that are mean proportionals between k and $2k$. Then (using our modern perspective in which k, x, y, and $2k$ denote both the line segments and their lengths), we have $k/x = x/y = y/2k$, or, equivalently, $x/k = y/x = 2k/y$. Call this common latter ratio w. It follows that $2k = wy = w^2x = w^3k$, and hence $w^3 = 2$ and $w = \sqrt[3]{2}$. This tells us that $x = kw = k\sqrt[3]{2}$. It follows that a cube with side x has volume $2k^3$, twice the volume of a cube with side k. Of course, an ancient Greek mathematician would have used a purely geometric argument instead of this algebraic argument.

Once Hippocrates established this result, efforts to solve the doubling-the-cube problem were redirected to the problem of finding two mean proportionals. There were several successful approaches to this problem using so-called “mechanical methods.”

These involved techniques such as moving or rotating some geometric object until some other condition (such as having some movable line pass through some fixed point) is satisfied. We will later see one such example, and then consider a solution involving a construction in three dimensions. It would have been preferable to use the Euclidean tools in two dimensions, but as we shall see, this was not to be!

The unsolvability of the three classical construction problems using Euclidean tools

What did Euclid mean by a construction in the plane? His first three postulates are the basis for what became, and still is, the prevailing perspective. His first three postulates are:

1. A line segment can be drawn joining any two points.
2. Any line segment can be extended infinitely in a straight line.
3. Given any line segment, a circle can be drawn having the segment as radius and one endpoint as center.

Although Euclid never referred to a "straightedge" or to a "compass," it is not hard to see that such constructions are precisely those allowed by his three postulates. This led to the perspective that "construction in the plane" means "construction with straightedge and compass." In other words, a geometric object is considered to be "constructible" if and only if it can be constructed using the Euclidean tools.

Many such constructions appear in the *Elements* and are taught in standard high school geometry classes. Three typical examples of such constructions are

- Constructing an equilateral triangle having a given line segment as one of its sides (Euclid's first proposition).
- Bisecting a given line segment (Euclid's tenth proposition).
- Drawing a line through a given point, perpendicular to a given line not containing the point (Euclid's twelfth proposition).

We note that the following three problems, which are similar to the three classical geometric construction problems, are solvable using the Euclidean tools:

1. Bisecting an arbitrary angle.
2. Squaring a rectangle. (That is, constructing a square having the same area as a given rectangle.)
3. Doubling a square.

Construction 1 is straightforward and appears as Proposition 9 of Book I of Euclid's *Elements*. Construction 2 is more interesting and somewhat harder. It is Proposition 14 of Book II of the *Elements*. Construction 3 is trivial since a side of the doubled square is equal to a diagonal of the original square.

All attempts at solving the three classical geometric construction problems using the Euclidean tools met with failure. It took more than 2000 years, but we now know why: it is impossible to accomplish these constructions using only the Euclidean tools. This was shown for the trisecting-an-angle problem and for the doubling-the-cube problem by Pierre Wantzel in 1837, and for the squaring-the-circle problem by Ferdinand von Lindemann in 1882. For the proofs, see, for example, Fraleigh [2].

As noted previously, we can rephrase the doubling-the-cube problem in modern terms: given a line segment of length k, construct a line segment of length $\sqrt[3]{k}$. (Recall that the ancient Greeks were pure, or synthetic, geometers; they would not have associated real numbers with line segments in this way). This was the approach that Wantzel used to show that the doubling-the-cube problem cannot be solved with the Euclidean tools. We now review the main ideas of his proof.

We work in the plane and we wish to define $2D$ constructible circles, lines, points, and real numbers. We begin by assuming that we are given two distinct points and the line segment between them. This segment is our first constructible line segment and these two points are our first constructible points. Then, we use only the Euclidean tools to obtain new:

1. constructible circles, obtained by using the compass with one constructible point as the center and another constructible point as a point on the circle;
2. constructible lines (and half lines and line segments), obtained by using two constructible points and the straightedge; and
3. constructible points, obtained as the intersection of two constructible lines, two constructible circles, or a constructible line and a constructible circle.

More formally, we can define the set of all $2D$ constructible circles, lines, and points as either:

1. the smallest set that contains the given two points and line segment and is closed under a finite number of applications of 1, 2, and 3 above, or
2. the union of a countable sequence of sets, the first of which includes only the original constructible two points and line segment, and each subsequent set includes the previous set together with all circles, lines, and points constructible from that previous set by using just one of the above rules one time.

Finally, we say that a real number r is *constructible* if and only if there is a constructible line segment of length $|r|$, where we define our unit length to be the length of the original constructible line segment. Formulation 2 above tells us that the set of $2D$ constructible objects is countable since it is the countable union of finite sets, and thus the set of $2D$ constructible real numbers is countable.

It is trivial to see that the set of constructible numbers is closed under addition since, given two constructible numbers, we can simply put two corresponding line segments end-to-end. The resulting line segment has length equal to the sum of the two given constructible numbers. It is true, but less obvious, that the set of constructible numbers satisfies the other field conditions. This field includes $\mathbb{Q}$ since it includes the number 1 by definition, is closed under addition and multiplication, and contains multiplicative inverses of each of its nonzero elements.

Wantzel showed that the set of all constructible real numbers is the smallest field that contains the rational numbers and is closed under square roots of its nonnegative numbers. This is the field $\mathbb{Q}[\sqrt[2]{\ }]$. Some examples of constructible real numbers are:

$$17, \quad \frac{5}{9}, \quad \sqrt[8]{5}, \quad \sqrt{\frac{5}{7} - \sqrt[4]{\frac{2}{3}}}.$$

There was nothing in ancient Greek mathematics resembling modern algebraic field theory, and it is algebraic field theory that supplies the crucial piece of the argument: $\sqrt[3]{2}$ is not in the field $\mathbb{Q}[\sqrt[2]{\ }]$. Before summarizing Wantzel's argument, we note that there is helpful intuition that provides some insight into why $\sqrt[3]{2}$ is not in the field $\mathbb{Q}[\sqrt[2]{\ }]$.

Beginning with $\mathbb{Q}$, we obtain new constructible numbers by using pairs of points to form lines and circles. We then look at intersections of lines with lines, of circles with circles, and of lines with circles, to get new constructible points, which are then used to form new lines and circles. We obtain new constructible real numbers as the lengths of new constructible line segments (and as the negatives of these lengths). At each stage of this process, the relevant algebra involves solving linear and/or quadratic equations. Solving such equations can never produce a new cube root (that is, a cube root that is not in $\mathbb{Q}$). Hence, $\sqrt[3]{2}$ is not in $\mathbb{Q}[\sqrt[2]{\ }]$.

In summary, the formal argument showing that $\sqrt[3]{2}$ is not in the field $\mathbb{Q}[\sqrt[2]{\ }]$ is as follows: The minimal polynomial of $\sqrt[3]{2}$ over the rationals has degree 3, whereas the minimal polynomial over the rationals of any constructible real number has degree a power of 2. (See Fraleigh [2]).

We may now complete the argument. Since the field $\mathbb{Q}[\sqrt[2]{\ }]$ is closed under multiplication and inverses, it follows that for any real number k in $\mathbb{Q}[\sqrt[2]{\ }]$, we have that $k\sqrt[3]{2}$ is not in $\mathbb{Q}[\sqrt[2]{\ }]$. Thus, $k\sqrt[3]{2}$ is not constructible. Therefore, given a cube with side k, it is not possible, using only the Euclidean tools, to construct a line segment such that a cube having that segment as a side has twice the volume of the original cube.

Most modern mathematicians would say that, in light of Wantzel's proof, the doubling-the-cube problem is not solvable (and similarly, due to Wantzel's other proof and the one by Lindemann referred to earlier, that the trisecting-an-angle problem and the squaring-the-circle problem are not solvable). However, not all of the ancient Greeks would have agreed! Indeed, all three of these problems were "solved" by the ancient Greeks, albeit by using something beyond the standard Euclidean tools.

In light of these impossibility results, if we are to solve these problems we need to loosen the requirement that we use only the $2D$ Euclidean tools and work only in two dimensions. There are two natural ways to loosen these requirements:

1. we can work in two dimensions, with the Euclidean tools, but also make use of other tools as well, or
2. we can work with something like the Euclidean tools, but in higher dimensions.

In each of the next two sections, we consider a solution to the doubling-the-cube problem. In the next section we use approach 1, and in the section after that we explore approach 2. In both cases, we continue to use the $2D$ Euclidean tools, but we allow additional tools as well.

A mechanical solution to the doubling-the-cube problem

We now present a so-called "mechanical" solution to the doubling-the-cube problem that is in the spirit of those actually used by ancient Greek mathematicians. In particular, our use of multiple similar triangles is comparable to a technique used by Eratosthenes to solve the doubling-the-cube problem. (See [3, Vol. 1, pp. 258–260]). It involves a construction in two dimensions, but uses something in addition to the Euclidean tools.

Consider Figure 1.* The angles PQR, PRS, and PST are all right angles. It follows that triangles PQR, PRS, and PST are similar, since they each have a right angle and they share a common angle at P. This tells us that

$$PQ : PR = PR : PS = PS : PT.$$

Thus, line segments PR and PS are mean proportionals for line segments PQ and PT. If we now imagine that PQ is the side of the cube that we wish to double and

*Note that the online version of this article has color diagrams.

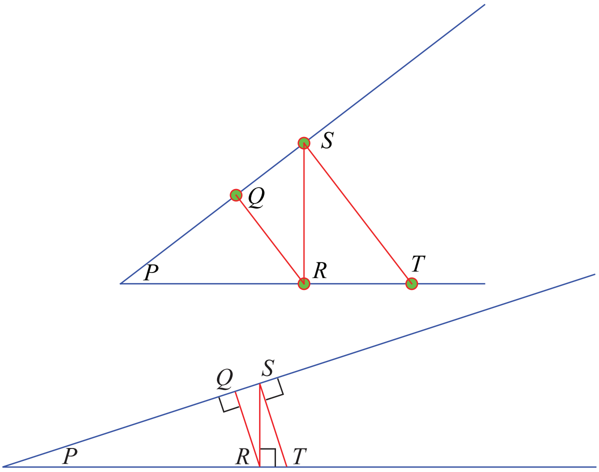


Figure 1 A solution to the doubling-the-cube problem employing methods beyond the Euclidean tools.

that $|PT| = 2|PQ|$, then PR is the side of the cube having twice the volume of the original cube.

How do we arrange things so that PQ is the side of the cube that we wish to double and $|PT| = 2|PQ|$? This is the "mechanical" part of this "mechanical method." We can imagine fixing point Q's position on one of the half-lines of the angle at P so that PQ is a side of the cube that we wish to double, and varying the angle at P from a very small angle to an angle just slightly less than a right angle. As we do so, points R, S, and T move along their half-lines in such a way as to preserve the right angles PQR, PRS, and PST. A small angle at P, as in the bottom figure in Figure 1, will result in a point T such that PT is only slightly longer than PQ, and a nearly right angle will result in a point T such that PT is many times longer than PQ. By the intermediate value theorem, there is an angle in-between so that $|PT| = 2|PQ|$. This is the angle illustrated on the top of Figure 1. For that angle, $|PR| = \sqrt[3]{2}|PQ|$, and thus the cube with side PR has twice the volume of the cube with side PQ, and we have doubled the cube.

As in the previous section, suppose that we have specified a unit length. (This is a necessary step in order to define which real numbers are constructible in our new setting, using a mechanical device.) Choose Q so that $|PQ| = 1$. Next, instead of adjusting the angle at P so that $|PT| = 2|PQ| = 2$, we adjust the angle so that $|PT| = c$ for some real number $c > 1$. Then, PR and PS are mean proportionals between PQ and PT. This tells us that $|PR| = \sqrt[3]{c}$. Thus, this mechanical method allows us to construct the cube root of any previously constructed $c > 1$. For $0 < c < 1$, we reverse the roles of points Q and T, fixing T so that $|PT| = 1$, let the points Q, R, and S move along their half-lines, and adjusting the angle at P so that $|PQ| = c$. Then, PS and PR are two mean proportionals between PT and PQ, and thus $PS = \sqrt[3]{c}$. This establishes that this mechanical method allows us to construct the cube root of any real number that was previously constructed.

The ancient Greeks found many mechanical solutions to the doubling-the-cube problem. See Heath [3].

Does this solve the doubling-the-cube problem? Plato answered this question with an emphatic "no!" According to the ancient biographer and essayist Plutarch [3, Vol. I,

p. 287], Plato's perspective on all such mechanical solutions, which involved moving parts rather than geometric constructions of immovable objects (what we would refer to as constructions using the Euclidean tools), was that

> the good of geometry is thereby lost and destroyed, as it is brought back to things of sense instead of being directed upward and grasping at eternal and incorporeal images.

What kinds of constructions might Plato have preferred? We address this in the next section.

A solution to the doubling-the-cube problem involving a construction in three dimensions

We now present a solution to the doubling-the-cube problem due to Archytas. It involves generalizing the Euclidean tools to three dimensions. Archytas of Tarentum was a mathematician, philosopher, and political leader. He was a friend of Plato, and some scholars argue that he may be one of Plato's models for the Philosopher King, as described in his *Republic*. Sir Thomas Heath, one of the foremost modern historians of ancient Greek mathematics, describes this solution to the doubling-the-cube problem as follows:

> The solution by Archytas is the most remarkable of all, especially when his date is considered (first half of the fourth century BC), because it is not a plane construction but a bold construction in three dimensions, determining a certain point as the intersection of three surfaces of revolution ... ([3, Vol. 1, pp. 246–247])

Of course, it may seem that performing a $3D$ construction to double the cube is only natural since a cube is a $3D$ object.

Next, we outline this construction:

1. See Figure 2.
   (a) Fix some plane in space.
   (b) Draw a circle in this plane.
   (c) Choose points A and C on the circle so that AC is a diameter.
   (d) Choose a third point B on the circle. (We shall describe later how B, which will be a constructible point, should be chosen.) Draw a semicircle with AC as diameter, but in a plane perpendicular to the plane of ABC.
   (e) Rotate this semicircle about the line through A that is perpendicular to the plane of ABC. This creates a half-torus with inner radius zero.
2. See Figure 3. Draw a half-cylinder with semicircle ABC as its base, on the same side of the plane of ABC as the half-torus.
3. See Figure 4. Rotate half-line AB about line AC to form a right circular cone.

Figure 5 shows these three surfaces together.

We are interested in the intersection point of these three surfaces. In Figure 6, we first show just the half-torus and the cone, indicate their intersection curve, and then show how this curve intersects the half-cylinder. Of course, this is the point of intersection of the three surfaces. Let this point be P, and let Q be the point at which the line through P perpendicular to the original plane intersects that plane. We note that Q is

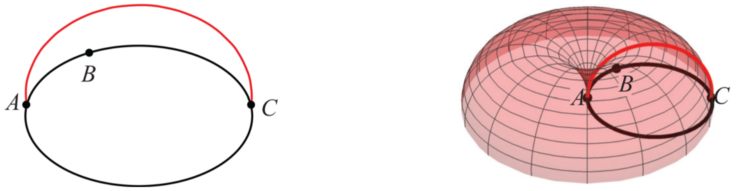


Figure 2 Step 1 of Archytas′ $3D$ solution to the doubling-the-cube problem.

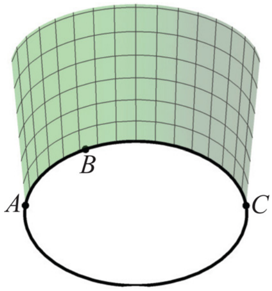


Figure 3 Step 2 of Archytas′ solution.

on circle ABC since P is on the half-cylinder drawn with ABC as base. See Figure 7. It can be shown that

$$AB : AQ = AQ : AP = AP : AC.$$

For the proof, see Heath [3, Vol. 1, pp. 247–249]. This reference includes both a purely geometric proof of the sort that any ancient Greek mathematician might have used, and a modern approach using analytic geometry.

This tells us that AQ and AP are mean proportionals for AB and AC.

Next, let us suppose that the cube we wish to double has sides of length k. If we perform this construction with $|AC| = 2k$ and $|AB| = k$, then a cube with side AQ will have twice the volume of a cube with side k, as desired.

We invite the reader to speculate as to how Archytas might have been led to use a torus, a cylinder, and a cone in his construction. This certainly solves the doubling-the-cube problem, but is this solution by Archytas a constructible solution? It is far from clear what we mean by "construction" in our present $3D$ context. How does this method relate to Euclid's three postulates, which led to our notion of straightedge-and-compass constructions in the plane? We shall consider this question later.

Generalizing the mechanical procedure: The adjustable angle root device

We have previously presented a mechanical method to double the cube. It involved a mechanical device that enabled us to find two mean proportionals between any two line segments. We also noted that this technique allows us to find the cube root of any previously constructed number.

Can we generalize this method to find roots other than cube roots? The device we introduced (Figure 1) involves three line segments inside an angle, and this allowed us to find cube roots. We can expand this device to n segments instead of 3, for $n > 3$. The natural and obvious conjecture is that this yields nth roots. This conjecture turns out to be true, and it is easy to establish.

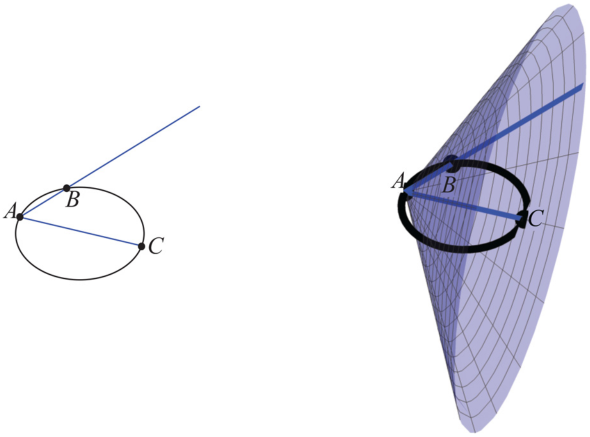


Figure 4 Step 3 of Archytas' solution.

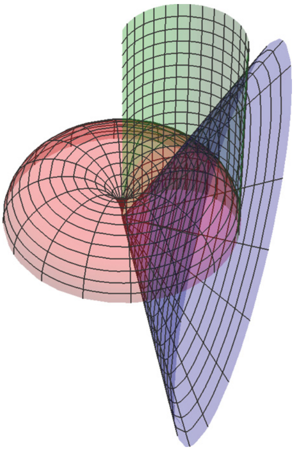


Figure 5 The three surfaces used in Archytas' solution shown in one diagram.

Let us call these mechanical devices *adjustable angle root devices*, or *aards*. More specifically, we shall call the aard with n segments inside the angle an $aard(n)$. We have shown that aard(3) enables us to construct the cube root of any previously constructed number. In terms of mean proportionals, aard(3) enables us to construct two mean proportionals between any two previously constructed line segments. Next, we illustrate how, for any $n > 3$, we can use aard(n) to construct $n - 1$ mean proportionals between any two previously constructed line segments, and thus to construct the nth root of any number that we have already constructed.

It is easy to see how to generalize the aard(3) device shown in Figure 1. This is shown in Figure 8, which illustrates aard(7). The generalization from aard(3) is completely straightforward. We have

$$PQ : PQ_1 = PQ_1 : PQ_2 = \cdots = PQ_6 : PT.$$

Precisely as we saw earlier, we choose Q so that $|PQ| = 1$ and, for any real number $c > 1$, we fix the angle at P so that $|PT| = c$. Then PQ_1 is the first of the six mean proportionals for PQ and PT, and thus $|PQ_1| = \sqrt[7]{c}$.

For $0 < c < 1$, we proceed as earlier by choosing T so that $|PT| = 1$ and then fixing the angle at P so that $|PQ| = c$. Then, $|PQ_6| = \sqrt[7]{c}$.

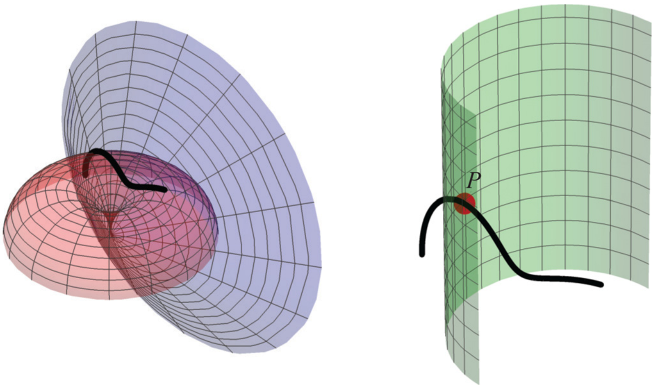


Figure 6 Finding the point of intersection of the three surfaces.

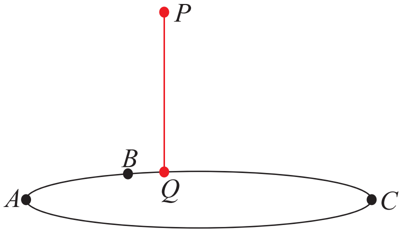


Figure 7 The point Q is on circle ABC.

When n is even, aard(n) looks a bit different from when n is odd. As we see in Figure 8, when n is odd, Q and T are on different legs of the angle at P. However, if n is even, then Q and T are on the same leg of the angle at P. It is easy to see that this presents no new difficulty, and aard(n) functions just fine in this case. However, we shall never need to use aard(n) for n even since, as we shall see, if n is even then aard(n) never gives us anything new that we could not obtain using aard(m) for some $m < n$. In particular, whereas aard(2) allows us to find square roots, we are able to find square roots using just the $2D$ Euclidean tools. Thus, aard(2) provides us with nothing new.

How can we describe the set of real numbers that are constructible if we allow ourselves to use aard(n) for various values of n? For $n \geq 2$, we say that a real number is aard(n)-constructible if we can use aard(n) together with the Euclidean tools to construct a line segment having length equal to (the absolute value of) that real number. But what does it mean to "use aard(n)"?

Referring again to Figure 8, we described using aard(7) by fixing point Q so that $|PQ| = 1$ and then varying the angle at P until $|PT| = c$ for some specified c. Then we shall say that all of the other line segments in aard(7), that is,

$$PQ1,\ PQ2,\ \ldots,\ PQ_6,\ QQ_1,\ Q_1Q_2,\ \ldots,\ Q_6T,$$

are each aard(n)-constructible line segments, and therefore the lengths of each of these line segments is an aard(n)-constructible real number (as is the negative of each of these numbers).

We shall allow ourselves to ignore certain points in the aard(n). For example (still referring to Figure 8), we can choose Q so that $|PQ| = 1$, ignore Q_6 and T, and then vary the angle at P until $|PQ_5| = c$ for some $c > 1$. Of course, we shall also allow ourselves to reverse the roles of Q and T, as described above, when $0 < c < 1$. For

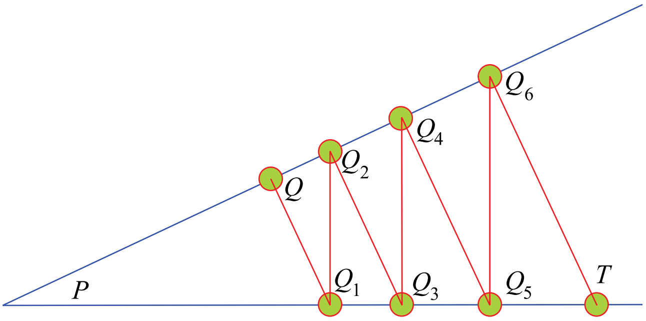


Figure 8 The adjustable angle root device aard(7).

example, we can choose T so that $|PT| = 1$, ignore Q, Q_1, and Q_2, and then vary the angle at P until $|PQ_3| = c$.

What real numbers are aard(n)-constructible? We know that the set of such numbers includes $\mathbb{Q}$ (since we are still allowing ourselves to use the $2D$ Euclidean tools). For any positive integer n, we have seen that aard(n) allows us to construct the nth root of any number previously constructed. Does aard(n) provide anything more than just nth roots?

For any previously constructed real number c, aard(7) gives us $\sqrt[7]{c} = c^{1/7}$ and (looking at PQ_2, PQ_3, PQ_4, PQ_5, and PQ_6 in Figure 8), we see that aard(7) also provides us with

$$c^{2/7}, \quad c^{3/7}, \quad c^{4/7}, \quad c^{5/7}, \quad c^{6/7}.$$

However, since our $2D$ Euclidean tools alone allow us to multiply, these other six numbers need not concern us; once we know that $c^{1/7}$ is aard(7)-constructible, we also know that $c^{2/7}$, $c^{3/7}$, $c^{4/7}$, $c^{5/7}$, $c^{6/7}$ are aard(7)-constructible. However, aard(7) does allow us to construct something besides 7th roots. For example, it also allows us to compute 5th roots, since we can fix Q so that $|PQ| = 1$, ignore Q_6 and T, and adjust the aard(7) so that $|PQ_5| = c$, for some previously constructed c. Then $|PQ_1| = c^{1/5}$.

It is not hard to see that nothing beyond kth roots for $k \leq n$ can be constructed using aard(n), and we conclude that the set of all real numbers constructible using aard(n) is the smallest field that contains the rationals and is closed under kth roots for every $k \leq n$. For $n = 7$, this is the field

$$\mathbb{Q}[\sqrt[2]{}, \sqrt[3]{}, \sqrt[4]{}, \sqrt[5]{}, \sqrt[6]{}, \sqrt[7]{}].$$

We next note that we need only consider aard(n) for prime numbers n. For example, if we wish to construct the 6th root of some previously constructed number c, we can do so using aard(3) since we can construct square roots and cube roots with aard(3), and $\sqrt[6]{c} = \sqrt[3]{\sqrt[2]{c}}$. Thus, we need only use aard(n) for n a prime number, and we can therefore simplify our field notation. For example, when we write

$$\text{“}\mathbb{Q}[\sqrt[2]{}, \sqrt[3]{}, \sqrt[4]{}, \sqrt[5]{}, \sqrt[6]{}, \sqrt[7]{}]\text{”},$$

we do not need to include $\sqrt[6]{}$ since, for any previously constructed c, we have that $\sqrt[6]{c}$ is already present in $\mathbb{Q}[\sqrt[2]{}, \sqrt[3]{}]$. Thus, we need only consider fields of the form

$$\mathbb{Q}[\sqrt[p_1]{\ },\sqrt[p_2]{\ },\ldots,\sqrt[p_n]{\ }]$$

for primes $p_1, p_2, \ldots, p_n$.

If $p_1, p_2, p_3, \ldots$ is a listing of all the prime numbers, then we shall let

$$\mathbb{Q}[\sqrt[p_1]{\ },\sqrt[p_2]{\ },\sqrt[p_3]{\ },\ldots]$$

denote the smallest field that includes $\mathbb{Q}$ and is closed under p_1 roots, p_2 roots, p_3 roots, and so on. Equivalently,

$$\mathbb{Q}[\sqrt[p_1]{\ },\sqrt[p_2]{\ },\sqrt[p_3]{\ },\ldots] = \mathbb{Q} \cup \mathbb{Q}[\sqrt[p_1]{\ }] \cup \mathbb{Q}[\sqrt[p_1]{\ },\sqrt[p_2]{\ }] \cup \mathbb{Q}[\sqrt[p_1]{\ },\sqrt[p_2]{\ },\sqrt[p_3]{\ }] \cup \ldots.$$

From our discussion above, we see that this field can also be described as the field of all real numbers that can be constructed by beginning with the rationals, and then using only the $2D$ Euclidean tools and any aard(p), for p a prime. It is the countable union of countable sets, and is therefore countable.

It is convenient to include $\sqrt[2]{\ }$ in our notation above, even though the construction of square roots requires only the $2D$ Euclidean tools, not an aard. We can view aard(2) as giving us an alternative method to compute square roots.

Generalizing Archytas′ construction: Postulates for constructions in three or more dimensions

We earlier presented Archytas' $3D$ construction, which solved the doubling-the-cube problem. We now ask: is Archytas' solution in the spirit of Euclidean constructions in two dimensions? The answer we must give is "yes and no." On the one hand, it is obviously something quite different from a construction using only the Euclidean tools since it takes place in three dimensions and Euclid only gives postulates for constructions in two dimensions. On the other hand, the methods used in this approach, such as erecting a line through a given point perpendicular to a given plane, or rotating a semicircle about a given line to form a half-torus, certainly appear to be natural generalizations of $2D$ Euclidean constructions, such as drawing a line through a given point perpendicular to a given line, or rotating a given point about another point to form a circle. Also, in contrast to the mechanical solutions we presented, Archytas' construction results in immovable objects. It is the kind of construction of which Plato might approve.

It is reasonable to ask why Euclid did not include postulates for the construction of geometric objects in three dimensions. The easy, but incorrect, response would be that Euclid only gave postulates for two dimensions because these were the only sorts of constructions with which he was concerned. While it is true that most of the geometric constructions in the *Elements* are $2D$ constructions, two of the most historically important topics in the *Elements* involve objects in $3D$ space. In Book XII of the *Elements*, Euclid analyzes the volumes of cones, pyramids, and cylinders, using the method of exhaustion, which can be thought of as a precursor to modern integral calculus. In Book XIII of the *Elements*, Euclid constructs the five Platonic solids (the cube, octahedron, tetrahedron, icosahedron, and dodecahedron). After completing the five constructions, he proves that these are the only regular polyhedra.

Another explanation one might propose for the lack of postulates for three dimensions in the *Elements* is that one can imagine physically performing $2D$ constructions with straightedge and compass, but it is much harder to imagine actually performing

$3D$ constructions. There are two problems with this explanation. One is that Euclid never talks about actually performing constructions with straightedge and compass, but only about the *existence* of lines and circles in the plane, based on his postulates. The existence of lines, cones, spheres, and other $3D$ objects is not different in kind from the existence of $2D$ objects. The second problem with this explanation is that, if we wish to actually imagine doing $2D$ constructions with straightedge and compass, there is no reason why we cannot also imagine some sort of technological device that allows us, for example, to construct a sphere in space with a given center and containing a given point. The existence (or lack) of technological devices for such constructions should certainly not influence our mathematical thinking. Plato would be horrified if it did!

In his $3D$ constructions, Euclid uses what most would agree are reasonable operations, such as erecting a line though a given point perpendicular to a given plane, and revolving a semicircle around its diameter to form a sphere, but he never gives postulates for these $3D$ constructions. We view it as rather odd that Euclid performs these steps without postulates or rules of any sort that specify what sorts of constructions are to be allowed. This is in sharp contrast to the $2D$ situation, where Euclid begins, at the very start of the *Elements*, with postulates that clearly state what is allowed in $2D$ constructions, and then goes on to use these postulates in subsequent geometric constructions. When he begins his $3D$ constructions, he starts constructing without any statement concerning what is an allowable step. This seems to us to be a gap in the *Elements*. In this section, we discuss postulates to fill this gap.

We begin by considering postulates for constructions in three dimensions, and then we briefly consider generalizing to more than three dimensions. We wish to define $3D$ constructible points, lines, circles, curves, surfaces, and real numbers. In order to generalize to three dimensions Euclid's postulates for two dimensions, we propose the following:

> We assume that we are given three points that are the vertices of an equilateral triangle. (This generalizes the two points and associated line segment of Euclid's first postulate). This yields, by definition,
>
> 1. our first $3D$ constructible plane (that is, the plane determined by the triangle),
> 2. our first three $3D$ constructible line segments, and
> 3. our first three $3D$ constructible points.

How do we use these objects to generate additional $3D$ constructible objects? We wish to generalize to three dimensions Euclid's notion of using two points to determine a line or a circle in two dimensions. We do this according to the following rules:

1. The usual Euclidean $2D$ constructions of lines and circles from two points can be done in any constructible plane.
2. As in the $2D$ case, any two $3D$ constructible points determine a $3D$ constructible line.
3. Any three noncollinear $3D$ constructible points determine a $3D$ constructible plane.
4. The intersection of $3D$ constructible objects is a $3D$ constructible object. (The intersection could be a point, line, conic section, etc.)
5. Any object obtained by rotating a $3D$ constructible object about a $3D$ constructible line is $3D$ constructible.

We can define the collection of $3D$ constructible objects more formally, as we earlier did for $2D$ constructability, as either

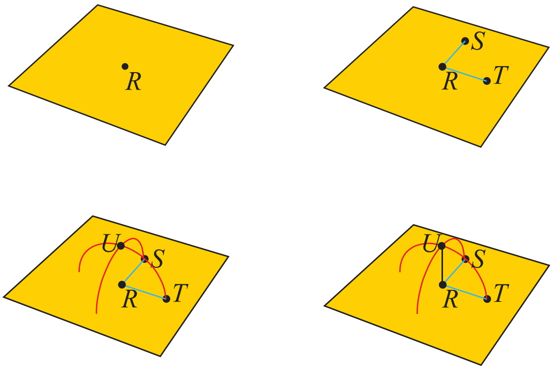


Figure 9 A $3D$ construction, using our five rules.

1. the smallest set that contains the given three points, three line segments, and plane, and is closed under a finite number of applications of the five rules listed above; or
2. the union of a countable sequence of sets, the first of which includes only the original $3D$ constructible three points, three line segments, and plane, and each subsequent set includes the previous set together with all objects $3D$ constructible from that previous set by using just one of the above rules one time.

Finally (again, as with $2D$ constructions), we say that a real number r is $3D$ constructible if and only if there is a $3D$ constructible line segment of length $|r|$, where we define our unit length to be the length of one of the three original $3D$ constructible line segments.

Precisely as in two dimensions, formulation 2 implies that the set of $3D$ constructible objects is countable, and thus the set of $3D$ constructible real numbers is countable.

Since the $2D$ Euclidean tools alone allow us to add, multiply, and find multiplicative inverses, it is not hard to see that the set of $3D$ constructible real numbers satisfies the field conditions. Hence, the set of $3D$ constructible real numbers is a field.

Next, we present a simple example of a $3D$ construction, using our five rules. Suppose we are given a $3D$ constructible plane and a $3D$ constructible point R on that plane, and we wish to construct a line perpendicular to the given plane, through point R. See Figure 9.

Let S be any other $3D$ constructible point on the given plane. (There must be other $3D$ constructible points on this plane besides R, since otherwise the plane would not be $3D$ constructible.) Using standard Euclidean $2D$ methods, we connect points R and S with a line segment and find a point T on the plane such that line segments RS and RT are perpendicular to each other and are of equal length.

Next, we construct two circles. For one, we rotate the $3D$ constructible point S about the $3D$ constructible line segment RT, and for the other, we rotate the $3D$ constructible point T about the $3D$ constructible line segment RS. It is straightforward to show that these two $3D$ constructible circles have two points of intersection, one on each side of the plane. Let U be one such point of intersection. Then U is a $3D$ constructible point and the line RU is $3D$ constructible and is perpendicular to the given plane, as desired.

Some additional examples of $3D$ constructible objects are the following:

1. A $3D$ constructible sphere, obtained by rotating a $3D$ constructible circle about a $3D$ constructible diameter of that circle.

2. A $3D$ constructible plane, obtained by rotating a $3D$ constructible line about a $3D$ constructible line that intersects it and is perpendicular to it.
3. A $3D$ constructible cone, obtained by rotating a $3D$ constructible line about a $3D$ constructible line that intersects it and is not perpendicular to it.
4. A $3D$ constructible cylinder, obtained by rotating a $3D$ constructible line about a $3D$ constructible line that is parallel to it.

We note that since the set of $3D$ constructible objects includes some (but not all) planes and cones, it also includes some (but not all) conic sections.

It is straightforward to see that Archytas' construction follows our five rules for $3D$ constructions. Thus, we can say that the doubling-the-cube problem is solvable by $3D$ construction.

More generally, we recall that, in our discussion of Archytas' method, for any chosen values of AB and AC, we were able to $3D$ construct mean proportionals AQ and AP such that

$$AB : AQ = AQ : AP = AP : AC.$$

If we choose $|AC| = c > 1$ and $|AB| = 1$, then $|AQ| = \sqrt[3]{c}$. If $0 < c < 1$, then we reverse the roles of AB and AC, as we have done previously. This tells us that the field of all $3D$ constructible real numbers is closed under cube roots, and thus every element of $\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }]$ is $3D$ constructible. We conjecture that our $3D$ constructions add nothing beyond cube roots, and hence that the reverse inclusion also holds:

Conjecture 1. *The field $\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }]$ is precisely the field of $3D$ constructible real numbers.*

We close by briefly considering higher dimensions. Even though we lose our geometric intuition when we move beyond three dimensions, it is common to attempt to generalize structures and techniques from three to higher dimensions. Our present situation invites us to do exactly that.

The three relevant questions for constructions in n-dimensions for $n > 3$, are

1. With what geometric objects do we start?
2. How do we create new nD constructible objects?
3. How can we describe the set of nD constructible geometric objects and real numbers?

We shall make some preliminary comments on these questions, but shall leave the serious work to the interested reader. We first note that the answer to question 1 seems clear, but the answers to questions 2 and 3 seem harder.

With what geometric object do we start? Recall that in two dimensions, we started with two constructible points and the constructible line segment between them, and in three dimensions, we started with three equidistant constructible points (which form an equilateral triangle) and the constructible line segments and plane determined by these points. Hence, it seems natural that in four dimensions, we start with four equidistant constructible points (which form a regular tetrahedron) and the constructible line segments, $2D$ planes, and $3D$ hyperplane determined by these points. ("Hyperplane" is the natural generalization of "plane" to higher dimensions. In this case, it is just the $3D$ space in which the tetrahedron exists, embedded in $4D$ space). In general, for $n + 1$-dimensions, we start with the regular n-simplex, which is an n-dimensional object in

$n+1$-dimensional space. A line segment is a regular 1-simplex, an equilateral triangle is a regular 2-simplex, and a regular tetrahedron is a regular 3-simplex.

How do we create new n-constructible objects? The first four of the five rules we listed for creating new 3-D constructible objects generalize in a natural way to higher dimensions. The fifth rule for creating new 3-D constructible objects concerns rotations. What exactly do we mean by rotations in higher dimensions? We invite the reader to consider this.

Open Question 1. *What are the allowable rotations for nD constructions when $n \geq 3$?*

Of course, we shall say that a real number r is nD constructible if and only if there is an nD constructible line segment of length $|r|$, where we define our unit length to be the length of one of the original nD constructible line segments. What can be said about the size of the set of nD constructible real numbers, even if we have not precisely defined how nD objects get constructed? First, as in two and three dimensions, if we view the collection of nD constructible objects as a set built up from a finite collection of nD constructible objects (that is, the appropriate beginning points, line segments, etc.) in countably many stages, we see that there are countably many nD constructible objects, and therefore countably many nD constructible real numbers.

We know that the collection of $2D$ constructible real numbers is $\mathbb{Q}[\sqrt[2]{\ }]$, and we conjectured that the collection of $3D$ constructible real numbers is $\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }]$. This suggests that the set of $4D$ constructible real numbers is $\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }, \sqrt[4]{\ }]$. This may be true, but it is less significant than might first appear since a fourth root is a square root of a square root, and hence $\mathbb{Q}[\sqrt[2]{\ }]$, the set of $2D$ constructible real numbers, is already closed under fourth roots (of nonnegative numbers). Thus, even though we have not given a complete definition of $4D$ constructability, we ask the following:

Open Question 2. *Does $4D$ constructability give us any real numbers beyond what we are given by $3D$ constructability? In other words (assuming the truth of our earlier conjecture asserting that the field of $3D$ constructible real numbers is the field $\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }]$), are there $4D$ constructible real numbers that are not in the field $\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }]$?*

If, indeed, nothing new is obtained when we move from the $3D$ to the $4D$ context, then it seems reasonable to conjecture that the hierarchy of n-constructible real numbers looks precisely like the hierarchy of aard(n)-constructible real numbers discussed previously. In other words:

Conjecture 2. *Fix any $n > 1$ and any real number r, and let p be the largest prime number less than or equal to n. Then*

1. *The number r is nD constructible if and only if it is pD constructible.*
2. *The number r is nD constructible if and only if it is aard(n)-constructible.*

Can we construct more real numbers if we broaden our notion of "construction" still further? We earlier saw that for any prime p, the set of all real numbers constructible using aard(p) is the field

$$\mathbb{Q}[\sqrt[2]{\ }, \sqrt[3]{\ }, \dots, \sqrt[p]{\ }],$$

and thus the set of all real numbers constructible using aard is

$$\mathbb{Q}[\sqrt[p_1]{\ },\ \sqrt[p_2]{\ },\ \sqrt[p_3]{\ }, \dots],$$

the union of all such fields over all primes p. Assuming the truth of our conjecture, this tells us that

$$\mathbb{Q}[\sqrt[p_1]{\ },\ \sqrt[p_2]{\ },\ \sqrt[p_3]{\ }, \dots]$$

is also the collection of all real numbers that are nD constructible for some value of n. We know that

$$\mathbb{Q}[\sqrt[p_1]{\ },\ \sqrt[p_2]{\ },\ \sqrt[p_3]{\ }, \dots]$$

is countable. Since there are uncountably many real numbers, it follows (continuing to assume the truth of the conjecture) that there are many real numbers that are not constructible using either our aards or our nD constructions. Thus we close with an open question:

Open Question 3. *Are there other methods that will allow us to "construct" (whatever that means in some new context) additional real numbers.*

Acknowledgments We would like to thank Jeffrey Hatley and Karl Zimmermann for helpful discussions about this paper and for making useful comments on an earlier draft. We also thank two anonymous referees and the editor for their helpful suggestions.

REFERENCES

[1] Euclid-Heath, T. L. (1956). *The Thirteen Books of Euclid's Elements*, 2nd ed. New York: Dover.

[2] Fraleigh, J. B. (2003). *A First Course in Abstract Algebra*, 7th ed. London: Pearson.

[3] Heath, T. L. (1981). *From Thales to Euclid: A History of Greek Mathematics*, Vol. 1. New York: Dover.

Summary. It is known that the three classical geometric construction problems introduced by the ancient Greeks: trisecting an angle, squaring a circle, and doubling a cube, cannot be solved using the Euclidean tools. However, ancient Greek mathematicians solved these three problems using other means. We present solutions to the doubling-the-cube problem using ideas that go beyond the Euclidean tools, and we consider generalizations to higher dimensions.

JULIUS BARBANEL received his Ph.D. from the State University of New York, Buffalo, in 1979. He spent almost all of his academic career at Union College, from which he retired in 2015. He began his mathematical research in set theory and later studied fair division. He also developed interests in ancient Greece, and in particular their mathematics. He enjoys cycling and cross-country skiing, and is presently trying to learn ancient Greek.

Zindler Points of Triangles

ALLAN BERELE
DePaul University
Chicago, IL 60614
aberele@depaul.edu

STEFAN CATOIU
DePaul University
Chicago, IL 60614
scatoiu@depaul.edu

We set the stage with two classical theorems on area bisecting lines in plane convex sets. The first is due to Zindler [10] (1920) and was popularized by Courant and Robbins [8] (1941); it is the first theorem in convex geometry. The second is due to Buck and Buck [6] (1949); it is one of the earliest sixpartite problems, the genre that led to the foundation of the modern theory of equipartitions. For more sixpartite results, see our earlier articles [1, 3, 4] and the one by Ceder [7].

Theorem (Zindler)**.** In any plane convex set, there are two perpendicular lines dividing the set into four parts of equal area.

Theorem (Buck and Buck)**.** In any plane convex set, there are three concurrent lines dividing the set into six parts of equal area.

Both theorems are proved by continuity, using basic properties of area bisecting lines of convex sets, such as: (1) in each direction there is a unique area bisector; (2) through each point on the perimeter there passes a unique area bisector; and (3) if a point moves continuously counterclockwise around the perimeter, then so does the other endpoint of the area bisecting line through it.

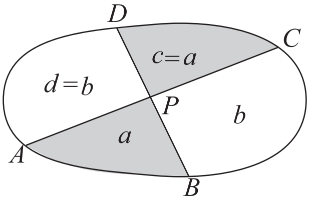


Figure 1 Proof of Zindler's theorem.

Proof of Zindler's theorem. Let A, B, C, D be four points arranged in counterclockwise order on the perimeter, such that AC is an area bisector and BD is the area bisector perpendicular to it, as shown in Figure 1. Let P be the intersection of AC and BD. Denote by a, b, c, d, respectively, the areas of the sectors APB, BPC, CPD, DPA. Then AC and BD are area bisectors precisely when $a + b = c + d$ and $a + d = b + c$. This is equivalent to $a - c = d - b$ and $a - c = b - d$, with solution $c = a$ and $d = b$. Clearly, if AC has directional angle θ, then BD has directional angle $\theta + \frac{\pi}{2}$. Moreover, the points A, B, C, D are continuous functions of θ, and so are the amounts $a = a(\theta)$, $b = b(\theta)$ and $f(\theta) := a(\theta) - b(\theta)$. Since

$$f\left(\frac{\pi}{2}\right) = a\left(\frac{\pi}{2}\right) - b\left(\frac{\pi}{2}\right) = b(0) - a(0) = -f(0),$$

Math. Mag. **95** (2022) 482–495. doi:10.1080/0025570X.2022.2127301 © Mathematical Association of America
MSC: 51M04

This article was originally published with errors, which have now been corrected in the online version. Please see Correction (http://dx.doi.org/10.1080/0025570X.2022.2146384).

by the intermediate value property, f has a root τ in the interval $[0, \frac{\pi}{2}]$. It follows that $a(\tau) = b(\tau)$, and the result follows easily from here. ■

Buck and Buck's theorem also follows from the intermediate value theorem using a subtler argument based on three lines instead of two.

The proofs of these two theorems are charming, but they do not give much insight about these area bisecting lines or their points of intersection, such as how many there are or how to find them. One basic case to consider would be triangles. How do these theorems play out in this case? In Buck and Buck's theorem, there is only one way to divide a triangle into six equal pieces using three concurrent lines: by using the three medians (see our article [2]).

The case of Zindler's theorem is different. Consider the example of equilateral triangles: Each median is a line of symmetry and, together with an area bisecting line perpendicular to it, divides the triangle into four equal parts. Let us call a pair of perpendicular lines dividing a triangle into four equal parts a *Zindler pair* and the intersection point of a Zindler pair a *Zindler point*. Then an equilateral triangle has at least three Zindler points and it is reasonable to expect that triangles close to equilateral will have nearly as many. We will prove that this is indeed the case and more.

This article follows Zindler's theorem in the case of a triangle. Our main result, Theorem 2, shows that a triangle close to equilateral will have two or three Zindler points, and other triangles will have only one. In particular, the equilateral triangle has exactly three Zindler points.

Preliminary remarks

A good insight on how our main theorem works is given by its particular, more explicit version for isosceles triangles:

Theorem 1. *Let an isosceles triangle have base a and height h. Then the triangle will have three Zindler points if $\sqrt{3}/2 \le h/a < 8/9$, two Zindler points if $h/a = 8/9$, and one Zindler point otherwise.*

In terms of being close to equilateral, this theorem can be restated in these equivalent ways: An isosceles triangle has two or three Zindler points if either the summit angle is between 60° and $2\tan^{-1} 9/16 \approx 58.7°$; or if the base angles are between 60° and $\tan^{-1} 16/9 \approx 60.6°$; or if the ratio of the lateral side to the base is between 1 and $\frac{1}{18}\sqrt{337} \approx 1.02$. See Figure 2 for an example with three Zindler pairs, and note that one Zindler pair involves the symmetry axis while the other two are mirror images about this axis. The case with two Zindler points is obtained from this example by fixing a and increasing h. That is, by making the triangle move as far as possible from equilateral. In this way, the two Zindler points that are mirror images of each other will get closer to each other, so that when $a/h = 8/9$, they will coincide on the symmetry axis; see Figure 8. At the same time, the Zindler pairs corresponding to the two moving Zindler points will become a single Zindler pair of area bisectors making 45° angles with the base.

Back to scalene triangles, given a Zindler pair of lines in a triangle, one closed side will intersect both of them. Let us say in this case that the Zindler pair or the Zindler point is *supported* by this side. In a scalene triangle, no Zindler pair is supported by two sides, as we will show. We proved that each Zindler point lies on the intersection of only one Zindler pair in a previous paper [5].

The following theorem describes how many Zindler pairs there are supported by a given side of a triangle. It implies that every triangle has at most three Zindler pairs

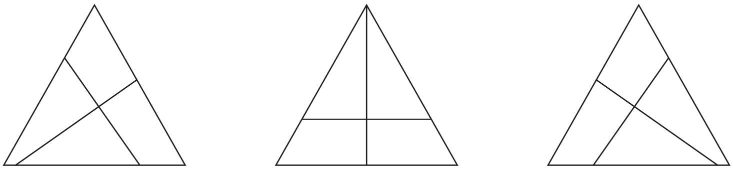


Figure 2 An isosceles triangle with three Zindler pairs.

and determines how many there are. The third part of the theorem refers to the three-sided region $\mathcal{D}$ in Figure 3. In this figure, the vertices are $(0, \sqrt{3})$ connected by the y-axis to $(0, 16/9)$, connected by a certain algebraic curve we will describe later to $\left(3 - 2\sqrt{2}, \sqrt{8\sqrt{2} - 8}\right)$, which connects back to $(0, \sqrt{3})$ along an arc of the circle with center $(1, 0)$ and radius 2. The triangle will be isosceles if A is on the lower or left edge of $\mathcal{D}$, so we need not consider those cases here.

Theorem 2. *In a scalene triangle, no Zindler pair is supported by two sides. Moreover, we have the following:*

1. *No Zindler pair is supported by the longest side.*
2. *Exactly one Zindler pair is supported by the side of middle length.*
3. *Let BC be the shortest side of a triangle with vertices $A = (a, h)$, $a \geq 0$, $h > 0$, $B = (-1, 0)$ and $C = (1, 0)$. Then BC will support two Zindler pairs if A is in the interior of the region $\mathcal{D}$ in Figure 3; one Zindler pair if it is on the upper edge of the region, but not the upper endpoint; and no Zindler pairs in any other case.*

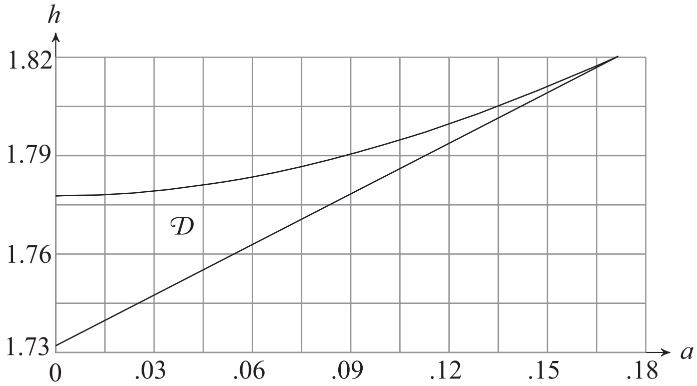


Figure 3 Region $\mathcal{D}$ described by the vertex A of a triangle ABC when a Zindler pair is supported by the shortest side BC, where $B = (-1, 0)$, $C = (1, 0)$ and $AB \geq AC$.

This theorem implies that if a triangle has more than one Zindler point, then the ratio of the longest to the shortest side is at most (approximately) 1.082, that the largest angle is at most (approximately) 65.5° and the smallest angle is at least (approximately) 57.2°. The extreme values correspond to vertex A close to the upper right vertex of the region $\mathcal{D}$ in Figure 3. All of which means that the triangle is close to equilateral.

Theorem 2 implies that the number of Zindler points in a scalene triangle is 1 plus the number of Zindler points supported by the shortest side. As an immediate consequence of this we get:

Corollary 3. *Every triangle has between one and three Zindler points, and it has more than one if and only if it is close to equilateral, as specified in Theorems 1 and 2.*

The goal of the paper is both to prove the main theorem and to better understand the region $\mathcal{D}$ to which it refers. Theorem 2 and its proof provide a method to compute the Zindler lines by solving algebraic equations. The two properties of a pair of area bisecting lines of a triangle that makes them a Zindler pair—namely that they are perpendicular and that they divide the area into fourths—are translated in the next section into two equations. The system of these two equations is then solved, leading to the proof of Theorem 2. We then show that in general the Zindler pairs of a triangle cannot be constructed by compass and straightedge. The final section interprets the upper side of the region $\mathcal{D}$ in Figure 2 as an envelope of a family of circles.

Algebraic description of Zindler points

In a triangle ABC, the median AA_1 is an area bisecting line. By property (3) for area bisecting lines of general convex sets outlined in the introduction, every other area bisecting line with endpoint on BC will either connect a point on BA_1 with one on AC, or a point on CA_1 with one on AB. In particular, if a Zindler pair consists of two area bisectors with endpoints on the same half of BC, then the pair would be supported by two sides of the triangle. Equivalently, there would be a side of the triangle at which neither lines of the Zindler pair intersect. The next lemma says that this cannot happen.

Lemma 4. *In every triangle, each Zindler pair intersects all three sides.*

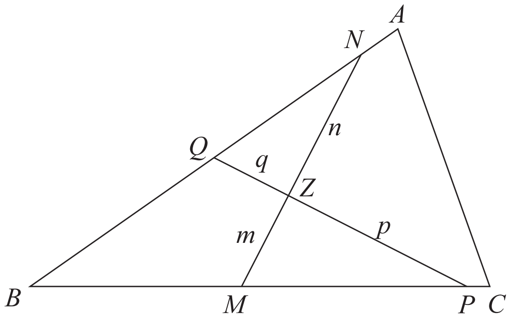


Figure 4 Proof of Lemma 4.

Proof. Suppose not. Then there is a triangle ABC with a Zindler point Z supported by two sides, say AB and BC. Let MN and PQ be a Zindler pair through Z, with $M, P \in BC$ and $N, Q \in AB$, as shown in Figure 4, and suppose that none of the points M, N, P, Q is a vertex of the triangle. Denote by m, n, p, q respectively the lengths of the segments MZ, NZ, PZ, QZ. The triangles ZNQ and ZMP have equal area. This area is greater than the area of triangle QZM and is at least the area of triangle NZP, and so

$$qn = pm > qm \qquad \text{and} \qquad qn = pm \geq pn.$$

The inequalities $qn > qm$ and $qn \geq pn$ imply that $n > m$ and $q \geq p$, while the inequalities $pm > qm$ and $pm \geq pn$ imply that $p > q$ and $m \geq n$, two contradictions. ■

Consider the triangle with vertices

$$A = (a, h), \qquad B = (-1, 0), \qquad C = (1, 0),$$

with $a \geq 0, h > 0$, and look at area bisecting lines intersecting BC that are part of a Zindler pair supported by BC (see Figure 5). First, consider such lines DE joining $D = (-t, 0)$, for $t > 0$, to a point E on AC. Since $DC = t + 1$, such a line will be an area bisector when E has y-coordinate $h/(t + 1)$. The condition that E is on AC will then make its x-coordinate $(t + a)/(t + 1)$, implying that DE has equation

$$y = \frac{h}{t^2 + 2t + a}(x + t) \tag{1}$$

By a similar computation, the area bisecting line joining $(s, 0)$ to a point on AB will have equation

$$y = -\frac{h}{s^2 + 2s - a}(x - s). \tag{2}$$

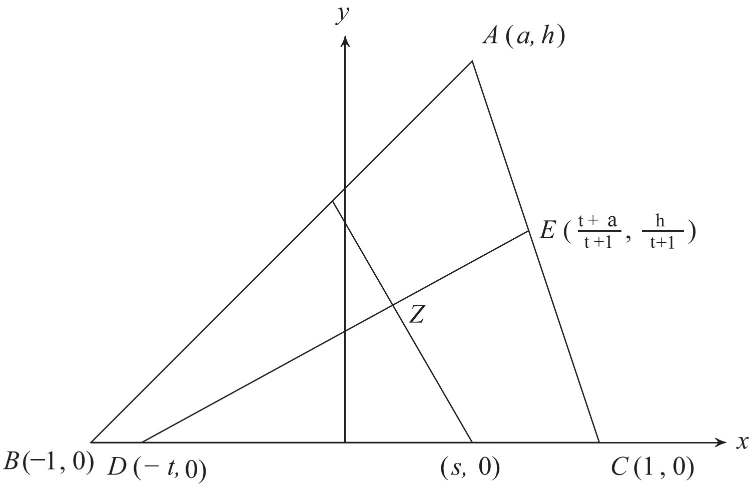


Figure 5 Coordinates of a Zindler pair.

In order for equations (1) and (2) to form a Zindler pair they must be perpendicular, and, together with the x-axis, they must make a triangle with area $h/4$. The lines will be perpendicular when the product of their slopes is -1. By equations (1) and (2), this is equivalent to

$$(s^2 + 2s - a)(t^2 + 2t + a) = h^2. \tag{3}$$

As for the area condition, the triangle will have base $s + t$ and so must have height $\frac{h}{2(s+t)}$. The height will be the y-coordinate of the intersection of the two lines, namely,

$$\frac{(t^2 + 2t + a)y}{h} - t = -\frac{(s^2 + 2s - a)y}{h} + s$$

and so

$$y = h(s + t)(s^2 + t^2 + 2s + 2t)^{-1}.$$

It follows that the two lines divide the triangle into equal parts when

$$2(s+t)^2 = s^2 + t^2 + 2s + 2t,$$

or

$$s^2 + 4st + t^2 - 2s - 2t = 0. \tag{4}$$

Note that equation (4) has no dependence on a or h. Its graph is a hyperbola with one branch going through the square $[0, 1] \times [0, 1]$, intersecting the top edge of the square at $(\sqrt{2} - 1, 1)$ and the right edge at $(1, \sqrt{2} - 1)$. Hence, we can sharpen the requirement $0 \leq s, t \leq 1$ to

$$\sqrt{2} - 1 \leq s, t \leq 1. \tag{5}$$

The hyperbola is symmetric about the line $s = t$, which is its major axis, and has vertex at $(2/3, 2/3)$. It follows that $s + t$, which we will need later, is minimal at $(2/3, 2/3)$ and maximal at both $(\sqrt{2} - 1, 1)$ and $(1, \sqrt{2} - 1)$. Therefore,

$$\frac{4}{3} \leq s + t \leq \sqrt{2}. \tag{6}$$

We summarize the results of this section in a lemma.

Lemma 5. *The number of Zindler pairs of a triangle ABC supported by the side BC is the number of solutions to the system of equations (3) and (4) with each of s and t between $\sqrt{2} - 1$ and 1. Geometrically, it is the number of intersection points of the graphs of equations (3) and (4) in the square* $[0, 1] \times [0, 1]$.

These intersection points, that is, the solutions to the system of equations (3) and (4), are studied in the next section.

Intersection points

The curve given by equation (4) depends only on s and t, whereas equation (3) depends on the four variables s, t, a, and h. In the (s, t) plane, we view the intersection of the graph of equation (4) with the 1×1 square as a fixed curve Γ and the intersection of the graph of equation (3) with the 1×1 square as a family of curves $\Gamma(a, h)$. By Lemma 5, the number of Zindler pairs of a triangle ABC, with $A = (a, h)$, $B = (-1, 0)$ and $C = (1, 0)$, that are supported by BC is the number of intersection points of Γ and $\Gamma(a, h)$. Such a count is done in the next section. This section has a different focus, which we now describe.

In general, how could the number of intersection points between a fixed curve C and a family of curves $C(x)$ change? Intuitively, there are two ways: Either at a point of tangency or at an endpoint.

When two tangent circles are slightly moving away or toward each other, their number of intersection points changes from 1 to either 0 or 2. When two line segments intersecting at one of their endpoints move slightly away or toward each other, their number of intersection points either stays constant at 1, or it changes. In this section, we would like to understand these two situations for our two curves, that is, to determine all points (a, h) where the intersection of Γ and $\Gamma(a, h)$ contains either tangency points (s, t) or points (s, t) that lie on the border of the 1×1 square. Specifically, we are describing the following subsets of the first quadrant in variables a and h:

$$\Lambda_1 = \{(a, h) \mid \Gamma \cap \Gamma(a, h) \text{ contains tangency points } (s, t)\},$$
$$\Lambda_2 = \{(a, h) \mid \Gamma \cap \Gamma(a, h) \text{ contains points } (s, t) \text{ with } s \text{ or } t = 1\}.$$

The result of Lemma 6 is that Λ_1 is a closed curve consisting of two arcs of circles, and the result of Lemma 8 and the paragraph following it is that Λ_2 is the algebraic curve bordering the region $\mathcal{D}$ in Figure 3, whose equation is now given explicitly in parametric form.

Lemma 6. *If $s = 1$ in equation (4), then $t = \sqrt{2} - 1$ and (a, h) in equation (3) is on the circle with center $(1, 0)$ and radius 2, and if $t = 1$, then $s = \sqrt{2} - 1$ and (a, h) is on the circle with center $(-1, 0)$ and radius 2.*

Proof. If $s = 1$ then equation (4) becomes $t^2 + 2t - 1 = 0$, whose only positive root is $\sqrt{2} - 1$. Substituting these values into equation (3) gives

$$(3 - a)(1 + a) = h^2 \qquad \text{or} \qquad a^2 + h^2 - 2a = 3.$$

Adding 1 to both sides to complete the square gives the circle with center $(1, 0)$ and radius 2, which is the first half of the lemma. The second half is similar. ■

Turning now to the question of when the curves are tangent, we use implicit differentiation on each considering s as a function of t. Equation (3) yields

$$(2s + 2)(t^2 + 2t + a)ds/dt + (s^2 + 2s - a)(2t + 2) = 0,$$

and equation (4) yields

$$(2s + 4t - 2)ds/dt + (4s + 2t - 2) = 0.$$

Hence, the two curves will have the same slope when

$$(2s + 2)(t^2 + 2t + a)(4s + 2t - 2) = (2s + 4t - 2)(s^2 + 2s - a)(2t + 2),$$

which simplifies to

$$(t - s)(t^2 s + ts^2 + t^2 + s^2 + s + t - 2) + 2a(t^2 + ts + s^2 + t + s - 1) = 0. \tag{7}$$

This gives us another lemma:

Lemma 7. *If the curves Γ defined by equation (4) and $\Gamma(a, h)$ defined by equation (3) are tangent at a point, then equation (7) must hold at that point.*

We now turn to the solutions of the system of these three equations.

Lemma 8. *The curves Γ defined by equation (4) and $\Gamma(a, h)$ defined by equation (3) are tangent in the 1×1 square only if $0 \le a \le 3 - 2\sqrt{2}$, in which case there is a unique h for which they are tangent.*

Proof. By Lemma 7, we need to find a common solution between equations (3), (4), and (7) in the 1×1 square. Noting that equation (4) is symmetric in s and t, we let $S = s + t$ and $P = st$. Then equation (4) is equivalent to

$$S^2 + 2P - 2S = 0 \qquad \text{or} \qquad P = S - \frac{1}{2}S^2. \tag{8}$$

Equation (7) can now be written as an equation in S and P using

$$\begin{aligned} t^2s + ts^2 + t^2 + s^2 + s + t - 2 &= SP + S + S^2 - 2P - 2 \\ &= S(S - \tfrac{1}{2}S^2) + S + S^2 - 2(S - \tfrac{1}{2}S^2) - 2 \\ &= -\tfrac{1}{2}S^3 + 3S^2 - S - 2, \end{aligned}$$

and

$$\begin{aligned} t^2 + ts + s^2 + t + s - 1 &= S^2 - P + S - 1 \\ &= S^2 - (S - \tfrac{1}{2}S^2) + S - 1 = \tfrac{3}{2}S^2 - 1. \end{aligned}$$

Equation (7) now becomes

$$(t-s)(-\frac{1}{2}S^3 + 3S^2 - S - 2) + 2a(\frac{3}{2}S^2 - 1) = 0.$$

The variable $S = s + t$ has a minimum of 4/3 obtained at (2/3, 2/3) and a maximum of $\sqrt{2}$ obtained at both $(\sqrt{2} - 1, 1)$ and $(1, \sqrt{2} - 1)$ (see equation (6)). It is not hard to see that the coefficients of $(t - s)$ and $2a$ are both positive there, so in order for a to be positive $t - s$ must be negative. Hence,

$$t - s = -\sqrt{s^2 - 2st + t^2} = -\sqrt{S^2 - 4P},$$

which by equation (8) equals $-\sqrt{3S^2 - 4S}$. Substituting into equation (7) and solving for a yields

$$a = -\frac{(\frac{1}{2}S^3 - 3S^2 + S + 2)\sqrt{3S^2 - 4S}}{3S^2 - 2}. \tag{9}$$

We claim that a is a strictly increasing function of S on the interval $[4/3, \sqrt{2}]$. To see this, we take the derivative. The denominator of the derivative is $(3S^2 - 2)^2$, which is positive, so we need only consider the numerator, which equals the negative of

$$\begin{aligned} &(\tfrac{3}{2}S^2 - 6S + 1)(3S^2 - 4S)^{1/2}(3S^2 - 2) \\ &\qquad + (\tfrac{1}{2}S^3 - 3S^2 + S + 2)\tfrac{1}{2}(3S^2 - 4S)^{-1/2}(6S - 4)(3S^2 - 2) \\ &\qquad - (\tfrac{1}{2}S^3 - 3S^2 + S + 2)(3S^2 - 4S)^{1/2}(6S), \end{aligned}$$

which in turn equals $(3S^2 - 4S)^{-1/2}$ times

$$9S^6 - 36S^5 + 6S^4 + 56S^3 - 36S^2 + 8,$$

which we denote as $F(S)$ and which we want to show is negative on $[4/3, \sqrt{2}]$. A computation shows that

$$F(S + 1) = 9S^6 + 18S^5 - 38S^4 - 100S^3 - 57S^2 - 6S + 7,$$

which has only two sign changes in its coefficients. By Descartes's rule of signs, $F(S + 1)$ has at most two positive roots, hence $F(S)$ has at most two roots greater than 1. Now $F(1.2) \approx 2.7$ and $F(1.3) \approx -2.9$, so $F(S)$ has a root between 1.2 and 1.3, and $F(3) = -505$ and $F(4) = 4552$, so $F(S)$ has a root between 3 and 4. It follows that $F(S)$, and hence da/dS, is never zero. In fact, it is always strictly positive on the interval $[4/3, \sqrt{2}]$, and so a strictly increasing, as claimed. It has a minimum of 0 obtained at $S = 4/3$ and a maximum $3 - 2\sqrt{2}$ obtained at $S = \sqrt{2}$. ■

The set of (a, h) such that Γ and $\Gamma(a, h)$ are tangent in the 1×1 square can be described parametrically as follows: s takes values from $\sqrt{2} - 1$ to 1, as per equation (5), but since $t - s$ must be negative, s is further restricted to $[2/3, 1]$. Then $t = 1 - 2s + \sqrt{3s^2 - 2s + 1}$, from solving equation (4). From this, we can solve for a using equation (7) and for h using equation (3). The endpoints of the curve are $(0, 16/9)$, achieved when $s = 2/3$, and $(3 - 2\sqrt{2}, \sqrt{8\sqrt{2} - 8})$, achieved when $s = 1$.

Proofs of the main theorems

Referring to Figure 6, the first quadrant is divided into four regions, which we label $\mathcal{A}$–$\mathcal{D}$, by three curves, the first two of which are assured by Lemma 6 and the third by Lemma 8 and the paragraph following it:

- one is an arc of a circle with center $(-1, 0)$ and radius 2, corresponding to (a, h) for which Γ and $\Gamma(a, h)$ intersect at $t = 1$;
- one is an arc of a circle with center $(1, 0)$ and radius 2, corresponding to (a, h) for which Γ and $\Gamma(a, h)$ intersect at $s = 1$;
- and one corresponds to (a, h) at which Γ and $\Gamma(a, h)$ are tangent.

The first two curves form the curve Λ_1 and the third is the curve Λ_2 defined at the beginning of Section 2.

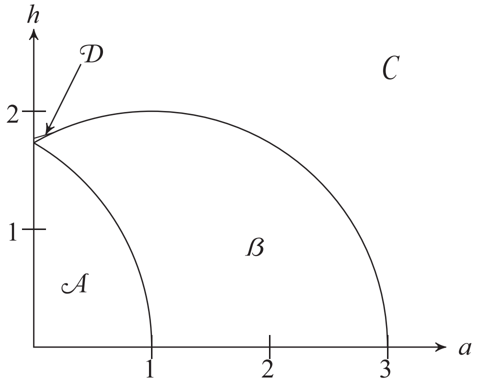


Figure 6 The four regions.

Recall that a function defined on a connected open set is constant if and only if it is locally constant, that is, constant in a neighborhood of each point of its domain.

Lemma 9. *The number of intersection points of Γ and $\Gamma(a, h)$ is constant in each of the connected open sets $\mathcal{A}$–$\mathcal{D}$.*

Proof. We want to show that the number of intersection points of Γ and $\Gamma(a', h')$ is constant in an open ball $B((a, h), \varepsilon)$ centered at (a, h) with radius ε. We will show separately that it is nondecreasing in such an open ball and nonincreasing in one.

Let Γ and $\Gamma(a, h)$ intersect at the points $P_1, \ldots, P_n$, none of which are endpoints of Γ and such that Γ and $\Gamma(a, h)$ do not have equal slopes at any of them. In order to see that the intersection points are continuous functions of (a, h), we apply the implicit function theorem to the function $F : \mathbb{R}^4 \to \mathbb{R}^2$, defined by

$$F(x_1, x_2, y_1, y_2) = \\ (y_1^2 + 4y_1y_2 + y_2^2 - 2y_1 - 2y_2, (y_1^2 + 2y_1 - x_1)(y_2^2 + 2y_2 + x_1) - x_2^2).$$

Then $F(P_i, a, h) = 0$, for each i. The use of the implicit function theorem at the points (P_i, a, h) requires the Jacobian

$$|J(F)| = |\partial(F_1, F_2)/\partial(y_1, y_2)|$$

to be nonzero there, which is equivalent to the hypothesis that Γ and $\Gamma(a, h)$ have different slopes at P_i. By the implicit function theorem, there exists continuous functions $g_i(x_1, x_2)$ from a neighborhood of (a, h) to $\mathbb{R}^2$ such that $g_i(a, h) = P_i$ and $F(x_1, x_2, g_i(x_1, x_2)) = (0, 0)$, for each (x_1, x_2) in that neighborhood, which means that $g_i(x_1, x_2)$ is in $\Gamma \cap \Gamma(x_1, x_2)$. Because of continuity, there exists an $\varepsilon > 0$, such that for each $(a', h') \in B((a, h), \varepsilon)$ the points $g_i(a', h')$, for $i = 1, \ldots, n$, can be made arbitrarily close to P_i, hence they are n distinct intersection points of Γ and $\Gamma(a', h')$. This completes the nondecreasing part of the proof.

To prove that the number of intersection points is nonincreasing in a neighborhood of (a, h), assume for a contradiction that there exists a sequence of points (a_i, h_i) converging to (a, h) such that each curve $\Gamma(a_i, h_i)$ intersects Γ in at least $n + 1$ distinct points. Say

$$\mathcal{Q}^{(i)} = \left(Q_1^{(i)}, \ldots, Q_{n+1}^{(i)}\right)$$

is an $(n + 1)$-tuple of such points. Since Γ (and therefore Γ^{n+1}) is a compact set, there would be a convergent subsequence, namely an $i_1 < i_2 < \cdots$ such that the $(n + 1)$-tuples $\mathcal{Q}^{(i_\alpha)}$ converge to a $\mathcal{Q} = (Q_1, \ldots, Q_{n+1})$. These Q_j would all be in $\Gamma \cap \Gamma(a, h)$, possibly some coincident, possibly some at the endpoints. But this is impossible: Γ and $\Gamma(a, h)$ do not intersect at the endpoints $t, s = 1$; and if, say $P_1 = Q_1 = Q_2$ then the chords $Q_1^{(i_\alpha)}Q_2^{(i_\alpha)}$ would converge to a common tangent to Γ and $\Gamma(a, h)$ at P_1 contradicting the assumption that they have different slopes at P_1. ■

By Lemma 9, if A_1 and A_2 are in the same one of the connected open sets $\mathcal{A}$–$\mathcal{D}$, then triangles A_1BC and A_2BC have the same number of Zindler pairs supported by BC. We now set about the task of computing these four numbers.

In order to compute the number of intersection points of Γ and $\Gamma(a, h)$ in each of these regions, one may use any convenient (a, h) point in it. In the next lemma, we compute the number of intersection points of Γ and $\Gamma(a, h)$ in regions $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ by considering the case of $h = 0$. Of course, it does not make much sense to talk about a triangle ABC with zero height, but it makes perfectly good sense to talk about the number of intersection points of Γ and $\Gamma(a, 0)$. By Lemma 9 this number will be the same as the number of intersection points of Γ and $\Gamma(a', h')$ for nearby points (a', h').

Lemma 10. *The curves Γ and $\Gamma(a, h)$ intersect in one point if (a, h) is in region $\mathcal{B}$ and in no points if (a, h) is in regions $\mathcal{A}$ or $\mathcal{C}$.*

Proof. Setting $h = 0$, equation (3) becomes $(s^2 + 2s - a)(t^2 + 2t + a) = 0$. Since a and t are positive, the second factor is nonzero. Thus, $s^2 + 2s - a = 0$, which implies $(s + 1)^2 = a + 1$. By equation (5), $\sqrt{2} - 1 \leq s \leq 1$, implying that $1 \leq a \leq 3$. Moreover, for each a in this range the equation $(s + 1)^2 = a + 1$ has one positive solution s. ■

This lemma yields the first two parts of Theorem 2, which we here demote to a lemma.

Lemma 11. *In any scalene triangle, there are no Zindler pairs supported by the longest side and exactly one supported by the side of intermediate length.*

Proof. First, let ABC be a triangle with longest side BC. We can create a coordinate system in which $B = (-1, 0)$, $C = (1, 0)$, and $A = (a, h)$, and we may assume without loss of generality that a and h are both nonnegative. Then the hypothesis that BC is the longest side of the triangle implies that the vertex A lies in region $\mathcal{A}$. By Lemma 10, Γ and $\Gamma(a, h)$ have no intersection points, and so by Lemma 5 there are no Zindler pairs supported by BC. The second case, in which BC is the middle side, is similar, except that in this case the vertex A will lie in region $\mathcal{B}$, and so there will be one Zindler pair supported by BC. ■

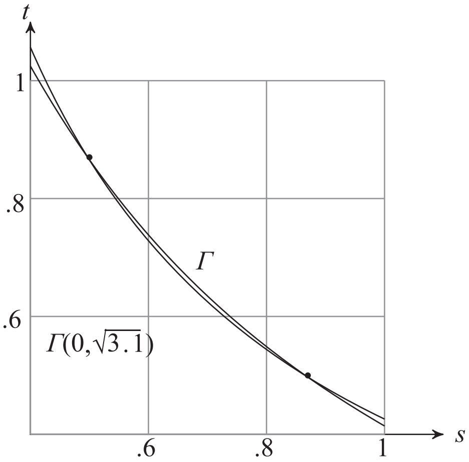


Figure 7 Γ and $\Gamma(0, \sqrt{3.1})$ intersect in two points.

There remains the case of region $\mathcal{D}$, a larger drawing of which is in Figure 3. In order to count the number of intersection points of Γ and $\Gamma(a, h)$ for (a, h) in $\mathcal{D}$, we will consider the case of $a = 0$, namely isosceles triangles. See Figure 7 for the curves Γ and $\Gamma(0, \sqrt{3.1})$.

Lemma 12. *The curves* Γ *and* $\Gamma(0, h)$ *intersect in two points for* $\sqrt{3} \le h < 16/9$ *and at one point for* $h = 16/9$.

Proof. The two curves are given by the equations

$$s^2 + 4st + t^2 - 2s - 2t = 0 \qquad \text{and} \qquad (s^2 + 2s)(t^2 + 2t) = h^2.$$

As in the proof of Lemma 8, we set $S = s + t$ and $P = st$ and recall that $S \in [4/3, \sqrt{2}]$ by (6). The first equation yields $P = S - \frac{1}{2}S^2$, which is the same as equation (8), and the second becomes $P^2 + 2SP + 4P = h^2$. Substituting the first equation into the second and simplifying leads to

$$S^4 - 8S^3 + 4S^2 + 16S - 4h^2 = 0. \tag{10}$$

We claim that $F(S) = S^4 - 8S^3 + 4S^2 + 16S$ is monotonic on the interval $[4/3, \sqrt{2}]$. For

$$F'(S) = 4S^3 - 24S^2 + 8S + 16,$$

and this cubic changes sign between -1 and 0, between 0 and 1.2, and between 5 and 6, and so cannot change sign in $[4/3, \sqrt{2}]$. Since $F(4/3) = 1024/81$ and $F(\sqrt{2}) = 12$, $4h^2$ must be between these two numbers , implying that h is between $16/9$ and $\sqrt{3}$.

Each value of $S = s + t$ and $P = st$ gives two values for the pair (s, t), unless $s = t$, which happens at $(2/3, 2/3)$. ■

Lemma 12 implies that if vertex A lies in the interior of region $\mathcal{D}$, then triangle ABC has two Zindler points supported by BC. This completes the proof of Theorems 1 and 2, except for the cases in which A lies on the boundary between two of the regions in Figure 6. Assume A lies on one of the circular arcs. Then triangle ABC is isosceles with BC a lateral side. In this case, there is an intersection point of Γ and $\Gamma(a, h)$ with s or t equal to one, corresponding to the Zindler pair in which one line is a median. If there were another intersection point P, then P would not be an endpoint or a point of tangency, implying there would be an open ball $B((a, h), \delta)$ such that for $(a', h') \in B((a, h), \delta)$ the curves Γ and $\Gamma(a', h')$ would intersect at a point near P. But this is impossible since they do not intersect for (a', h') in $\mathcal{A}$ or $\mathcal{C}$. Finally, on the curve separating $\mathcal{C}$ and $\mathcal{D}$, the curves Γ and $\Gamma(a, h)$ are tangent, and the shortest side BC supports one Zindler pair (or two coincident ones) in addition to the pair supported by the middle side. This completes the proofs of Theorems 1 and 2.

Non-constructibility

In this section, we prove that in general the Zindler point of a triangle is not constructible by straightedge and compass. In order to do this, we will prove that the solution to equation (10) does not lie in an iterated quadratic extension of the rational numbers in the case of $h^2 = 3.1$. In this case, after clearing fractions, equation (10) becomes

$$5S^4 - 40S^3 + 20S^2 + 80S - 62 = 0 \tag{11}$$

which we call $g(S)$. Before starting the proof, we recall a theorem of Dedekind, which will be our main tool [9, pp. 302–304].

Theorem. Let $f(x) \in \mathbb{Z}[x]$ be an irreducible, degree n polynomial, and let p be a prime not dividing the leading coefficient. Assume that $f(x)$ factors modulo p as $f_1 \ldots f_m$ with no factor a constant multiple of another, and with degrees $d_1, \ldots, d_m$. Then the Galois group of $f(x)$, considered as a subgroup of the symmetric group S_n, contains an element of cycle type $(d_1, \ldots, d_m)$.

Here is the main theorem of this section:

Theorem 13. *The Zindler lines of a triangle are in general not constructible by straightedge and compass.*

Proof. Consider the triangle with vertices $(\pm 1, 0)$ and $(0, \sqrt{3.1})$. If all the Zindler lines of this triangle were constructible, then the solution of equation (11) would lie in an iterated quadratic extension of $\mathbb{Q}$ and so the Galois group would have order a power of 2. Note that $g(S)$ is irreducible by Eisenstein's criterion, using the prime $p = 2$.

Reducing mod 3, equation (11) becomes

$$2S^4 + 2S^3 + 2S^2 + 2S + 1 = 0,$$

which factors as

$$(S + 2)(2S^3 + S^2 + 2) = 0.$$

The second factor is a cubic with no roots in $\mathbb{Z}_3$ and is therefore irreducible. It follows from Dedekind's theorem that the Galois group of the splitting field of $g(S)$ contains a 3-cycle and so has order a multiple of 3. ■

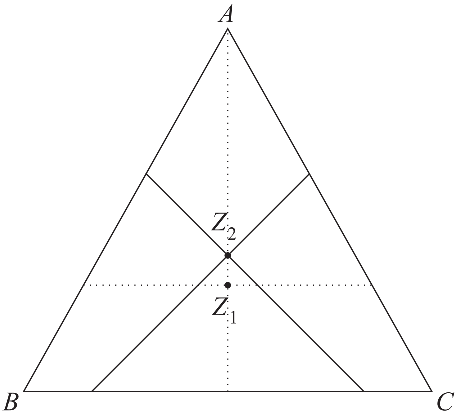


Figure 8 An isosceles triangle with two Zindler points.

Figure 8 shows the isosceles triangle ABC with $A = (0, 16/9)$, $B = (-1, 0)$ and $C = (1, 0)$, its two Zindler points Z_1 and Z_2, and their corresponding Zindler pairs, one of which makes 45° angles with the base. One easy computation is that $Z_1 = \left(0, \frac{16}{9} - \frac{8}{9}\sqrt{2}\right)$. A good exercise would be to follow the general computations we had earlier in the paper to deduce that $Z_2 = (0, \frac{2}{3})$. In particular, this isosceles triangle has two constructible Zindler points.

We leave as an open question to the interested reader the problem of finding a non-isosceles triangle having a constructible Zindler point.

Another point of view

Given s between $\sqrt{2} - 1$ and 1, there is a unique positive t so that (s, t) will be a solution to (4). Substituting these values into (3) we get the set of (a, h) so that the triangle with vertices $B, C = (\pm 1, 0)$ and $A = (a, h)$ has a Zindler pair going through $(s, 0)$ and $(-t, 0)$. These points (a, h) lie on the circle

$$a^2 + h^2 + (t^2 + 2t - s^2 - 2s)a = (s^2 + 2s)(t^2 + 2t).$$

In Figure 9 (Right) we draw a number of these (semi) circles.

The reader can see the highlighted semicircles with centers B and C and radius 2, and that the half plane is divided into four regions: In the intersection $\mathcal{A}$ of the two circles, where BC would be the largest side of triangle ABC, there are no circular

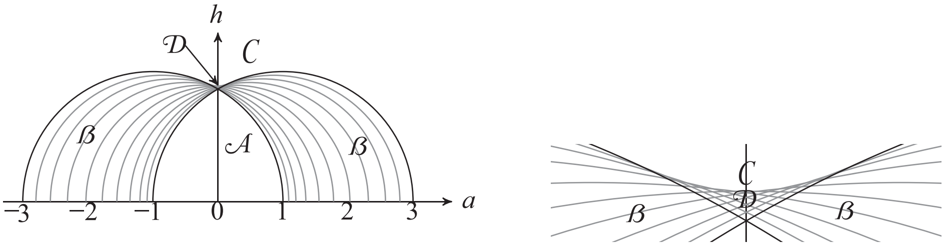


Figure 9 Family of semicircles with close-up of region $\mathcal{D}$.

arcs; in the regions $\mathcal{B}$ contained in just one of the two highlighted circles, where BC would be the middle side, there is one arc through each point; in $\mathcal{C}$, most of the region outside the circles, where BC would be the smallest side, there are no arcs; and in the small region $\mathcal{D}$ just above the highlighted circles near the y-axis there are two arcs through each point. A larger version of region $\mathcal{D}$ is shown in Figure 9 (Right), and an even larger version of the right half of $\mathcal{D}$ (without the circular arcs) is shown in Figure 3.

Acknowledgments We thank the referee for his/her careful reading and suggestions.

REFERENCES

[1] Berele, A., Catoiu, S. (2017). The perimeter sixpartite center of a triangle. *J. Geom.* 108(3): 861–868. doi.org/10.1007/s00022-017-0380-4.

[2] Berele, A., Catoiu, S. (2018). The centroid as a nontrivial area bisecting center of a triangle. *Coll. Math. J.* 49(1): 27–34. doi.org/10.1080/07468342.2017.

[3] Berele, A., Catoiu, S. (2018). Nonuniqueness of sixpartite points. *Amer. Math. Monthly* 125(7): 638–642. http://dx.doi.org/10.1080/00029890.2018.1467191.

[4] Berele, A., Catoiu, S. (2020). The Fermat-Torricelli theorem in convex geometry. *J. Geom.* 111(2): Paper No. 22, 21pp. doi.org/10.1007/s00022-020-00535-6.

[5] Berele, A., Catoiu, S. (2022.) The pentagonal pizza conjecture. *Amer. Math. Monthly*. 129(5): 445–453. https://doi.org/10.1080/00029890.2022.2038005

[6] Buck, R. C., Buck, E. F. (1949). Equipartition of convex sets. *Math. Mag.* 22(4): 195–198. MR0029521

[7] Ceder, J. G. (1964). Generalized sixpartite problems. *Bol. Soc. Mat. Mex.* 9(2): 28–32.

[8] Courant, R., Robbins, H. (1941). *What Is Mathematics?* New York: Oxford Univ. Press.

[9] Jacobson, N. (1985). *Basic Algebra I*, 2nd. ed. San Francisco: Freeman.

[10] Zindler, K., Über konvexe Gebilde, I(1920), II(1921), III(1922). *Monatsh. Math.* (I) 30: 87–102; (II) 31: 25–56; (III) 32: 107–138. DOI: 10.1007/BF01699908

Summary. Zindler's theorem of 1920 says that each planar convex set admits two perpendicular lines that divide it into four parts of equal area. Call the intersection of the two lines a Zindler point. We show that each triangle admits either one, two or three Zindler points, and we classify all triangles according to these three numbers.

ALLAN BERELE (MR Author ID: 35020) received his Ph.D. from the University of Chicago and is a professor at DePaul University in Chicago. His main research interest is in algebras with polynomial identities, although he fell in love with Euclidean geometry many years before he ever heard of p. i. algebras.

STEFAN CATOIU (MR Author ID: 632038) received his Ph.D. from the University of Wisconsin-Madison and is an associate professor at DePaul University in Chicago. His research interest includes noncommutative algebra, real analysis, geometry, number theory and elementary mathematics.

The Difference of Two Polygonal Numbers

DAEYEOL JEON
Department of Mathematics Education
Kongju National University, 56
Gongjudaehak-ro, Gongju-si
Chungcheongnam-do 314-701, South Korea
dyjeon@kongju.ac.kr

HEONKYU LEE
Department of Mathematics Education
Kongju National University, 56
Gongjudaehak-ro, Gongju-si
Chungcheongnam-do 314-701, South Korea
poscokoala@naver.com

The numbers

$$1, 4, 9, 16, 25, 36, 49, \ldots$$

are called *square numbers*, for reasons made clear by Figure 1. We see that 15 is the difference of two square numbers: $15 = 16 - 1$. In how many ways can 15 be written as the difference of two square numbers?

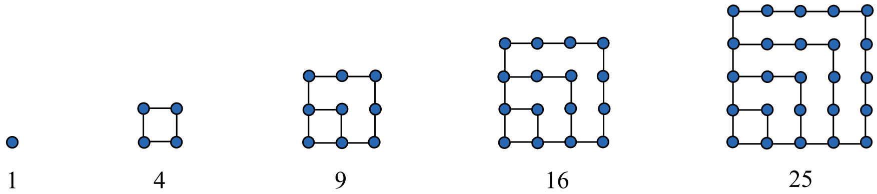


Figure 1 Square numbers.

We generalize this problem from square numbers to arbitrary polygonal numbers, where a *polygonal number* is a number that can be represented as dots arranged in the shape of a regular polygon. A polygonal number is also called an *m-gonal number* if it corresponds to a regular m-gon. In particular, we call polygonal numbers triangular, square, pentagonal, or hexagonal numbers when $m = 3, 4, 5,$ or 6, respectively. Our generalization is to ask: which natural numbers can be expressed as the difference of two m-gonal numbers? If they can be so expressed, then in how many different ways? For simplicity, we always assume that 0 is the 0th polygonal number.

Let $P(m, k)$ denote the kth m-gonal number. Then the difference of two consecutive m-gonal numbers $P(m, i + 1) - P(m, i)$ can be computed by counting the dots on the border of the larger polygon that do not overlap with the dots on the border of the preceding smaller polygon.

If there are i dots on one side of the smaller polygon, then there are $(m - 3)i + (i + 1)$ dots on the nonoverlapping borders of the larger polygon. Thus,

$$P(m, i + 1) - P(m, i) = (m - 2)i + 1, \tag{1}$$

and hence

Math. Mag. **95** (2022) 496–501. doi:10.1080/0025570X.2022.2126261 © Mathematical Association of America
MSC: Primary 11B83, Secondary 05A17

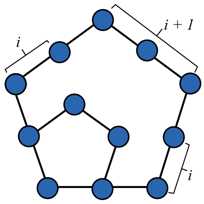


Figure 2 The difference of two pentagonal numbers.

$$\begin{aligned} P(m,k) &= \sum_{i=0}^{k-1} \{P(m,i+1) - P(m,i)\} \\ &= \sum_{i=0}^{k-1} \{(m-2)i+1\} = \frac{k\{(m-2)(k-1)+2\}}{2}. \end{aligned} \tag{2}$$

The difference of two polygonal numbers

Let us state our problem more formally:

(1) Which natural numbers can be expressed as the difference of two m-gonal numbers?

(2) How many different ways are there?

The first few numbers that can be expressed as the difference of two square numbers are as follows:

$$1, \quad 3, \quad 4, \quad 5, \quad 7, \quad 8, \quad 9, \quad 11, \quad 12, \quad 13, \quad \ldots$$

We can see that the numbers exhibit a pattern: they are not congruent to 2 modulo 4. It is an interesting problem to prove that this holds for every number which can be expressed as the difference of two square numbers. Kang et al. [2] proved that if n is not congruent to 2 modulo 4, then n can be expressed in

$$\left\lfloor \frac{\tau\left(\frac{n}{4}\right)+1}{2} \right\rfloor \qquad \text{or} \qquad \left\lfloor \frac{\tau(n)+1}{2} \right\rfloor$$

different ways as the difference of two square numbers depending on whether n is divisible by 4 or n is odd, respectively. Here, $\tau(n)$ is the number of positive divisors of n and $\lfloor\cdot\rfloor$ is the greatest integer function. For example, when $n = 15$ there are $\left\lfloor \frac{\tau(15)+1}{2} \right\rfloor = 2$ expressions of 15 as the difference of two square numbers. Indeed,

$$15 = P(4,4) - P(4,1) = P(4,8) - P(4,7).$$

The first question is not meaningful for triangular numbers because every natural number can be expressed as the difference of two triangular numbers: equation (1) implies that $n = P(3,n) - P(3,n-1)$.

However, the second question is very interesting for triangular numbers. The difference of two triangular numbers is equal to a sum of consecutive numbers. It is known that the number of different expressions of a natural number n as a sum of consecutive numbers is equal to the number of odd divisors of n [1]. Thus, the number of different expressions of n as the difference of two triangular numbers is equal to the number of odd divisors of n. For example 15 has 4 odd divisors, and

$$\begin{aligned} 15 &= P(3,5) - P(3,0) = P(3,6) - P(3,3) \\ &= P(3,8) - P(3,6) = P(3,15) - P(3,14). \end{aligned}$$

Now consider which numbers can be expressed as the difference of two pentagonal numbers. The first few such numbers are as follows:

$$1, \quad 4, \quad 5, \quad 7, \quad 10, \quad 11, \quad 12, \quad 13, \quad 16, \quad 17, \quad 19, \quad 20, \quad \ldots \tag{3}$$

In contrast to triangular and square numbers, it is difficult to find any modulus pattern among them. This can be attributed to the fact that such numbers depend on a varying modulus. Indeed, the numbers in equation (3) satisfy the following congruences:

$$\begin{aligned} 4 \equiv 1 &\equiv P(5,1) \pmod{3 \cdot 1}, \\ 5 \equiv 5 &\equiv P(5,2) \pmod{3 \cdot 2}, \\ 7 \equiv 1 &\equiv P(5,1) \pmod{3 \cdot 1}, \\ 10 \equiv 1 &\equiv P(5,1) \pmod{3 \cdot 1}, \\ 11 \equiv 5 &\equiv P(5,2) \pmod{3 \cdot 2}, \\ &\vdots \end{aligned}$$

Kim et al. [3] proved that a positive integer n can be expressed as the difference of two pentagonal numbers if and only if $n \equiv P(5,l) \pmod{3l}$, with $n \geq P(5,l)$ for some positive integer l. We extend their result to the cases of m-gonal numbers with $m \geq 5$. Figure 3 shows that

$$\begin{aligned} P(m,k+2) - P(m,k) &= (m-3)k + (m-2)(k+1) + (k+2) \\ &= (m-2)(2k) + m \\ &= (m-2)(2k) + P(m,2). \end{aligned}$$

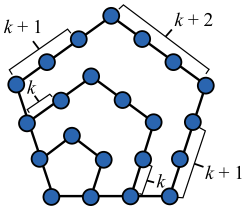


Figure 3 The difference of two pentagonal numbers.

Similarly, for a nonnegative integer k and a positive integer ℓ, we can use equation (1) and a telescoping series to compute the following:

$$\begin{aligned} P(m,k+\ell) - P(m,k) &= (m-3)k + \sum_{i=1}^{\ell-1}(m-2)(k+i) + (k+\ell) \\ &= (m-2)\ell k + P(m,\ell). \end{aligned} \tag{4}$$

This leads to the following result:

Theorem 1. *Suppose $m \geq 5$. A natural number n can be expressed as the difference of two m-gonal numbers if and only if*

$$n \equiv P(m, \ell) \pmod{(m-2)\ell} \tag{5}$$

with $n \geq P(m, \ell)$ for some positive integer ℓ.

Moreover, it is equivalent to

$$n \equiv \begin{cases} \ell & \pmod{(m-2)l}, \quad \ell \text{ is odd}, \\ -\frac{\ell}{2}(m-4) & \pmod{(m-2)\ell}, \quad \ell \text{ is even}. \end{cases} \tag{6}$$

with $n \geq P(m, \ell)$ for some positive integer ℓ.

Proof. Equation (5) follows from equation (4). By using equation (1), one can easily check that

$$P(m, \ell) \equiv \begin{cases} \ell & \pmod{(m-2)\ell}, \quad \ell \text{ is odd}, \\ -\frac{\ell}{2}(m-4) & \pmod{(m-2)\ell}, \quad \ell \text{ is even}, \end{cases} \tag{7}$$

and so equation (6) follows from equations (5) and (7). ■

For example, when $m = 5$ and $\ell = 1$, the first case of equation (6) turns out to be $n \equiv 1 \pmod 3$ with $n \geq 1$. Hence, $1, 4, 7, 11, \ldots$ can be expressed as the difference of two pentagonal numbers. Also, when $m = 6$ and $\ell = 2$, the second case of equation (6) turns out to be $n \equiv 6 \pmod 8$, with $n \geq 6$. Hence $6, 14, 22, 30, \ldots$ can be expressed as the difference of two hexagonal numbers.

The number of different expressions

Now consider the second problem. From now on, we always assume that n can be expressed as the difference of two m-gonal numbers. From equations (2) and (4), we have that

$$n = P(m, k + l) - P(m, k) = \begin{cases} \ell \left\{(m-2)\left(k + \frac{\ell-1}{2}\right) + 1\right\}, & \ell \text{ is odd}, \\ \frac{\ell}{2}\left\{(m-2)(2k + \ell - 1) + 2\right\}, & \ell \text{ is even}, \end{cases} \tag{8}$$

which shows that the number of expressions of n as the difference of two m-gonal numbers is determined by the divisors of n in the curly brackets in equation (8). The divisor of n in the curly bracket of the first case of equation (8) is congruent to 1 modulo $m - 2$, which implies $n \equiv \ell \pmod{(m-2)\ell}$. Hence, the first case of equation (6) follows from the first case of equation (8). The divisor of n in the curly bracket of the second case of equation (8) is congruent to $2(m-2)k - (m-4)$ modulo $(m-2)\ell$, which implies that $n \equiv -\frac{\ell}{2}(m-4) \pmod{(m-2)\ell}$. Hence, the second case of equation (6) follows from the second case of equation (8).

Now suppose that $n \equiv \ell \pmod{(m-2)\ell}$ for some odd integer ℓ. Then $n = \ell\{(m-2)t + 1\}$ for some integer t. If we let $d = (m-2)t + 1$, then d is a divisor of n which satisfies the following:

$$d \equiv 1 \pmod{m-2} \qquad \text{and} \qquad \frac{n}{d} = \ell \text{ is odd.}$$

Moreover, from equation (8),

$$n \geq P(m, \ell) = \frac{\ell\{(m-2)(\ell-1)+2\}}{2}. \tag{9}$$

By substituting $\ell = \frac{n}{d}$ in equation (9), we have the following inequality:

$$(m-2)n \leq 2d^2 + (m-4)d.$$

Next, suppose that $n \equiv -\frac{\ell}{2}(m-4) \pmod{(m-2)l}$ for some even integer ℓ. Then $n = \frac{\ell}{2}\{2(m-2)t - (m-4)\}$ for some integer t. If we let $d = 2(m-2)t - (m-4)$, then d is a divisor of n that satisfies

$$d \equiv m \pmod{2(m-2)}.$$

By substituting $\ell = \frac{2n}{d}$ in (9), we get

$$2(m-2)n \leq d^2 + (m-4)d.$$

Therefore, we have the following result:

Theorem 2. *Suppose $m \geq 5$. Then the number of different expressions of n as the difference of two m-gonal numbers is equal to the number of divisors d of n satisfying one of the following:*

(a) $d \equiv 1 \pmod{m-2}$, *$\frac{n}{d}$ is odd and $(m-2)n \leq 2d^2 + (m-4)d$.*

(b) $d \equiv m \pmod{2(m-2)}$ *and $2(m-2)n \leq d^2 + (m-4)d$.*

Proof. From the discussion prior to Theorem 2, we know that the first case of equation (6) gives a divisor d of n which satisfies (a), and the second case of equation (6) gives a divisor d of n which satisfies (b). Thus, it suffices to show that each divisor d of n satisfying either (a) or (b) gives an expression of n as the difference of two m-gonal numbers.

First, we suppose that a divisor d of n satisfies (a). Then $d = (m-2)a + 1$ for some integer a. If we let $\ell = \frac{n}{d}$ which is odd, then we have

$$n = \ell d = \ell\{(m-2)a+1\}. \tag{10}$$

On the other hand, by dividing the inequality of (a) by d, we have

$$(m-2)\ell \leq 2d + (m-4) = 2(m-2)a + (m-2),$$

which implies that $a \geq \frac{\ell-1}{2} \geq 0$. By comparing equation (10) and the first case of equation (8), we can take

$$k = a - \frac{\ell-1}{2} \geq 0. \tag{11}$$

Thus, we have an expression $n = P(m, k+\ell) - P(m, k)$.

Next, we suppose that a divisor d of n satisfies (b). Then $d = 2(m-2)b + m$ for some integer b. If we let $\ell = \frac{2n}{d}$, which is even, then we have

$$n = \frac{\ell}{2}d = \frac{\ell}{2}\{(m-2)(2b+1)+2\}. \tag{12}$$

On the other hand, by dividing the inequality of (b) by d, we have

$$(m-2)\ell \leq d + (m-4) = 2(m-2)b + 2(m-2),$$

which implies that $b \geq \frac{l}{2} - 1 \geq 0$. By comparing equation (12) and the second case of equation (8), we can take

$$k = b - \frac{\ell}{2} + 1 \geq 0, \tag{13}$$

then we have an expression $n = P(m, k + \ell) - P(m, k)$. ■

For example, consider $m = 5$ and $n = 46$. The divisors 46 and 23 satisfy (a) and (b) of Theorem 2, respectively. Thus, 46 has two different expressions as the difference of two pentagonal numbers by Theorem 2.

If we take $d = 46$ in (a) of Theorem 2, then $\ell = 1$ and $46 = 1 \cdot (3 \cdot 15 + 1)$. Hence, $a = 15$ in equation (10). By equation (11), $k = 15$, and we have

$$46 = P(5, 16) - P(5, 15) = 376 - 330.$$

If we take $d = 23$ in (b) of Theorem 2, then $l = 4$ and $46 = 2 \cdot (3 \cdot 7 + 2)$. Hence, $b = 3$ in equation (12). By equation (13), $k = 2$. Hence, we have

$$46 = P(5, 6) - P(5, 2) = 51 - 5.$$

For example, suppose $m = 6$ and $n = 85$. The divisors of 85 are 1, 5, 17, and 85, which satisfy the first two conditions of (b) of Theorem 2. However, only 17 and 85 satisfy the last condition. Thus, 85 has two different expressions as the difference of two hexagonal numbers, by Theorem 2.

If we take $d = 17$ in (a) of Theorem 2, then $\ell = 5$ and $85 = 5 \cdot (4 \cdot 4 + 1)$. Hence, $a = 4$ in equation (12). By equation (11), $k = 2$, and we have

$$85 = P(6, 7) - P(6, 2) = 91 - 6.$$

If we take $d = 85$ in (a) of Theorem 2, then $\ell = 1$ and $85 = 1 \cdot (4 \cdot 21 + 1)$. Hence, $a = 21$ in equation (12). By equation (11), $k = 21$, and we have

$$85 = P(6, 22) - P(6, 21) = 946 - 861.$$

REFERENCES

[1] Hirschhorn, M. D., Hirshchorn, P. M. (2005). Partitions into consecutive numbers. *Math. Mag.* 78(5): 396–397. DOI: 10.2307/30044200

[2] Kang, H., Kim, D., Shin, J., Yoon, J., Lee, S., Cho, S., Hwang, S. (2015). A research on the difference of two figurate numbers, (in Korean) *J. Sci. Edu. Gifted*. 7: 150–157.

[3] Kim, Y., Choi, D., Choi, J. (2019). A research on the difference of two pentagonal numbers, (in Korean) *J. Sci. Edu. Gifted*. 11: 289–295.

Summary. We determine all natural numbers that can be expressed as the difference of two m-gonal numbers. For each such number, we determine the number of possible expression as the difference of two m-gonal numbers.

DAEYEOL JEON received his PhD from the Department of Mathematical Science at KAIST (Korea Advanced Institute of Science and Technology). He worked as a researcher at KIAS (Korea Institute for Advanced Study) and studied Diophantine geometry. After leaving KIAS, he joined the faculty at Kongju National University to teach mathematics and continue his research in Number Theory.

HEONKYU LEE graduated from the Department of Biology Education at Kongju National University and obtained a master's degree at the Department of Mathematics Education at the University under the supervision of professor Daeyeol Jeon.

Unified Proofs of Three Fundamental Properties of Continuous Functions

DANIEL DANERS
The University of Sydney
New South Wales 2006, Australia
daniel.daners@sydney.edu.au

Every rigorous course on calculus or introductory analysis will establish a number of fundamental properties of continuous functions on closed and bounded intervals in $\mathbb{R}$. They are the *intermediate value theorem*, the *extreme value theorem* and for the very ambitious the *uniform continuity*. Browsing through popular textbooks on advanced calculus, they are often proved by rather different methods; see for instance Spivak [**6**]. In books on analysis, they are often proved using concepts from topology not available at an introductory level.

Our aim is to establish a building block that is consistently used to prove all three facts by only building on a few core concepts. The concepts we use are the completeness of the real numbers as manifested by the *least upper bound axiom* and the ε-δ definition of continuity. All proofs are accompanied by diagrams that convincingly illustrate the main idea and provide a guide on how to construct a proof. All proofs are direct proofs rather than the quite common contradiction proofs.

The main building block

Our basic building block is the core argument often used to establish the intermediate value theorem; see for instance Spivak [**6**, Theorem 7-1, p. 133], Adams & Essex [**1**, Appendix 3] or Ellis & Gulick [**3**, A9-11]. We make that building block explicit.

Building Block. *Suppose that $[a, b] \subseteq \mathbb{R}$ is a closed and bounded interval and that $f : [a, b] \to \mathbb{R}$ is continuous. Fix $t_0 \in [a, b)$ and let $c > f(t_0)$. Define*

$$t_1 := \sup\bigl\{t \in (t_0, b) \colon f(s) < c \textit{ for all } s \in [t_0, t)\bigr\}. \tag{1}$$

Then one of the following two alternatives must occur:

(i) *$t_1 = b$ and $f(b) \le c$;*
(ii) *$t_0 < t_1 < b$ and $f(t_1) = c$.*

The number t_1 is the largest number so that the graph $y = f(t)$ stays strictly below the level $y = c$ for all $t \in [t_0, t_1)$. It may happen that this is the case for all $t \in [t_0, b)$, in which case alternative (i) holds. If not, then by continuity we expect the graph $y = f(t)$ to cross or touch the line $y = c$, as illustrated in Figure 1.

The above statement, together with Figure 1, already suggest how to construct a proof. We define the set

$$A := \{t \in (t_0, b) \colon f(s) < c \text{ for all } s \in [t_0, t)\}$$

representing the largest interval with left endpoint t_0 where the graph $y = f(t)$ stays strictly below the line $y = c$. Then $t_1 := \sup(A)$ is the right endpoint of that interval.

Math. Mag. **95** (2022) 502–508. doi:10.1080/0025570X.2022.2125725 © Mathematical Association of America
MSC: Primary 26A15, Secondary 26A06

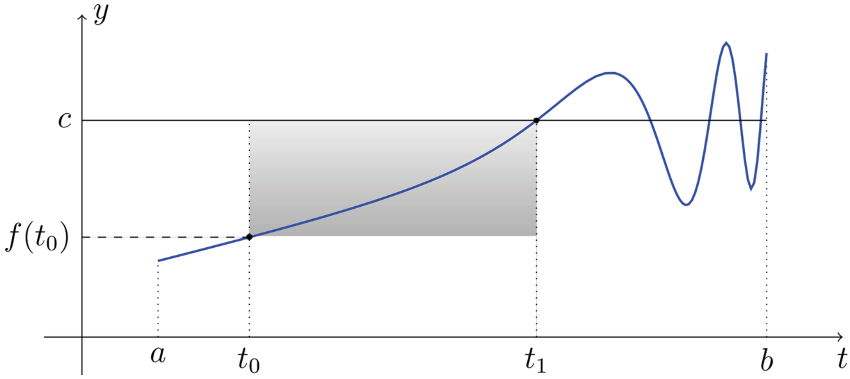


Figure 1 First point of intersection giving alternative (ii).

We need to show that $t_1 \in (t_0, b]$. As a first step, we argue that A is non-empty and bounded from above. The boundedness is immediate because $A \subseteq [a, b]$. To show that A is non-empty, we use the continuity of f at t_0 to show that $f(t) < c$ for t close to b. By assumption,

$$\varepsilon := c - f(t_0) > 0,$$

so by continuity there exists $\delta > 0$ so that $0 \leq t - t_0 < \delta$ implies $|f(t) - f(t_0)| < \varepsilon$. In particular

$$f(t) - f(t_0) < \varepsilon = c - f(t_0),$$

and thus, $f(t) < c$ whenever $0 < t - t_0 < \delta$. Hence, $(t_0, t_0 + \delta) \subseteq A$, which is to say that A is non-empty. By the least upper bound axiom, $t_1 = \sup A$ exists and $t_0 < t_1 \leq b$.

We next use the continuity of f to show that $f(t_1) \leq c$ if $t_1 = b$, and $f(t_1) = c$ if $t_1 < b$. To do so, first note that by the definition of t_1 as a supremum over the set A, we have

$$f(t) < c, \quad \text{for all } t \in [t_0, t_1), \tag{2}$$

and if $t_1 < b$ we also have:

$$\text{For every } \delta > 0, \text{ there exists } s \in [t_1, t_1 + \delta) \text{ so that } f(s) \geq c. \tag{3}$$

Now fix $\varepsilon > 0$. Then there exists $\delta > 0$ such that $|f(t) - f(t_1)| < \varepsilon$ whenever $t \in [a, b]$ and $|t - t_1| < \delta$. In particular, using equation (2), we have

$$t \in (t_1 - \delta, t_1] \implies c - f(t_1) > f(t) - f(t_1) > -\varepsilon.$$

Hence, $f(t_1) < c + \varepsilon$. Since the above argument works for every $\varepsilon > 0$, we conclude that $f(t_1) \leq c$, proving (i). If $t_1 < b$, then by equation (3) there exists $s \in [t_1, t_1 + \delta)$ such that

$$c - f(t_1) \leq f(s) - f(t_1) < \varepsilon.$$

Hence, $f(t_1) > c - \varepsilon$. Since the above arguments work for every choice of $\varepsilon > 0$, we deduce that $f(t_1) \geq c$. Together with the fact that $f(t_1) \leq c$ we conclude that $f(t_1) = c$.

The intermediate value theorem

By making suitable assumptions, the building block previously established almost directly gives the intermediate value theorem. The theorem states that the graph of f intersects the line $y = c$ at least once whenever c lies strictly between $f(a)$ and $f(b)$.

Intermediate Value Theorem. *Suppose that $f\colon [a, b] \to \mathbb{R}$ is continuous, and that either $f(a) < c < f(b)$ or $f(b) < c < f(a)$. Then there exists $t \in (a, b)$ such that $f(t) = c$.*

We only need to look at the first case since the second follows by considering $-f$. By assumption, $f(a) < c$. Hence, we can apply the building block with $t_0 = a$ and find $t_1 \in (a, b]$ satisfying one of the two alternatives. The alternative $t_1 = b$ can be excluded since by assumption $c < f(b)$ and hence, by continuity, there exists $\delta > 0$ such that

$$|f(t) - f(b)| < \varepsilon := f(b) - c.$$

In particular $t_1 \in (a, b)$ and $f(t_1) = c$, as required.

The extreme value theorem

The extreme value theorem states that a continuous function on a closed and bounded interval has a maximum and a minimum. Traditionally, the theorem is proved in two parts: first show the boundedness, then the existence of a maximum and minimum. Our approach is based on iteratively applying the building block and combines the two parts into one.

Extreme Value Theorem. *Suppose that $f\colon [a, b] \to \mathbb{R}$ is continuous. Then there exist $t_m, t_M \in [a, b]$ such that*

$$f(t_m) \leq f(t) \leq f(t_M)$$

for all $t \in [a, b]$.

We call $f(t_M)$ the maximum and $f(t_m)$ the minimum of f on $[a, b]$, or more generally the extreme values of f.

We only prove the existence of a maximum. The existence of a minimum follows by looking at the maximum of $-f$. We set

$$M := \sup\{f(t)\colon t \in [a, b]\}.$$

Then either $M < \infty$ or $M = \infty$, we do not know which at this stage. If $M = f(a)$, then we are done, and $t_M = a$ is as required. If $M > f(a)$ we choose a strictly increasing sequence

$$f(a) := c_0 < c_1 < c_2 < \cdots$$

so that

$$M = \sup\{c_k\colon k \geq 0\}.$$

If $f(a) < M < \infty$, we can, for instance, choose $c_0 = f(a)$ and inductively define

$$c_{k+1} = \frac{c_k + M}{2},$$

for $k \geq 0$. If $M = \infty$, then we can, for instance, choose $c_k := f(a) + k$ for all $k \in \mathbb{N}$. The idea of the proof is to use the building block to get a sequence (t_k) in (a, b), where the graph of f for the first time crosses the line $y = c_k$. Then we show that the maximum of f is attained at

$$t_M := \sup\{t_k : k \in \mathbb{N}\},$$

as shown in Figure 2.

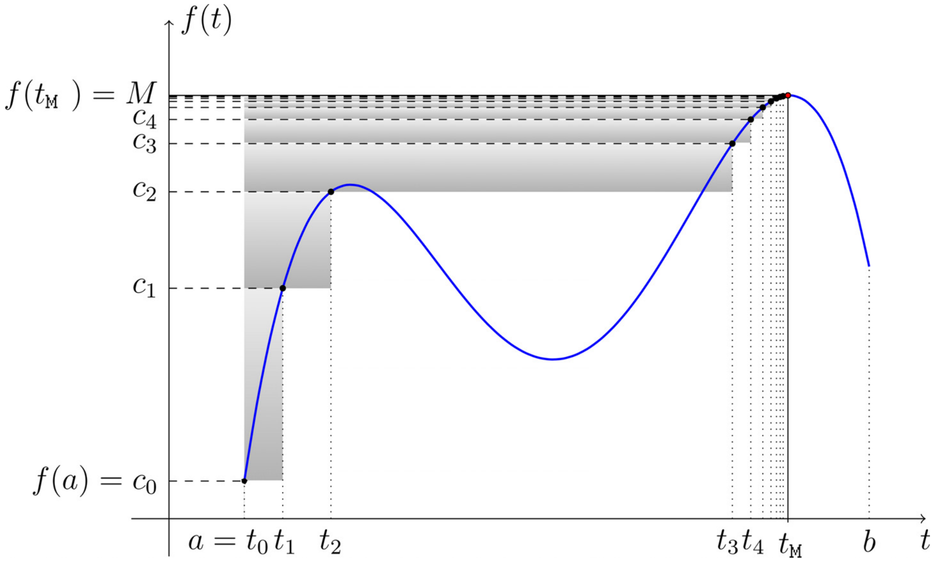


Figure 2 Illustration of the proof of the extreme value theorem.

To carry out this plan, for $k \geq 1$, use the building block with $t_0 = a$ and $c = c_k$. Since $c_k < M$, the definition of a supremum implies the existence of $t \in [a, b]$ such that $c_k < f(t) < M$, which means that $t_k < b$ for all $k \in \mathbb{N}$. By the building block, we have $f(t_k) = c_k$. Since $c_{k+1} > c_k$, we have $t_k < t_{k+1} < b$ for all $k \in \mathbb{N}$. By the least upper bound axiom,

$$t_M := \sup\{t_k : k \in \mathbb{N}\} \leq b$$

exists. We need to show that $M < \infty$ and that $f(t_M) = M$. For that we make use of the continuity of f.

By the definition of M, we know that $f(t_M) \leq M$. To show the opposite inequality, we use the continuity of f at t_M. Fix $\varepsilon > 0$, and let $\delta > 0$ be such that

$$|f(t) - f(t_M)| < \varepsilon$$

whenever $t \in [a, b]$ satisfies $|t - t_M| < \delta$. By the definition of t_M as a supremum over the increasing sequence $(t_k)_{k \geq 0}$, there exists k_0 such that $0 < t_M - t_k < \delta$ for all $k \geq k_0$. Hence,

$$c_k - f(t_M) = f(t_k) - f(t_M) < \varepsilon,$$

for all $k \geq k_0$. In particular, $c_k < f(t_M) + \varepsilon$ for all $k \geq k_0$. As a consequence, the increasing sequence (c_k) is bounded, and by the definition of a supremum, we have

$$M \leq f(t_M) + \varepsilon.$$

In particular, M is finite. Since the above argument works for all $\varepsilon > 0$, it follows that $M \leq f(t_M) < \infty$. Together with the inequality we proved already, we have $M = f(t_M)$, so M is a maximum.

Uniform continuity

The uniform continuity is not usually accessible with the same tools as the intermediate value theorem since the most common proofs rely on notions of compactness. There is one that relies on the extreme value theorem for functions of several variables and hence is not suitable in the current context; see Daners [2]. As it turns out, there is a proof that fits very well with the theme of this exposition. By definition, $f\colon [a, b] \to \mathbb{R}$ is uniformly continuous if for every $\varepsilon > 0$, a number $\delta > 0$ can be chosen independently of $x \in [a, b]$ so that

$$0 \leq |t - s| < \delta \implies |f(t) - f(s)| < \varepsilon.$$

In contrast to the definition of continuity, the δ is not allowed to depend on t. We prove a statement that is equivalent: given $\varepsilon > 0$ consider the horizontal lines $y = f(a) + k\varepsilon$, $k \in \mathbb{N}$, and show that the graph of f can move from a line to a neighboring line at most finitely many times, as shown in Figure 3. The formulation of uniform continuity considered here is exactly the one required to prove the Riemann integrability of continuous functions on $[a, b]$. Our argument is essentially the one published by Heine [4, p. 188] in 1872. However, the argument goes back to Dirichlet's lectures from 1854, which were only published in 1904 in Lejeune-Dirichlet [5, Section 2].

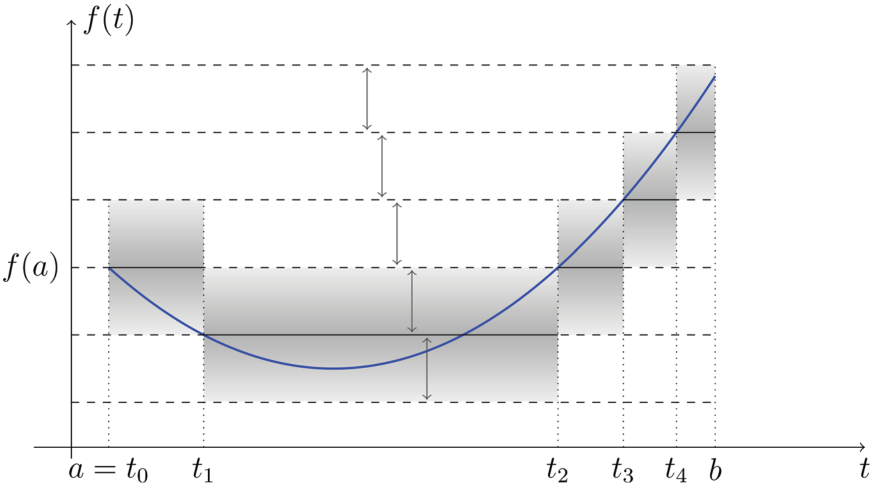


Figure 3 Construction of a partition with maximum oscillation 2ε.

Uniform Continuity. *Suppose that $f\colon [a, b] \to \mathbb{R}$ is continuous. Then for every $\varepsilon > 0$, there exists a partition $a = t_0 < t_1 < \cdots < t_n = b$ of the interval $[a, b]$ with the following properties:*

(i) *For every $k = 1, \ldots, n$ we have $|f(t) - f(t_{k-1})| < \varepsilon$ for all $t \in (t_{k-1}, t_k)$;*

(ii) *For every $k = 1, \ldots, n - 1$ we have $|f(t_k) - f(t_{k-1})| = \varepsilon$.*

Figure 3 tells us how to proceed. We need to find the first instance where the graph $y = f(t)$ crosses or touches either $y = f(a) + \varepsilon$ or $y = f(a) - \varepsilon$. Hence, we look at the function $g_1 \colon [a, b] \to \mathbb{R}$ given by

$$g_1(t) := |f(t) - f(t_0)|.$$

Since g_1 is continuous, we can apply our building block with $t_0 = a$ and set

$$t_1 := \sup\bigl\{t \in [a, b] \colon |f(s) - f(t_0)| < \varepsilon \text{ for all } s \in (a, t)\bigr\}.$$

There are two possibilities: either $t_1 = b$ and $g_1(s) < \varepsilon$ for all $s \in [t_0, b)$, or $t_1 < b$ and

$$g_1(t_1) = |f(t_1) - f(t_0)| = \varepsilon.$$

In the first case, we are done. In the second case, we repeat the argument and set

$$t_2 := \sup\left\{t \in [t_1, b] \colon |f(s) - f(t_1)| < \varepsilon \text{ for all } s \in (t_1, t)\right\}.$$

Applying the building block to $g_2(t) := |f(t) - f(t_1)|$, we conclude that either $t_2 = b$ or $t_1 < t_2 < b$. In the first case, we are done, and in the second case, we have

$$|f(t_2) - f(t_1)| = \varepsilon.$$

We repeat the previous step again, replacing t_1 by t_2; see Figure 3. In fact, proceeding more formally, we can set $t_0 := a$ and define inductively

$$t_k := \sup\left\{t \in [t_{k-1}, b] \colon |f(s) - f(t_{k-1})| < \varepsilon \text{ for all } s \in (t_{k-1}, t)\right\}.$$

for $k \geq 1$ as long as $t_{k-1} < b$, making use of the building block at every step. If $t_k = b$ we are done, otherwise $t_k < b$ and

$$|f(t_{k+1}) - f(t_k)| = \varepsilon. \tag{4}$$

There are two possibilities: either there exists $k \in \mathbb{N}$ so that $t_k = b$, or $t_k < b$ for all $k \in \mathbb{N}$. In the first case we are done. To exclude the second case we set

$$\tilde{t} := \sup_{k \in \mathbb{N}} t_k \leq b$$

and show that f cannot be continuous at $\tilde{t}$. Fix any $\varepsilon > 0$, and let $\delta > 0$. By the definition of $\tilde{t}$, there exists $k \in \mathbb{N}$ so that $0 < \tilde{t} - t_k < \delta$. Then also $0 < \tilde{t} - t_{k+1} < \delta$ and hence

$$\varepsilon = |f(t_{k+1}) - f(t_k)| \leq |f(t_{k+1}) - f(\tilde{t})| + |f(\tilde{t}) - f(t_k)|.$$

Therefore,

$$|f(t_{k+1}) - f(\tilde{t})| \geq \frac{\varepsilon}{2} \quad \text{or} \quad |f(\tilde{t}) - f(t_k)| \geq \frac{\varepsilon}{2}.$$

Since this argument is valid for any $\varepsilon > 0$ and $\delta > 0$, the function f cannot be continuous at $\tilde{t}$. This concludes the proof of the uniform continuity.

We only wanted to work with the definition of a supremum and the continuity of f. However, if we decide to work with sequential continuity, then there is an alternative to the last argument. We note that $t_k \to \tilde{t}$ and $t_{k+1} \to \tilde{t}$ as $k \to \infty$, and hence by equation (4) we have

$$0 < \varepsilon = \lim_{k \to \infty} |f(t_{k+1}) - f(t_k)| = |f(\tilde{t}) - f(\tilde{t})| = 0,$$

which is impossible.

We conclude by demonstrating that the statement proved above really implies the uniform continuity. Choose $\varepsilon > 0$. By what we proved, there is a partition $a = t_0 < t_1 < \cdots < t_n = b$ having properties (i) and (ii), but with ε replaced by $\varepsilon/3$. We set

$$\delta := \min_{k=1,\dots,n} (t_k - t_{k-1})$$

and let $t, s \in [a, b]$ so that $0 < t - s < \delta$. By the definition of δ, there exists $k \in \{1, \dots, n\}$ so that either

$$t_{k-1} \leq s < t \leq t_k$$

or

$$t_{k-1} < s < t_k < t < t_{k+1}.$$

In the first case

$$|f(t) - f(s)| \leq |f(t) - f(t_{k-1})| + |f(s) - f(t_{k-1})| \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} < \varepsilon.$$

In the second case

$$\begin{aligned} |f(t) - f(s)| &\leq |f(t) - f(t_k)| + |f(t_k) - f(t_{k-1})| + |f(s) - f(t_{k-1})| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Hence f is uniformly continuous.

REFERENCES

[1] Adams, R. A., Essex, C. (2017). *Calculus: A Complete Course*, 9th ed. Toronto: Pearson Canada.

[2] Daners, D. (2015). Uniform continuity of continuous functions on compact metric spaces. *Amer. Math. Monthly* 122(6): 592. doi.org/10.4169/amer.math.monthly.122.6.592

[3] Ellis, R., Gulick, D. (1994). *Calculus with Analytic Geometry*, 5th ed. Fort Worth: Saunders College Publishing.

[4] Heine, E. (1872). Die Elemente der Functionenlehre. *J. Reine Angew.Math.* 74: 172–188 (German). doi.org/10.1515/crll.1872.74.172

[5] Lejeune-Dirichlet, P. G. (1904). *Vorlesungen über die Lehre von den einfachen und mehrfachen bestimmten Integralen*. Braunschweig:F. Vieweg & Sohn (German). Edited by Gustav Arendt. https://openlibrary.org/books/OL22881156M

[6] Spivak, M. (2006). *Calculus*, 3rd ed. Cambridge: Cambridge University Press.

Summary. We provide a unified approach to three fundamental properties of continuous functions on closed and bounded intervals: the intermediate value theorem, and the uniform continuity theorem. We prove all three using the same building block, only making use of the least upper bound axiom and the $\epsilon - \delta$ definition of continuity.

DANIEL DANERS (MR Author ID: 325132, ORCID 0000-0002-0122-3789) grew up in Switzerland. He studied and obtained a Ph.D. at the University of Zürich in 1992. After a number of postdoctoral years in various places he finally settled with his family in Australia, teaching at the University of Sydney.

Summation Formulas, Generating Functions, and Polynomial Division

ETHAN BERKOVE
Lafayette College
Easton, PA 18042
berkovee@lafayette.edu

MICHAEL A. BRILLESLYPER
Florida Polytechnic University
Lakeland, FL 33805
mbrilleslyper@floridapoly.edu

One of the exciting aspects of mathematics is when familiar objects appear in unexpected places. For us, this happened while considering a polynomial division problem, when we noticed Fibonacci numbers popping up as coefficients. Digging deeper, we found a connection between our problem and generating functions, which led to a beautiful way of calculating partial sums, and hence to a general construction for Fibonacci identities and more.

Fibonacci identities are a common staple when learning induction. Both authors of this article remember proving that

$$\sum_{j=0}^{n} f_j = f_{n+2} - 1.$$

What is particularly nice about this example is that it is easy to guess a formula for the sum by working through some examples. However, what do you do when the series being summed has a more complicated form? In this article, we hope to convince the reader that with a little time and patience, one can find an explicit formula for many sums of this type, not just for Fibonacci numbers, but for all sorts of interesting number sequences.

Fibonacci numbers and long division

The Fibonacci numbers $\{f_n\}$ are among the most studied of all integer sequences, arising with surprising frequency in mathematics as well as in other fields. This deceptively simple sequence begins:

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots,$$

where each subsequent term is the sum of the two previous terms. One can define the Fibonacci numbers by setting $f_0 = 0$, $f_1 = 1$, and then applying the recurrence relation

$$f_n = f_{n-1} + f_{n-2}, \quad \text{or equivalently,} \quad f_n - f_{n-1} - f_{n-2} = 0, \tag{1}$$

for $n \geq 2$. As an integer sequence, the Fibonacci numbers have such a rich structure that they have their own journal, the *Fibonacci Quarterly*. They satisfy hundreds of identities, many of which can be found in Koshy's comprehensive text [10]. Here is an example [10, Formula 28.1]. Say you are interested in finding the sum

$$S_n = \sum_{j=1}^{n} j f_j = f_1 + 2f_2 + \cdots + nf_n. \tag{2}$$

Math. Mag. **95** (2022) 509–519. doi:10.1080/0025570X.2022.2127302 © Mathematical Association of America
MSC: Primary 05A15, Secondary 11B39

It is easy to calculate that $S_1 = 1$, $S_2 = 3$, $S_3 = 9$, and $S_4 = 21$. However, it is much harder to determine, even with more information, that

$$S_n = nf_{n+2} - f_{n+3} + 2.$$

Of course, once you have a formula for S_n it is easy to confirm it using induction. But how can one find these formulas in the first place?

Our story begins with a chance observation by one of the authors, who noticed something interesting while doing polynomial long division as part of a problem in complex analysis:

$$\begin{array}{rrrrrrr}
 & & & x^3 & +x^2 & +2x & +3 \\ \cline{2-7}
x^2-x-1 \big) & x^5 & & & & & \\
 & -x^5 & +x^4 & +x^3 & & & \\ \cline{2-4}
 & & x^4 & +x^3 & & & \\
 & & -x^4 & +x^3 & +x^2 & & \\ \cline{3-5}
 & & & 2x^3 & +x^2 & & \\
 & & & -2x^3 & +2x^2 & +2x & \\ \cline{4-6}
 & & & & 3x^2 & +2x & \\
 & & & & -3x^2 & +3x & +3 \\ \cline{5-7}
 & & & & & 5x & +3
\end{array}$$

Observe that the polynomial coefficients in the quotient are Fibonacci numbers, as are the coefficients in the remainder. A little exploration should convince you that this does not happen by chance. Recall that $f_0 = 0$. Then a proof by induction, for example, shows that dividing x^{n+1} by $x^2 - x - 1$ yields a polynomial quotient

$$Q(x) = f_0 x^n + f_1 x^{n-1} + f_2 x^{n-2} + \cdots + f_{n-1}x + f_n$$

with remainder

$$R(x) = f_{n+1}x + f_n,$$

so that

$$x^{n+1} = (x^2 - x - 1)Q(x) + R(x). \tag{3}$$

We can solve for $Q(x)$ in equation 3:

$$\begin{aligned} Q(x) &= f_0 x^n + f_1 x^{n-1} + f_2 x^{n-2} + \cdots + f_{n-1}x + f_n \\ &= \frac{x^{n+1} - R(x)}{x^2 - x - 1} \\ &= \frac{x^{n+1} - f_{n+1}x - f_n}{x^2 - x - 1}. \end{aligned} \tag{4}$$

Before we examine why the Fibonacci numbers occur in this problem, let us see how we can use the two descriptions of $Q(x)$ in equation 4 to build a variety of summation formulas. One thing we can do is evaluate equation 4 at various values of x. For example, when $x = 1$ we obtain one of the most well-known Fibonacci identities, the one at the start of this article:

$$\sum_{j=0}^{n} f_j = f_{n+2} - 1. \tag{5}$$

Substituting $x = 2$ and fiddling a little yields a less familiar identity, one with weighted coefficients which surprisingly provides another way to count binary strings of length $n + 1$:

$$f_{n+1} + f_{n+2} + \sum_{j=0}^{n} 2^{n-j} f_j = 2^{n+1}.$$

(The reader is encouraged to check out Identity 10 in Benjamin and Quinn [1] for an explanation of this interpretation.) We can build an identity with alternating terms by substituting negative values for x, like $x = -1$:

$$\sum_{j=0}^{n} (-1)^{n-j} f_j = (-1)^{n+1} + f_{n+1} - f_n.$$

Noting that $f_{n+1} - f_n = f_{n-1}$, we multiply by $(-1)^n$ and simplify. In this way, we can rewrite the identity in a nicer form which is equivalent to an identity in Koshy's book [10, Chapter 5, Identity 20]:

$$\sum_{j=0}^{n} (-1)^j f_j = (-1)^n f_{n-1} - 1.$$

We also get interesting formulas when we evaluate equation 4 at complex numbers. Setting $x = i$, we derive two identities simultaneously by equating real and complex parts. We assume that n is even, so that $(i)^{2n-j} = (-i)^j$:

$$\begin{aligned}\sum_{j=0}^{2n} (i)^{2n-j} f_j &= \sum_{j=0}^{n} (i)^{2j} f_{2j} - \sum_{j=1}^{n} (i)^{2j-1} f_{2j-1}\\ &= \frac{i - f_{2n} - f_{2n+1} i}{-2 - i}\\ &= \frac{1}{5}(f_{2n+2} + f_{2n} - 1) + \frac{i}{5}(f_{2n+1} + f_{2n-1} - 2).\end{aligned}$$

We used the Fibonacci recurrence relation from equation 1 to simplify the last line. The result is two alternating sums of even- and odd-numbered Fibonacci numbers when n is even:

$$\sum_{j=0}^{n} (-1)^j f_{2j} = \frac{1}{5}(f_{2n+2} + f_{2n} - 1),$$

and

$$\sum_{j=1}^{n} (-1)^j f_{2j-1} = \frac{1}{5}(f_{2n+1} + f_{2n-1} - 2).$$

The sequences that arise from the closed forms, $\{0, 2, 15, 104, 714, \dots\}$ and $\{1, 9, 64, 441, \dots\}$, appear in the OEIS as sequences A081018 and A049684. However, at the time of the writing of this article, neither alternating series was mentioned in its entry, and the listed closed forms were different from ours. They yield the Fibonacci identities

$$\frac{1}{5}(f_{2n+2} + f_{2n} - 1) = f_n f_{n+1},$$

and

$$\frac{1}{5}(f_{2n+1}+f_{2n-1}-2)=(f_n)^2.$$

Proving these directly is an interesting, nontrivial exercise.

We can get another weighted Fibonacci identity by taking the derivative of equation 4 and setting $x=1$. This can be done either by hand or by using a program like *Mathematica*. The result is

$$\begin{aligned}
&\sum_{j=1}^{n-1}(n-j)f_j\\
&\quad=\left.\frac{(x^2-x-1)\big((n+1)x^n-f_{n+1}\big)-(2x-1)(x^{n+1}-f_{n+1}x-f_n\big)}{(x^2-x-1)^2}\right|_{x=1}\\
&\quad=-\left((n+1)-f_{n+1}\right)-(1-f_{n+1}-f_n)\\
&\quad=-(n+1)+f_{n+1}-1+f_{n+2}\\
&\quad=f_{n+3}-n-2 \quad \text{(see Koshy [10, Chapter 28]).} \qquad (6)
\end{aligned}$$

In a similar way, we can derive a sum formula with $(n-j)^2$ terms as the weightings by taking the derivative of equation 4, multiplying by x, and taking the derivative a second time.

We also note that equation 6 resembles equation 2, with j replaced by $n-j$. Although the sum in equation 2 may feel more natural, it is closely related to this one through a clever trick mentioned by Koshy [9]: as $(n-j)f_j+jf_j=nf_j$,

$$\begin{aligned}
\sum_{j=0}^{n}jf_j&=n\sum_{j=0}^{n}f_j-\sum_{j=0}^{n}(n-j)f_j\\
&=n(f_{n+2}-1)-(f_{n+3}-n-2)\quad \text{using equations 5 and 6}\\
&=nf_{n+2}-f_{n+3}+2.
\end{aligned}$$

This is the sum of the series S_n in equation 2.

Some of the series manipulations we have just used may feel familiar. They show up, for example, in many calculus texts in series problems associated to the geometric sum formula

$$1+x+x^2\cdots+x^n=\frac{x^{n+1}-1}{x-1}. \qquad (7)$$

The hard part about doing similar analyses for generic sums is that we need an expression like equation 4 to work with. Clearly, such an expression is a handy tool for finding summation formulas. But how do we do this in general?

Generating functions

A generating function for a sequence is a power series whose coefficients are the sequence of interest. Using a standard technique (see Brualdi [2, Chapter 7] or Wilf [13, Chapter 1]), let us build the generating function,

$$G_f(x)=\sum_{j=0}^{\infty}f_jx^j,$$

for the Fibonacci numbers. We will also use $f_n - f_{n-1} - f_{n-2} = 0$, the Fibonacci recurrence relation in equation 1. The form of equation 1 suggests that we write out the series $G_f(x)$, multiply it by $-x$ and $-x^2$, and add the three series together:

$$\begin{aligned} G_f(x) &= f_0 + f_1x + f_2x^2 + f_3x^3 + f_4x^4 + \cdots \\ -xG_f(x) &= - f_0x - f_1x^2 - f_2x^3 - f_3x^4 - \cdots \\ -x^2G_f(x) &= - f_0x^2 - f_1x^3 - f_2x^4 - \cdots \end{aligned} \tag{8}$$

Equation 1 implies that on the right hand side only the two leftmost columns have nonzero sums, so the sum of equations 8 simplifies to

$$(1 - x - x^2)G_f(x) = f_0 + (f_1 - f_0)x = x.$$

Now we can easily solve for the generating function,

$$G_f(x) = \frac{x}{1 - x - x^2}. \tag{9}$$

We call equation 9 the *concise form* of $G_f(x)$. We make a few observations that apply in general to generating functions $G(x)$ that arise from linear, homogeneous recurrences with constant coefficients, like equation 1.

1. $G(x)$ will always be a rational function. If we wish to evaluate $G(x)$ at some value of x, equality only holds when x is inside the radius of convergence.
2. The numerator of $G(x)$ will always have strictly smaller degree than the denominator.
3. The denominator of $G(x)$ will always have 1 as the constant term.

The long division problem we saw earlier looks a lot like equation 9, the generating function for the Fibonacci numbers, and indeed there is a connection. However, the generating function in equation 9 is an infinite sum, and we only need finitely many terms for our summation formulas. We write

$$G_f(x) = f_0 + f_1x + \cdots + f_nx^n + R_n(x), \tag{10}$$

and use generating function techniques a second time to find a concise form for the remainder terms,

$$R_n(x) = \sum_{j=n+1}^{\infty} f_jx^j.$$

We remove a factor of x^{n+1} from all the terms in $R_n(x)$. Then

$$R_n(x) = x^{n+1}\sum_{j=0}^{\infty} f_{j+n+1}x^j,$$

which shows that $R_n(x)$ is just a shifted version (by x^{n+1}) of the generating function for the sequence $f_{n+1}, f_{n+2}, \ldots$ This shifted sequence still satisfies the recurrence in equation 1, so we can use equations 8 again to find a concise form for its generating function.

$$R_n(x) = x^{n+1}\left(\frac{f_{n+1} + (f_{n+2} - f_{n+1})x}{1 - x - x^2}\right) = \frac{f_{n+1}x^{n+1} + f_nx^{n+2}}{1 - x - x^2}.$$

(See Hansen [5] for a different derivation for this expression.) We then substitute this expression for $R_n(x)$ into equation 10 and, incorporating equation 9, solve for the polynomial part $G_f(x) - R_n(x)$, which we call $Q_1(x)$:

$$Q_1(x) = f_0 + f_1 x + \cdots + f_n x^n = \frac{x - f_{n+1}x^{n+1} - f_n x^{n+2}}{1 - x - x^2}. \tag{11}$$

Equation 11 for $Q_1(x)$ is analogous to equation 7 and closely resembles equation 4 for $Q(x)$, although the terms of $Q_1(x)$ have the form $f_j x^j$ rather than $f_j x^{n-j}$. The polynomials $Q(x)$ and $Q_1(x)$ are *reciprocal polynomials* of each other, and satisfy

$$Q_1(x) = x^{\deg Q(x)} Q\left(\frac{1}{x}\right).$$

We will denote the reciprocal polynomial of $Q(x)$ by $Q^R(x)$.

Mystery solved! The quotient polynomial in our long division problem is the reciprocal polynomial of the difference of generating functions. One implication is that we could have used either expression, $Q(x)$ or $Q^R(x)$, to evaluate the sums we considered in the opening section. Going forward, we will use the generating function approach for evaluating sums since it provides a construction that works for any sequence determined by a recurrence relation like equation 1. However, we note that the equivalent polynomial division form is useful too; it is a finite expression, which means that we do not have to worry about convergence issues in our series manipulations. (This is also discussed in Niven's pretty survey article [11].) Moreover, it provides a way to build cool division problems to impress your mathematician friends.

Some examples

In the opening section, we focused on a variety of summation formulas involving Fibonacci numbers that arose from formulas like equations 4 or 11. In this section, we illustrate the results from the previous section by finding summation formulas associated to two sequences defined by other recurrence relations.

Tribonacci numbers $\{t_n\}$ The Tribonacci numbers are relatives of the Fibonacci numbers. They start off with $t_0 = t_1 = 0$, $t_2 = 1$ and, as their name suggests, satisfy the recurrence

$$t_n - t_{n-1} - t_{n-2} - t_{n-3} = 0. \tag{12}$$

That is, the nth Tribonacci number in the sequence is the sum of the previous three terms, in contrast to the previous two terms for Fibonacci numbers. The first few terms of the Tribonacci sequence are $\{0, 0, 1, 1, 2, 4, 7, 13, \dots\}$.

To find the concise form of the generating function for the Tribonacci numbers, $G_t(x)$, we follow the approach outlined in the previous section and build a system analogous to equations 8 using $G_t(x)$, $-xG_t(x)$, $-x^2G_t(x)$, and $-x^3G_t(x)$. The result is

$$G_t(x) = \frac{t_0 + (t_1 - t_0)x + (t_2 - t_1 - t_0)x^2}{1 - x - x^2 - x^3} = \frac{x^2}{1 - x - x^2 - x^3}. \tag{13}$$

Similarly, and using equation 12 to simplify, the remainder term has the form

$$R_n(x) = x^{n+1}\left(\frac{t_{n+1} + (t_{n+2} - t_{n+1})x + t_n x^2}{1 - x - x^2 - x^3}\right).$$

This means that we can write

$$\sum_{j=0}^{n} t_j\, x^j = \frac{x^2 - t_{n+1}x^{n+1} - (t_{n+2} - t_{n+1})x^{n+2} - t_n x^{n+3}}{1 - x - x^2 - x^3}. \tag{14}$$

Let us evaluate this expression at $x = 1$ to sum the Tribonacci numbers. The right-hand side will give the formula for the sum:

$$\sum_{j=0}^{n} t_j = \frac{1 - t_{n+1} - (t_{n+2} - t_{n+1}) - t_n}{-2} = \frac{1}{2}(t_{n+2} + t_n - 1).$$

This formula is known, but not well-known. It has appeared in a couple of recent papers: in an article by Choi and Jo [3], where the formula is derived using a tabular format and recurrence relations, and in one by Kiliç [7], using a matrix approach.

We can get other identities by using different values of x. For example, assume that n is even and set $x = i$ (so $1 - i - i^2 - i^3 = 2i$). When we equate real and complex parts and simplify we get the alternating sum identities:

$$\sum_{j=0}^{n}(-1)^j t_{2j} = \frac{1}{2}(t_{2n+2} - t_{2n+1} - 1)$$

$$\sum_{j=1}^{n}(-1)^j t_{2j-1} = \frac{1}{2}(t_{2n+1} - t_{2n}).$$

These identities can be found in a paper of Frontczak [4], where they are derived using an explicit formula for Tribonacci numbers and a telescoping series argument.

Finally, we can take the derivative of both sides of equation 14 and evaluate them at $x = 1$. The result is the weighted sum of Tribonacci numbers

$$\sum_{j=1}^{n} j\, t_j = \frac{1}{2}(1 + nt_n - t_{n+1} + (n-1)t_{n+2}).$$

The sequence $\{f_{3n}\}$ We examine a subsequence of $\{f_n\}$ as a further illustration of the utility of our methods. The sequence of every third Fibonacci number beginning with $f_0 = 0$ is $\{0, 2, 8, 34, \ldots\}$. By repeatedly using the Fibonacci recurrence in equation 1, we can derive a recurrence relation for f_{3n} with $n \geq 2$ in terms of previous Fibonacci numbers:

$$f_{3n} = f_{3n-1} + f_{3n-2} = 2f_{3n-2} + f_{3n-3} = 3f_{3n-3} + 2f_{3n-4}. \tag{15}$$

We can do something similar starting with f_{3n-6}, but working this time in the other direction.

$$f_{3n-6} = f_{3n-4} - f_{3n-5} = 2f_{3n-4} - f_{3n-3}. \tag{16}$$

We equate the expressions for $2f_{3n-4}$ from equations 15 and 16 to get the recurrence

$$f_{3n} - 4f_{3n-3} - f_{3n-6} = 0.$$

Following our example from the previous section, we find the concise form of the generating function $G_{f_{3n}}(x)$ for the sequence $\{f_{3n}\}$ and an expression for $R_n(x)$:

$$G_{f_{3n}}(x) = \frac{f_0 + (f_3 - 4f_0)}{1 - 4x - x^2} = \frac{2}{1 - 4x - x^2},$$

and

$$R_n(x) = x^{n+1}\left(\frac{f_{3n+3} + f_{3n}x}{1 - 4x - x^2}\right).$$

The formula that results when we take the difference $G_{f_{3n}}(x) - R_n(x)$ is

$$\sum_{j=0}^{n} f_{3j}\, x^j = \frac{2 - f_{3n+3}x^{n+1} - f_{3n}x^{n+2}}{1 - 4x - x^2}.$$

At $x = 1$, this produces the identity

$$\sum_{j=0}^{n} f_{3j} = \frac{1}{4}(f_{3n+3} + f_{3n} - 2),$$

which is a special case of a theorem by Koshy [8, Theorem 5.11]. We leave it to the reader to find other summation formulas using this sequence and the techniques described in the opening section. We also note that it is possible to derive variations of equations 15 and 16 to build recurrences for subsequences $\{f_{kn}\}$ for other $k > 1$. (See also Koshy [10, Chapter 19, Problem 17].)

The Hadamard product

At this point, we can build summation formulas and weighted summation formulas for any series whose concise form is a rational generating function. In this section, we show how our methods can be adapted to find various identities involving sums of products of terms arising from two generating functions. To do this, we need a way to determine the concise form of a generating function such as

$$\sum_{j=0}^{n} f_j t_j x^j, \tag{17}$$

whose coefficients are the product of Fibonacci and Tribonacci numbers. This will greatly expand the collection of sequences we can work with. We remark that, at least initially, we do not know the recurrence relation for the coefficients in equation 17. The *Hadamard product* of the generating functions $F(x) = \sum_{j=0}^{\infty} f_j x^j$ and $G(x) = \sum_{j=0}^{\infty} g_j x^j$ is the generating function

$$F(x) * G(x) = \sum_{j=0}^{\infty} f_j g_j x^j.$$

One can prove that when $F(x)$ and $G(x)$ have concise forms that are rational functions, then so does the generating function $F(x) * G(x)$ (see Stanley [12, Proposition 4.2.5]). We will show how to do this explicitly, following an approach by Yuan [14].

Before we do so, we introduce a bit more background on generating function solutions, welcoming back our friend the reciprocal polynomial. We write the concise forms of $F(x)$ and $G(x)$ as $\frac{P_f(x)}{Q_f(x)}$ and $\frac{P_g(x)}{Q_g(x)}$ respectively, where $\deg Q_f(x) = m_1$ and $\deg Q_g(x) = m_2$. Denote by $\{r_k\}$ the set of m_1 complex roots of $Q_f(x)$'s reciprocal polynomial, $Q_f^R(x)$. One can show, using the partial fraction decomposition for $\frac{P_f(x)}{Q_f(x)}$ and geometric series, that when the roots are distinct, $f_n = \sum_{j=1}^{m_1} c_j r_j^n$ for some complex constants c_j. A slightly more complicated result holds when roots are repeated [2, Section 7.4]. Similarly, let $\{s_l\}$ denote the set of m_2 complex roots of $Q_g^R(x)$. Then the coefficients of the generating function for the Hadamard product $F(x) * G(x)$ can be written as a weighted sum of powers of the $m_1 m_2$ products $\{r_k s_l\}$. Therefore, the denominator of the concise form for $F(x) * G(x)$ will be the polynomial whose reciprocal polynomial has $\{r_k s_l\}$ as roots.

We recast the problem into one of linear algebra. By observation 3 after equation 9 in the generating function section, the denominator of a generating function, $Q(x)$, always has 1 as its constant term, so $Q^R(x)$ is a monic polynomial, say

$$Q^R(x) = x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0.$$

The polynomial $Q^R(x)$ has an associated companion matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{m-1} \end{bmatrix},$$

from which $Q^R(x)$ can be extracted as the characteristic polynomial, $\det(xI - A)$. Therefore, the roots of $Q^R(x)$ are precisely the eigenvalues of A. This observation allows us to build separate matrices with eigenvalues $\{r_k\}$ and $\{s_l\}$, but we still need a way to build a matrix whose eigenvalues are the products $\{r_k s_l\}$.

There is another construction from linear algebra that is just the thing we need: the Kronecker product. Given an $m \times n$ matrix A and a $p \times q$ matrix B, the *Kronecker product* (or tensor product) of A and B is the $mp \times nq$ matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

One of the useful properties of the Kronecker product is that if the eigenvalues of A and B are $\{r_k\}$ and $\{s_l\}$, respectively, then $\{r_k s_l\}$ are precisely the eigenvalues of $A \otimes B$ [6, Theorem 4.2.12]. This means that the reciprocal polynomial of $\det(xI - A \otimes B)$ will be the denominator of the concise form for $F(x) * G(x)$.

We demonstrate this technique on the generating function in equation 17. From equation 9, the denominator of the concise form of the rational generating function for the Fibonacci numbers is $1 - x - x^2$, so its reciprocal polynomial is $x^2 - x - 1$. Similarly, for Tribonacci numbers, the reciprocal polynomial is $x^3 - x^2 - x - 1$ by equation 13. The associated companion matrices are

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

The Kronecker product $A \otimes B$ is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

with associated characteristic polynomial

$$Q^R(x) = x^6 - x^5 - 4x^4 - 5x^3 - 2x^2 + x - 1,$$

so the denominator of the concise form of the rational generating function for equation 17 is $Q(x) = 1 - x - 4x^2 - 5x^3 - 2x^4 + x^5 - x^6$. We can determine the numerator $P(x)$ via multiplication using the equation

$$\frac{P(x)}{Q(x)} = \sum_{j=0}^{\infty} f_j t_j x^j = f_0 t_0 + f_1 t_1 x + f_2 t_2 x^2 + \cdots .$$

On the right-hand side, the generating function for the sequence $\{f_j t_j\}$ starts off as $x^2 + 2x^3 + 6x^4 + 20x^5 + 56x^6 + \cdots$. By observation 2 after equation 9 in the generating function section, the numerator of the generating function will have a maximum degree of 5, so we only need a partial calculation.

$$\begin{aligned}(1 - x - 4x^2 - 5x^3 - 2x^4 + x^5 - x^6)(x^2 + 2x^3 + 6x^4 + 20x^5 + 56x^6)\\ = x^2 + x^3 + x^5 - 169x^7 - 335x^8 + \cdots = x^2 + x^3 + x^5 + O(x^6).\end{aligned}$$

We conclude that

$$\sum_{j=0}^{\infty} f_j t_j x^j = \frac{x^2 + x^3 + x^5}{1 - x - 4x^2 - 5x^3 - 2x^4 + x^5 - x^6}. \tag{18}$$

One consequence of equation 18 is that the sequence $\{a_j\} = \{f_j t_j\}$ satisfies the recurrence relation

$$a_n = a_{n-1} + 4a_{n-2} + 5a_{n-3} + 2a_{n-4} - a_{n-5} + a_{n-6}.$$

We can take this recurrence relation as well as the concise form in equation 18 and apply our results on generating functions to find new summation formulas. Here is one that results from setting $x = -1$:

$$\begin{aligned}\sum_{j=0}^{n} (-1)^j f_j t_j = 1 + (-1)^{n+1}(-f_{n+2}t_{n+2} + 3f_{n+3}t_{n+3}\\ + 2f_{n+4}t_{n+4} + 2f_{n+5}t_{n+5} - f_{n+6}t_{n+6}).\end{aligned}$$

We end this section with a couple of remarks for readers who know a bit more about generating functions.

1. The methods outlined in this section work when a polynomial denominator from the concise form of the generating function has repeated roots since the multiplicity of eigenvalues is preserved in the Kronecker product [6, Theorem 4.2.12]. The resulting construction may result in a polynomial whose degree is strictly larger than the minimum necessary for the generating function of the Hadamard product. In such cases, the concise form of the generating function will not be in lowest terms.
2. Given a generating function $G(x)$, it is well-known that $\frac{1}{1-x}G(x)$ is a generating function whose terms are partial summations of the sequence associated to $G(x)$. However, this observation does not allow one to easily determine a formula for those partial summations.

3. One can remove a step from the procedure in this section by noting that $\det(I - xA \otimes B)$ is the reciprocal polynomial of $\det(xI - A \otimes B)$.

We have found the generating function framework a powerful tool, able to provide summation formulas in a uniform way for a wide range of interesting examples. Now it is your turn—happy hunting!

REFERENCES

[1] Benjamin, A., Quinn, J. (2003). *Proofs that Really Count*. Washington, DC: Math. Assoc. Amer.

[2] Brualdi, R. (2010). *Introductory Combinatorics*, 5th ed. Upper Saddle River: Pearson Prentice Hall.

[3] Choi, E., Jo, J. (2015). On partial sum of Tribonacci numbers, *Int. J. Math. Math. Sci.*, Art. ID 301814. doi.org/10.1155/2015/301814

[4] Frontczak, R. (2018). Sums of Tribonacci and Tribonacci-Lucas numbers. *Int. J. Math. Math. Anal.* 12(1): 19–24. DOI: 10.12988/ijma.2018.712153

[5] Hansen, R. (1972). Generating Identities for Fibonacci and Lucas triples, *Fib. Quart.* 10: 571–578. DOI: 10.12988/ijma.2018.712153

[6] Horn, R., Johnson, C. (1991). *Topics in matrix analysis*. Cambridge: Cambridge Univ. Press.

[7] Kiliç, E. (2008). Tribonacci sequences with certain indices and their sums, *Ars Combin.* 86: 13–22.

[8] Koshy, T. (1998). New Fibonacci and Lucas identities, *Math. Gazette* 82(495): 481–484. doi.org/10.2307/3619903

[9] Koshy, T. (2000). Weighted Fibonacci and Lucas sums, *Math. Gazette* 85(502): 93–96. doi.org/10.2307/3620481

[10] Koshy, T. (2001). *Fibonacci and Lucas Numbers with Applications*. New York: Wiley-Interscience.

[11] Niven, I. (1969). Formal power series. *Amer. Math. Monthly* 76(8): 871–889. doi.org/10.1080/00029890.1969.12000359

[12] Stanley, R. (2012). *Enumerative Combinatorics*, Vol. 1, 2nd ed. Cambridge: Cambridge Univ. Press.

[13] Wilf, H. (2006). *generatingfunctionology*, 3rd ed. Wellesley: A. K. Peters, Ltd.

[14] Yuan, Q. (2017). Algorithm for computing Hadamard product of two rational generating functions. Available at: https://math.stackexchange.com/q/2520666. Last accessed September 2022.

Summary. We describe a general method that finds closed forms for partial sums of power series whose coefficients arise from linear recurrence relations. These closed forms allow one to derive a vast collection of identities involving the Fibonacci numbers and other related sequences. Although motivated by a polynomial long division problem, the method fits naturally into a standard generating function framework. We also describe an explicit way to calculate the generating function of the Hadamard product of two generating functions, a construction on power series which resembles the dot product. This allows one to use the method for many examples where the recurrence relation for the coefficients is not initially known.

ETHAN BERKOVE (MR Author ID: 608283) received his Ph.D. from the University of Wisconsin, Madison, and has taught at Lafayette College since 1999. He enjoys working collaboratively on engaging math problems wherever he finds them. His interests include hiking, biking, and reading (when he can find the time). He would like to thank his colleagues, Jonathan Bloom and Gary Gordon, as well as anonymous referees, for comments and conversations which greatly improved the exposition and content of this paper. He lives in Easton, Pennsylvania with his wife and two sons.

MIKE BRILLESLYPER (MR Author ID: 994411) received his Ph.D. from the University of Arizona. After a 21-year career on the faculty at the U. S. Air Force Academy, he is now the department chair of applied mathematics at Florida Polytechnic University. He enjoys teaching a variety of courses and collaborating with colleagues on interesting problems, many of which start out as undergraduate research projects. Mike and his wife MaryAnn (an amazing musician and teacher) enjoy exploring their new state and following the adventures of their two daughters: Emma and Meg.

A Geometric Generalization of the Pythagorean Means

GREG MARKOWSKY
Monash University
Clayton VIC 3800, Australia
greg.markowsky@monash.edu

DYLAN PHUNG
Yarra Valley Grammar School
Ringwood VIC 3134, Australia
dkp1@yvg.vic.edu.au

DAVID TREEBY
Monash University
Clayton VIC 3800, Australia
david.treeby@gmail.com

Consider two positive numbers a and b. The three classical *Pythagorean means* are the *arithmetic mean* $A(a, b)$, the *geometric mean* $G(a, b)$, and the *harmonic mean* $H(a, b)$, defined as

$$A(a,b) = \frac{a+b}{2}, \quad G(a,b) = \sqrt{ab}, \quad H(a,b) = \frac{2ab}{a+b}.$$

It is well known that these means are ordered according to the so-called inequality of the means:

$$\min(a,b) \le H(a,b) \le G(a,b) \le A(a,b) \le \max(a,b),$$

with equality holding if and only if $a = b$. The inequality can be illustrated by comparing segment lengths in Figure 1 [1, p. 8]. However, this figure does not constitute a self-evident proof since one must first confirm that these lengths actually correspond to the means indicated.

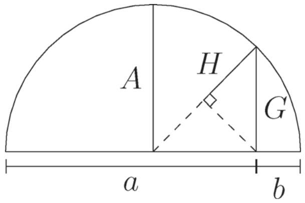


Figure 1 A visual illustration of the inequality of the means.

In this paper we consider an alternative geometric construction that makes the inequality of the means intuitively obvious. This construction was first considered by Gordon and Dietel [3], who showed that each of these classic means correspond naturally to the geometric centroid of a bounded region in the plane. Our contribution is to provide a short proof that substantiates the inequality. We should note that Stolarsky [7] gives a longer proof of a more general result in his seminal paper investigating generalized means. This approach was further refined in a more abstract paper by Leach and Sholander [5].

Math. Mag. **95** (2022) 520–524. doi:10.1080/0025570X.2022.2125729 © Mathematical Association of America
MSC: Primary 26E60, Secondary 26D99

We should also mention that various authors have published papers on generalizations of the Pythagorean means, often with a geometrical flair. A recent example of this is given by Gordon [4], where geometric intuition features prominently. Finally, Pearce [6] explores similar ideas related to convexity and integration.

The generalized mean

For real number s, define $f_s(x) = x^s$. We consider the region

$$\{(x, y) \in \mathbb{R}^2 : a \leq x \leq b, 0 \leq y \leq f_s(x)\},$$

shown in Figure 2.

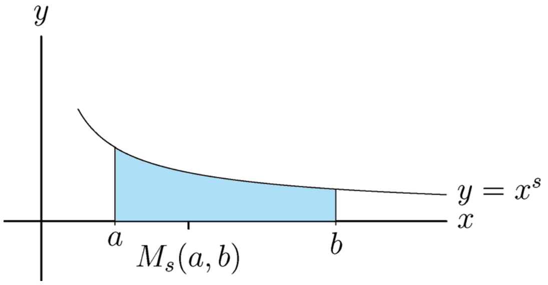


Figure 2 The generalized mean as a geometric centroid.

If $a < b$, then the x-coordinate of this region's centroid is

$$M_s(a, b) = \frac{\int_a^b x f_s(x)\, dx}{\int_a^b f_s(x)\, dx} = \frac{\int_a^b x^{s+1}\, dx}{\int_a^b x^s\, dx}.$$

We also define $M_s(a, a) = a$. This accords with the fact that $\lim_{t \to a} M_s(a, t) = a$. Gordon and Dietel observed that $\{M_s(a, b)\}_{s \in \mathbb{R}}$ is a one parameter family of means that generalizes each of the of classic means. Indeed, the reader can readily confirm that when s is -3,$-\frac{3}{2}$, or 0 we obtain

$$H(a, b) = M_{-3}(a, b), \qquad G(a, b) = M_{-3/2}(a, b), \qquad A(a, b) = M_0(a, b).$$

If $s = -1$, then

$$M_{-1}(a, b) = \frac{b - a}{\ln b - \ln a}.$$

This is known as the *logarithmic-mean* of a and b, and it is used often in various fields of engineering. We denote it by $L(a, b)$. Furthermore, if we further define

$$M_{-\infty}(a, b) = \lim_{s \to -\infty} M_s(a, b) \quad \text{and} \quad M_\infty(a, b) = \lim_{s \to \infty} M_s(a, b),$$

then it is a straightforward calculation to show that

$$M_{-\infty}(a, b) = \min(a, b) \quad \text{and} \quad M_\infty(a, b) = \max(a, b).$$

It is not difficult to show that $M_s(a, b)$ satisfies a range of properties that we would hope for any mean to possess:

1.	Value preservation	$M_s(a, a) = a$
2.	Homogeneity	$M_s(ra, rb) = rM_s(a, b)$
3.	Invariance under exchange	$M_s(a, b) = M_s(b, a)$
4.	Averaging	$\min(a, b) \leq M_s(a, b) \leq \max(a, b)$

In summary, we see that each of the classic means can be interpreted as the geometric centroid of a region in the plane, and each of these is illustrated in Figure 3. From this figure, the inequality of the means is apparent:

$$\min(a, b) \leq H(a, b) \leq G(a, b) \leq L(a, b) \leq A(a, b) \leq \max(a, b).$$

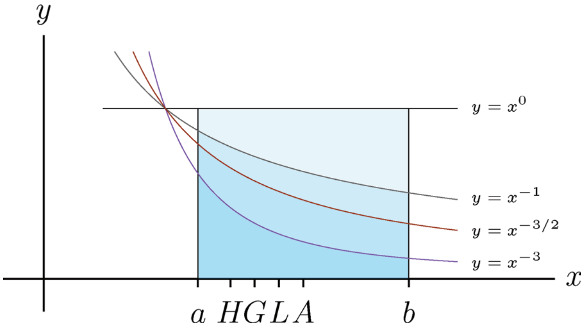


Figure 3 Each of the means is the centroid of one the regions.

Note that decreasing s has the effect of concentrating points closer to a, and this serves to shift the centroid in the same direction. However, while it is intuitively clear that $M_s(a, b)$ increases monotonically in s, a proof of this fact is not entirely straightforward.

Monotonicity of the generalized mean

To prove that $M_s(a, b)$ increases monotonically in s, we require a preliminary definition. We call a twice differentiable function **convex**[*] on an interval if $f''(t) \geq 0$ for all points t in the interval. That convex functions have increasing gradients makes the following result intuitive. Here, we show that such functions have "increasing differences," in the sense given by Lemma 1, and illustrated in Figure 4.

Lemma 1. Suppose that f is twice differentiable and convex on an interval I. Then for any $d \geq 0$ and $s \leq t$, we have that:

$$f(s + d) - f(s) \leq f(t + d) - f(t)$$

for all points $s, t, s + d, t + d$ in I.

Proof. As $f''(x) \geq 0$, the function f' is increasing. Therefore, if $g(x) = f(x + d) - f(x)$, then $g'(x) = f'(x + d) - f'(x) \geq 0$, in which case g' is an increasing function. We conclude that $g(s) \leq g(t)$, which completes the proof. ■

[*] It is, of course, possible to give a definition of convexity that does not mention differentiability, but this is not required for our purposes.

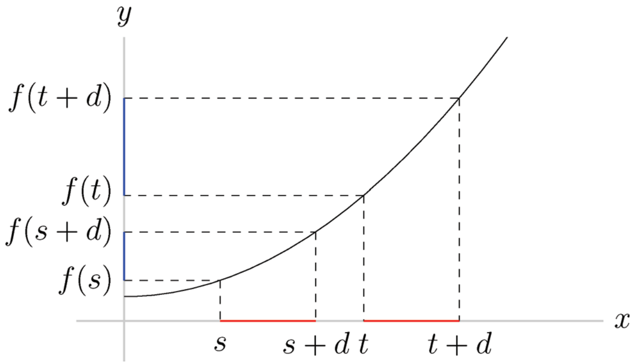


Figure 4 The increasing differences of a convex function.

With this simple lemma at hand, we now prove the monotonicity result that yields the inequality of the means. The existing proof of this result can be found in [7, p. 89], and although this requires only elementary calculus, our innovation is to shorten the proof by a simple application of the Cauchy-Schwarz inequality for integrals. That is, if f and g are continuous on the closed interval $[a, b]$, then

$$\left(\int_a^b f(x)\cdot g(x)\,dx\right)^2 \leq \int_a^b (f(x))^2\,dx \cdot \int_a^b (g(x))^2\,dx.$$

Theorem 1. Let a, b be positive real numbers. If $s \leq t$, then $M_s(a, b) \leq M_t(a, b)$.

Proof. We first define $f(t) = \int_a^b x^t\,dx$, noting that $M_t(a, b) = \frac{f(t+1)}{f(t)}$. We aim to show that if $s \leq t$, then

$$\frac{f(s+1)}{f(s)} \leq \frac{f(t+1)}{f(t)}.$$

Equivalently,

$$(\log\circ f)(s+1) - (\log\circ f)(s) \leq (\log\circ f)(t+1) - (\log\circ f)(t).$$

That is, it is sufficient to show that $\log\circ f$ has increasing differences. By Lemma 1, this is guaranteed if we can show that $\log\circ f$ is convex. However, by twice differentiating $\log\circ f$, we see that

$$(\log\circ f)''(x) \geq 0 \Leftrightarrow f''(t)f(t) \geq (f'(t))^2.$$

We will prove the later inequality. To this end, by twice differentiating under the integral* we obtain

$$\begin{aligned} f''(t)\cdot f(t) &= \frac{d^2}{dt^2}\left(\int_a^b x^t\,dx\right)\cdot\int_a^b x^t\,dx \\ &= \int_a^b x^t(\log x)^2\,dx \cdot \int_a^b x^t\,dx \\ &= \int_a^b (x^{\frac{t}{2}}\log x)^2\,dx \cdot \int_a^b (x^{\frac{t}{2}})^2\,dx. \end{aligned}$$

*This is permissible by the Leibniz integral rule, the preconditions of which are readily met as $g(x, t) = s^t$ is continuously differentiable. See Conrad [2] for an excellent exposition on this topic.

This final product is then estimated using the Cauchy-Schwarz Inequality, giving

$$f''(t) \cdot f(t) \geq \left(\int_a^b x^t \log x \, dt \right)^2 = (f'(t))^2.$$

This both proves that $\log \circ f$ is convex and completes the proof. ■

Acknowledgment The authors would like to thank an anonymous referee for valuable comments that have improved the paper.

REFERENCES

[1] Alsina, C., Nelsen, R. (2006). *Math Made Visual: Creating Images for Understanding Mathematics*. Washington, D.C.: Mathematical Association of America.

[2] Conrad, K. (2018). *Differentiating under the integral sign*. https://kconrad.math.uconn.edu/blurbs/analysis/diffunderint.pdf Last accessed September 2022.

[3] Dietel, B., Gordon, R. (2003). Using tangent lines to define means. *Math. Mag.* 76(1): 52–61. doi.org/10.1080/0025570X.2003.11953947

[4] Gordon, R. (2019). Geometric problems leading to power means. *Math. Gazette.* 103(557): 318–323. doi.org/10.1017/mag.2019.67.

[5] Leach, E., Sholander, M. (1978). Extended mean values. *Amer. Math. Monthly*. 85(2): 84–90. doi.org/10.1080/00029890.1978.11994526.

[6] Pearce, C., Pečarić, J., Šimić, V. (1998). Stolarsky means and Hadamard's inequality. *J. Math. Anal. App.* 220(1): 99–109. doi.org/10.1006/jmaa.1997.5822

[7] Stolarsky, K. (1975). Generalizations of the logarithmic mean. *Math. Mag.* 48(2): 87–92. doi.org/10.1080/0025570X.1975.11976447

Summary. Each of the Pythagorean means corresponds to the centroid of a region in the Cartesian plane. We show how this insight leads to a short proof of a result that generalizes the HM-GM-AM inequality.

GREG MARKOWSKY obtained his Ph.D. at the City University of New York (CUNY) and is now a lecturer at Monash University. His research interests include complex analysis, probability, graph theory, and good old calculus, as is used in this paper.

DYLAN PHUNG is a student at Yarra Valley Grammar School.

DAVID TREEBY is a research associate at Monash University, where he obtained his Ph.D. He enjoys digging for gems along the boundary of classical physics and mathematics. He also teached at Scotch College, Melbourne.

From Uniform Boundedness to the Boundary Between Convergence and Divergence

EHSSAN KHANMOHAMMADI
Schenectady, NY 12308
ehssan@pm.me

OMID KHANMOHAMADI
Seattle, WA 98164
omidmath@pm.me

Three of the fundamental ideas Stefan Banach introduced in functional analysis together lead to his discovery of three fundamental results [6, 7, section VI.84]. The ideas were *abstract points* (functions as points, leading to operators and function spaces), *abstract sizes* (norms of functions, leading to distances between functions), and *abstract limits* (limits of sequences of functions, leading to completeness of function spaces). The results were the *uniform boundedness principle*, the *open mapping theorem*, and the *closed graph theorem*, which are all interrelated, in the sense that in complete normed vector spaces (known as Banach spaces), Baire's category theorem leads to several equivalences between qualitative properties (e.g., finiteness, surjectivity, regularity) and quantitative properties (e.g., estimates) of continuous (or equivalently bounded) linear operators [11, section 1.7]. One of these equivalences is captured by the uniform boundedness principle, also known as the Banach-Steinhaus theorem. In this article, we introduce a dual of the uniform boundedness principle which does *not* require *completeness* and gives an indirect means for testing the boundedness of a set. The dual principle, although known to analysts and despite its applications in establishing results such as the Hellinger-Toeplitz theorem, is often missing from elementary treatments of functional analysis. In Example 1, we indicate a connection between the dual principle and a question in the spirit of du Bois-Reymond regarding the boundary between convergence and divergence of sequences. This example is intended to illustrate why the statement of the principle is natural and clarify what the principle claims and what it does not.

Unbounded sets in normed spaces

We begin with a proposition of linear algebraic flavor about the relation between unbounded subsets of a normed space and the linear functionals on that space. In what follows, we shall assume that all vector spaces are over the field $\mathbb{R}$, and all linear maps between them are real-linear, although our results carry over easily to the field of complex numbers.

Proposition 1. *Let S be an unbounded subset of a normed vector space X. Then there exists a linear functional $\phi\colon X \to \mathbb{R}$ whose restriction to S has an unbounded image in $\mathbb{R}$.*

We include a simple proof here for the sake of completeness.

Proof. First assume X is finite dimensional and let $n = \dim X$. Then since all norms on any finite-dimensional normed space are equivalent, we will assume that the norm of X is induced by an inner product. Fix an orthonormal basis $\{e_1, \ldots, e_n\}$ with respect

Math. Mag. **95** (2022) 525–531. doi:10.1080/0025570X.2022.2126254 © Mathematical Association of America
MSC: Primary 40A05, Secondary 46-01

to this inner product for X, and for $i = 1, \dots, n$ let $\mathrm{Proj}_{e_i}: X \to \mathbb{R}$ denote the scalar projection onto the ith coordinate:

$$\mathrm{Proj}_{e_i}(c_1e_1 + \cdots + c_ne_n) = c_i \quad \text{for any } c_1, \dots, c_n \in \mathbb{R}.$$

We claim that for some value of i, the restriction of Proj_{e_i} to S has an unbounded image. Indeed, if we had

$$\sup_{x \in \mathrm{Proj}_{e_i} S} |x| \leq M_i \quad \text{with } M_i \geq 0 \text{ for all } i = 1, \dots, n,$$

then it would follow that $\sup_{s \in S} \|s\| \leq \sqrt{\sum M_i^2} < \infty$, contrary to the unboundedness of S.

The same argument proves the proposition that if S (or an unbounded subset of S) is contained in a finite-dimensional subspace of X, or equivalently, if $\dim \mathrm{Span}\, S < \infty$.

So we will assume that S is not contained in any finite-dimensional subspace of X and then proceed to find an infinite linearly independent subset $\{b_1, b_2, \dots\}$ of S and extend $\{b_1, b_2, \dots\}$ to a possibly uncountable Hamel basis B of X. Now define a function ϕ on B by

$$\phi(b) = \begin{cases} k & \text{if } b = b_k, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for } b \in B,$$

and extend ϕ linearly to a functional, also denoted by ϕ, on the entire space X. By construction, the restriction of ϕ to $\{b_1, b_2, \dots\} \subset S$ has an unbounded image. This completes the proof. ■

The restriction of the functional ϕ to S in the above proof had an unbounded image, as we required. However, the functional itself might also be "unbounded" or discontinuous, an unwelcome phenomenon in analysis. Therefore, we can ask whether unboundedness of S can be captured by a *continuous* linear functional. Although it may not be clear a priori, this question is closely related to the famous uniform boundedness principle in analysis. We explore this relation in the next two sections. As we shall see, a central role in this regard is played by the notion of the *operator norm*, denoted by $\|T\|_{\mathrm{op}}$, of a linear map of normed spaces $T: X \to Y$. We say that T is *bounded* when the operator norm defined by

$$\|T\|_{\mathrm{op}} = \sup_{\|x\|_X \leq 1} \|Tx\|_Y$$

is finite. In words, T is said to be bounded (as a function) if the image of the unit ball under T is bounded (as a set). A simple observation that shows the importance of this definition in analysis is that boundedness and continuity are equivalent properties for linear maps of normed spaces.

Uniform boundedness principle

Before introducing its dual, let us first give a "quantitative" version of the uniform boundedness principle itself.

Theorem 1 (Uniform boundedness principle)**.** *Let X be a Banach space and let Y be a normed space. Consider a family F of bounded linear operators $T: X \to Y$. If F is pointwise bounded, then it is uniformly bounded.*

If F is not uniformly bounded, then there exists a point $x \in X$ and a sequence (T_n) of operators in F satisfying

$$\|T_{n+1}\|_{op} > \|T_n\|_{op}, \quad \|T_{n+1}x\|_Y > \|T_n x\|_Y$$

for all n, and $\|T_n x\|_Y \to \infty$.

Proof. Suppose F is not uniformly bounded. Then we can find a sequence (T_n) of nonzero operators in F such that

$$\|T_{n+1}\|_{\text{op}} \geq 4^{2n+1}\|T_n\|_{\text{op}}.$$

Choose unit vectors x_n such that

$$\|T_n x_n\|_Y \geq \frac{1}{2}\|T_n\|_{\text{op}}.$$

For any $a, b \in X$, by the triangle inequality, at least one of the two inequalities

$$\|a + b\|_X \geq \|b\|_X \qquad \text{and} \qquad \|a - b\|_X \geq \|b\|_X$$

must hold. Thus, we may define a vector x by

$$x = \sum_{k=1}^{\infty} \sigma(k) 4^{-k} x_k,$$

where $\sigma(k)$ takes its values from $\{\pm 1\}$ and is defined recursively so that

$$\left\| \sum_{k=1}^{n} \sigma(k) 4^{-k} T_n x_k \right\|_Y \geq \|4^{-k} T_n x_n\|_Y \geq \frac{1}{2} 4^{-n} \|T_n\|_{\text{op}}.$$

Note that the series defining x is absolutely convergent and hence convergent by completeness of X. The triangle inequality then implies that

$$\begin{aligned} \|T_n x\|_Y &\geq \left\| \sum_{k=1}^{n} \sigma(k) 4^{-k} T_n x_k \right\|_Y - \left\| \sum_{k=n+1}^{\infty} \sigma(k) 4^{-k} T_n x_k \right\|_Y \\ &\geq \frac{1}{2} 4^{-n} \|T_n\|_{\text{op}} - \frac{1}{3} 4^{-n} \|T_n\|_{\text{op}} = \frac{1}{6} 4^{-n} \|T_n\|_{\text{op}}. \end{aligned} \tag{1}$$

Since $\|x\|_X \leq \frac{1}{3}$, we have $\|T_n x\|_Y \leq \frac{1}{3}\|T_n\|_{\text{op}}$ and hence equation (1) yields

$$\begin{aligned} \|T_{n+1} x\|_Y &\geq \frac{1}{6} 4^{-n-1} \|T_{n+1}\|_{\text{op}} \\ &\geq \frac{1}{6} 4^{-n-1} 4^{2n+1} \|T_n\|_{\text{op}} = \frac{1}{6} 4^n \|T_n\|_{\text{op}} \\ &> \frac{1}{3} \|T_n\|_{\text{op}} \geq \|T_n x\|_Y, \end{aligned}$$

as desired. ■

Remark. Over the years, there have been numerous proofs of the uniform boundedness principle. These proofs may be categorized into those which use Baire's category theorem ("non-elementary" proofs) and those which do not ("elementary" proofs). Out of the "elementary" proofs, the "simple" ones are of special interest; they usually make use of a “gliding hump argument,” such as the ones given by [3, p. 51] or [10]. Carothers [3] reports that the original proof of the principle by Steinhaus and his protege Banach must have been an elementary proof of this kind, but apparently it was lost during the war. The nonelementary proof that survived was suggested as an alternative proof by Saks, who refereed their paper! Some other elementary proofs (such as the one given by Riesz and Sz.-Nagy [9, p. 63]) make use of a “nested ball” argument, similar to the argument used in the proof of Baire’s category theorem. The advantage of the proof given above is its “constructive” nature (as opposed to most other proofs that are proofs by contradiction) which allows us to give a quantitative version of the uniform boundedness principle that we shall use in proving Theorem 2.

Norms in the codomain and its dual Since the dual involves the codomain Y, let us say a few words about Y in the uniform boundedness principle, which is merely a normed space. One of the easy consequences of the Hahn-Banach theorem is the duality between the definitions of the norm in Y and in its dual Y^* consisting of *bounded* linear maps $y^*\colon Y \to \mathbb{R}$. More precisely, for any $y^* \in Y^*$,

$$\|y^*\|_{\text{op}} = \sup_{\|y\| \le 1} |y^*(y)|,$$

and for any $y \in Y$,

$$\|y\| = \sup_{\|y^*\|_{\text{op}} \le 1} |y^*(y)|,$$

where in the second equality the supremum is attained.

Remark. The most basic examples of normed spaces are, of course, the scalar fields $\mathbb{R}$ and $\mathbb{C}$. It is perhaps interesting to note that the uniform boundedness principle for linear functionals (i.e., in the special case that the codomain is a scalar field) implies the same theorem for all linear operators using the above remark about the computation of norms. To see this, let F be a family of bounded operators $T\colon X \to Y$ between normed spaces with X complete. Suppose F is pointwise bounded so that $\|Tx\| \le M_x$ for some $M_x \ge 0$ depending on each x in the unit ball of X. Then for each y^* in the unit ball of Y^*, the functional $y^* \circ T$ is bounded and $|(y^* \circ T)(x)| \le M_x$. Thus by the uniform boundedness principle for functionals applied to the family $\{y^* \circ T \mid T \in F, y^* \in Y^* \text{ with } \|y^*\|_{\text{op}} \le 1\}$, we conclude that $|(y^* \circ T)(x)| \le M$ for some $M \ge 0$ independent of x. Taking the supremum over y^*, we obtain $\|Tx\| \le M$, as claimed.

A dual for the uniform boundedness principle

The appearance of the Hahn-Banach theorem, which is applicable to general (i.e., not necessarily complete) normed spaces, in the last section is not completely accidental. It suggests the idea that the uniform boundedness principle might have some applications in the context of general normed spaces as well. Our Hahn-Banach argument proves, in particular, the following theorem, which is where Hahn (1879–1934), Banach (1892–1945), and Steinhaus (1887–1972) meet, posthumously!

Theorem 2 (Dual for the uniform boundedness principle)**.** *Let S be a subset of a normed space X. If $\phi(S)$ is bounded for each $\phi \in X^*$, then S is bounded.*

If S is unbounded, then there exists $\phi \in X^*$ and a sequence (s_n) in S satisfying $\|s_{n+1}\|_X > \|s_n\|_X$, $|\phi(s_{n+1})| > |\phi(s_n)|$ for all n, and $|\phi(s_n)| \to \infty$.

Theorem 2 can be thought of as a dual for the uniform boundedness principle since the boundedness of $\phi(S)$ can be rephrased as the finiteness of $\sup_{s\in S}|\phi(s)|$. Note that this theorem gives an affirmative answer to the question that we raised after Proposition 1.

Proof. Since $\phi(S)$ is bounded, for $\phi \in X^*$,

$$\sup_{s\in S}|s^{**}(\phi)| = \sup_{s\in S}|\phi(s)| < \infty.$$

This shows that the hypotheses of the uniform boundedness principle are satisfied for $F = \{s^{**} \colon X^* \to \mathbb{R} \mid s \in S\}$, thanks to the fact that the dual of any normed space is complete. Therefore, by the uniform boundedness principle and the fact that the map $x \mapsto x^{**}$ is an isometry from X into the Banach space X^{**}, we have that

$$\sup_{s\in S}\|s\|_X = \sup_{s\in S}\|s^{**}\|_{X^{**}} < \infty,$$

as desired. The last assertion follows from the second part of Theorem 1. ■

Let us finish with a question about a possible strengthening of Theorem 2 that we shall pick up in the next section.

Question 1. Suppose $S = \{s_1, s_2, \ldots\}$ is a subset of a normed space X such that

$$\|s_{n+1}\|_X > \|s_n\|_X$$

for each n, and $\|s_n\|_X \to \infty$. Can we necessarily find a functional $\phi \in X^*$ satisfying $|\phi(s_{n+1})| > |\phi(s_n)|$ for each n, and $|\phi(s_n)| \to \infty$?

Boundary between convergence and divergence

We begin with a question concerning the boundary between convergence and divergence of series that first appeared in the work of Abel [1], Dini [4], and du Bois-Reymond [5].

Question 2. Suppose $\sum_{n=1}^{\infty} x_n$ is a convergent series with positive terms. Does there exist a sequence (y_n) such that $y_n \to \infty$ and $\sum_{n=1}^{\infty} x_n y_n < \infty$? Similarly, suppose $\sum_{n=1}^{\infty} x_n$ is a divergent series with positive terms. Does there exist a sequence (y_n) such that $y_n \to 0$ and $\sum_{n=1}^{\infty} x_n y_n = \infty$?

The answer to both of these, as is well-known, is affirmative [2, 8]. That is to say, there is neither a fastest convergent series nor a slowest divergent series. One can, of course, make analogous claims about sequences and, for instance, easily show that there is no slowest divergent sequence. Generalizing this, we pose a more restrictive question about the boundary between convergence and divergence of sequences.

Question 3. Suppose (x_n) is a sequence of numbers diverging to infinity. Does there exist a sequence (y_n) such that $\sum_{n=1}^{\infty} y_n < \infty$ and $x_n y_n \to \infty$? What if we require $(y_n) \in \ell^p$?

Let us provide a quick comparison of the claims made in Questions 2 and 3.

	Given	Wanted Convergence	Wanted Divergence
Q2	$\sum_{n=1}^{\infty} x_n = \infty$	$y_n \to 0$	$\sum_{n=1}^{\infty} x_n y_n = \infty$
Q3	$x_n \to \infty$	$\sum_{n=1}^{\infty} y_n < \infty$	$x_n y_n \to \infty$

Example 1. Let $x_n = \sqrt{n}$ for $n = 1, 2, \dots$. Now we ask whether there exists a sequence $(y_n) \in \ell^2$ such that $x_n y_n \to \infty$ as $n \to \infty$. What makes the sequence (x_n) worth studying in this context is the fact that $(1/\sqrt{n^{1+\epsilon}}) \in \ell^2$ for all $\epsilon > 0$, and $(1/\sqrt{n^{1-\epsilon}}) \notin \ell^2$ for all $\epsilon \geq 0$. Before answering this question, we indicate its connection with Theorem 2, the dual for the uniform boundedness principle.

Let (x_n) be a sequence of numbers such that $|x_{n+1}| > |x_n|$ for all n and $|x_n| \to \infty$ and let $\{e_1, e_2, \dots\}$ be the standard orthonormal basis for ℓ^2. Define a set S by

$$S = \{x_1 e_1, x_2 e_2, \dots\}.$$

Then S is an unbounded subset of ℓ^2 and hence, by Theorem 2, we can find $\phi \in (\ell^2)^*$ and a sequence (s_n) in S satisfying

$$\|s_{n+1}\|_2 > \|s_n\|_2, \quad |\phi(s_{n+1})| > |\phi(s_n)|$$

for all n, and

$$|\phi(s_n)| \to \infty.$$

But since $|x_{n+1}| > |x_n|$, we must have $s_k = x_{n_k} e_{n_k}$ for a subsequence (x_{n_k}) of (x_n). Thus,

$$|\phi(x_{n_{k+1}} e_{n_{k+1}})| > |\phi(x_{n_k} e_{n_k})|$$

for all n, and

$$|\phi(x_{n_k} e_{n_k})| \to \infty.$$

To reveal the connection between the dual for the uniform boundedness principle and Question 3, we use the Riesz representation theorem to establish the existence of a sequence $y = (y_n) \in \ell^2$ such that $\phi(x) = \langle x, y \rangle$ for all $x \in \ell^2$. Thus, for each k we find

$$\phi(x_{n_k} e_{n_k}) = \langle x_{n_k} e_{n_k}, y \rangle = x_{n_k} \langle e_{n_k}, y \rangle = x_{n_k} y_{n_k}.$$

In conclusion, Theorem 2 implies the existence of a square-summable sequence $(y_k) \in \ell^2$ such that

$$|x_{n_{k+1}} y_{k+1}| > |x_{n_k} y_k| \quad \text{and} \quad |x_{n_k} y_k| \to \infty$$

as $k \to \infty$ for a *subsequence* (x_{n_k}) of (x_n). But this statement is rather obvious! For instance, in the case of $(x_n) = (\sqrt{n})$ the subsequence (x_{n_k}) defined by $x_{n_k} = \sqrt{k^4} = k^2$ and the sequence $(y_k) = (\frac{1}{k})$ has the desired properties.

This cannot be done, however, if we do not allow the passage to subsequences as in Questions 1 and 3. To see this, we return to $(x_n) = (\sqrt{n})$ and let (y_n) be any sequence such that $x_n y_n \to \infty$. Then

$$x_n^2 y_n^2 = n y_n^2 \to \infty.$$

But since the harmonic series $\sum_{n=1}^{\infty} n^{-1}$ is divergent, and we are assuming that $y_n^2 / n^{-1} \to \infty$, the limit comparison theorem for series implies that $\sum_{n=1}^{\infty} y_n^2$ must also be divergent, i.e., $(y_n) \notin \ell^2$.

Questions of this nature arise in Fourier analysis and provide a means for measuring regularity of functions. For instance, for a periodic function $f \in L^2(\mathbb{T})$, we have

$\widehat{f} \in \ell^2(\mathbb{Z})$, and if $f \in C^k(\mathbb{T})$, $k \geq 0$, then $\widehat{f} \in o(n^{-k})$, where $\widehat{f}(n)$ is the nth Fourier coefficient of f given by

$$\widehat{f}(n) = \int_0^1 f(x)e^{-2\pi inx}\, dx$$

for each $n \in \mathbb{Z}$. See the Riesz-Fischer theorem and the Riemann-Lebesgue lemma [11]. Thus, "the smoother the function, the faster the decrease of its Fourier coefficients." Indeed, a standard application of integration by parts shows that if $f \in C^k(\mathbb{T})$, then $(n^k \widehat{f}(n)) \in \ell^2(\mathbb{Z})$. It would be nice if the converse were true, but it is false. It turns out that a weaker form of the converse is true, but we shall not state it here. Instead, we invite the interested reader to explore how this set of ideas leads to the definition of a Sobolev space.

REFERENCES

[1] Abel, N. H. (1828). Note sur le mémoire de Mr. L. Olivier No. 4. du second tome de ce journal, ayant pour titre "Remarques sur les séries infinies et leur convergence". *Journal für die Reine und Angewandte Mathematik*, 3(1): 79–82.

[2] Ash, J. M. (1997). Neither a worst convergent series nor a best divergent series exists. *Coll. Math. J*. 28(4): 296–297. doi.org/10.1080/07468342.1997.11973879

[3] Carothers, N. L. (2005). *A Short Course on Banach Space Theory*. Cambridge: Cambridge Univ. Press.

[4] Dini, U. (1867). Sulle serie a termini positivi. *Ann. Univ. Toscana*. 9:41–76.

[5] Du Bois-Reymond, P. (1873). Eine neue Theorie der Convergenz und Divergenz von Reihen mit positiven Gliedern. *Journal für die reine und angewandte Mathematik*, 76: 61–91. DOI: 10.1515/crll.1873.76.61

[6] Gowers, T., Barrow-Green, J., Leader, I., eds. (2008). *The Princeton Companion to Mathematics*. Princeton: Princeton Univ. Press.

[7] Kałuża, R. (1996). *The Life of Stefan Banach*. Boston: Birkhäuser.

[8] Knopp, K. (1928). *Theory and Application of Infinite Series*. Glasgow: Blackie & Son.

[9] Riesz, F., Sz.-Nagy, B. (1990). *Functional Analysis*. New York: Dover.

[10] Sokal, A. D. (2011). A really simple elementary proof of the uniform boundedness theorem. *Amer. Math. Monthly*. 118(5): 450–452. DOI: 10.4619/amer.math.monthly.118.05.450

[11] Tao, T. (2010). *An epsilon of room, I: Real analysis*, vol. 117 of *Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society.

Summary. We introduce a dual of the uniform boundedness principle that does not require completeness and gives an indirect means for testing the boundedness of a set. The dual principle, although known to analysts and despite its applications in establishing results such as the Hellinger-Toeplitz theorem, is often missing from elementary treatments of functional analysis. We give an example showing a connection between the dual principle and a question in the spirit of du Bois-Reymond regarding the boundary between convergence and divergence for sequences. This example is intended to illustrate why the statement of the principle is natural and clarify what the principle claims and what it does not.

EHSSAN KHANMOHAMMADI (MR Author ID: 916345) received his Ph.D. in Mathematics from The Pennsylvania State University. His research interests include spectral theory, harmonic analysis, and mathematical physics.

OMID KHANMOHAMADI (MR Author ID: 861820) has earned a Master's in Mechanical & Aerospace Engineering as well as Master's and Ph.D. in Applied & Computational Mathematics. His research has spanned a wide range of topics, including numerical integration on manifolds, spectral methods for PDEs and inverse problems, numerical analysis of nonnormal operators, and high performance computational software design.

Like a Sparrow, Small but Complete

FANG CHEN
Oxford College of
Emory University
Oxford, GA 30054
fchen2@emory.edu

Dedicated to the memory of Svetoslav Savchev, who inspired generations of students to be interested in mathematics in Bulgaria, Argentina and many other countries.

"What is it like to engage in pure mathematics research?" First and second year college students often ask this question. A formal answer may be readily given, but the best way to understand the answer is to explore it. At increasingly early points in their education, more and more undergraduate students are participating in basic research in social science, the natural sciences, and applied mathematics through programming tools and hands-on experiments. However, research in pure mathematics can be intimidating because students have been provided few opportunities to engage in substantial mathematical activities at an elementary level.

Such opportunities would allow students to experience the process of independent mathematical thinking. They would prompt one to ask, investigate, and solve nontrivial questions. As a result, students would develop interests, abilities, and confidence.

At the core of these opportunities are appropriate problems. These problems should contain typical features of mathematical investigations, such as conjectures and proofs, examples and counterexamples. Of equal importance, the problems should involve interesting ideas and useful techniques. Ideally, such problems should require minimum knowledge and be capable of staging and scaling to accommodate students with different levels of involvement and capabilities.

In this article, we propose such a problem. It illustrates a Chinese proverb describing something that is small but complete, "Little as it is, a sparrow has all its vital organs." The mathematical content is elementary and accessible to undergraduate students. The problem is interesting on its own, but its educational value is most pertinent, and we will structure the exposition accordingly.

Apart from exemplifying a few important features of a mathematical setting, the journey through this problem allows one to experience how actual mathematical work is done. Readers are invited to participate actively as we proceed.

From prehistoric man to the present

We start with an old folklore problem:

> 101 weights with integer masses have total mass 200. Prove that they can be divided into two groups of mass 100.

One solution uses a classic idea, which the world-famous mathematician Erdős found too minor to describe as classic. He believed the idea was known already to prehistoric men. Let the weights have positive integer masses $a_1, a_2, \dots, a_{101}$. Form the sums

$$S_1 = a_1,\ S_2 = a_1 + a_2,\ \dots,\ S_{101} = a_1 + a_2 + \cdots + a_{101} = 200.$$

Math. Mag. **95** (2022) 532–544. doi:10.1080/0025570X.2022.2125255 © Mathematical Association of America
MSC: Primary 97D50, Secondary 11A99

Look at their remainders upon division by 100. There are 101 sums $S_1, S_2, \dots, S_{101}$ and 100 different remainders modulo 100. It follows that at least two sums with different indices have the same remainder. Let S_i and S_j be such sums, with $i < j$. Then the difference $S_j - S_i = a_{i+1} + a_{i+2} + \cdots + a_j$ is divisible by 100. On the other hand, notice that $S_j - S_i$ contains at least one summand and at most 100. It follows that $S_j - S_i$ is an integer strictly between 0 and 200. Because this integer is a multiple of 100, the only possibility is $S_j - S_i = 100$. In other words, the masses $a_{i+1}, a_{i+2}, \dots, a_j$ add up to 100. Then all remaining masses also add up to 100, which solves the problem.

For convenience, we will refer interchangeably either to a collection of weights with integer masses or to a positive integer sequence. Sequences may have repeated terms (in fact they must have, in the cases of interest to us). The *length* of a (finite) sequence is the number of its terms. All sequences in what follows have positive integer terms, so we will sometimes call them just sequences for brevity. For an arbitrary positive integer n, we consider sequences with sum $2n$. If such a sequence can be divided into two parts with sum n in each part, we call it *separable*. Otherwise it is *inseparable*. Given the definitions, the prehistoric man's argument proves a more general fact:

> For an arbitrary positive integer n, every positive integer sequence with sum $2n$ and length $\ell \geq n + 1$ is separable.

What about shorter lengths? To avoid trivialities, let us assume that in the sequel the sequences considered are such that all their terms *do not exceed* n. This is reasonable since a term greater than n would make the sequence trivially inseparable. It is easy to see that shorter lengths no longer guarantee separability, at least not for *all* n. For example, the sequence α consisting of n twos is not separable if n is odd. In this case n would have to be a sum of 2's, and therefore even, which is false. The same example does not work for even n. The sequence α is easily separated into two parts with $n/2$ twos in each.

Let us find out more about the even case. We take specific values of n and try to find inseparable sequences with length as large as possible, by trial and error. The outcome is discouraging if we expect examples of length close to n. For $n = 100$, the longest inseparable sequence we managed to come up with has length 67: one term 2 and 66 terms 3. This sequence is inseparable because the sum of any part of the sequence has remainder 0 or 2 modulo 3, while 100 is 1 modulo 3. Many other even values are similar to $n = 100$, such as $n = 80$ and $n = 70$. The longest inseparable sequences we found have the same structure: all terms are 3s, except one term which is 1 or 2. The respective lengths are therefore roughly equal to $2n/3$. Worse yet, not even this much can be done with numbers like $n = 90$. The most we achieve is an inseparable sequence of length only 46: one term 1, one term 3, and 44 terms 4. You should check for yourself that the latter sequence is inseparable.

Of course, our failure does not prove that "long" examples do not exist. They may simply have a structure too complicated for us to notice. Nevertheless, we feel that the case of n even is more complicated than the introductory problem suggests, and we have to approach it consistently. It is natural to state a clear objective:

> Let $n \geq 3$ be a given integer. Determine the least integer ℓ_n such that each sequence with positive integer terms not exceeding n, sum $2n$ and length at least ℓ_n is separable.

In other words, *all* sequences with sum $2n$ and length at least ℓ_n must be separable, and *at least one* sequence with sum $2n$ and length $\ell_n - 1$ is inseparable. We state the problem for $n \geq 3$ because for $n = 1, 2$, all sequences with terms not exceeding n and sum $2n$ are trivially separable.

We already know that $\ell_n = n + 1$ for n odd. As for n even, examples of length close to n do not turn up. So we wonder: Can it be the case that *parity* makes such a difference? Can one have $\ell_n = n + 1$ for n odd and ℓ_n roughly equal to $2n/3$ for n even? If yes, this suggests nontrivial work. In similar settings, the difficulty of a question grows very fast as the sequence length decreases.

At this point, some readers might think of the so-called zero-sum problems in finite abelian groups (see Gao and Geroldinger [2] for a survey). Indeed, this is how the question under discussion emerged for us as we tried to translate the group setting into the language of positive integer sequences, hoping that such a point of view would be useful. While our hopes turned out to be ungrounded, the problem and its solution are interesting and substantial in their own right.

The question seems hard if considered directly, even with specific values of n and ℓ. A highly capable group of university students, among them former International Mathematical Olympiad medalists, tried to do the case $n = 70$. All they managed to come up with was a heavily case-based solution that revealed no general insight. The case $n = 60$ appeared as the last question on the individual part of the 12th Hungary–Israel Binational Mathematical Competition (2001). The solution presented by Gueron [3] proves the result for any n divisible by 6. It uses a notion called a *universal sequence*, which has also been called a *behaving sequence* [4] and by different names in other papers. This elegant notion, though without a standard name, has been used extensively by researchers in various fields. However, it is unfamiliar to most students. Indeed the proof given by Gueron [3] seems to be written only for those who are well versed in related techniques. In what follows, we will unravel the problem and make it readily accessible.

An elementary and systematic approach

We have stated the problem in the general case, with the prehistoric result being the first step. Let us search for a systematic approach leading to the complete solution. In what follows, n is an arbitrary positive integer unless declared otherwise.

The trivial algorithm We first describe a procedure for dividing an arbitrary collection of weights into two groups. Because it is quite simple, we call it the *trivial algorithm* (or just the *algorithm*, for brevity).

Let α be a collection of weights with integer masses not exceeding n of total mass $2n$. Start placing the weights on the pans of a two-pan balance, one at a time, in decreasing order of their masses. Each time the current weight w is to be placed on the pan containing the smaller mass (select one arbitrary if there is a tie), *provided that after placing w this pan has mass not exceeding n*. In other words, each pan is assumed to have capacity n which is not to be exceeded at any step.

Suppose the latter condition cannot be satisfied at a certain moment, that is, placing the current weight would overload the lighter pan. The first time this occurs, with a weight t, the algorithm stops and we say that it fails. If so, we call t the *critical weight* or the *critical term* of α. Observe that since all terms are at most n, the trivial algorithm cannot fail at the first two steps. Clearly, if all weights can be placed according to the rules, the procedure yields a partition of the system into two groups of mass n.

Before proceeding further, let us consider specific examples to make sure that we understand the trivial algorithm. We apply it to the sequences

$$\alpha_1 = 5, 5, 4, 3, 2, 1, \quad \alpha_2 = 6, 3, 3, 3, 3, 2, \quad \text{and} \quad \alpha_3 = 5, 3, 2, 2, 2, 2, 2, 2,$$

and we look at the outcomes. All three sequences have sum 20, which corresponds to $n = 10$. For α_1, the contents of the two pans are, consecutively:

$$(5, 0), \quad (5, 5), \quad (9, 5), \quad (9, 8), \quad (9, 10), \quad (10, 10).$$

The algorithm terminates successfully, yielding a separation of the system into two groups of mass 10. For α_2 we have:

$$(6, 0), \quad (6, 3), \quad (6, 6), \quad (9, 6), \quad (9, 9).$$

Now the trivial algorithm fails because the last weight 2 cannot be placed on either pan without exceeding its capacity 10. Thus, the algorithm does not yield a separation. No wonder, because α_2 is not separable. For α_3, the development is like this:

$$(5, 0), \quad (5, 3), \quad (5, 5), \quad (7, 5), \quad (7, 7), \quad (9, 7), \quad (9, 9).$$

Then the algorithm fails to give a separation for the same reason as with α_2, the last 2 cannot be placed on any pan. On the other hand, α_3 can clearly be separated into parts 5, 3, 2 and 2, 2, 2, 2, 2.

These are the three possible outcomes. If the trivial algorithm terminates for a sequence, then the sequence is separable and the procedure has given us an actual separation. But if the algorithm fails, this does not imply inseparability. The sequence may very well be separable, perhaps even with an obvious separation. Even if this is true, the trivial algorithm will, of course, not deviate from its rigid rules to produce a separation. Thus, the question about separability is resolved (positively) if the algorithm terminates, but nothing definitive can be said if it fails.

Here is a rough idea for how to prove that a certain sequence is separable. First, we apply the trivial algorithm and immediately conclude separability if the procedure terminates. If not, we focus on the reasons for the failure. It turns out that substantial information can be obtained by looking at those reasons, in which the critical term of the sequence plays a crucial role. Based on the conclusions, we then find a way to apply the trivial algorithm again to reach separability. Let us make the following observations, which result from analyzing the situation when the trivial algorithm fails.

Observation 1. *Let the trivial algorithm fail for a positive integer sequence α with terms not exceeding n, of length ℓ and sum $2n$, with a critical term t. Then:*

(i) *The sum of the terms after t in the decreasing arrangement of α does not exceed $t - 2$. In particular, t is never equal to* 1.

(ii) *α contains at least* 3 *terms each $\geq t$. In particular, $\ell \geq 3$.*

(iii) *Consider the quadratic function $g(x) = x^2 - (\ell + 3)x + (2n + 2)$. Then, t being a critical term implies $g(t) \geq 0$.*

Proof. Because t is a critical term, it cannot be placed on the lighter pan without overloading, but then it cannot be placed on the other pan either. Hence, the unused capacity of each pan is at most $t - 1$ at the step when the algorithm fails. So, the unused capacity of the whole balance is at most $2t - 2$. On the other hand, this unused capacity equals the sum of all terms not yet placed. These are t itself and the later terms, implying that the sum of the later terms does not exceed $(2t - 2) - t = t - 2$. In particular, t cannot be 1, or else the sum of all later terms would be at most $t - 2 = 1 - 2 < 0$ which is impossible. This proves (i).

Part (ii) is due to the assumption that all weights have mass at most n. The first two weights (the heaviest ones) are placed on different pans, and neither overloads its pan.

Therefore, neither is critical. Thus, the critical term t comes in later, and the first two are $\geq t$. The claim follows.

Part (iii) is the most substantial observation. By (i), the sum A of the terms after t satisfies $A \leq t-2$. Then the number of these terms is $\leq A$ (as each one is ≥ 1). The terms before t have sum $2n-t-A$; each one is $\geq t$, hence their number is $\leq (2n-t-A)/t$. Since there are no more than $(2n-t-A)/t$ terms before t and no more than A terms after it, the length ℓ satisfies

$$\ell \leq \frac{2n-t-A}{t}+1+A=\frac{2n}{t}+A(1-1/t).$$

Note that $1-(1/t)>0$ since $t>1$, as remarked in observation (i). So, by $A \leq t-2$ we obtain

$$\ell \leq \frac{2n}{t}+(t-2)(1-1/t)=\frac{2n+(t-2)(t-1)}{t},$$

which is equivalent to

$$t^2-(\ell+3)t+(2n+2) \geq 0.$$

■

Prehistoric man revisited and a general approach Now we go back to sequences of length $\geq n+1$ and apply the trivial algorithm.

Proposition 1. *The trivial algorithm terminates successfully for each positive integer sequence with sum $2n$ and length $\ell \geq n+1$.*

Proof. First, notice that in a positive integer sequence with sum $2n$ and length at least $n+1$, no term can exceed n. Now suppose, for a contradiction, that the trivial algorithm fails for one such sequence α. Denote the critical term by t, and consider the quadratic function $g(x)=x^2-(\ell+3)x+(2n+2)$. Compute the value of g at $x=2$:

$$g(2)=2^2-2(\ell+3)+(2n+2)=-2\ell+2n.$$

Since $\ell \geq n+1$, $g(2)<0$.

Now, g has a positive leading coefficient (equal to 1), so its graph is an upward-opening parabola. The condition $g(2)<0$ implies that g has two distinct real roots x_1 and x_2, with $x_1<2<x_2$. Next, the graph is symmetric with respect to the vertical line $x=(\ell+3)/2$, so that $g(x)=g(\ell+3-x)$ for all real x. In particular we have

$$g(\ell+1)=g(2)<0,$$

and since $\ell+1 \geq 2$, it follows that

$$x_1<2 \leq \ell+1<x_2.$$

On the other hand, by observation (iii), the critical term t satisfies $g(t) \geq 0$. Therefore, $t \leq x_1$ or $t \geq x_2$, which implies $t<2$ or $t>\ell+1$.

Since t is a positive integer, we infer that $t=1$ or $t \geq \ell+2$. However, $t=1$ is impossible by observation (i), and $t \geq \ell+2$ is also impossible. Otherwise, because α has at least three terms $\geq t$ (observation (ii)), its sum $2n$ would satisfy

$$2n \geq 3t \geq 3(\ell+2) \geq 3(n+3)>2n.$$

A contradiction is reached, and the proof is finished. ■

Formally speaking, we do not need Proposition 1. The prehistoric man's argument told us that length $\geq n+1$ guarantees separability. But we do need a consistent approach to the whole question of which $\ell \geq n+1$ seems to be the easiest step. There must be a proof for $\ell \geq n+1$ that generalizes naturally. Hopefully we now have such a proof; let us try to read out what it says.

The problem certainly obeys some internal logic of its own. Apparently it subdivides into cases, and for the simplest one $\ell \geq n+1$, the proof of Proposition 1 tells us that the logic goes as follows:

$$\begin{array}{c}\text{length condition}\\ (\ell \geq n+1)\end{array} \Rightarrow g(2) < 0 \Rightarrow t \leq 1 \text{ or } t \geq \ell+2 \Rightarrow \text{a contradiction} \qquad (1)$$

The key is $g(2) < 0$. Now we are looking at the second simplest case. We say "second" as we do not know if it is the only case left. So, if our thinking is to the point, the important feature here should be $g(3) < 0$. This leads to the scheme

$$\text{length condition (?)} \Rightarrow g(3) < 0 \Rightarrow t \leq ? \text{ or } t \geq ? \Rightarrow \text{end (?)} \qquad (2)$$

The above is only a guess, yet a plausible one. To make it precise, we need to replace the question marks in equation (2) by actual values. This is easy; one basically just reproduces the proof of Proposition 1, with 3 playing the role of 2. In particular, we will obtain a length condition, which is badly needed.

For the first ?, compute $g(3) = -3\ell + 2n + 2$. We see that $g(3) < 0$ if and only if $3\ell \geq 2n+3$, that is, $\ell \geq 2n/3 + 1$. This must be the restriction on the length needed for the even case. We are relieved to see that it agrees with our experiments. All inseparable sequences we managed to find have length just a bit less than $(2n/3) + 1$.

For the second and third ?, observe again that the graph of g is an upward-opening parabola. Hence, the condition $g(3) < 0$ implies that g has distinct real roots x_1 and x_2, and that $x_1 < 3 < x_2$. We again have that $g(x) = g(\ell + 3 - x)$ for all x, by symmetry with respect to the line $x = (\ell+3)/2$. Therefore, $g(\ell) = g(3) < 0$, and since $3 \leq (\ell+3)/2$ (this is equivalent to $\ell \geq 3$ which holds by observation (ii)), we have $x_1 < 3 \leq \ell < x_2$. On the other hand, $g(t) \geq 0$ by observation (iii). It follows that $t \leq 2$ or $t \geq \ell + 1$ (as t is an integer), and the second and the third ? are 2 and $\ell + 1$, respectively.

The last ? in equation (2) remains, and it will not be "a contradiction" as in equation (1). All we can say is that $t \geq \ell + 1$ cannot hold. Indeed, recalling that α has at least three terms $\geq t$, we would obtain that its sum $2n$ satisfies $2n \geq 3t \geq 3\ell + 3 \geq 2n + 6$, by the new length condition $3\ell \geq 2n + 3$. Thus, $t \geq \ell + 1$ is impossible. However, $t \leq 2$ cannot be rejected. Therefore "a contradiction" in equation (1) will become $t \leq 2$ in equation (2). Since t is a positive integer and $t \neq 1$, the latter just means $t = 2$.

We emphasize that $t = 2$ can actually occur. For instance, take $n = 10$ and the sequence

$$\alpha = 3, 3, 2, 2, 2, 2, 2, 2, 2$$

with sum $2n = 20$. The length 9 of α satisfies $\ell \geq (2n/3) + 1$. And the trivial algorithm fails with α, ending up unable to place the last 2. Thus $t = 2$.

In summary, the chain (2) has flesh and blood already:

$$\begin{array}{c}\text{length condition}\\ (\ell \geq 2n/3 + 1)\end{array} \Rightarrow g(3) < 0 \Rightarrow t \leq 2 \text{ or } t \geq \ell + 1 \Rightarrow t = 2 \qquad (3)$$

Notice that in order to obtain equation (3), we actually proved the following:

Lemma 1. *If the trivial algorithm fails for a positive integer sequence of terms not exceeding n, with sum $2n$ and length $\ell \geq (2n/3) + 1$, with a critical term t, then $t = 2$.*

Moreover, if the situation in Lemma 1 occurs, then $t = 2$ must be the last term in the decreasing arrangement of the sequence (as the terms after it must have sum $\leq 2 - 2 = 0$ by observation (i)). Therefore, the sequence has no terms 1.

The case *n* even and the missing idea Our guess is that length $\ell \geq (2n/3) + 1$ is likely to guarantee separability for even n. Let us start the proof.

Let n be even, and let α be a sequence of sum $2n$ and length $\ell \geq (2n/3) + 1$. We want to prove that it is separable. As already explained, the first thing to do is to apply the trivial algorithm to α. The task is complete if the algorithm terminates. If not, Lemma 1 implies that the critical term t is 2 and it is the last term in the decreasing arrangement of α. This means that when the algorithm fails, all terms except $t = 2$ are on the two pans, hence the sum is exactly $n - 1$ on each pan at the critical moment. This information is most welcome, and we seem to be on the right track. Yet we have to admit still feeling pretty nervous at this point, and for good reason. The condition that n is even has not been used so far. This is unsurprising since all we did was look at the trivial algorithm in which the parity of n is completely irrelevant.

On the other hand, the guess would be false without assuming n even—simply because $\ell \geq (2n/3) + 1$ is not enough to imply separability for *all* n. We know that the stronger inequality $\ell \geq n + 1$ is necessary for n odd. Therefore, we are still missing something crucial, a consideration using the fact that n is even. Of what nature could it be? How can one possibly continue?

Let us take another look at the failure of the algorithm for a specific sequence we already know. Take $n = 10$ again and the sequence

$$\alpha = 3, 3, 2, 2, 2, 2, 2, 2, 2$$

with sum $2n = 20$. Recall that the length condition $\ell \geq (2n/3) + 1$ holds. The trivial algorithm fails with α, exactly as explained above, as it is unable to place the last 2. The reason is that the only odd terms 3, 3 go to different pans at steps 1 and 2. Hence, each pan has an odd mass already after the first two steps, which is preserved from this point on. Consequently, no even mass is ever obtained on a pan, and in particular $n = 10$ cannot be obtained. But if having the two 3s on different pans is the reason for the failure, why not prevent it from happening? No, I do not mean by changing the algorithm, but by changing the sequence. What causes trouble are the two odd terms 3. Combine them together and imagine they form a new term equal to $3 + 3 = 6$; the point being that 6 is even. A new sequence is obtained,

$$\beta = 6, 2, 2, 2, 2, 2, 2, 2,$$

with even terms and with the same sum $2n = 20$. Clearly, if β is separable, then so is α. So let us try the trivial algorithm on β. It readily terminates; here is the development:

$$(6, 0), \quad (6, 2), \quad (6, 4), \quad (6, 6), \quad (8, 6), \quad (8, 8), \quad (10, 8), \quad (10, 10).$$

We simply forced the two odd terms 3 to go together, so that, at any moment, the sum of the weights on each pan is always even. It worked!

Essentially the same situation occurs in the general case when the trivial algorithm fails. At the critical moment, there is mass exactly $n - 1$ on each pan, which is odd since n is even, making it impossible to have room for the last 2. On the other hand, one can easily ensure even mass on each pan at all times. Just divide the odd terms into pairs and replace every pair by its sum, which is even. Then try the trivial algorithm for the new sequence. True, we still have to make sure that this always yields a separation. But it feels like the missing idea is already here. Let us go ahead with it.

Since α has sum $2n$, which is even, it contains an even number of odd terms. Let us form a new sequence β by leaving the even terms of α untouched, grouping all the odd terms into pairs (in an arbitrary way), and then replacing each pair with the sum of the two terms in it. Observe that β has at least one term 2, for instance the critical term $t = 2$ of α, which is also a term of β. In view of this even term 2 present in α, the length ℓ_β of β satisfies

$$\ell_\beta \geq 1 + (\ell - 1)/2 = (\ell + 1)/2 \geq n/3 + 1$$

(note that ℓ_β is a minimum when all terms of α except the 2 are odd, so that they are all paired up). We want to show that the trivial algorithm separates β, but if we hope to use Lemma 1 then the length condition $\ell_\beta \geq (n/3) + 1$ is not enough.

We need a technique: divide the terms of β by 2 to obtain another new sequence β' with the same length as β and sum n. We may do this since β has even terms. Notice that β' contains a 1, as β contains a 2. Define $n' = n/2$. This is an integer since n is even, the parity of n is used at last! Then β' has sum $2n'$. Now, the assumption $\ell \geq (2n/3) + 1$ implies that the analogous inequality holds for n' and the length $\ell_{\beta'}$ of β':

$$\ell_{\beta'} = \ell_\beta \geq \frac{\ell + 1}{2} \geq \frac{n}{3} + 1 = 2n'/3 + 1.$$

Let us check that terms of β' do not exceed n'. Suppose there is a term in β' greater than n'. Then there are two odd terms in α whose sum is greater than n, and thus at least $n + 2$ because n is even. Therefore, the remaining $\ell - 2$ terms have sum $\leq n - 2$. On the other hand, each one of them is at least 2 since α does not contain a 1, implying that the same sum is no smaller than

$$2(\ell - 2) \geq 2\left(\frac{2n}{3} - 1\right) = \frac{4n}{3} - 2.$$

This leads to the inequality $n - 2 \geq (4n/3) - 2$ which is false.

We are almost done. Suppose the trivial algorithm fails for β'. Then Lemma 1 applies, implying that the critical term of β' is $t' = 2$. However, this is impossible by observation (i) because β' contains a term 1. Therefore, the trivial algorithm separates β' and hence separates β. It follows that α is separable.

So our guess was right, we obtained what we expected.

Proposition 2. *Let n be an even positive integer. Each positive integer sequence of terms not exceeding n, with sum* $2n$ *and length* $\ell \geq (2n/3) + 1$*, is separable.*

Now we summarize our accomplishments. For a real number x, let $\lceil x \rceil$ denote the least integer greater than or equal to x. By Proposition 2, having length

$$\ell \geq \left\lceil \frac{2n}{3} \right\rceil + 1$$

ensures separability for n even. We believe that $\lceil 2n/3 \rceil + 1$ is actually the least possible such length, that is, $\ell_n = \lceil \frac{2n}{3} \rceil + 1$, for all even n. The latter will follow if there is at least one example of an inseparable sequence with length $\lceil 2n/3 \rceil$.

Several such examples are known to us, for certain even values of n. Recall that we found an inseparable sequence for $n = 100$, consisting of 66 terms 3 and one term 2. The same example generalizes to each n with remainder 1 divided by 3. Take the sequence consisting of $(2n-2)/3$ terms 3 and one term 2. Its sum is $2n$ and its length is $(2n+1)/3 = \lceil 2n/3 \rceil$. Furthermore, the sequence is inseparable. A similar example exists for each n with remainder 2 divided by 3: the sequence consisting of $(2n-1)/3$ terms 3 and one term 1. Summarizing the results so far, we have obtained

$$\ell_n = \left\lceil \frac{2n}{3} \right\rceil + 1 \quad \text{for all even } n \text{ that are not divisible by 3.}$$

The last case There is a single case remaining: n even and divisible by 3, that is, n is divisible by 6. All we need is an example of an inseparable sequence with length $\lceil 2n/3 \rceil$, and then we will be completely done.

Alas, such an example does not exist. Remember our attempts with $n = 90$, where we only found an inseparable sequence with length 46, roughly half of 90. The same happens with other values of n divisible by 6. One starts suspecting that ℓ_n is much less than $\lceil 2n/3 \rceil + 1$ for all multiples of 6.

This indeed turns out to be the case, meaning that the problem is even more stubborn than we thought. Moreover, the case n a multiple of 6 takes longer to deal with than the one we solved, n even and not divisible by 3. Fortunately, no new insight is needed.

Let us go back to the scheme analogous to chain (2):

$$\text{length condition (?)} \Rightarrow g(?) < 0 \Rightarrow t \leq ? \text{ or } t \geq ? \Rightarrow \text{end (?)} \tag{4}$$

There is no doubt that $g(4)$ will be the next relevant value. Computation gives $g(4) = -4\ell + 2n + 6$, and the inequality $g(4) < 0$ yields the new length condition $4\ell \geq 2n + 7$, or $\ell \geq (n/2) + 2$. Analyzing the parabola g and noticing that $g(4) = g(\ell - 1) < 0$ ($\ell \geq 5$ is assumed here) yields that $g(t) \geq 0$ implies $t \leq 3$ or $t \geq \ell$. Like before, the second alternative is impossible. This is left as an exercise.

Hence, the first alternative $t \leq 3$ remains, that is, $t = 2$ or $t = 3$. Notice that $\ell \geq 5$ holds by the length condition $\ell \geq n/2 + 2$ if $n \geq 6$. For small values of n: $n \leq 5$, by using observation (ii), one can show directly that $t = 2$. Therefore, our new chain takes the following form:

$$\begin{array}{c}\text{length condition}\\ (\ell \geq n/2 + 2)\end{array} \Rightarrow g(4) < 0 \Rightarrow t \leq 3 \text{ or } t \geq \ell \Rightarrow t = 2 \text{ or } t = 3 \tag{5}$$

We have also proven the following analogous version of Lemma 1:

Lemma 2. *If the trivial algorithm fails for a positive integer sequence of terms not exceeding n, with sum $2n$ and length $\ell \geq (n/2) + 2$ (that is $4\ell \geq 2n + 8$), with a critical term t, then $t = 2$ or $t = 3$.*

We believe that $\ell \geq (n/2) + 2$ is likely to guarantee separability for n divisible by 6 and that $\ell = (n/2) + 2$ is the least possible such length. The latter will follow if there is at least one example of an inseparable sequence with length $(n/2) + 1$. Here is such an example: a sequence consisting of $(n/2) - 2$ ones, one $n/2$ and two terms of $(n/2) + 1$. The problem will be solved if we manage to show the following proposition. The proof is taken from Chen [1], and we include it here for completeness.

Proposition 3. *Let n be a positive integer divisible by 6, and let α be a sequence with positive integer terms not exceeding n and sum $2n$. Suppose that the length ℓ of α satisfies $\ell \geq (n/2) + 2$ (that is $4\ell \geq 2n + 8$). Then α is separable.*

Proof. We proceed as in Proposition 2. The first step is to try the algorithm on α. The task is complete if it succeeds, so suppose the algorithm fails. By Lemma 2, the critical term t of α is $t = 2$ or $t = 3$.

Now we divide the argument into two cases:

(a) α has at least two 2s;

(b) α has at most one 2 and at least two 3s.

One of them must hold. Indeed, if there is at most one 2 (i.e., (a) fails), then look at the critical term t. For $t = 2$, there are no 1s in α, and there is at most one 2 by assumption. If there were at most one 3, the sum of α would be no smaller than

$$2 + 3 + 4(\ell - 2) = 4\ell - 3 \geq 2n + 5.$$

Similarly, for $t = 3$ there are no 2s in α, and at most one 1 (by observation (i)). If there is at most one 3, then

$$2n \geq 1 + 3 + 4(\ell - 2) = 4\ell - 4 \geq 2n + 4.$$

(a) α has at least two 2s. Let us pair up the odd terms in α and proceed as in the proof of Proposition 2 to obtain β'. Write $n' = n/2$. This is an integer since n is even. Then β' has sum $2n'$. Because of the two 2s in α, β' contains at least two 1s, and the length $\ell_{\beta'}$ of β' satisfies the inequality

$$\ell_{\beta'} \geq 2 + \frac{\ell - 2}{2} = \frac{\ell + 2}{2}.$$

It follows that

$$4\ell_{\beta'} \geq 2\ell + 4 \geq (n + 4) + 4 = n + 8 = 2n' + 8.$$

Finally, each term of β' is $\leq n' = n/2$ since every two terms of α add up to a sum not exceeding n. Otherwise, if there are two terms of α whose sum is greater than n, then the remaining $\ell - 2$ terms of α will have sum less than n. On the other hand, we have $t = 2$ in case (a) ($t = 3$ would imply no 2s), so that 1s are not to be found in α. It turns out that $2(\ell - 2) < n$ which contradicts $4\ell \geq 2n + 8$.

It remains to show that the trivial algorithm works for β'. Suppose not, and let t' be the critical term. We saw that $4\ell_{\beta'} \geq 2n' + 8$, so that Lemma 2 applies. Hence, $t' = 2$ or $t' = 3$. However, both are impossible since β' has two 1s.

(b) α has at most one 2 and at least two 3s. The sum $2n$ of α is divisible by 3 because n is divisible by 3. Thus, the terms not divisible by 3 can be partitioned into groups of size 2 or 3, with the sum of each group a multiple of 3. Replace the terms in each group by their sum, then divide by 3 the terms of the sequence obtained. Set $n' = n/3$, which is an integer. The result is a positive integer sequence α' with sum $2n' = 2n/3$.

Because of the two 3s in α, there are at least two 1s in α', and the length $\ell_{\alpha'}$ of α' satisfies the inequality

$$\ell_{\alpha'} \geq 2 + \frac{\ell - 2}{3} = \frac{\ell + 4}{3}.$$

Hence,

$$4\ell_{\alpha'} \geq \frac{4\ell + 16}{3} \geq \frac{2n + 24}{3} = 2n' + 8.$$

Finally, no terms of α' exceeds $n' = n/3$. This is left as an exercise.

Now the trivial algorithm works for α'. For suppose not. Then since $4\ell_{\alpha'} \geq 2n' + 8$, Lemma 2 applies, implying that the critical term t' of α' would be 2 or 3. Both are impossible since α' has two 1s. This completes (b) and the main proof. ■

In summary, we have determined ℓ_n for all values of $n \geq 3$: $\ell_n = n + 1$ if n is odd, $\ell_n = \lceil 2n/3 \rceil + 1$ for all even n that are not divisible by 3, and $\ell_n = n/2 + 2$ for all n divisible by 6. This result is a complete response to the objective stated in the second section.

A generalization

Probably the first thought of extending the question is to consider a sequence with sum kn and dividing it into k parts of equal sum. Let us state a definition. A positive integer sequence with terms not exceeding n and sum kn is called *k-separable* if it can be divided into k parts each with sum n. Otherwise it is called *k-inseparable*. Here is the objective:

> Let $k \geq 2$ be an integer. For an arbitrary integer $n \geq 3$, determine the least integer $\ell_k(n)$ such that each sequence with positive integer terms not exceeding n, sum kn, and length at least $\ell_k(n)$, is k-separable.

Let us postpone any discussion of whether such a generalization is meaningful. As an introduction to research, it is a good exercise with interesting twists. The reader is invited to apply the ideas from the $k = 2$ case to the problem for $k = 3, 4$, and beyond. Details are contained in Chen [1]. Here we include some hints to aid one's investigation.

The trivial algorithm needs obvious modification. Like before, we analyze the situation when the algorithm fails, and we obtain information analogous to Observation 1. In particular, we derive the function

$$g_k(x) = (k-1)x^2 - (\ell + 2k - 1)x + (kn + k)$$

and conclude that t being a critical term implies $g_k(t) \geq 0$.

A first length condition comes from $g_k(2) < 0$: $\ell > \frac{k(n+1)}{2} - 1$. Let $\lfloor x \rfloor$ denote the greatest integer less than or equal to x. Using the same argument as in the proof of Proposition 1, we can show that

$$\ell \geq \left\lfloor \frac{k(n+1)}{2} \right\rfloor$$

guarantees k-separability. If n is odd, a k-inseparable sequence (due to parity) of length $\frac{k(n+1)}{2} - 1$ is not hard to find. The sequence consisting of $\frac{k(n-1)+2}{2}$ terms of 2 and $k - 2$ terms of 1 is an example. It follows that

$$\ell_k(n) = \left\lfloor \frac{k(n+1)}{2} \right\rfloor$$

for any odd n and $k \geq 2$.

As expected, the picture for n even is a little more complicated. We get a second length condition from $g_k(3) < 0$: $\ell > \frac{k(n+4)}{3} - 2$. Is this close to $\ell_k(n)$ for even n?

When $k = 3$, we see that

$$\frac{k(n+4)}{3} - 2 = n + 2,$$

and we can prove that length $\ell > n + 2$ implies 3-separability for even n. However, when looking for a 3-inseparable sequence of length $n + 2$, the only one that turned up is the sequence consisting of $n-1$ terms of 3 and three terms of 1. Moreover, it is 3-inseparable only when n has a remainder of 2 divided by 3, but 3-separable otherwise (if n has a remainder of 0 or 1 divided by 3). These results indicate that as in the case $k=2$, the remainder of n divided by 3 also seems to make a difference. The question is whether the difference is as small as a constant or as large as a different coefficient of n (as in the case of $k=2$).

In order to gain an intuition for a plausible guess, we look for more examples and think about the reasons for such sequences being inseparable. It is not hard to find 3-inseparable sequences of length $n+1$. Here is one for any even n: the sequence consisting of three terms of $n/2+1$, one $n/2$ and $n-3$ terms of 1. It is 3-inseparable because there are too many large terms that are at least $n/2$. This example can be generalized for any $k \geq 2$ and even n. Indeed, due to capacity, there can be at most k terms larger than $n/2$. Hence, the sequence with k terms of $(n/2+1)$, one $n/2$ and $\frac{n}{2}(k-1)-k$ terms of 1 is k-inseparable. Its length is $\frac{n}{2}(k-1)+1$.

Observe that

$$\frac{n}{2}(k-1)+1 > \frac{k(n+4)}{3}-2$$

is equivalent to $(n-8)(k-3) > 6$. If we assume that n is large enough, then the inequality holds when $k \geq 4$. This suggests that for $k \geq 4$ and even n, we have that $\ell > (k(n+4)/3)-2$ is not enough to guarantee k-separability. We suspect that

$$\ell \geq \frac{n}{2}(k-1)+2$$

is sufficient, which turns out to be almost the case.

Activities such as constructing examples, formulating guesses, and confirming with arguments are typical for mathematical work. After several rounds of conjecture and proof—the actual path varies with individuals—there is little doubt that a persistent mind will reach the right conclusions. Working out all the details might require new considerations, such as estimating the number of terms with small values, rejecting large values of the critical term, and taking care of special cases. Some of these involve standard techniques known to researchers in additive number theory. For a beginner, thinking through such problems provides a chance to get acquainted with such matters in a natural way, and to develop useful skills for analyzing similar situations.

Moving forward

Like a sparrow, the problem presented in this article is small but complete, and with some potential. Working through the problem, one learns to ask questions, experiment, come up with guesses, and tell the difference between trivial and essential. One also practices evaluating conditions, constructing examples, applying previously worked approaches, and inventing new techniques to overcome obstacles.

Once the problem is solved in full generality, for all $k \geq 2$, typically students are satisfied. An experienced mind knows that it is worth asking more questions to further increase understanding. Indeed, it is quite common to see the solutions without fully grasping the essence of the matter. Additional questions may also lead to other interesting investigations. We conclude with some comments to start the reader's thinking.

- For n even, why is it that the answer for $k \geq 4$ is simpler than the one for $k=3$, and that the case $k=2$ turns out to be the most intriguing? One might see evidence of this in the proofs, but that could be approach-dependent. Are there intrinsic reasons for $k \geq 4$ to be easier? One way to understand it is to investigate examples with critical lengths.

- Independently, one might want to construct diverse examples anyway. Apart from clarifying the picture, it might raise new questions. For instance, about the two reasons for being inseparable, namely not enough odd terms and too many large terms, if we relax these conditions, can the least length be significantly shorter? More precisely, when n is odd, consider sequences with at least k odd terms. Can a shorter length guarantee k-separability? If n is even and $k \geq 4$, can the bound $\frac{n}{2}(k-1)+2$ be improved if we limit the number of large terms?
- We may raise another kind of question. It might not be a surprise that the "generalization" considered turned out not to be more complicated than $k = 2$. Perhaps looking at $k = 2$ from a different perspective will bring to light more interesting problems. For example, for a fixed n, consider positive integer sequences with sum kn (where k is an integer less than n), such that no proper nonempty subsequences has a sum divisible by n. As a related observation, there are also the "inverse questions," namely characterizing inseparable sequences. For example, with $k = 2$, describe all inseparable sequences, say of length $\geq n/2 + 2$.

  Another generalization is to form a sum other than n. We consider positive integer sequences with sum S, where the objective is to represent a certain T between 1 and S as a subsequence sum. Much of what holds for our setting ($S = 2n$ and $T = n$) also holds for the general case. But there are obvious complications, such as restrictions on the sequences to avoid trivialities. It seems that these variants can be solved without essential new ideas.

Acknowledgments The author wishes to thank Shuxuan (Annie) Luo, a first-year undergraduate student at Oxford College of Emory University, for reading part of the manuscript as a participant in the Summer Oxford Research Scholars Program in 2020.

REFERENCES

[1] Chen, F. (2021). From a Hungary-Israel contest problem. (2021). *Integers*. 21(#A44): 13. http://math.colgate.edu/̃integers/v44/v44.pdf

[2] Gao, W. D., Geroldinger, A. (2006). Zero-sum problems in finite abelian groups: a survey. *Expo. Math.* 24(4): 337–369. doi.org/10.1016/j.exmath.2006.07.002

[3] Gueron, S. (2004). *Hungary-Israel Mathematics Competition: The First Twelve Years.* Bruce, ACT: Australian Math. Trust Publishing.

[4] Savchev, S., Chen, F. (2007). Long zero-free sequences in finite cyclic groups. *Discrete Math.* 307(22): 2671–2679. doi.org/10.1016/j.disc.2007.03.049

Summary. By investigating an interesting number theory problem and its generalization, a reader will experience the process of research in pure mathematics. The content is completely elementary and accessible to any university students. We hope that the article will serve as an invitation to mathematical research.

FANG CHEN has a B.A. from Bryn Mawr College and a Ph.D. in mathematics from Yale University. She is an associate professor at Oxford College of Emory University.

Counting Islands in Nurikabe

JACOB A. BOSWELL
Missouri Southern State University
Joplin, MO 64801
boswell-j@mssu.edu

JACOB N. CLARK
Missouri Southern State University
Joplin, MO 64801
jnc@jnclark.org

CHIP CURTIS
Missouri Southern State University
Joplin, MO 64801
curtis-c@mssu.edu

Nurikabe (ぬりかべ) is a class of puzzles constructed over a grid, with some of the tiles in the grid containing positive integers. Each cell in the grid represents either water or land. The goal is to shade cells in the grid in such a way that, with the shaded cells representing water and the unshaded cells representing land,

1. each cell with a number is one of a set, called an *island*, of exactly that many contiguous cells (where two cells are adjacent if they share an edge);
2. the water cells are all contiguous; and
3. there are no two-by-two blocks of water cells (*pools*).

The name Nurikabe is shared with that of a spirit of Japanese folklore. A Nurikabe is a spirit that takes the form of an invisible wall which blocks the passage of travelers at night [**2**, pp. 140–141]. The connected winding water tiles evoke this mysterious creature, blocking the movement of our island tiles.

The puzzle's first documented appearance found by the authors is a submission to a collaborative puzzle corner (オモパ) in issue 33 of the Japanese puzzle magazine *Puzzle Communication* (パズル通信ニコリ) in 1991 [**4**–**6**]. The puzzle appears to have caught on relatively quickly, and appeared in the World Puzzle Championships in 1998 [**1**,**4**].

The rules of Nurikabe in action

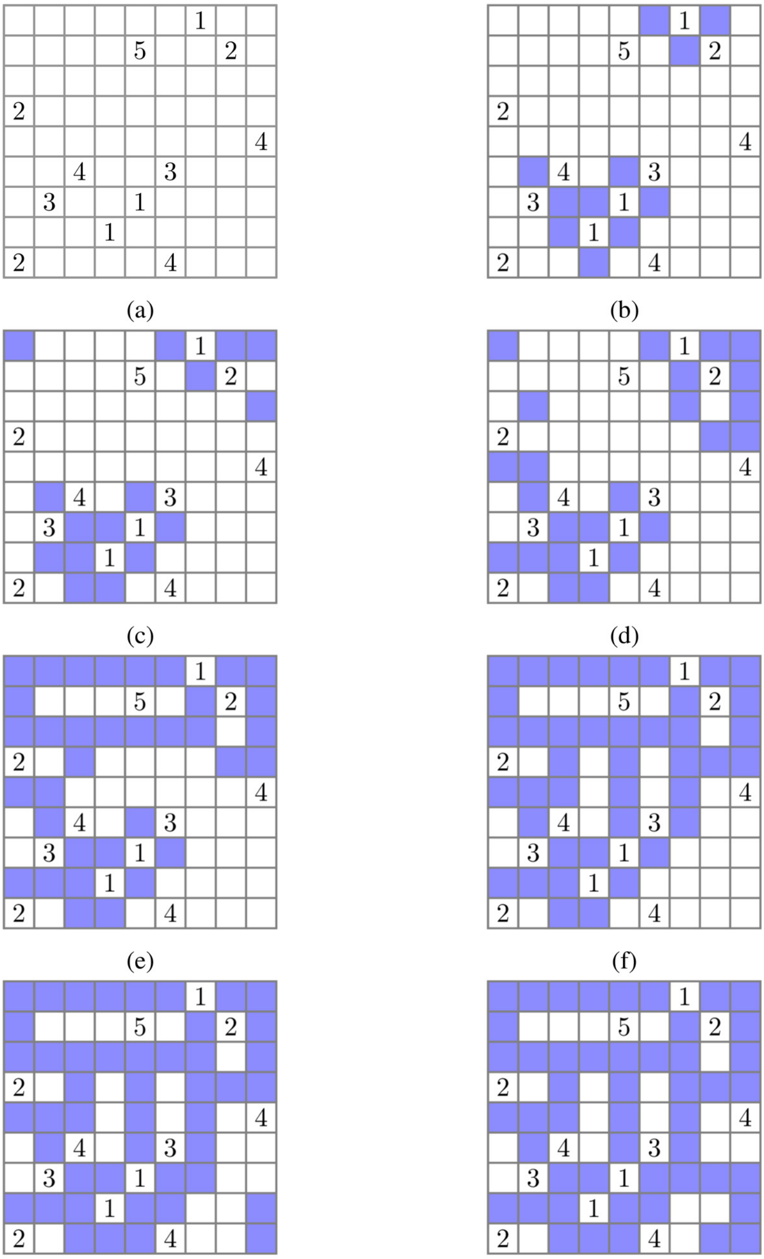
To demonstrate the rules outlined above, as well as solution strategies, we solve the 9×9 puzzle presented in Figure 1a.* In what follows, we use (row, column) numbering to label the cells, starting from the upper-left corner.

Four basic principles, which follow from the rules of the game, get us started.

- The *one-cell principle* says that each cell adjacent to a one-tile island must be water.
- Suppose there is a two-cell island for which it has already been determined that the second cell must occur in one of two adjacent directions. The *two-cell principle* says that the cell diagonally-adjacent to the 2 and between the possible cells must then be water since otherwise we would have adjacent islands (which would then not be distinct islands).

Math. Mag. **95** (2022) 545–553. doi:10.1080/0025570X.2022.2127306 © Mathematical Association of America
MSC: Primary 05B99, Secondary 00A08

*Note that the online version of this article has color diagrams.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 1 The process of solving a Nurikabe puzzle.

- The *diagonal principle* says that if two numbers are in diagonally-adjacent cells, then the other two cells of the 2×2 block they form must be water, again to avoid adjacent islands.
- The *reachability principle* says that if there is a cell that cannot be reached by any island, then it must be water.

The one-cell principle and the diagonal principle give Figure 1b; the two-cell principle and the reachability principle then give Figure 1c. Pool avoidance requires that the two-cell island based in the (9, 1)-cell must continue into the (9, 2)-cell. This, in turn, determines the location of the three-cell island just above it. Applying the two-cell principle to the 2 in the (4, 1)-spot and placing water in the (2, 9)-spot to ensure connected water determines the position of the nearby two-cell island. Ensuring the water remains connected forces the (4, 9)-spot to be water, and gives Figure 1d.

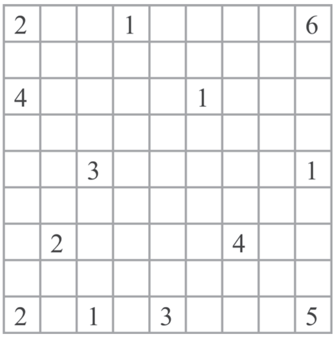
We now claim that the five-cell island must be horizontal. To see this, note that to avoid pools, at least one of the four cells in the upper-left 2×2 block of the puzzle must be land, as must one of the cells (2, 6) and (3, 6). The only way to simultaneously achieve both of these is for the five-cell island to reach from (2, 2) to (2, 6). Filling in water around the five-cell island, and also shading in cell (3, 1) to connect the water, determines the orientation of the two-cell island based in cell (4, 1). Shading around this island gives us Figure 1e.

Avoiding pools in rows three and four, while ensuring that the water in cell (5, 2) connects to other water in the puzzle determines the location of the central three-cell and four-cell islands. The result is Figure 1f.

Connecting the water in the lower left portion of the puzzle to other water in the puzzle requires water in cells (9, 5) and (8, 6). Pool avoidance then requires that the four-cell island based at (9, 6) reaches cell (8, 7) and that the four-cell island based at (5, 9) reaches cell (5, 8). This in turn implies that cell (7, 7) must be water to maintain water connectivity. Now no island can reach cells (8, 9) or (9, 9), thus requiring water in those spots. This gives Figure 1g.

Consideration of the possible locations for the four-cell islands in the lower-right portion of the puzzle gives the completed solution, found in Figure 1h.

We invite the reader to try solving the Nurikabe puzzle in Figure 2.



2			1					6
4					1			
		3						1
	2					4		
2		1		3				5

Figure 2 A Nurikabe puzzle for the reader to solve.

Nurikabe with one-tile islands

A natural question is how many Nurikabe puzzles exist for a given grid size. Collaboration with undergraduate students led to the calculation of the number of pictures satisfying the rules of Nurikabe for small grid sizes. No obvious formula emerged. One explanation for the complexity involved is that one can view the islands and water in a Nurikabe puzzle as polyominoes. Given that, as far as we can tell, no formula has been established for the number of polyominoes of a given size, it seems likely that it will be difficult to find a formula for the number of Nurikabe puzzles.

Accordingly, we now consider the much narrower problem of computing, as a function of n, the minimum and maximum number of islands that can occur in a Nurikabe puzzle of size $n \times n$, containing only one-cell islands. We note that from the point of view of someone solving the puzzle, such puzzles leave almost nothing to do, since every cell without a 1 in the starting puzzle must be water. In considering this problem, however, we establish some techniques that we hope can be generalized.

In this section, we establish a formula for the minimum as well as an upper bound for the number of islands; in the next section, we show that this upper bound is sharp.

Theorem 1. *The minimum number of islands present in an $n \times n$ Nurikabe puzzle having only one-tile islands is $\left\lfloor \frac{n}{2} \right\rfloor^2$.*

Proof. We proceed in cases.

For n even, the $n \times n$ grid is made up of $(\frac{n}{2})^2$ blocks of size 2×2. Each of these blocks must contain an island in order to avoid having a pool. The minimum number of islands is therefore at least $(\frac{n}{2})^2$. For equality, we note that placing islands in the upper left corner of each of these 2×2 blocks, that is, in the (i, j)-positions where i and j are odd, achieves a Nurikabe puzzle.

For $n = 1$, one water tile is a Nurikabe puzzle having $0 = \left\lfloor \frac{1}{2} \right\rfloor^2$ islands.

For n odd, $n > 1$, the $n \times n$ grid contains the $(n-1) \times (n-1)$ grid obtained by deleting the first row and column. By the argument given in the even case, this must contain at least $(\frac{n-1}{2})^2 = \left\lfloor \frac{n}{2} \right\rfloor^2$ islands. To see equality, we note that placing islands in the (i, j)-positions where i and j are even, achieves a Nurikabe puzzle. ■

Now we turn our attention to the maximum number of islands a Nurikabe puzzle having only one-tile islands may contain. When placing as many islands as possible into a puzzle, the rule against pools plays a smaller role, but, intuitively, having too many islands would violate the continuity of the water. If we view the water cells of a Nurikabe puzzle as a graph, we can apply results from graph theory.

We let the water tiles be the vertices of a graph. Two vertices are connected by an edge if and only if they correspond to adjacent water tiles. The rule guaranteeing the continuity of the water ensures that, for a Nurikabe puzzle, this is a connected graph. As this graph is planar, we may apply Euler's formula [3, p. 65] which, after removing the infinite, external face, gives $v - e + f = 1$, where v, e, and f are the number of vertices, edges, and faces, respectively.

In the setting of Nurikabe puzzles having only one-tile islands, we see that $v = n^2 - k$, where k represents the number of islands. If the entire grid contained water tiles, then the number of edges would be $2n(n-1)$. Each corner island removes two of these connections, each non-corner border island removes three connections, and each interior island removes four connections. This changes Euler's formula into $(n^2 - k) - (2n(n-1) - 2\#(\text{corners}) - 3\#(\text{non-corner borders}) - 4\#(\text{interior islands})) + f = 1$. This can be rewritten as

$$k = \frac{1}{3}((n-1)^2 + \#(\text{corner}) + \#(\text{border}) - f).$$

Note that in the previous, "border" includes corner islands as well. To understand what faces would be translated into in this context, we give a definition.

Definition. A set of interior island tiles all connected to one another by a set of diagonal adjacencies is called an *archipelago* if none of its member tiles is diagonally connected to an island tile outside the set.

Faces correspond to loops of water tiles. The rule against pools eliminates the possibility of a face that does not have any islands interior to it. We see then that each face corresponds to the unique archipelago it contains. Incorporating this with what we had previously gives

$$k = \frac{(n-1)^2 + \#(\text{corner}) + \#(\text{border}) - \#(\text{archipelagos})}{3}. \tag{1}$$

The number of corner islands is clearly at most four. We seek an upper bound for the difference between the number of border islands and the number of archipelagos in terms of n, in order to have an upper bound for the number of islands, k.

First, we consider the case when n is odd. Since two island tiles on the border cannot be adjacent to one another in a Nurikabe puzzle having only one-tile islands, the number of border islands is at most half of the number of border tiles. Thus, the number of border island tiles is at most $2n-2$. Note, however, that if there are $2n-2$ island tiles on the border then, because of the rule about all of the water being connected, they must be in the $(1, i)$, $(i, 1)$ (n, i), and (i, n)-positions, where i is odd. Adjacency to these island tiles, along with the requirement of the connectedness of the water, then implies that all tiles in the second row, second column, $(n-1)$st row, and $(n-1)$st column are water. If this is not the complete puzzle, i.e., if $n > 3$, any other island tiles would form into at least one archipelago. This shows that $\#(\text{border}) - \#(\text{archipelagos}) \leq 2n-1$. Substituting this into equation (1) gives

$$k \leq \frac{n^2+2}{3} \text{ for } n \text{ odd}, n > 3. \tag{2}$$

If n is even, the number of border island tiles is at most $2n-4$. To see this, note that because island tiles cannot be adjacent and because water tiles cannot be isolated, the three tiles making up one corner of the border can contain at most one island tile. Thus, the twelve tiles making up the four corners of the border can contain at most four island tiles. The remainder of the border is comprised of four segments, each containing $n-4$ tiles. These can each be at most half land, giving at most $2n-8$ island tiles on these border segments. Putting that together with the corners gives at most $2n-4$ island tiles on the border.

The number of archipelagos is at least zero. Substituting this into equation (1) gives

$$k \leq \frac{n^2+1}{3} \text{ for } n \text{ even}. \tag{3}$$

Combining equation (2) and equation (3), we define the following function on the natural numbers:

$$E(n) = \begin{cases} 4 & n = 3 \\ \lfloor \frac{n^2+2}{3} \rfloor & n \text{ is odd}, n \neq 3 \\ \lfloor \frac{n^2+1}{3} \rfloor & n \text{ is even}, \end{cases}$$

$$= \begin{cases} 4 & n = 3 \\ \frac{n^2}{3} & 3 \mid n, n \neq 3 \\ \frac{n^2+2}{3} & 3 \nmid n, n \text{ odd} \\ \frac{n^2-1}{3} & 3 \nmid n, n \text{ even}. \end{cases}$$

Summarizing all of the results from the water graph we have:

Theorem 2. *The number of islands in an $n \times n$ Nurikabe puzzle having only one-tile islands is at most $E(n)$.*

In the next section, we provide an algorithm to reach this upper bound, $E(n)$.

Achieving the upper bound

To show that we can achieve the upper bound $E(n)$ for all $n \times n$ Nurikabe, we first verify it for $1 \leq n \leq 10$ and for $n = 11, 13$, and 15. The appendix illustrates an optimal puzzle (one with $E(n)$ island tiles) for all of these sizes except $n = 11$, which is shown in the center of Figure 3. We then show that we can extend optimal $n \times n$ puzzles to optimal puzzles of size $(n + 6) \times (n + 6)$, letting us obtain all larger square puzzles.

To extend an optimal $n \times n$ puzzle to an optimal puzzle of size $(n + 6) \times (n + 6)$, we surround the $n \times n$ puzzle by three rings of cells, each with a width of one tile. We will show that, if the border tiles of the $n \times n$ puzzle appear in a prescribed pattern, we can choose the location of island tiles in the rings in such a way that the resulting $(n + 6) \times (n + 6)$ puzzle is an optimal Nurikabe puzzle whose border follows the same pattern. Therefore, the process can be iterated. We consider the cases of n odd and n even separately.

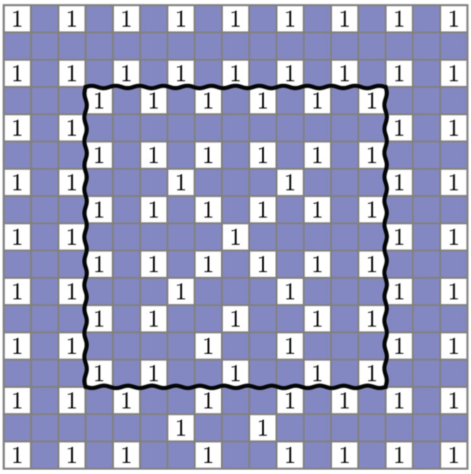


Figure 3 Odd border extension, $n = 11$ to $n = 17$.

First, consider the case of n odd. In the $n \times n$ puzzle, we assume that all four corners are islands and that the west, north, and east borders have alternating land and water tiles. We assume that the southern border has a seven-tile segment LWWLWWL (where L represents land and W represents water) located at least two tiles from each corner but is otherwise alternating.

The rings to be added are formed as follows: On the west, north, and east, the outermost and innermost rings have alternating land and water tiles, with an island in each corner. On these sides, the middle ring is all water. On the south side, the outer and inner rings match, each containing a seven-tile segment LWWLWWL as before, but shifted one cell to the west of that corresponding seven-tile segment in the $n \times n$ puzzle. As before, the rest of the south side of those rings alternates land and water. The inner ring has exactly two land tiles, the first placed one cell to the left of the middle L in the seven-tile segment, and the other placed one cell to the left of the right-most L in the seven-tile segment. This procedure is illustrated in Figure 3.

The procedure for the case of n even is similar. In the $n \times n$ puzzle, we again assume all four corner tiles are islands. We further assume that each border has $n/2$ land tiles, which means the land and water tiles alternate, except for a four-tile segment LWWL on each side.

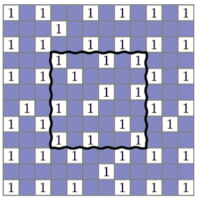


Figure 4 Even border extension, $n = 6$ to $n = 12$.

In adding rows and columns, as before, the innermost matches the outermost. As with the border of the $n \times n$ puzzle, these contain on each side a four-tile segment LWWL, but are otherwise alternating.

The four-tile segments are placed with an offset of one cell from the corresponding segment on the $n \times n$ border. The middle ring has a total of four tiles, one on each side. These should be placed to align with the first water tile in the clockwise orientation of each LWWL segment of the innermost ring. See Figure 4.

We claim that in both the odd and even cases, the resulting puzzles are both legal and optimal. To see this, we note that this process introduces no pools and that placement of the two-tile water segments and the placement of the islands in the middle rows and columns ensures that all of the new water tiles connect to the water tiles of the original puzzle. Since the water tiles of the original puzzle are connected, this ensures that we have created a Nurikabe puzzle.

The number of new island tiles for n odd is

$$(2(n+6)-3)+2+(2(n+2)-3) = 4n+12$$

and for n even is

$$(2(n+6)-4)+4+(2(n+2)-4) = 4n+12.$$

The number of island tiles in the new puzzle is therefore

$$\begin{aligned} E(n)+4n+12 &= \begin{cases} \frac{n^2}{3}+4n+12 & 3 \mid n \\ \frac{n^2+2}{3}+4n+12 & 3 \nmid n,\, n \text{ odd} \\ \frac{n^2-1}{3}+4n+12 & 3 \nmid n,\, n \text{ even} \end{cases} \\ &= \begin{cases} \frac{(n+6)^2}{3} & 3 \mid n \\ \frac{(n+6)^2+2}{3} & 3 \nmid n,\, n \text{ odd} \\ \frac{(n+6)^2-1}{3} & 3 \nmid n,\, n \text{ even} \end{cases} \\ &= E(n+6). \end{aligned}$$

Thus, we can conclude with a theorem.

Theorem 3 (Main Result). *There are at most $E(n)$ islands in a $n \times n$ Nurikabe puzzle containing only one-tile islands, and this bound is sharp. Namely, there exists at least one $n \times n$ Nurikabe puzzle for which there are exactly $E(n)$ islands.*

Although the Nurikabe that we are generating with this procedure are lacking in symmetry, this is not characteristic of all optimal puzzles. It can be shown that for infinitely many odd values of n, in particular, when $n = 2^k + 3$, there is a perfectly symmetric optimal puzzle of size $n \times n$ whose border consists of alternating land and water tiles. However, this is not the case for every odd n. For example, no optimal

13 × 13 puzzles have symmetry. Furthermore, none of the optimal 13 × 13 puzzles have a completely alternating border.

Further questions

The results above show that for each n there is at least one optimal puzzle of size $n \times n$. We have found several instances in which multiple fundamentally different optimal puzzles exist. This leads to the question of how many distinct optimal puzzles of a given size exist, up to symmetry.

We hope that the techniques found here can be applied to the problem of determining the maximum and minimum number of island tiles in Nurikabe puzzles with other island sizes. Early cases to consider include that of only two-cell islands and that of a mix of one- and two-cell islands. We note that as the island sizes grow, the increased variety in the island shapes adds to the challenge.

Once one moves away from only one-tile islands, it becomes necessary to distinguish between a tiling that satisfies the rules of Nurikabe from the unique solution of a puzzle. For example, of the thirteen colorings of a 2 × 2 grid that satisfy the rules of Nurikabe, only five are the unique solutions of a puzzle. Determining which tilings are are the unique solution of a Nurikabe puzzle provides another layer of complexity.

Appendix

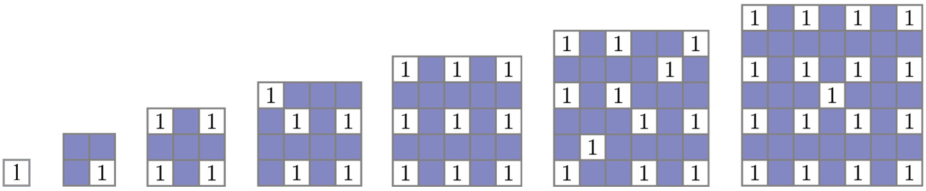


Figure 5 Optimal $n \times n$ Nurikabe, $n = 1, 2, 3, 4, 5, 6$, and 7.

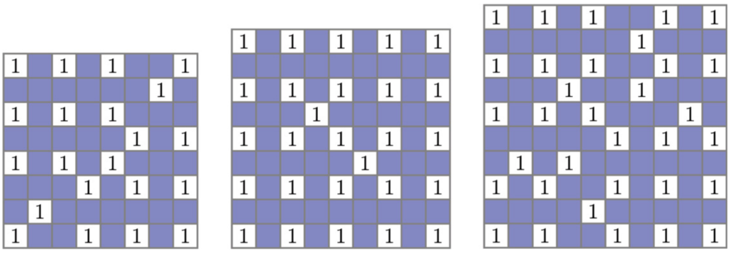


Figure 6 Optimal $n \times n$ Nurikabe, $n = 8, 9$, and 10.

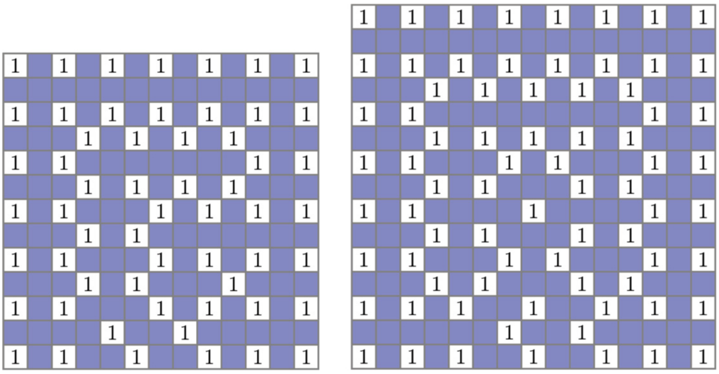


Figure 7 Optimal $n \times n$ Nurikabe, $n = 13$ and 15.

Acknowledgment We thank the referee and Editor for insightful comments that substantially improved the content and clarity of this paper. Thanks also to Killian Meehan of KUIAS and Michio Yoshiwaki of RIKEN AIP, who assisted with translation of the Japanese documents regarding the first recorded appearance of Nurikabe.

REFERENCES

[1] Antonick, G. (2015). Wei-Hwa Huang and Nurikabe puzzles. Available at: https://wordplay.blogs.nytimes.com/2015/02/23/wei-hwa/ Last accessed September 2022.

[2] Foster, M. D. (2015). *The Book of Yokai: Mysterious Creatures of Japanese Folklore*. Oakland: Univ. of California Press.

[3] Voloshin, V. I. (2009). *Introduction to Graph Theory*. New York: Nova Science.

[4] オモロパズル大全集. Available at: https://web.archive.org/web/20040112061310/http://www.nikoli.co.jp/storage/addition/omopadaizen/ Last accessed September 2022.

[5] オモパリスト. Available at: http://nikoli.co.jp/ja/publication/various/nikoli/omopalist/ Last accessed November 2019

[6] 1991年のニコリを振り返る. Available at: http://nikoli.co.jp/ja/misc/30anniversary/looking_back_1991/ Last accessed December 2019.

Summary. We present an introduction to the Japanese pencil puzzle Nurikabe and to its basic solution strategies. Further, we establish formulas for the minimum and maximum number of islands in a Nurikabe puzzle made up of one-tile islands.

JACOB A. BOSWELL (MR Author ID: 1164508) earned a Ph.D. from Purdue University in 2015. He is an assistant professor of mathematics at Missouri Southern State University. His academic interests include mathematical problem solving and commutative algebra. Outside of mathematics, he enjoys video and board games, playing music, and playing pickleball.

JACOB N. CLARK (ORCID: 0000-0002-6142-0602) earned his Ph.D. from the University of Missouri-Columbia in 2019. He is an assistant professor of Mathematics at Missouri Southern State University, alongside his coauthors. His academic interests include the mathematics of games and puzzles, (topological) data analysis, and machine learning. Beyond academic pursuits, he enjoys swimming, cooking, and the outdoors.

CHARLES "CHIP" CURTIS earned a Ph.D. from the University of Washington in 1994. He is a professor of mathematics at Missouri Southern State University. He enjoys mathematical problem solving, hiking, and playing the piano.

Why Quaternions and Octonions Exist

FRANKLIN R. GOULD
Central Connecticut State University
New Britain, CT 06050
fgould@wesleyan.edu

Let (a, b) be an ordered pair of real numbers. What are the coordinates of a vector, expressed in terms of a and b, that will be perpendicular to (a, b) for all a and b (under the usual dot product)? It is immediately clear that any solution will have to be some multiple of the vector $(-b, a)$. It is also clear that the two vectors (a, b) and $(-b, a)$ will span the two-dimensional vector space as long as a and b are not both zero.

Question 1. *For what dimensions n can one find n permutations of n variables with appropriately placed minus signs so that, viewed as the coordinates of n vectors in a real n-dimensional vector space (with an inner product), they form n mutually orthogonal vectors that span the space, unless all n variables are zero?*

To put it another way, "Given an *arbitrary* point in n-dimensional Euclidean space, for which values of n is it possible to find $n - 1$ additional points whose coordinates are each a permutation, up to factors of -1, of the original, so that the lines connecting these points to the origin are mutually perpendicular?" That may sound like an innocent question, but the answer has far-reaching consequences. We will show that Question 1 has a positive solution when $n = 2$, 4, or 8, but not for $n = 3$, and not for any n that is not a power of 2. In fact, there can be no solutions for n equal to any power of 2 greater than 8 because, as we shall see shortly, that would imply the existence of a composition algebra with real dimension greater than 8. We will find that it suffices to show that a solution with $n = 16$ leads to a contradiction. However, as I have not found a simple way to demonstrate this with the methods used in these notes, that task will be left as a challenge to the reader.

The fact that a solution exists for $n = 2$ leads to an algebraic identity, as follows: Let v be an arbitrary vector. If $ab \neq 0$, then we can write:

$$v = c(a, b) + d(-b, a).$$

Then the square of the length of v is given by

$$v \cdot v = c^2(a, b) \cdot (a, b) + d^2(-b, a) \cdot (-b, a) = (c^2 + d^2)(a^2 + b^2)$$

since the cross term, $2(a, b) \cdot (-b, a)$, is zero.

On the other hand, we can also write $v = (ac - bd, bc + ad)$. By computing the square of the length of v directly, we obtain $v \cdot v = (ac - bd)^2 + (bc + ad)^2$, yielding the identity

$$(c^2 + d^2)(a^2 + b^2) = (ac - bd)^2 + (bc + ad)^2.$$

This simple algebraic identity can be verified directly, implying that it must be true in any commutative ring. For example, if we restrict a, b, c, and d to the integers, this tells us that if the integers p and q are each the sum of two squares, then the product pq is also the sum of two squares.

What we have just shown for the case $n = 2$ can easily be generalized to the following statement.

Math. Mag. **95** (2022) 554–567. doi:10.1080/0025570X.2022.2126252 © Mathematical Association of America
MSC: 11R52

Claim 1. *Given n permutations of n real variables so that, as the coordinates of n vectors up to a fixed set of minus signs, the vectors are mutually perpendicular for all values of the variables, there is an n-square ring identity that gives the product of two sums of n squares as another sum of the squares of n quantities that are bilinear in the two original sets of variables.*

To see this, we just need to generalize our previous argument. Assume that the matrix U is a solution to Question 1; that is, the first column u is the vector of variables $(a, b, c, \dots)$ and the remaining $n-1$ columns are permutations, up to minus signs, so that the n columns are mutually perpendicular. Let $v = (r, s, t, \dots)$ be another arbitrary column vector. Now notice that the matrix product $w = Uv$ is a third vector that is a linear combination of the columns in U with coefficients consisting of the components of v. Then the square of the length, $|w|^2$, is given by the matrix product

$$w^t w = (Uv)^t Uv = v^t (U^t U) v = v^t (|u|^2 I_n) v = |u|^2 |v|^2.$$

This is an "n-square identity," where the product of two sums of squares is again expressed as a sum of squares. However, the existence of the identity is a consequence of the existence of the n mutually orthogonal vector variables with the given properties. The converse statement is also true, that when such an identity exists, the specified set of n orthogonal vectors can be formed, but we will not attempt to show this directly.

John H. Conway and Derek A. Smith [**1**] point out what are probably the most important consequences of the algebraic identities that follow from the positive solutions to our question, namely, the existence of the complex numbers, the quaternions, and the octonions! For example, with a, b, c, and d as real numbers, the product of a pair of complex numbers is illustrated by the equation

$$(a + ib)(c + id) = (ac - bd) + i(ad + bc),$$

so our identity is saying that the norm of the product is equal to the product of the norms. This fact is what guarantees that the complex numbers are a division ring. In fact, the complex numbers are, of course, a field as well because the product is commutative.

In their book, Conway and Smith state that the existence of the 1-, 2-, 4-, and 8-square identities is in some sense equivalent to the existence of the reals, the complex numbers, the quaternions and the octonions. In Chapter Six they give a very nice and general proof of the following theorem due to Hurwitz:

Theorem 1 (Hurwitz)**.** *The only composition algebras are $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$, and $\mathbb{O}$.*

By "composition algebra" they mean an algebra with a norm (a mapping from the algebra to the reals satisfying the triangle inequality) that also satisfies $|xy| = |x|\,|y|$. Their proof does not assume any other properties of the norm, but, in fact, the norm (or its square) can be chosen to be quadratic in real multiples of the vector x, so that the corresponding inner product is bilinear.

Having shown that each solution to Question 1 leads to an n-square identity, we will focus most of our attention on the direct connection between solutions to Question 1 and composition algebras over the reals. We are making the case that the solutions to Question 1 are the best way of viewing the facts that make both the n-square identities and the composition algebras possible. To me, at least, they seem less mysterious than the bare algebraic identities, and in that form they can be used to directly define concrete implementations of the composition algebras themselves, as we will demonstrate.

Defining an algebra from a solution to question 1.

Let $v = v_1 = (x_1, x_2, \ldots, x_n)$ be an n-tuple, and suppose that we have a prescription for defining $v_2, \ldots, v_n$ to form $n - 1$ additional vectors that are mutually orthogonal under the usual dot product, where each of the vectors is a permutation of the original n variables up to appropriately inserted minus signs. By using this prescription, we can turn each vector into a matrix by putting v_1 in column 1 and arranging the remaining $n - 1$ vectors into the remaining columns in some prescribed order. This gives us a *linear* map $v \mapsto M(v) := V$ from n-tuples to $n \times n$ matrices. If v and u are two vectors, we can define the product

$$v \star u := Vu,$$

where "$\star$" represents the vector product we have defined, while adjacent terms such as Vu stand for the usual matrix product between a pair of rectangular arrays with compatible dimensions.

The structure we obtain with "$\star$" as a product has the properties of an algebra because the distributive law is satisfied. To see this, note that the prescription that converts a vector into a matrix is a homomorphism of vector addition (thinking of the matrix space as an n^2-dimensional vector space). Of course, matrix multiplication is distributive. The algebra may not be commutative because matrix multiplication is not in general commutative. In fact, an algebra so defined might not even be associative! This happens if the prescription M for turning a vector into a matrix is not a homomorphism from vector multiplication to matrix multiplication. To see this explicitly, suppose that $a \mapsto A$, $b \mapsto B$, et cetera, according to our prescription. Then, clearly, $a \star (b \star c) = ABc$ by our definition. Now suppose that $a \star b = Ab := d$, say. Then if $d \mapsto D \neq AB$, it follows that $(a \star b) \star c \neq a \star (b \star c)$.

We show that if we can convert a vector space into an algebra in this way, then the usual norm obtained from the dot product is a homomorphism of multiplication, which means that our algebra is, in fact, a *composition* algebra. Let u and v be vectors with $u \mapsto U$ and $v \mapsto V$ under our prescription. Then we define the *norm* of v to be the dot product of v with itself, given by

$$|v|^2 := v^t v = {x_1}^2 + {x_2}^2 + \cdots + {x_n}^2,$$

where we take $v^t v$ to be the matrix product of the row vector v^t by the column vector v. Note that the column vectors in V are mutually perpendicular and all have the same length, giving us $V^t V = |v|^2 I_n$, where I_n is the unit matrix. This implies that

$$|v \star u|^2 = (Vu)^t(Vu) = u^t(V^tV)u = |v|^2 u^t I_n u = |v|^2|u|^2,$$

as claimed.

Finding solutions to question 1

Let us now try to see what algebras we can construct using this prescription. Let $v = (x_1, x_2, \ldots, x_n)$ be an n-tuple that we consider to be a column vector. We want to find an additional $n - 1$ n-tuples, each of which is some permutation of these n variables with negative signs inserted appropriately in front of some of them in such a way that the vectors are mutually orthogonal for all values of the variables. If we are successful, we arrange the n column vectors in an $n \times n$ array, that is, each column is one of the mutually perpendicular vectors.

It turns out that the easiest way to find such solutions is to assume that one exists for some undetermined dimension n, and then rearrange it until either (1) The first m rows and the first m columns form a solution and/or (2) a contradiction is obtained. We assume that we have a solution V with n columns and rows:

$$V = \begin{bmatrix} x_1 & \pm x_{i_1} & . & . & . & \pm x_{k_1} \\ x_2 & \pm x_{i_2} & . & . & . & \pm x_{k_2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_n & \pm x_{i_n} & . & . & . & \pm x_{k_n} \end{bmatrix}.$$

However, for illustration purposes, it will help to be able to examine an actual solution with $n = 8$ in order to better visualize the conditions that must be satisfied:

$$V = \begin{bmatrix} a & -t & -v & d & -r & c & s & -b \\ b & v & -t & c & s & -d & r & a \\ c & r & -s & -b & -t & -a & -v & -d \\ d & -s & -r & -a & v & b & -t & c \\ r & -c & d & v & a & -t & -b & -s \\ s & d & c & t & -b & v & -a & r \\ t & a & b & -s & c & r & d & -v \\ v & -b & a & -r & -d & -s & c & t \end{bmatrix}.$$

For the n columns to be orthogonal for all values of the n variables, it is clear that we must have the following:

1. *The Sudoku Property:* No row can contain the same variable more than once, (similar to a Sudoku array). To see this, note that if x_k occurs in two different columns in the same row, then the dot product of those two columns is $\pm {x_k}^2$ when $x_j = 0$ for all $j \neq k$.
2. *The Matching Pairs Property:* For each pair of entries in the same row from two different columns, there is, of necessity, a "matching pair" from the same two columns in a different row that exactly cancels the first pair when we compute the dot product of the pair of columns. This amounts to the requirement that for each pair of variables in a given row, the same two variables occur in the same two columns of a different row, forming the four corners of a sub-array. Two opposite corners of the sub-array match if and only if the other two opposite corners "anti-match." For example, if row 1 has x_5 in column 2 and x_7 in column 3, then some other row must have either x_7 in column 2 and $-x_5$ in column 3 or else $-x_7$ in column 2 and x_5 in column 3. To see this, just set the remaining variables to zero.

From the Matching Pairs Property, we see that n must be even, with the exception of the trivial case, $n = 1$.

It is clear that many different arrangements are possible for essentially one solution. Our next step is to rearrange our solution so that the format is standardized, and so that the first m rows and columns form a sub-matrix comprising a smaller solution. We are free to make any of the following changes:

1. Redefine the variables so that the first column has no minus signs in front of the variables (which we have already done).

2. Change the order of the column vectors.
3. Simultaneously reorder the components of the column vectors (or, equivalently, reorder the rows of the matrix).
4. Multiply any column or row by -1.

Since x_1 must occur in every row of the matrix, we order the column vectors so that $\pm x_1$ occurs along the diagonal. Next, for each $k \geq 2$, if the kth entry in the kth column is $-x_1$, we multiply that column vector by -1. In addition, we can write the matrix V in the following form:

$$V = x_1 I_n + V',$$

where the matrix V' contains only the variables $\pm x_2$ through $\pm x_n$ as entries, with 0s down the diagonal. We can also say that the matrix V' must be anti-symmetric. This follows from the Matching Pairs Property, where one of each of the matching pairs is on the diagonal (now that x_1 is always on the diagonal).

So far, we can say that a vector $v = (a, b, c, d, r, s, t, \ldots)$ in our n-dimensional normed algebra, is mapped to a matrix $M(v) := V$ of the form

$$V = \begin{bmatrix} a & -b & -c & -d & -r & -s & -t & . \\ b & a & . & . & . & . & . & . \\ c & . & a & . & . & . & . & . \\ d & . & . & a & . & . & . & . \\ r & . & . & . & a & . & . & . \\ s & . & . & . & . & a & . & . \\ t & . & . & . & . & . & a & . \\ . & . & . & . & . & . & . & . \end{bmatrix},$$

where $V - aI_n$ is anti-symmetric. It is important to note that selecting the variable $x_1 = a$ to be singled out in this way was an arbitrary choice.

Properties of the algebra derived from a solution

Even at this stage of rearrangement of our solution, we can say that any solution to Question 1 with $n > 1$ contains two smaller solutions—one with $n = 1$ and also our initial example with $n = 2$. With a little work, we can say a good deal more about *any* solution to Question 1 in this form:

Let i_k be the unit basis vector in the kth component direction. Then,

1. There is a two-sided identity element for the $\star$ operation, namely $i_1 = (1, 0, \ldots, 0)$. When using this format, we call the first component of a vector the *real* component.
2. For $k > 1$ we have,

$$i_k \star i_k = (0, \ldots, 1, \ldots)^2 = (-1, 0, 0, \ldots, 0) = -i_1.$$

3. If we define the conjugate, $\overline{u}$, of u to be the first column of U^t, then the conjugate reverses the sign of the non-real components and leaves the real component unchanged. It follows immediately that $\overline{\overline{u}} = u$.
4. $\overline{u} \star u = u \star \overline{u} = |u|^2 i_1$.

5. If $M(u) = U$, then $M(\overline{u}) = U^t$.
6. If $j, k > 1$ and $j \neq k$ then $i_j \star i_k = -i_k \star i_j$
7. $\overline{u \star v} = \overline{v} \star \overline{u}$ (a consequence of 2 and 6).

To illustrate how all of the conclusions 1 through 7 follow without knowing any more details about the map $u \mapsto M(u)$, we examine conclusion 6 a little more closely. Suppose that we want to verify that $i_2 \star i_5 = -i_5 \star i_2$. Our solution to Question 1 looks something like the following (with dots where the array elements are yet to be determined):

$$V = \begin{bmatrix} a & -b & -c & -d & -r & -s & -t & . \\ b & a & . & . & . & . & \mp r & . \\ c & . & a & . & . & . & . & . \\ d & . & . & a & . & . & . & . \\ r & . & . & . & a & . & \pm b & . \\ s & . & . & . & . & a & . & . \\ t & \pm r & . & . & \mp b & . & a & . \\ . & . & . & . & . & . & . & . \end{bmatrix}.$$

Notice that in row 7 we have inserted the pair $\pm r, \ldots, \mp b$ as the "matching pair" for $-b, \ldots, -r$ in row 1. It is true that at this point we do not actually know in which row the matching pair will occur, but the result does not depend on that detail. In addition, we know that the matching pair will be reflected with the opposite signs in column 7. Then from the definition of the product we have

$$i_2 \star i_5 = \begin{bmatrix} 0 & -1 & -0 & -0 & -0 & -0 & -0 & . \\ 1 & 0 & . & . & . & . & \mp 0 & . \\ 0 & . & 0 & . & . & . & . & . \\ 0 & . & . & 0 & . & . & . & . \\ 0 & . & . & . & 0 & . & \pm 1 & . \\ 0 & . & . & . & . & 0 & . & . \\ 0 & \pm 0 & . & . & \mp 1 & . & 0 & . \\ . & . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ . \end{bmatrix} = \mp i_7,$$

and similarly,

$$i_5 \star i_2 = \begin{bmatrix} 0 & -0 & -0 & -0 & -1 & -0 & -0 & . \\ 0 & 0 & . & . & . & . & \mp 1 & . \\ 0 & . & 0 & . & . & . & . & . \\ 0 & . & . & 0 & . & . & . & . \\ 1 & . & . & . & 0 & . & \pm 0 & . \\ 0 & . & . & . & . & 0 & . & . \\ 0 & \pm 1 & . & . & \mp 0 & . & 0 & . \\ . & . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ . \end{bmatrix} = \pm i_7.$$

Two additional features of these algebras are worth mentioning here:

- The *dot product* or *inner product* between two vectors can be expressed using the algebraic operations. Since we have that $|u|^2 i_1 = (\overline{u} \star u)$ for the norm, we can use the following relationship:

$$u \cdot v = \frac{1}{2}(|u+v|^2 - |u|^2 - |v|^2) \quad \Longrightarrow \quad (u \cdot v) i_1 = \frac{1}{2}(\overline{u} \star v + \overline{v} \star u).$$

- We can also define the *cross product*. Given the vectors $u = a i_1 + u'$ and $v = b i_1 + v'$, with u' and v' being the non-real components, we have

$$u \star v = (ab - u \cdot v) i_1 + b u' + a v' + u' \times v'.$$

  where the cross product is defined by this equation.

It is clear from the properties of $\star$ that $u' \times v'$ also has real component zero and is antisymmetric in u' and v'. One can also show that $u' \times v'$ is orthogonal to both u' and v', with $|u' \times v'| = |u'|\,|v'|\,|\sin\theta|$, where θ is the angle between u' and v'. Thus, the cross product defines an algebra on the $(n-1)$-dimensional subspace of a solution to Question 1. This algebra is definitely *not* a composition algebra because, for example, $u' \times u' = 0$. This algebra is not associative (unless it is trivial), even though in the case $n - 1 = 3$, the corresponding 4-dimensional $\star$ algebra (the quaternions) is associative under the $\star$ product. Unlike the $\star$ algebra, the $\times$ algebra has full democracy among all elements with the same length, that is, there is an automorphism of the algebra taking any element to any other element of the same length.

Finding solutions to question 1 when $n > 2$

Now, suppose that our hypothetical solution to Question 1 is for $n > 2$. In order to further standardize the format to avoid duplicate solutions, we permute the components of the column vectors so that the matching pairs in the first two column vectors are adjacent—as is already the case for the pair $(a, -b)$ in row 1 followed by the pair (b, a) in row 2. For example, suppose we happen to have the variables appearing in the following arrangement:

$$V = \begin{bmatrix} a & -b & -c & -e & . & . & -d & . & . \\ b & a & \mp d & \mp f & . & . & \mp c & . & . \\ c & \pm d & a & . & . & . & . & . & . \\ e & \pm f & . & a & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ d & \pm c & . & . & . & . & a & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \end{bmatrix}.$$

We can begin the process by exchanging the rows beginning with e and d. However, in order to maintain the variable a along the diagonal, we follow this with exchanging the *columns* headed by $-e$ and $-d$, resulting in:

$$V = \begin{bmatrix} a & -b & -c & -d & . & . & -e & . & . \\ b & a & \mp d & \mp c & . & . & \mp f & . & . \\ c & \pm d & a & . & . & . & . & . & . \\ d & \pm c & . & a & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ e & \pm f & . & . & . & . & a & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \end{bmatrix}.$$

We can also ensure that matching pairs in rows 3 and 4 of the first two columns are arranged so that the one in column 2 with the negative sign precedes the one with the positive sign—we simply perform one more row/column swap, if necessary. we can continue this procedure so that the matching pairs of variables are in adjacent rows all the way down columns 1 and 2, with the pair containing the negative sign preceding the one that does not.

All of this is to say that, up to an equivalence, if there exists a solution to Question 1 for $n > 2$, then that solution can be reformatted as follows (again, with dots in places where the matrix elements are yet to be determined):

$$V = \begin{bmatrix} a & -b & -c & -d & -r & -s & -t & -v & . & . \\ b & a & d & -c & s & -r & v & -t & . & . \\ c & -d & a & b & . & . & . & . & . & \\ d & c & -b & a & . & . & . & . & . & . \\ r & -s & . & . & a & b & . & . & . & . \\ s & r & . & . & -b & a & . & . & . & . \\ t & -v & . & . & . & . & a & b & . & \\ v & t & . & . & . & . & -b & a & . & . \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . \end{bmatrix}.$$

All of the entries shown are forced by the format decisions we have made up to this point, including the placement of the as and bs down the diagonal as far as the matrix extends. In addition, we can now conclude the following:

1. If Question 1 has a solution for $n > 2$, then it has essentially unique solutions for $n = 1,\ 2$, and 4. In fact, the solutions are displayed above and are easily verified to indeed be solutions.
2. Question 1 has no solution for $n = 3$.

Before going on to investigate $n > 4$, we will need the following result:

Theorem 2 (Sub-matrix condition)**.** *In the solution to Question 1 as formatted above, each 2×2 sub-matrix must contain just two variables, with each variable on one of the diagonals.*

This theorem is proved using the "matching pairs" property by pairing variables within a 2 × 2 submatrix with a or b from two different 2 × 2 blocks along the diagonal. Once that is established, we can go on to arrange a hypothetical solution to Question 1 with $n > 4$ so that the variables occur in 4 × 4 blocks. The rearrangement is accomplished by swapping *adjacent pairs* of rows with each other, along with *adjacent pairs* of columns.

All of this is left as an exercise. We only point out that the procedure can be iterated by doubling the number of adjacent variables each time, allowing us to conclude that n must be a power of 2 for any solution to Question 1. It turns out to be a more or less fruitless exercise because no solutions exist beyond $n = 8$. In what follows, we will derive the $n = 8$ solution(s) less systematically.

The reals and the complex numbers

The reals, of course, are a trivial solution, with $n = 1$. The complex numbers are the case we already described, with $n = 2$ and the vector-to-matrix assignment given by

$$\begin{bmatrix} a \\ b \end{bmatrix} \mapsto \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

The product of two vectors is given by

$$\begin{bmatrix} a \\ b \end{bmatrix} \star \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} ac - bd \\ ad + bc \end{bmatrix}.$$

Under our mapping $u \mapsto M(u)$, the product is mapped to the matrix

$$\begin{bmatrix} ac - bd & -ad - bc \\ ad + bc & ac - bd \end{bmatrix}.$$

But this is exactly what we obtain if we represent *both* factors by their corresponding matrices before computing the product:

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} \cdot \begin{bmatrix} c & -d \\ d & c \end{bmatrix} = \begin{bmatrix} ac - bd & -ad - bc \\ ad + bc & ac - bd \end{bmatrix}.$$

This means that $u \mapsto M(u)$ is actually a two-dimensional real representation of the complex numbers, with

$$a + ib \;=\; \begin{bmatrix} a \\ b \end{bmatrix} \;\mapsto\; \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

In this representation, we see that if $\alpha \mapsto A$ then $\overline{\alpha} \mapsto A^t$. We see that when $u \mapsto M(u)$ is a representation, the transpose properties are directly reflected in the properties of conjugation. It is also easy to check that the complex numbers defined in this way are commutative and are therefore a field. The cross product associated with the complex numbers is trivial since it always yields the zero vector.

The quaternions

Now let us see what happens for $n = 4$. As we noted earlier, the matrix of orthogonal column vectors assigned to the 4-vector $u = (a, b, c, d)$ is completely determined, once we have standardized the format for $M(u)$ up to choosing a diagonal element and an ordering for the first two columns:

$$u = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \quad \mapsto \quad M(u) := U = \begin{bmatrix} a & -b & -c & -d \\ b & a & d & -c \\ c & -d & a & b \\ d & c & -b & a \end{bmatrix}.$$

As we have seen, quaternion products do not commute in general. However, we can be assured that associativity holds because the map $u \mapsto M(u)$ again produces a matrix representation of the algebra. To more easily check that this is indeed the case, we separate the matrix into 2×2 components:

$$U = \left[\begin{array}{cc|cc} a & -b & -c & -d \\ b & a & d & -c \\ \hline c & -d & a & b \\ d & c & -b & a \end{array}\right].$$

We notice that each component is in the form of a complex number in our real two-dimensional representation:

$$U = \left[\begin{array}{c|c} \alpha & -\overline{\beta} \\ \hline \beta & \overline{\alpha} \end{array}\right],$$

where the bar over the entries represents *complex* conjugation. Then if $u = (a, b, c, d)$ and $v = (r, s, t, v)$, we have

$$U = \left[\begin{array}{c|c} \alpha & -\overline{\beta} \\ \hline \beta & \overline{\alpha} \end{array}\right] \quad V = \left[\begin{array}{c|c} \rho & -\overline{\sigma} \\ \hline \sigma & \overline{\rho} \end{array}\right] \quad UV = \left[\begin{array}{c|c} \alpha\rho - \overline{\beta}\sigma & -\alpha\overline{\sigma} - \overline{\beta}\rho \\ \hline \overline{\alpha}\sigma + \beta\overline{\rho} & \overline{\alpha\rho} - \beta\overline{\sigma} \end{array}\right].$$

which is exactly the matrix to which the first column would be mapped, as claimed.

The quaternionic conjugate is given by the transpose of U (which is also the complex conjugate/transpose—or *Hermetian conjugate*—of the matrix, when viewed as a 2×2 complex matrix):

$$\overline{U} = \left[\begin{array}{c|c} \overline{\alpha} & \overline{\beta} \\ \hline -\beta & \alpha \end{array}\right] \qquad \overline{U}U = \left[\begin{array}{c|c} \overline{\alpha}\alpha + \overline{\beta}\beta & 0 \\ \hline 0 & \overline{\alpha}\alpha + \overline{\beta}\beta \end{array}\right].$$

Again, we have $\overline{(UV)} = \overline{V}\,\overline{U}$, reflecting both the transpose property and the fact that quaternions do *not* commute. To compare this with the usual implementation, we make the assignments:

$$1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad i = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad j = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad k = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then, after using $u \mapsto M(u)$ as defined above, we find that all the products of i, j, and k with each other, have the opposite signs of the conventional ones. To obtain the usual correspondence, we can switch the definitions of i and j, for example, or use any odd permutation.

If we examine the cross product on the non-real components of a quaternion, we find that it is just the normal three-dimensional cross product.

The octonions.

If we assume that a solution to Question 1 exists for $n > 4$, then we can continue to rearrange the rows and columns—this time in pairs. We find that we can arrange for each 4×4 sub-matrix to contain just four variables, while maintaining the one variable along the diagonal. In addition, each of those 4×4 sub-matrices is composed of four 2×2 sub-matrices, as before. However, the details within those 2×2 blocks are not completely determined. Even so, we can conclude that the next larger solution, if it exists, is for eight mutually orthogonal column vectors. We have a situation something like the following, with six 2×2 sub-matrices in $M(u)$ that are only partially determined, where u is the first column of the 8×8 matrix:

$$V = \left[\begin{array}{cc|cc||cc|cc}
a & -b & -c & -d & -r & -s & -t & -v \\
b & a & d & -c & s & -r & v & -t \\
\hline
c & -d & a & b & \mp t/v & \mp v/t & \mp r/s & \mp s/r \\
d & c & -b & a & \mp v/t & \mp r/v & \mp s/r & \mp r/s \\
\hline\hline
r & -s & \pm t/v & \pm v/t & a & b & \mp c/d & \mp d/c \\
s & r & \pm v/t & \pm t/v & -b & a & \mp d/c & \mp c/d \\
\hline
t & -v & \pm r/s & \pm s/r & \pm c/d & \pm d/c & a & b \\
v & t & \pm s/r & \pm r/s & \pm d/c & \pm c/d & -b & a
\end{array}\right].$$

However, we will see that once we fill in one of the six cells in a way that is compatible with our orthogonality conditions, the other five will then be determined. We might guess a quaternionic arrangement in the bottom left 4×4 block. That would at least guarantee that the first four columns are mutually orthogonal. Then we would have the following arrangement after accounting for the anti-symmetry of the non-diagonal part of the matrix:

$$V \neq \left[\begin{array}{cc|cc||cc|cc}
a & -b & -c & -d & -r & -s & -t & -v \\
b & a & d & -c & s & -r & v & -t \\
\hline
c & -d & a & b & t & -v & -r & s \\
d & c & -b & a & v & t & -s & -r \\
\hline\hline
r & -s & -t & -v & a & b & . & . \\
s & r & v & -t & -b & a & . & . \\
\hline
t & -v & r & s & . & . & a & b \\
v & t & -s & r & . & . & -b & a
\end{array}\right].$$

Upon closer inspection, we see that while the first four columns are mutually perpendicular, columns 5 through 8 cannot be orthogonal to columns 3 and 4 unless $b = 0$. On the other hand, notice what happens if we swap the bottom half of columns 3 and 4 and the second half of rows 3 and 4. This maintains the anti-symmetry, and it is still the case that the first four columns are mutually perpendicular, and likewise for the first four rows. When we attempt to fill the remaining two 2×2 cells, there is one forced solution:

$$V = \left[\begin{array}{cc|cc||cc|cc}
a & -b & -c & -d & -r & -s & -t & -v \\
b & a & d & -c & s & -r & v & -t \\
\hline
c & -d & a & b & v & t & -s & -r \\
d & c & -b & a & t & -v & -r & s \\
\hline\hline
r & -s & -v & -t & a & b & d & c \\
s & r & -t & v & -b & a & c & -d \\
\hline
t & -v & s & r & -d & -c & a & b \\
v & t & r & -s & -c & d & -b & a
\end{array}\right].$$

We notice two things about the six new 2×2 entries into the matrix:

1. Unlike the other entries, these new ones are no longer in the form of our real representation of a complex number.
2. The variables in the new 2×2 entries have "traded diagonals" as compared with their counterparts in the remainder of the matrix.

There are no 8×8 solutions with all 2×2 entries in the complex format. However, there does exist an alternative solution with all eight variables on consistent diagonals throughout. This can be obtained by making the following substitutions within those six new cells only:

$$t \mapsto v, \quad v \mapsto -t, \quad r \mapsto s, \quad s \mapsto -r, \quad c \mapsto d, \quad d \mapsto -c,$$

giving us the solution:

$$V = \left[\begin{array}{cc|cc||cc|cc}
a & -b & -c & -d & -r & -s & -t & -v \\
b & a & d & -c & s & -r & v & -t \\
\hline
c & -d & a & b & -t & v & r & -s \\
d & c & -b & a & v & t & -s & -r \\
\hline\hline
r & -s & t & -v & a & b & -c & d \\
s & r & -v & -t & -b & a & d & c \\
\hline
t & -v & -r & s & c & -d & a & b \\
v & t & s & r & -d & -c & -b & a
\end{array}\right].$$

This is the solution that we will use to define octonion multiplication. As we warned, when two such matrices are multiplied, the columns remain mutually orthogonal (since multiples of orthogonal matrices form a group under multiplication), but columns 2 through 8 in the matrix product are no longer just permutations of the variables in column 1, up to a sign. Instead, they are more complicated linear combinations of those variables.

Fortunately, it is not necessary to lay out the general form for a $\star$ product in order to produce a multiplication table for the octonions. Let $i_1, i_2, \ldots, i_8$ be the unit basis vectors in our defining basis. Each of them maps to a matrix according to the above prescription; for example, $i_1 = (1, 0, 0, 0, 0, 0, 0, 0) \mapsto I_8$ (the identity matrix). We can then compute product $i_k \star i_j$ as the matrix for i_k times the column vector i_j. We find that this amounts to using the above matrix of variables $a, b, c, \ldots$ as a kind of multiplication table itself. For example, to multiply $i_3 \star i_5$:

1. Since "r" is in the fifth row of column 1, find "$\pm c$" in the 5th column (headed by $-r$) and note which row it's in (row 7) and the sign $(+1)$.
2. This means that $i_3 \star i_5 = i_7$.

Note that $i_5 \star i_3$ must have the opposite sign because the matching pair in row 7 for the pair $-c$ and $-r$ in row 1 must have opposite signs. Here is the resulting multiplication table:

$\star$	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
i_1	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
i_2	i_2	$-i_1$	$-i_4$	i_3	$-i_6$	i_5	$-i_8$	i_7
i_3	i_3	i_4	$-i_1$	$-i_2$	i_7	$-i_8$	$-i_5$	i_6
i_4	i_4	$-i_3$	i_2	$-i_1$	$-i_8$	$-i_7$	i_6	i_5
i_5	i_5	i_6	$-i_7$	i_8	$-i_1$	$-i_2$	i_3	$-i_4$
i_6	i_6	$-i_5$	i_8	i_7	i_2	$-i_1$	$-i_4$	$-i_3$
i_7	i_7	i_8	i_5	$-i_6$	$-i_3$	i_4	$-i_1$	$-i_2$
i_8	i_8	$-i_7$	$-i_6$	$-i_5$	i_4	i_3	i_2	$-i_1$

Once we determine the multiplication table, then everything else is determined by the distributive law. This time, we find that the octonion $\star$ product is no longer associative, and thus the function M can no longer be a representation. However, property 7 must still hold true.

One way to see the equivalence of all non-real directions is to arbitrarily choose two orthogonal vectors u and v, both with vanishing real components. Then $w = u \star v = u \times v$ will turn out to be a third non-real direction orthogonal to the first two; and the four vectors i_1, u, v, w will generate a sub-algebra isomorphic to the quaternions.

We have based the above multiplication table for the octonions on our "standard" form for the matrix showing our special set of orthogonal vectors. This was useful for developing the overall picture of the normed algebras. However, if we were to focus on the intrinsic symmetries of the octonions alone, then we would choose a different ordering and naming convention for an orthonormal basis. The choice made by Conway and Smith can be emulated by making these substitutions:

$$i_1 \mapsto i_\infty = 1, \quad i_2 \mapsto i_1, \quad i_3 \mapsto i_2, \quad i_4 \mapsto -i_4$$

$$i_5 \mapsto i_0, \quad i_6 \mapsto i_3 \quad i_7 \mapsto -i_6 \quad i_8 \mapsto -i_5.$$

Then, for i_0 through i_6 we have:

$$i_k \star i_{k+1} = i_{k+3} \quad \text{with subscript addition mod 7.}$$

This determines the entire multiplication table when combined with cyclic permutations and the fact that two perpendicular vectors anti-commute when their real components are zero.

Acknowledgements My thanks to Douglass L. Grant and his December 2018 article in this Magazine, "Proving Euler's Four-Square Lemma Using Linear Algebra," where he uses the algebraic identity derived from the $n = 4$ case to obtain the corresponding number-theoretic fact about the sum of four squares [2]. It was this article that led me down the path that resulted in these notes.

Special thanks to Roger Vogeler, professor of mathematics at Central Connecticut State University. He has helped to greatly improve the both the accuracy and the readability of these notes.

Of course, the book *On Quaternions and Octonions* by John H. Conway and Derek A. Smith has been an invaluable resource [1]. I was greatly saddend to hear of Conway's passing in April of 2020. His insights and expository style will be greatly missed.

REFERENCES

[1] Conway, J. H., Smith, D. A. (2003). *On Quaternions and Octonions.* Boca Raton: CRC Press.

[2] Grant, D. L. (2018). Proving Euler's four-square lemma using linear algebra. *Mathematics Magazine* 91(4): 384–385.

Summary. There is a simple combinatorial anomaly, making possible some special linear algebra and thereby some special geometry, that occurs only in dimensions 1, 2, 4, and 8. The consequences are wide ranging and in particular lead to the existence of the complex numbers, the quaternions and the octonions. This article explains why the anomaly exists only in these dimensions using elementary linear algebra.

FRANK GOULD graduated from Berea Colege in 1965 with a degree in physics. He subsequently earned an MA in physics at Syracuse University. After working for many years at Aetna as a programmer and performance analyst, he returned to graduate school at Wesleyan University, earning a PhD in topological groups in 2009. He also enjoys all kinds of puzzles and hiking mountain trails.

Editor's Notes

JASON ROSENHOUSE
James Madison University
Editor, *Mathematics Magazine*

- *Mathematics Magazine* prints in black and white, but authors often send in color diagrams. Usually this does not present a problem. We use the color diagrams in the online version of the article. In most cases, the color diagrams are still perfectly readable when rendered in black and white, meaning there is no need to prepare separate black and white diagrams for the print magazine. Unfortunately, that was not the case with the article, “A Few Ripe Red-Blue Cherries” by Scott Andrew Herman, published in the June 2022 issue. When the red and blue cherries of the title were rendered in black and white, the resulting shades of gray appeared identical, rendering the print version of the article all but unreadable. I would urge anyone interested in this article to have a look at the online version, where everything appears as it should. If anyone reading this does not have access to the online version, then simply write to me directly at rosenhjd@jmu.edu, and I will send a copy to you. I greatly regret the error in allowing this printing issue to slip through.

- In the December 2021 issue, we ran an article called “A Frameless 2-Coloring of the Plane Lattice” by C. W. Groetsch and Jeffrey Shallit. It has since come to light that Theorem 3 of that paper was stated, without proof, in a 1965 paper by Hao Wang [1]. We thank Jean-Paul Allouche for calling this to our attention.

REFERENCES

[1] Wang, H. (1965). Games, logic and computers. *Sci. Amer*. 213(5): 98–107.

Math. Mag. **95** (2022) 568. doi:10.1080/0025570X.2022.2136438 © Mathematical Association of America

PROOFS WITHOUT WORDS

Golden Tiles and Arctangent Identities

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

The *golden rectangle* (a $\phi \times 1$ rectangle where ϕ denotes the *golden ratio*) has a property that enables us to use it in plane tilings and then overlay right triangles to derive arctangent identities such as the following:

$$\arctan 2 = 2\arctan\frac{1}{\phi}; \tag{1}$$

$$\frac{\pi}{4} = \arctan\frac{1}{\phi} + \arctan\frac{1}{2\phi+1}. \tag{2}$$

If we place two copies of an $x \times 1$ rectangle adjacent to one another, one in "landscape" orientation and the other in "portrait" orientation, as shown in Figure 1, then the three vertices marked ∘ are collinear if and only if the rectangle is golden.

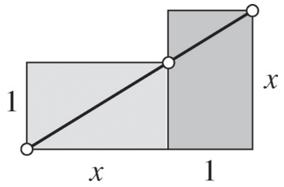


Figure 1 We have that $x > 0$ and $\frac{x}{x+1} = \frac{1}{x}$ if and only if $x = \phi$.

Figures 2 and 3 now provide visual proofs of equations (1) and (2), respectively.

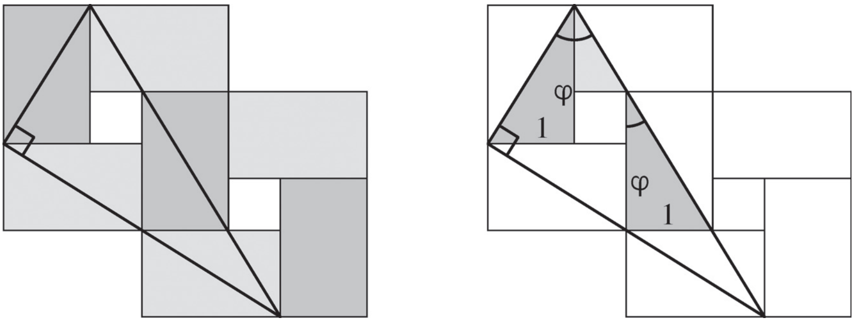


Figure 2 Proof of equation (1).

Math. Mag. **95** (2022) 569–570. doi:10.1080/0025570X.2022.2125732 © Mathematical Association of America

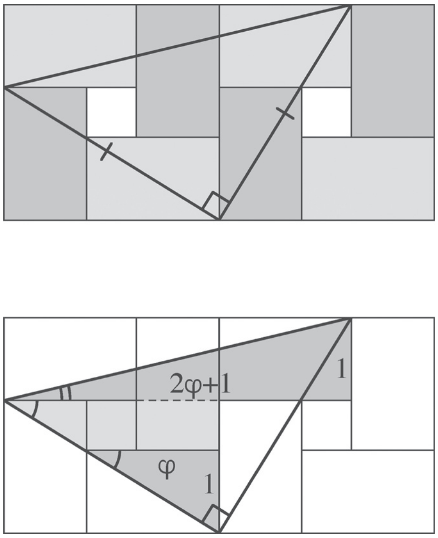


Figure 3 Proof of equation (2).

For another proof without words of equation (1), see Plaza [1]. A third identity,

$$\arctan\frac{1}{2} = \arctan\phi - \arctan\frac{1}{\phi},$$

can be found in Figure 2 by considering the smaller acute angle in the large right triangle.

Acknowledgments The author wishes to thank an anonymous reviewer and the Editor for helpful suggestions on an earlier draft of this note.

REFERENCES

[1] Plaza, A. (2017). Proof without words: Arctangent of two and the golden ratio. *Math. Mag.* 90(3): 179. doi.org/10.4169/math.mag.90.3.179

Summary. We use plane tilings with golden rectangles to derive arctangent identities.

ROGER B. NELSEN (MR Author ID: 237909) is a professor emeritus at Lewis & Clark College, where he taught mathematics and statistics for 40 years.

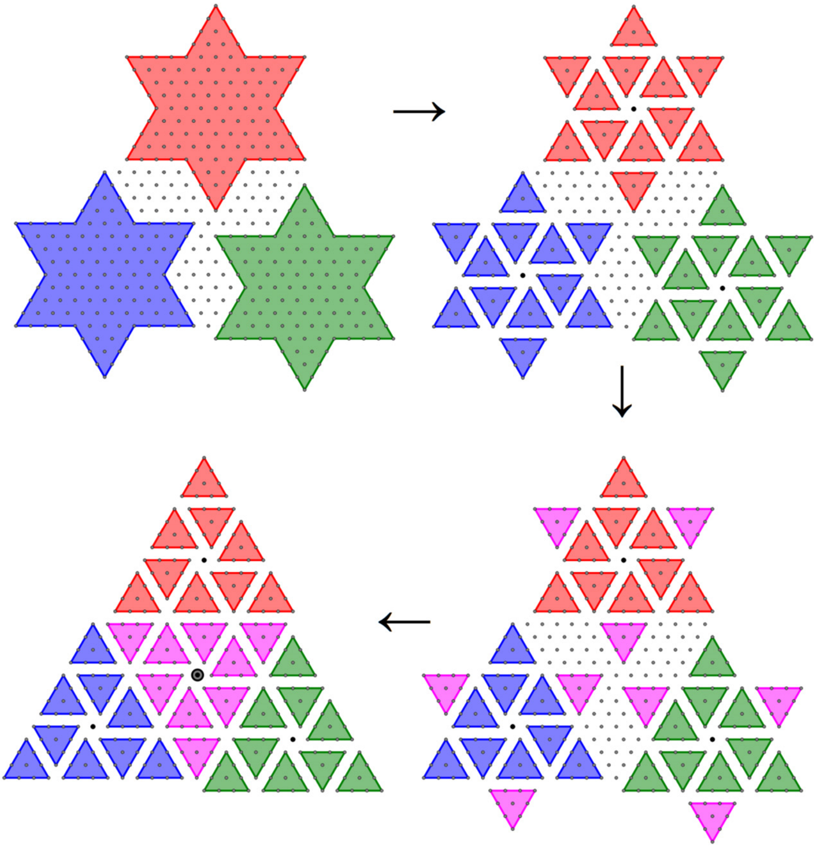
A Relationship of Triangular and Star Numbers

GUNHAN CAGLAYAN
New Jersey City University
Jersey City, NJ 07305
gcaglayan@ncju.edu

A *star number*, given by the formula $S_n = 6n(n-1)+1$, is the figurate number represented by the number of dots in a regular hexagram.

Theorem 1. *Let S_n and T_n represent the nth star number and nth triangular number, respectively. Then we have that $4T_{3n+1} = 3S_{n+1} + 1$.*

Our proof illustrates the case $n = 4$, implying that $T_{13} = 91$ and $S_5 = 121$.



Summary. We give a visual proof for an identity relating triangular and star numbers.

GUNHAN CAGLAYAN (MR Author ID: 1116420) teaches mathematics at New Jersey City University. His main interests are visual mathematics and student learning through modeling and visualization.

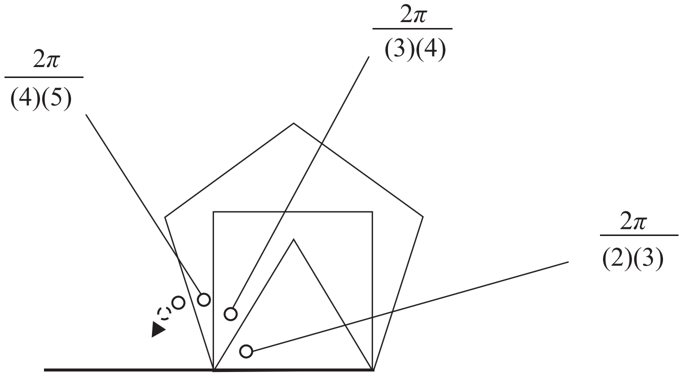
Math. Mag. **95** (2022) 571. doi:10.1080/0025570X.2022.2126159 © Mathematical Association of America

The Series of Reciprocals of Triangular Numbers

PAUL STEPHENSON
Magic Mathworks Travelling Circus
Essen, Germany 45144
pstephenson1@me.com

The triangular numbers are defined by the formula $T_n = \frac{n(n+1)}{2}$, from which it follows that the triangle number reciprocals are given by $1/T_n = \frac{2}{n(n+1)}$. Recall that the interior angle of a regular n-gon is $\frac{(n-2)\pi}{n}$. This implies that the difference between the angle of the regular $(n+1)$-gon and the regular n-gon is $\frac{2\pi}{n(n+1)}$.

Theorem 1. *The infinite sum of the triangle number reciprocals is* 1.



$$\frac{2\pi}{(2)(3)} + \frac{2\pi}{(3)(4)} + \frac{2\pi}{(4)(5)} + \cdots = \pi$$

$$\frac{2}{(2)(3)} + \frac{2}{(3)(4)} + \frac{2}{(4)(5)} + \cdots = 1$$

Other visual proofs of this result can be found in the book by Nelsen [1].

REFERENCES

[1] Nelsen, R. (2020). *Proofs Without Words*, Vol. 1. Providence, RI: MAA Press.

Summary. We provide a visual proof that the infinite sum of the reciprocals of the triangular numbers is 1. Such visual proofs usually stack line segments or rectangular blocks. Here we stack angles.

PAUL STEPHENSON taught mathematics and physics for two decades before creating a touring math lab structured around the multiple embodiment principle of the late Z. P. Dienes. The Magic Mathworks Traveling Circus, registered in the U. K. as a not-for-profit company, has now been on the road for thirty years.

Math. Mag. **95** (2022) 572. doi:10.1080/0025570X.2022.2126173 © Mathematical Association of America

PROBLEMS

Proposals

To be considered for publication, solutions should be received by May 1, 2023.

2156. *Proposed by Cezar Lupu, Tsinghua University, Beijing, China.*

Let $ABCD$ be a convex quadrilateral in the plane with vertices having rational coordinates. Let P be a point in its interior having rational coordinates such that

$$m\angle PAB = m\angle PBC = m\angle PCD = m\angle PDA = q\pi, \text{ with } q \in \mathbb{Q}.$$

Show that $ABCD$ is a square. Give an example to show that the condition that $q \in \mathbb{Q}$ cannot be dropped.

2157. *Proposed by Philippe Fondanaiche, Paris, France.*

Consider two sequences. One is the number of digits in the base 2 representation of 10^k, $k = 1, 2, \ldots$, and the other is the number of digits in the base 5 representation of 10^k, $k = 1, 2, \ldots$. Show that every integer greater than 1 appears in exactly one of the two sequences. Which sequence contains 2023?

2158. *Proposed by the Missouri State University Problem Solving Group, Missouri State University, Springfield, MO.*

(a) Arrange the integers from 1 to 15 (inclusive) in a row so that the sum of any two adjacent numbers is a perfect square.

(b) Find the smallest positive integer n such that the integers from 1 to n can be arranged in a circle so that the sum of any two adjacent numbers is a perfect square. Justify your answer.

Math. Mag. **95** (2022) 573–582. doi:10.1080/0025570X.2022.2126649 © Mathematical Association of America

We invite readers to submit original problems appealing to students and teachers of advanced undergraduate mathematics. Proposals must always be accompanied by a solution and any relevant bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. Submitted problems should not be under consideration for publication elsewhere.

Proposals and solutions should be written in a style appropriate for this MAGAZINE.

Authors of proposals and solutions should send their contributions using the Magazine's submissions system hosted at http://mathematicsmagazine.submittable.com. *More detailed instructions are available there. We encourage submissions in PDF format, ideally accompanied by* LaTeX *source. General inquiries to the editors should be sent to* mathmagproblems@maa.org.

2159. *Proposed by George Stoica, Saint John, NB, Canada.*

Two players, A and B, alternately throw a pair of dice with A going first. Let $a, b \in \{2, 3, \dots, 12\}$ be fixed. Player A wins by having a roll worth a points before player B has a roll worth b points. Otherwise, player B wins.

What is the probability that player A wins?

2160. *Proposed by Gregory Dresden, Washington & Lee University, Lexington, VA.*

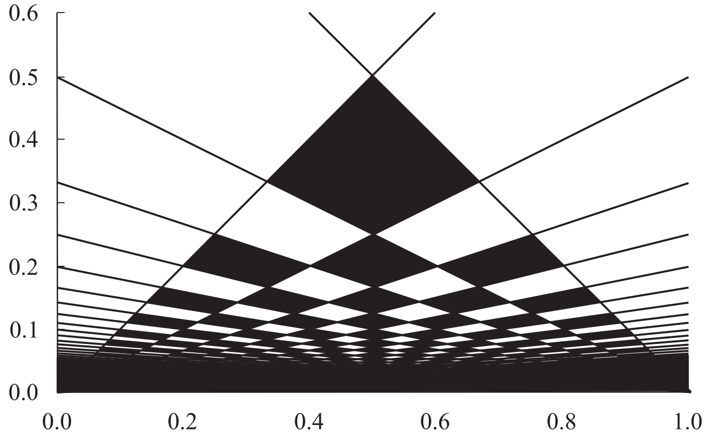
Consider the lines

$$y = x/1,\ y = x/2,\ y = x/3,\ y = x/4, \dots$$

and the lines

$$y = (1 - x)/1,\ y = (1 - x)/2,\ y = (1 - x)/3,\ y = (1 - x)/4, \dots,$$

which intersect to form an infinite number of quadrilaterals. Starting with the lozenge at the top, shade every other quadrilateral, as shown in the figure.



Find the total area of all the shaded quadrilaterals.

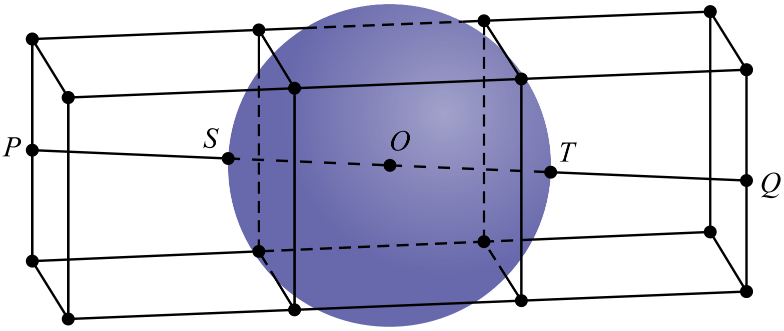
Quickies

1125. *Proposed by Albert Natian, Los Angeles Valley College, Valley Glen, CA.*

Let m and n be two arbitrary positive integers. Show that there is a positive integer x such that the decimal representation of mx contains the decimal representation of n as a substring.

1126. *Proposed by Tran Quang Hung, Hanoi, Vietnam.*

Three unit cubes are placed face-to-face as shown in the figure. A sphere with center O is tangent to all of the edges of the central cube. Let P be the midpoint of one of the sides of a face parallel to, but disjoint from, the two common faces. Let Q be the reflection of P through O. Denote the points of intersection of line PQ and the sphere by S and T, where S is between P and T. Determine ST/PS.



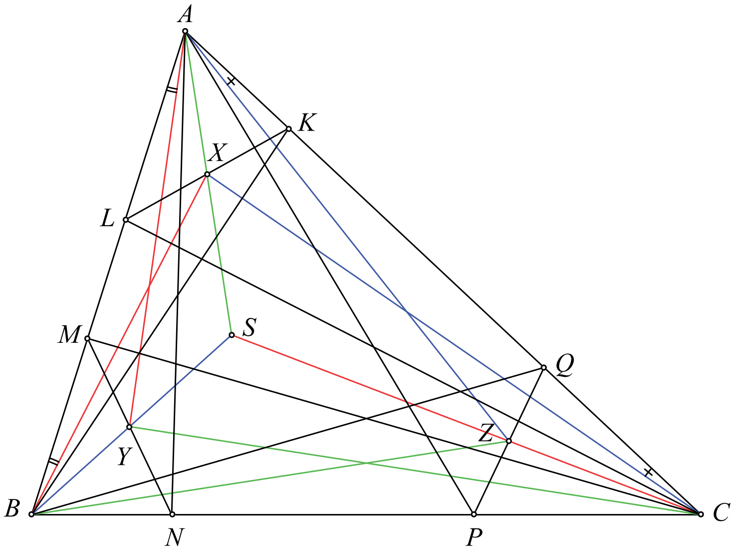
Solutions

A property of the symmedian point **December 2021**

2131. *Proposed by Tran Quang Hung, Hanoi, Vietnam.*

Recall that a symmedian is the reflection of a median through a vertex across the angle bisector passing through that vertex. The three symmedians of a triangle meet in a point known as the symmedian (or Lemoine or Grebe) point. Let ABC be a triangle with symmedian point S. Let X, Y, and Z be points lying on segments SA, SB, and SC, respectively, such that $\angle XBA \cong \angle YAB$ and $\angle XCA \cong \angle ZAC$. Prove that $\angle ZBC \cong \angle YCB$.

Solution by Do Van Quyet, Vinh Phuc, Vietnam.
Recall that line ℓ_1 is said to be anti-parallel to line ℓ_2 with respect to lines m_1 and m_2 if the opposite angles in the quadrilateral formed by the four lines are supplementary.



Let the anti-parallel line to BC with respect to sides AC and AB passing through X meet those sides at K and L, respectively.

Let the anti-parallel line to AC with respect to sides BA and BC passing through Y meet those sides at M and N, respectively.

Let the anti-parallel line to AB with respect to sides CB and CA passing through Z meet those sides at P and Q, respectively.

Note that quadrilaterals $BCKL$, $CAMN$, and $ABPQ$ are cyclic.

A key property of a symmedian through a vertex is that it bisects any anti-parallel to the opposite side with respect to the adjacent sides. Therefore, X, Y, and Z are the midpoints of segments KL, MN, and PQ, respectively.

We have

$$\angle ALK \cong \angle ACB \text{ (since } BCKL \text{ is cyclic),}$$

and

$$\angle ACB \cong \angle BMN \text{ (since } CAMN \text{ is cyclic).}$$

Therefore, $\angle ALK \cong \angle BMN$ and consequently, $\angle BLX \cong \angle AMY$ (supplementary angles). We are given that $\angle XBA \cong \angle YAB$, so

$$\triangle AMY \sim \triangle BLX \text{ by the AA criterion.}$$

Since X and Y are the midpoints of KL and MN, respectively, we deduce that $\triangle AMN \sim \triangle BLK$. Therefore, $\angle LBK \cong \angle MAN$. Now

$$\angle MAN \cong \angle MCN$$

since $CAMN$ is cyclic and the angles are subtended by the same arc. Therefore, $\angle LBK \cong \angle MCN$.

A similar argument shows that $\angle LCK \cong \angle QBP$.

We have

$$\angle LBK \cong \angle LCK,$$

since $BCKL$ is cyclic and the angles are subtended by the same arc.

From the three congruences directly above, we obtain $\angle MCN \cong \angle QBP$. Now

$$\angle QPC \cong \angle BAC \text{ (because } CAMN \text{ is cyclic)}$$

and

$$\angle BAC \cong \angle MNB \text{ (because } ABPQ \text{ is cyclic).}$$

Thus

$$\angle QPC \cong \angle MNB, \text{ and therefore, } \angle BPQ \cong \angle MNC \text{ (supplementary angles).}$$

Hence,

$$\triangle CMN \sim \triangle BQP \text{ by the AA criterion.}$$

Since Y and Z are the midpoints of MN and PQ, respectively, $\triangle BZP \sim \triangle CYN$. Therefore, $\angle ZBC = \angle YCB$, as we wished to show.

Also solved by Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Nandan Sai Dasireddy (India), Michael Goldenberg & Mark Kaplan, Volkhard Schindler (Germany), and the proposer.

Buffon's tetrahedron **December 2021**

2132. *Proposed by the Missouri State University Problem Solving Group, Missouri State University, Springfield, MO.*

A regular tetrahedral die with sides of length 1 is tossed onto a floor having a family of parallel lines spaced 1 unit apart. What is the probability that the die lands on a line?

Solution by the Eagle Problem Solvers, Georgia Southern University, Statesboro, GA and Savannah, GA.

Since the tetrahedron is regular, every configuration of the bottom triangular face on the floor is equally likely. In other words, the probability we seek is the same as the probability of a randomly tossed equilateral triangle landing on a line. Orient the parallel lines horizontally and use the usual cartesian coordinate system. We can give the vertical coordinate of any point on the floor as a real number in the interval $[0, 1)$, representing the distance to the closest horizontal line below, or passing through, the given point. Let y represent the vertical coordinate of the lowest point of the triangular face. Let θ represent the angle with smallest nonnegative measure between the sides of the triangle containing the lowest point and the positive x-axis. Then

$$0 \le y < 1 \qquad \text{and} \qquad 0 \le \theta < \frac{2\pi}{3}.$$

Thus, a random toss of the equilateral triangle corresponds to a random selection of a point (θ, y) from the rectangle

$$\left[0, \frac{2\pi}{3}\right) \times [0, 1).$$

If we rotate around a vertex fixed on a horizontal line, then the vertical coordinate of the highest vertex will be

$$\sin\left(\frac{\pi}{3} + \theta\right) \text{ for } 0 \le \theta \le \frac{\pi}{3}, \quad \text{and} \quad \sin\theta \text{ for } \frac{\pi}{3} \le \theta < \frac{2\pi}{3}.$$

Thus, the triangle will miss all horizontal lines if and only if

$$0 < y < 1 - \sin\left(\frac{\pi}{3} + \theta\right)$$

for $0 \le \theta \le \frac{\pi}{3}$ and

$$0 < y < 1 - \sin\theta$$

for $\frac{\pi}{3} \le \theta < \frac{2\pi}{3}$.

The area of this region in the rectangle is given by

$$\int_0^{\pi/3} \left[1 - \sin\left(\frac{\pi}{3} + \theta\right)\right] d\theta + \int_{\pi/3}^{2\pi/3} (1 - \sin\theta)\, d\theta = \frac{2\pi}{3} - 2.$$

Thus, the probability that the tetrahedral die misses all lines is

$$\frac{\frac{2\pi}{3} - 2}{\frac{2\pi}{3}} = 1 - \frac{3}{\pi},$$

and the probability that the die lands on a line is

$$\frac{3}{\pi} \approx 0.95493.$$

Editor's Note. Michael Vowe points out that a more general result is known (and has been rediscovered multiple times): if d is the distance between the lines, and p is the perimeter of a convex polygon, then the probability the polygon lands on a line is $p/(\pi d)$ as long as the diameter of the polygon is less than or equal to d. See: Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. New York: McGraw-Hill, pp. 251–255.

Also solved by Jacob Boswell & Chip Curtis, Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Owen Byer and the Calculus II class at Eastern Mennonite University, Robert Calcaterra, Stephen J. Herschkorn, José Heber Nieto (Venezuela), Didier Pinchon (France), Volkhard Schindler (Germany), Randy K. Schwartz, Michael Vowe (Switzerland), and the proposers. There were three incomplete or incorrect solutions.

An infinite series involving the tangent function December 2021

2133. *Proposed by Péter Kórus, University of Szeged, Szeged, Hungary.*

Evaluate the infinite sum

$$\sum_{k=1}^{\infty} 2^{-k} \tan\left(2^{-k}\right).$$

Solution by Seán M. Stewart, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Observe that for $x \in (0, \pi/2)$ we have

$$2\cot(2x) - \cot(x) = 2 \cdot \frac{\cot^2(x) - 1}{2\cot(x)} - \cot(x) = -\frac{1}{\cot(x)} = -\tan(x).$$

Setting $x = 2^{-k}$ in this trigonometric identity and multiplying both sides by 2^{-k} we obtain

$$\frac{1}{2^k} \tan\left(\frac{1}{2^k}\right) = -\frac{1}{2^k} \cot\left(\frac{1}{2^k}\right) - \frac{1}{2^{k-1}} \cot\left(\frac{1}{2^{k-1}}\right).$$

Consider the nth partial sum

$$S_n = \sum_{k=1}^{n} 2^{-k} \tan\left(2^{-k}\right).$$

From the equation above, we can write this partial sum as

$$\begin{aligned} S_n &= \sum_{k=1}^{n} \left[\frac{1}{2^k} \cot\left(\frac{1}{2^k}\right) - \frac{1}{2^{k-1}} \cot\left(\frac{1}{2^{k-1}}\right)\right] \\ &= -\cot(1) + 2^{-n} \cot\left(2^{-n}\right) \end{aligned}$$

since the sum telescopes. Therefore, the required sum is

$$\sum_{k=1}^{\infty} 2^{-k} \tan\left(2^{-k}\right) = \lim_{n\to\infty} S_n = -\cot(1) + \lim_{n\to\infty} 2^{-n} \cot\left(2^{-n}\right).$$

Letting $u = 2^{-n}$, we have

$$\lim_{n\to\infty} 2^{-n}\cot\left(2^{-n}\right) = \lim_{u\to 0^+} u\cot(u) = \lim_{u\to 0^+} \frac{u}{\tan(u)} = 1.$$

Therefore,

$$\sum_{k=1}^{\infty} 2^{-k}\tan\left(2^{-k}\right) = 1 - \cot(1).$$

Also solved by Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Paul Bracken, Brian Bradie, Robert Calcaterra, Hongwei Chen, CMC 328, Bruce Davis, Prithwijit De (India), Noah Garson (Canada), Subhankar Gayen (India), G. Greubel, Lixing Han, Mark Kaplan, Kelly McLenithan, Albert Natian, José Nieto (Venezuela), Northwestern University Math Problem Solving Group, Shing Hin Jimmy Pa (China), Didier Pinchon (France), Angel Plaza & Francisco Perdomo (Spain), Michael Reid, Henry Ricardo, Celia Schacht, Volkhard Schindler (Germany), Vishwesh Ravi Shrimali (India), Albert Stadler (Switzerland), Michael Vowe (Switzerland), and the proposer. There were three incomplete or incorrect solutions.

Questions about nilpotent matrices December 2021

2134. *Proposed by Antonio Garcia, Strasbourg, France.*

Let $N \in M_n(\mathbb{R})$ be a nilpotent matrix. In what follows, $X \in M_n(\mathbb{R})$.

(a) Show that there is always an X such that $N = X^2 + X - I$.

(b) Show that if n is odd, there is no X such that $N = X^2 + X + I$.

(c) Show that if $n = 2$ and $N \neq 0$, there is no X such that $N = X^2 + X + I$.

(d) Give examples, when $n = 4$, of an $N \neq 0$ and an X such that $N = X^2 + X + I$ and of an N with no X such that $N = X^2 + X + I$.

Solution by the Case Western Reserve University Problem Solving Group, Case Western Reserve University, Cleveland, OH.

(a) We claim that if M is a nilpotent matrix, then $I + M$ has a square root. Consider the formal power series

$$\sqrt{1+x} = \sum_{i=0}^{\infty} \binom{1/2}{i} x^i.$$

If $M^k = 0$, we set $x = M$ and obtain

$$\sqrt{1+M} = \sum_{i=0}^{k-1} \binom{1/2}{i} M^i.$$

Returning to the problem, we may rewrite the condition as

$$I + \frac{4}{5}N = \left(\frac{2\sqrt{5}}{5}X + \frac{\sqrt{5}}{5}I\right)^2.$$

Since $\frac{4}{5}N$ is nilpotent, we can solve for X using the claim above.

(b) Assume to the contrary that there exists such an X. Since N is nilpotent,

$$N^k = (X^2 + X + I)^k = 0$$

for some k. This implies that $(\lambda^2 + \lambda + 1)^k$ is a polynomial multiple of the minimal polynomial of X. Therefore X cannot have any real eigenvalues, since the eigenvalues of X are the roots of the minimal polynomial, and $(\lambda^2 + \lambda + 1)^k$ has no real roots. However, n is odd, which guarantees that X has a real eigenvalue. This is a contradiction.

(c) Suppose there exists such an X. Since we are in dimension two,

$$N^2 = (X^2 + X + I)^2 = 0.$$

This implies that $(\lambda^2 + \lambda + 1)^2$ is a polynomial multiple of the minimal polynomial of X. Since $N = X^2 + X + I \neq 0$, $(\lambda^2 + \lambda + 1)^2$ must be the minimal polynomial of X. The characteristic polynomial of X must have degree 2, and also must be a multiple of the minimal polynomial. But the minimal polynomial has degree 4. This is a contradiction.

(d) Let

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

It is straightforward to verify that the characteristic polynomial of X is

$$\lambda^4 + 2\lambda^3 + 3\lambda^2 + 2\lambda + 1 = \left(\lambda^2 + \lambda + 1\right)^2.$$

Let $N = X^2 + X + I$. One readily verifies that $N \neq 0$ and by the Cayley-Hamilton theorem, $N^2 = 0$. This solves the first part of the problem.

For the second part of the problem, let

$$N = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and assume there were an X such that $N = X^2 + X + I$. Since $N^4 = 0$, the minimal polynomial of X must be a polynomial multiple of $(\lambda^2 + \lambda + 1)$. Because $N^k \neq 0$ for $k < 4$, $(\lambda^2 + \lambda + 1)^4$ must be the minimal polynomial. The characteristic polynomial of X must be a multiple of the minimal polynomial and also must have degree 4. But $(\lambda^2 + \lambda + 1)^4$ has degree 8. This is a contradiction.

Also solved by Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Jacob Boswell & Chip Curtis, Paul Budney, Robert Calcaterra, Lixing Han, Eugene A. Herman, Sonebi Omar (Morroco), Didier Pinchon (France), Michael Reid, and the proposer.

An exponential generating function — December 2021

2135. *Proposed by Băetu Ioan, "Mihai Eminescu" National College, Botoşani, Romania.*

For $k \in \mathbb{Z}^+$, let $a_n(k)$ denote the number of elements $\sigma \in S_n$, the group of all permutations on an n-element set, such that $\sigma^k = e$, the identity element. We take $a_0(k) = 1$

by convention. Find a closed form for the exponential generating function

$$f_k(x) = \sum_{n=0}^{\infty} \frac{a_n(k)x^n}{n!}.$$

Solution by Jacob Boswell and Chip Curtis, Missouri Southern State University, Joplin, MO.

Let $\mathbb{N} = \{0, 1, 2, \ldots\}$. A permutation σ satisfies $\sigma^k = e$ if and only if all of its disjoint cycles have lengths which are factors of k. Let $k_1, k_2, \ldots, k_r$ be the distinct factors of k. We note that the number of permutations of jk objects that are a product of j k-cycles is given by $(jk)!/k^j j!$. Breaking permutations with $\sigma^k = e$ into a product having j_i k_i-cycles, we see that

$$\begin{aligned} a_n(k) &= \sum_{\substack{(j_i)\in\mathbb{N}^r \\ \sum j_i k_i = n}} \binom{n}{j_1k_1,\, j_2k_2, \ldots, j_rk_r} \frac{(j_1k_1)!}{k_1^{j_1} j_1!} \cdots \frac{(j_rk_r)!}{k_r^{j_r} j_r!} \\ &= \sum_{\substack{(j_i)\in\mathbb{N}^r \\ \sum j_i k_i = n}} \frac{n!}{k_1^{j_1} \cdots k_r^{j_r} \cdot j_1! \cdots j_r!}, \end{aligned}$$

where

$$\binom{n}{i_1, i_2, \ldots, i_r} = \frac{n!}{i_1! i_2! \cdots i_r!}$$

is a multinomial coefficient. Thus,

$$\begin{aligned} f_k(x) &= \sum_{n=0}^{\infty} \left(\sum_{\substack{(j_i)\in\mathbb{N}^r \\ \sum j_i k_i = n}} \frac{1}{k_1^{j_1} \cdots k_r^{j_r} j_1! \cdots j_r!} \right) x^n \\ &= \left(\sum_{j_1}^{\infty} \frac{x^{j_1 k_1}}{k_1^{j_1} j_1!} \right) \cdots \left(\sum_{j_r}^{\infty} \frac{x^{j_r k_r}}{k_r^{j_r} j_r!} \right) \\ &= \prod_{i=1}^{r} \exp\left(\frac{x^{k_i}}{k_i} \right) = \exp\left(\sum_{d|k} \frac{x^d}{d} \right). \end{aligned}$$

Editor's Note. Albert Stadler notes that this result appears in an old paper of Chowla, Herstein, and Scott: Chowla, S., Herstein, I. N., Scott, W. R. (1952). The solutions of $x^d = 1$ in symmetric groups. *Norske Vid. Selsk.* 25: 29–31.

Also solved by Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), CMC 328, Reiner Martin (Germany), José Heber Nieto (Venezuela), Michael Reid, and the proposer.

Answers

Solutions to the Quickies from page 574.

A1125. Let $m = 2^a 5^b C$, where a and b are nonnegative integers and $\gcd(C, 10) = 1$. Let d be the number of decimal digits in n. Since C and 10 are relatively prime, there exists a positive integer k such that

$$k10^d + 1 \equiv 0 \pmod{C}.$$

Letting $t = \max(a, b)$ and

$$x = \frac{k10^d + 1}{C} 2^{t-a} 5^{t-b} n$$

gives a solution to the problem.

For example, if $m = 1234$ and $n = 1111$, then $a = 1$, $b = 0$, $C = 617$, and $d = 4$. Solving $k10^4 + 1 \equiv 0 \pmod{617}$ gives $k = 429$ as one solution and hence

$$x = \frac{4290001}{617} \cdot 5 \cdot 1111 = 38623915.$$

Checking, we find that

$$38623915 \cdot 1234 = 47661911110,$$

as desired.

A1126. We have

$$PQ = \sqrt{1^2 + 3^2} = \sqrt{10}.$$

Since the sphere is tangent to all of the edges of the central cube, its diameter is the length of a face diagonal, so

$$ST = \sqrt{2}.$$

Therefore,

$$\frac{ST}{PS} = \frac{2ST}{PQ - ST} = \frac{2\sqrt{2}}{\sqrt{10} - \sqrt{2}} = \frac{\sqrt{5} + 1}{2}.$$

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

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

Bayer, Jonas, et al., Mathematical proof between generations, https://arxiv.org/abs/2207.04779.

Dierk Schleicher (Aix-Marseille Université) details the case of a 50-year-old "result" in dynamical systems whose proof is flawed, which cannot be fixed by the methods applied, and which remains an open question—yet many papers since have used the result or its flawed methodology. Other papers in this "collage" collection of essays reflect further on mathematical proof: Yuri Matiyasevich (Steklov Institute of Mathematics at St. Petersburg) reasserts that in 25 years mathematical journals will demand that papers be accompanied by proofs verified by a computer (his own proof of the undecidability of Hilbert's tenth problem was verified in a formalization by students of Schleicher, as related in another essay). Leslie Lamport (École Normale Supérieure de Paris) claims that one-third of all published mathematical papers contain significant errors and urges the use of hierarchically structured proofs. Christoph Benzmüller (Otto-Friedrich-Universität Bamberg) notes that ideally a proof should consist of both a human-oriented argument and a machine formal proof. Kevin Buzzard (Imperial College London) notes that "Computers certainly can't directly read the crap we write in our papers (and humans often can't read them either)" but urges making formalization "fun" by making it game-like. Lawrence Paulson (University of Cambridge) points out that current proof assistants still face obstacles in supporting mathematics.

Pitici, Mircea (ed.), *The Best Writing on Mathematics 2021*, Princeton University Press, 2022; xv + 287 pp + 16 pp color plates, $24.95(P). ISBN 978-0-691-22570-8.

Among the topics in this 12th volume of exemplary writing are accounts of "lockdown" mathematics (done by locked-up famous mathematicians—is isolation conducive to productivity?), 3D-printed models of dynamical systems, the dangers of "dark data," the bicycle paradox (push back the lower pedal—which way does the bicycle go?), and how arithmetic lessons should feature mathematical reasoning rather than just "getting the answer." Unfortunately, the publisher has discontinued this annual series. Despite there being 25K members in the MAA, 30K in the AMS, and the book being priced for the popular market, the volumes have sold only a few thousand each, mainly to libraries. Still, continuing the series would be a worthy project for the MAA, the AMS, or another publisher in furthering popular understanding of mathematics.

Nahin, Paul J., *In Pursuit of Zeta-3: The World's Most Mysterious Unsolved Math Problem*, Princeton University Press, 2022; xx + 320 pp, $26.95. ISBN 978-0-691-20607-3.

Is there a simple formula/expression for the case $k = 3$ of the sum $\zeta(k) = \sum_{n=1}^{\infty} \frac{1}{n^k}$? This book is a historical account of the mathematicians and their work involved in addressing that question, aimed at students who have studied calculus. However, the preface, which features some "tests" of the reader's abilities, may scare some readers away or discourage them unduly. The conundrum about $\zeta(3)$ is that Euler found that $\zeta(2n) = \frac{a}{b}\pi^{2n}$, with a and b integers depending on n, but there is no known formula for any $\zeta(2n+1)$. Analogy would suggest that $\zeta(3)$ might be a rational multiple of π^3, or perhaps a multiple by some other combination of certain numbers (e, $\ln 2$, etc.); but no such multiple has been found. The book includes *lots* of computations, as well as challenge problems for the reader (with solutions), some Chuck Norris(!) math jokes, and appendices with further details. The author's concluding advice is to be persistent and "slog away" at the problem.

Math. Mag. **95** (2022) 583–584. doi:10.1080/0025570X.2022.2126650 © Mathematical Association of America

Schutt, Randy, Probit and wealth inequality—how random events and the laws of probability are partially responsible for wealth inequality, *Chance* 35(1) (2022) 18–25.

Author Schutt uses coin flips and die rolls to model the vicissitudes of life for a population that starts with equal wealth, thereby demonstrating that subsequent wealth is eventually determined in part by random events—lucky breaks (higher salaries, windfalls, inheritances) or the opposite (natural catastrophes, personal disasters, illnesses). Under such models, the distribution of wealth tends toward the probit distribution, the quantile function for the standard normal distribution, i.e., the functional inverse of its cumulative distribution function. "These examples show that any society subject to random economic costs and benefits will develop some wealth inequality produced solely by the laws of probability." Just before his arrest, Christ quoted a verse from the Old Testament, "There will always be poor people in the land" (Deuteronomy 15:10) (but then reminded his disciples that they would not always have him). Some have taken that observation, and some could take Schutt's research, as an excuse not to try to address poverty. But Schutt concludes in a moral vein: "If a society values fairness, then it must, in some way, mitigate the wealth inequality caused by these natural events"—thereby echoing the verse following the one that Christ quoted: "Therefore I command you to be openhanded toward your brothers and toward the poor and needy in your land" (Deuteronomy 15:11).

Strogatz, Steven, with Colin Adams and Lisa Piccirillo, Untangling why knots are important, https://www.quantamagazine.org/printmc_cid=a3c2d0c9cf&mc_eid=942d0c52e3.

This lively 44-minute podcast (transcript available) from *Quanta Magazine* discusses knots, their invariants, and knotting and unknotting in higher-dimensional spaces, together with the solution of a 50-year-old problem involving the Conway knot. Other mathematics podcasts in the magazine's new "The Joy of Why" series, with Steven Strogatz as host, address whether computers can be mathematicians and how mathematicians know that proofs are correct.

Rosenthal, Jeffrey S., For kicks, we came up with a fairer way to determine the World Cup draw, https://www.thestar.com/opinion/contributors/2022/03/30/for-kicks-we-came-up-with-a-fairer-way-to-determine-the-world-cup-draw.htm.

By the time that you read this, the soccer World Cup group stage will be over, with the final match to be just before Christmas. Author Rosenthal reflects on the restrictions placed on assigning teams to the eight groups: "each group must include one team from each of four 'pots' (based on team rankings), no group can have more than two teams from Europe nor more than one from South America, etc." The assignment by FIFA is done by randomly selecting a team and placing it in the next available group that does not violate any restrictions. The result is that valid complete assignments are not equally likely. Rosenthal and Gareth Roberts have developed alternative assignment algorithms that generate equally-likely valid assignments.

Thiffeault, Jean-Luc, The mathematics of burger flipping, https://arxiv.org/abs/2206.13900.

OK, so you're cooking burgers. Should you flip them once, or more often? Author Thiffeault presents a model for cooking involving flipping; the rate of cooking depends on a linear operator and the fixed point of a map. Flipping repeatedly can reduce cooking time by about 30%, since doing so leaves little time for the top surface to cool off; flipping just twice reduces time by 8%. The author solves the heat equation for relevant values of the parameters and prescribes just when to do any flips.

Sibley, Thomas Q., *Thinking Algebraically: An Introduction to Abstract Algebra*, MAA Press, 2021; xiii + 478 pp, $85(P) ($63.75 MAA or AMS member). ISBN 978-1-4704-6030-3.

This textbook for a course in abstract algebra proceeds from realizing the utter disconnect between students' high school experience of algebra (manipulate symbols, solve equations) and the emphasis in abstract algebra on structures and their properties. The book starts by identifying properties of number systems familiar to students (including modular arithmetics), investigates mappings (isomorphism, homomorphism), introduces cyclic and abelian groups, and then explores rings. Further chapters feature vector spaces, Galois theory, and topics in group theory and ring theory (symmetry groups, Sylow theorems, lattices, Boolean algebras). There are abundant exercises, plus biographical sketches of many of the mathematicians involved. (Disclosure: Long ago, the author was a teaching assistant for me, and he has been a friend ever since.)