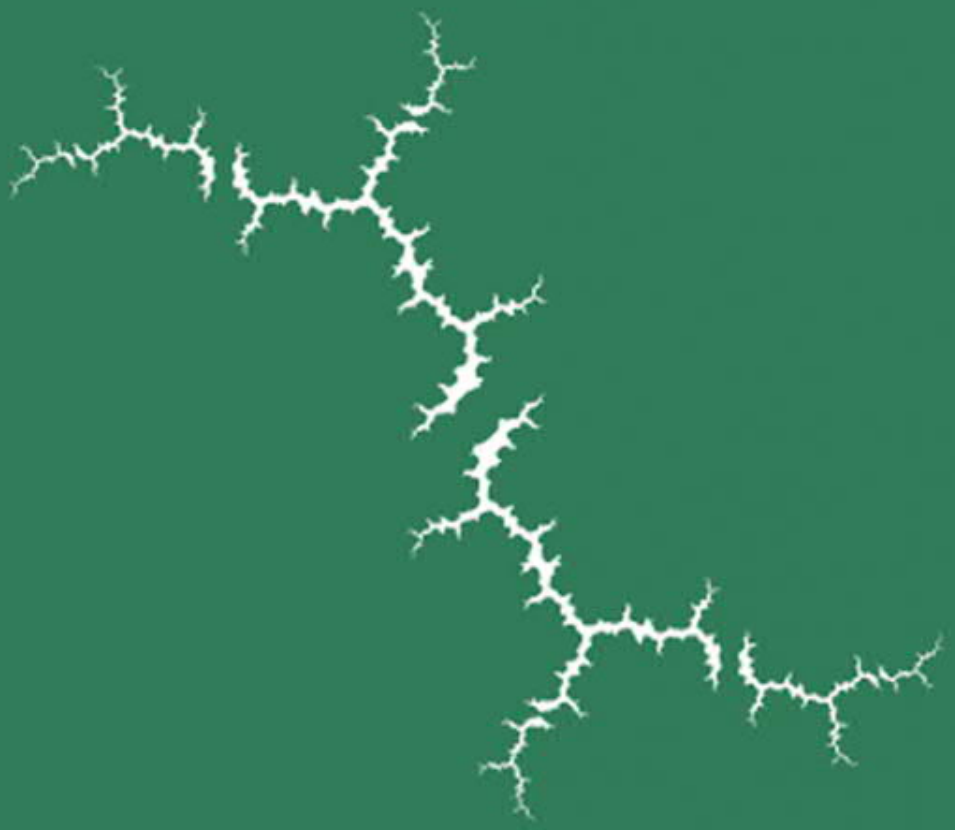
Vol. 92, No. 1, February 2019

MAA

MATHEMATICS MAGAZINE



- Julia sets–from slices of higher dimensional graphs
- Change of basis matrix for Laguerre polynomials is its own inverse
- TRIBUS: a puzzle for the 2019 issues of the MAGAZINE
- Should the definition of a ring include the multiplicative identity?

An Official Publication of the **Mathematical Association of America**

LETTER FROM THE EDITOR

Welcome to 2019! This issue begins with an article that considers complex dynamics and the projections of four-dimensional graphs to the plane. In the article, Julia Barnes and Lisbeth Schaubroeck examine the contour plots of the projected graphs and always see Julia sets.

The Laguerre polynomials, named from French mathematician Edmond Laguerre, form one of the classical families of orthogonal polynomials. Maureen Carroll and Elyn Rykken show that the transition matrix corresponding to the basis of Laguerre polynomials is its own inverse. Along the way, they provide a historical basis (pun intended), and the Gram–Schmidt process makes a cameo.

You may think of the mathematics of origami as being a more recent area of study. However, almost a century ago, Italian mathematician Margherita Piazolla Beloch studied an origami fold and its associated geometry. In the next article, Jorge Lucero analyzes the conditions of when Beloch's fold can occur.

In last year's February issue of THIS MAGAZINE, Arthur Befumo and Jonathan Lenchner determined conditions for when certain one-deficient three- and higher-dimensional boards could be tiled by L-trominoes. Befumo and Lenchner are back at it, having more fun with L-trominoes, tiling two-deficient boards.

Turning to geometry, John J. Wetzel and Wacharin Wichiramala determine a circular sector that contains a congruent copy of each unit arc. Their result is in the spirit of Leo Moser's "worm" problem, a conjecture that states that a circular sector with angle $\pi/6$ and radius 1 contains a congruent copy of each unit arc in the plane.

In the next article, Jürgen Grahl and Shahar Nevo demonstrate some surprising phenomena from elementary calculus related to oscillating functions. For example, they demonstrate that a continuously differentiable function with a strict minimum doesn't have to be decreasing to the left nor increasing to the right of the minimum. Besides comparing several definitions of inflection point, they provide other examples and counterexamples that will be useful to those teaching or learning calculus.

Should the definition of ring include the existence of a multiplicative identity? Bjorn Poonen argues that it should because it is part of what associativity should be. He considers the consequences of requiring the identity and addresses counterarguments.

A well-known problem in graph theory is to enumerate the number of spanning trees of a graph. What about adding one more edge so that the spanning subgraph has one cycle? Jacob Siehler counts such unicycle structures using the matrix tree theorem.

In 2019, every issue of THE MAGAZINE will have a TRIBUS PUZZLE, created by David Nacin. Try an easy one in this issue. Harder puzzles will appear later in the year. Brittany Shelton and Andrés Caicedo selected the puzzle among a number of submissions. Look for their introduction to the puzzle in this issue.

Interspersed throughout the issue are three proofs without words, one each by Ángel Plaza, Adrian Chunpong Chu, and Cherng-tiao Perng. Also, Poo-Sung Park provides an easy proof of the irrationality of $\sqrt{2}$. The issue wraps with the Problems and Reviews sections. Let me offer an editor's note: look for a correction of a typo to Problem 2052 at the beginning of the Problems section.

Finally, let me share some big news. Jason Rosenhouse is now editor-elect for the MAGAZINE. Jason is a Professor of Mathematics at James Madison University. You may know him for his stint as the Book Review editor at *The American Mathematical Monthly* or as the co-author with Laura Taalman of *Taking Sudoku Seriously*, or as the co-editor with Jennifer Beineke of the collections *The Mathematics of Various Entertaining Subjects*, Volumes I-III. To learn more about Jason's vision for the MAGAZINE, check out Jacqueline Jensen-Vallin's interview with Jason in the December 2018/January 2019 issue of MAA Focus (Rosenhouse Welcomed as Editor-elect of Mathematics Magazine, 38 (6): 15). Please join me in welcoming Jason to the MAGAZINE family.

Michael A. Jones, Editor

Any Way You Slice It, It Comes Up Julia Sets

JULIA A. BARNES
Western Carolina University
jbarnes@email.wcu.edu

LISBETH E. SCHAUBROECK
United States Air Force Academy
beth.schaubroeck@usafa.edu

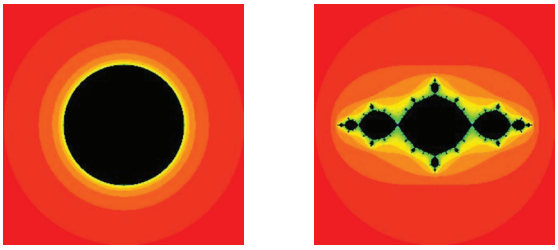
In "Emerging Julia Sets" from THIS MAGAZINE [1], along with co-authors, we considered functions that are obtained by composing certain types of complex functions with themselves many times. The real and imaginary parts of these functions are generated, level 2 and -2 contours of these surfaces are graphed, and Filled Julia Sets surprisingly appear in the middle of the contour plots uncrossed by the contours. As we compose the function with itself more times, the level 2 and -2 contours surround the Julia Set more closely, marking places where the function oscillates steeply around the center. Aside from indicating where the function oscillates, what do these contours signify? What if, instead, we looked at contours at other heights? Additionally, what if the surface itself were changed by projecting onto a line other than the real or imaginary axis? As it turns out, the answers to all of these questions are all related to simple straight lines.

Background definitions

In this paper, we focus our attention on polynomial functions of one complex variable, with all examples provided coming from the family $f_c(z) = z^2 + c$. We consider graphs of functions obtained by repeated composition of a function $f(z)$ with itself. This process is called *iteration* where $f^2(z) = f(f(z))$, $f^3(z) = f(f(f(z)))$, and so on. Thus, $f^n(z)$ denotes $f(z)$ composed with itself n times. The set $f^{-1}(z)$, called the *preimage* of z under $f(z)$, is the collection of points that map to z. Note that $f^{-1}(z)$ is not to be read as an inverse function because there will typically be more than one preimage for a given point. In addition, the set $f^{-n}(z)$ consists of all of the preimages of z under the function $f^n(z)$. For any point $z \in \mathbb{C}$, the set $\{z, f(z), f^2(z), \ldots, f^n(z), \ldots\}$ is called the *forward orbit* of z, while the set $\{z, f^{-1}(z), f^{-2}(z), \ldots, f^{-n}(z), \ldots\}$ is called the *backward orbit* of z. We will be particularly interested in when these sets are bounded or not; a set S is called *bounded* if there is an $M \in \mathbb{R}$ such that for every $s \in S$, $|s| < M$.

The *Filled Julia Set*, $K(f)$, consists of all points whose forward orbit is a bounded set. Consequently, the *complement of the Filled Julia Set*, the set of points not in the Filled Julia Set, consists of the points whose forward orbit is not a bounded set. Points in the complement of the Filled Julia Set iterate toward infinity. The boundary, or edge, of the Filled Julia Set is called the *Julia Set*, $J(f)$. Two common Filled Julia Sets, those

Math. Mag. **92** (2019) 3–16. doi:10.1080/0025570X.2019.1538715
MSC: Primary 37F50, Secondary 37F10; 30D05
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.



$f_0(z) = z^2$ $f_{-1}(z) = z^2 - 1$

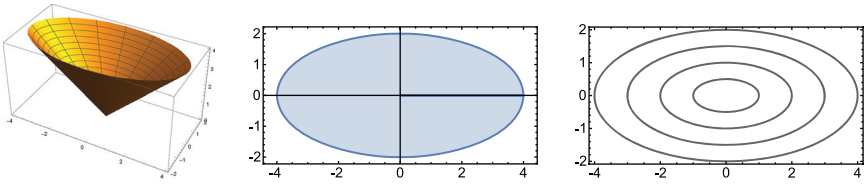
Figure 1 The Filled Julia Sets for two complex polynomials.

for $f_0(z) = z^2$ and $f_{-1}(z) = z^2 - 1$, are shown in Figure 1; in each, the Filled Julia Set is black and the Julia Set is the boundary of the Filled Julia Set. Also, the Filled Julia Set is an *invariant set* under iteration, meaning that $f^n(K(f)) = K(f) = f^{-n}(K(f))$ for all $n \in \mathbb{N}$. This is because the definition of $K(f)$ relies on the long-term forward behavior of points under iteration, which does not change as you iterate f forward or backward. For a point $z \in K(f)$, we know that the forward orbit is a bounded set. It follows that the forward orbits of $f(z)$ and $f^{-1}(z)$ must still be bounded. Similarly, the complement of the Filled Julia Set and the Julia Set are both invariant under iteration.

Polynomial functions, $f(z)$, are all complex *analytic*, meaning that at every point z, the derivative $f'(z)$ exists at z as well as on any small disc containing z. Also, the standard calculus definition for *critical points* still holds: z_0 is a critical point for $f(z)$ if and only if $f'(z_0)$ is 0. To find the critical points for $f^n(z)$, we simply use the chain rule. A point z_0 is a critical point for $f^n(z)$ if and only if $(f^n(z_0))' = f'(f^{n-1}(z_0)) \cdot f'(f^{n-2}(z_0)) \cdots f'(z_0) = 0$. This means that the critical points for $f^n(z)$ are the critical points for $f(z)$, along with the first $n-1$ preimages of the critical points for $f(z)$. Specifically, the critical points for $f^n(z)$ are $\{z_0, f^{-1}(z_0), \ldots, f^{-(n-1)}(z_0) : z_0$ is a critical point for $f(z)\}$.

Projections

Imagine that you are holding an elliptical cone as in Figure 2a; the cone can be described by the points (x, y, z) such that $z = \sqrt{x^2 + 4y^2}$. What if you wanted to represent this cone in two dimensions instead? You have several options. You could look at the cone from the positive z-axis and project it down to a screen on the xy-plane as in Figure 2b. This is the plot showing only the contour at a height of 4; a more standard contour plot with intermediate curves represented is shown in Figure 2c. For another vantage point, you could stand on the negative y-axis and look at the cone from there and you would see the shadow of the object onto a screen placed at $y = 0$.



(a) (b) (c)

Figure 2 (a) A cone, (b) projection of the cone onto the xy-plane, and (c) its contour diagram.

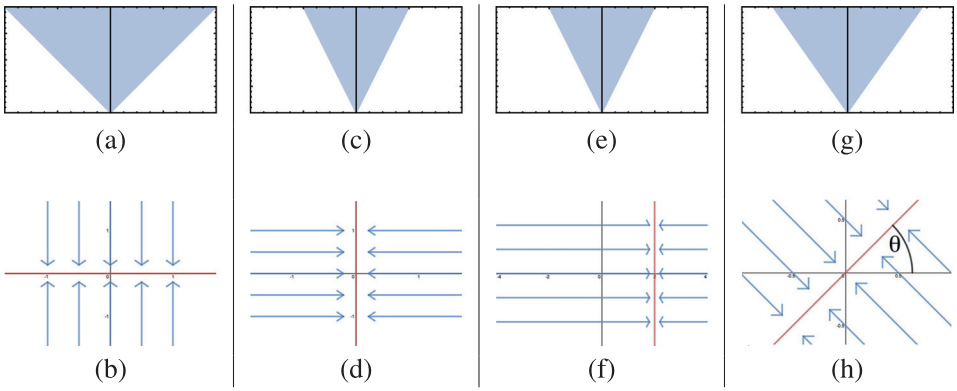


Figure 3 (a) The image of the cone projected onto the xz-plane and (b) a diagram indicating the direction in which the cone was projected; (c) the projection of the cone onto the yz-plane with (d) corresponding diagram of the direction of projection; (e) the projection of the cone onto a screen placed at $x = 2$ and (f) its corresponding diagram; (g) the projection onto a screen placed at the line through the origin at an angle $\theta = \pi/4$ with (h) its corresponding diagram.

This is the projection onto the xz-plane, as seen in Figure 3a. The diagram in Figure 3b indicates the direction in which the cone is projected onto the screen. Similarly, you could stand on the positive x-axis and look at the cone from there. This would be the projection onto the yz-plane, or the shadow of the object onto a screen placed at $x = 0$ as shown in Figure 3c; again Figure 3d indicates the direction of the projection onto this new screen. What if we move further over on the positive x-axis and project to the plane $x = 2$; that is, we move the screen from $x = 0$ to $x = 2$ as seen in Figure 3f? Does this change the image any? No, because the shadow does not change, as seen in Figure 3e. Now, imagine that screen were placed back at $y = 0$, and then rotated by $\pi/4$ radians; you are standing directly in front of the screen from a distance of 2 units from the origin along $y = -x$ as seen in Figure 3h. The projected image is shown in Figure 3g. This is a very different projection of the cone from the ones described above, although for this example, the image looks similar.

What if we use this same idea to bring the image of the four-dimensional graph of a complex function down to three dimensions? One natural way to do that is to compute the real part or imaginary part of the function. Recall that complex functions can be written as $f : \mathbb{C} \to \mathbb{C}$ where $f(x + iy) = u(x, y) + iv(x, y)$; the functions $u(x, y)$ and $v(x, y)$ are the *real part* and *imaginary part* of $f(x + iy)$, respectively. Both $u(x, y)$ and $v(x, y)$ are real valued and would have graphs that are three-dimensional surfaces. These surfaces are analogous to the projections of the cone in Figures 3a and c, as well as the surfaces explored [**1**] and [**2**]. In these papers, instead of looking at the three-dimensional surfaces, we and our co-authors looked at contour diagrams of the projections. For example, in Figures 4a, c, and e, we see the surfaces of the real parts of $f_0(z) = z^2$, $f_0^2(z)$, and $f_0^3(z)$. Notice that these surfaces are wavy and don't have any maximum or minimum values because all of the critical points are saddle points [**1**]. Figures 4b, d, and f show the level 2 contours for the same surfaces. These are the cross sections obtained if we were to slice the surfaces at a height of 2.

Mapping functions forward and backward

One thing we do when we explore complex functions is look at where points map and what points map to them. For example, what are the preimages of 2 via the function

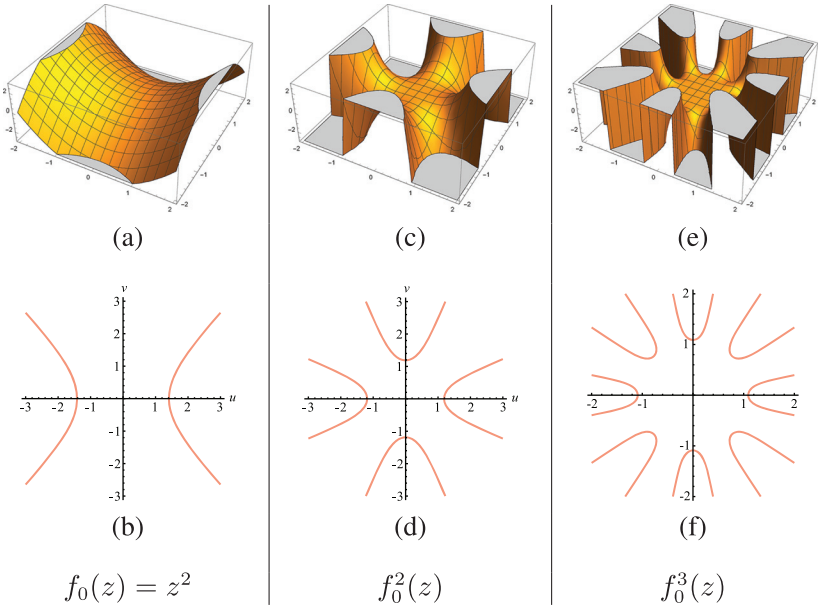


Figure 4 Surfaces of the real parts of iterates and their corresponding level 2 contours for (a)–(b) $f_0(z) = z^2$; (c)–(d) $f_0^2(z)$; (e)–(f) $f_0^3(z)$.

$f_0(z)$, i.e., what is the full set of solutions to $z^2 = 2$? This is a question from high school algebra; both $\sqrt{2}$ and $-\sqrt{2}$ map to 2. Now, what are the preimages of $\sqrt{2}$ and $-\sqrt{2}$? Solving $z^2 = \sqrt{2}$ is still fairly easy, giving us $2^{1/4}$ and $-2^{1/4}$. We need complex numbers to find the rest of the preimages, obtaining $2^{1/4}i$ and $-2^{1/4}i$. Note that all of $\{2^{1/4}, -2^{1/4}, 2^{1/4}i, -2^{1/4}i\}$ map to 2 via $f_0^2(z)$.

What if we continue looking for preimages of 2 under $f_0^n(z)$ as n approaches infinity? We can plot these preimages as seen in Figure 5. Notice that even for small values of n, the preimages appear to be approaching the unit circle. Indeed, the preimages of 2 are relatively straight forward to compute and are the complex roots of unity times $2^{1/2^n}$ which approach the unit circle as n goes to infinity.

What if we do the same thing with another function, such as $f_{-1}(z) = z^2 - 1$ seen in Figure 6? Now the values of the preimages are hard to compute by hand, but we are able to use a computer algebra system to plot these points. This time, the points are not approaching the unit circle, but a different familiar shape. They appear to be approaching the Julia Set for $f_{-1}(z)$. Since the Julia Set for $f_0(z)$ is the unit circle, this same behavior occurred with $f_0(z)$ as well. In fact, this is always true for any function $f_c(z)$; the preimage of 2, or any other number outside of the Filled Julia Set, will approach $J(f_c)$, which follows from a corollary to a theorem found in [**5**, p. 71] as stated in Theorem 1.

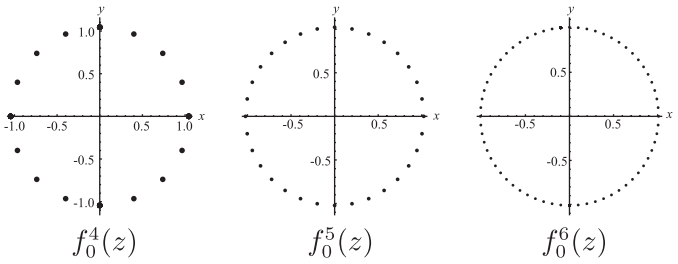


Figure 5 Preimages of 2 under several iterates of $f_0(z) = z^2$.

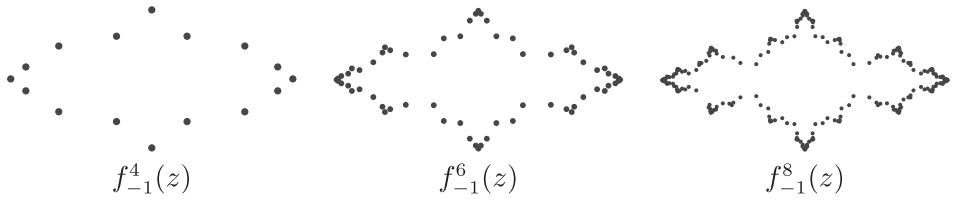


Figure 6 Preimages of 2 under several iterates of $f_{-1}(z) = z^2 - 1$.

Theorem 1. Let $f(z)$ be a complex polynomial of degree at least 2 and let $z_0 \in \mathbb{C}$ and $z_0 \notin J(f)$. Then the following two statements are true about the preimages, $f^{-n}(z_0)$.

a) Let $B = \cup_{n=1}^{\infty} f^{-n}(z_0)$. If B is infinite, then $J(f)$ is contained in the closure of B. Additionally, $J(f)$ does not have any elements in B; therefore, all elements of $J(f)$ are limit points of B. That is, for every element of $J(f)$, there is a sequence of points in B that converge to it.

b) Let $U_\varepsilon = \{z \in \mathbb{C} : |z - w| < \varepsilon \text{ for some } w \in J(f)\}$. If for all $z \notin J(f)$, $f^n(z)$ does not have a subsequence of points that converge to z_0, then there is some $N > 0$ such that $f^{-n}(z_0) \subset U_\varepsilon$ for all $n > N$.

In Part a) of Theorem 1, the condition that B be infinite is not terribly restrictive. For polynomials, there is at most one complex number z_0 where B could be finite. For example, with $f_0(z) = z^2$, the number 0 is the only such point [**5**, p. 65].

Part b) of Theorem 1 looks at the condition that for all $z \notin J(f)$, $f^n(z)$ does not have a subsequence of points that converges to z_0. This is saying that there are no points in the complement of the Julia Set such that $f^n(z)$ *accumulates* at z_0. For example, consider the Julia Set for $f_0(z)$, the unit circle. All points outside of the unit circle iterate to infinity while all points inside the unit circle iterate to 0. Therefore, $z_0 = 0$ is the only complex number for which there exists $z \notin J(f_0)$ where $f_0^n(z)$ accumulates at z_0. For another example, notice that when iterating $f_{-1}(z) = z^2 - 1$, $0 \to -1 \to 0 \to -1 \to \cdots$ creating a *cycle* with 2 elements. All points $z \notin J(f_{-1})$ either iterate to infinity or they approach the cycle $0 \to -1 \to 0 \to -1 \to \cdots$ [**5**, Sect. 1.6]. In this case, the only complex numbers z_0 with points $z \notin J(f_{-1})$ accumulating to z_0 are -1 and 0. Similar behavior occurs for all other functions in the family $f_c(z)$, with points iterating either to infinity or to a finite cycle.

If the conditions of both parts of Theorem 1 are met, the set $f^{-n}(z_0)$ approximates the Julia Set, and as n approaches infinity, the approximation becomes better and better. This is the basis behind the Random Backwards Iteration Algorithm or versions of the Inverse Iteration Method often used to generate images of Julia Sets without first looking at a Filled Julia Set [**4**, **6**, **9**, **10**]. For our examples in Figures 5 and 6, observe that the conditions of Theorem 1 hold for $z_0 = 2$, so these preimages approximate the Julia Set.

So far, we have been looking at the behavior of preimages of individual points. What if we look instead at images and preimages of curves? Consider $f_0(z) = z^2$ again. If we think of z as a number in polar form, then squaring z will square its radius and double its angle. If we look at the portion of the unit circle in the first quadrant as shown in the leftmost part of Figure 7 and square it, all angles double and the image is the top half of the unit circle, as in the center of Figure 7. Continuing in this fashion, squaring the top half of the unit circle will produce the whole unit circle, as in the rightmost part of Figure 7. Therefore, the quarter circle (and actually any quarter circle on the unit circle) maps to the whole unit circle via $f_0^2(z)$, and the quarter circle is part of the preimage of the unit circle via $f_0^2(z)$.

Let's look at a different shape. A common question in a standard complex variables course is to find the preimage of a vertical line, such as $x = 2$. The preimages of the

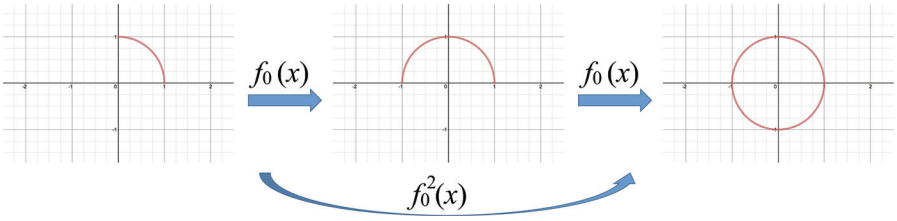


Figure 7 Mapping portions of a circle under the function $f_0(z) = z^2$.

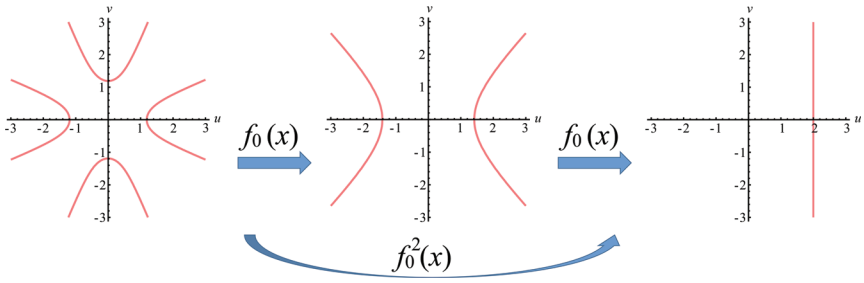


Figure 8 First and second preimages of $x = 2$ under the function $f_0(z) = z^2$.

line $x = 2$ are the two parabolas shown in the center of Figure 8. Then, we can use any computer algebra system to generate the preimage for each of the two parabolas as seen in the leftmost part of Figure 8. Note that these preimages will then map to $x = 2$ via $f_0^2(z)$. But wait! These images in Figure 8 appear to be the same images we saw in Figure 4. As it turns out, for any polynomial $f(z)$, the level 2 contours of $\mathrm{Re}(f^n(z))$ are precisely the preimages of $x = 2$ under $f^n(z)$!

How do we know that the preimages of $x = 2$ are the contour lines in question? Let's consider what the level 2 contours for $\mathrm{Re}(f(z))$ represent. In multivariable calculus, the level 2 contours of a function are the collection of all points in the domain where the surface has a height of 2, or equivalently, where the value of the function is 2. As we consider surfaces whose domain is $\mathbb{C}$, the level 2 contours for $\mathrm{Re}(f(z))$ are still the points in the domain where $\mathrm{Re}(f(z))$ is 2, or equivalently the points z such that $f(z)$ lies on the line $x = 2$. It follows that the level 2 contour is the preimage of $x = 2$ via $f(z)$. Since this is true for any polynomial, it is true for any iterate $f^n(z)$ as well; that is, the level 2 contours for the real part of $f^n(z)$ are the preimages of the line $x = 2$ via $f^n(z)$.

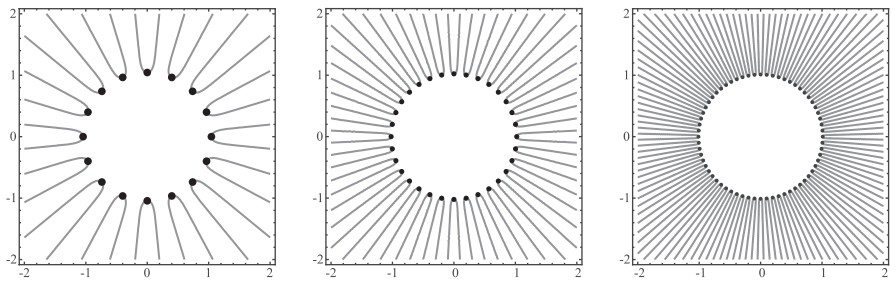


Figure 9 Level 2 contours of the real part of $f_0^n(z)$, along with the preimages of 2, for several values of n.

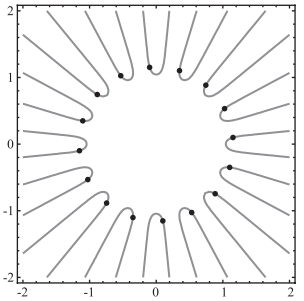


Figure 10 The preimages of the point $z = 2 + 10i$ along with the level 2 contours under $f_0^4(z)$.

If we plot preimages of $x = 2$ under higher iterates as in Figure 9, we begin to see the Julia Set in the center. The dots on these contours are the preimages of the number 2. Therefore, the appearance of the Julia Set in the middle follows naturally. Note that it isn't just the preimages of 2 that approximate the Julia Set; actually every point along $x = 2$ satisfies the hypotheses of Theorem 1, so the preimages of every point will eventually approximate the Julia Set. For example, Figure 10 shows the preimages of $2 + 10i$ along with the preimages of $x = 2$ which still converge to the Julia Set even though they do not lie on the bend of the contour lines. Furthermore, the line $x = 2$ is completely outside of the Filled Julia Set for any function $f_c(z)$ in the family [7]. Since the complement of the Filled Julia Set is invariant under iteration, the preimages of $x = 2$ will also remain outside of the Filled Julia Set. Thus, these contour lines will never actually intersect the Julia Set; they just get arbitrarily close to it. This is why we see the Julia Set appearing in the center. Figure 11 shows the preimages of $x = 2$ via $f_{-1}(z)$; notice that the contours behave much like those in Figure 9.

Even though we see the Julia Set in the center of all images in Figures 9–11 there are still unbounded rays outside of the Filled Julia Set. At first glance, this may seem to contradict the fact that all points on the line $x = 2$ will eventually approach the

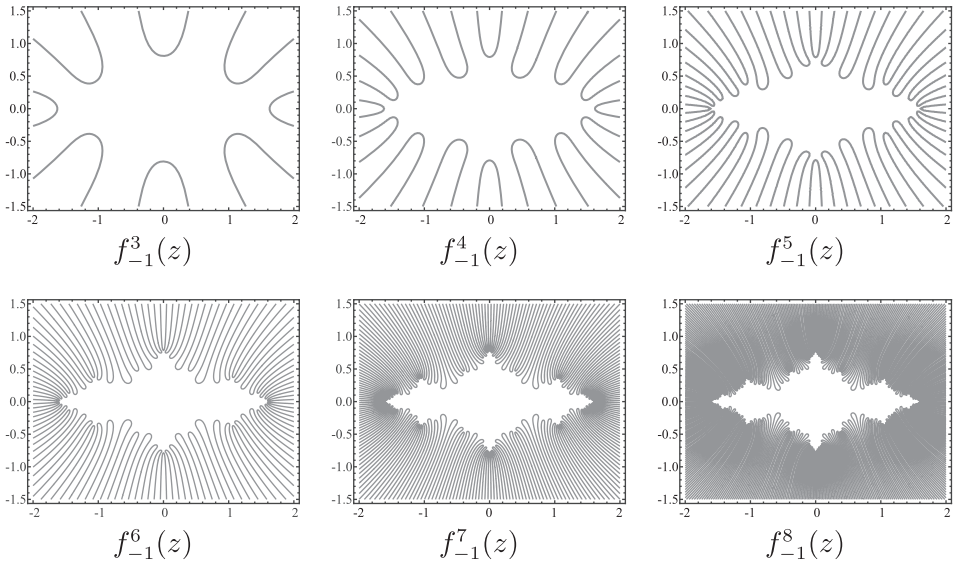


Figure 11 Level 2 contours of the real part of $f_{-1}^n(z)$ for several values of n as the plots begin to resemble the Julia Set in Figure 1.

Julia Set as we plot more and more preimages. It is true that if we choose any fixed point on $x = 2$, even with a very large imaginary part, its preimages will eventually approach the Julia Set. However, if we consider the set of points $\{2, 2 + i, 2 + 2i, \ldots, 2 + ni, \ldots\}$ and look at the preimages of that set, there will always be points arbitrarily far from the origin; this guarantees the existence of rays that reach out to infinity. This demonstrates the mysterious nature of infinity.

What if we want to plot preimages of another vertical line $x = a$ with $a \neq 2$? If the line remains outside of the Filled Julia Set, the same behavior will occur. The preimages of the real number a will approach the Julia Set and the corresponding line will bend and approach the Julia Set. The only difference is that convergence happens at a different rate because one starting line is further from the origin than the other.

What if $x = a$ intersects with the Filled Julia Set? By the invariance under iteration of the Filled Julia Set, its complement, and the Julia Set, the portions of $x = a$ in each of these three sets remain in these sets. Suppose every point $z_0 \notin J(f_c)$ on the line $x = a$ satisfies the conditions of Theorem 1. That is, the set of preimages of z_0 is infinite, and there are no points in the complement of the Julia Set that accumulate to z_0 under iteration. For example, in Figure 12, we see the preimages of $x = 0.01$ via the function $f_0(z) = z^2$. This resembles the preimages we have been seeing; curves bend in the middle with rays going to infinity, while the centers approach the Filled Julia Set. What is different is that some portions of the curve now lie inside the Filled Julia Set; the points inside the Filled Julia Set will still converge to the Julia Set because of Theorem 1, but convergence to the Julia Set comes from within the Filled Julia Set.

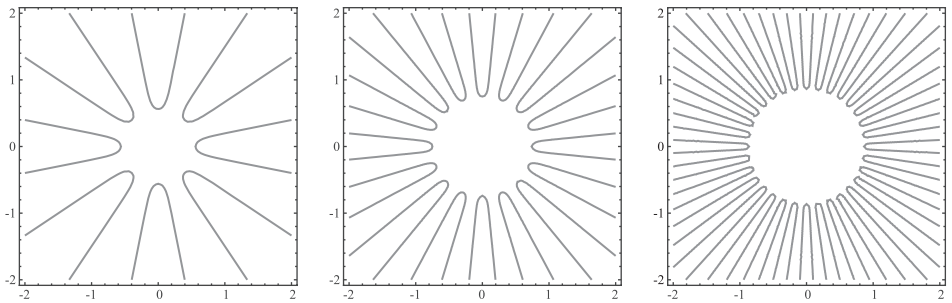


Figure 12 Level 0.01 contours of the real part of the function $f_0^n(z)$ for several values of n.

Not every line $x = a$ that intersects the Filled Julia Set has preimages like those in Figure 12. Consider the images in Figure 13; the preimages for $x = 0$ under $f_{-1}^n(z)$ for various values of n. As the number of iterates increase, the curves around the edges of the plot do still bend near the Julia Set with rays reaching out to infinity. However, there are extra features on the inside.

Recall that some points from the Filled Julia Set for $f_{-1}(z) = z^2 - 1$ accumulate at 0 and -1 under iteration; in fact, there are an infinite number of points that map to 0 and -1 under iteration as indicated in the tree diagram in Figure 14. In addition, there are points near each number on this tree diagram that will iterate toward the cycle of 0 and -1. Therefore, the hypothesis of Theorem 1b is not met at the points $z_0 = 0$ and $z_0 = -1$. This produces some extra points in the contour plot that are not going to approach the Julia Set.

Notice that the level 0 contour lines in Figure 13 now have crossings in the middle. These crossings are at some of the critical points for $f_{-1}^n(z)$ and display the classic

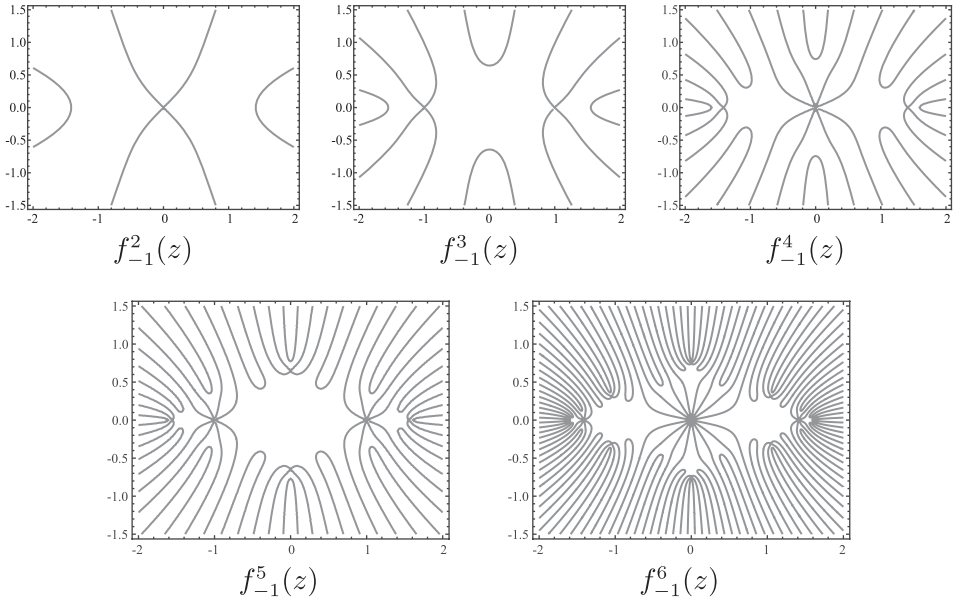


Figure 13 Level 0 contours of the real part of $f^n_{-1}(z)$.

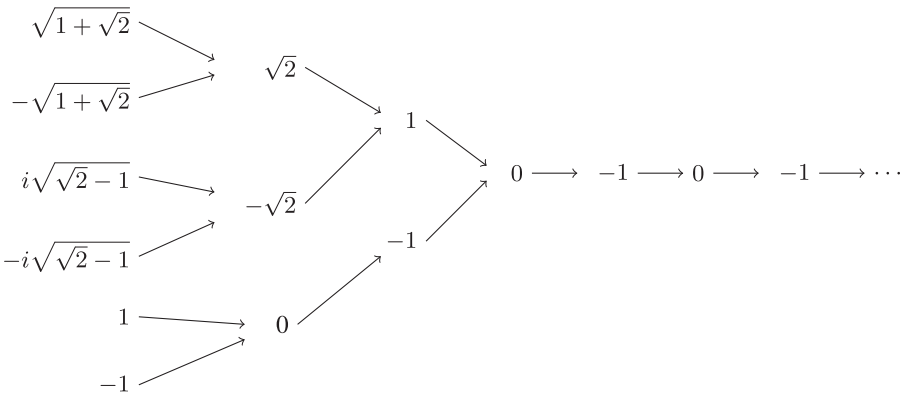


Figure 14 Part of the forward and backward orbit of 0 via the function $f_{-1}(z)$ – note that this diagram expands to the left, picking up an infinite number of additional preimages.

contour shape for saddle points. Recall that the critical points for any polynomial function $f^n(z)$ are the critical points of $f(z)$ along with the first $n-1$ preimages of the critical points for $f^n(z)$. For the function $f_{-1}(z)$ and its first few iterates, we obtain the following critical points, as seen below.

Function	Critical Points
$f_{-1}(z)$	0
$f^2_{-1}(z)$	$0, 1, -1$
$f^3_{-1}(z)$	$0, 1, -1, \sqrt{2}, -\sqrt{2}$
$f^4_{-1}(z)$	$0, 1, -1, \sqrt{2}, -\sqrt{2}, \sqrt{1+\sqrt{2}}, -\sqrt{1+\sqrt{2}}, i\sqrt{\sqrt{2}-1}, -i\sqrt{\sqrt{2}-1}$

As we compare the critical points to the contour plots in Figure 13, we see that not all critical points appear in the level 0 contours. This is because we only see critical

points whose value is 0 after n iterations. For $f_{-1}^2(z)$, the only critical point we see is 0 because both 1 and -1 map to -1 after 2 iterations. For $f_{-1}^3(z)$, the only critical points that maps to 0 after 3 iterations are 1 and -1. The other critical points map to -1 under 3 iterations. For $f_{-1}^4(z)$, the critical points we see are the ones that map to 0 under 4 iterations; they are 0, $\sqrt{2}$, and $-\sqrt{2}$. The rest of the critical points map to -1. We are still beginning to see the Julia Set appearing, but we will always have extra decorations in the middle; in particular, we will always have somewhat of a flower shape at 0 for plots corresponding to any even iterate.

Recall that -1 also fails to meet the criteria for Theorem 1b. If we plot the preimages of $x = -1$ we see images similar to those in Figure 13. The difference is that the critical points that appear in those contour plots are the ones that map to -1 under n iterations instead of the ones mapping to 0 under n iterations.

Other perspectives

So far, we have only considered taking a complex function and projecting the output onto the real axis, creating a surface in three-space. Suppose we instead project to a line, ℓ_θ, that is at an angle of θ from the real axis as shown in Figure 3h for the cone image. How will the resulting surface differ from when we projected onto the real axis? Notice that ℓ_θ is simply a rotation of the real axis by θ. To simplify the computations for plotting the surface resulting from projecting the image of a function onto ℓ_θ, we compose the output of the function with a rotation and then take the real part. That is, for a function $f = u + iv$, we evaluate $\text{Re}(e^{-i\theta}(u + iv)) = \text{Re}((\cos(\theta) - i\sin(\theta))(u + iv)) = \cos(\theta)u + \sin(\theta)v$. Note that if $\theta = 0$, this is the real part of $f(z)$, and if $\theta = \pi/2$, this is the imaginary part of $f(z)$.

This new surface has the same critical points as the surface obtained by computing the real part. Why? This new surface is $\text{Re}(e^{-i\theta} f(z))$. It was shown in [**1**] that the critical points of an analytic function are the same as the critical points for the real part. Therefore, the critical points for this new surface are the places where $g'(z) = 0$ for $g(z) = e^{-i\theta} f(z)$. But, $g'(z) = e^{-i\theta} f'(z)$, so the critical points of $f(z)$ and the critical points of $g(z)$ are the same. Since $g(z)$ is a polynomial, there are a finite number of critical points, hence they are isolated. These isolated critical points can be classified as maximums, minimums, or saddle points. Since $f(z)$ is analytic, $g(z)$ is as well. Also, the real part of any analytic function must satisfy the *maximum modulus principle* [**8**, **11**], which states that if we look at the real part of the function on any disc, then the maximum or minimum must occur along the boundary. Therefore, taking a disc around any critical point that is small enough to exclude all other critical points, we see that the critical point cannot be a maximum or minimum. Therefore, all critical points for this new surface are still saddle points, and we should expect the surface to be wavy with no maximum or minimum values.

Consider the surface obtained from projecting $f_{-1}^3(z)$ onto the line at an angle of $\pi/4$ from the real axis. This surface is shown in Figure 15a and its corresponding level 2 contours in Figure 15b. These diagrams seem quite similar to the surface and level 2 contours when we project the same function onto the real axis, as seen in Figures 15c and d, respectively. Are they the same? To investigate, let's look at the level 2 contours for both projections on the same axes as seen in Figure 16. This plot indicates that they are not the same. What mathematical concepts are being illustrated in these images?

When we project to the real axis and plot level 2 contours, we are simply looking at the preimages of the line $x = 2$. When we project to the line $\ell_{\pi/4}$ and plot level 2 contours, we are looking at the preimages of a line perpendicular to $\ell_{\pi/4}$ whose distance

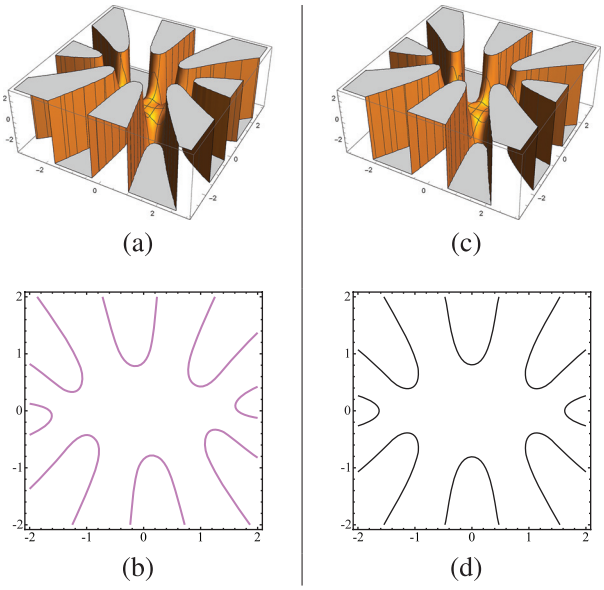


Figure 15 (a) The surface of $f_{-1}^3(z)$ projected onto $\ell_{\pi/4}$ along with (b) its level 2 contours; (c) the surface of $f_{-1}^3(z)$ projected onto the real axis, along with (d) its level 2 contours.

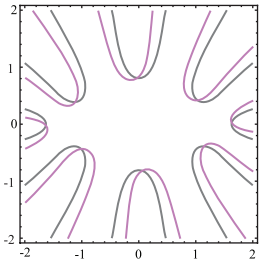


Figure 16 Overlay of the level 2 contours from Figure 15.

from the origin is 2, as depicted in Figure 17. Notice that the line perpendicular to $\ell_{\pi/4}$ in the diagram and $x = 2$ meet at an angle of $\pi/4$. Since the function we are considering is analytic, it preserves angles [**8**, **11**]; the contours from the two different projections in Figure 15 also meet at angles of $\pi/4$. The same idea holds for any angle θ, including the case where $\theta = \pi/2$, creating the projection onto the imaginary axis.

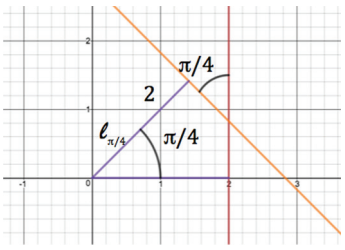
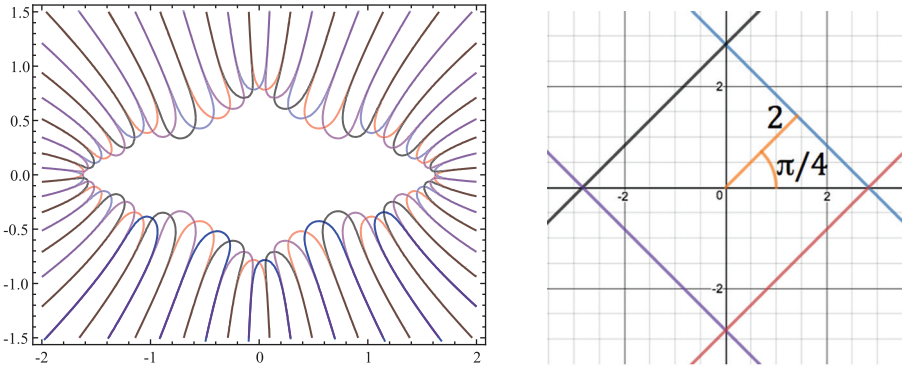


Figure 17 The line $\ell_{\pi/4}$ and the line perpendicular to $\ell_{\pi/4}$ whose distance from the origin is 2.



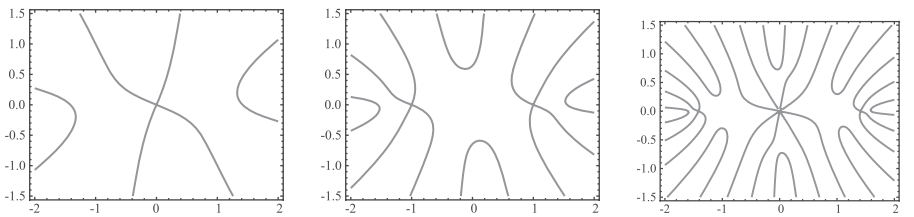
(a) (b)

Figure 18 (a) Level 2 contours of the surface obtained by projecting $f_{-1}^4(z)$ onto $\ell_{\pi/4}, \ell_{3\pi/4}, \ell_{5\pi/4}, \ell_{7\pi/4}$, and (b) the projection lines corresponding to the level contours in (a).

This illustrates the well-known result that the contours of the real and imaginary part of an analytic function meet at right angles [**8**, **11**]. For more images comparing projections onto the real and imaginary axes, see [**1**] and [**2**].

Recall that every complex number that is not in the Filled Julia Set has the property that its preimages approach the Julia Set according to Theorem 1. If we rotate the line $x = 2$ about the origin by any angle θ, the resulting line will remain outside of the Filled Julia Set for $f_{-1}(z)$ [**6**]. Consequently, the level 2 contours for a projection onto ℓ_θ will approach the Julia Set in the same fashion as they did for the real projection. Figure 18a shows the level 2 contours for $f_{-1}^4(z)$ using projections from four different angles; all of these contours bend and approach the Julia Set. Note that the contours in Figure 18a are the preimages of the lines shown in Figure 18b.

We have been looking at projecting onto a line ℓ_θ and creating level 2 contours; these contours came from rotating the line $x = 2$ by $-\theta$ about the origin and then looking at the preimages of that line. What happens if instead we start with $x = a$ for $a \neq 2$, rotate it by $-\theta$ about the origin, and create its new preimages? The behavior described for preimages of $x = a$ above is similar and based on whether the rotated image of $x = a$ satisfies the hypotheses of Theorem 1 or not. If the rotated line satisfies the hypotheses of Theorem 1, then the images of the associated contour plots appear similar to those in Figures 9 and 11. If the hypothesis of Theorem 1b is not met, then we obtain results similar to Figure 13, with the same critical points as if $x = a$ were not rotated. However, the contour lines are rotated depending on the angle θ. See Figure 19 for an example of this situation. Note that the contours have to twist to connect to the same critical points we saw in the earlier figures.



(a) (b) (c)

Figure 19 Projecting onto $\ell_{\pi/4}$, the level 0 contours of (a) $f_{-1}^2(z)$, (b) $f_{-1}^3(z)$, and (c) $f_{-1}^4(z)$.

What if we project onto a line that does not go through the origin? We can, but this does not add anything new. Why? This is like the cone example in Figure 3e when we moved the screen to a new location parallel to the original. It does not change the image at all. If we project onto a line parallel to ℓ_θ and find the line perpendicular to that line which is a distance of 2 from the origin, we would get precisely the same line as if we found a line perpendicular to ℓ_θ at a distance of 2 from the origin. Therefore, we would obtain exactly the same level 2 contours as when we projected to ℓ_θ.

External rays

For people already familiar with complex dynamics, the images in this paper resemble pictures of external rays. If we look at points in the region outside of the Filled Julia Set in any of our images, these points all map toward infinity under iteration. In fact, as points approach infinity, their behavior under iteration is much like that of z^d where d is the degree of the polynomial. There is a conformal (angle preserving) function, $\phi(z)$, called the Böettcher function that maps the complement of the Filled Julia Set to the region outside the unit circle and satisfies $z^d = \phi(f(\phi^{-1}(z)))$, or equivalently $f(z) = \phi^{-1}((\phi(z))^d)$; note d in this case is a power, and not an iterate. *External rays* are the images, under ϕ^{-1}, of the radial lines outside of the unit circle. That is, for any ray $\gamma_\theta = \{re^{i\theta} : r > 1\}$, the external ray is $\phi^{-1}(\gamma_\theta)$.

In the case of $f_{-1}(z) = z^2 - 1$, the Böettcher function maps the real axis to the real axis. Consequently, the image under $\phi(z)$ of the vertical line $L = \{x + iy : x = 2\}$ is perpendicular to γ_0 on the positive real axis. As we look at preimages of γ_0 under z^2 we gain γ_π on the negative real axis, then $\gamma_{\pi/2}$ and $\gamma_{3\pi/2}$ both on the imaginary axis, and so on. Meanwhile, the preimages of $\phi(L)$ under z^2 will be perpendicular to each γ_θ. Then, ϕ^{-1} of each γ_θ produces an external ray, while ϕ^{-1} of the preimages of $\phi(L)$ are the level curves as seen in Figure 11. Since $\phi(z)$ and z^2 are conformal mappings, it follows that these external rays are perpendicular to the level curves; each external ray begins at the Julia Set, crosses the level curve once, and continues toward infinity.

For other functions, $f_c(z)$, similar behavior occurs with external rays, although the reals don't always map to the reals under $\phi(z)$.

Extending to rational maps

Does this explanation of contours hold for other functions besides polynomials? Actually, Theorem 1 is true for the class of complex rational functions $R(z) = P(z)/Q(z)$ where $P(z)$ and $Q(z)$ are polynomials, the degree of $P(z)$ is at least 2 more than the degree of $Q(z)$, and there is some $M \in \mathbb{R}$ such that for all $z \in \mathbb{C}$ with $|z| > M$, $\lim_{n\to\infty} R^n(z) = \infty$. In addition, the critical point formulas are the same, and are still all saddle points. It follows that somewhat similar behavior will occur when contour plots are created. However, the images become more complicated when poles and preimages of poles appear, as seen in the images found in [**1**, **3**].

Consequently, for a function $f(z)$ that is either a complex polynomial function or a rational function from the class described above, there is a way to see a glimpse of the graph of $f^n(z)$. The output of $f^n(z)$ can be projected onto almost any line in $\mathbb{C}$, creating a surface. When contour plots of the surface are computed, an interesting property arises. Almost any way you slice it, it comes up with a Julia Set in the middle.

Acknowledgements The authors thank Clinton Curry and Elizabeth Russell for helpful discussions related to the content of this paper.

REFERENCES

[1] Barnes, J., Curry, C., Russell, E., Schaubroeck, L. (2015). Emerging Julia sets. *Math. Mag.* 88(2): 91–102.

[2] Barnes, J., Curry, C., Schaubroeck, L. (2010). Real and imaginary parts of polynomial iterates. *New York J. Math*. 16: 749–761.

[3] Barnes, J., Kreahling, W., Schaubroeck, B. (2017). *Coloring Book of Complex Function Representations*. Washington, DC: Mathematical Association of America.

[4] Barnsley, M. (1988). *Fractals Everywhere*. Boston, MA: Academic Press.

[5] Beardon, A.F. (1991). *Iteration of Rational Functions*. New York, NY: Springer-Verlag.

[6] Devaney, R. (1992). *A First Course in Chaotic Dynamical Systems*. Boulder, CO: Westview Press.

[7] Devaney, R. (2003). *An Introduction to Chaotic Dynamical Systems*, 2nd ed. Boulder, CO: Westview Press.

[8] Mathews, J., Howell, R. (2011). *Complex Analysis for Mathematics and Engineering*, 6th ed. Burlington, MA: Jones & Bartlett Learning.

[9] Peitgen, H., Jürgens, H., Saupe, D. (1992). *Chaos and Fractals: New Frontiers of Science*. New York, NY: Springer-Verlag.

[10] Peitgen, H., Richter, P. (1986). *The Beauty of Fractals: Images of Complex Dynamical Systems*. New York, NY: Springer-Verlag.

[11] Zill, D., Shanahan, P. (2013). *Complex Analysis: A First Course with Applications*. Jones & Bartlett Learning.

Summary. Functions from a complex plane to itself are difficult to visualize because of their four-dimensional nature. In the paper "Emerging Julia Sets" published in the April 2015 issue of *Mathematics Magazine*, the authors explore the connections between the real and imaginary parts of complex functions and their corresponding Filled Julia Sets. In this paper, we expand on these ideas by looking more closely at what the actual contour plots of the projected graphs mean. In addition we explore what happens when we slice the complex plane in locations other than the real and imaginary axes and project the four-dimensional graphs onto those slices. We describe precisely what the contours in each projection mean for complex polynomials of degree at least 2 and why we always see Julia Sets in the images. We end by discussing how these concepts extend to complex rational maps.

JULIA A. BARNES (MR Author ID: 613418) received her Ph.D. in mathematics from the University of North Carolina at Chapel Hill in 1996 and is now a professor of mathematics at Western Carolina University. Her mathematical interests include complex dynamical systems, ergodic theory, organizing math treasure hunts, developing hands-on teaching ideas, and searching for interesting mathematical coloring book images. She is also an associate director for Project NExT and enjoys hiking and playing racquetball in her spare time.

LISBETH E. SCHAUBROECK (MR Author ID: 663310) earned her Ph.D. at the University of North Carolina at Chapel Hill in 1998 and is now a professor of mathematical sciences at the United States Air Force Academy. Her professional interests include complex variables, undergraduate knot theory, mentoring new faculty members, and searching for interesting mathematical coloring book images. She lives with her husband and two sons in Colorado Springs, where they especially enjoy bowling and target archery.

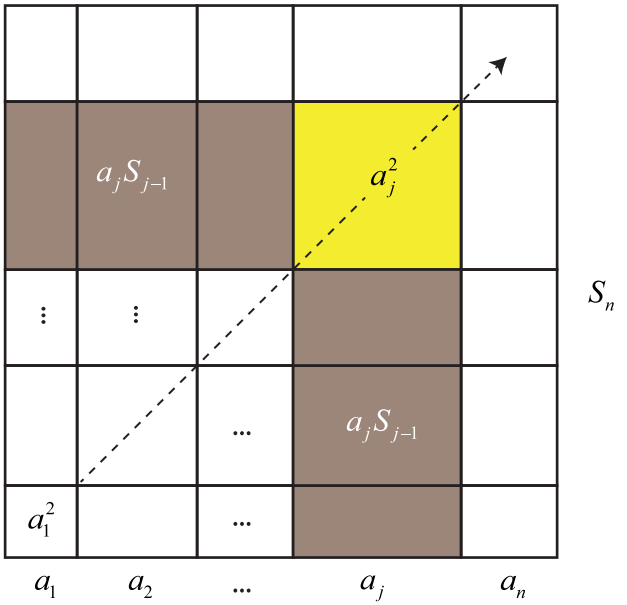
Proof Without Words: The Square of a Sum

ÁNGEL PLAZA
Universidad de Las Palmas de Gran Canaria, Spain
angel.plaza@ulpgc.es

Theorem. *For any finite sequence of positive numbers* $\left(a_j\right)_{j=1}^{n}$ *whose partial sums are* $\left(S_j\right)_{j=1}^{n}$ *we have* $\sum_{j=1}^{n}\left(a_j^2+2a_jS_{j-1}\right)=S_n^2$.

Notice that for $j=1$, S_{j-1} is defined as 0, thus $S_1^2=\sum_{j=1}^{1}\left(a_1^2+2a_1S_0\right)=a_1^2$.

Proof.



Anonymous referees have suggested some precursors of this proof, such as the square of two numbers in [2], a charming video for the square of three numbers in [3], and Fry's iconic proof without words for the sum of cubes [4].

REFERENCES

[1] Treeby, D. (2016). Further physical derivations of Fibonacci summations. *Fibonacci Quart.* 54(4): 327–334.
[2] Nelsen, R.B. (1993). *Proofs without words. Exercises in visual thinking*. Washington, DC: The Mathematical Association of America, p. 3.
[3] https://www.youtube.com/watch?v=w4UrDkDzkI4.
[4] Fry, A. L. (1985). Proofs without words: sum of cubes. *Math. Mag.* 58(1): 11.

Summary. The square of a sum of a finite sequence of positive numbers is written as the sum of a certain sequence involving partial sums and the general term.

ÁNGEL PLAZA (MR Author ID: 350023) received his master's degree from Universidad Complutense de Madrid in 1984 and his Ph.D. from Universidad de Las Palmas de Gran Canaria in 1993, where he is a Full Professor in Applied Mathematics. He is interested in mesh generation and refinement, combinatorics and visualization support in teaching and learning mathematics.

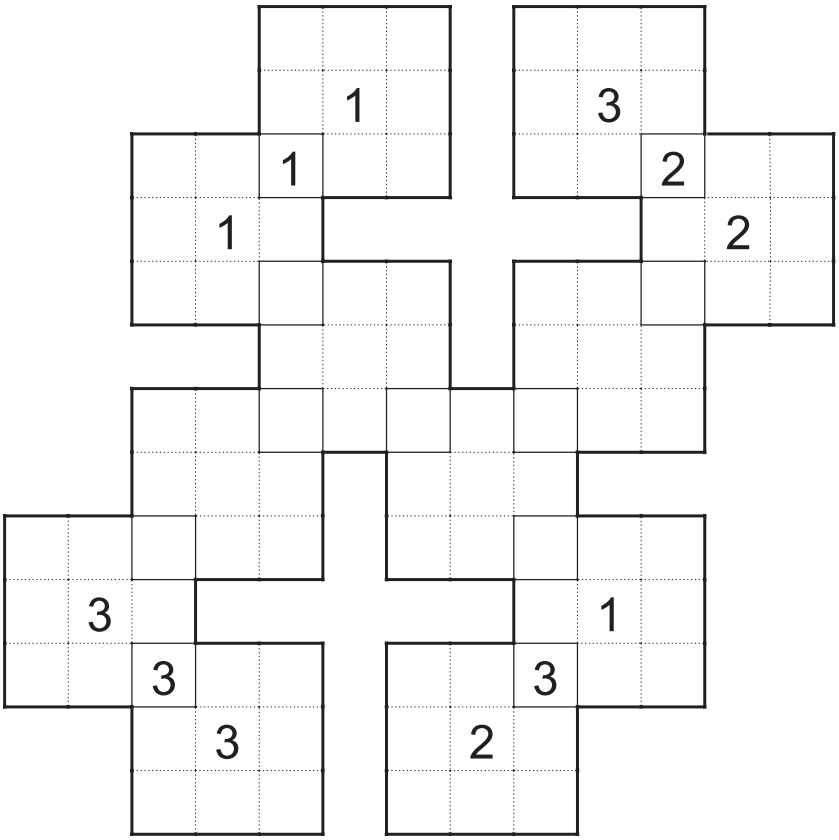
Math. Mag. **92** (2019) 17. doi:10.1080/0025570X.2018.1541686
MSC: Primary 05A19
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

Introducing TRIBUS

In each issue of THE MAGAZINE in 2019, readers will be treated to a TRIBUS puzzle. Mathematician and puzzle enthusiast David Nacin is the creator of TRIBUS as well as countless other mathematically rich puzzles. We solved many different puzzles in the process of finding the right one for this year's offering. We think that TRIBUS contains the right balance of entertainment and mathematics. As you solve each puzzle, you will undoubtedly discover small lemmas that help along the way. Try your hand at this issue's installment below, and enjoy! For more of David's puzzles, check out his puzzle collection at http://quadratablog.blogspot.com.

— Brittany C. Shelton, Albright College, Reading, PA
Andrés E. Caicedo, Mathematical Reviews, Ann Arbor, MI

TRIBUS puzzle



How to play. Fill each of the three-by-three squares with either a 1, 2, or 3 so that each number appears exactly once in each column and row. Some cells apply to more than one square, as the squares overlap. Each of the three-by-three squares must be distinct. The solution can be found on page 31.

— David Nacin, William Paterson University, Wayne, NJ (nacind@wpunj.edu)

The Involutory Laguerre Transition Matrix

MAUREEN T. CARROLL
University of Scranton
Scranton, PA 18510
maureen.carroll@scranton.edu

ELYN RYKKEN
Muhlenberg College
Allentown, PA 18104
elynrykken@muhlenberg.edu

It's unusual for a matrix to be its own inverse. Indeed, a quick poll suggests that many mathematicians do not even know the name for a matrix with this curious property. (Pssst! It's involutory.) So, we thought it was interesting to happen upon one such nontrivial matrix unsuspectingly. When this matrix provided a gateway to a sequence of involutory matrices, one of any size, we sat up and took notice. The matrices with this unexpected property surfaced while working with sets of orthogonal polynomials, specifically, Laguerre polynomials. In this paper, we construct the family of Laguerre polynomials and discuss its context within the classical families of orthogonal polynomials. Once we have provided the necessary basics, we prove the surprising result that the transition matrix corresponding to a basis of these polynomials is its own inverse.

Named for French mathematician Edmond Laguerre (1834–1886), the Laguerre polynomials form a family of orthogonal polynomials. Any such family is a sequence of polynomials of increasing degree which are mutually orthogonal relative to a specified inner product on the vector space of continuous functions on $[a, b]$. Specifically, if $p_0, p_1, p_2, \ldots$ is a sequence of polynomials where $deg(p_i) = i$ for each i, and $\langle p_i, p_j \rangle = 0$ whenever $i \neq j$, then $\{p_i\}$ is a sequence of orthogonal polynomials. Studied for over two centuries as an important tool to solve problems arising in approximation theory, probability and differential equations, the classical families hail from the nineteenth century and bear the familiar names of Adrien-Marie Legendre (1752–1833), Carl Jacobi (1804–1851), Pafnuty Chebyshev (1821–1894), Charles Hermite (1822–1901) and Laguerre. Orthogonality in all of the classical families of polynomials is given by an inner product of the form $\langle p, q \rangle = \int_a^b p(x)q(x)w(x)dx$, where the interval and the positive and continuous weight function, $w(x)$, determine the family. The interval of integration may be open, half-open, finite or infinite, but $\int_a^b f(x)w(x)dx$ must converge for every f when the integral is improper. In particular, for Laguerre polynomials we have $\langle p, q \rangle = \int_0^\infty p(x)q(x)e^{-x}dx$. Since orthogonality is preserved under scalar multiplication, we have a small degree of freedom when choosing a polynomial of a given degree. For the Laguerre polynomials, with $p_0 = 1$ and $p_1 = ax + b$, solving $\langle p_0, p_1 \rangle = 0$ gives $a = -b$. So, for example, we may take $p_1 = 1 - x$ or any scalar multiple of $1 - x$.

While these families are named for mathematicians who worked with them, as is often the case with the naming of theorems, the moniker does not necessarily correspond with the original explorer. Both Joseph-Louis Lagrange (1736–1813) and Niels Henrik Abel (1802–1829) worked on the Laguerre polynomials well before Laguerre himself did, though neither's work was published until the late nineteenth century. Lagrange encountered these polynomials in 1765 in the course of solving a differential equation related to fixed wire oscillations, but his work did not appear in print until 1867 [**6**]. Though Abel's work was not published until 1881, he produced a generating

Math. Mag. **92** (2019) 19–23. doi:10.1080/0025570X.2018.1541688
MSC: Primary 15B36, Secondary 15A09

function for these polynomials in 1826, namely $f(x, v) = (1 - v)^{-1} e^{-\frac{xv}{1-v}}$ [1]. Even Chebyshev worked with Laguerre polynomials before Laguerre, using these polynomials to approximate an arbitrary function in a paper that appeared in 1859 [14]. The mathematician whose name *is* attached to these polynomials published his work a full twenty years after Chebyshev [7]! Like Lagrange, Laguerre was studying polynomial solutions to a differential equation. Ironically, however, the polynomials that Laguerre studied were very similar but not identical to those that now bear his name. Readers interested in tracing the history of Laguerre polynomials should consult [1, 2, 6–10, 14]. Those looking for material on classical orthogonal polynomials may find [3–5, 11, 12] or [13] useful.

While any of the classical families can be defined in a number of ways, each can be characterized as satisfying a particular three-term recurrence relation and a second-order differential equation. The Laguerre polynomials that we generate in the following section satisfy both the differential equation

$$xy'' + (1 - x)y' + ny = 0, \tag{1}$$

and the recurrence relation

$$L_{n+1}(x) = (2n + 1 - x)L_n(x) - n^2 L_{n-1}(x), \tag{2}$$

for $n \geq 1$ where $L_0 = 1$ and $L_1 = 1 - x$. Though we could use the recurrence relation or Abel's generating function to produce the Laguerre polynomials, we prefer to take a more constructive approach that provides insight into their orthogonality in the section to follow.

Generating Laguerre polynomials

Any family of orthogonal polynomials can be produced using the Gram–Schmidt (Jorgen Gram (1850–1916) and Erhard Schmidt (1876–1959)) process on the standard basis for $\mathbb{P}_n$. For the Laguerre polynomials we use the inner product $\langle p, q \rangle = \int_0^\infty p(x)q(x)e^{-x}dx$, and take $p_0 = 1$. As linear algebra students will recall, for p_1 the Gram–Schmidt process requires that we find the component of x that is orthogonal to the space spanned by p_0. Letting $W_0 = \text{Span}\{p_0\}$, we have

$$p_1 = x - \text{proj}_{W_0} x = x - \frac{\langle x, 1 \rangle}{\langle 1, 1 \rangle} 1 = x - \frac{1}{1} 1 = x - 1.$$

Likewise, for p_2, we find the component of x^2 orthogonal to the space $W_1 = \text{Span}\{p_0, p_1\}$. This gives

$$\begin{aligned} p_2 = x^2 - \text{proj}_{W_1} x^2 &= x^2 - \frac{\langle x^2, 1 \rangle}{\langle 1, 1 \rangle} 1 - \frac{\langle x^2, x - 1 \rangle}{\langle x - 1, x - 1 \rangle} (x - 1) \\ &= x^2 - \frac{2}{1} 1 - \frac{4}{1}(x - 1) = x^2 - 4x + 2. \end{aligned}$$

For $p_3 = x^3 - \text{proj}_{W_2} x^3$ with $W_2 = \text{Span}\{p_0, p_1, p_2\}$, we have

$$p_3 = x^3 - 6 \cdot 1 - 18(x - 1) - 9(x^2 - 4x + 2) = x^3 - 9x^2 + 18x - 6.$$

Traditionally, the Laguerre polynomials have leading coefficients that alternate in sign. Thus, multiplying every other polynomial from our construction by -1 gives the Laguerre basis $\{1, -x + 1, x^2 - 4x + 2, -x^3 + 9x^2 - 18x + 6, \dots\}$. This family of

polynomials satisfies both the differential equation and the recurrence relation given in equations (1) and (2), and is consistent with Abel's generating function. (Try it!) It is interesting to note that, in his investigation of the integral $\int_x^\infty t^{-1}e^{-t}dt$, Laguerre studied a differential equation similar to equation (1) which produces a slightly different set of polynomials, namely $\{1, x+1, x^2+4x+2, x^3+9x^2+18x+6, \dots\}$. The standard Laguerre polynomials often found in the literature today have a constant term of one, and thus, satisfy a slightly different recurrence relation than equation (2).

While the Gram–Schmidt construction is a natural way to generate the polynomials from scratch, we will need a closed-form solution for the nth degree Laguerre polynomial, $L_n(x)$, in the proof of our main result. This can be generated by finding polynomial solutions to equation (1). Using a power series of the form $y = \sum a_i x^i$, we have $y' = \sum i a_i x^{i-1}$ and $y'' = \sum i(i-1)a_i x^{i-2}$. With these substitutions, the differential equation gives a recursive formula for the coefficients, $a_{i+1} = -\frac{(n-i)}{(i+1)^2}\, a_i$, where we are free to choose the constant term, a_0. Taking $a_0 = n!$ to match the polynomials produced by the Gram–Schmidt process above, we can use induction to show that the coefficients of $L_n(x)$ are given by $a_i = (-1)^i \frac{n!}{i!}\binom{n}{i}$ for all positive integers i. Thus, we have

$$L_n(x) = \sum_{i=0}^{n} \frac{n!}{i!}\binom{n}{i}(-x)^i. \tag{3}$$

Alternatively, this closed formula can also be derived by using a variation of Abel's generating function and the power series expansion for e^x.

Notice that the Gram–Schmidt construction guarantees orthogonality, which in turn ensures linear independence. Thus, any finite collection $\{p_0, p_1, p_2, \dots, p_n\}$ forms a basis for $\mathbb{P}_n$, and we can consider how to calculate coordinate vectors in the following section.

The Laguerre transition matrix

In any vector space, a change in basis produces a different coordinate vector for any vector q relative to the new basis. One of the benefits of working with an orthogonal basis is the relative ease of coordinate vector calculation. In particular, the coordinates of q relative to an orthogonal basis $\mathcal{B} = \{b_1, b_2, \dots, b_n\}$ are given by $c_i = \frac{\langle q, b_i\rangle}{\langle b_i, b_i\rangle}$ since $\langle q, b_i\rangle = \langle\, c_1b_1 + c_2b_2 + \cdots + c_nb_n,\ b_i\,\rangle = c_1\langle b_1, b_i\rangle + c_2\langle b_2, b_i\rangle + \cdots + c_n\langle b_n, b_i\rangle = c_i\langle b_i, b_i\rangle$, as orthogonality ensures that $\langle b_i, b_j\rangle = 0$ for all $i \neq j$. For example, with basis $\mathcal{B}' = \{1, 1-x, 2-4x+x^2\}$ and $q = 1-2x+3x^2$, calculating c_3 using this formula gives $c_3 = \frac{\langle q, b_3\rangle}{\langle b_3, b_3\rangle} = \frac{12}{4} = 3$.

Without orthogonality, we can produce the coordinates of q by solving a matrix equation. For example, given $\mathcal{B}'$ as above and the standard basis $\mathcal{B} = \{1, x, x^2\}$ for $\mathbb{P}_2$, to determine $[\,q\,]_{\mathcal{B}'} = \langle c_1, c_2, c_3\rangle$, we note that $q = 1-2x+3x^2 = c_1(1) + c_2(1-x) + c_3(2-4x+x^2) = (c_1+c_2+2c_3)(1) + (-c_2-4c_3)x + c_3x^2$. Rewritten as a matrix equation, we have $A[\,q\,]_{\mathcal{B}'} = [\,q\,]_{\mathcal{B}}$, where

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & -1 & -4 \\ 0 & 0 & 1 \end{bmatrix},$$

and coordinate vectors $[\,q\,]_{\mathcal{B}}$ and $[\,q\,]_{\mathcal{B}'}$ are written as 3×1 matrices. Solving this system gives $[\,q\,]_{\mathcal{B}'} = \langle 5, -10, 3\rangle$. Matrix A translates between coordinate vectors of different bases, and thus, is called the *transition matrix* from $\mathcal{B}'$ to $\mathcal{B}$. Since the

columns of A consist of basis vectors, every transition matrix is invertible, making A^{-1} the transition matrix from $\mathcal{B}$ to $\mathcal{B}'$.

We discovered that the transition matrix from the Laguerre polynomials to the standard basis for $\mathbb{P}_n$ has an interesting and unusual property, one that is not found in the other classical orthogonal families of Legendre, Chebyshev or Hermite. To illustrate this, consider the 4×4 transition matrix from the Laguerre basis $\{1, 1-x, 2-4x+x^2, 6-18x+9x^2-x^3\}$ to the standard basis for $\mathbb{P}_3$,

$$A_4 = \begin{bmatrix} 1 & 1 & 2 & 6 \\ 0 & -1 & -4 & -18 \\ 0 & 0 & 1 & 9 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

It surprised us to find that $A_4^{-1} = A_4$. With the fifth Laguerre polynomial, $24 - 96x + 72x^2 - 16x^3 + x^4$, the 5×5 upper-triangular transition matrix A_5 is formed from A_4 by adding the fifth column $[24, -96, 72, -16, 1]^T$. It is easy to check that A_5 is also its own inverse! In fact, this property holds for any Laguerre transition matrix, and the following is a proof of this surprising and beautiful property.

Theorem. *Every $n \times n$ Laguerre transition matrix is involutory.*

Proof. An invertible matrix A is its own inverse if and only if $A^2 = I$. In order for an upper triangular matrix to satisfy the equation $A^2 = I$, each entry on the main diagonal must be ± 1. Thus, we will use the Laguerre polynomials where each leading coefficient is ± 1. Using equation (3), this means that the i, j^{th} entry of the $n \times n$ transition matrix is the coefficient of x^{i-1} in the polynomial of degree $j-1$. Hence, the transition matrix is given by

$$A_{i,j} = (-1)^{i-1} \frac{(j-1)!}{(i-1)!} \binom{j-1}{i-1}$$

for $1 \le i \le j \le n$, and $A_{i,j} = 0$ elsewhere. Therefore, A is upper-triangular. Since the product of upper-triangular matrices is upper-triangular, we need only consider the i, j^{th} entries of A^2 where $i \le j$ in order to verify that $A^2 = I$.

First let's consider the entries of A^2 where $i = j$. By equation (3), $A_{i,i} = \pm 1$. Since A is upper-triangular, the main diagonal of A^2 consists of squares of the corresponding diagonal entries of A, that is, $(A_{i,i})^2 = (\pm 1)^2 = 1$.

Next, for the entries where $i < j$, we have

$$\begin{aligned} A_{i,j}^2 &= \sum_{r=i}^{j} A_{i,r} A_{r,j} \\ &= \sum_{r=i}^{j} (-1)^{i-1} \frac{(r-1)!}{(i-1)!} \binom{r-1}{i-1} (-1)^{r-1} \frac{(j-1)!}{(r-1)!} \binom{j-1}{r-1} \\ &= \sum_{r=i}^{j} (-1)^{i+r} \frac{[(j-1)!]^2}{[(i-1)!]^2} \left(\frac{1}{(j-r)!(r-i)!} \right) \\ &= \frac{[(j-1)!]^2}{[(i-1)!]^2} \sum_{r=i}^{j} (-1)^{i+r} \left(\frac{1}{(j-r)!(r-i)!} \right). \end{aligned}$$

Setting $k = r - i$, we rewrite the summation term as $\sum_{k=0}^{j-i} (-1)^k \left(\frac{1}{(j-i-k)!k!} \right)$.

With $m = j - i$, this can be simplified further to

$$\sum_{k=0}^{m}(-1)^k\left(\frac{1}{(m-k)!k!}\right) = \frac{1}{m!}\sum_{k=0}^{m}(-1)^k\binom{m}{k} = \frac{1}{m!}[1+(-1)]^m .$$

Notice that, since $i < j$, then m is positive and this sum is 0. Hence, we have shown that $A^2 = I$, and thus, $A = A^{-1}$, as desired. ■

As a final note, let's reconsider the computation of coordinate vector c for a polynomial v relative to any non-standard basis $\{b_i\}$ for $\mathbb{P}_{n-1}$. In general, this requires solving $Ac = v$, a linear system of n equations with n unknowns, where A is the transition matrix. If we only need to find one set of coordinates, then row reduction is easier than finding A^{-1}. When the basis is orthogonal, it is preferable to use the inner product, exploiting the fact that $\langle b_i, b_j\rangle = 0$ whenever i differs from j. Hence, as previously noted, the coordinate vector components are $c_i = \frac{\langle v, b_i\rangle}{\langle b_i, b_i\rangle}$, which greatly simplifies the computation of c in the general orthogonal case. However, it's trivial to find c when the transition matrix is involutory as in the Laguerre case. Here, Av yields the coordinate vector with no extra effort, thus eclipsing the usual computational advantage to working with an orthogonal basis.

Acknowledgments We thank our linear algebra students for posing the question. We also thank the referees and, especially, the editor for his helpful suggestions.

REFERENCES

[1] Abel, N. H. (1881). *Oeuvres Complète*, Vol. 2, Grøndahl, Christiania (Norway).
[2] Chebyshev, P. L. (1962). *Oeuvres*, Vol. 1. New York: Chelsea Pub. Co.
[3] Doman, B. G. S. (2016). *The Classical Orthogonal Polynomials*. Hackensack, NJ: World Scientific Pub.
[4] Hilbert, D., Courant, R. (1953). *Methods of Mathematical Physics*, Vol. 1. New York: Interscience Pub.
[5] Jackson, D. (1941). *Fourier Series and Orthogonal Polynomials*. Washington, DC: The Mathematical Association of America. Dover reprint, 2004.
[6] Lagrange, J. L. (1867). *Oeuvres*, Vol. 1. Paris: Gauthier-Villars.
[7] Laguerre, E. N. (1898). *Oeuvres*, Vol. 1. Paris: Gauthier-Villars.
[8] Mawhin, J., Ronveaux, A. (2010). Schrödinger and Dirac equations for the hydrogen atom, and Laguerre polynomials, *Arch. Hist. Exact Sci.* 64(4): 429–460.
[9] Murphy, R. (1833). On the inverse method of definite integrals, with physical applications. *Trans. Camb. Phil. Soc.* 4: 353–408.
[10] Murphy, R. (1835) On the inverse method of definite integrals, with physical applications. *Trans. Camb. Phil. Soc.* 5: 113–148, 315–393.
[11] Sansone, G. (1959). *Orthogonal Functions*. New York: Interscience Publishers.
[12] Shohat, J., Hille, E., Walsh, J. L. (1940). A bibliography of orthogonal polynomials. *Bull. Natl. Res. Council* 103.
[13] Szegö, G. (1939). *Orthogonal Polynomials*. Providence, RI: American Mathematical Society.
[14] Szegö, G. (1982). *Gábor Szegö: Collected Papers*, Askey, R., ed. Boston, MA: Birkhäuser.

Summary. Named for French mathematician Edmond Laguerre (1834–1886), the Laguerre polynomials form one of the classical families of orthogonal polynomials. We construct this family of polynomials before presenting the surprising result that the transition matrix corresponding to a basis of Laguerre polynomials is its own inverse.

MAUREEN T. CARROLL (MR Author ID: 659601) is a professor of mathematics at the University of Scranton. Her most recent contribution to this *Magazine* was a crossword puzzle appearing in 2017. She has been collaborating with her coauthor since meeting in MAA's Project NExT. Their latest project, *Geometry: The Line and the Circle*, was published last month in the AMS/MAA Textbook Series.

ELYN RYKKEN (MR Author ID: 627637) is a professor of mathematics at Muhlenberg College. She is a St. Olaf alum who earned her Ph.D. in dynamical systems from Northwestern University. In addition to mathematics, she enjoys cooking, traveling and bicycling.

Existence of a Solution for Beloch's Fold

JORGE C. LUCERO
University of Brasília
Brasília, DF 70910-900, Brazil
lucero@unb.br

Almost a century ago, Italian mathematician Margherita Piazolla Beloch studied the following fold operation and its associated geometry [3, 9].

Beloch's fold. Given points P and Q and lines m and n on a sheet of paper, fold the paper along a straight line so that P is placed onto m and Q onto n simultaneously (Figure 1).

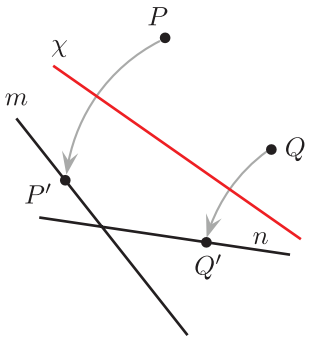


Figure 1 Beloch's fold along line χ.

In her work, Beloch showed that the fold may be applied to solve arbitrary cubic equations. The operation is considered today as one of the so-called "axioms" of origami (the Japanese art of paper folding), which are combinations of alignments between given points and lines on a sheet of paper that may be achieved with a single fold [1, 2, 4, 7, 9]. For reference, a standard version of the complete set of axioms is listed in Table 1 [2]. It has been shown that Beloch's fold is the most powerful of these axioms, in the sense that it includes all the others as particular cases [7]. Due to its association with a cubic equation, the operation may have up to three solutions. In fact, it has been applied to solve construction problems related to cubic equations, such as trisecting arbitrary angles [8], duplicating the cube [12] and constructing heptagons [5].

As an example of the utility of Beloch's fold, Figure 2 shows how to find the cube root of an arbitrary positive real number a [4, 9]: (1) Set a Cartesian system of coordinates x, y and mark points $P(0, -1)$ and $Q(-a, 0)$. (2) Trace (e.g., by folding) lines $y = 1$ and $x = a$, denoted by m and n, respectively. (3) Fold along a line χ so as to place P onto m and Q onto n. (4) The x intercept of line χ (point R) is $\sqrt[3]{a}$.

Note that QOS, POR and SOR in Figure 2 are similar right triangles. Therefore,

$$\frac{|\overline{QO}|}{|\overline{SO}|} = \frac{|\overline{SO}|}{|\overline{RO}|} = \frac{|\overline{RO}|}{|\overline{PO}|}.$$

Math. Mag. **92** (2019) 24–31. doi:10.1080/0025570X.2018.1526591
MSC: Primary 51M15
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

Number	Axiom
1	Given two points P and Q, we can fold a line connecting them.
2	Given two points P and Q, we can fold P onto Q.
3	Given two lines m and n, we can fold m onto n.
4	Given a point P and a line m, we can make a fold perpendicular to m passing through P.
5	Given two points P and Q and a line m, we can make a fold that places P onto m and passes through Q.
6	Given two points P and Q and two lines m and n, we can make a fold that places P onto m and places Q onto n.
7	Given a point P and two lines m and n, we can make a fold perpendicular to m that places P onto n.

TABLE 1: "Axioms" of origami constructions [2].

Eliminating $|\overline{SO}|$ yields

$$|\overline{RO}|^3 = |\overline{QO}||\overline{PO}|^2,$$

and substituting $|\overline{QO}| = a$ and $|\overline{PO}| = 1$ produces $|\overline{RO}| = \sqrt[3]{a}$.

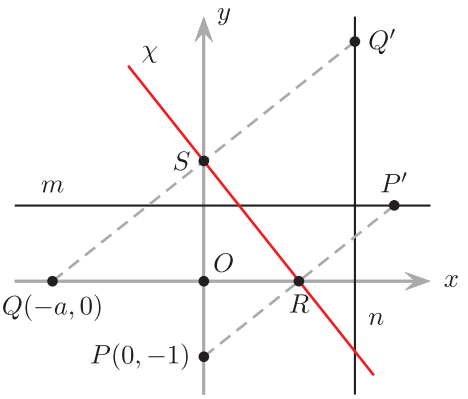


Figure 2 Construction for finding the cube root of a number $a > 0$.

However, depending on the specific configuration of points and lines, the fold operation may be impossible to perform. Therefore, it is sometimes stated in a form similar to "whenever possible, fold the paper so that..." [1, 9]. Naturally, the question arises: When is Beloch's fold possible? Further, it is regularly assumed that P and Q are distinct, as well as lines m and n, and that the points are not initially located on the respective lines. Then, one may also ask what happens if those conditions are not met, does the operation still have a solution? A complete answer to the previous questions does not seem to be reported in the literature, and this article provides a simple and general one: a solution exists if and only if the distance between P and Q is not smaller than the distance between m and n. Here, the distance between a pair of lines is taken as the distance between the corresponding point sets. Hence, if m and n are nonparallel then their distance is zero and the fold is always possible. This fact was already deduced by Martin [11] under the restriction that P and Q are distinct. In the

same work, he also indicated that if m and n are parallel then the fold may not have a solution. In addition, the particular cases of $P = Q$ and $m = n$ were discussed by Geretschläger [4, 6], who also argued that a solution always exists if m and n intersect.

The present analysis provides a general treatment encompassing all possible configurations of the given points and lines.

Analysis of solutions of Beloch's fold

Reflection of points Folding a sheet of paper along a straight line superposes the paper on one side of the fold line to the other side, and the superposition may be modeled as a reflection across the fold line (Figure 3) [11].

Definition 1. Given a line χ, the reflection $\mathcal{F}_\chi$ in χ is the mapping on the set of points in the plane such that for point P

$$\mathcal{F}_\chi(P) = \begin{cases} P & \text{if } P \in \chi, \\ P' & \text{if } P \notin \chi \text{ and } \chi \text{ is the perpendicular bisector of segment } \overline{PP'}. \end{cases}$$

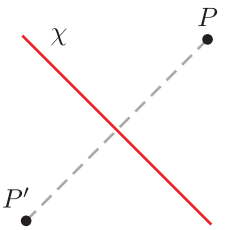


Figure 3 Reflection of point P in line χ.

Degenerate cases The following lemmas are related to degenerate cases of Beloch's fold, i.e., when the given points or lines are equal, or one of the points is already on a line (the relation is explained in Theorem). Demonstrations of the lemmas may be found in the literature; e.g., [1, 7, 10, 11]. Nevertheless, they are provided here for clarity of the analysis.

Lemma 1. *Given a point P and a line m, with $P \notin m$, a fold line reflects P onto m if and only if the line is tangent to a parabola with focus P and directrix m.*

Proof. Without loss of generality, choose a Cartesian system of coordinates x, y so that P is located at $(0, k)$ and line m is $y = -k$, where k is a constant (Figure 4). Also, let $P' = \mathcal{F}_\chi(P)$ be located at $(t, -k)$, where t is a free parameter.

The fold line χ is perpendicular to $\overline{PP'}$ and therefore has a slope of $t/(2k)$. Further, χ passes through the midpoint of $\overline{PP'}$, located at $(t/2, 0)$. Then, χ has an equation

$$y = \frac{t}{2k}\left(x - \frac{t}{2}\right). \tag{1}$$

Next, consider point T located at the intersection of χ with a vertical line through P'. Its coordinates are given by equation (1) with $x = t$, which produces $(t, t^2/(4k))$. Those coordinates describe parametrically a parabola Ψ with focus at $(0, k)$ and directrix $y = -k$. Further, the slope of a tangent to Ψ at point T is $y'(t) = t/(2k)$, which is the same slope of χ. Therefore, χ is a line tangent to Ψ at point T.

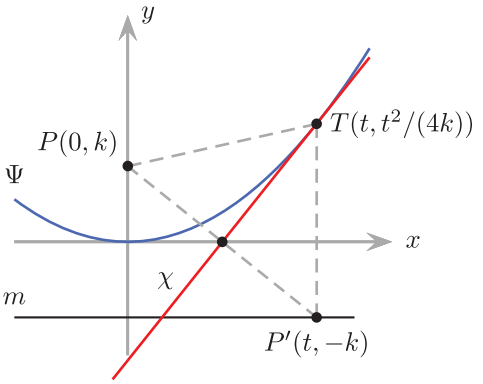


Figure 4 Reflection of a point P onto a line m. χ is the fold (reflection) line, and Ψ is a parabola with focus P and directrix m.

Also, equation (1) describes any tangent to Ψ at an arbitrary point T; therefore, any tangent is a fold line that reflects P onto m. ■

Lemma 2. *Given a point P and a line m, with $P \in m$, any fold line that passes through P or is perpendicular to m reflects P onto m.*

Proof. The lemma follows from Definition 1. Any fold line χ through P reflects P onto itself; therefore, $\mathcal{F}_\chi(P) = P \in m$. Further, any fold line ξ perpendicular to m is a perpendicular bisector of a segment $\overline{PQ}$, where Q is a point in m. Therefore, $\mathcal{F}_\xi(P) = Q \in m$. ■

Lemma 3. *Given points P, Q, and a line m, with $P \notin m$, a fold line reflects P onto m and passes through Q if and only if the distance between Q and P is not smaller than the distance between Q and m.*

Proof. By Lemma 1, any fold line is tangent to a parabola with focus P and directrix m. Assume the same parabola Ψ of Figure 4 and a point Q at the position (x_q, y_q). Replacing the coordinates of Q into equation (1) produces the quadratic equation

$$t^2 - 2x_q t + 4y_q = 0. \tag{2}$$

The discriminant of equation (2) is $\Delta = 4x_q^2 - 16y_q$, and $\Delta = 0$ yields $y_q = x_q^2/4$, which implies $Q \in \Psi$. Since Ψ is the location of points that are equidistant from P and m, we conclude that the fold operation has a unique solution when Q is equidistant to P and m, two solutions when Q is closer to m (i.e., $\Delta > 0$), and no solution when Q is closer to P (i.e., $\Delta < 0$). Figure 5 shows an example for the case of two solutions. ■

Lemma 4. *Given point P and lines m and n, with $P \notin m$, a fold line perpendicular to n reflects P onto m if and only if m is not parallel to n.*

Proof. Let $m \parallel n$ ($m \nparallel n$) represent that m is parallel (not parallel) to n. If $P \notin m$, according to Lemma 1 any fold line must be tangent to a parabola with focus P and directrix m. Again, assume the same parabola Ψ of Figure 4, and a line n with equation $ax + by + c = 0$. The fold line χ has a slope $t/(2k)$, and two cases are possible:

1. If $m \nparallel n$, then $a \neq 0$. Therefore, $t/(2k) = -b/a$, which has a unique solution for t (Figure 6).
2. If $m \parallel n$, then n is parallel to the directrix of Ψ and cannot be perpendicular to any tangent to the parabola. ■

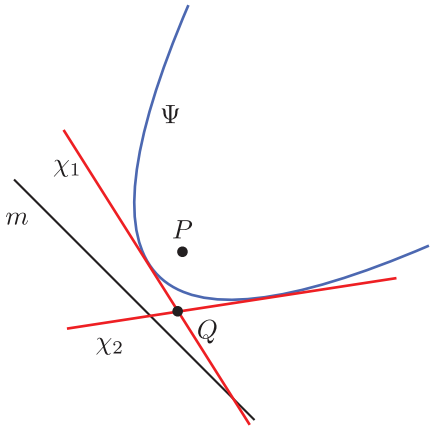


Figure 5 The fold lines χ_1 and χ_2 reflect P onto m and pass through Q. Curve Ψ is the parabola with focus P and directrix m.

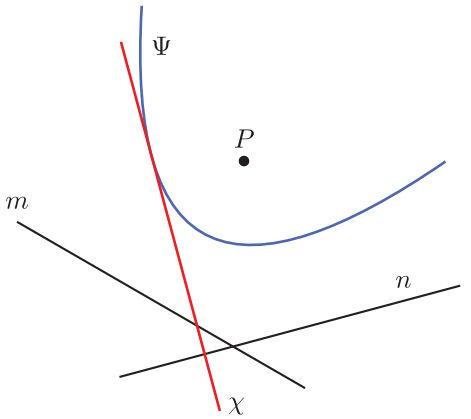


Figure 6 The fold line χ reflects P onto m and is perpendicular to n. Curve Ψ is the parabola with focus P and directrix m.

Main case This case is the regular assumed configuration of given points and lines in Beloch's fold.

Lemma 5. *Given points P and Q and lines m and n, with $P \notin m$, $Q \notin n$, and $P \neq Q$ or $m \neq n$, a fold line places P onto m and Q onto n if and only if the distance between m and n is smaller than the distance between P and Q.*

Proof. According to Lemma 1, any fold line must be tangent to both a parabola Ψ with focus P and directrix m, and a parabola Θ with focus Q and directrix n. For the parabola Ψ of Figure 4, a point Q is located at (x_q, y_q), and its reflection $Q' = \mathcal{F}_\chi(Q)$ at (x'_q, y'_q). Then, segment $\overline{QQ'}$ has a slope $(y_q - y'_q)/(x_q - x'_q)$.

The fold line χ, given by equation (1), is perpendicular to $\overline{QQ'}$. Therefore,

$$\frac{t}{2k} = -\frac{x_q - x'_q}{y_q - y'_q}. \tag{3}$$

Further, χ passes through the midpoint of $\overline{QQ'}$, which is located at

$$((x_q + x'_q)/2, (y_q + y'_q)/2).$$

Substituting those coordinates into equation (1) produces

$$2k(y_q + y'_q) = t\left(x_q + x'_q - t\right). \tag{4}$$

Finally, eliminating t from equations (3) and (4) yields

$$(y_q + y'_q)(y_q - y'_q)^2 = -(x_q^2 - x_q'^2)(y_q - y'_q) - 2k(x_q - x'_q)^2. \tag{5}$$

For a given line n, the coordinates of Q' satisfy an equation of the form

$$ax'_q + by'_q + c = 0, \tag{6}$$

where a, b and c are constants.

Equations (5) and (6) may be solved for x'_q and y'_q. Substituting this solution into equation (3) gives t, which defines the fold line χ in equation (1). Two cases are possible:

1. If $m \parallel n$, then Q' is on a horizontal line and so $y'_q = -c/b$. In that case, equation (5) is quadratic in x'_q and may have zero to two solutions.
2. If $m \nparallel n$, solving equation (6) for x'_q or y'_q and replacing in equation (5) produces a cubic equation with one to three solutions. An example for the latter case is shown in Figure 7.

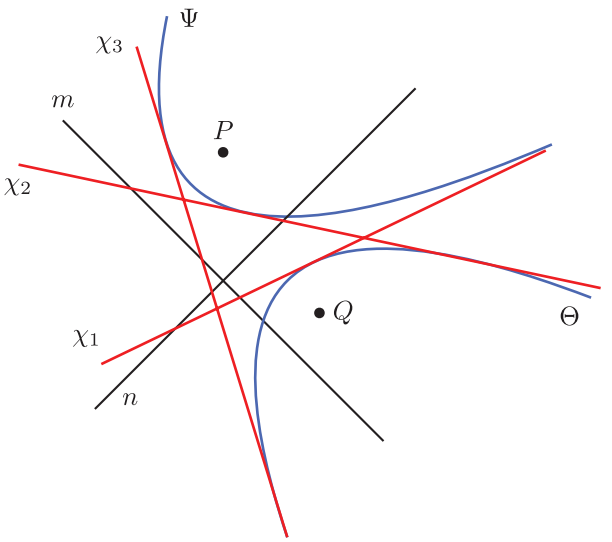


Figure 7 The fold lines χ_1, χ_2 and χ_3 reflect P and Q onto m and n, respectively. Curves Ψ and Θ are the parabolas with focus P, directrix m, and focus Q, directrix n, respectively.

Thus, the operation may not have a solution only in the case of $m \parallel n$. Re-arranging equation (5) produces

$$(2k - y_q + y'_q)(x_q - x'_q)^2 + 2x_q(y_q - y'_q)(x_q - x'_q) + (y_q + y'_q)(y_q - y'_q)^2 = 0,$$

which is a quadratic equation in $(x_q - x'_q)$. The discriminant is

$$\Delta = 4x_q^2(y_q - y'_q)^2 - 4(2k - y_q + y'_q)(y_q + y'_q)(y_q - y'_q)^2,$$

and letting $\Delta = 0$ produces

$$x_q^2 + (y_q - k)^2 = (y_q' + k)^2. \tag{7}$$

The left side of equation (7) is the squared distance between P and Q, and the right side is the squared distance between m and n. Therefore, the operation has a solution if and only if the former is not smaller than the latter, which results in $\Delta \geq 0$. That condition also includes the case of $m \nparallel n$, in which the distance between m and n is zero. ■

Main theorem

Theorem. *Given points P, Q, and lines m, n, a fold line that places P on m and Q on n exists if and only if the distance between P and Q is not smaller than the distance between m and n.*

Proof. Let us consider exhaustively all possible cases. To simplify the explanation, the distance between objects α and β (points or lines) is denoted as $d(\alpha, \beta)$.

1. $P \notin m$ and $Q \notin n$:
   (a) $P \neq Q$ or $m \neq n$: This is the main case treated by Lemma 5. A solution exists if and only if $d(P, Q) \geq d(m, n)$.
   (b) $P = Q$ and $m = n$: This case reduces to reflecting point P onto line m, treated by Lemma 1, which always has a solution. Note that $d(P, Q) = d(m, n) = 0$.
2. $P \notin m$ and $Q \in n$: By combining Lemmas 1 and 2, this case reduces to reflecting P onto m with a fold line through Q or perpendicular to n.
   (a) $P \neq Q$ or $m \neq n$:
      i. $m \nparallel n$: By Lemma 4, a solution perpendicular to n always exists. Since $d(m, n) = 0$, then $d(P, Q) \geq d(m, n)$.
      ii. $m \parallel n$ (including the case $m = n$): By Lemma 4, there is no solution perpendicular to n. However, by Lemma 3 there is a solution that reflects P onto m and passes through Q if and only if $d(P, Q) \geq d(m, Q) = d(m, n)$.
   (b) $P = Q$ and $m = n$: The conditions of this case form a contradiction and therefore it is impossible.
3. $P \in m$ and $Q \notin n$: This case is similar to Case 2, with the roles of P and Q, and m and n exchanged.
4. $P \in m$ and $Q \in n$: By Lemma 2, a fold line passing through both P and Q reflects P onto m and Q onto n. Such a line always exists and therefore this case always has a solution. If $m \parallel n$ (including $m = n$), then $d(P, Q) = d(m, n)$. If $m \nparallel n$, then $d(P, Q) \geq d(m, n) = 0$.

As detailed above, the fold has a solution in all cases that $d(P, Q) \geq d(m, n)$, and it does not have a solution when that condition is not satisfied. ■

REFERENCES

[1] Alperin, R. C. (2000). A mathematical theory of origami constructions and numbers. *New York J. Math.* 6: 119–133.

[2] Alperin, R. C., Lang, R. J. (2006). One-, two-, and multi-fold origami axioms. In: Lang, R. J., ed. *Origami 4: Fourth International Meeting of Origami Science, Mathematics and Education*. Natick, MA: A K Peters, pp. 371–393.

[3] Beloch, M. P. (1936) Sul metodo del ripiegamento della carta per la risoluzione dei problemi geometrici. *Periodico di Mathematiche Ser. 4*, 16: 104–108.
[4] Geretschläger, R. (1995). Euclidean constructions and the geometry of origami. *Math. Mag.* 68: 357–371.
[5] Geretschläger, R. (1997) Folding the regular heptagon. *Crux Mathematicorum.* 23: 81–88.
[6] Geretschläger, R. (2008). *Geometric Origami.* Shipley: Arbelos.
[7] Ghourabi, F., Kasem, A., Kaliszyk, C. (2013). Algebraic analysis of Huzita's origami operations and their extensions. In: Ida, T., Fleuriot, J., eds. *Automated Deduction in Geometry*. Berlin: Springer, pp. 143–160.
[8] Hull, T. (1996). A note on "impossible" paper folding. *Amer. Math. Monthly.* 103: 240–241.
[9] Hull, T. C. (2011). Solving cubics with creases: The work of Beloch and Lill. *Amer. Math. Monthly.* 118: 307–315.
[10] Lucero, J.C. (2017). On the elementary single-fold operations of origami: reflections and incidence constraints on the place. *Forum Geometricorum*. 17: 207–221.
[11] Martin, G. E. (1998). *Geometric Constructions.* New York, NY: Springer.
[12] Messer, P. (1986). Problem 1054. *Crux Mathematicorum.* 12: 284–285.

Summary. A fundamental operation in geometric constructions by paper folding is: given two points and two lines on a sheet of paper, fold the paper so that each point is placed onto one of the lines. This article analyzes conditions of existence of the fold operation and shows that it may be realized if and only if the distance between the given points is not smaller than the distance between the given lines.

JORGE C. LUCERO (MR Author ID: 630437) graduated from the National University of Córdoba, Argentina, in 1987, and received a PhD in Engineering from Shizuoka University, Japan, in 1993. Currently, he is Professor of Computer Science at the University of Brasília, Brazil, specializing in applied and computational mathematics. In his spare time, he enjoys folding origami.

More Fun with L-Trominoes in Three and Higher Dimensions

ARTHUR BEFUMO
Yale University
New Haven, CT 06520
arthur.befumo@yale.edu

JONATHAN LENCHNER
IBM Research Africa
Catholic University of East Africa
Nairobi, Kenya
jonathan.lenchner@ke.ibm.com

Anyone familiar with the game Tetris is familiar with polyominoes—they are finite collections of edge-connected, equal-sized, squares in the plane. In three dimensions, we adjust the definition and instead of edge-connected squares consider face-connected cubes. The simplest possible nonstraight polyomino is the L-tromino, 2D and 3D versions of which are shown in Figure 1.

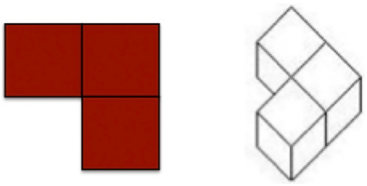


Figure 1 2D and 3D L-trominoes.

In 1954, Solomon Golomb introduced the notion of a polyomino and in the same article [**6**] showed via a beautifully simple inductive argument that if you remove a square from a chess board of size $2^N \times 2^N$ then the resulting board can always be tiled by L-trominoes. For those who have never seen it, we challenge you to reproduce this argument now.

In the 1980s, Chu and Johnsonbaugh [**4**, **5**] completely characterized which 2D $M \times N$ chess boards are tilable by L-trominoes if you remove an arbitrary square. In [**3**], we extended this work to three dimensions, showing that arbitrary 3D rectangular boards with one cube removed, of size $K \times L \times M$, where $KLM \equiv 1 \pmod 3$ and $K, L, M > 1$ are L-tilable.

Boards with two squares or cubes removed, called 2-deficient boards, are especially interesting because it is easy to see that there are **no** rectangular $N \times M$ boards that are generically tilable in the sense that they can be tiled with L-trominoes regardless of the squares removed—just remove two squares that effectively isolate a corner square. However, in 2008, Starr [**8**] showed that all 3D cubical boards of size N^3 for $N \equiv 2 \pmod 3$ with two cubes removed are L-tilable.

The present work began when the first author, Befumo, sent an email to the second author, Lenchner, suggesting that perhaps it would be possible to always split

Math. Mag. **92** (2019) 32–40. doi:10.1080/0025570X.2019.1537416
MSC: Primary 52C22, Secondary 05B45; 05B50
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

a 2-deficient 3D board into two 1-deficient boards and separately tile those with L-trominoes using the results of [3]. We suggest that the reader ponder this possibility to consider why such a strategy might or might not work, and how it could provide a roadmap to an ultimate proof.

This article extends Starr's result to show that arbitrary 2-deficient 3D boards of size $K \times L \times M$, where $KLM \equiv 2 \pmod 3$, and $K, L, M > 1$ are L-tilable. As in the case where one cube was removed, we extend the result to all higher dimensions. A sketch of the results in this paper was previously presented in [2].

L-Trominoes in 3D

We shall require the following result of Boltyanski and Soifer [7, Theorem 20.1], the proof of which is quite simple and is suggested as an exercise to the reader.

Lemma 1. *Any board of size $K \times L \times M$ where $3 \mid KLM$ and $K, L, M \geq 2$ is L-tilable.*

We restate the main result from [3], which we shall make frequent use of. Our over-arching strategy, whenever possible, is to break our 2-deficient board into two 1-deficient boards and L-tile these 1-deficient boards separately.

Theorem 1 (Befumo and Lenchner [3])**.** *A $(K \times L \times M) - 1$ board, where $KLM \equiv 1 \pmod 3$ and $K, L, M > 1$ is always L-tilable.*

In addition, we utilize one of the results of Starr's [8] mentioned in the introduction that kicked off the topic of 2-deficient boards in 3D.

Theorem 2 (Starr [8])**.** *Any 3D cubical board of size N^3 for $N \equiv 2 \pmod 3$ with two cubes removed is L-tilable.*

If we denote our board in general terms as a $(K \times L \times M) - 2$ board, where $KLM \equiv 2 \pmod 3$ and $K, L, M > 1$, then the argument will be a triple induction on K, L and M. In this argument, there are four essentially "prime" cases that need to be argued independently to kick off the induction; these are the cases where all values of $K, L, M \in \{2, 5\}$, so the cases $(2 \times 2 \times 2) - 2$, $(2 \times 2 \times 5) - 2$, $(2 \times 5 \times 5) - 2$, and $(5 \times 5 \times 5) - 2$.

The "Prime" Cases. Of the cases $(2 \times 2 \times 2) - 2$, $(2 \times 2 \times 5) - 2$, $(2 \times 5 \times 5) - 2$, and $(5 \times 5 \times 5) - 2$, the cases $(2 \times 2 \times 2) - 2$ and $(5 \times 5 \times 5) - 2$ are handled by Starr's Theorem 2.

Lemma 2. *A $(2 \times 2 \times 5) - 2$ board is L-tilable.*

We think of the associated $2 \times 2 \times 5$ board as two parallel 2×5 board as in Figure 2. We establish this lemma via a series of successively more sophisticated observations, the first two of which are left to the reader.

Observation 1. A $(2 \times 5) - 1$ board is L-tilable as long as the missing square is not in the center column.

Observation 2. If a $(2 \times 2 \times 5) - 2$ board is missing a cell from each board then it is L-tilable if neither cell is in the center column.

Observation 3. If a $(2 \times 2 \times 5) - 2$ board is missing a corner cell then it is L-tilable

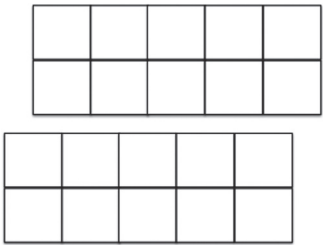


Figure 2 A $2 \times 2 \times 5$ board visualized as two parallel 2×5 boards.

Proof. It the board is missing two corner tiles then, as long as the two corner tiles do not come from the same row in the same board, we can slice the $2 \times 2 \times 5$ board into two 2×5 boards (in one of the two possible ways) such that each of the 2×5 boards are missing a corner square and be done by Observation 2. If only one of the missing cells comes from a corner cell, or if the two missing cells are both corner cells from the same row in the same board, we can assume one of the missing cells is the cell in Figure 3 and then place a first L-tromino as indicated in the same figure, leaving a $(2 \times 2 \times 4) - 1$ board, which is L-tilable by Theorem 1. ■

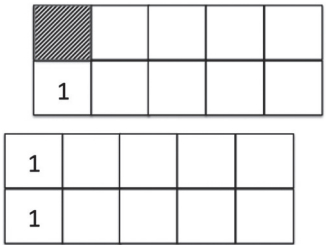


Figure 3 A $2 \times 2 \times 5$ board with exactly one corner cell removed, or two corner cells removed but the two cells being opposite cells on the same row of the same board. Placing an L-tromino in the cells labeled with the number 1, leaves a $(2 \times 2 \times 4) - 1$ board.

Observation 4. If both missing cells from a $(2 \times 2 \times 5) - 2$ board come from the same column then the board is L-tilable.

Proof. The case where they both come from an end column is covered by Observation 3. If they both come from a second-to-end column then the board can be factored into the disjoint union $(2 \times 2 \times 2) - 2 \sqcup 2 \times 2 \times 3$, i.e. of a $(2 \times 2 \times 2) - 2$ board and a $2 \times 2 \times 3$ board, the first of which is L-tilable by Theorem 2 and the second of which is L-tilable by Lemma 1.

Lastly, suppose both missing cells come from the middle column. Slicing the board appropriately we may assume that the missing tiles come from separate 2×5 boards. Use an additional L-tromino on the "top" 2×5 board to fill the third and fourth columns, and use another L-tromino on the "bottom" 2×5 board to fill the second and third columns, as in Figure 4. The remainder of the tiling is depicted in Figure 5. ■

Returning now to the proof of Lemma 2, the only remaining possibilities are where the two missing cells are either (a) from the 2nd and 4th columns on the same, say "top" 2×5 board, or (b) from the 2nd and 3rd columns (up to symmetry) on the same

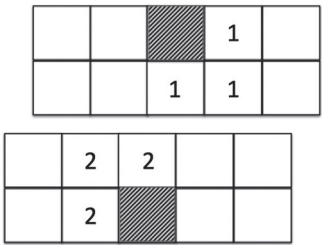


Figure 4 A $2 \times 2 \times 5$ board with two cells removed from the center column. We place a first additional L-tromino in the "top" 2×5 board to fill the third and fourth columns, and another on the "bottom" 2×5 board to fill the second and third columns.

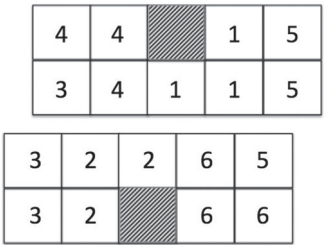


Figure 5 Completion of the tiling of a $2 \times 2 \times 5$ board with two cells removed from the center column.

or different boards. In case (a) we add L-trominoes to the "top" board to completely fill the first four columns and the use an additional L-tromino to complete the 5th column and take a single cell out of the 5th column of the "bottom" 2×5 board, which can then be L-tiled using Observation 1. We are thus left only with case (b). If now the missing cells are from the 2nd and 3rd columns on the same, say "top" board, then we can place three L-trominoes similarly to case (a) to again get an L-tilable $(2 \times 5) - 1$ board. Thus the only case left is where we are missing a cell from the 2nd column of the "top" board, say, and another cell missing from the 3rd column of the "bottom" board. But in this case we can place an L-tromino on the top board so that the 2nd and 3rd columns are filled, and place another L-tromino on the bottom board so that its 3rd and 4th columns are filled, leaving essentially the same situation as in Figure 4, and we can finish up as described in Figure 5. An arbitrary $(2 \times 2 \times 5) - 2$ board can thus be tiled and Lemma 2 is proved.

Lemma 3. *A* $(2 \times 5 \times 5) - 2$ *board is L-tilable.*

Proof. The first part of the proof is a bit of a foreshadowing of the inductive portion of the proof in the next section. As before we think of the $2 \times 5 \times 5$ board as two parallel 5×5 boards.

Suppose first that the two missing cells are both contained in the same $2 \times 4 \times 4$ board. We show that an arbitrary $(2 \times 4 \times 4) - 2$ board can be L-tiled. We can think of the $2 \times 4 \times 4$ board as four $2 \times 2 \times 2$ boards. Either the two missing cells are in separate $2 \times 2 \times 4$ boards or both cells come from a single $2 \times 2 \times 2$ cell. In the former case, the $(2 \times 2 \times 4) - 1$ boards can separately be L-tiled using Theorem 1. In the latter case, we can L-tile the $(2 \times 2 \times 2) - 2$ board using Starr's Theorem 2 leaving three $2 \times 2 \times 2$ boards in an L-shape. Call these boards A, B, and C with A connected to B and B connected to C. Now place an L-tromino so that it covers two cells of A and one cell of B and another L-tromino so that it covers two cells of C and a cell of B distinct from that covered by the previous L-tromino. This leaves A, B, and

C as $(2 \times 2 \times 2) - 2$ boards that can be L-tiled using Starr's Theorem 2. Now that the $(2 \times 4 \times 4) - 2$ portion of the $(2 \times 5 \times 5) - 2$ board can be L-tiled, L-tile the rest as in Figure 6.

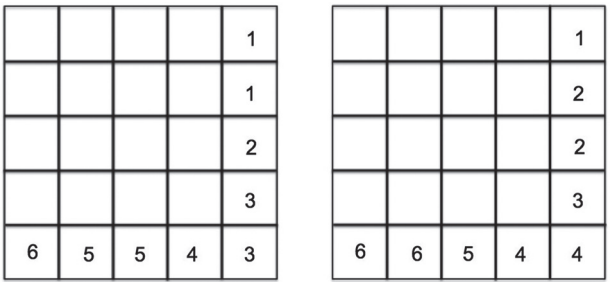


Figure 6 Completion of the tiling of a $(2 \times 5 \times 5) - 2$ board under the assumption that the two missing cells are from some $2 \times 4 \times 4$ subboard.

If the two missing cells are not both contained in the same $2 \times 4 \times 4$ board then the board can be factored into the disjoint union of a $(2 \times 5 \times 4) - 1$ board and a $(2 \times 5 \times 1) - 1$ and we write that $(2 \times 5 \times 5) - 2 = (2 \times 5 \times 4) - 1 \sqcup (2 \times 5 \times 1) - 1$. We know how to L-tile a $(2 \times 5 \times 4) - 1$ by Theorem 1 and we also know that the only problematic $(2 \times 5 \times 1) - 1 = (2 \times 5) - 1$ boards are when the missing tile is in the center column (Observation 1)—see Figure 7.

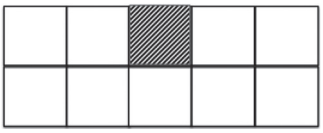


Figure 7 The only impossible to tile $(2 \times 5) - 1$ board.

Now in the factoring $(2 \times 5 \times 5) - 2 = (2 \times 5 \times 4) - 1 \sqcup (2 \times 5 \times 1) - 1$, the only way we can be forced into having a "bad" $(2 \times 5 \times 1) - 1$ board, i.e., a board like that in Figure 7 is if the two missing cells are either opposite central cells on distinct 5×5 boards or on the same 5×5 board, i.e., the two situations depicted in Figure 8.

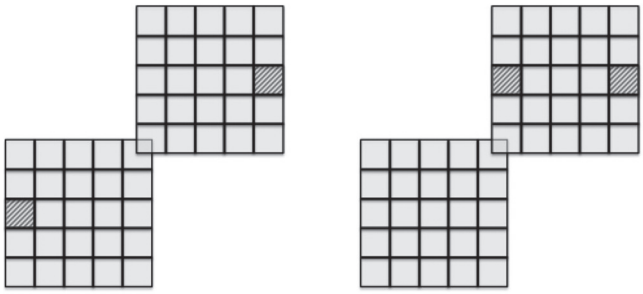


Figure 8 The two remaining problematic $(2 \times 5 \times 5) - 2$ boards: case (a) where the missing opposite central cells are on different 5×5 boards, and case (b) where they are on the same 5×5 board.

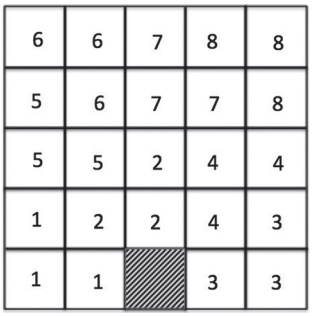


Figure 9 Demonstration that a $(5 \times 5) - 1$ board with a missing central edge square is L-tilable.

In case (a) we can actually tile each of the $(5 \times 5) - 1$ sub-boards independently. Although a $(5 \times 5) - 1$ board is not always tilable, such a board with a middle edge square as the missing square is in fact L-tilable, as Figure 9 illustrates. Case (b) is actually quite tricky but a tiling can be arrived at as indicated in Figure 10. ■

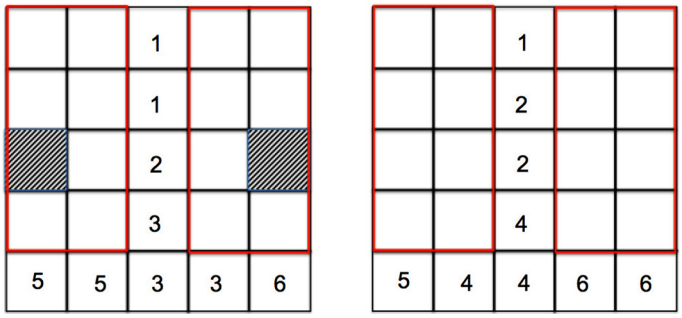


Figure 10 Demonstration of an L-tiling of a $(5 \times 5) - 2$ board where the two missing tiles are from opposite central edge squares on the same 5×5 sub-board: First, break the board into two $(4 \times 2 \times 2) - 1$ sub-boards and two-dimensional "T." Then L-tile the two $(4 \times 2 \times 2) - 1$ sub-boards using Theorem 1 and L-tile the "T" as shown.

Completing the argument in 3D. Again, if we denote our board in general terms as a $(K \times L \times M) - 2$ board, where $KLM \equiv 2 \pmod 3$ and $K, L, M > 1$ then the argument will be a triple induction on K, L, and M.

First let us suppose that there is just one of the quantities from among the K, L, M that is not in $\{2, 5\}$, and without loss of generality suppose it is K. Since $K \equiv 2 \pmod 3$, we must have $K \geq 8$. Now either some $(K - 3) \times L \times M$ sub-board contains both missing cells and we break the board up into a disjoint union of this $(K - 3) \times L \times M$ board and a $3 \times L \times M$ board—for short we shall refer to this split using the shorthand $K \times L \times M = (K - 3) \times L \times M \sqcup 3 \times L \times M$, or we can break the board up into the disjoint union $(K - 4) \times L \times M \sqcup 4 \times L \times M$ where each sub-board contains one missing cell. In the former case we use induction to L-tile the 2-deficient $(K - 3) \times L \times M$ board and Lemma 1 to L-tile the $3 \times L \times M$ board. In the latter case, we can tile each of the two 1-deficient boards using Theorem 1.

Next suppose that some (at least) two of the $K, L, M \notin \{2, 5\}$ and argue by double induction. Say the two are L, M, i.e., $L, M \in \{4, 7, 8, 10, \ldots\}$.

If now $L \equiv 1 \pmod 3$, write

$$L = L_0 + L_1, \text{ where } L_0, L_1 \equiv 2 \pmod 3, \text{ i.e., } L_0 = L - 2, L_1 = 2,$$

while if $L \equiv 2 \pmod 3$, write

$$L = L_0 + L_1, \text{ where } L_0, L_1 \equiv 1 \pmod 3, \text{ i.e., } L_0 = L - 4, L_1 = 4.$$

Similarly write $M = M_0 + M_1$ using the same convention.

Then we have two ways of splitting the solid, either

$$K \times L \times M = K \times L_0 \times M \sqcup K \times L_1 \times M, \text{ or} \tag{1}$$

$$= K \times L \times M_0 \sqcup K \times L \times M_1 \tag{2}$$

where $KL_0M, KL_1M, KLM_0, KLM_1 \equiv 1 \pmod 3$, since $KLM \equiv 2 \pmod 3$ and none of these factors are congruent to 0 (mod 3)—the summands must therefore each be congruent to 1 (mod 3). Additionally, we have the further decomposition of $K \times L \times M$ into

$$K \times L_0 \times M_0 \sqcup K \times L_0 \times M_1 \sqcup K \times L_1 \times M_0 \sqcup K \times L_1 \times M_1, \tag{3}$$

where $KL_0M_0, KL_0M_1, KL_1M_0, KL_1M_1 \equiv 2 \pmod 3$

We illustrate these decompositions in Figure 11 for the case $K \times L \times M = 2 \times 4 \times 7$, which we visualize as two parallel copies of a 4×7 board. If we make both the

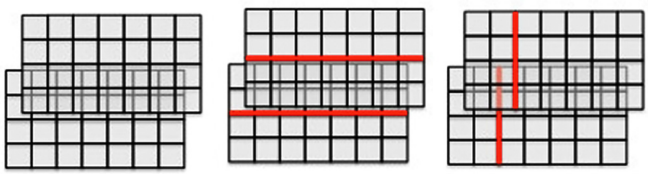


Figure 11 A $2 \times 4 \times 7$ board visualized as two parallel 4×7 boards (left). The same board visualized with a cut corresponding to the decomposition $2 \times 4 \times 7 = 2 \times 2 \times 7 \sqcup 2 \times 2 \times 7$ (center) and $2 \times 4 \times 7 = 2 \times 4 \times 2 \sqcup 2 \times 4 \times 5$ (right).

cuts illustrated in this figure we get a decomposition $2 \times 4 \times 7 = 2 \times 2 \times 2 \sqcup 2 \times 2 \times 2 \sqcup 2 \times 2 \times 5 \sqcup 2 \times 2 \times 5$, which corresponds to the more general equation (3).

Now note that if we remove two cells from the $K \times L \times M$ board that either they belong one to each side of the split associated with equations (1) or (2), OR both cells belong to one of the smaller sub-boards corresponding to equation (3). In the first case we can tile the two sub-boards, and hence the entire $(K \times L \times M) - 2$ board using Theorem 1. In the latter case, the small sub-board can be L-tiled by induction and it remains to L-tile the remaining three sub-boards. We denote these sub-boards, for convenience by A, B, C with A connected to B and B connected to C.

We L-tile $A \sqcup B \sqcup C$ as follows. Place an L-tromino so that it takes two cells from the face of A that is adjacent to B and one cell from the face of B that is adjacent to A. Analogously take two cells out of the face of C that is adjacent to B and once cell from the face of B that is adjacent to C, taking care not to take the same cell as earlier (there are at least 4 cells in these faces so this cannot be a problem). If we take away the cells with these trominoes, A, B, and C are each left with two cells removed and then can be further L-tiled by induction, completing the tiling and the argument. We can thus state the following theorem.

Theorem 3. *A $(K \times L \times M) - 2$ board, where $KLM \equiv 2 \pmod 3$ and $K, L, M > 1$ is always L-tilable.*

L-Trominoes in Higher Dimensions

We can use the same trick as was used in [3] to embed a 2-deficient board in 4D as a 2-deficient board in 3D such that all adjacencies of cells in the 3D board are adjacencies in the 4D board. Figure 12 provides an illustration of how one can embed $4 \times 2 \times 2 \times 5$

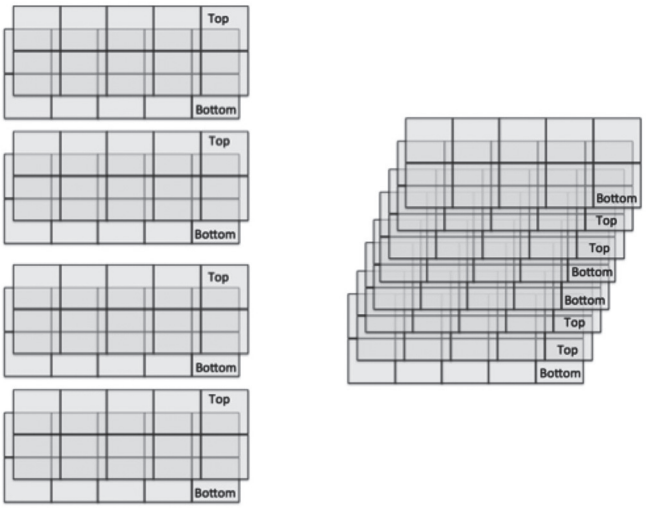


Figure 12 On the left, a $4 \times 2 \times 2 \times 5$ board thought of as four parallel $2 \times 2 \times 5$ boards with squares in a given location on each of the respective boards thought of as being connected to the associated square on the board above and board below, if such boards exist. If we reorder the boards of the top $2 \times 2 \times 5$ board and slide them over the second-from-top $2 \times 2 \times 5$ board, and reorder the boards of the third-from-top board and do the same, we get an embedding of the $4 \times 2 \times 2 \times 5$ board into an $8 \times 2 \times 5$ board such that any two squares that are adjacent in the $8 \times 2 \times 5$ board were adjacent in the $4 \times 2 \times 2 \times 5$ board prior to the embedding.

board into an $8 \times 2 \times 5$ board. This embedding can be extended inductively to higher and higher dimensions. For example, in 5D, if we have a $K_1 \times K_2 \times K_3 \times K_4 \times K_5$ board, we think of it as K_1 parallel $K_2 \times K_3 \times K_4 \times K_5$ boards. Each $K_2 \times K_3 \times K_4 \times K_5$ board can be thought of as a $K_2K_3 \times K_4 \times K_5$ board with fewer adjacencies, but of all of whose adjacencies are adjacencies in the higher dimensional board, and hence all K_1 such boards can analogously be thought of as a $K_1K_2K_3 \times K_4 \times K_5$ board with the same property. We thus obtain the following generalization of Theorem 3.

Theorem 4. *A* $(K_1 \times \cdots \times K_N) - 2$ *board, where* $K_1 \cdots K_N \equiv 2 \pmod 3$, *for* $N \geq 3$ *and some three of the* $K_i > 1$, *is always L-tilable.*

Future Work

It is not possible to L-tile an arbitrary 3-deficient board in 3D since one can remove three cells and isolate the corner cell. However, we conjecture that it is always possible to L-tile a 3-deficient board in four and higher dimensions for board-sizes of the needed modularity, and wonder if the isolation of the corner cell is the only way to obstruct an L-tiling in 3D.

As a result of [3] and the present work we see that an L-tromino is remarkably efficient for tiling purposes. In [1] we showed that the solid 7-omino obtained by removing a cell from a solid $2 \times 2 \times 2$ cubical array can be used to tile an arbitrary $(2^N \times 2^N \times 2^N) - 1$ board. Is the tiling of $(2^N \times 2^N \times 2^N) - 1$ boards an anomaly or can this 7-omino tile much more general boards, and perhaps all rectangular boards of the needed modularity in four and higher dimensions?

To conclude, we pose the following problem: Find a polyomino of dimension k that cannot be used to tile any rectangular board in dimension k or higher.

REFERENCES

[1] Befumo, A., Lenchner, J. (2014). Extensions of Golomb's tromino theorem. Presented at The 24th Annual Fall Workshop in Computational Geometry, University of Connecticut, Storrs, CT, October 31.
[2] Befumo, A., Lenchner, J. (2016). Tiling two-deficient rectangular solids with trominoes in three and higher dimensions. Presented at The 26th Annual Fall Workshop in Computational Geometry, CUNY Graduate Center, New York, NY, October 27.
[3] Befumo, A. Lenchner, J. (2018). Tiling one-deficient rectangular solids with trominoes in three and higher dimensions, *Math. Mag.* 91: 62–69.
[4] Chu, I., Johnsonbaugh, R. (1985–86). Tiling boards with trominoes, *J. Recreat. Math.* 18: 188–193.
[5] Chu, I., Johnsonbaugh, R. (1986). Tiling deficient boards with trominoes, *Math. Mag.* 59: 34–40.
[6] Golomb, S. W. (1954). Checker boards and polyominoes, *Amer. Math. Monthly.* 61: 675–682.
[7] Soifer, A. (2010). *Geometric Etudes in Combinatorial Mathematics*, 2nd ed. New York: Springer.
[8] Starr, N. (2008). Tromino tiling deficient cubes of side length 2^n, *Geombinatorics* XVIII (2): 72–87.

Summary. In a recent article for THIS MAGAZINE [3], it was shown that if any single cell is removed from a $K_1 \times \cdots \times K_N$ N-dimensional chess board, for $N \geq 3$, where $K_1 \cdots K_N \equiv 1 \pmod 3$ and at least three of the $K_i > 1$, then the remaining board can always be tiled using solid L-trominoes of the appropriate dimension. In this article we show that the same is true when two cells are removed from such chess boards, as long as $K_1 \cdots K_N \equiv 2 \pmod 3$.

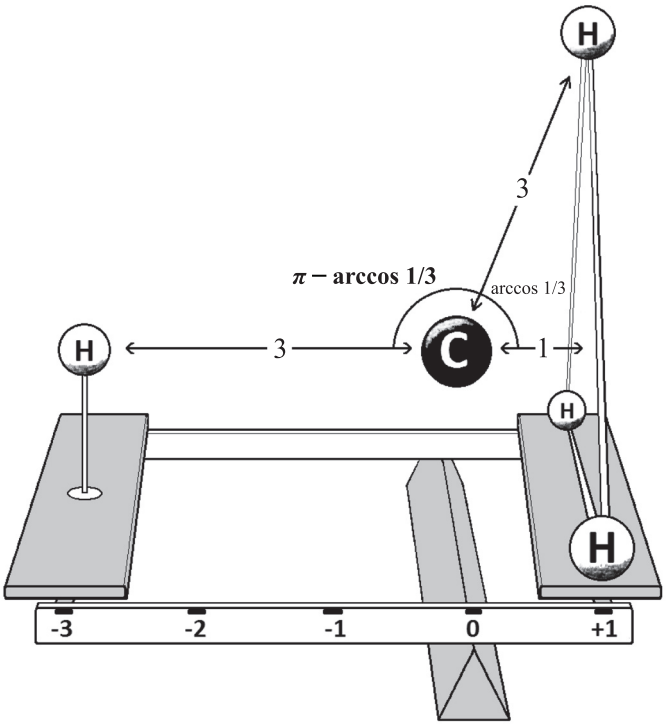
ARTHUR BEFUMO (MR Author ID: 1183745) at age 14 simultaneously enrolled in the University of Montana and Hellgate High School, opting to take all of his math classes at the University. During his junior year Arthur was first introduced to Golomb's Tromino theorem. He is currently pursuing a B.S. in Computer Science and Mathematics at Yale University, Class of 2019. In his free time Arthur fences, composes electronic music, and avoids complicated integration like the plague.

JONATHAN LENCHNER (MR Author ID: 782105) earned a Ph.D. in mathematics late in life from Polytechnic University (now part of the NYU Courant Institute) in 2008. He has spent almost 20 years at IBM and became Chief Scientist of IBM's African research labs, one in Nairobi, the other in Johannesburg, in May of 2016. He learned about Golombs theorem from Arthur on a family visit, while Arthur was still in high school.

Proof Without Words: Tetrahedral Bond Angle

ADRIAN CHUNPONG CHU
The Chinese University of Hong Kong
adrianchu001@gmail.com

From this steady configuration of four identical white balls (supported by weightless sticks) on a balance, simulating the methane molecule CH_4, we conclude that the angle subtended by any two vertices of a regular tetrahedron through its center is $\pi - \arccos 1/3$.



Summary. The angle subtended by any two vertices of a regular tetrahedron through its center is shown to measure $\pi - \arccos 1/3$.

ADRIAN CHUNPONG CHU (MR Author ID: 1240970) is an undergraduate student at The Chinese University of Hong Kong. He is interested in minimal surface theory and complex geometry.

Math. Mag. **92** (2019) 41. doi:10.1080/0025570X.2018.1538278
MSC: Primary 51M04

Sectorial Covers for Unit Arcs

JOHN E. WETZEL
1601 Lakeside Drive Unit A
Champaign, IL 61821
jewetz@comcast.net

WACHARIN WICHIRAMALA
Chulalongkorn University
Bangkok 10330, Thailand
wacharin.w@chula.ac.th

In 1966, Moser [5] asked for the smallest (convex) set in the plane that contains a congruent copy of each planar arc of unit length (see also Moser [6]), a problem known as "the classic worm problem" that remains unsolved.

More generally, one can seek a planar set, perhaps of prescribed shape, that can accommodate a congruent copy of each of the arcs in any prescribed collection and whose area, or thickness, or diameter, or whatever, is extreme in some sense. Little is known in general about such so-called "worm problems," although there is an extensive literature on a wide variety of special cases. (See, e.g., [10], [12], [2, Section 7.4] and [3].) A set that contains a congruent copy of every arc in a family is called a *cover* for that family. Very few worm problems have been solved, because it is generally quite difficult to show that a set is a cover. Indeed, if one calls the set a *can* and the arcs *worms*, then posing a worm problem is truly opening a can of worms.

The smallest convex cover currently known for the family $\mathcal{F}$ of all unit arcs in the plane was constructed in 2006 by Wang [9]. Its area is about 0.2709. In the early 1970s, it was conjectured ([10], [8, p. 161] and [11, p. 358]) that the 30° circular sector with radius 1, which with area about 0.2618 is a little smaller than Wang's cover, is also a cover for $\mathcal{F}$. This conjecture remains unsettled, although partial results have been established including the fact that every *convex* unit arc (and consequently every drapeable unit arc) fits (see [7]).

Concerning sectors, an early result was given in 1973 [10, Theorem 5], where it is shown that for $0° < \vartheta \le 90°$ a sector of angle ϑ and radius $\frac{1}{2}\csc\frac{1}{2}\vartheta$ can accommodate each unit arc. Hence, a 30° sector with radius at least $r = \frac{1}{2}\csc 15° \approx 1.932$ is a cover for the family $\mathcal{F}$ of all unit arcs in the plane.

Here, we use a generalization of a useful lemma in [1] to improve this result and show that a circular sector with angle ϑ, $0° < \vartheta < 90°$, and radius $\csc 2\vartheta$ is a cover for $\mathcal{F}$. In particular, a 30° sector with radius $r = \frac{2}{3}\sqrt{3} \approx 1.155$ is a cover for $\mathcal{F}$.

Preliminaries. To show that a compact set is a cover for the family of all open arcs of unit length, it suffices to show that it contains a congruent copy of every open simple polygonal arc of unit length (cf. [13, Corollary 5]).

Let $\Pi(r, \vartheta)$ be the circular sector of radius r and central angle ϑ. Suppose an open (i.e., different endpoints), simple (i.e., not self-intersecting), polygonal arc γ of unit length is given, parametrized by arclength, with endpoints $P = \gamma(0)$ and $Q = \gamma(1) \neq P$. Denote by H the closed convex hull of γ.

If a ray supports the hull H of a polygonal arc, then it meets that hull only in isolated points and closed intervals.

Write $X \prec Y$ (or $Y \succ X$) when X precedes Y in the parametric ordering of points on γ, and $X \preceq Y$ (or $Y \succeq X$) if the equality is permitted; and write $A \sim B \sim C$ when

Math. Mag. **92** (2019) 42–46. doi:10.1080/0025570X.2018.1523648
MSC: Primary 52A10, Secondary 51N20; 52C15
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

B is parametrically between A and C, i.e., when $A \prec B$ and $B \prec C$, *or* $C \prec B$ and $B \prec A$. When $X \prec Y$, we denote the subarc of γ from X to Y by γ_{XY}.

Coulton and Movshovich [**1**, Theorem 5.1, p. 86ff] showed in 2006 that an open simple polygonal arc (not a line segment) has parallel support lines ℓ and m and points A, B, and C so that both A and C lie on one line, B lies on the other, and $A \sim B \sim C$. We need an analogous result for arcs in an angle.

For coterminal rays u and v write $\angle u, v$ for both the size (i.e., degree measure) of the angle and also for the closed angular region bounded by the rays. Throughout let γ be an open simple polygonal arc of unit length.

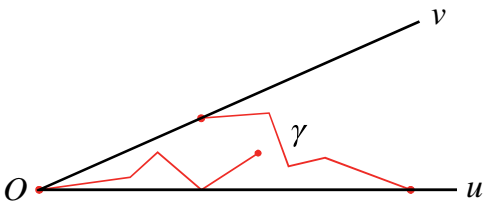


Figure 1 The arc γ fits in $\angle u, v$ but not properly.

We say that γ *fits properly in* $\angle u, v$ if it fits in $\angle u, v$ with a point on each of the rays u and v but it cannot fit in $\angle u, v$ touching the vertex O. Figure 1 shows an example of an arc γ that fits in $\angle u, v$ but does not fit properly.

Definition. An arc γ *in* $\angle u, v$ is in Λ-position if it can be placed in $\angle u, v$ in such a way that there are three different points A, B, C, of γ with A and C on one ray, B on the other ray, and $A \sim B \sim C$. We say that the points A, B, C of γ are in Λ-position.

Lemma. *If* γ *fits properly in* $\angle u, v$, *then it is in* Λ*-position in* $\angle u, v$.

Proof. There is no loss of generality to suppose that γ touches u before it reaches v. Let X be the *last* point of γ on u for which γ_{PX} does not touch v, and let Y be the first point on γ on v. Then $P \preceq X \prec Y \preceq Q$. The subarc γ_{XY} divides the interior of $\angle u, v$ into two open sets, and without loss of generality we suppose that P lies on the same side as the vertex O. There are different situations depending on whether the endpoint Q is on the same or opposite side of γ_{XY}.

Case 1. Suppose the endpoints P and Q are on opposite sides of γ_{XY}. Regarding the given angle as rigid but movable, like a pair of angle calipers with fixed opening and rays u and v as sides, we turn $\angle u, v$ about the hull H in the sense that moves v toward P, keeping the rays u and v as support rays of H (and consequently of γ), and we watch what happens. Denote the moving angle $\angle u', v'$. Either (a) one ray, u' or v', meets ∂H in a segment AC and the other ray meets ∂H at a point B of γ_{AC}, in which case A, B, and C are in Λ-position (Figure 2); or (b) one ray meets ∂H in a segment AC and γ_{AC} does not meet the other ray, in which case the turning can be continued; or (c) both rays meet ∂H at single points of γ, in which case the rotation can again be continued. Because γ has just finitely many vertices and the rotation is cyclical, the process must find points A, B, and C in Λ-position.

Case 2. Suppose both P and Q are on the same side of γ_{XY} as O. Denote the moving angle by $\angle u', v'$ as before. Turn $\angle u, v$ about the hull H in the sense that moves v toward P. Then one of the following two possible situations must occur (Figure 3).

(a) If the turned angle $\angle u', v'$ has one ray that supports H with a contact point A on γ_{PX} and a contact point C on γ_{YQ}, and if the opposite ray has a contact point B on γ_{XY}, then A, B, and C are in Λ-position in $\angle u', v'$.

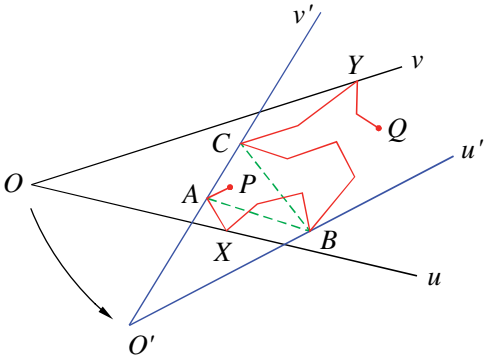


Figure 2 Endpoints on opposite sides of γ_{XY}.

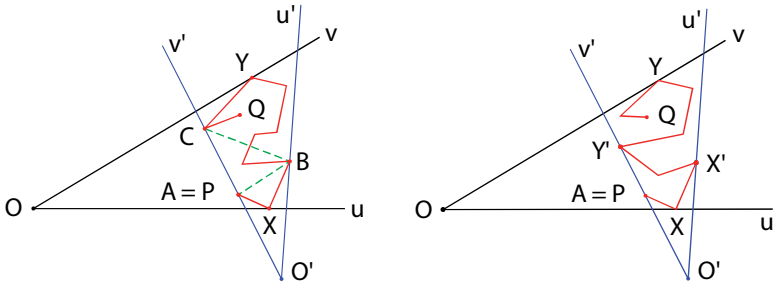


Figure 3 Same side, cases (a) and (b), left and right, respectively.

(b) If for the turned angle $\angle u', v'$, P and Q are on opposite sides of $\gamma_{X'Y'}$, where X' is the last point of γ on u' and Y' is the first point of γ on v', then the desired conclusion follows from Case 1.

Similar arguments work when both P and Q are initially across γ_{XY} from O. ■

Our main result follows from this lemma.

Theorem. *The circular sector* $\Pi(r, \vartheta)$ *with angle* ϑ *in* $(0°, 90°]$ *and radius* $r = \csc 2\vartheta$ *is a cover for* F.

Proof. It is enough to show that simple polygonal arcs fit. (Detailed proofs of this basic result can be found in [4] and [13]). Let U and V be the points on u and v, respectively, so that $OU = OV = r$, and let Γ be the minor arc of radius r with endpoints U and V. Let γ be an open unit polygonal arc in $\angle u, v$. If γ touches O, then it obviously fits in $\Pi(1, \vartheta) \subseteq \Pi(r, \vartheta)$.

By way of contradiction, assume that γ does not fit in $\Pi(r, \vartheta)$. Then γ does not touch O, and according to the lemma there are points A, B, and C on γ in Λ-position. We choose the notation so that A and C are on u, with A between O and C.

The distance h from V to u is $r \sin \vartheta$. If B lies across Γ from O, then $1 = \ell(\gamma) > 2h = 2r \sin \vartheta$, so that $r < 1/(2 \sin \vartheta) < \csc 2\vartheta$, a contradiction.

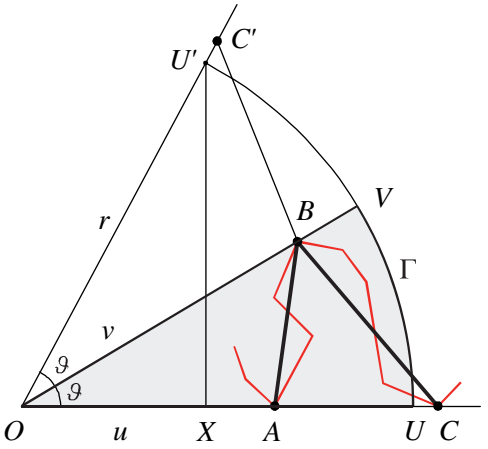


Figure 4 $O \sim U \sim C$.

Otherwise, B lies on the segment OV. If C lies beyond Γ on u (Figure 4), let C' be the reflection of C across the ray v and X the projection of U' on u. Then

$$1 = r \sin 2\vartheta = U'X < AB + BC' = AB + BC \le \ell(\gamma) = 1,$$

a contradiction. So A and C lie on the segment OU and B lies on the segment OV.

Because u and v support γ and γ does not fit in Π, there is a point D of γ in the angle $\angle u, v$ across Γ from O. There are just two possible locations for D: (1) between B and C or (2) beyond C on γ.

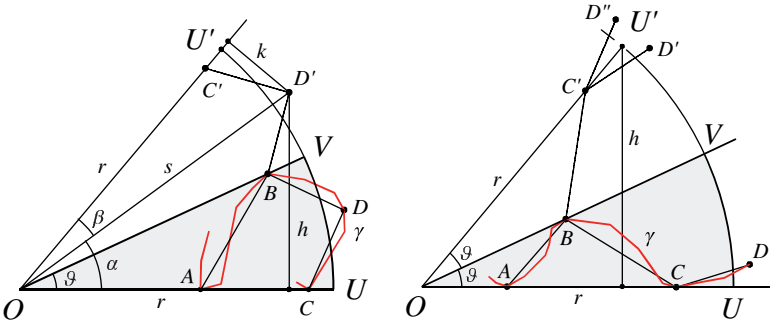


Figure 5 Case 1, on the left; Case 2, on the right.

Case 1. Suppose there is a point D between B and C and across Γ from O (Figure 5, left). Let OU' and $BC'D'$ be the mirror images of OU and the polygonal arc BCD in the line OV, and let $\alpha = \angle UOD'$, $\beta = \angle D'OU'$, and $s = OD'$. Note that $s > r$. Let h and k be the distances from D' to the ray u and the reflected ray u', respectively. Then $AB + BD' \ge h = s \sin\alpha$ and $C'D' \ge k = s \sin\beta$. Consequently, we have the contradiction

$$\begin{aligned} 1 &\le r \sin 2\vartheta < r \sin\alpha + r \sin\beta < h + k \\ &\le \big(AB + BD'\big) + D'C' = AB + BD + DC \le \ell(\gamma) = 1. \end{aligned}$$

Case 2. Suppose there is a point D on γ beyond C and across Γ from O (Figure 5, right). Let OU' and $BC'D'$ be the mirror images of OU and the polygonal arc BCD

in the line OV, and let $C'D''$ be the mirror image of the segment $C'D'$ in the line OU'. Because $O'' > r$ and $\operatorname{dist}(D'', u) > k$, we have the contradiction

$$1 = \ell(\gamma) \geq AB + BC + CD = AB + BC' + C'D'' > h = r \sin 2\vartheta = 1.$$

This completes the argument. ■

Acknowledgments We thank Pongbunthit Tonpho for his assistance and Dr. Y. Movshovich for her help and encouragement on this work. Wichiramala is partially supported by the 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

REFERENCES

[1] Coulton, P., Movshovich, Y. (2006). Besicovitch triangles cover unit arcs. *Geom. Dedicata.* 123: 79–88. doi.org/10.1007/s10711-006-9107-7
[2] Falconer, K. J. (1985). *The Geometry of Fractal Sets.* Cambridge, UK: Cambridge University Press.
[3] Füredi, Z., Wetzel, J. E. (2011). Covers for closed curves of length two. *Period. Math. Hungar.* 63: 1–17. doi.org/10.1007/s10998-011-7001-z
[4] Maki, J. M., Wetzel, J. E., Wichiramala, W. (2005). Drapeability. *Discret. Comput. Geom.* 34: 637–657. doi.org/10.1007/s00454-005-1189-8
[5] Moser, L. (undated). Poorly formulated unsolved problems in combinatorial geometry, 9pp., mimeographed, undated, but about 1966.
[6] Moser, W. O. J. (1991). Problems, problems, problems. *Discret. Appl. Math.* 31: 201–225. doi.org/10.1007/s00454-005-1189-8
[7] Movshovich, Y., Wetzel, J. E. (2017). Drapeable unit arcs fit in the unit 30° sector. *Adv. Geom.* 17: 497–506.
[8] Norwood, R., Poole, G., Laidacker, M. (1992). The worm problem of Leo Moser. *Discret. Comput. Geom.* **7**: 153–162. doi.org/10.1007/BF02187832
[9] Wang, W. (2006). An improved upper bound for worm problem. *Acta Math. Sin. (Chin. Ser.)* 49: 835–846.
[10] Wetzel, J. E. (1973). Sectorial covers for curves of constant length. *Canad. Math. Bull.* 16: 367–375.
[11] ——— (2013). Fits and covers. *Math. Mag.* 76: 349–363, 398.
[12] ——— (2013). Bounds for covers of unit arcs. *Geombinatorics.* XXIII: 116–122.
[13] Wetzel, J. E., Wichiramala, W. (2010). A covering theorem for families of sets in R^d. *J. Combin.* 1: 69–76. doi.org/10.4310/JOC.2010.v1.n1.a5

Summary. It has been conjectured in conjunction with Leo Moser's "worm" problem that a circular sector with angle 30° and radius 1 contains a congruent copy of each unit arc in the plane. If this is true, this sector would be the smallest such set currently known. This conjecture remains unsettled.

We show that for $0° < \vartheta < 90°$ the circular sector with vertex angle ϑ and radius $r = \csc 2\vartheta$ contains a congruent copy of each unit arc. For $\vartheta = 30°$, this means that the radius $r = \frac{2}{3}\sqrt{3} \approx 1.1548$ suffices. Our arguments depend on an interesting extension to angles of a result of Coulton and Movshovich about parallel support lines of an arc.

JOHN E. WETZEL (MR Author ID: 182185) did his undergraduate work at Purdue University, received his Ph.D. in mathematics from Stanford University in 1964, and spent his academic career at the University of Illinois, from which he retired in 1999. Always interested in classical geometry, he wonders about the ways in which shapes can fit in other shapes.

WACHARIN WICHIRAMALA (MR Author ID: 694359) did his undergraduate work at Chulalongkorn University, received his Ph.D. in mathematics from University of Illinois at Urbana–Champaign in 2002, and has spent his academic career at Chulalongkorn University. After proving the planar triple bubble conjecture in his doctoral thesis, he studied and fell in love with Moser's worm problem. Since then he has been cursed to obsess with solving Wetzel's conjectures on the problem.

Oscillating Functions that Disprove Misconceptions on Real-Valued Functions

JÜRGEN GRAHL
University of Würzburg
97074 Würzburg, Germany
grahl@mathematik.uni-wuerzburg.de

SHAHAR NEVO
Bar-Ilan University
Ramat-Gan 52900, Israel
nevosh@math.biu.ac.il

When preparing a lecture on basic calculus you are sometimes reminded of Hamlet's famous words "There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy." In this spirit, the purpose of this paper is to draw the reader's attention to some surprising observations on differentiable functions of one variable which have arisen from our own experience as lecturers. Some of them seem to be little-known or even unknown, others can already be found in at least one of the many excellent calculus textbooks, but are included since we consider them to be particularly instructive. We hope that this note is of interest to faculty who want to flavor their lectures with some unexpected or even weird examples and to students who want to learn about such examples. In this context we also refer the reader to the books of Appell [2], Gelbaum and Olmsted [7], and Rajwade and Bhandari [10] which contain rich collections of counterexamples and peculiarities from real analysis.

The unifying theme in our counterexamples is the use of "oscillating" functions like $x \mapsto x^k \sin \frac{1}{x^m}$ (where $k \geq 0$, $m \geq 1$) and related functions. They are a popular source of enlightening counterexamples in basic calculus, see, for example, [7, Chapter 3]) and for a more detailed discussion of the properties of those functions [2, Section 4.5], [3, 4]. For instance, the function $x \mapsto x^2 \sin \frac{1}{x^2}$ shows that derivatives are not necessarily continuous, that they can even be unbounded on compact intervals, and that differentiable functions need not be rectifiable on compact intervals, i.e., their graphs can have infinite length. From the point of view of complex analysis, the interesting properties of these functions are of course related to the fact that $z \mapsto z^k \sin \frac{1}{z^m}$ has an essential singularity at $z = 0$.

We present some further applications of functions of this kind that disprove popular misconceptions on the behavior of differentiable real functions. Some of those misconceptions seem to arise from the very same origin: Many students tend to think of differentiable functions as being piecewise monotonic and are unaware of the fact that the monotonicity behavior might change infinitely many times within a compact interval. Naturally, oscillating functions like $x \mapsto \sin \frac{1}{x}$ are apt to show those students that their imagination might mislead them. This central idea will appear in several of our examples, either obviously (Examples 2 and 3) or in a somewhat "disguised" form (Example 5), in the sense that not the function but its derivative has accumulating changes of monotonicity (and hence accumulating critical points).

Example 1. Expressing monotonicity in terms of derivatives is straightforward. A differentiable function $f : I \to \mathbb{R}$ on an interval I is nondecreasing if and only if $f'(x) \geq 0$ for all $x \in I$. Things are a bit less straightforward for *strict* monotonicity.

Math. Mag. **92** (2019) 47–57. doi:10.1080/0025570X.2018.1537415
MSC: Primary 26A06, Secondary 26A24; 97I40

The condition $f'(x) > 0$ for all $x \in I$ is sufficient but not necessary for f being strictly increasing. In fact, as known from elementary calculus, f is strictly increasing if and only if $f'(x) \geq 0$ for all $x \in I$ and if every proper interval $J \subseteq I$ contains a point x such that $f'(x) > 0$.

In particular, there must be strictly increasing functions which have infinitely many critical points within a compact interval. An example is provided by the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \int_0^x |t| \cdot \left(1 + \cos\frac{1}{t}\right) dt \qquad \text{for all } x \in \mathbb{R}$$

which is continuously differentiable and whose critical points are 0 and the points $\frac{1}{(2n+1)\pi}$ (where $n \in \mathbb{N}$) which accumulate at the origin. See Figure 1(a).

In [8, p. 186] a similar example is given. Define $f : [0, 1] \to \mathbb{R}$ by $f(0) = 0$ and

$$f(x) := x \cdot (2 - \sin(\log(x)) - \cos(\log(x))) \qquad \text{for } 0 < x \leq 1.$$

An easy calculation shows that this function is strictly increasing with critical points at $e^{-2\pi n}$ ($n \in \mathbb{N}$). However, at $x = 0$ it is only continuous, but not differentiable.

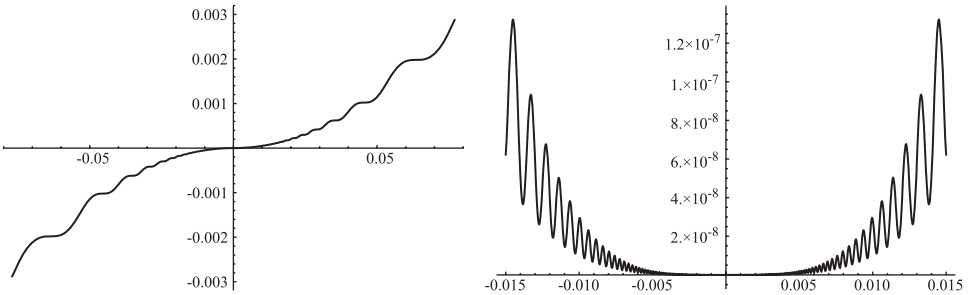


Figure 1 (a) $x \mapsto \int_0^x |t| \cdot \left(1 + \cos\frac{1}{t}\right) dt$ (b) $x \mapsto x^4\left(2 + \sin\frac{1}{x}\right)$.

It's worth noting that the monotonicity criterion mentioned above can be extended as follows. Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Assume that there is a countable set S such that f is differentiable on $[a, b] \setminus S$ and $f'(x) > 0$ for all $x \in [a, b] \setminus S$. Then f is strictly monotonically increasing on $[a, b]$.

An amazingly short proof for this classical result was given by Zalcman [12].

Example 2. If $f : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function with a strict local minimum at some point x_0, then most students tend to expect that there is a small neighborhood $(x_0 - \delta, x_0 + \delta)$ such that f is decreasing on $(x_0 - \delta, x_0]$ and increasing on $[x_0, x_0 + \delta)$.

The function [7, p. 36]

$$f(x) := \begin{cases} x^4\left(2 + \sin\frac{1}{x}\right) & \text{for } x \in \mathbb{R} \setminus \{0\}, \\ 0 & \text{for } x = 0 \end{cases}$$

shows that this is incorrect (Figure 1(b)). The function f has a strict local minimum at 0, and f is even continuously differentiable on $\mathbb{R}$ with $f'(0) = 0$, but f' assumes both positive and negative values in any neighborhood of 0 as we can easily see from

$$f'(x) = x^2 \cdot \left(8x + 4x \sin\frac{1}{x} - \cos\frac{1}{x}\right) \qquad \text{for } x \neq 0.$$

Example 3. In some sense, our next example is a counterpart to the last one which dealt with the monotonicity behavior near a critical point. Now we look at the situation that $f : \mathbb{R} \to \mathbb{R}$ is differentiable and $f'(x_0) \neq 0$ for some $x_0 \in \mathbb{R}$. This condition does **not** imply that f is monotonic in some neighborhood of x_0. This is illustrated by the function [7, p. 37]

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) := \begin{cases} x + \alpha x^2 \sin \frac{1}{x^2} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0, \end{cases} \qquad \text{where } \alpha \neq 0$$

which is differentiable everywhere with $f'(0) = 1 \neq 0$, but is not monotonic (hence not one-to-one) on any neighborhood of 0 since f' assumes both positive and negative values there. In fact, f' is unbounded in both directions on each such neighborhood (see Figure 2). Of course this example also shows that an accumulation point of critical points of a differentiable function doesn't have to be a critical point itself.

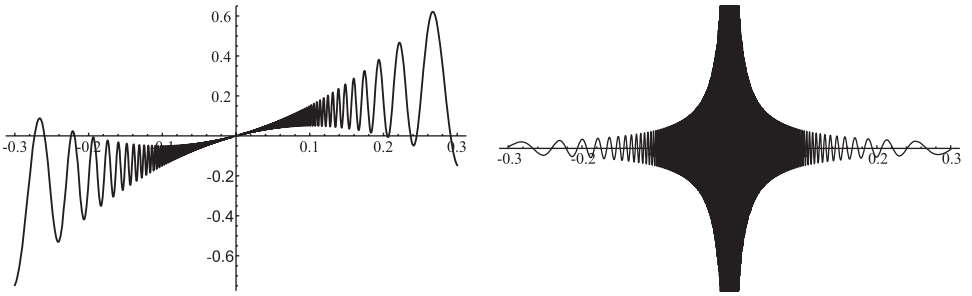


Figure 2 $x \mapsto x + 5x^2 \sin \frac{1}{x^2}$ and its derivative.

Example 4. By Darboux's intermediate value theorem for derivatives, if a derivative f' is not continuous at some point x_0, then x_0 has to be an oscillation point of f' (i.e. there is a certain interval J of positive length such that f' assumes all values in J in arbitrarily small neighborhoods of x_0). In the usual examples for this situation (like $x \mapsto x^2 \sin \frac{1}{x}$), f' oscillates in both directions, above and below $f'(x_0)$, i.e., for arbitrary small neighborhoods U of x_0, $f'(x_0)$ is an inner point of $f'(U)$.

The function

$$f(x) := x^3 + \int_0^x \left| \cos \frac{1}{t} \right|^{1/|t|} dt$$

(a slight variation of the function from Example 1) is an example of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ which satisfies $f'(0) = 0 < f'(x)$ for all $x \in \mathbb{R} \setminus \{0\}$ and whose derivative f' is not continuous at 0, i.e., f' oscillates only above $f'(0)$ (see Figure 3). In other words, f' has even a strict minimum at the origin. Here, $f'(0) = 0$ is not obvious, so we provide the details.

Proof. The function $t \mapsto \left|\cos \frac{1}{t}\right|^{1/|t|}$ is bounded and continuous except at 0, so by Lebesgue's criterion it is integrable (even in the sense of Riemann) on compact intervals. Therefore, f is well defined. For $x \neq 0$ it is clear from the fundamental theorem of calculus that $f'(x) = 3x^2 + \left|\cos \frac{1}{x}\right|^{1/|x|} > 0$.

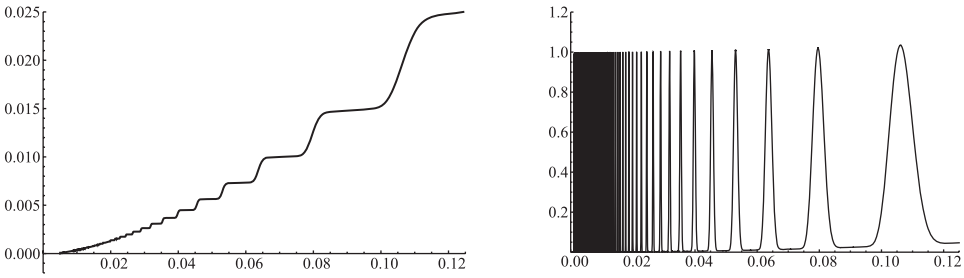


Figure 3 $x \mapsto x^3 + \int_0^x \left|\cos \frac{1}{t}\right|^{1/|t|} dt$ and its derivative (Note the different scales in both figures.).

So it remains to show that $f'(0) = 0$. For this purpose, we substitute $u := \frac{1}{t}$ to write $f(x)$ in the form

$$f(x) = x^3 + \int_{1/x}^{\infty} \frac{|\cos u|^u}{u^2}\, du \qquad \text{for } x > 0$$

and consider the points

$$a_k := k\pi, \qquad b_k := \left(k + \frac{1}{k^{1/4}}\right) \cdot \pi, \qquad c_k := \left(k + 1 - \frac{1}{k^{1/4}}\right) \cdot \pi \qquad (k \in \mathbb{N}).$$

For $k \geq 16$, we have $a_k \leq b_k \leq c_k \leq a_{k+1}$,

$$\int_{a_k}^{b_k} \frac{|\cos u|^u}{u^2}\, du \leq \frac{b_k - a_k}{a_k^2} = \frac{1}{\pi \cdot k^{9/4}}$$

and

$$\int_{c_k}^{a_{k+1}} \frac{|\cos u|^u}{u^2}\, du \leq \frac{a_{k+1} - c_k}{c_k^2} \leq \frac{1}{\pi \cdot k^{9/4}}.$$

Furthermore, using the estimate $\cos y \leq 1 - \frac{y^2}{4}$ for $0 \leq y \leq \frac{\pi}{2}$, for $k \geq 16$, we obtain

$$\int_{b_k}^{c_k} \frac{|\cos u|^u}{u^2}\, du \leq \frac{c_k - b_k}{b_k^2} \cdot \left(\cos \frac{\pi}{k^{1/4}}\right)^{b_k} \leq \frac{1}{\pi k^2} \cdot \left(1 - \frac{\pi^2}{4\sqrt{k}}\right)^{k\pi}.$$

Here, $\lim_{k\to\infty} \left(1 - \frac{\pi^2}{4\sqrt{k}}\right)^{\sqrt{k}\pi} = e^{-\pi^3/4} < 1$, so if we set $\alpha := \frac{1}{2} \cdot \left(1 + e^{-\pi^3/4}\right)$, then there exists some $N_0 \geq 16$ such that

$$\left(1 - \frac{\pi^2}{4\sqrt{k}}\right)^{\sqrt{k}\pi} \leq \alpha \qquad \text{for all } k \geq N_0,$$

and we get

$$\int_{b_k}^{c_k} \frac{|\cos u|^u}{u^2}\, du \leq \frac{1}{\pi k^2} \cdot \alpha^{\sqrt{k}} \qquad \text{for all } k \geq N_0.$$

Combining these estimates, for all $N \geq N_0 - 1$, we deduce

$$0 \leq \int_{(N+1)\pi}^{\infty} \frac{|\cos u|^u}{u^2}\, du$$

$$= \sum_{k=N+1}^{\infty} \left(\int_{a_k}^{b_k} \frac{|\cos u|^u}{u^2}\, du + \int_{b_k}^{c_k} \frac{|\cos u|^u}{u^2}\, du + \int_{c_k}^{a_{k+1}} \frac{|\cos u|^u}{u^2}\, du \right)$$

$$\leq \sum_{k=N+1}^{\infty} \left(\frac{2}{\pi \cdot k^{9/4}} + \frac{1}{\pi k^2} \cdot \alpha^{\sqrt{k}} \right) \leq \frac{2}{\pi} \int_N^{\infty} \frac{dx}{x^{9/4}} + \frac{1}{\pi} \int_N^{\infty} \frac{\alpha^{\sqrt{x}}}{\sqrt{x}}\, dx$$

$$= \frac{8}{5\pi} \cdot \frac{1}{N^{5/4}} + \frac{2}{\pi} \int_{\sqrt{N}}^{\infty} \alpha^y\, dy = \frac{8}{5\pi} \cdot \frac{1}{N^{5/4}} + \frac{2\alpha^{\sqrt{N}}}{\pi \log \frac{1}{\alpha}}.$$

Now let some $x > 0$ be given. Without loss of generality, we may assume $x < \frac{1}{N_0 \pi}$. Then there exists some $N_x \in \mathbb{N}$, $N_x \geq N_0$ such that $\frac{1}{(N_x+1)\pi} < x \leq \frac{1}{N_x \pi}$. Here as $x \to 0+$ we have $N_x \to \infty$, so we obtain

$$0 \leq \frac{f(x)}{x} \leq x^2 + (N_x + 1)\pi \cdot \int_{N_x \pi}^{\infty} \frac{|\cos u|^u}{u^2}\, du$$

$$\leq x^2 + \frac{8}{5} \cdot \frac{N_x + 1}{(N_x - 1)^{5/4}} + \frac{2}{\log \frac{1}{\alpha}} \cdot (N_x + 1) \cdot \alpha^{\sqrt{N_x - 1}} \quad \to 0 \text{ as } x \to 0+.$$

This shows that $\frac{f(x)}{x} \to 0$ as $x \to 0+$. In view of $f(-x) = -f(x)$ the same holds for $x \to 0-$. So f is differentiable at 0 with $f'(0) = 0$.

It is clear that f' is not continuous at 0. ■

This example can be easily modified such that f' is even unbounded near the origin (and still oscillates only in the positive direction, i.e., between 0 and ∞), by setting

$$f(x) := \int_0^x \frac{1}{\sqrt{|t|}} \cdot \left| \cos \frac{1}{t} \right|^{1/t^2} dt.$$

Furthermore, functions with the properties discussed in this item can also be constructed as integral functions of piecewise linear functions, having increasingly thinner and increasingly higher peaks which accumulate at the origin but this might be considered less elegant than the examples discussed above.

Example 5. Examples 1–4 dealt with accumulating critical points of a differentiable function (where the functions in Examples 2 and 3 had infinitely many changes of monotonicity while those in Examples 1 and 4 were still monotonic). Now we will encounter a situation where not the function, but its derivative has accumulating critical points—which, of course, means that the function switches infinitely many times between convexity and concavity.

Such an example occurs in a natural way when comparing the various definitions of an *inflection point* found in calculus textbooks. (For a more detailed discussion, we refer to [**10**, Section 5.5].) Maybe the most common definition (see, for example, [**8**, p. 195], [**9**, p. 147] and [**11**, p. 186]) is the following one.

(D1) A continuous function $f : [a, b] \to \mathbb{R}$ has an inflection point at $x_0 \in (a, b)$ if f is strictly convex on one side of x_0 (more precisely, in a certain interval $(x_0 - \delta, x_0)$ resp. $(x_0, x_0 + \delta)$) and strictly concave on the other side.

According to this definition also a function like

$$f(x) := \begin{cases} x^2 & \text{for } x < 0 \\ \sqrt{x} & \text{for } x \geq 0 \end{cases} \tag{1}$$

has an inflection point at $x = 0$ though it is not differentiable there.

Sometimes (see, for example, [**1**, p. 388]) the following definition can be found.

(D2) A function $f : [a, b] \to \mathbb{R}$ has an inflection point at $x_0 \in (a, b)$ if f is differentiable at x_0 and if the graph of f is strictly above the tangent line $x \mapsto f(x_0) + f'(x_0) \cdot (x - x_0)$ on one side of x_0 and strictly below this tangent line on the other side.

This definition does not apply to the function in equation (1), but in the case of functions which are differentiable at x_0, it is more general than the first definition. For example, if we set

$$f(x) := x^3 + \operatorname{sgn}(x) \cdot x^2 \sin^2 \frac{1}{x} = \begin{cases} x^3 + x^2 \sin^2 \frac{1}{x} & \text{for } x > 0, \\ 0 & \text{for } x = 0, \\ x^3 - x^2 \sin^2 \frac{1}{x} & \text{for } x < 0, \end{cases}$$

then f is differentiable at $x = 0$ with $f'(0) = 0$, $f(x) > 0$ for all $x > 0$ and $f(x) < 0$ for all $x < 0$, so f has an inflection point at $x = 0$ in the sense of (D2). However, f is neither convex nor concave in $(0, \delta)$ for any $\delta > 0$, and the same holds in the intervals $(-\delta, 0)$. See Figure 4.

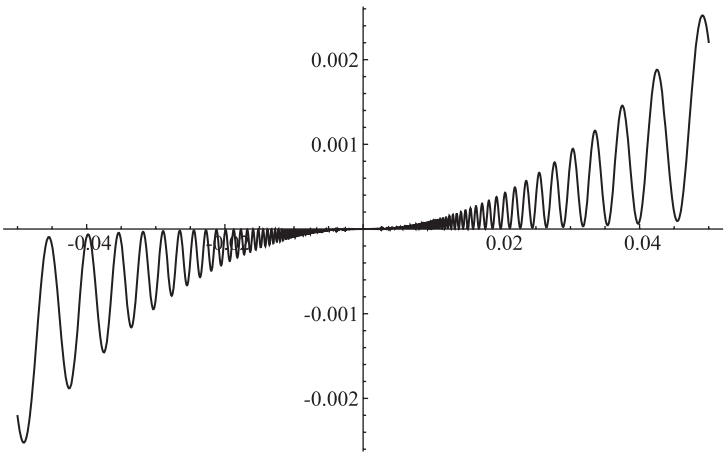


Figure 4 $x \mapsto x^3 + \operatorname{sgn}(x) \cdot x^2 \sin^2 \frac{1}{x}$.

Definition (D2) even applies to functions which are not continuous in any neighborhood of the inflection point. For example, with this definition the function

$$f(x) := \begin{cases} x^3 & \text{for } x \in \mathbb{Q}, \\ x^5 & \text{for } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

(whose graph looks like the union of the graphs of $x \mapsto x^3$ and of $x \mapsto x^5$, though this is of course a visual illusion) has an inflection point at $x = 0$.

Finally, some authors (see, for example, [**5**, p. 150]) prefer the following definition which is, however, the least general one since it is restricted to functions that are differentiable everywhere (not just at the inflection point).

(D3) A differentiable function $f : [a, b] \to \mathbb{R}$ has an inflection point at $x_0 \in (a, b)$ if f' has a strict local extremum at x_0.

For functions that are differentiable everywhere, definition (D3) is more general than (D1) but less general than (D2). More precisely, if $f : [a, b] \to \mathbb{R}$ is differentiable, then the following holds.

- If $x_0 \in (a, b)$ is an inflection point in the sense of (D1), then x_0 is also an inflection point in the sense of (D3) (since strict convexity resp. concavity can be described in terms of the first derivative being strictly increasing resp. decreasing).
- If x_0 is an inflection point in the sense of (D3), then x_0 is also an inflection point in the sense of (D2) as easily seen by considering the function

$$d(x) := f(x) - f(x_0) - f'(x_0)(x - x_0)$$

which under the conditions of (D2) is strictly monotonic with a zero at x_0, hence changes its sign at $x = x_0$.

That an inflection point in the sense of (D2) need not be an inflection point in the sense of (D3) is again illustrated by the function from Figure 4. To find an example for an inflection point in the sense of (D3), but not in the sense of (D1), we recall that a strict local minimum doesn't necessarily mean a change from monotonically decreasing to monotonically increasing as we know from Example 2. In fact, the desired example is provided by an antiderivative of a function as in Example 2, say the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) := \int_0^x t^2 \left(1.1 + \sin \frac{1}{t}\right) dt \qquad \text{for } x \in \mathbb{R}.$$

Here we have slightly modified the integrand in order to make the desired phenomenon more apparent in the graph. But g' still has the very same properties as the function f from Example 2: It has a strict extremum at the origin, but isn't decreasing on one side nor decreasing on the other side of the extremum.

Then f is differentiable with $f'(x) > f'(0) = 0$ for all $x \neq 0$, so (D3) is satisfied for $x_0 = 0$. However, (D1) is violated since f'' changes its sign in arbitrary small neighborhoods of x_0 (see Figure 5).

Example 6. There are certain analogies between infinite series and improper Riemann integrals of the form $\int_0^\infty f(t)\,dt$. For example, if $f : [0, \infty) \to [0, \infty)$ is monotoni-

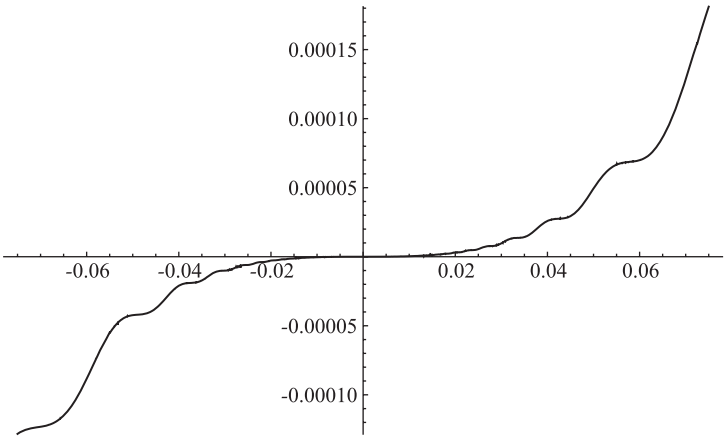


Figure 5 $x \mapsto \int_0^x t^2 \left(1.1 + \sin \frac{1}{t}\right) dt$.

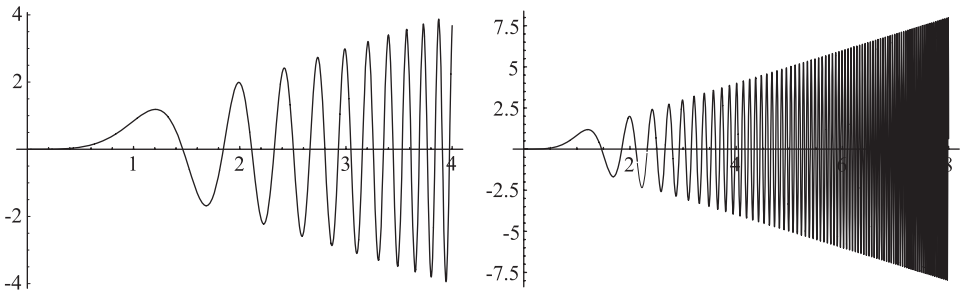


Figure 6 $x \mapsto x \sin(x^3)$ in the intervals $[0, 4]$ and $[0, 8]$.

cally decreasing, then the convergence of the improper integral $\int_0^\infty f(t)\,dt$ is equivalent to the convergence of the series $\sum_{n=0}^\infty f(n)$.

Now if $\sum_{n=0}^\infty a_n$ is a convergent series, $(a_n)_n$ necessarily converges to 0. This might (mis)lead to the conjecture that if the improper Riemann integral $\int_0^\infty f(t)\,dt$ converges (where f is a continuous function) then $\lim_{t\to\infty} f(t) = 0$. But this is wrong. On the contrary, the function

$$f(x) := x \sin(x^3)$$

(Figure 6) shows that $f(x)$ can be even unbounded as $x \to \infty$ (cf. [**7**, p. 46]). Here, to show the convergence of $\int_0^\infty f(t)\,dt$ we substitute $y := t^3$ which gives

$$\int_1^R f(t)\,dt = \int_1^{R^3} \frac{\sin y}{3y^{1/3}}\,dy \qquad \text{for all } R > 0$$

and observe that the improper integral $\int_1^\infty \frac{\sin y}{3y^{1/3}}\,dy$ converges as we can see by

(1) using integration by parts,

(2) applying Leibniz' alternating series test, essentially by showing that the sequence of the quantities $a_n := \int_{n\pi}^{(n+1)\pi} \frac{|\sin y|}{y^{1/3}}\,dy$ is decreasing to 0 for $n \to \infty$, (For details see [**8**, p. 300].) or

(3) using an integral version of Dirichlet's test of convergence which can be stated as follows [**6**]: *Let g be continuous and h be monotonically decreasing on $[a, \infty)$. Assume that $I(x) := \int_a^x g(t)\,dt$ is bounded on $[a, \infty)$ and that $lim_{x\to\infty} h(x) = 0$. Then the improper integral $\int_a^\infty g(x)h(x)\,dx$ converges.*

Of course, for the phenomenon described in this example it is crucial once more that f is a kind of "oscillating" function. Actually, if $f : [0, \infty) \to [0, \infty)$ is decreasing (or, more generally, non-increasing), then the convergence of $\int_0^\infty f(t)\,dt$ of course enforces $\lim_{t\to\infty} f(t) = 0$.

On the other hand, the function f in our counterexample can even be chosen to be positive. To see this, a near relative of the derivative of the function in Example 4 proves to be helpful. Specifically, we choose an arbitrary $\varepsilon > 0$ and define

$$f(x) := \frac{1}{(x+1)^2} + \log(x+1) \cdot |\cos x|^{x^{2+\varepsilon}} \qquad \text{for } x \geq 0.$$

Then f is positive and unbounded on $[0, \infty)$, and by a reasoning similar to Example 4 we shall see that the improper integral $\int_0^\infty f(t)\,dt$ exists.

Proof. We consider the points

$$a_k := k\pi, \qquad b_k := \left(k + \frac{1}{k^{1+\varepsilon/4}}\right) \cdot \pi, \qquad c_k := \left(k + 1 - \frac{1}{k^{1+\varepsilon/4}}\right) \cdot \pi$$

for $k \in \mathbb{N}$. Then for $k \geq 2$ we have $a_k \leq b_k \leq c_k \leq a_{k+1}$, hence we can estimate

$$\int_{a_k}^{b_k} \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt \leq (b_k - a_k) \cdot \log(b_k + 1) \leq \frac{\pi \log(\pi k + 3)}{k^{1+\varepsilon/4}},$$

$$\int_{c_k}^{a_{k+1}} \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt \leq \frac{\pi \log(\pi k + 5)}{k^{1+\varepsilon/4}},$$

$$\begin{aligned}\int_{b_k}^{c_k} \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt &\leq (c_k - b_k) \cdot \log(\pi k + 5) \cdot \left(\cos \frac{\pi}{k^{1+\varepsilon/4}}\right)^{b_k^{2+\varepsilon}} \\ &\leq \pi \cdot \log(\pi k + 5) \cdot \left(1 - \frac{\pi^2}{4k^{2+\varepsilon/2}}\right)^{(k\pi)^{2+\varepsilon}}.\end{aligned}$$

From $\lim_{k\to\infty} \left(1 - \frac{\pi^2}{4k^{2+\varepsilon/2}}\right)^{k^{2+\varepsilon/2}} = e^{-\pi^2/4}$, $\frac{\pi^2}{4} > 1$ and $\lim_{k\to\infty} k^2 e^{-k^{\varepsilon/2}} = 0$, we see that for k sufficiently large we have

$$\int_{b_k}^{c_k} \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt \leq \pi \log(\pi k + 5) \cdot e^{-\pi^{2+\varepsilon} \cdot k^{\varepsilon/2}} \leq e^{-k^{\varepsilon/2}} \leq \frac{1}{k^2}.$$

Here the series $\sum_{k=1}^{\infty} \frac{\pi \log(\pi k+5)}{k^{1+\varepsilon/4}}$ and $\sum_{k=1}^{\infty} \frac{1}{k^2}$ are convergent. Therefore the improper integral $\int_0^\infty \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt$ exists, hence so does $\int_0^\infty f(t)\, dt$. ■

Example 7. When talking about differentiable functions $f : [a; \infty) \to \mathbb{R}$ with a certain asymptotic behavior near ∞, say $\lim_{x\to\infty} f(x) = 0$, the question arises how the derivative f' behaves near ∞. Probably some students will think that $f'(x)$ is bound to tend to 0 as $x \to \infty$.

However, there's no good reason to think so, as you can illustrate by the following analogy: If a small particle is trapped in a small room and we make this room smaller and smaller, this particle might still be able to move arbitrarily fast within this room (as long as it is able to accelerate and decelerate within "very short" times).

In fact, a simple counterexample is provided by an appropriate antiderivative of the function in Example 6, i.e., by

$$f(x) := -\int_x^\infty t \sin(t^3)\, dt;$$

see Figure 7. Here in view of Example 6 $\lim_{x\to\infty} f'(x)$ does not exist, and f' is even unbounded near ∞. The same can be achieved by the more explicit example

$$f(x) := \frac{\sin(x^3)}{x}.$$

At first sight, it seems to be more difficult to give a counterexample where f is even monotonic, i.e., f' doesn't change signs. But here once again Example 6 comes to our

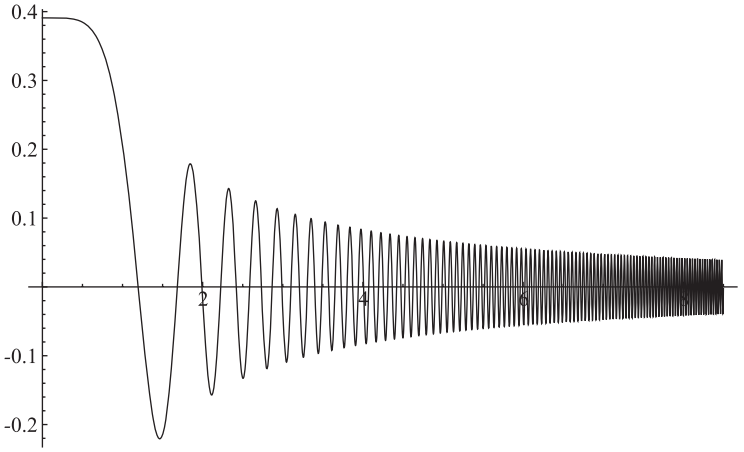


Figure 7 $x \mapsto -\int_x^\infty t \sin(t^3)\, dt$.

aid. In particular, for arbitrary $\varepsilon > 0$ and $x > 0$, we define

$$f(x) := \int_0^x \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt.$$

Then f is differentiable on $(0, \infty)$ with non-negative derivative $f'(x) = \log(x+1) \cdot |\cos x|^{x^{2+\varepsilon}}$. The zeros of f' are isolated, so f is strictly increasing, and $f'(x)$ does not tend to 0 for $x \to \infty$; in fact, f' is even unbounded. Furthermore, as we have seen in Example 6, the limit

$$\lim_{x\to\infty} f(x) = \int_0^\infty \log(t+1) \cdot |\cos t|^{t^{2+\varepsilon}}\, dt$$

exists.

The graph of f and f' (for $\varepsilon = \frac{1}{100}$) is shown in Figure 8.

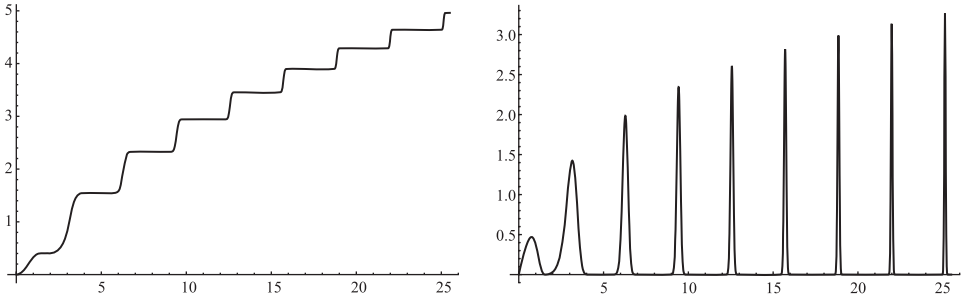


Figure 8 $x \mapsto \int_0^x \log(t+1) \cdot |\cos t|^{t^{2.01}}\, dt$ and its derivative.

Acknowledgment We'd like to thank the anonymous referees for their valuable suggestions which contributed to improving the exposition of the paper. Furthermore, we are very grateful to Professor Stephan Ruscheweyh and Professor Lawrence Zalcman for several helpful comments.

REFERENCES

[1] Apostol, T. (1961). *Calculus*, Vol. 1, New York, NY: Blaisdell Publishing Company.

[2] Appell, J. (2009). *Analysis in Examples and Counterexamples. An Introduction to the Theory of Real Functions*. Berlin: Springer.

[3] Appell, J. (2011). Some counterexamples for your calculus course. *Analysis (München)* 31: 1–12.

[4] Appell, J., Roos, A.-K. (2014). Continuity, monotonicity, oscillation, variation: obvious and surprising aspects (Stetigkeit, Monotonie, Oszillation, Variation: Naheliegendes und Überraschendes). *Math. Semesterber.* 61: 233–248.

[5] Erwe, F. (1962). *Differential- und Integralrechnung. Band 1: Elemente der Infinitesimalrechnung und Differentialrechnung*, Mannheim: B.I.-Hochschultaschenbücher, Bibliographisches Institut.

[6] Flatto, L. (1976). *Advanced Calculus*, Baltimore, MD: Williams & Wilkins.

[7] Gelbaum, B. R., Olmsted, J. M. H. (1964). *Counterexamples in Analysis*, San Francisco, CA: Holden-Day Inc.

[8] Köhler, G. *Analysis*, Lemgo: Heldermann.

[9] Königsberger, K. (1992). *Analysis*, Vol. 1, 2nd ed. Berlin: Springer.

[10] Rajwade, A. R., Bhandari, A. K. (2007). *Surprises and Counterexamples in Real Function Theory*. Texts and Readings in Mathematics, Vol. 42, New Delhi: Hindustan Book Agency.

[11] Salas, S., Hille, E. (1995). *Calculus. One and several variables (Calculus. Einführung in die Differential- und Integralrechnung*. Heidelberg: Spektrum.

[12] Zalcman, L. (1988). Positive derivatives and increasing functions. *Elem. Math.* 43: 120–121.

Summary. We discuss some surprising phenomena from basic calculus related to oscillating functions. Among other things, we see that a continuously differentiable function with a strict minimum doesn't have to be decreasing to the left nor increasing to the right of the minimum, we present a function f whose derivative is discontinuous at one point x_0, but oscillates only above $f'(x_0)$ (i.e., f' has a strict minimum at x_0), and we compare several definitions of inflection point.

JÜRGEN GRAHL (MR Author ID: 671443), born in 1972, received his Ph.D. in 2002 under the supervision of Prof. S. Ruscheweyh and is now a lecturer at the University of Würzburg. His field of research is complex analysis, in particular Nevanlinna theory and normal families.

SHAHAR NEVO (MR Author ID: 683995), born in 1966, received his Ph.D. in 2000 under the supervision of Prof. L. Zalcman. He's now associate professor at Bar Ilan University. His main research interests lie in the field of complex analysis and include normal and quasi-normal families as well as operator theory.

Why All Rings Should Have a 1

BJORN POONEN
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
poonen@math.mit.edu

Should the definition of ring require the existence of a multiplicative identity 1?

Emmy Noether, when giving the modern axiomatic definition of a commutative ring, in 1921, did not include such an axiom [**15**, p. 29]. For several decades, algebra books followed suit [**16**, §3.1], [**18**, I.§5]. But starting around 1960, many books by notable researchers began using the term "ring" to mean "ring with 1" [**7**, 0.(1.0.1)] [**14**, II.§1], [**17**, p. XIV], [**1**, p. 1]. Sometimes a change of heart occurred in a single person, or between editions of a single book, always toward requiring a 1: compare [**11**, p. 49] with [**13**, p. 86], or [**2**, p. 370] with [**3**, p. 346], or [**4**, I.§8.1] with [**5**, I.§8.1]. Reasons were not given; perhaps it was just becoming increasingly clear that the 1 was needed for many theorems to hold; some good reasons for requiring a 1 are explained in [**6**].

But is either convention more *natural*? The purpose of this article is to answer yes, and to give a reason: existence of a 1 is a part of what associativity should be.

Total associativity

The usual associative law is about reparenthesizing *triples*: $(ab)c = a(bc)$ for all a, b, c. Why are there not also ring axioms about reparenthesizing n-tuples for $n > 3$? It is because they would be redundant, implied by the law for triples. The whole point of associativity is that it lets us assign an unambiguous value to the product of any finite sequence of two or more terms.

By why settle for "two or more"? Cognoscenti do not require *sets* to have two or more elements. So why restrict attention to sequences with two or more terms? Most natural would be to require *every* finite sequence to have a product, even if the sequence is of length 1 or 0. This suggests the following definition.

Definition. A *product* on a set A is a rule that assigns to each finite sequence of elements of A an element of A, such that the product of a 1-term sequence is the term.

Next, let us explain why the usual associative law is insufficient to regulate such products. The usual associative law, although it involves three elements at a time, is a condition on a *binary* operation; that is, it constrains only the 2-fold products. But the new definition of product provides a value also to the 4-term sequence $abcd$, and so far there is no axiom to require this value to be the same as $((ab)c)d$ built up using the product on pairs repeatedly. We need a stronger associativity axiom to relate all the products of various lengths. This motivates the following definition.

Definition. A product is *totally associative* if each finite product of finite products equals the product of the concatenated sequence (for example, $(abc)d(ef)$ should equal the 6-term product $abcdef$).

Math. Mag. **92** (2019) 58–62. doi:10.1080/0025570X.2018.1538714
MSC: Primary 16-01, Secondary 16-03

Note that the finite products in this definition are not required to involve two or more terms; indeed, the definition would be more awkward if it spoke only of "finite products of two or more finite products of two or more terms each."

As argued at the beginning of this section, a product is more natural than a binary operation, insofar as it does not assign preferential status to the number 2. Similarly, total associativity, although less familiar than associativity, is more natural in that a law applicable to all tuples is more natural than a law applicable only to triples; after all, the number 3 is not magical either. Hence the ring axioms should be designed so that they give rise to a totally associative product. Now the key point is the following theorem, whose proof will be sketched at the end of this section.

Theorem. *A binary operation extends to a totally associative product if and only if it is associative and admits an identity element.*

What?! Where did that identity element come from? The definition of totally associative implies the equations

$$(abc)d = abcd$$
$$(ab)c = abc$$
$$(a)b = ab$$
$$()a = a.$$

The last equation, which holds for any a, shows that the empty product () is a left identity. Similarly, () is a right identity, so () is an identity element.

Thus the natural extension of associativity demands that rings should contain an empty product, so it is natural to require rings to have a 1. But occasionally one does encounter structures that satisfy all the axioms of a ring except for the existence of a 1. What should they be called? Happily, there is an apt answer, suggested by Louis Rowen [13, p. 155]: *rng*! (Other suggestions include *pseudo-ring* [5, I.§8.1] and *(associative) $\mathbb{Z}$-algebra* [6, Appendix A].) As our reasoning explains and as Rowen's terminology suggests, it is better to think of a rng as a ring with something missing than to think of a ring with 1 as having something extra.

Sketch of proof of the theorem. Given a binary operation that extends to a totally associative product, the argument above shows that it admits an identity element, and the usual associative law $(ab)c = a(bc)$ follows too since total associativity implies that both sides equal the 3-term product abc.

Conversely, given a binary operation $*$ that is associative and admits a 1, define

$$a_1 a_2 \cdots a_n := \begin{cases} 1, & \text{if } n = 0; \\ a_1, & \text{if } n = 1; \\ (a_1 a_2 \cdots a_{n-1}) * a_n, & \text{(inductively) if } n \geq 2. \end{cases}$$

This is a product that extends $*$, and an involved but standard inductive argument shows that the usual associative law for $*$ can be used over and over to reparenthesize any finite product of finite products into the product of the concatenated sequence; for the details, see [5, I.§1.2, Théorème 1, and §2.1]. In other words, this extension of $*$ is a totally associative product. ∎

Counterarguments

Here we mention some arguments for *not* requiring a 1, in order to rebut them.

- "Algebras should be rings, but Lie algebras usually do not have a 1."

  Lie algebras, which are objects used in more advanced mathematics to study the group of invertible $n \times n$ real matrices and its subgroups [8], are usually not associative either. We require a 1 only in the presence of associativity. It is accepted nearly universally that ring multiplication should be associative, so when the word "algebra" is used in a sense broad enough to include Lie algebras, it is understood that algebras have no reason to be rings.
- "An infinite direct sum of nonzero rings does not have a 1."

  By definition, if $A_1, A_2, \ldots$ are abelian groups, say, then the *direct product* $\prod_{i=1}^{\infty} A_i$ is the set of tuples $(a_1, a_2, \ldots)$ with $a_i \in A_i$ for each i, while the *direct sum* $\bigoplus_{i=1}^{\infty} A_i$ is the subgroup consisting of those tuples satisfying the additional condition that there are only finitely many i for which a_i is nonzero. Direct sums are typically defined for objects like vector spaces and abelian groups, for which the set of homomorphisms between two given objects is an abelian group, for which quotients exist, and so on. Rings fail to have these properties, whether or not a 1 is required: the quotient of a ring by a subring is no longer a ring (there is no natural way to multiply elements of $\mathbb{R}/\mathbb{Z}$, for instance). So it is strange even to speak of a direct sum of rings. Instead one should speak of the direct product, which does have a 1, namely the tuple with a 1 in every position.
- "If a 1 is required, then function spaces like the space $C_c(\mathbb{R})$ of continuous functions $f : \mathbb{R} \to \mathbb{R}$ that are 0 outside some unspecified bounded interval will be disqualified."

  This is perhaps the hardest to rebut, given the importance of function spaces. But many such spaces are ideals in a natural ring (e.g., $C_c(\mathbb{R})$ is an ideal in the ring $C(\mathbb{R})$ of *all* continuous functions $f : \mathbb{R} \to \mathbb{R}$), and fail to include the 1 only because of some condition imposed on their elements. So one can say that they, like the direct sums above and like the rng of even integers, *deserve* to be ousted from the fellowship of the ring. In any case, however, these function spaces still qualify as $\mathbb{R}$-algebras.

Further counterarguments can be found in the preface to [9].

Rngs should not be banished completely, because there are applications for which rngs are more convenient than rings: see [12, pp. 31–36] and [10], for instance. The latter gave the first example of an infinite finitely generated group all of whose elements have finite order.

Implications

Once the role of the empty product is acknowledged, other definitions that seemed arbitrary become natural.

- A ring homomorphism $A \to B$ should respect finite products, so in particular it should map the empty product 1_A to the empty product 1_B.
- A subring should be closed under finite products, so it should contain the empty product 1.
- An ideal is prime if and only if its complement is closed under finite products. This explains why the unit ideal (1) in a ring is never considered to be prime: if (1) were a prime ideal, then its complement $\emptyset$ would be closed under finite products and in particular would contain the empty product 1; but $\emptyset$ does not contain 1.
- The argument that rings should have a 1 involved only one binary operation, multiplication, so the same argument explains also why monoids are more natural than

semigroups. (A *semigroup* is a set with an associative binary operation, and a *monoid* is a semigroup with a 1.)

There are also applications of these ideas in higher mathematics. For example, there is the notion of a *category*, which has objects and morphisms satisfying certain axioms modeled on the properties of groups and homomorphisms. One of these axioms lets one compose morphisms $f\colon A \to B$ and $g\colon B \to C$ to produce $g \circ f$ when the target object of f (here called B) matches the source object of g. Another axiom asserts the existence of identity morphisms. Now we can see why the axiom about identity morphisms is natural: it arises as a special case of composing a chain of morphisms. More specifically, given objects $A_0, \ldots, A_n$ and a chain of morphisms

$$A_0 \xrightarrow{f_1} A_1 \xrightarrow{f_2} \cdots \xrightarrow{f_n} A_n,$$

one wants to be able to form the composition, even when $n = 0$, and in the $n = 0$ case the composition is the identity morphism from A_0 to A_0.

Final comments

It would be ridiculous to introduce the definition of ring to beginners in terms of totally associative products. But it is nice to understand why certain definitions should be favored over others.

Acknowledgments I thank Keith Conrad, Greg Marks, Gigel Militaru, Johan Öinert, and Lance Small for many helpful comments. I thank also the referees for many suggestions that improved this note. This research was supported in part by National Science Foundation grants DMS-1069236 and DMS-1601946 and grants from the Simons Foundation (#340694 and #402472 to Bjorn Poonen).

REFERENCES

[1] Atiyah, M. F., Macdonald, I. G. (1969). *Introduction to Commutative Algebra*. Reading, MA: Addison-Wesley Publishing Co.
[2] Birkhoff, G., Mac Lane, S. (1953). *A Survey of Modern Algebra*, Rev. ed. New York, NY: Macmillan Co.
[3] Birkhoff, G., Mac Lane, S. (1965). *A Survey of Modern Algebra*, 3rd ed. New York, NY: Macmillan Co.
[4] Bourbaki, N. (1958). *Éléments de mathématique. Part I. Les structures fondamentales de l'analyse. Livre II. Algèbre. Chapitre I. Structures algébriques*, Actual. Sci. Ind., no. 1144, Paris: Hermann.
[5] Bourbaki, N. (1970). *Éléments de mathématique. Algèbre. Chapitres 1 à 3*, Paris: Hermann (English transl., *Algebra I. Chapters 13, Elements of Mathematics*, Berlin: Springer-Verlag, 1998. Reprint of the 1989 English translation).
[6] Conrad, K. (2013). Standard definitions for rings. http://www.math.uconn.edu/ kconrad/blurbs/ringtheory/ringdefs.pdf
[7] Grothendieck, A. (1960). Éléments de géométrie algébrique. I. Le langage des schémas, *Inst. Hautes Études Sci. Publ. Math.*, No. 4, 228 pp.
[8] Erdmann, K., Wildon, M. J. (2006). *Introduction to Lie Algebras*, Springer Undergraduate Mathematics Series. London: Springer-Verlag London, Ltd.
[9] Gardner, B. J., Wiegandt, R. (2004). *Radical Theory of Rings*, Monographs and Textbooks in Pure and Applied Mathematics, Vol. 261. New York, NY: Marcel Dekker, Inc.
[10] Golod, E. S. (1964). On nil-algebras and finitely approximable p-groups, Russian, *Izv. Akad. Nauk SSSR Ser. Mat.*, 28: 273–276.
[11] Jacobson, N. (1951). *Lectures in Abstract Algebra. Vol. I. Basic Concepts*. New York, NY: D. Van Nostrand Co., Inc.
[12] Jacobson, N. (1979). *Lie Algebras*, Republication of the 1962 original. New York, NY: Dover Publications, Inc.

[13] Jacobson, N. (1985). *Basic Algebra, I*, 2nd ed. New York, NY: W. H. Freeman and Company.
[14] Lang, S. (1965). *Algebra*. Reading, MA: Addison-Wesley Publishing Co., Inc.
[15] Noether, E. (1921). Idealtheorie in Ringbereichen. *Math. Ann.*, 83(1–2): 24–66.
[16] van der Waerden, B. L. (1966). *Algebra. Teil I*, Siebte Auflage. Heidelberger Taschenbücher, Band 12 (1970 English version. *Algebra*, Vol 1. Translated by F. Blum and J.R. Schulenberger, New York, NY: Frederick Ungar Publishing Co) Berlin: Springer-Verlag.
[17] Weil, A. (1967). *Basic Number Theory*, Die Grundlehren der mathematischen Wissenschaften, Band 144. New York, NY: Springer-Verlag.
[18] Zariski, O., Samuel, P. (1975). *Commutative Algebra*, Vol. 1. With the cooperation of I. S. Cohen; Corrected reprinting of the 1958 edition; Graduate Texts in Mathematics, No. 28. New York, NY: Springer-Verlag.

Summary. We argue that the definition of ring should require the existence of a multiplicative identity 1 because this requirement is part of what associativity should be. We also address counterarguments and explore some implications of our argument.

BJORN POONEN (MR Author ID: 250625) is the Claude Shannon Professor of Mathematics at MIT. His research focuses mainly on number theory and algebraic geometry, and his expository writing earned him the MAA Chauvenet Prize. Nineteen mathematicians have completed a Ph.D. thesis under his guidance. He serves the mathematical community in various ways, for instance as managing editor of *Algebra & Number Theory*. Finally, he has been singing choral music for over 30 years.

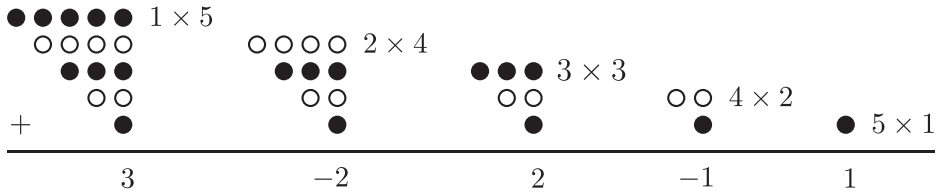
Proof Without Words: Alternating Sums of Products with Increasing and Decreasing Factors

CHARLES F. MARION
Yorktown Heights, NY 10598
charliemath@optonline.net

We substantiate the following pattern using an image.

$$\begin{aligned}
(1 \times 1) &= 1\\
(1 \times 3) - (2 \times 2) + (3 \times 1) &= 2\\
(1 \times 5) - (2 \times 4) + (3 \times 3) - (4 \times 2) + (5 \times 1) &= 3\\
(1 \times 7) - (2 \times 6) + (3 \times 5) - (4 \times 4) + (5 \times 3) - (6 \times 2) + (7 \times 1) &= 4\\
&\vdots\\
[1 \times (2n+1)] - [2 \times (2n)] + \cdots - [(2n) \times 2] + [(2n+1) \times 1] &= n+1.
\end{aligned}$$

For example, to evaluate the sum for $n = 2$, let $\bullet$ and $\circ$ be positive and negative in the following diagram.



Summary. Wordlessly, we show that alternating sum of products of increasing and decreasing positive integers resolve themselves into the positive integers.

CHARLES F. MARION (MR Author ID: 1211614) taught mathematics at Lakeland High School (Shrub Oak NY) for 32 years. He continues to relish the opportunities retirement has presented to wander about and explore the mathematical landscape in greater depth.

Math. Mag. **92** (2019) 63. doi:10.1080/0025570X.2018.1541687
MSC: Primary 05A19

How Many Unicycles on a Wheel?

JACOB SIEHLER
Gustavus Adolphus College
Saint Peter, MN 56082
jsiehler@gustavus.edu

We are in the world of graph theory here, not circus acts, and we have something interesting to enumerate; let's have a look at what we'll be counting. On the left of Figure 1 is the wheel graph W_4. It is shown alongside several, though far from all, of its *spanning unicycles*.

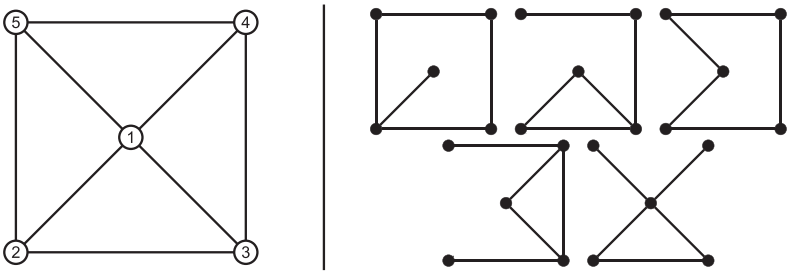


Figure 1 Wheel graph W_4 and a few of its spanning unicycles.

To define our terms: For $n \geq 3$, the *wheel graph* W_n consists of n vertices connected in a cycle (the rim), together with one additional vertex (the hub) which is connected to all of them. The standard drawing of the wheel has the rim vertices equally spaced around a circle with the hub at the center.

A *spanning unicycle* in a graph is a subset of the edges which leaves the vertices connected and contains exactly one cycle. Think of it as using graph edges to build a "ring road" through some or all of the vertices, and then choosing just enough additional edges to reach any remaining vertices off the ring. If you do this on a graph with n vertices, you will find you always use exactly n edges. At this point, you might like to pause and work out how many spanning unicycles can be found in W_4 before proceeding to the following theorem, our main result, which discloses the number of unicycles in any wheel graph.

Theorem. *The wheel graph W_n (on $n+1$ vertices) contains $n \cdot F_{2n-1}$ spanning unicycles, where F_{2n-1} denotes the $(2n-1)$th Fibonacci number.*

So, for example (and to make sure we agree about indices), the wheel graph W_4 in Figure 1 contains $4 \cdot 13 = 52$ spanning unicycles, because we set $F_1 = F_2 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for $n \geq 0$. In the following sections, we shall prove this little, specialized counting theorem about unicycles by applying a bigger, more general counting theorem about trees.

Math. Mag. **92** (2019) 64–69. doi:10.1080/0025570X.2018.1528096
MSC: Primary 05C30, Secondary 05C50

Kirchhoff and spanning trees

If we remove an edge from the cycle of a spanning unicycle, what remains is a connected graph with *no* cycles; this is known as a *spanning tree* in the original graph. Since the idea of a spanning tree is likely to be more familiar than a unicycle to most readers, we could turn things around and say that a spanning unicycle is just what you get when you add one more edge to a spanning tree (necessarily forming one cycle).

Kirchhoff's matrix-tree theorem, which we will state in a moment, says that the answer to the question, "How many spanning trees are in my graph?" is given by a determinant. In order to state the theorem, we need to know that the *Laplace matrix* of a graph is a square matrix L with one row and one column for each vertex of the graph. The (i, j) entry is given by

$$l_{ij} = \begin{cases} \text{degree of vertex } i, & \text{if } i = j \\ -n, & \text{if there are } n \text{ edges between vertex } i \text{ and vertex } j, \end{cases}$$

where the degree of a vertex is simply the number of edges touching that vertex. For example, the Laplace matrix of our example W_4 is (with the rows and columns ordered according to the labeling of the vertices in Figure 1)

$$\begin{pmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 3 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ -1 & -1 & 0 & -1 & 3 \end{pmatrix}.$$

Remarkably, this simple matrix representation of the graph data reduces the enumeration of spanning trees to a straightforward calculation.

Kirchhoff's matrix-tree theorem. *The number of spanning trees in a graph is equal to the absolute value of the determinant of any minor obtained by deleting any one row and column of the graph's Laplace matrix.*

Thus, according to the theorem, by deleting the first row and column from the matrix above, we can compute that W_4 contains

$$\det \begin{pmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix} = 45 \text{ spanning trees.}$$

Kirchhoff's theorem is a gem, and if you haven't encountered it before, I recommend drawing a few small graphs with their spanning trees, and working out the appropriate determinant to confirm your count. You can find an insightful proof of the theorem in Matoušek's book [**9**]. An article by Benjamin and Cameron [**1**] also presents it along with other applications of determinants to enumeration. In what follows, it will be useful to keep in mind that the theorem holds true even in graphs that may have multiple edges between two vertices.

Although the problem of counting unicycles seems to be only "one edge away" from the problem of counting spanning trees, there is no simple analog of the matrix-tree theorem for unicycles. However, we shall be optimistic and examine the relationship between the two counting problems.

Given a graph G with a specified cycle C, let us define a "collapsed" graph G/C which replaces all of C with a single new vertex, c. The idea is conveyed in Figures 2

and 3. The edges of G between vertices of C vanish in G/C. Any edge from a vertex v outside C to a vertex in C becomes an edge from v to c in G/C. This may result in multiple edges from v to c, as seen in Figure 2.

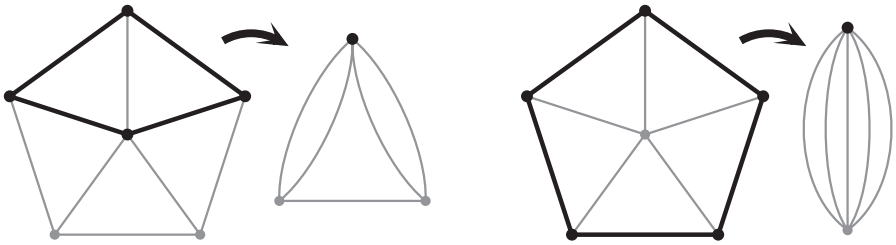


Figure 2 Collapsing different cycles in W_5.

Suppose we have drawn a spanning unicycle on G with cycle C. Collapsing C makes the cycle disappear, leaving a spanning tree in G/C. On the other hand, any spanning tree in G/C can be turned into a spanning unicycle in G with cycle C just by adding the edges of C. These two functions, from unicycles in G to trees in G/C and back again, are inverses of one another, and so we have a bijection.

Lemma 1. *The number of spanning unicycles in G with cycle C is equal to the number of spanning trees in the collapsed graph G/C.*

In principle, this gives a strategy for counting spanning unicycles in a graph G: For each cycle C, use the matrix-tree theorem to count spanning trees in G/C, and add up the results. In practice, this strategy is only as good as our ability to enumerate the cycles in G and organize the sum. Fortunately, in wheel graphs, the cycles are simple and the high amount of symmetry will make the summation easy.

Collapsing and counting

Consider a spanning unicycle on the $(n+1)$-vertex wheel W_n, where $n \geq 3$. Its cycle must be either:

(a) the rim, or

(b) Some k (for $2 \leq k \leq n$) consecutive vertices on the rim, together with the hub, which we call a (hub+k)-cycle.

In case (a), there are simply n ways to complete the unicycle by choosing a spoke to connect the hub to the rim.

In case (b), for any given k, there are n of these (hub+k)-cycles. Since they are all related by rotational symmetries of W_n in its standard drawing, the collapsed graph G/C does not depend (up to isomorphism) on which of them we call C and collapse.

Figure 3 shows the collapse of a (hub+2)-cycle in W_6. The Laplace matrix of the collapsed graph is

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \times & 3 & -1 & 0 & 0 \\ \times & -1 & 3 & -1 & 0 \\ \times & 0 & -1 & 3 & -1 \\ \times & 0 & 0 & -1 & 3 \end{pmatrix}.$$

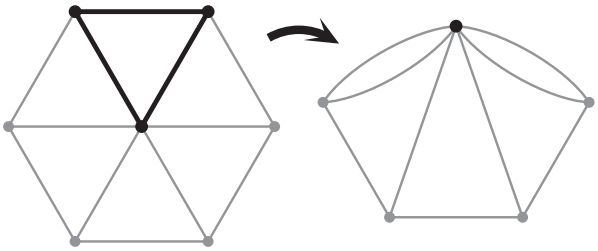


Figure 3 Collapsing any (hub+2)-cycle in W_6.

The $\times$'s represent entries corresponding to the new vertex c, which will be irrelevant as we will discard that row and column to apply the matrix-tree theorem. Now, this particular example nicely illustrates the general case: If we collapse a (hub+k)-cycle in W_n, order the remaining vertices consecutively around the rim, and ignore the row and column of the collapsed vertex, we get an $(n-k)\times(n-k)$ minor in the Laplace matrix with 3's on the main diagonal and -1's above and below it. To apply the matrix-tree theorem, we will need to evaluate the determinant of a matrix in this form.

Let M_k denote the $k\times k$ tridiagonal matrix with 3's on the diagonal and -1's on the sub- and superdiagonal, where

$$M_1=\begin{pmatrix}3\end{pmatrix},\quad M_2=\begin{pmatrix}3&-1\\-1&3\end{pmatrix},\quad M_3=\begin{pmatrix}3&-1&0\\-1&3&-1\\0&-1&3\end{pmatrix},\quad \text{and so on.}$$

This is where the Fibonacci numbers enter. We see that $\det M_1=3=F_4$ and $\det M_2=8=F_6$, and the pattern continues, as described in the following lemma.

Lemma 2. *For all $k\geq 1$, $\det M_k=F_{2k+2}$.*

This result is a special case of more general results about determinants of tridiagonal matrices [2, 10]. A short paper by Fielder [4] on the subject of enumerating trees also includes this particular case. Still, we can outline a short proof as follows: by expanding the determinant of M_k along the first row, we see that

$$\det M_k=3\det M_{k-1}-(-1)\det\left(\begin{array}{c|ccc}-1&-1&0&\cdots\\\hline 0&&&\\\vdots&&(M_{k-2})&\\0&&&\end{array}\right)$$

$$=3\det M_{k-1}-\det M_{k-2}\quad\text{for each } k\geq 3,$$

and the even-index Fibonacci numbers satisfy the same recurrence (this follows quickly from the definition of the Fibonacci sequence). Since $\det M_1=F_4$ and $\det M_2=F_6$ and the two sequences satisfy the same second-order recurrence relation, they agree for all k.

The following lemma is a routine mathematical induction exercise using the defining relation of the Fibonacci sequence.

Lemma 3. *For all $n\geq 1$, $\displaystyle\sum_{j=1}^{n}F_{2j}=F_{2n+1}-1$.*

With that, we have all we need to establish our main result.

Proof of the theorem. By Lemma 1 and the matrix-tree theorem, $\det M_{n-k}$ represents the number of spanning unicycles in W_n that contain a particular (hub+k)-cycle. Adding the n rim-unicycles to all the (hub+k)-unicycles, we find the total number of unicycles in W_n is

$$n + \sum_{k=2}^{n} n \det M_{n-k} = n + n \sum_{j=0}^{n-2} \det M_j = n + n \sum_{j=0}^{n-2} F_{2j+2}$$

$$= n \left(1 + \sum_{j=1}^{n-1} F_{2j} \right) = n \cdot F_{2n-1}.$$

Lemma 2 replaces the determinant with a Fibonacci number in the second equality, and Lemma 3 provides the evaluation of the sum in the final equality. ■

Go further with unicycles

The sequence $\{n \cdot F_{2n-1}\}$ which we have obtained here appears as A117202 in the On-Line Encyclopedia of Integer Sequences [**11**], where a different combinatorial connection to wheel graphs is noted (without proof). It is curious that the same numbers also count certain *acyclic* subgraphs of W_n.

The reader who wants practice applying the matrix-tree theorem may take it as an exercise to find the number of spanning trees in W_n. The answer to this is also found in the OEIS, as sequence A004146 [**11**] (you will find connections to many other problems there as well). There are further examples of spanning tree calculations for interesting graph families in the articles of Haghighi and Bibak [**7**] and Hilton [**8**].

Finally, it is worth mentioning that the number of spanning unicycles (or spanning trees) in a graph can be obtained by an appropriate evaluation of the graph's Tutte polynomial. See Bollobás [**3**] for a good introduction. Bollobás mentions the application to spanning trees; the application to unicycles is noted in [**5**] (in the comments following Theorem 7).

Computing the entire Tutte polynomial and then evaluating it is overkill; you can use the deletion and contraction operations which define it to write a simpler recursive formula that is specialized to the unicycle-counting problem. However, dedicated software exists for the purpose of computing Tutte polynomials [**6**], so that is one way to explore the unicycle counting problem experimentally and make discoveries and conjectures. Explore an interesting graph family (for example, the prism graphs or Möbius ladders), and you will likely find an interesting sequence counting the unicycles.

REFERENCES

[1] Benjamin, A. T., Cameron, N. T. (2005). Counting on determinants. *Amer. Math. Monthly.* 112(6): 481–492. doi.org/10.2307/30037518

[2] Boesch, F. T., Prodinger, H. (1986). Spanning tree formulas and Chebyshev polynomials. *Graphs Combin.* 2(1): 191–200. doi.org/10.1007/BF01788093

[3] Bollobás, B. (1998). *Modern Graph Theory*. Graduate Texts in Mathematics, Vol. 184, New York: Springer. doi.org/10.1007/978-1-4612-0619-4

[4] Fielder, D. C. (1974). Fibonacci numbers in tree counts for sector and related graphs. *Fibonacci Quart.* 12: 355–359.

[5] Fey, A., Levine, L., Wilson, D. B. (2010). Approach to criticality in sandpiles. *Phys. Rev. E.* 82: 031121, 14pp. doi.org/10.1103/PhysRevE.82.031121

[6] Haggard, G., Pearce, D. J., Royle, G. (2010). Computing Tutte polynomials. *ACM Trans. Math. Softw.* 37: Article 17, 24pp. doi.org/10.1145/1824801.1824802

[7] Haghighi, M. H., Bibak, Kh. (2012). The number of spanning trees in some classes of graphs. *Rocky Mt. J. Math.* 42(4): 1183–1195. doi.org/10.1216/RMJ-2012-42-4-1183
[8] Hilton, A. J. W. (1974). Spanning trees and Fibonacci and Lucas numbers. *Fibonacci Quart.* 12(3): 259–262.
[9] Matoušek, J. (2010). *Thirty-three Miniatures: Mathematical and Algorithmic Applications of Linear Algebra*. Providence, RI: American Mathematical Society.
[10] Narayan, D. A., Cahill, N. D., D'Errico, J. R., Narayan, J. Y. (2002). Fibonacci determinants. *College Math. J.* 33: 221–225. doi.org/10.2307/1559033
[11] OEIS Foundation Inc. (2011). The on-line encyclopedia of integer sequences, http://oeis.org

Summary. The enumeration of spanning trees in a graph is simply accomplished by a determinant (and a great theorem). But what happens when you add one more edge to a spanning tree? The resulting "unicycle" structures in a graph are harder to count, but we explore the problem in the family of wheel graphs, where the enumeration leads to a tidy answer and some old familiar friends.

JACOB SIEHLER (MR Author ID: 719629) is currently an Assistant Professor at Gustavus Adolphus College. His mathematical interests include braids, knots, and low-dimensional topology, counting absolutely everything, and when not actively counting things, coming up with new things to count.

A Variant of the Proof of the Pythagorean Theorem by Huygens

CHERNG-TIAO PERNG
Norfolk State University
ctperng@nsu.edu

Theorem (Pythagorean Theorem)**.** *Let ΔCBA be a right triangle with hypotenuse c and two legs a and b. Then $c^2 = a^2 + b^2$.*

Proof.

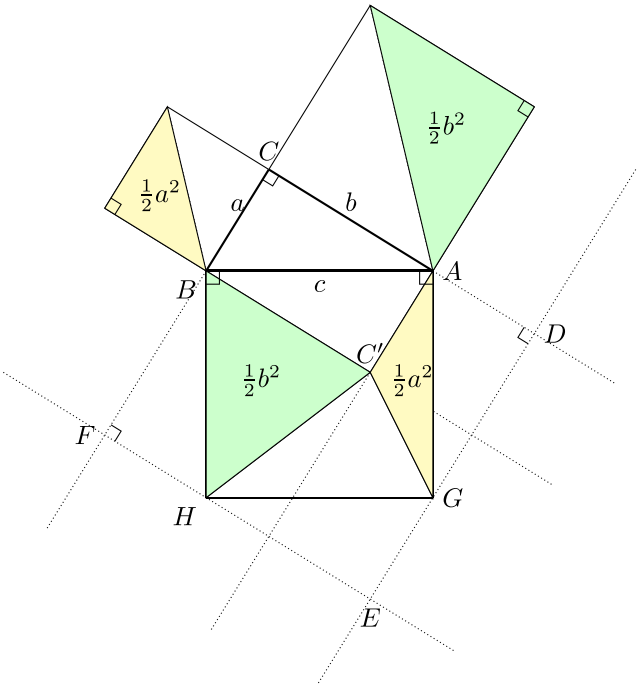


Figure 1 $\frac{1}{2}a^2 + \frac{1}{2}b^2 = \frac{1}{2}c^2 \Rightarrow a^2 + b^2 = c^2$.

A few remarks are in order.

- The proof is wordless and it is inspired by Huygens's proof of the Pythagorean theorem (see the geometric proof number 31 in [**1**, p. 118] or [**2**]).
- In Figure 1, if one looks at the square CFED and ignores what's inside the square BHGA, this gives the algebraic proof number 35 in [**1**, p. 49].
- If one looks at the diamond-shaped polygon BFEDA and extends the line C′E, this gives a slightly different proof (for the geometric proofs, number 77 (p. 147), number 84 (p. 151), number 85 (p. 151), number 42 (p. 125), or number 92 (p. 156) in [**1**]). The essential idea is similar to what Euclid employed in his proof: triangles (or parallelograms) with equal base and height have the same area.

Math. Mag. **92** (2019) 70–71. doi:10.1080/0025570X.2018.1541685
MSC: Primary 51M04
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

REFERENCES

[1] Loomis, E. S. (1968). *The Pythagorean Proposition: Its Demonstrations Analyzed and Classified and Bibliography of Sources for Data of the Four Kinds of "Proofs"*. Reissued second edition (second edition published in 1940). Reston, VA: National Council of Teachers of Mathematics, Inc.

[2] McCarthy, J. P. (1936). Huygens' proof of the theorem of Pythagoras. *Math. Gazette.* 20(240): 280–281.

Summary. We give a variant (proof without words) of the proof of the Pythagorean theorem by Huygens.

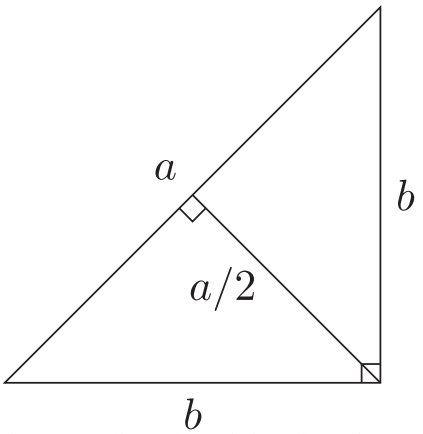
CHERNG-TIAO PERNG (MR Author ID: 984612) received a doctoral degree in pure mathematics (in 2005) and a master's degree in computer science (in 2003) from the University of Pennsylvania. After one year at ECPI College of Technology in Virginia Beach in 2005, he has been teaching and doing research at Norfolk State University; currently he is a tenured Professor in the Department of Mathematics. His main interests include number theory, modular forms, algebraic geometry, actuarial science and game theory. Other than teaching and doing mathematics, he enjoys reading and spending time with family.

An Easy Proof of the Irrationality of $\sqrt{2}$

POO-SUNG PARK, Kyungnam University, Changwon 51767, Korea; pspark@kyungnam.ac.kr

There exist many beautiful proofs of the irrationality of $\sqrt{2}$ (see [**1**] for a rather complete collection) which use a wide variety of approaches. I blend geometric and algebraic elements to present a new short, simple proof.

Assume that $\sqrt{2} = a/b$, where a, b are positive integers and a/b is in lowest terms. Then, $a^2 = 2b^2$, which implies the existence of a minimal right isosceles triangle with hypotenuse length a and legs length b (this is the starting point in Apostol's proof, which appears as Proof 7 in [**1**]), and also forces a to be even.



From this it may be concluded that the altitude to the hypotenuse splits the original right isosceles triangle into two right isosceles triangle with smaller integer side lengths (b for the hypothenuse and $a/2$ for the legs). But this contradicts the minimality of a and b. Thus, $\sqrt{2}$ cannot be rational.

Acknowledgment. The author appreciates the referee's kind and valuable comments.

REFERENCE

[1] Bogomolny, A. (2019). Square root of two is irrational. cut-the-knot.org/proofs/sq_root.shtml.

Summary. If $\sqrt{2}$ is rational, one can construct an isosceles right triangle of minimal integer side lengths. This implies that the hypotenuse is even, leading to a contradiction of the minimality.

POO-SUN PARK (MR Author ID: 803154) obtained a Ph.D. in mathematics from Seoul National University in 2005. He is an associate professor in the Department of Mathematics Education at Kyungnam University, South Korea. His research focuses on the arithmetic theory of quadratic forms. He is also interested in recreational mathematics.

PROBLEMS

EDUARDO DUEÑEZ, *Editor*
University of Texas at San Antonio

EUGEN J. IONAŞCU, *Proposals Editor*
Columbus State University

JOSÉ A. GÓMEZ, Facultad de Ciencias, UNAM, Mexico; *CODY PATTERSON*, University of Texas at San Antonio; *RICARDO A. SÁENZ*, Universidad de Colima, Mexico; *ROGELIO VALDEZ*, Centro de Investigación en Ciencias, UAEM, Mexico; *Assistant Editors*

Correction to Problem 2052

There is a typographical error in Problem 2052 from the October 2018 issue of THE MAGAZINE. Point O_2 should be the circumcenter of triangle $\triangle ACE$ (not of triangle $\triangle ACD$ as published). We apologize for this error and we thank the readers who contacted us to let us know of the error.

Proposals

To be considered for publication, solutions should be received by July 1, 2019.

2061. *Proposed by Florin Stanescu, Şerban Cioiculescu school, Găeşti, Romania.*

Three complex numbers a, b, c satisfy

$$|a| = |b| = |c| = 1 \qquad \text{and} \qquad a^3 + b^3 + c^3 = 2abc.$$

Prove that a, b, c are vertices of an isosceles triangle on the complex plane.

2062. *Proposed by Enrique Treviño, Lake Forest College, Lake Forest, IL.*

For every positive integer n, let $f(n)$ denote the number of occurrences of the digit 2 in the sequence $1, 2, \ldots, n$ of integers written in base 10. (For instance, $f(25) = 9$ because the digit 2 appears once in 2, 12, 20, 21, 23, 24, 25 and twice in 22.)

(*i*) Find a positive integer n such that $f(n) = n$.
(*ii*) Are there infinitely many solutions to $f(n) = n$?

Math. Mag. **92** (2019) 72–78. doi:10.1080/0025570X.2018.1544816.

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 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.

2063. *Proposed by Ovidiu Furdui and Alina Sîntămărian, Technical University of Cluj-Napoca, Cluj-Napoca, Romania.*

Evaluate

$$\sum_{n=0}^{\infty}\sum_{k=1}^{\infty}\frac{(-1)^{n+k-1}}{(n+k)^2}.$$

2064. *Proposed by Ioan Băetu, Botoşani, Romania.*

Characterize those integers $n \geq 2$ such that the ring $\mathbb{Z}_n$ of integers modulo n has a subset F that is a field under the operations of addition and multiplication induced from $\mathbb{Z}_n$. [Note that the unity i of such a field F need not be the unity 1 of $\mathbb{Z}_n$.]

2065. *Proposed by Su Pernu Mero, Valenciana GTO, Mexico.*

Let $\mathcal{Q}$ be a cube centered at the origin of $\mathbb{R}^3$. Choose a unit vector (a, b, c) uniformly at random on the surface of the unit sphere $a^2 + b^2 + c^2 = 1$, and let Π be the plane $ax + by + cz = 0$ through the origin and normal to (a, b, c). What is the probability that the intersection of Π with $\mathcal{Q}$ is a hexagon?

Quickies

1087. *Proposed by Michel Bataille, Rouen, France.*

Let $H_0 = 0$ and $H_n = \sum_{k=1}^{n} k^{-1}$ for $n \in \mathbb{N}$. Given positive integers m, n, prove that

$$\sum_{k=1}^{n} k(H_{m+k} - H_{k-1}) + \sum_{k=1}^{m} k(H_{n+k} - H_{k-1}) = mn + m + n.$$

1088. *Proposed by Luke Harmon and Greg Oman, University of Colorado, Colorado Springs, CO.*

A binary operation $*$ on a set S is *injective* if, for all $a, b, c, d \in S$, the equality $a * b = c * d$ implies $a = c$ and $b = d$. Is there an infinite set with an associative and injective binary operation?

Solutions

Nonarchimedean convexity and an integral inequality **February 2018**

2036. *Proposed by Dan Stefan Marinescu, Hunedoara City and Leonard Giugiuc, Drobeta Turnu-Severin, Romania.*

Let a and b be real numbers with $a < b$. Let $f : [a, b] \to \mathbb{R}$ be a continuous function such that $f(tx + (1 - t)y) \leq \max\{f(x), f(y)\}$ for all $x, y \in [a, b]$ and $t \in [0, 1]$. Prove that if $f(a) = 0$ and $\int_a^b f(x)\,dx = 0$ then $\int_a^b f(x)g(x)\,dx \geq 0$ for all increasing functions $g : [a, b] \to \mathbb{R}$.

Solution by Tom Jager, Calvin College, Grand Rapids, MI.
Let $L = \sup\{x \in [a, b] : f(x) \leq 0\}$. (Since $f(a) = 0$, the supremum above does exist.) We have $f(L) \leq 0$ by continuity of f; thus, for $x \in [a, L]$ we have $f(x)$

$\leq \max\{f(a), f(L)\} \leq 0$. If $L < b$ then from the definition of L we see that $f(x) > 0$ for $x \in (L, b]$. It follows that, for g increasing, we have $f(x)[g(x) - g(L)] \geq 0$ for all $x \in [a, b]$; thus,

$$0 \leq \int_a^b f(x)[g(x) - g(L)]\,dx = \int_a^b f(x)g(x)\,dx - g(L)\int_a^b f(x)\,dx = \int_a^b f(x)g(x)\,dx,$$

since $\int_a^b f(x)dx = 0$ by assumption. (Note that fg is integrable since f is continuous and g increasing on $[a, b]$, thus both are measurable and bounded and so is their product.)

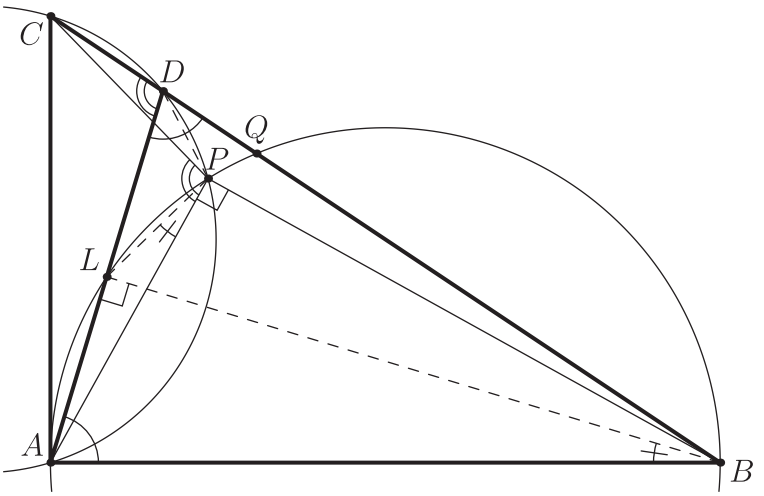
Also solved by Robert Calcaterra, James Duemmel, Robert L. Doucette, Dmitry Fleischman, Eugene A. Herman, Elias Lampakis (Greece), Andrew Lu, Kangrae Park (South Korea), Isaac Wass, and the proposer. There were 3 incomplete or incorrect solutions.

An isosceles triangle inscribed in a right triangle **February 2018**

2037. *Proposed by Ioana Mihăilă, Cal Poly Pomona, Pomona, CA.*

A point D lies on the hypothenuse $\overline{BC}$ of a right triangle $\triangle ABC$ so that $AB = BD$. Let P be the point on the circumcircle of $\triangle ADC$ such that $\angle APB$ is a right angle, and let L be the midpoint of $\overline{AD}$. Show that $\overline{PC}$ is perpendicular to $\overline{PL}$.

Solution by Elton Bojaxhiu (Germany) and Enkel Hysnelaj (Australia).



Since $AB = BD$, the altitude $\overline{BL}$ of the isosceles triangle $\triangle ABD$ is perpendicular to its base $\overline{AD}$ at its midpoint L, and moreover $\angle LDB = \angle LAB$. Quadrilateral $ALPB$ is cyclic since $\angle ALB = \pi/2 = \angle APB$ by construction of P, so we have $\angle APL = \angle ABL$. Since $APDC$ is also cyclic, we have $\angle APC = \angle ADC$ as well. Hence,

$$\angle LPC = \angle APC - \angle APL = \angle ADC - \angle ABL = (\pi - \angle LDB) - \angle ABL =$$

$$= \pi - (\angle LAB + \angle ABL) = \angle ALB = \frac{\pi}{2}.$$

Thus, $\overline{PC}$ is perpendicular to $\overline{PL}$.

Also solved by Michel Bataille (France), Robert Calcaterra, Prithwijit De (India), Andrea Fanchini (Italy), Dmitry Fleischman, Kyle Gatesman, Marty Getz & Dixon Jones, Michael

Goldenberg & Mark Kaplan, Jiseong Ha & Wonjoon Lee (Korea), Eugene A. Herman, Sea Hwang (Korea), Gayoon Jeong (Korea), Soo-Young Kim (Korea), Elias Lampakis (Greece), Kee-Wai Lau (China), Weiping Li, Graham Lord, Peter McPolin (Northern Ireland), José H. Nieto & Laura Queipo (Venezuela), Mingyu Park (Korea), Michael Reid, Ioannis D. Sfikas (Greece), Achilleas Sinefakopoulos (Greece), Neculai Stanciu & Titu Zvonaru (Romania), Michael Vowe (Switzerland), and the proposer. There was 1 incomplete or incorrect solution.

A golden limit of random variables **February 2018**

2038. *Proposed by Eugène Delacroix, Lycée Thérèse d'Avila, France and Su Pernu Mero, Valenciana GTO, Mexico.*

Given any real-valued random variable X, let A_{11}, A_{12}, A_{21}, A_{22} be independent random variables that have the same distribution as X, and let

$$\widetilde{X} = \min_i \max_j A_{ij} = \min\{\max\{A_{11}, A_{12}\}, \max\{A_{21}, A_{22}\}\}.$$

(Although $\widetilde{X}$ does not directly depend on X but rather on the variables A_{ij}, its probability distribution is uniquely determined by that of X.) Define a sequence $\{X_0, X_1, \ldots, X_n, \ldots\}$ recursively by $X_0 = X$ and $X_{n+1} = \widetilde{X_n}$. Prove that, as $n \to \infty$, X_n tends in distribution to a discrete random variable Z taking at most two values. Characterize the distribution of Z in terms of the distribution of X.

Solution by Stephen J. Herschkorn, College of Staten Island, CUNY, Staten Island, NY. Let $F : \mathbb{R} \to [0, 1]$ be the cumulative distribution function (CDF) of X (i.e., $F(x) = \Pr[X \le x]$ for $x \in \mathbb{R}$). Let $\gamma = (\sqrt{5} - 1)/2 = 0.618\ldots$ be the reciprocal of the golden ratio, and let

$$a = \inf\{x : F(x) \ge \gamma\}, \qquad b = \sup\{x : F(x) \le \gamma\}.$$

We show that Z has CDF

$$G = \gamma \mathbf{1}_{[a,\infty)} + (1 - \gamma)\mathbf{1}_{[b,\infty)}.$$

Since F is monotonically increasing and right-continuous, we have $a \le b$ and $a = \min\{x : F(x) \ge \gamma\}$, with $a = b$ iff $F(x) = \gamma$ for at most one x. Otherwise, $F(x) = \gamma$ for $x \in [a, b)$, and G has the same property: It is the distribution function of a random variable taking only the distinct values a, b with probability γ and $1 - \gamma$, respectively, or else the single value $a = b$ with probability 1.

To prove that G is the CDF of Z, define a sequence (F_n) of CDFs by $F_0 = F$ and $F_{n+1} = \widetilde{F_n}$ for $n \ge 0$, where $\widetilde{H} = 2H^2 - H^4$. If A_{11}, A_{12}, A_{21}, A_{22} are independent with common CDF F_n, then $B_1 = \max\{A_{11}, A_{12}\}$ and $B_2 = \max\{A_{21}, A_{22}\}$ are independent with CDF F_n^2 (since $\Pr[B_i \le x] = \Pr[A_{i1} \le x]\Pr[A_{i2} \le x]$), and so $\min\{B_1, B_2\}$ has CDF $F_{n+1} = 1 - [1 - F_n^2]^2 = 2F_n^2 - F_n^4 = \widetilde{F_n}$ (since $\Pr[\min\{B_1, B_2\} > x] = \Pr[B_1 > x]\Pr[B_2 > x]$). It follows by induction that F_n is the CDF of X_n for all n.

The function $\tau(x) = 2x^2 - x^4$ maps $[0, 1]$ to itself; it has fixed points $x = 0, \gamma, 1$ in $[0, 1]$. The factorization $x - \tau(x) = x(x - 1)(x - \gamma)(x + \gamma^{-1})$ shows that $0 < \tau(x) < x$ if $0 < x < \gamma$, and $x < \tau(x) < 1$ if $\gamma < x < 1$. Observing that $\widetilde{H}(x) = \tau(H(x))$, it follows that

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0, & F(x) < \gamma; \\ \gamma, & F(x) = \gamma; \\ 1, & F(x) > \gamma. \end{cases}$$

This limit is an increasing function that agrees with G at all points of continuity thereof, which suffices to prove that G is the CDF of Z.

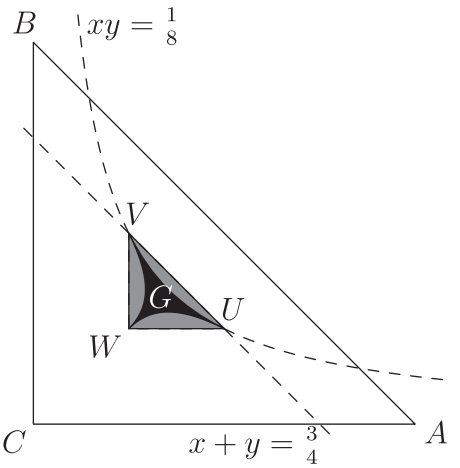
Also solved by Elton Bojaxhiu (Germany) and Enkel Hysnelaj (Australia), Robert Calcaterra, Robert Doucette, Dmitry Fleischman, Kyle Gatesman, Northwestern University Math Problem Solving Group, and the proposer.

The locus of midpoints of median lines of a triangle **February 2018**

2039. *Proposed by Baris Burcin DEMIR, Ali Naili Erdem Anatolian High School, Ankara, Turkey.*

Given a triangle $\triangle ABC$, let $\mathcal{M}$ be the locus of all midpoints P of segments $\overline{DE}$ that divide $\triangle ABC$ into equivalent (i.e., equal-area) parts, where both D and E lie on some side (or possibly a vertex) of $\triangle ABC$. Compute the ratio of the area of the region enclosed by $\mathcal{M}$ to the area of $\triangle ABC$.

Solution by Kyle Gatesman (student), Johns Hopkins University, Baltimore, MD.



We remark that the ratio is the same for all triangles because any two triangles are related by an affine transformation that necessarily preserves ratios of areas. We show that the ratio is $(3/4)\ln 2 - 1/2$ (about 2%).

The area of a region $\mathcal{R}$ will be denoted $[\mathcal{R}]$. We use barycentric coordinates $[x : y : z]$ (where $x + y + z = 1$) on the plane of triangle $\triangle ABC$. First consider the case when $D = [u : 0 : 1 - u]$ lies on $\overline{AC}$ $(0 \le u \le 1)$ and $E = [0 : v : 1 - v]$ on $\overline{BC}$ $(0 \le v \le 1)$. By hypothesis, $uv = [\triangle CDE]/[\triangle ABC] = \frac{1}{2}$. The midpoint of $\overline{DE}$ is $P = [u/2 : v/2 : 1 - (u + v)/2]$. Thus, one piece $\mathcal{M}_C$ of $\mathcal{M}$ is the curve consisting of points $P = [x : y : z]$ with $xy = \frac{1}{8}$ and $\frac{1}{4} \le x, y \le \frac{1}{2}$. The case when D, E lie on $\overline{AB}$ and $\overline{BC}$ (resp., on $\overline{AB}$ and $\overline{AC}$) gives another piece $\mathcal{M}_B : xz = \frac{1}{8}$, and $\frac{1}{4} \le x, z \le \frac{1}{2}$ (resp., $\mathcal{M}_A$: $yz = \frac{1}{8}$ and $\frac{1}{4} \le y, z \le \frac{1}{2}$) of the locus $\mathcal{M}$.

It follows from the equations and inequalities defining $\mathcal{M}_C$ that $x + y \le \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$, so $z = 1 - (x + y) \le \frac{1}{4}$, hence $z \le x$ and $z \le y$ (with strict inequalities except at the endpoints $U = [\frac{1}{2} : \frac{1}{4} : \frac{1}{4}]$ and $V = [\frac{1}{4} : \frac{1}{2} : \frac{1}{4}]$ of $\mathcal{M}_C$). Thus, $\mathcal{M}_C$ is a subset of triangle $\triangle UGV$ (where $G = [\frac{1}{3} : \frac{1}{3} : \frac{1}{3}]$ is the barycenter of $\triangle ABC$) and lies in its interior except for the vertices U and V themselves. We conclude that $\mathcal{M}$ is a simple closed path that starts at $W = [\frac{1}{4} : \frac{1}{4} : \frac{1}{2}]$, follows $\mathcal{M}_B$ to U, then $\mathcal{M}_C$ to V, and finally $\mathcal{M}_A$ back to W.

Hence, the area enclosed by $\mathcal{M}$ equals that of $\triangle UVW$ minus the sum of the areas of each of the three following hyperbolic sectors: $\mathcal{H}_C$ enclosed by $\mathcal{M}_C$ and $\overline{UV}$, $\mathcal{H}_B$ enclosed by $\mathcal{M}_B$ and $\overline{UW}$, and $\mathcal{H}_A$ enclosed by $\mathcal{M}_A$ and $\overline{VW}$. Since every permutation of the vertices of $\triangle ABC$ is realized by an (area-preserving) affine transformation, these three hyperbolic sectors are equivalent. Without loss of generality, let A, B and C have Cartesian coordinates $(1, 0)$, $(0, 1)$, and $(0, 0)$, respectively. Then $\triangle UVW$ has vertices $U = (\frac{1}{2}, \frac{1}{4})$, $V = (\frac{1}{4}, \frac{1}{2})$ and $W = (\frac{1}{2}, \frac{1}{2})$, while $\mathcal{H}_C$ is bounded by the hyperbola $xy = \frac{1}{8}$ and the line $x + y = \frac{3}{4}$, so

$$\frac{[\mathcal{H}_C]}{[\triangle ABC]} = \frac{\int_{1/4}^{1/2} \left(\frac{3}{4} - x - \frac{1}{8x}\right) dx}{1/2} = \frac{3}{16} - \frac{1}{4} \ln 2.$$

Since $[\triangle UVW]/[\triangle ABC] = (1/32)/(1/2) = 1/16$, the area enclosed by $\mathcal{M}$ is

$$\frac{[UVW] - [\mathcal{H}_A] - [\mathcal{H}_B] - [\mathcal{H}_C]}{[\triangle ABC]} = \frac{[UVW]}{[\triangle ABC]} - 3\frac{[\mathcal{H}_C]}{[\triangle ABC]} = \frac{1}{16} - 3\left(\frac{3}{16} - \frac{1}{4} \ln 2\right)$$
$$= \frac{3}{4} \ln 2 - \frac{1}{2}.$$

Editor's Note. The locus $\mathcal{M}$ is the envelope of all area-bisecting segments $\overline{DE}$; it is known as the *(area-bisecting) deltoid* of $\triangle ABC$. J. A. Dunn, J. E. Pretty, Halving a triangle, *Math. Gaz.* **56** (1972) 105–108. We thank Celia Schacht for bringing to our attention a recent article in this MAGAZINE that states many classical properties of the area-bisecting deltoid as motivation for the study of the perimeter-bisecting deltoid. A. Berele, S. Catoiu, Bisecting the Perimeter of a Triangle, *Math. Mag.* **91**:2 (2018) 121–133, doi.org/10.1080/0025570X.2017.1418589.

Also solved by Elton Bojaxhiu (Germany) and Enkel Hysnelaj (Australia), Robert Calcaterra, Timothy Craine, Robert Doucette, Elias Lampakis (Greece), José Nieto (Venezuela), Michael Reid, Celia Schacht, Michael Vowe (Switzerland), and the proposer.

The binomial theorem and m-th roots of matrices — February 2018

2040. *Proposed by George Stoica, Saint John, New Brunswick, Canada.*

Fix a positive integer n. Let A be an $n \times n$ complex matrix such that $A^n = 0$. For any complex number $z \neq 0$ and positive integer m, prove that there exists a matrix B such that $B^n = 0$ and $A + z^m I = (B + zI)^m$, where I denotes the $n \times n$ identity matrix.

Editor's Note. Due to a typographic error, the condition $B^n = 0$ above was incorrectly printed as $B^m = 0$ in the February 2018 Problems column. This mistake was brought to our attention by several readers. We offer our apologies.

Solution by Robin Chapman, University of Exeter, Exeter, UK.
Recall that for any real or complex number α and nonnegative integer k the binomial coefficient $\binom{\alpha}{k}$ is defined by

$$\binom{\alpha}{k} = \frac{\alpha(\alpha - 1) \ldots (\alpha - k + 1)}{k!},$$

with the usual convention $\binom{\alpha}{0} = 1$. By the binomial theorem, the formal power series

$$F_m(T) = (1 + T)^{1/m} = \sum_{k=0}^{\infty} \binom{1/m}{k} T^k$$

satisfies

$$[F_m(T)]^m = 1 + T$$

as an identity of formal power series. Let

$$F_m^{(n)}(T) = \sum_{k=0}^{n-1} \binom{1/m}{k} T^k$$

be the truncation of $F_m(T)$ after n terms. Thus, $F_m^{(n)}(T) \equiv F_m(T) \pmod{T^n}$ (i.e., $F_m^{(n)}(T) - F_m(T) = T^n H(T)$ for some power series H). It follows that

$$[F_m^{(n)}(T)]^m \equiv [F_m(T)]^m = 1 + T \pmod{T^n}. \tag{1}$$

Given $z \neq 0$, since $A^n = 0$ by hypothesis, evaluating (1) above at $T = z^{-m}A$ in the complex polynomial algebra $\mathbb{C}[A]$ gives

$$(I + z^{-1}B)^m = I + z^{-m}A, \tag{2}$$

where $B = z[F_m^{(n)}(z^{-m}A) - I]$. Multiplying both sides of equation (2) by z^m we see that $A + z^m I = (B + zI)^m$. Since $F_m^{(n)}(T) \equiv 1 \pmod{T}$, we see that $B \equiv 0 \pmod{A}$, hence $B^n \equiv 0 \pmod{A^n}$. Since $A^n = 0$, we have $B^n = 0$.

Also solved by Robert Calcaterra, Robin Chapman (UK), Dmitry Fleischman, Eugene Herman, Tom Jager, Michael Reid, Jeffrey Stuart, and the proposer.

Answers

Solutions to the Quickies from page 73.

A1087. Let $S(m, n) = \sum_{k=1}^{n} k(H_{m+k} - H_{k-1})$. Note that $H_k = \int_0^1 \sum_{j=1}^{k} t^{j-1} dt$. Thus,

$$S(m,n) = \sum_{k=1}^{n} \left[k \cdot \sum_{j=k}^{m+k} \int_0^1 t^{j-1}\, dt \right] = \sum_{k=1}^{n} \left[\int_0^1 kt^{k-1} \cdot \sum_{j=0}^{m} t^j\, dt \right]$$

$$= \int_0^1 \left[\sum_{k=1}^{n} kt^{k-1} \cdot \sum_{j=0}^{m} t^j \right] dt = \int_0^1 \left[\sum_{j=0}^{m} t^j \right] d\left(\sum_{k=0}^{n} t^k \right).$$

Integrating by parts, we obtain

$$S(m,n) = \left(\sum_{j=0}^{m} t^j \cdot \sum_{k=0}^{n} t^k \right) \Bigg|_0^1 - \int_0^1 \sum_{k=0}^{n} t^k\, d\left(\sum_{j=0}^{m} t^j \right) = (m+1)(n+1) - 1 - S(n,m).$$

The result follows immediately.

A1088. The answer is no. We show that a nonempty set S with an associative and injective binary operation must have only 1 element. Let $*$ be associative and injective on $S \neq \emptyset$. Let $a, b \in S$ be arbitrary. By associativity, have $a * (b * a) = (a * b) * a$. By injectivity, we have $a = a * b$ and $b * a = a$. Swapping a and b, we also have $a * b = b$; thus, $a = a * b = b$ for all $a, b \in S$, showing that S has exactly one element a.

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.

Roe, John, Russ deForest, and Sara Jamshidi, *Mathematics for Sustainability*, Springer, 2018; xix + 523 pp, $69.99. ISBN 978-3-319-76659-1.

This is a most revolutionary book, and I would love to teach a course using it! There is much interest in what is variously termed quantitative literacy (QL), quantitative reasoning (QR), and critical thinking (CT). This book is the first in a series of Texts for Quantitative Critical Thinking, whose topics have "high sophistication, low prerequisite" and are about "thinking, not computation and procedure." The chapters are about measurement, flows, networks, change, risk, and decision-making, with eight case studies, each denominated as analysis (is recycling worth it?) or advocacy (nuclear power?). QL/QR/CT has little to do with college-level mathematics, so appropriately the "mathematics" in this book is mainly the arithmetic of converting units and plugging values into algebraic and logarithmic expressions. For example, exponential growth is covered in terms of power functions, with no mention of the exponential function. Exercises include computational problems, but the main emphasis is on student writing in response to questions such as "What do you think that means? Do you agree? What do *you* think?". In that respect, the book brings home to students the issues of sustainability, conveys numerous insights about environmental issues, and may convince them that quantitative considerations are important in considering them. The mathematical prerequisite for the book is only high school algebra, so that the book might seem appropriate for high school. But learning from this book would require more maturity, both conceptual and experiential, and involve more subtlety, than many seniors in high school yet possess; also, teaching a course from this book would require more familiarity with natural and social science than many teaching assistants are confident about. The book's philosophy is encapsulated in "**your first loyalty should be to the truth**" (p. 411, bolded there). That is a highly laudable ideal, particularly important in an era of denial of "inconvenient" objective facts, accusing those who bring them up as being fake-news conspirators, and insistence on person-centered loyalty over truth. I do not want to fail to mention that the book's layout is simply beautiful.

Cheng, Eugenia, *The Art of Logic in an Illogical World*, Basic Books, 2018; xii + 305 pp, $27. ISBN 978-1-5416-7248-2.

In college, I took a required course in "Philosophical Logic" from the Dept. of Philosophy. Already knowing symbolic logic (propositional and predicate calculi) made parts of the course easy for me. Author Cheng has written a book for a course that I would title "Sociological Logic." Venn diagrams, implication arrows, and even a few trees and vertex-edge graphs appear; but she avoids symbolic logic—no truth tables, no deduction schemata. Instead, in easy-reading first-person style, she explores the role of logic in life, drawing examples from contemporary issues (sexual harassment, police shootings, white privilege, airline removal of a passenger). Not least, she discusses webs of causation, "how to be right," truth vs. illumination, how to convince people, paradoxes, gray areas in life, perceptions, emotional responses to logical arguments, and manipulation—and how to have a genuinely good argument with another person.

Burger, Edward B., *Making Up Your Own Mind: Thinking Effectively through Creative Puzzle-Solving*, Princeton University Press, 2019; x + 115 pp, $19.95. ISBN 978-0-691-18278-0.

I have lamented elsewhere[1] the lack of evidence for learning mathematics as a way to "train the mind." Ed Burger is the author of mathematics textbooks and also *The Five Elements of Effective Thinking* (2012) (with Michael Starbird). This new booklet may forge a missing link in the argument for the value of learning mathematics. It shows how to use solving logical and mathematical puzzles (even if "whimsical") to teach effective thinking in all realms of life. This goal accords with a commitment at Southwestern University (he is its president) that every course offer "intentional opportunities" to think "through" the material—that is, "to discover the utility and power of that thinking beyond the subject itself." All the value of his course based on this book is indeed "beyond the subject itself"—since ultimately the subject (puzzles and riddles) doesn't really matter in life. "The ultimate goal is not to solve the riddle at hand, but rather to apply multiple practices of effective thinking.... Solving the puzzle is like receiving the diploma—it's **not** *the thing*.... *The thing* is the journey itself...." Burger explains his "five elements of effective thinking," followed by eight collections of three puzzles each (easy, medium, challenging), a chapter (printed upside-down) of prompts for each puzzle to the principles of effective thinking, and a chapter (printed mirror-imaged) of "insights" into them. The over-arching theme is to reflect: What principles did I apply? What insights were provoked? How can I see the puzzle in a different way? A final chapter details how he teaches the course, which includes assignments to write careful solutions, to turn in examples of applications of effective thinking in assignments for other courses, and to compose reflective essays.

Duchin, Moon, Geometry v. gerrymandering, *Scientific American* 319 (5) (November 2018) 49–53.

Kueng, Richard, Dustin G. Mixony, and Soledad Villar, Fair redistricting is hard. https://arxiv.org/abs/1808.08905.

Stephanopoulos, Nicholas O., and Eric M. McGhee, Partisan gerrymandering and the efficiency gap, *University of Chicago Law Review* 82 (2015) 831–900. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2457468.

The U.S. states will redraw their electoral districts after the 2020 Census. What does it mean for a redistricting plan to be fair, and how can we tell? Legal means that each district *must* have approximately the same population. Also, districts *should* be geographically "compact," a concept in search of a definition. Author Duchin rejects an "eyeball test" of the shape of a district in favor of comparing the plan to the universe of all plans to see if it is an outlier in the number of seats produced for each party in the previous election. Kueng et al. prove that finding a "fair" redistricting among legal plans (with some notion of compactness) is NP-hard. Duchin's solution is to do a Markov chain Monte Carlo random walk through the space of plans; ergodic theory guarantees that if you walk long enough, the plans that you encounter will be "representative" of the entire space, and hence you can assess whether a proposed plan is an outlier. Duchin does not make explicit what it means for a plan to be fair; Kueng et al. offer a weak form of proportionality; Stephanopoulos and McGhee suggest an "efficiency gap" measure of partisan symmetry; and Germany and other European countries mandate quasi-proportional representation in their legislatures (if a party gets a fraction f of the votes—above some threshold—it gets approximately fraction f of the seats).

Sackrowitz, Harold, The point(s)-after-touchdown decision revisited, *Chance* 31 (3) (2018) 29–36.

In 2015, the U.S. National Football League changed placement of the ball for the one-point conversion (kick) after touchdown. Author Sackrowitz revisits an earlier article of his about strategy to attempt a one-point or a two-point conversion, based on the possibility that now "going for two" has an expected value greater than one point (there are not enough data to say for sure). (Unfortunately, there are no references for earlier papers cited or alluded to (including the author's earlier piece); editors should insist on such documentation.)

[1] Mathematics: Unreasonably ineffective?, *The UMAP Journal of Undergraduate Mathematics and Its Applications* 39 (1) (2018) 1–4.