LETTER FROM THE EDITOR

I am writing this letter shortly after receiving the sad news that Richard K. Guy has passed away at the age of 103. Among many other accomplishments, he was the recipient of a Lester R. Ford award for mathematical exposition and a coauthor, with John Conway and Elwyn Berlekamp, of the classic *Winning Ways for Your Mathematical Plays*. The chessplayers in the audience might remember him as a former editor of the endgames column in the *British Chess Magazine*, and cocreator of the now standard GBR code for categorizing endgame studies. (He was the "G." The other two letters represent Hugh Blandford and John Roycroft.)

Let me tell you about the one time I got to work with Richard Guy.

I was one of the editors, with Jennifer Beineke, of the proceedings volumes for the biennial MOVES conferences hosted by the Museum of Mathematics in New York City. The conferences are dedicated to research in recreational mathematics. ("MOVES" is an acronym for "The Mathematics of Various Entertaining Subjects.") Richard Guy participated in the conference, and we were very excited when he sent us a contribution for the book.

Guy's paper was a masterful elucidation of various difficult topics in Euclidean geometry, but it needed a fair amount of revision and copy editing before we could publish it. For example, I felt that he was assuming considerable prior familiarity with obscure jargon and theorems, and I suggested lengthening the paper by explicitly including some of this background. The ensuing conversation involved us sending numerous drafts back and forth, before we arrived at the final, published, version. Guy was incredibly easy-going and fun to work with during this whole process.

There was one point in the paper where Guy included a diagram that frankly looked like it was drawn by a crazy person. It showed some thirty-two different circles with various straight lines slashing across them. I spent a significant part of an afternoon struggling to understand what the diagram was meant to illustrate. When I finally saw the point I was amazed, for it was really quite beautiful. However, I felt that the reader deserved some warning before being punched in the face by this diagram. I inserted a parenthetical telling readers they might have to stare at the diagram for a while before the point became clear.

Guy replied that he did not feel the parenthetical was necessary. "For most readers," he said, "no amount of staring at that diagram will be much help."

I am still smiling at that remark to this day.

Of course, anything that can induce a smile these days is worth contemplating. As I write this, the outbreak of the COVID-19 virus, and the ensuing disruptions and quarantines, have now affected literally everyone. Those of us who work in schools of any kind have been scrambling to adapt our classes to online formats.

Perhaps, then, I can offer this issue of *Mathematics Magazine* as a brief and meager respite from the grim news of the day. We have an excellent and varied assortment of mathematical exposition for you to consider.

Our lead article, by Hans Humenberger, explores the tricky subject of slicing up triangular pizzas. The problem is not so much the slicing itself, but rather constructing the slices so that they bisect both the pizza and the crust (the area and the perimeter, in other words). Humenberger starts with a classic version of the problem, which admits a wonderful, visual solution, and then proceeds to consider several variations.

Math. Mag. **93** (2020) 162–163. doi:10.1080/0025570X.2020.1748426

Geometry and probability come together in our next article, by T. Kyle Petersen and Bridget Eileen Tenner. Their starting point is the famous "broken stick" problem—if a stick of fixed length is broken at two interior points, thereby forming three pieces, what is the probability the three pieces can be formed into a triangle? This problem is quickly generalized to breaking the stick at $k-1$ interior points and then trying to make a k-gon from the resulting pieces, and the resulting complications are both beautiful and engrossing.

The number theorists in the audience will enjoy the article by Rebecca L. Jayne and Robb T. Koether. This time it is the famous locker problem that is extended and generalized. You know the problem I mean—One thousand students numbered sequentially from 1 to 1000 are sent one at a time past 1000 lockers that are all initially closed. If each student reverses all the lockers that are multiples of her number, then which lockers are open at the end? Jayne and Koether have many insightful things to say about this problem and its various generalizations.

If you prefer abstract algebra then you can have a go at Alan Beardon's article, which helps to clarify a tricky question in the theory of complex exponentiation—if z and w are complex numbers, should z^w be regarded as a single number, or as a whole set of numbers? Fans of linear algebra can sink their teeth into Jeff Suzuki's contribution, which discusses the perennially popular topic (in some quarters!) of finding eigenvectors and eigenvalues without the use of determinants.

From Travis Kulhanek and Vadim Ponomarenko we have an explanation of why surprises in knockout tournaments are actually unsurprising, and from Martin Lukarevski we have an ingenious short proof of a well-known inequality in Euclidean geometry. We round out the issue with Problems, Reviews, and a pair of Proofs Without Words.

Jason Rosenhouse, Editor

ARTICLES

Fair Sharing of Triangular Pizzas

HANS HUMENBERGER
Faculty of Mathematics
University of Vienna, Austria
hans.humenberger@univie.ac.at

An email from Grégoire Nicollier pointed me on the one hand to a beautiful problem of elementary geometry, and on the other hand to a beautiful solution [5]. The problem was about sharing a pizza fairly: Imagine a pizza shaped like an equilateral triangle. If the pizza is shared fairly, then everybody should get the same amount of "area" and the same amount of "crust," and this was achieved by making the cuts in the way described in the problem's solution. Of course, in reality there are no such pizzas, and it was not clear how to produce the solution's indicated cuts in reality. The problem of fair pizza division has received considerable attention from mathematicians (see, e.g., Carter and Wagon [1], Humenberger [3], and Mabry and Deiermann [4]), but the theory and reasoning behind the solution are more important than the illusion of modeling real life situations.

Nicollier's result is presented as follows: If an equilateral triangular pizza is divided by six straight cuts going from an arbitrary interior point to the vertices and to the sides at right angles, then two people share the pizza and the crust fairly by taking alternate slices.

The diagram in Figure 1 makes everything clear "without words."

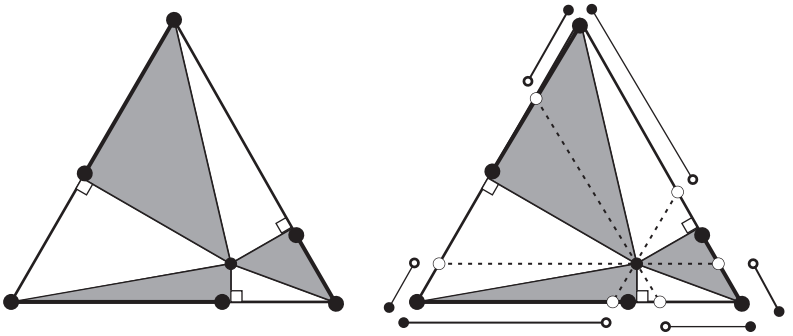


Figure 1 Nicollier's solution to the fair-sharing problem for equilateral triangle pizzas. The left side shows the six cuts arising from an arbitrary interior point. The right side makes clear that both the area (pizza) and perimeter (crust) are bisected by this procedure.

Such proofs without words are useful in teaching situations. The task for students would be to find the corresponding words of explanation. Maybe this method for

Math. Mag. **93** (2020) 164–174. doi:10.1080/0025570X.2020.1742553
MSC: Primary 51M04
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

handling proofs should be used more often. To *find* comparably simple, but nevertheless substantial, proofs is something completely different, but to interpret them can be a good exercise for students: What can be seen in the figure? What does it tell us? Why does the picture prove the statement (theorem)? In this case the situation is sufficiently straightforward that students at school or university could prove the corresponding theorem by themselves (i.e., without a proof figure offered to them), but to interpret an ingenious figure, to find the missing words, may also be an interesting task.

What about non-equilateral triangles?

Presumably, in non-equilateral triangles things will not be as simple, but it is natural to ask further questions.

First, we introduce the following terminology:

- *Area-bisecting property*: The total area of the gray pizza pieces equals the total area of the white pizza pieces.
- *Crust-bisecting property*: The total crust length of the gray pizza pieces equals the total crust length of the white pizza pieces.

We now ask: Are there interior points with the area-bisecting property and the crust-bisecting property in the case of a non-equilateral triangle pizza and using the tiling method above? If yes, what are they, and if not, why not?

Let us now leave the pizza context and stick to pure geometry.

One can quickly find a partial answer to the question (we deliberately do not ask for the locus of all such points, which is a more difficult question, discussed below). In the case of an isosceles triangle, the method works for points on the axis of symmetry. In the case of a scalene triangle, the incenter I is such a point, as is the circumcenter O, if it lies in the interior of the triangle, as it does for acute triangles.

(Recall that the incenter and circumcenter of a triangle are, respectively, the centers of the inscribed and circumscribed circles of the triangle. Later we will use the fact that the incenter is the intersection point of the three angle bisectors.)

Let us now restrict to just one of the bisecting properties. The case of the area-bisecting property is more complex, but the investigation of the crust-bisecting property is an interesting problem appropriate for students.

Let us state the problem formally:

Problem 1. *Given a non-equilateral triangle ABC, are there points P in the interior of the triangle having the crust-bisecting property? If yes, find all such points, and if not, why not?*

Figure 2 illustrates the sort of point P we are looking for.

For students, there are several possibilities for exploring the situation.

1. One can use Dynamic Geometry Software (DGS), such as GeoGebra and Sketchpad and use its tools for measuring. One can experiment by moving the point P and monitoring the value of $|\text{crust gray}| - |\text{crust white}|$. For which choices of P does this difference vanish? The students will notice that there are many such points and that they seem to lie on a straight line. But how do we describe that straight line?
2. On the other hand, there are two very special and famous points for which the crust-bisecting property can be quickly seen and proven: the circumcenter O and the incenter I. In this approach a program like GeoGebra can again help. The students can experimentally affirm that all points of the straight line IO seem to have this property.

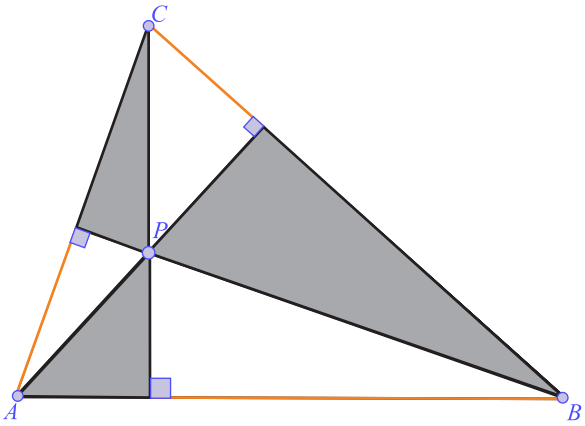


Figure 2 Which points P have the crust-bisecting property?

Through careful experimentation, one will find that all the points on the straight line IO have the crust-bisecting property and that other points do not have it. For our problem, only points in the interior of the triangle are relevant. The straight line IO is shown in Figure 3. Note that in an obtuse triangle, O lies outside the triangle.

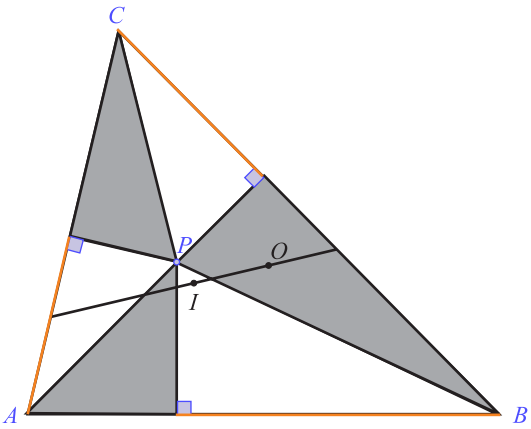


Figure 3 The straight line joining the incenter I to the circumcenter O.

At this point we should have a look back: In the case of an equilateral triangle, the points O and I coincide, and there is no unique straight line joining them. We can thus view any point P within the triangle as lying, vacuously, on the line IO, and therefore our result can be said to apply to equilateral triangles as well. Moreover, in an isosceles triangle, the line joining O and I is precisely the axis of symmetry, which makes our claim plausible in this case as well.

In the following, we denote by g the straight line joining O and I in a non-equilateral triangle ABC.

The proof of our claim comes in two parts:

1. All points on g have the crust-bisecting property.
2. No other points have this property.

If r, s, and t represent various lengths of crust, then the changes in these lengths arising from changes in our choice of interior point will be denoted by Δr, Δs, and Δt. Since I and O have the crust-bisecting property, it is clear that in executing the

transition from I to O along the straight line g, the total crust change $\Delta r + \Delta s + \Delta t$ vanishes. This is shown in Figure 4.

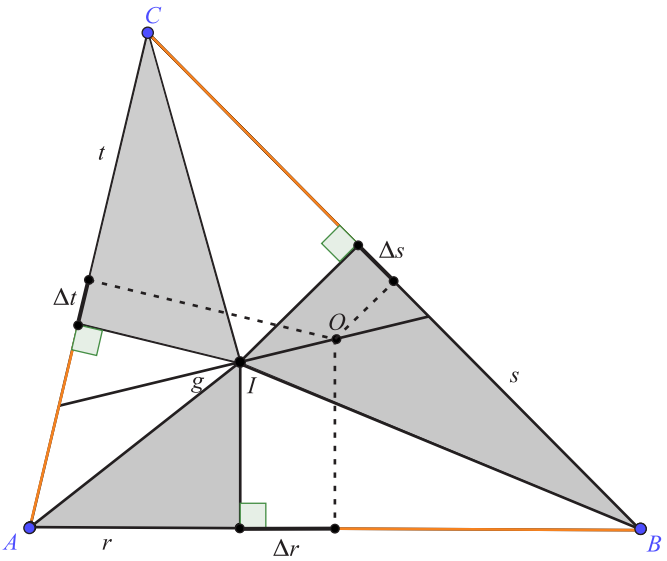


Figure 4 We find that $\Delta r + \Delta s + \Delta t = 0$ when $P = I$ is moved to $P = O$. In this transition, we have that $\Delta r > 0$ and $\Delta s, \Delta t < 0$.

By using the theorem of proportional segments, we can argue as follows: If P (starting at $P = I$) moves to another point $P \neq I \in g$ with $\overrightarrow{IP} = k \cdot \overrightarrow{IO}$, then the changes $\Delta r, \Delta s, \Delta t$ are all multiplied by k. That is, the new corresponding changes are $k \cdot \Delta r, k \cdot \Delta s, k \cdot \Delta t$. The sum of these changes is

$$(k \cdot \Delta r) + (k \cdot \Delta s) + (k \cdot \Delta t) = k(\Delta r + \Delta s + \Delta t) = 0,$$

which completes the proof of statement 1.

For the proof of statement 2 we introduce a functional point of view: Let $f(P) = r + s + t$ be the sum of the lengths of the gray crusts. Suppose we know of two different points $P_1 \neq P_2$, with corresponding function values $f(P_1) = c_1$ and $f(P_2) = c_2$. If we move both points by the same vector $\vec{v}$ and define

$$P_1 + \vec{v} = Q_1 \quad \text{and} \quad P_2 + \vec{v} = Q_2$$

then we have that

$$f(Q_1) = c_1 + \Delta f \quad \text{and} \quad f(Q_2) = c_2 + \Delta f.$$

This follows because the vector $\vec{v}$ has unique components in the directions of the triangle's sides, and these components determine $\Delta r, \Delta s$, and Δt, and thus also Δf, regardless of the point from which the process starts.

This phenomenon has two important consequences.

First, points on lines h parallel to g all have the same f-values because to every such parallel line one can draw two equal vectors $\vec{v}$ from I and from O, terminating at Q_1 and Q_2. This is shown in Figure 5. This phenomenon can easily be confirmed by DGS experiments: When moving a point on a straight line parallel to g, the values of the function f do not change. Furthermore, one will observe that the values of f are smaller than the semiperimeter of the triangle on one side of g and bigger on the other. It remains to be shown that different parallel lines have different f-values. Once this step is accomplished, we will have proven that the parallel lines to g are a division of

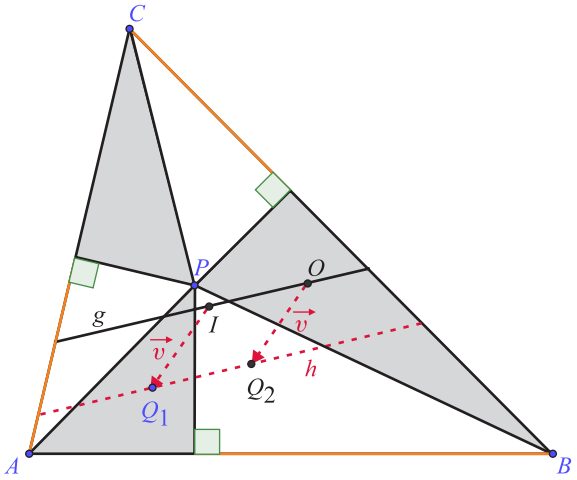


Figure 5 Equal f values on lines h parallel to g.

the triangle into regions with constant f-values, though different lines have different values. The idea is illustrated in Figure 6.

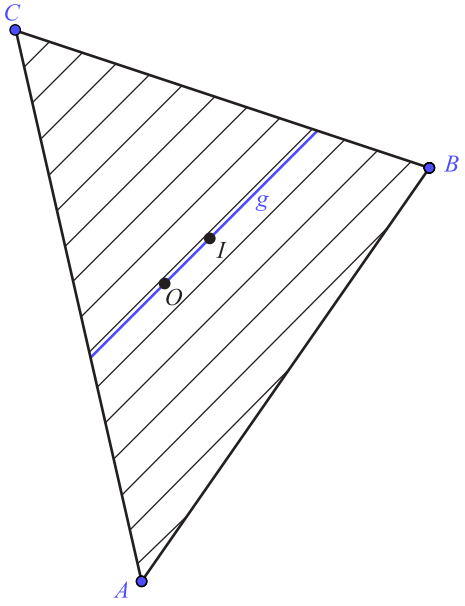


Figure 6 Points on different parallel lines to g have different f-values.

Second, if we translate by the vector $2\vec{v}$, then the change Δf of the function is two times the change produced via translation by $\vec{v}$. That means that if the distance of a parallel line p_2 to g is twice the distance of another parallel line p_1, then the change Δf for points on p_2 is twice the change for points on p_1.

Therefore, if we can show that there exists a line parallel to g with different f-values, then we are done. Proving this turns out to be easy for non-equilateral triangles.

In a non-equilateral triangle there are always at least two angle bisectors that do not coincide with altitudes. Let us assume that the angle at A is one such. When moving P from its initial position at I along this angle bisector, the function f changes. This is shown in Figure 7. We then have $f(I) = r + s + t$, and

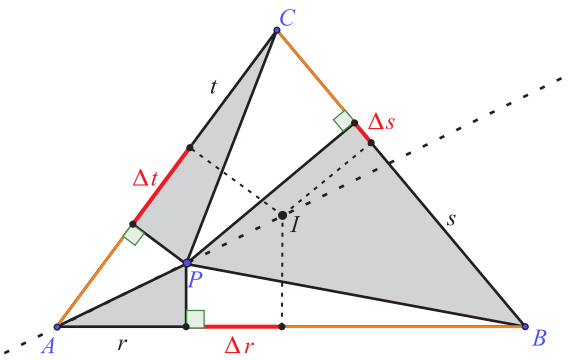


Figure 7 Translating P along the bisector of the angle at A. Note that the starting position was when $P = I$.

$$f(P) = (r + \Delta r) + (s + \Delta s) + (t + \Delta t).$$

Since P lies on the bisector of angle A, we get $\Delta r = -\Delta t$. (In Figure 7 we have $\Delta r < 0$.) And since this angle bisector is not an altitude, we can conclude that $\Delta s \neq 0$. Thus, $f(P) \neq f(I)$ which proves item 2.

Slicing triangular pizzas with cevians

We have already noted that the generalization of the original proof without words to the case of scalene triangles for area-bisection, instead of crust-bisection, is more difficult. The locus sought turns out to be the so called *Stammler hyperbola*, a special conic section through O and I, see Embacher and Humenberger [**2**]. One can see easily that both the circumcenter O and the incenter I do have this property. However, a full solution in this context is not so easy. For our purposes here, let us simply accept this result without proof. We can then give an answer to the simultaneous area and crust problem: The only common points of the Stammler hyperbola and the line joining I and O are I and O themselves. Therefore, $\{O, I\}$ is the solution to the simultaneous area and crust problem.

We will consider instead a different problem related to area-bisection. Rather than use the slicing procedure described by Nicollier, we shall use one based on cevians.

As background, recall that a *cevian* is a line segment connecting a vertex of a triangle with a point on the opposite side. The English word "cevian" comes from the Italian mathematician Giovanni Ceva (1647–1734). Medians and angle bisectors are examples of cevians.

We can now state our problem:

Problem 2. *Suppose you are given an arbitrary triangle ABC and an arbitrary point P inside the triangle. The three cevians through P divide the triangle into* 6 *smaller triangles, which we alternately color gray and white (as shown in Figure 8). Are there*

points P in the interior of the triangle possessing the area-bisecting property? That is, are there points P for which $|\text{area gray}| = |\text{area white}|$? If yes, describe all such points, and if not, why not?

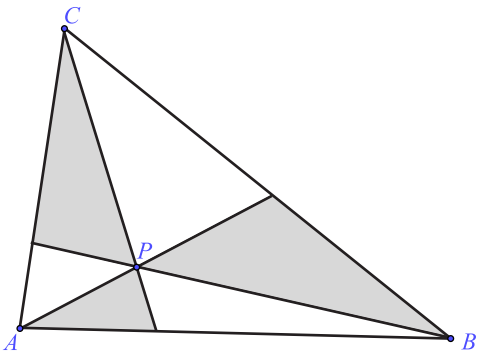


Figure 8 Which points P in the interior of the triangle have the area-bisecting property?

Besides doing DGS experiments, one can develop an intuition for the solution in the following way: If P comes near to a side's midpoint, then four out of the six smaller triangles become arbitrarily small. Only two triangles, with equal area, remain, one of them white and the other gray. Furthermore, if P is the centroid, then it is a standard result that all six smaller triangles (3 white, 3 gray) have equal area. Hence, we already have the centroid and the sides' midpoints as possible positions of P for the area-bisecting property. It is not a very big step to the conjecture that the desired locus of points consists of the points on the medians. This locus is illustrated in Figure 9. DGS experiments readily lend support to this conjecture.

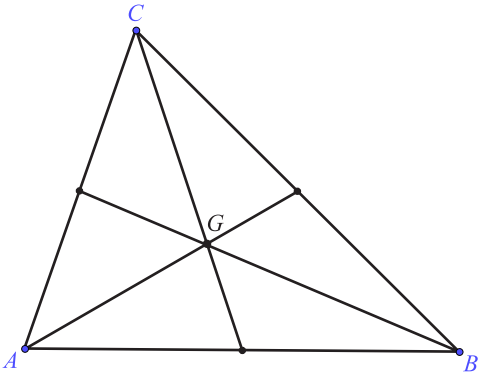


Figure 9 Are the medians the solution to Problem 2?

Now we turn to proving this phenomenon. Once again, we have two statements to prove:

1. All points on the medians do have the area-bisecting property.

2. No other points have this property.

We will restrict the position of P to the interior of the triangle so that there are always six smaller triangles. Therefore, we exclude the midpoints of the sides and the vertices themselves.

In preparation for the proof we remind the reader of some standard results from Euclidean geometery.

Lemma. 1. *In a triangle, the parallels to a side of the triangle—and only these—are bisected by the corresponding median.*

2. *For $P \in m_c$, where m_c is the median from vertex c, let D and E be the intersection points of the cevians through P with the triangle sides BC and AC. Then ED is parallel to AB. In other words, the intersection of the diagonals of quadrilateral PECD lies on the median m_c. This is illustrated in Figure 10.*

We can now prove statement 1. If P lies on a median, say m_c (as shown in Figure 10) then the two triangles $AM_{AB}P$ (gray) and $M_{AB}BP$ (white) have the same area. Since ED is parallel to AB and since the triangles EPF and FPD also have equal area (this follows from the lemma), we conclude that triangles APE (white) and PBD (gray) have the same area, as do triangles EPC (gray) and PDC (white). This completes the proof of statement 1.

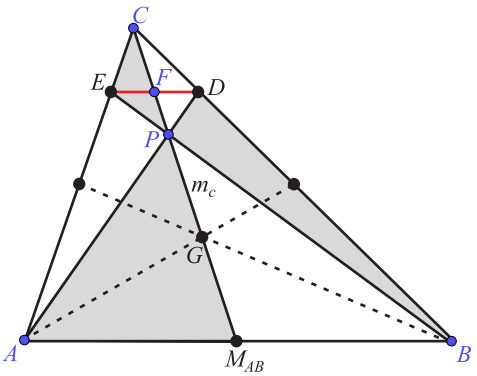


Figure 10 If $P \in m_c$, then the gray and white triangles have equal areas. The quadrilateral $PECD$ and its diagonals illustrate item 2 in the lemma.

Proving the second part is not so easy. We will need Ceva's theorem. Figure 11 shows three cevians of a triangle with endpoints D, E, and F. With the notation as presented in the figure, Ceva's theorem says that the cevians are concurrent if and only if

$$\left(\frac{a_1}{a_2}\right)\left(\frac{b_1}{b_2}\right)\left(\frac{c_1}{c_2}\right) = 1.$$

Equivalently, $a_1b_1c_1 = a_2b_2c_2$.

Let us now assume that P does not lie on a median. Suppose, without loss of generality, that P lies in the interior of the triangle "left hand side, above" (in the other regions the proof would work analogously). We will prove that P then cannot have the area-bisecting property.

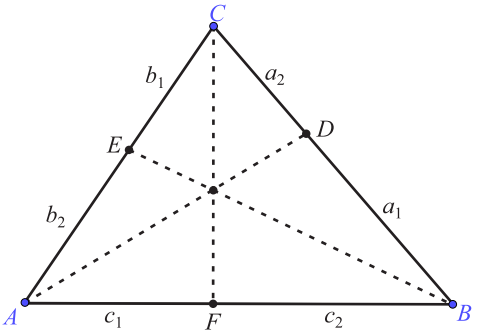


Figure 11 The set-up for Ceva's theorem.

We connect P and B by a straight line. The resulting intersection point with the median m_c we denote by P_1. This is shown in Figure 12.

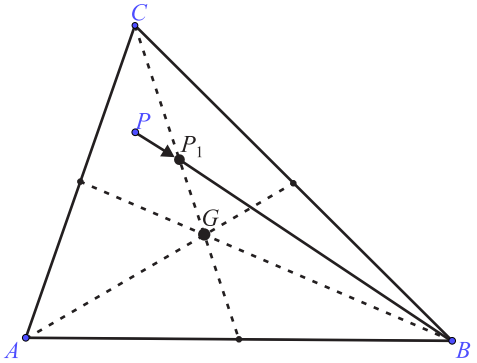


Figure 12 The transition $P \to P_1$.

What happens to the white and gray areas due to the transition $P \to P_1$? If we can show that the area changes (of white and gray resp.) are not 0, then we have also proven that P cannot have the area-bisecting property. This follows because we have established that P_1 *does* have the area-bisecting property.

In Figure 13, the areas that change their color in the direction white $\to$ gray during the transition $P \to P_1$ are marked with a checkerboard pattern. The areas that change their color in the direction gray $\to$ white are hatched. It now suffices to prove that the area of triangle AHL differs from that of the triangle $FM_{AB}C$.

The fraction of the area of triangle ABC occupied by triangles AHL and $FM_{AB}C$ is given, respectively, by the ratios $\dfrac{|FM_{AB}|}{|AB|}$ and $\dfrac{|LH|}{|BC|}$. We ask, why does this inequality hold:

$$\frac{|FM_{AB}|}{|AB|} > \frac{|LH|}{|BC|}? \tag{1}$$

To answer this, we consider the more detailed Figure 14. We define the notation

$$a = \frac{|BC|}{2} \qquad b = \frac{|AC|}{2} \qquad c = \frac{|AB|}{2},$$

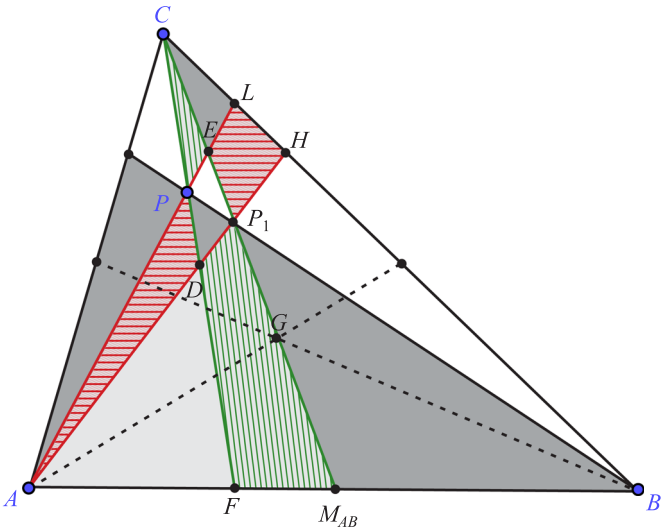


Figure 13 The color changes white → gray are shown with a checkerboard pattern, while the changes gray → white are shown with a hatched pattern.

and we let I be the point of intersection of the cevian CP with the median at B. We can use Lemma to get the parallelisms

$$FJ \,||\, M_{AB}M_{BC} \quad \text{and} \quad KH \,||\, M_{AC}M_{BC}.$$

Then, by using the theorem of proportional segments we get

$$\frac{u}{c} = \frac{x}{a} \tag{2}$$

and

$$\frac{v}{b} = \frac{x+z}{a}. \tag{3}$$

It could happen that the line segment LH extends into the line segment JM_{BC} (i.e., that H is between the points J and M_{BC}). We would then have $z < 0$, but it would still be true that $x + z > 0$ because H lies somewhere "above" M_{BC} (recall that we are assuming that P lies somewhere in the triangle "left side, above.") The other introduced quantities x, y, u, and v are always positive. By using equation (2), we find that inequality (1) is equivalent to $x/a > y/a$, and therefore to the inequality $x > y$.

Now we use Ceva's theorem (applied to the intersection point P of the cevians) to obtain

$$(a + x + z + y)(b - v)(c - u) = (a - x - z - y)(b + v)(c + u).$$

Solving for y yields:

$$y = -x - z + a\left(\frac{bu + cv}{bc + uv}\right).$$

This implies that $x > y$ is true if and only if

$$2x + z > \frac{abu + acv}{bc + uv}.$$

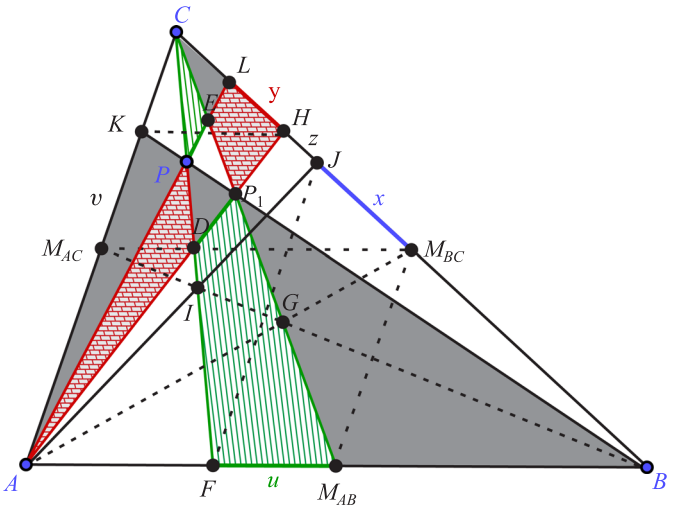


Figure 14 The proof of inequality (1) now reduces to the question: Why does $x > y$ hold?

From equation (2) we have that $au = cx$. Thus, our inequality is equivalent to

$$(2x + z)(bc + uv) > bcx + acv.$$

Dividing both sides by bc gives

$$(2x + z)\left(1 + \frac{uv}{bc}\right) > x + \frac{av}{b}.$$

Combined with equation (3) we get

$$(2x + z)\left(1 + \frac{uv}{bc}\right) > 2x + z,$$

which is clearly true. This completes the proof of statement 2.

REFERENCES

[1] Carter, L., Wagon, S. (1994). Proof without words: fair allocation of a pizza. *Math. Mag.* 67(4): 267. doi.org/10.1080/0025570X.1994.11996228

[2] Embacher, F., Humenberger, H. (2019). A note on the Stammler hyperbola. *Amer. Math. Monthly*. 126(9): 841–844. doi.org/10.1080/00029890.2019.1644125

[3] Humenberger, H. (2015). Dividing a pizza into equal parts—an easy job? *Math. Enthus.* 12(1–3): 389–403.

[4] Mabry, R., Deiermann, P. (2009). Of cheese and crust: a proof of the pizza conjecture and other tasty results. *Amer. Math. Monthly*. 116(5): 423–438. doi.org/10.1080/00029890.2009.11920956

[5] Nicollier, G. (2015). Half issues in the equilateral triangle and fair pizze sharing. *Math. Mag.* 88(5): 337. doi.org/10.4169/math.mag.88.5.337

Summary. How can we slice up a triangular pizza so that two people taking alternate slices receive the same amount of pizza and the same amount of crust? For equilateral triangles, there is an elegant "proof without words" solution to this problem. We investigate several related questions with regard to non-equilateral triangles.

HANS HUMENBERGER has been a professor of mathematics education at the University of Vienna since 2005. In this position, he is responsible for the education of students who want to become teachers at secondary schools or high schools (grade 5–12, 11–18-year-old students). His primary interests include teaching mathematics as a process, teaching mathematical modeling, problem solving, and elementary mathematics.

Broken Bricks and the Pick-up Sticks Problem

T. KYLE PETERSEN*
DePaul University
Chicago, IL 60604
tpeter21@depaul.edu

BRIDGET EILEEN TENNER†
DePaul University
Chicago, IL 60604
bridget@math.depaul.edu

There is a classical probability exercise about forming a triangle from pieces of a stick:

> **The broken stick problem—classical version.** Consider a stick of fixed length. Pick two distinct interior points on the stick, independently and at random, and cut the stick at these two points. What is the probability that the resulting three pieces form a triangle?

For example, if the stick has length 1, then breaking the stick into segments of lengths $1/10$, $3/7$, and $1 - (1/10) - (3/7) = 33/70$ will produce a triangle, whereas breaking it into segments of lengths $1/10$, $3/8$, and $1 - (1/10) - (3/8) = 21/40$ will not produce a triangle, as shown in Figure 1.

Figure 1 Two breakings of a stick into three pieces, one of which can form a triangle and one of which cannot.

The classical broken stick problem apparently dates back to the 1854 exam on pure and applied mathematics at Cambridge University (later known ast the Mathematical Tripos) [**4**]. It can be answered by a nice argument in geometric probability, showing that we produce a triangle with probability $1/4$.

This problem generalizes naturally to arbitrary polygons, as follows.

> **The broken stick problem—general version.** Consider a stick of fixed length and a positive integer $k \geq 3$. Pick $k - 1$ distinct interior points on the stick, independently and at random, and cut the stick at these $k - 1$ points. What is the probability that the resulting k pieces form a k-gon?

The scenario of the broken stick problem has applications to a number of other fields [**6**]. Another application is that, due to Proposition 2, the general broken stick problem is related to a k-candidate plurality election in which no candidate wins a majority of the votes. Many other generalizations and related discussions have appeared [**1**, **3**, **5**, **7**, **8**].

*Research partially supported by Simons Foundation Collaboration Grant for Mathematicians 353772.
†Research partially supported by Simons Foundation Collaboration Grant for Mathematicians 277603.
Math. Mag. **93** (2020) 175–185. doi:10.1080/0025570X.2020.1736888
MSC: 00, 05
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

The generalized broken stick problem has an elegant answer, as shown by D'Andrea and Gómez [2].

Theorem 1. *Take a stick of fixed length and a positive integer $k \geq 3$. Pick $k-1$ distinct interior points on the stick, independently and at random, and cut the stick at these $k-1$ points. The probability that the resulting k pieces form a k-gon is*

$$1 - \frac{k}{2^{k-1}}.$$

The theorem can be proved geometrically, and we present that argument here as motivation. Suppose, without loss of generality, that the stick has unit length, and consider, as the sample space, the interior of the unit simplex

$$\{(x_1, \ldots, x_k) \in (0,1)^k : \sum x_i = 1\}.$$

We call this the "sample space" because we interpret a point $(x_1, \ldots, x_k)$ in this space as describing the stick having been cut at the points

$$0 < x_1 < x_1 + x_2 < \cdots < x_1 + x_2 + \cdots + x_{k-1} < 1,$$

to create segments of lengths $x_1, x_2, \ldots, x_{k-1}$, and $x_k = 1 - (x_1 + \cdots + x_{k-1})$. It transpires (see Proposition 2) that the multiset

$$\{x_1, x_2, \ldots, x_k\}$$

of these lengths describes the side lengths of a k-gon, necessarily of unit perimeter, if and only if

$$x_i < \frac{1}{2} \quad \text{for all } i. \tag{1}$$

The x_i are positive and sum to 1, so inequality (1) can fail for at most one coordinate at a time. Saying, for example, that $x_k \geq 1/2$, is equivalent to requiring that the remaining coordinates satisfy the inequality

$$x_1 + x_2 + \cdots + x_{k-1} \leq 1/2.$$

This defines a subset of the sample space, which can be described as a contraction of the full simplex toward the corner $(0, \ldots, 0, 1)$:

$$(y_1, \ldots, y_k) \mapsto \left(\frac{y_1}{2}, \frac{y_2}{2}, \cdots, \frac{y_{k-1}}{2}, \frac{y_1 + \cdots + y_{k-1}}{2} + y_k \right).$$

The volume of this subset is $1/2^{k-1}$ of the volume of the full simplex. This argument holds for any of the k coordinates failing inequality (1). Hence, the proportion of the sample space that has some coordinate failing that inequality is $k/2^{k-1}$, and so the desired probability is, indeed, $1 - k/2^{k-1}$. Figure 2 depicts the cases $k = 3$ and $k = 4$.

The geometric probability argument for the broken stick problem is beautiful, but for two authors who spend most of their time counting things, a discrete version of the problem has great appeal. Thus, we consider an analogue of the problem in which the stick has integer length and can only be broken at integer increments. We are certainly not the first to take the discrete approach to a geometric probability problem. As reported by Goodman [4], this phrasing of the broken stick problem for triangles ($k = 3$) is nearly as old as the problem itself, going back to a work of Lemoine in 1875. Being sore-footed parents of young children, we think of this discrete version as a "(LEGO) brick analogue."

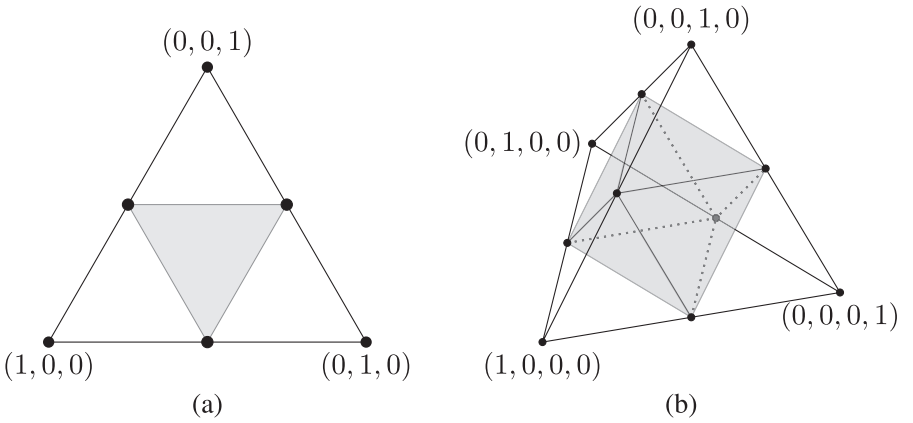


Figure 2 The geometric argument for $k = 3$ and $k = 4$. In (a), we see that the probability of making a triangle is 1/4 of the area of the sample space. In (b), we see the probability of making a quadrilateral is 1/2 of the volume of the sample space.

The broken brick problem. Let $n \geq k \geq 3$ be positive integers, and consider a stick of length n. Pick $k - 1$ distinct interior integer points on the stick, independently and at random, and cut the stick at these $k - 1$ points. What is the probability that the resulting k pieces form a k-gon?

The nice thing about this version of the problem is not only that it enables a combinatorial proof of D'Andrea and Gómez's result, but also that it allows students and even small children to experiment with the question. For example with a stick of $n = 10$ bricks, there are only 36 ways to break the stick into $k = 3$ pieces, and the experimenter can record how many of these breakings result in a triangle. See Figure 3.

Figure 3 Experimenting with sticks of LEGO bricks.

Polygonal inequalities

The classical broken stick problem is a reference to the triangle inequality: a multiset S of three positive numbers gives the side lengths of a triangle if and only if each

(potential) side length is less than the sum of the other two (potential) side lengths. Thus, the first step in solving generalizations of the broken stick problem is to find a k-gon analogue to the triangle inequality. More precisely, given a k-element multiset S of positive numbers (later we will require that they be integers), is there a k-gon whose side lengths are the elements of S? What properties of S must hold for such a k-gon to exist? The following result appears in D'Andrea and Gomez [2], but we include a proof here in order to make their "tweaking" explicit. In what follows, we write $\|S\|$ to denote the sum of the elements of S.

Proposition 2. Fix a positive integer $k \geq 3$ and a k-element multiset S of positive numbers. There exists a (convex) polygon whose side lengths are the elements of S if and only if $x < \|S\| - x$, or equivalently,

$$x < \frac{\|S\|}{2} \tag{2}$$

for each $x \in S$.

Proof. We proceed by induction on k, noting that the case $k = 3$ is precisely the triangle inequality. Assume, inductively, that the result holds for all j-gons, with $3 \leq j < k$.

Suppose that a k-element multiset S contains the side lengths of a k-gon P. Any diagonal of P will separate the region into two polygons Q_1 and Q_2, each with fewer than k sides, as shown in Figure 4. Thus each Q_i is subject to the inductive hypothesis,

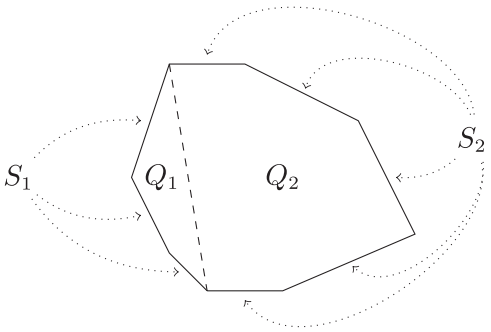


Figure 4 Decomposing a k-gon by drawing a diagonal.

producing collections of inequalities as in inequality (2). Let d be the length of the diagonal separating Q_1 from Q_2, and decompose $S = S_1 \cup S_2$ so that the side lengths of Q_1 are $T_1 = S_1 \cup \{d\}$, and the side lengths of Q_2 are $T_2 = S_2 \cup \{d\}$, as shown in Figure 4. Then by the inductive hypothesis,

$$x < \frac{\|T_i\|}{2}$$

for each $x \in T_i$ and $i \in \{1, 2\}$.

Let us see if the same holds with respect to S. Let $x \in S$. Then in particular, $x \in T_1 \cup T_2$ and $x \neq d$. Without loss of generality, suppose $x \in T_1$. First of all,

$$x < \frac{\|T_1\|}{2} = \frac{\|S_1\| + d}{2}.$$

Likewise,

$$d < \frac{\|T_2\|}{2} = \frac{\|S_2\| + d}{2}.$$

Therefore $d/2 < \|S_2\|/2$, and so

$$x < \frac{\|T_1\|}{2} = \frac{\|S_1\| + d}{2} < \frac{\|S_1\| + \|S_2\|}{2} = \frac{\|S\|}{2},$$

as desired.

Now suppose that $S = \{s_1, \ldots, s_k\}$ is a k-element multiset and that

$$x < \frac{\|S\|}{2} \tag{3}$$

for all $x \in S$. We want to construct a k-gon whose side lengths are the elements of S. To do this, our goal is to find a value d such that there is a triangle with sides of length $\{s_1, s_2, d\}$ and a $(k-1)$-gon with sides of length $\{s_3, s_4, \ldots, s_k, d\}$. Without loss of generality, suppose that $s_1 \geq s_2$, and that $s_k \geq s_i$ for all $i \in [3, k-1]$. Then, by the inductive hypothesis, this is equivalent to finding d such that

$$s_1 - s_2 < d < s_1 + s_2$$

and

$$s_k - \left(s_3 + \cdots + s_{k-1}\right) < d < s_3 + \cdots + s_{k-1} + s_k.$$

The only way to have no such d would be for these intervals to be disjoint, meaning that either

$$s_1 + s_2 \leq s_k - (s_3 + \cdots + s_{k-1})$$

or

$$s_3 + \cdots + s_{k-1} + s_k \leq s_1 - s_2.$$

However, the first of these would contradict inequality (3) for $x = s_k$, and the second would contradict it for $x = s_1$. Thus the intervals must indeed overlap and so there must be such a length d.

Then, by the inductive hypothesis, there is a triangle with sides of length $\{s_1, s_2, d\}$ and a $(k-1)$-gon with sides of length $\{s_3, s_4, \ldots, s_k, d\}$. Gluing these along their sides of length d produces a k-gon with side lengths $\{s_1, \ldots, s_k\} = S$, as shown in Figure 5. ■

There are some interesting things to note about the result and proof of Proposition 2. First, a given multiset S need not produce a unique polygon. For example, the multiset $\{x, x, x, x\}$ describes the side lengths of infinitely many rhombi. Also, the construction in the proof of Proposition 2 can be used to produce convex k-gons.

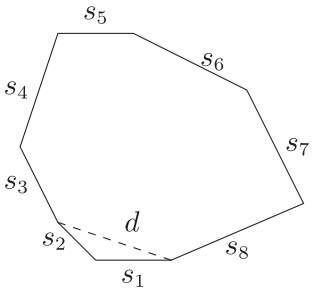


Figure 5 Constructing a k-gon, given inequalities on side lengths.

Broken bricks

Having characterized those multisets of positive numbers that can describe polygons, we now give our solution to the discrete analogue of the broken stick problem: the broken brick problem.

Theorem 3. *Let $n \geq k \geq 3$ be positive integers, and consider a stick of length n. Pick $k-1$ distinct interior integer points on the stick, independently and at random, and cut the stick at these $k-1$ points. The probability that the resulting k pieces form a k-gon is*

$$1 - \frac{k\binom{\lfloor n/2 \rfloor}{k-1}}{\binom{n-1}{k-1}}.$$

Proof. Fix values of n and k with $n \geq k \geq 3$. The sample space for the broken brick problem is the set of integer points

$$S = \{(x_1, \ldots, x_k) : x_i \in \mathbb{Z},\ x_i \geq 1,\ \text{and } \sum x_i = n\},$$

with the same interpretation as before. We will call each element of S, which represents a collection of the segments obtained from the full stick, an *inventory*.

In combinatorics, vectors satisfying the rules of membership in S are called *compositions* of n into k *parts*. The number of compositions of n into k parts is $\binom{n-1}{k-1}$, as shown by the following argument. Each x_i is positive, so

$$1 \leq x_1 < x_1 + x_2 < \cdots < x_1 + x_2 + \cdots + x_{k-1} < x_1 + x_2 + \cdots + x_k = n.$$

By setting $a_1 := x_1$, $a_2 := x_1 + x_2$, and so on up to

$$a_{k-1} := x_1 + x_2 + \cdots + x_{k-1},$$

the composition $(x_1, \ldots, x_k)$ determines (and is uniquely determined by) a set of integers $(a_1, \ldots, a_{k-1})$ satisfying

$$1 \leq a_1 < a_2 < \cdots < a_{k-1} < n.$$

Because the a_i are $k-1$ increasing integers between 1 and $n-1$, the number of such vectors is $\binom{n-1}{k-1}$.

Having established that $|S| = \binom{n-1}{k-1}$, we wish to show that the set of inventories that do *not* form a k-gon has cardinality $k\binom{\lfloor n/2 \rfloor}{k-1}$. To this end, we define sets containing the

"bad" points in S; that is, the inventories that do not describe k-gons. For $1 \le i \le k$, let

$$S_i := \{(x_1, \ldots, x_k) \in S : x_i \ge n/2\},$$

which is the set of all inventories that violate the k-gon inequality because the segment length x_i is longer than the sum of the other lengths. There can be at most one such segment in any inventory, so the sets $S_1, S_2, \ldots, S_k$ are pairwise disjoint. By Proposition 2, any bad point is contained in some S_i, and by symmetry they each have the same cardinality. Thus, the number of bad inventories is

$$|S_1 \cup S_2 \cup \cdots \cup S_k| = \sum_{i=1}^{k} |S_i| = k|S_1|.$$

It remains to prove that $|S_1| = \binom{\lfloor n/2 \rfloor}{k-1}$, which, again, comes down to counting compositions! Consider a point $(x_1, \ldots, x_k) \in S_1$. Thus $x_1 + \cdots + x_k = n$ and $x_1 \ge n/2$. Define

$$x_1' := x_1 - \lceil n/2 \rceil + 1 \ge 1.$$

Then the vector $(x_1', x_2, \ldots, x_k)$ has

$$x_1' + x_2 + \cdots + x_k = n - \lceil n/2 \rceil + 1 = \lfloor n/2 \rfloor + 1.$$

Therefore, inventories in S_1 are in bijection with compositions $(x_1', x_2, \ldots, x_k)$ of $\lfloor n/2 \rfloor + 1$ into k parts. There are $\binom{\lfloor n/2 \rfloor}{k-1}$ such compositions, which completes the proof. ∎

We can observe that as $n \to \infty$, the probability computed in Theorem 3 in the discrete setting approaches the probability computed in Theorem 1. We do this by noting that for large m and fixed j, we can approximate $\binom{m}{j}$ by the polynomial $\frac{1}{j!}m^j$:

$$\binom{m}{j} = \frac{1}{j!} m(m-1)\cdots(m-j+1) = \frac{1}{j!}\left(m^j + (\text{smaller powers of } m)\right).$$

Thus, for fixed k,

$$\lim_{n\to\infty} \frac{\binom{\lfloor n/2 \rfloor}{k-1}}{\binom{n-1}{k-1}} = \lim_{n\to\infty} \frac{\frac{1}{(k-1)!}\left((n/2)^{k-1} + (\text{smaller powers of } n/2)\right)}{\frac{1}{(k-1)!}\left((n-1)^{k-1} + (\text{smaller powers of } n-1)\right)} = \frac{1}{2^{k-1}},$$

and hence

$$1 - k \cdot \frac{\binom{\lfloor n/2 \rfloor}{k-1}}{\binom{n-1}{k-1}} \to 1 - \frac{k}{2^{k-1}},$$

which recovers Theorem 1.

This answer to the general broken stick problem tells us that the probability of forming a k-gon increases to 1 rapidly as k increases. This should match our intuition: a stick broken into a billion pieces that form a polygon should be akin to forming a circle from a pile of dust.

Pick-up sticks

The first author found D'Andrea and Gómez's result so delightful that, for a period of time, he told it to anyone who would listen. To grab a listener's attention, he would focus on the case $k = 4$, where the probability is $1 - (4/8) = 1/2$. Interestingly, most people found this probability to be extremely counter-intuitive, guessing that the probability would be much higher than $1/2$.

When challenged, the first author wrote a computer simulation to convince a skeptic. Disturbingly, the simulation recorded a success rate of 83%! As it turned out, rather than coding the "broken brick problem," he had instead implemented a simulation of the following problem, stated here in both continuous and discrete versions.

> **The pick-up sticks problem.** Let $k \geq 3$ be a positive integer. Select k sticks of lengths chosen from a bounded, uniform distribution of stick lengths. What is the probability that the resulting k sticks form a k-gon?
>
> **The pick-up bricks problem.** Let $n \geq 1$, $k \geq 3$ be positive integers. Select k sticks, each of which has length chosen from the uniform distribution on $\{1, 2, \ldots, n\}$. What is the probability that the resulting k sticks form a k-gon?

What the first author had naively assumed was that

> "pick up four random sticks"

was basically the same as

> "break a random stick into four random pieces,"

but they are not the same at all! Indeed, the solution to the pick-up bricks problem is as follows.

Theorem 4. *Let $n \geq 1$, $k \geq 3$ be positive integers. Pick k distinct sticks, each of which has length chosen from the uniform distribution on $\{1, 2, \ldots, n\}$. The probability that the resulting k sticks form a k-gon is*

$$1 - \frac{k\binom{n+1}{k}}{n^k}.$$

Proof. The sample space for the pick-up bricks problem is

$$[1, n]^k = \{(x_1, \ldots, x_k) : x_i \in \mathbb{Z} \text{ and } 1 \leq x_i \leq n \text{ for all } i\},$$

which clearly has cardinality n^k. As in the proof of Theorem 3, we will count the points in the sample space that do *not* form a k-gon, showing that there are $k\binom{n+1}{k}$ such "bad" points.

As in the earlier proof, denote the (disjoint) sets of bad points by

$$S_i := \{(x_1, \ldots, x_k) \in S : x_i \geq (x_1 + \cdots + x_k) - x_i\}.$$

Again, the total number of bad inventories is

$$|S_1 \cup S_2 \cup \cdots \cup S_k| = \sum_{i=1}^{k} |S_i| = k|S_1|,$$

and it remains to prove that $|S_1| = \binom{n+1}{k}$.

Let $(x_1, \ldots, x_k) \in S_1$, meaning that $n \geq x_1 \geq x_2 + \cdots + x_k$. To count such points, we make the following change of variables:

$$\begin{aligned} y_1 &:= x_1 + 1, \\ y_2 &:= x_2 + x_3 + x_4 + \cdots + x_k, \\ y_3 &:= x_3 + x_4 + \cdots + x_k, \\ &\vdots \\ y_k &= x_k. \end{aligned}$$

That is, $y_1 = x_1 + 1$ and $y_j = \sum_{i=j}^{k} x_i$ for $j \geq 2$. Since each x_i is positive, this gives a bijection between the points $(x_1, \ldots, x_k) \in S_1$ and the set of integer points $(y_1, \ldots, y_k)$ satisfying:

$$n + 1 \geq y_1 > y_2 > \cdots > y_k \geq 1.$$

That is, the y_i are simply k distinct numbers between 1 and $n + 1$, and thus there are $\binom{n+1}{k}$ such sets $\{y_1, \ldots, y_k\}$. This shows that $|S_1| = \binom{n+1}{k}$, and the theorem follows. ■

We can obtain the solution to the continuous pick-up sticks problem by considering the limit as $n \to \infty$ of the previous result. Using the same estimate for binomial coefficients as before, we have

$$\lim_{n\to\infty} \frac{k\binom{n+1}{k}}{n^k} = \frac{k}{k!} \cdot \lim_{n\to\infty} \frac{(n+1)^k + (\text{smaller powers of } n+1)}{n^k} = \frac{1}{(k-1)!},$$

which gives us the following result:

Corollary 5. Let $k \geq 3$ be a positive integer. Select k sticks of lengths chosen from a uniform distribution of stick lengths. The probability that the resulting k sticks form a k-gon is

$$1 - \frac{1}{(k-1)!}.$$

In light of this result, the 83% coming from the computer simulation of the pick-up sticks problem with $k = 4$ was approximating $1 - 1/3! = 5/6$. (Whew!)

Given the elegance of the answer appearing in Corollary 5, the reader may wonder if there is a geometric proof of the pick-up sticks problem, and indeed there is. Since each stick length is chosen independently, the sample space is now a cube. The volume of a region in the cube for which one of the sticks violates the k-gon inequality, $x_i \geq \|S\| - x_i$, can be shown to be $1/k!$ of the volume of the whole cube. As there are k such regions and these regions are disjoint, the complementary volume is

$$1 - k \cdot \frac{1}{k!} = 1 - \frac{1}{(k-1)!}.$$

See Figure 6 for an illustration of the case $k = 3$.

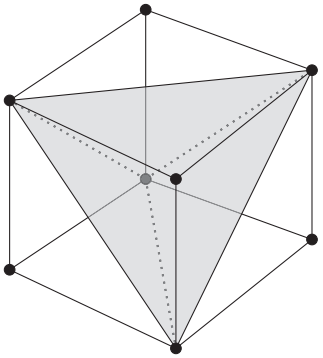


Figure 6 The geometric argument for the pick-up sticks problem with $k = 3$ sticks. The three corners missing from the cube are each $1/6$ of the volume of the entire cube. Thus the shaded region's volume is $1/2$ of the volume of the whole cube.

Four-gon conclusions

The mistake made when attempting to run a computer simulation of the broken brick problem was a happy accident because it led to another interesting, related question with a satisfying answer. The authors were then motivated to consider a whole host of related questions about combinations of picking up and breaking sticks.

Consider five different "four-gon" problems, each of which can be described in continuous and discrete settings.

- Stick(4): one stick breaks into four pieces.
- Stick(3,1): two sticks of random lengths, one of which breaks into three pieces.
- Stick(2,2): two sticks of random lengths, each of which breaks into two pieces.
- Stick(2,1,1): three sticks of random lengths, one of which breaks into two pieces.
- Stick(1,1,1,1): four sticks of random lengths.

The classical broken stick/brick problem is the scenario described by Stick(4). The pick-up sticks/bricks problem is Stick(1,1,1,1). What about the other three? Computer experiments suggest probabilities of forming a four-gon are, respectively, approximately 37%, 50%, and 61% in these scenarios. We invite interested readers to study these problems and find their own *four-gon conclusions.*

As a broader line of future inquiry, we remark that our labeling of the problems here generalizes to any integer partition. That is, for any integer partition

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 1$$

with $\sum \lambda_i = k$, we have:

The broken pick-up sticks problem. Stick (λ)**.** Consider a partition $\lambda = (\lambda_1, \ldots, \lambda_m)$ with

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 1,$$

with $\sum \lambda_i = k$. Pick up m sticks chosen from a uniform distribution of stick lengths. For each i, break the ith stick into λ_i pieces by choosing $\lambda_i - 1$ cut points independently at random. What is the probability that the resulting k pieces form a k-gon?

In this article, we have given answers to the problems corresponding to the extreme partitions (k) and $(1, 1, \ldots, 1)$. We encourage the reader to explain other interesting cases, and perhaps solve the problem in complete generality!

REFERENCES

[1] CrowdMath, P. A. (2018). The broken stick project. arXiv:1805.06512

[2] D'Andrea, C., Gómez E. (2006). The broken spaghetti noodle. *Amer. Math. Monthly*. 113: 555–557. doi.org/10.1080/00029890.2006.11920336

[3] Gardner, G. (2001). *The Colossal Book of Mathematics*. New York: W. W. Norton & Company.

[4] Goodman, G. S. (2008). The problem of the broken stick reconsidered. *Math. Intell*. 30: 43–49. doi.org/10.1007/BF02985378

[5] Ionaşcu, E. J., Prăjitură, G. (2013). Things to do with a broken stick. *Int. J. Geom.* 2: 5–30.

[6] Kong, L., Lkhamsuren, L., Turney, A., Uppal, A., Hildebrand, A. J. (2013). Random points, broken sticks, and triangles project report. https://faculty.math.illinois.edu/*sim*hildebr/ugresearch/brokenstick-spring2013report.pdf

[7] Lemoine, I. (1875). Sur une question de probabilités. *Bull. Sol. Math. France*. 1: 39–40.

[8] Poincaré, H. (1912). *Calcul des Probabilités*. Paris: Gauthier-Villars.

Summary. We generalize the well-known broken stick problem in several ways, including a discrete "brick" analogue and a sequential "pick-up sticks/bricks" version. The limit behavior of the broken brick problem gives a combinatorial proof of the broken stick problem. The pick-up version gives a variation on those scenarios, and we conclude by showing a greater context—namely, that the broken stick/brick problem and the pick-up sticks/bricks problem are two extremes in a family of interesting, and largely open, questions.

T. KYLE PETERSEN is a professor in the Department of Mathematical Sciences at DePaul University. He earned his Ph.D. in 2006 at Brandeis University under the supervision of Ira Gessel. After three years at the University of Michigan, he joined DePaul in 2009. His research is in combinatorics and related topics. He has three children and thousands of LEGO bricks.

BRIDGET EILEEN TENNER is a professor in the Department of Mathematical Sciences at DePaul University. She received her A.B. and A.M. degrees from Harvard University, and her Ph.D. from MIT under the supervision of Richard Stanley. She has been at DePaul University since 2007. Professor Tenner is a combinatorialist, focusing on enumerative, algebraic, and topological combinatorics. She has three children and thousands of LEGO bricks.

Complex Exponents and Group Theory

ALAN F. BEARDON
The University of Cambridge
Cambridge CB2 1TN, United Kingdom
afb@dpmms.cam.ac.uk

There are many publications (especially electronic ones) that introduce students to the meaning of z^w, where z and w are complex numbers. In many of these we are told that z^w (i.e., z "raised to the power" w) is a *set* of complex numbers, but, unfortunately, this plainly contradicts both our knowledge that $3^2 = 9$ (because 9 is not a set of numbers), and Euler's famous formula $e^{i\pi} = -1$ (because this implies that $e^{i\pi}$ is the single number -1). On the other hand, it is hard to understand the meaning of, for example, i^i in any way other than as a set of complex numbers. This expository note arose out of a desire to resolve this issue which, of course, is simply a matter of definition and notation. It is clear that what is needed are *different* definitions and notation for those situations in which z^w is a single complex number, and for those in which z^w is a set of complex numbers. Briefly, we shall use the notation z^w when this is a complex number, and use $[z]^w$ when this is a set. A good notation facilitates arguments, and once these terms are defined they should be rigorously adhered to. As informality is the cause of the problem, it must be abandoned immediately.

The two different definitions, and the associated notations, already exist in the language of group theory, although it seems that this is rarely (if ever) used to describe z^w in this way. For example, if we regard $1^{1/4}$ as a set, this set is $\{1, i, -1, -i\}$, which is a multiplicative group. Likewise, $16^{1/4}$, as a set, is the coset $2\{1, i, -1, -i\}$. More generally, the set Ω_n of nth roots of unity is a multiplicative group, and the set of solutions of $z^n = w$ is a coset of Ω_n. Despite this familiar example, it does not seem to be widely recognized that the same is true in much greater generality, and that this resolves the difficulties described above. Now once we have accepted that z^w can be a set, we must also accept that 3^2 can be a set too; in fact, in our new notation, $3^2 = 9$ and $[3]^2 = \{9\}$. This is no problem, for if g is an element of a group G, then g^2 is an element of G, and the singleton set $\{g^2\}$ is just the coset $g^2\{e\}$ with respect to the trivial subgroup $\{e\}$ of G. Thus, we will see that in *some* (but not all) circumstances, z^w can be defined (by various algebraic and analytic processes) as a single complex number, whereas *in all circumstances*, z^w arises as a coset (and therefore as a set) of some subgroup of a given group. What follows is an exposition of the two different interpretations of z^w from the perspective of group theory. From a teaching perspective, these ideas provide an unusual and elementary application of group theory to real and complex analysis.

In the next section, we briefly review the (small amount of) group theory that we shall use. Informally, z^w is $\exp(w \log z)$, so any progress necessarily depends on a clear understanding of the complex logarithm. For this reason, we will follow the group theory with a brief discussion of the complex exponential function, and the complex logarithm. We then discuss, and illustrate, the two possible interpretations of z^w.

Group theory

Throughout, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ are the additive groups of integers, real numbers, and complex numbers, respectively, while $\mathbb{R}^+$, $\mathbb{R}^*$, and $\mathbb{C}^*$ are the multiplicative groups of

Math. Mag. **93** (2020) 186–192. doi:10.1080/0025570X.2020.1736876
MSC: 30-01

positive real numbers, non-zero real numbers, and non-zero complex numbers, respectively. All of these groups are abelian, and these (and their subgroups) are the only groups that we shall use.

We need the idea of a coset. Let H be a subgroup of a group G, and define an equivalence relation $\sim$ on G by $f \sim g$ if and only if $fH = gH$. It is then easy to check that the equivalence classes are sets of the form $\{gh : h \in H\}$, which we write as gH. The set of cosets is denoted by G/H.

Now suppose that $\theta : G_1 \to G_2$ is a homomorphism from a group G_1 into a group G_2. The kernel K of θ is the set of g in G_1 that θ maps to the identity e_2 of G_2, and K is a subgroup of G_1. Under these circumstances, $gK = Kg$ for every g in G_1, and the set G/K of cosets forms a group in its own right with the operation

$$(fH, gH) \mapsto (fg)H.$$

Finally, take any w in G_2, and consider the set of solutions g in G_1 of the equation $\theta(g) = w$. If $w \notin \theta(G_1)$ then there is no solution. However, if $w \in \theta(G_1)$ then, by definition, there is some g_0 in G_1 such that $\theta(g_0) = w$, and it follows from this that *the set of all solutions of* $\theta(g) = w$ *is the coset* g_0K. This fact will be used repeatedly in what follows.

The exponential map

The exponential map $\exp : \mathbb{C} \to \mathbb{C}$ is defined for each complex z by the absolutely convergent power series

$$\exp z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots,$$

and for all complex numbers a and b, we have

$$\exp(a + b) = \exp(a) \exp(b). \tag{1}$$

This implies that $\exp(z) \exp(-z) \neq 0$, so that exp maps $\mathbb{C}$ into $\mathbb{C}^*$, and we now see that (1) simply expresses the fact that exp is a homomorphism of $\mathbb{C}$ into $\mathbb{C}^*$. The kernel of this homomorphism is $\{z : \exp(z) = 1\}$, and it is well known that this is the cyclic subgroup $2\pi i\mathbb{Z}$ of $\mathbb{C}$ that is generated by $2\pi i$.

The real logarithm

If x is real then

$$d(\exp x)/dx = \exp(x) = \left(\exp \tfrac{1}{2}x\right)^2 > 0,$$

so that $\exp x$ is a strictly increasing function of x. Clearly, $\exp x \to +\infty$ as $x \to +\infty$, and since $\exp(-x) = 1/\exp(x)$ we see that $\exp x \to 0$ as $x \to -\infty$. We conclude that the restriction of exp to $\mathbb{R}$ is an isomorphism of the group $\mathbb{R}$ onto the group $\mathbb{R}^+$. The inverse of this isomorphism is the *real logarithm function* $\ln : \mathbb{R}^+ \to \mathbb{R}$, and since (from algebra alone) this is necessarily also an isomorphism, we see that $\ln(a) + \ln(b) = \ln(ab)$ for all positive numbers a and b. No analytic proof of this formula is necessary. Note that we are using the notation $\ln(x)$ for the real logarithm of x in order to distinguish it from the complex logarithm $\log z$ of z which we will discuss later.

A complex power of a positive number

Later we shall discuss the definition of the complex logarithm as a set-valued function, but the introduction of the real logarithm ln as a (single-valued) function allows us to define a (single-valued) function t^w when t is positive and w is complex. This case is important in its own right for, as we shall see below, it arises in many branches of analysis.

Definition 1. If $t > 0$ and $w \in \mathbb{C}$, then $t^w = \exp\big(w \ln(t)\big)$.

This definition produces a complex-valued function $(t, w) \mapsto t^w$ on the space $(0, +\infty) \times \mathbb{C}$ which has the following properties:

(i) for all positive t, $t^0 = 1$;
(ii) for all complex w, $1^w = 1$;
(iii) for all positive t, and complex z and w, $t^{z+w} = t^z t^w$ (so $t^2 = t \times t$, etc.);
(iv) if $t > 0$ and n is a positive integer, then $t^{1/n}$ is the *positive* nth root of t;
(v) if $e = \exp(1)$ then, for all complex w, $e^w = \exp w$.

Note that (v) does need to be proved because the two terms e^w and $\exp w$ are defined in different ways. Also, if $w = \pi i$, then we do obtain Euler's result that

$$e^{i\pi} = \exp(i\pi) = \cos(\pi) + i \sin(\pi) = -1.$$

We now mention several situations in which Definition 1 is used in analysis. We shall not explore these situations any further, for our only objective here is to convince the reader of the importance of Definition 1. First, a Dirichlet series is a series of the form

$$\sum_{n=0}^{\infty} a_n e^{-\lambda_n z}, \tag{2}$$

where the a_n and z are complex numbers, and where

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n \to +\infty.$$

Generally speaking, (2) converges on a half-plane given by some inequality of the form $\mathrm{Re}(z) > \sigma$. If $\lambda_n = \ln(n)$, then (2) becomes $\sum_n a_n/n^z$, and with $a_n = 1$ this is the famous Riemann ζ-function $\sum_n 1/n^z$. If $\lambda_n = n$, then (2) reduces to a power series in e^{-z}.

Another occurrence of the function defined in Definition 1 is in the definition of the Gamma function. For each complex z, $\Gamma(z)$ is defined by

$$\Gamma(z) = \int_0^{\infty} t^{z-1} \exp(-t)\, dt.$$

Integration by parts shows that for each positive integer n,

$$\Gamma(n) = (n-1)!\, \Gamma(1).$$

Since $t^0 = 1$, we have

$$\Gamma(1) = \int_0^{\infty} t^0 \exp(-t)\, dt = \int_0^{\infty} \exp(-t)\, dt = 1, \tag{3}$$

so that, finally, $\Gamma(n) = (n-1)!$. Of course, this explains the otherwise mysterious definition that $0! = 1$.

A third example is the Laplace transform of a suitable function f. Given a function f defined on $[0, +\infty)$, the Laplace transform of f is the function F, of a complex variable s, given by

$$F(s) = \int_0^{+\infty} f(t)e^{-st}\,dt.$$

In many cases there is an inverse transform (i.e., we can recapture f from F), and this has many uses in differential equations and in many other applications of mathematics. Other examples in which the function t^z is used include the Beta-function

$$\mathrm{B}(z, w) = \int_0^1 t^{z-1}(1-t)^{w-1}\,dt, \tag{4}$$

where z and w are complex, the Fourier transform

$$\hat{f}(z) = \int_{-\infty}^{+\infty} f(t)e^{-2\pi itz}\,dt,$$

and the characteristic function of the probability distribution of a random variable X on $\mathbb{R}$ which is the expectation $E(e^{tX})$ of the random variable e^{tX}. In conclusion, the definition of t^z, where $t > 0$, plays an important role in analysis, and its existence as a single-valued function is based on the existence of the (single-valued) real logarithm function ln.

The complex logarithm

Intuitively, the logarithm of a complex number z is the "many-valued" inverse of the exponential map, but since (by definition) a function cannot be "many-valued," we must define it as a *set-valued* function.

Definition 2. Suppose that $z \in \mathbb{C}$. Then the complex logarithm of z is the set $\mathcal{L}(z)$, where

$$\mathcal{L}(z) = \{u \in \mathbb{C} \colon \exp(u) = z\}.$$

Observe that, according to this definition (and contrary to many accounts), the complex logarithm $\mathcal{L}(0)$ of 0 *is* defined, albeit as the empty set. Also, if $x > 0$ then $\ln(x)$ is just one of the elements of the set $\mathcal{L}(x)$. Although an interpretation of the complex logarithm through the geometry of the complex plane is interesting, and important, it is worth noting that we can also obtain information from group theory alone. Indeed, since exp maps $\mathbb{C}$ onto $\mathbb{C}^*$, and since $\mathcal{L}(w)$ is the solution set of $\exp(z) = w$, we see that if $w \neq 0$, then $\mathcal{L}(w)$ is a coset with respect to the kernel $2\pi i\mathbb{Z}$ of the homomorphism exp. Thus, from group theory alone, and without any mention of the argument of w, we see that

$$\mathcal{L}(w) = \{u + 2n\pi i : n \in \mathbb{Z}\},$$

where u is any complex number such that $\exp(u) = w$. Of course, if $r > 0$ and φ is real, then (1) implies that

$$\exp\left(\ln(r) + i\varphi\right) = \exp\left(\ln(r)\right)\,\exp(i\varphi) = r\exp(i\varphi),$$

so that $\ln(r) + i\varphi$ is one of the values in $\mathcal{L}\left(re^{i\varphi}\right)$. This group-theoretic approach to the complex logarithm is consistent with the usual view that is derived from the polar

form of complex numbers, but it avoids the seemingly self-contradictory idea of the "many-valued" function $\arg z$.

For subsets A and B of $\mathbb{C}$, we define $A + B$ to be

$$\{a + b\colon a \in A,\ b \in B\},$$

and we leave the reader to prove that for any non-zero complex numbers w_1 and w_2, we have

$$\mathcal{L}(w_1 w_2) = \mathcal{L}(w_1) + \mathcal{L}(w_2).$$

We emphasize that this does *not* say that two numbers are equal; it says that two sets are equal!

Before moving on, we should at least mention the popular idea of a principal value $\mathrm{Log}(z)$ of the complex logarithm, even though we have no use for it here. This function is defined on the "cut" complex plane $\mathbb{C}\backslash(-\infty, 0]$; that is, the complex plane $\mathbb{C}$ with the interval $(-\infty, 0]$ removed.

Definition 3. Suppose that $z \in \mathbb{C}\backslash(-\infty, 0]$. Then $z = |z|e^{i\theta}$ for some unique θ in $(-\pi, \pi)$, and the *principal branch of the logarithm* is the complex-valued function Log, where

$$\mathrm{Log}(z) = \ln(|z|) + i\theta.$$

Of course, if $x > 0$, then

$$\mathrm{Log}(x) = \ln(x).$$

In fact, we can now define a (single-valued) function z^w, where $z \in \mathbb{C}\backslash(-\infty, 0]$, by $\exp(w\mathrm{Log}(z))$, and this agrees with the earlier definition of z^w, $z > 0$.

A serious disadvantage of the function Log is that it is not defined at negative numbers (and there is no reason at all to denigrate negative numbers in this way). However, there is a simple, and useful, formula for $\arg z$ when $z \in \mathbb{C}\backslash(-\infty, 0]$ which is, perhaps, worth repeating here. If $z \in \mathbb{C}\backslash(-\infty, 0]$, and

$$z = x + iy = |z|e^{i\theta},$$

where $|\theta| < \pi$, then $|z| + x \neq 0$, so that

$$\frac{y}{|z| + x} = \frac{\sin\theta}{1 + \cos\theta} = \tan\tfrac{1}{2}\theta. \tag{5}$$

It follows that if

$$\tan^{-1}\colon \mathbb{R} \to (-\pi/2, \pi/2),$$

then

$$\theta = 2\tan^{-1}\left(\frac{y}{|z| + x}\right). \tag{6}$$

This is important because (given that $\tan^{-1}$ is a continuous function) it shows that θ is a *single-valued, continuous choice* of the argument of z on $\mathbb{C}\backslash(-\infty, 0]$, and many discussions in complex analysis depend on the existence of such a choice of $\arg z$.

Complex exponents z^w

We come now to the heart of this paper, and we shall use the notation $[z]^w$ for our interpretation of z^w as a set-valued function.

Definition 4. Given complex numbers z and w, we define the set $[z]^w$ by

$$[z]^w = \exp\big(w\mathcal{L}(z)\big) = \{\exp(w\lambda)\colon \lambda \in \mathcal{L}(z)\}. \tag{7}$$

We can immediately check that we now have a notation that genuinely distinguishes between numbers and sets; for example,

$$4^{1/2} = 2 \quad \text{and} \quad [4]^{1/2} = \{-2, 2\}$$

because

$$4^{1/2} = \exp\left(\tfrac{1}{2}\ln(4)\right) = \exp(\ln(2)) = 2;$$
$$[4]^{1/2} = \{\exp(\lambda/2)\colon \lambda = \ln(4) + 2m\pi i, m \in \mathbb{Z}\} = \{-2, 2\}.$$

The reader can now check that the function $[z]^w$ has the following properties:

(i) $[0]^w = \emptyset$ (because $\operatorname{Log}(0) = \emptyset$);
(ii) if $z \neq 0$ then $[z]^0 = [1]$;
(iii) if n is a positive integer, then $[z]^n$ is the singleton set $\{z^n\}$;
(iv) if $t > 0$, then t^w is one of the values in the set $[t]^w$.

For example, (iii) holds because if n is a positive integer, and $\exp(\eta) = z$, then

$$[z]^n = \{\exp(n[\eta + 2m\pi i])\colon m \in \mathbb{Z}\}$$
$$= \{\exp(n\eta)\} = \{(\exp\eta)^n\} = \{z^n\}.$$

At this point, to test the reader's understanding, we ask the reader to confirm that

$$(9/4)^{27/8} = (27/8)^{9/4},$$

but

$$[9/4]^{27/8} \neq [27/8]^{9/4},$$

the latter because the two sets do not have the same number of elements.

Our last result generalizes the fact that the set of solutions of an equation $z^n = w$ is a coset of the group of solutions of the equation $z^n = 1$.

Theorem 1. *Let w be any complex number. Then $[1]^w$ is a cyclic subgroup of $\mathbb{C}^*$ and, for any non-zero complex number z, $[z]^w$ is a coset with respect to the subgroup $[1]^w$. Moreover, the composition in the group $\mathbb{C}^*/[1]^w$ is given by*

$$([u]^w, [v]^w) \mapsto [uv]^w.$$

Proof. Let $\mu = \exp(2\pi i w)$. Then

$$[1]^w = \{\exp(w\zeta)\colon \exp\zeta = 1\}$$
$$= \{\exp(2\pi i n w)\colon n \in \mathbb{Z}\}$$
$$= \{\mu^n\colon n \in \mathbb{Z}\}.$$

Thus, $[1]^w$ is indeed the cyclic subgroup of $\mathbb{C}^*$ generated by μ.

Now suppose that $z \neq 0$, and choose some value, say ρ, in $\mathcal{L}(z)$. Then

$$\begin{aligned}[z]^w &= \{\exp(w[\rho + 2n\pi i]) : n \in \mathbb{Z}\} \\ &= \exp(w\rho)\,[1]^w,\end{aligned}$$

and this shows that $[z]^w$ is a coset of the group $[1]^w$ or, equivalently, that $[z]^w \in \mathbb{C}^*/[1]^w$. Finally, suppose that $\rho_j \in \mathcal{L}(z_j)$ for $j = 1, 2$. Then

$$[z_j]^w = \exp\left(w\rho_j\right)[1]^w,$$

so the group operation in the quotient group $\mathbb{C}^*/[1]^w$ is the operation $\star$, where

$$\begin{aligned}[z_1]^w \star [z_2]^w &= \exp(w\rho_1)\ [1]^w \star \exp(w\rho_2)\ [1]^w \\ &= (\exp(w\rho_1) \times \exp(w\rho_2))\ [1]^w \\ &= \exp(w[\rho_1 + \rho_2])\,[1]^w \\ &= [z_1 z_2]^w\end{aligned}$$

because $\rho_1 + \rho_2 \in \mathcal{L}(z_1 z_2)$. ■

Finally, the reader may like to verify the following statements:

(i) $[1]^w$ is the trivial group if and only if $w \in \mathbb{Z}$;
(ii) $[1]^w$ is a finite group if and only if $w \in \mathbb{Q}$;
(iii) $[1]^w \subset \{z : |z| = 1\}$ if and only if w is real;
(iv) the coset $[i]^i$ is a set of *real* numbers.

Summary. There is much confusion in the literature about the meaning of z^w, where z and w are complex numbers with $z \neq 0$. For example, we are often told that $e^{i\pi} = -1$ (a number), but that i^i is an infinite set. So which of these is true? Is z^w a set or a number? It cannot be both, and here we suggest how the inconsistent use of the universal notation z^w can be clarified and resolved by using elementary group theory.

ALAN F. BEARDON received his Ph.D. in 1964 after studying at Harvard University and Imperial College, London. He has taught at the University of Maryland and, in the UK, at the Universities of Kent at Canterbury and Cambridge. Since his retirement in 2007 he has taught regularly at the African Institute of Mathematical Sciences in Muizenberg, South Africa.

Surprises in Knockout Tournaments

TRAVIS KULHANEK
Univ. of California, Los Angeles
Los Angeles, CA 90095
travisk4747@ucla.edu

VADIM PONOMARENKO
San Diego State University
San Diego, CA 92182
vponomarenko@sdsu.edu

The NCAA Division I basketball tournament has 64 teams, which participate in $6 = \log_2 64$ rounds, with no ties permitted and the loser of each game eliminated. This is one well-known example of a knockout or single-elimination tournament. Other examples are large tennis tournaments like Wimbledon and the NFL playoffs. Knockout tournaments are of substantial interest to mathematicians and other scientists (see, e.g., Adler et al. [1], Prince, Smith, and Geunes [5], and Williams [7]). These are distinct from round-robin tournaments, which are of even greater interest (see, e.g., Charon and Hudry [2], Hudry [3], and Volkmann [6]), but which we will not consider here.

The (knockout) tournament winner ends unbeaten, and the tournament runner-up loses only in the final match. The tournament rankings of the other participants are unclear, so we will focus only on the top two possible tournament outcomes. The overall top player, whom we shall call Frankie, is more likely to finish first than second, being favored in the final match regardless of opponent. Our interest lies in the second-best player, whom we shall call Skylar.

Intuition might suggest that Frankie ought to finish first, and Skylar ought to finish second. In particular, one might expect that Skylar should be more likely to finish second than first. Is this intuition reliable? We can test this: In the aforementioned basketball tournament over the sixteen years 2004–2019, the men's second seed* ended up winning four times, and finishing second twice. Of course, this is a small sample size. We call a situation where Skylar is more likely to finish first than second a *surprise*, and we will characterize such surprises under certain assumptions. It will turn out that surprises are actually very likely.

Our main assumption is that all other tournament participants, whom we shall call Player, are indistinguishable from each other. This assumption dramatically simplifies the problem and allows us to consider arbitrarily large tournaments without worrying about an ever growing pool of probabilities. In our calculations, we condition on Skylar getting to the finals and facing either Frankie or one of the Players. If facing Frankie, Skylar is favored to finish second (no surprise); if facing a Player, Skylar is favored to finish first (surprise). The relative likelihood of these two situations depends on the probabilities involved.

Thus, for us there are just three probabilities of interest. Set p to be the probability that Skylar beats Frankie. We assume this to be less than half; after all, Frankie is better than Skylar. Set q to be the probability that Skylar beats Player. Set r to be the

*Although NCAA basketball tournament brackets give 1–16 rankings in each of four regions, these are prepared by the selection committee out of a 1–64 ranking of all the teams. We use that overall ranking here.

Math. Mag. **93** (2020) 193–199. doi:10.1080/0025570X.2020.1736892
MSC: Primary 00A08, Secondary 60C05; 00A69
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

probability that Frankie beats Player. Our probabilities should satisfy

$$0 \le p < \frac{1}{2} < q \le r \le 1.$$

For convenience, set $p' = 1 - p, q' = 1 - q, r' = 1 - r$. Our tournament will have n rounds before the championship and $n + 1$ total rounds. We assume that there are no byes. That is, there are 2^{n+1} players.

The two most common ways of designing a knockout tournament are by seeding the players, and randomly. These two tournament designs might be very similar in some ways (see Marchand [4]), but for us the calculations will be quite different, and we will treat them separately.

Seeded knockout tournaments

If the tournament is seeded, then Frankie and Skylar can only meet in the final round. The probability that Skylar finishes first is

$$q^n r^n p + q^n (1 - r^n) q,$$

where the first term corresponds to facing Frankie in the finals, and the second corresponds to Frankie losing before the finals. The probability that Skylar finishes second is

$$q^n r^n p' + q^n (1 - r^n) q'.$$

The surprise happens when the difference is positive, that is, when

$$q^n (r^n p + (1 - r^n) q - r^n p' - (1 - r^n) q') > 0.$$

We cancel the positive q^n, and note that

$$p - p' = 2p - 1 < 0$$

while

$$q - q' = 2q - 1 > 0.$$

We then combine terms to get

$$r^n (2p - 1) + (1 - r^n)(2q - 1) > 0,$$

and rearrange to get

$$(2q - 1) - 2r^n (q - p) > 0.$$

Hence, the surprise happens in a seeded tournament exactly when the following condition holds:

$$2r^n < \frac{2q - 1}{q - p} = 2 - \frac{1 - 2p}{q - p}. \tag{S}$$

Condition (S) has some interesting properties:

1. Decreasing r will decrease the left-hand side, leaving the right-hand side fixed. Hence, if Condition (S) holds, it will still hold after decreasing r.
2. Increasing n will similarly decrease the left-hand side, leaving the right-hand side fixed. Hence, if Condition (S) holds, it will still hold after increasing n.

3. For every p, q, r, so long as $r < 1$, there is some minimum n for which Condition (S) holds for that and all larger n.
4. On the other hand, for every p, q, n, if $r = 1$ then Condition (S) will never hold. Frankie will always be in the finals and be favored.
5. Increasing p will decrease $q - p$, and hence increase $\frac{2q-1}{q-p}$, leaving the left-hand side fixed. Hence, if Condition (S) holds, it will still hold after increasing p.
6. Taking $p = 0$, Condition (S) simplifies to $2r^n < 2 - \frac{1}{q}$. Hence, if $2r^n < 2 - \frac{1}{q}$, Condition (S) holds for all p.
7. Increasing q will increase $q - p$, decrease $\frac{1-2p}{q-p}$, and hence increase the right-hand side, leaving the left-hand side fixed. Hence, if Condition (S) holds, it will still hold after increasing q.
8. Fixing p, n and taking the limit as q, r approach $\frac{1}{2}$, the left-hand side approaches a positive constant while the right-hand side approaches 0. Hence, Condition (S) will not hold in the limit as $q, r \to \frac{1}{2}$.

For the special case $q = r$, corresponding to Skylar and Frankie having equal advantage over the Players, we can plot the surprise curves for various p, q, n, as shown in Figure 1. The area above each curve is the surprise region, where Condition (S) holds. Each curve passes through $(0.5, 0.5)$ and $(1, 0.5)$ because no surprise can happen for those values of r. Note that, curiously, increasing q might exit the surprise region—Skylar plays better, but is now less likely to finish first! This is because, in Figure 1, we are increasing r simultaneously with q.

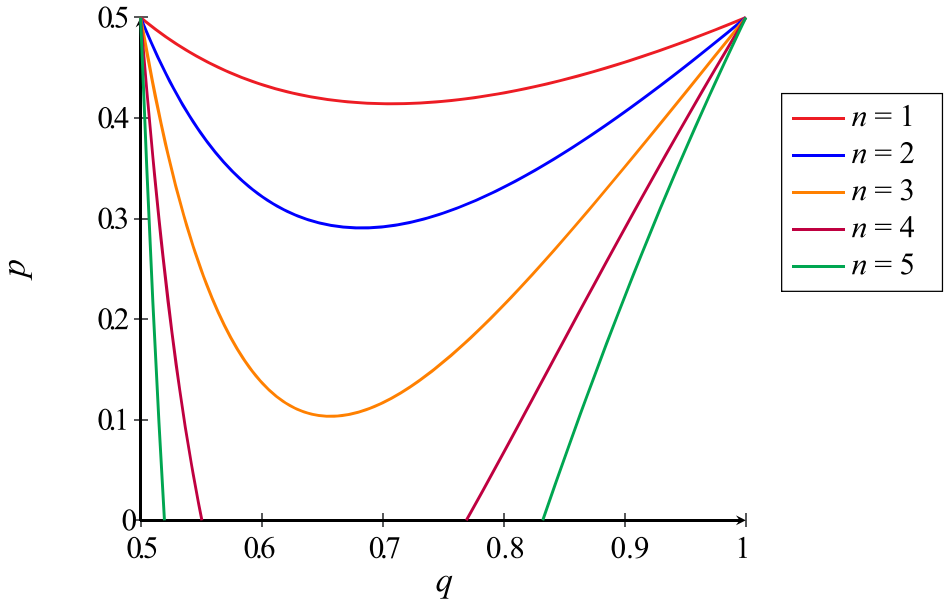


Figure 1 Surprise curves for the special case $q = r$, seeded tournament.

Note that for every n and every p, there will be a middle range of $q(= r)$ where the surprise holds. Outside this region the surprise will not hold.

For all n, p, with $q(= r)$ very close to 0.5, the surprise will not hold, as described in Property 8. The occasional times Skylar faces Frankie in the finals and is likely to lose outweighs the very slight advantage Skylar enjoys the rest of the time from facing Player in the finals.

For all n, p, with $q(=r)$ very close to 1, the surprise again will not hold. Both Frankie and Skylar are likely to make it to the finals, and thus Frankie's advantage over Skylar dominates the results.

Random knockout tournaments

For random knockout tournaments, the computation is more difficult. Frankie and Skylar might meet in the finals, or in round k, for $1 \le k \le n$. If they meet in round k, then Frankie must have won up to that point, in a field of 2^{k-1} players. If they meet in the finals, then Frankie must have already won in his half-tournament of 2^n players. However, these $n+1$ possibilities for when the players could meet are not equally likely. Set $m = 2^{n+1} - 1$. In a random tournament, the m positions other than Skylar's are all equally likely for Frankie. In 2^{k-1} of them, Frankie could meet Skylar in round k, and in 2^n of them, they could meet in the finals.

Hence, the probability that Skylar finishes first is

$$\sum_{k=1}^{n+1} \frac{2^{k-1}}{m} q^n r^{k-1} p + \sum_{k=1}^{n+1} \frac{2^{k-1}}{m} q^n (1 - r^{k-1}) q. \qquad \text{(Skylar\#1)}$$

Note that the first sum corresponds to Skylar beating Frankie at some point (including in the finals), while the second sum corresponds to Frankie losing before facing Skylar.

The probability that Skylar finishes second is

$$\frac{2^n}{m} q^n r^n p' + \sum_{k=1}^{n} \frac{2^{k-1}}{m} q^{n-1} r^{k-1} pq' + \sum_{k=1}^{n+1} \frac{2^{k-1}}{m} q^n (1 - r^{k-1}) q'. \qquad \text{(Skylar\#2)}$$

Here, the first term corresponds to Skylar losing to Frankie in the finals. The second term corresponds to Skylar beating Frankie in round k, but losing in the finals. The third term corresponds to Frankie losing prior to meeting Skylar and Skylar winning every round except the finals. The surprise happens when (Skylar#1) is greater than (Skylar#2). We multiply each by the positive $\frac{m}{q^{n-1}}$, and pull constants out of sums, to find the surprise equivalent to

$$qp \sum_{k=1}^{n+1} 2^{k-1} r^{k-1} + q^2 \sum_{k=1}^{n+1} 2^{k-1} (1 - r^{k-1}) > 2^n q r^n p' + pq' \sum_{k=1}^{n} 2^{k-1} r^{k-1} + qq' \sum_{k=1}^{n+1} 2^{k-1} (1 - r^{k-1}).$$

Each series is either geometric or the difference of two geometric series, and hence we can find their sums. After considerable rearrangement, we find the surprise happens in a random tournament exactly when the following condition holds:

$$2^{n+1} q(2q-1)(2r-1) + (2q-1)(2q - 2qr - p) > 2^n r^n (q-p)(4qr-1). \qquad \text{(R)}$$

Note that

$$2q - 1, 2r - 1, q - p, \quad \text{and} \quad 4qr - 1$$

are each positive. Considering the bounds on p, q, r, we find that

$$-\frac{1}{2} \leq 2q - 2qr - p \leq \frac{1}{2}.$$

Condition (R) shares some properties with Condition (S). For convenience, set

$$f(p,q,r,n) = 2^{n+1}q(2q-1)(2r-1) + (2q-1)(2q-2qr-p) - 2^n r^n (q-p)(4qr-1).$$

Condition (R) is equivalent to $f(p,q,r,n) > 0$. We calculate

$$f(p+\epsilon,q,r,n) - f(p,q,r,n) = -(2q-1)\epsilon + 2^n r^n (4qr-1)\epsilon \geq \epsilon(2q-1)\left(-1 + 2^n r^n \left(2r + \frac{2r-1}{2q-1}\right)\right) > 0.$$

Hence, if Condition (R) holds, it will still hold if we increase p. We calculate

$$\begin{aligned} f(p,q,r,n+1) - 2rf(p,q,r,n) &= 2^{n+1}q(2q-1)(2r-1)(2-2r) \\ &\quad + (1-2r)(2q-1)(2q-2qr-p) \\ &= (2r-1)(2q-1) \\ &\quad (q(1-r)(2^{n+2}-2) + p) \geq 0. \end{aligned}$$

Hence, if Condition (R) holds, it will still hold if we increase n. Further, if $r < 1$ then

$$\lim_{n\to\infty} \frac{f(p,q,r,n)}{2^n} = 2q(2q-1)(2r-1) > 0,$$

so there is some minimum n which will ensure that the surprise occurs for that and all larger n.

Hence, if the surprise holds for $p = 0, n = 1$, then it will hold independently of p, n. Unfortunately, no values of q, r meet this condition. Likewise for $p = 0, n = 2$. However, for $p = 0$ and for each $n \geq 3$, there is a region in the $q - r$ plane for which the surprise will hold independently of p. Note that each successive region includes all previous ones. These regions are plotted in Figure 2.

Unfortunately, varying q or r does not appear to respect Condition (R) in the same way as with Condition (S). We are able to prove that, considering each of these variables separately, the surprise will hold on a (possibly empty) interval.

The function $f(p,q,r,n)$, fixing p, r, n, is a quadratic polynomial in q, with leading coefficient

$$2^{n+2}(2r-1) + 4(1-r) - 2^{n+2}r^{n+1}.$$

This leading coefficient is positive for all $r \in (0.5, 1)$, so the parabola opens upward. Hence, $f(p,q,r,n) > 0$ will hold for all $q \in \mathbb{R}$, apart from some interval. This interval may intersect with (or contain all of) $(0.5, r]$.

Considering $f(p,q,r,n)$ as a function of r, we find that

$$\frac{\partial}{\partial r} f(p,q,r,n) = Ar^n + B,$$

for some real constants A, B. This has at most one positive zero. By the mean value theorem, $f(p,q,r,n)$ has at most one positive zero. Hence, the surprise will happen for r in some half-line intersected with $[q, 1]$.

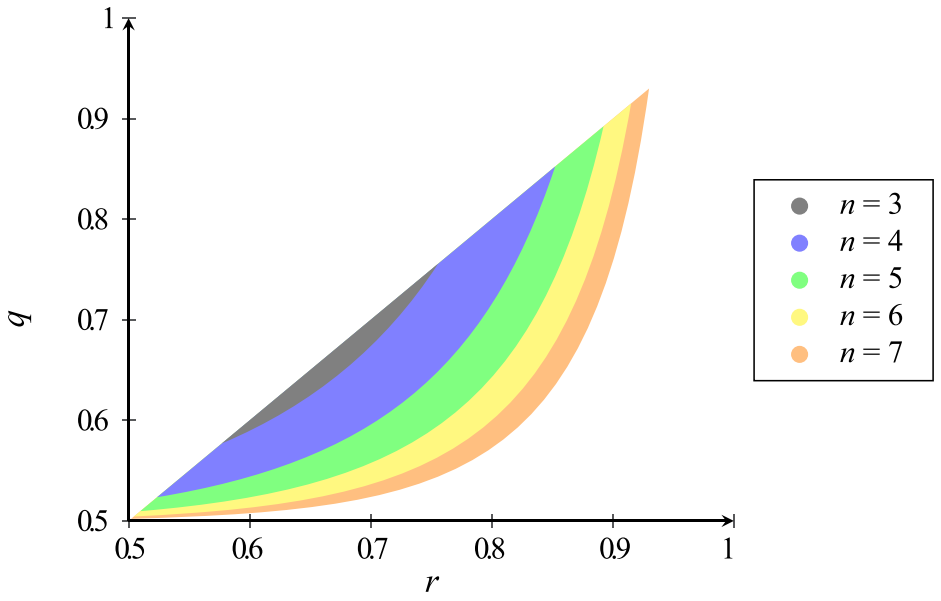


Figure 2 Surprise regions for $p = 0$, random tournament.

As before, we consider the special case of $q = r$. We have

$$\lim_{r\to\frac{1}{2}} \frac{f(p,q,r,n)}{2r-1} = \lim_{r\to\frac{1}{2}} 2^{n+1} r(2r-1) + 2r(1-r) - p - 2^n r^n (r-p)(2r+1)$$
$$= \frac{1}{2} - p - \left(\frac{1}{2} - p\right) \quad (2)$$

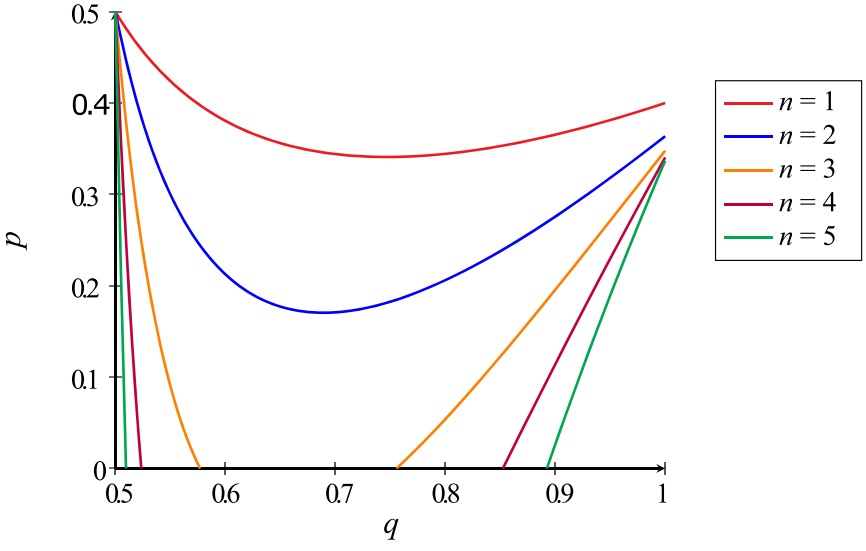


Figure 3 Surprise curves for the special case $q = r$, random tournament.

$$= -\frac{1}{2} + p < 0.$$

Hence, for q, r sufficiently close to $1/2$, Condition (R) fails to hold. On the other hand, if we take $q = r = 1$, then Condition (R) simplifies to

$$p > \frac{2^n}{3 \cdot 2^n - 1}.$$

Compare with the seeded tournament case, where for $r = 1$ the surprise is impossible. We plot the surprise curves in Figure 3.

We close by asking if a player even lower-ranked than second, might still be more likely to win the tournament than to finish second. Such a situation would be an even greater surprise.

REFERENCES

[1] Adler, I., Cao, Y., Karp, R., Peköz, E. A., Ross, S. M. (2017). Random knockout tournaments. *Oper. Res.* 65(6): 1589–1596. doi.org/10.1287/opre.2017.1657

[2] Charon, I., Hudry, O. (2010). An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Ann. Oper. Res.* 175(1): 107–158. doi.org/10.1007/s10479-009-0648-7

[3] Hudry, O. (2009). A survey on the complexity of tournament solutions. *Math. Soc. Sci.* 57(3): 292–303. doi.org/10.1016/j.mathsocsci.2008.12.002

[4] Marchand, E. (2002). On the comparison between standard and random knockout tournaments. *The Statistician*. 51(2): 169–178. doi.org/10.1111/1467-9884.00309

[5] Prince, M., Smith, J. C., Geunes, J. (2013). Designing fair 8- and 16-team knockout tournaments. *IMA J. Manag. Math.* 24(3): 321–336. doi.org/10.1093/imaman/dpr024

[6] Volkmann, L. (2007). Multipartite tournaments: a survey. *Discrete Math.* 307(24): 3097–3129. doi.org/10.1016/j.disc.2007.03.053

[7] Williams, V. V. (2016). Knockout tournaments. In: Brandt, F., Conitzer, V., Endriss, U., Lang, J., Procaccia, A. D., eds. *Handbook of Computational Social Choice*. New York: Cambridge Univ. Press, pp. 453–474.

Summary. Sometimes the second-best player in a knockout tournament is more likely to win that to finish second. We characterize this surprising situation, under certain natural conditions.

TRAVIS KULHANEK is an undergraduate student at the University of California, Los Angeles. He studies pure mathematics and plans to pursue higher education and become a math professor.

VADIM PONOMARENKO is a professor of mathematics at San Diego State University. He heads a long-running summer REU program, and has recently written an introduction-to-proofs/discrete math textbook.

Eigenvalues and Eigenvectors: Generalized and Determinant Free

JEFF SUZUKI
CUNY Brooklyn College
Brooklyn, NY 10001
jsuzuki@brooklyn.cuny.edu

Sheldon Axler's campaign [1] to rid elementary linear algebra of the determinant has a ready audience in anyone actually concerned with computational complexity. Only a sadist would require using Cramer's rule to solve systems of equations or the adjoint method to find inverses, and only a masochist would choose to use these methods. Fortunately, there are determinant-free alternatives to Cramer's rule for solving systems of equations, and to the adjoint method for finding inverses.

What about eigenvalues and eigenvectors, which seem to rely on the determinant to produce the characteristic polynomial? Axler [2] provides a determinant-free introduction to eigenvalues and eigenvectors, including a derivation of the characteristic polynomial. McWorter and Meyers [3] extend Axler's work and provides a determinant-free general approach to finding all eigenvalues and eigenvectors.

However, a perusal of the standard linear algebra texts shows this alternative approach nowhere to be found. We suspect the unpopularity of determinant-free methods is reflected by a student comment: "With the determinant, you don't have to think" [3, p. 24]. In particular, the method of McWorter and Meyers requires some heuristic decisions that require insight and intuition on the part of the user.

In the following, we build on the approach of McWorter and Meyers to produce a purely algorithmic and determinant-free solution to the eigenvalue-eigenvector problem. We show how all eigenvalues and eigenvectors can be found without using the determinant. Moreover, our approach allows us to introduce *generalized* eigenvectors in a natural fashion.

Preliminaries

The eigenvalue-eigenvector problem is typically introduced in the context of finding a nonzero vector $\mathbf{v}$ and a value λ where $A\mathbf{v} = \lambda\mathbf{v}$. Very quickly, this equation is transformed into $(A - \lambda I)\mathbf{v} = \mathbf{0}$, which has a nontrivial solution if and only if $|A - \lambda I| = 0$.

Leaving aside any philosophical or pedagogical objections to the determinant itself, there remains one unavoidable challenge: If A is an $n \times n$ matrix, the characteristic polynomial $|A - \lambda I|$ is necessarily an nth degree polynomial in λ, so in general it will be algebraically unsolvable. As a result, we often restrict ourselves to 2×2 matrices, very simple 3×3 matrices, or relegate the problem of finding eigenvalues and eigenvectors to a computer algebra system.

A determinant free introduction

Instead, consider a different approach. After the introduction of the eigenvalue-eigenvector problem, there are several obvious questions to ask before "How do

Math. Mag. **93** (2020) 200–212. doi:10.1080/0025570X.2020.1736874
MSC: 15-01, 97H60

we find eigenvalues and eigenvectors?" For example, we might ask whether a matrix could have more than one eigenvector, or more than one eigenvalue.

Once we have established the possibility that a matrix might have two or more distinct eigenvectors, the next natural question (in linear algebra, at least) is to ask whether they are linearly independent. It is relatively easy to prove:

Theorem 1. *Eigenvectors for distinct eigenvalues are linearly independent.*

Since the eigenvector for an $n \times n$ matrix will have n components, it follows that an $n \times n$ matrix will have at most n linearly independent eigenvectors.

This establishes an important requirement for any algorithm: it gives you a stopping point, since once you have found n linearly independent eigenvectors, you cannot find any more. At the same time, it establishes a new challenge: if you *do not* find n linearly independent eigenvectors, there is the possibility you have missed some.

So how do we find some (if not all) of the eigenvectors of A? The method of McWorter and Meyers begins as follows:

- Choose an arbitrary non-zero seed vector $\mathbf{u}$.
- Find the least value k for which the set

$$\mathcal{V}_k = \left\{\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^k\mathbf{u}\right\}$$

  is dependent. (Proving that such a k must exist is a good way to reinforce the ideas of basis and independence.)

Since k is the least value for which $\mathcal{V}_k$ is dependent, it follows that $A^k\mathbf{u}$ can be expressed as a linear combination of other vectors, so there is a nontrivial linear combination of the form

$$A^k\mathbf{u} + a_{k-1}A^{k-1}\mathbf{u} + \cdots + a_1A\mathbf{u} + a_0\mathbf{u} = \mathbf{0}.$$

We set

$$f(A) = A^k + a_{k-1}A^{k-1} + \cdots + a_1A + a_0I$$

to be the minimal polynomial of A with respect to $\mathbf{u}$. We note that $f(A)\mathbf{u} = \mathbf{0}$, and that the coefficient of A^k is equal to 1.

We claim:

Theorem 2. *The minimal polynomial of A with respect to* $\mathbf{u}$ *is unique.*

Proof. Assume A is an $n \times n$ matrix that acts on vectors in $\mathbb{F}^n$, where the components of A are drawn from the field $\mathbb{F}$.

By assumption, k is the least power for which

$$\mathcal{V}_k = \left\{\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^k\mathbf{u}\right\}$$

is dependent. Since these are vectors in $\mathbb{F}^n$, $k \leq n$.

Because the set is dependent,

$$x_{k+1}A^k\mathbf{u} + x_kA^{k-1}\mathbf{u} + x_{k-1}A^{k-2}\mathbf{u} + \cdots + x_2A\mathbf{u} + x_1\mathbf{u} = \mathbf{0}$$

has nontrivial solutions, which can be found by row reducing the matrix of column vectors

$$\begin{pmatrix}\mathbf{u} & A\mathbf{u} & A^2\mathbf{u} & \ldots & A^k\mathbf{u}\end{pmatrix}.$$

First, x_{k+1} must be a free variable. If it is not, then the $k+1$st column of the row echelon form of the coefficient matrix will have a non-zero coefficient. But since this is the last column of the coefficient matrix, it follows that $x_{k+1} = 0$. However, we have a nontrivial solution to the equation, which means that

$$x_k A^{k-1}\mathbf{u} + x_{k-1}A^{k-2}\mathbf{u} + \cdots + x_2 A\mathbf{u} + x_1\mathbf{u} = \mathbf{0}$$

for some $x_1, x_2, \ldots, x_k$ not all equal to zero. Consequently, $\mathcal{V}_{k-1}$ will be dependent, contradicting our assumption that k is the least $\mathcal{V}_k$ which is dependent.

Next, suppose there are other free variables. We can let $x_{k+1} = 0$, and as long as at least one of the other free variables is set to a non-zero value, we will obtain a non-trivial expression of the form

$$x_i A^{i-1}\mathbf{u} + x_{i-1}A^{i-2}\mathbf{u} + \cdots + x_2 A\mathbf{u} + x_1\mathbf{u} = \mathbf{0}$$

and so $\mathcal{V}_{i-1}$, with $i-1 < k$, will be a dependent set. Again, this contradicts our assumption that k is the least value for which $\mathcal{V}_k$ is dependent.

Thus, x_{k+1} is the only free variable in our system. Setting $x_{k+1} = 1$ gives us a unique solution to

$$A^k\mathbf{u} + x_k A^{k-1}\mathbf{u} + x_{k-1}A^{k-2}\mathbf{u} + \cdots + x_2 A\mathbf{u} + x_1\mathbf{u} = \mathbf{0}.$$

Consequently, the minimal polynomial of A with respect to $\mathbf{u}$ is unique. ■

Given the minimal polynomial, we can find a factorization. However, matrix multiplication is not in general commutative, necessitating the statement of

Lemma 1. *The factors of*

$$f(A) = (A - \lambda_1 I)(A - \lambda_2 I)\cdots(A - \lambda_k I)$$

commute.

We omit the proof, which is straightforward.

The significance of the minimal polynomial comes from the following result:

Theorem 3. *All roots of the minimal polynomial are eigenvalues.*

Proof. Let

$$f(A) = (A - \lambda_1 I)(A - \lambda_2 I)\cdots(A - \lambda_k I).$$

We know

$$(A - \lambda_1 I)(A - \lambda_2 I)(A - \lambda_3 I)\cdots(A - \lambda_k I)\mathbf{u} = \mathbf{0}.$$

If

$$(A - \lambda_2 I)(A - \lambda_3 I)\cdots(A - \lambda_k I)\mathbf{u} \neq \mathbf{0}$$

then it will be an eigenvector corresponding to eigenvalue λ_1. But by assumption, $f(A)$ is the minimal polynomial of A with respect to $\mathbf{u}$. Consequently, the polynomial produced by the factors of $f(A)$ excepting $(A - \lambda I)$ cannot, when applied to $\mathbf{u}$, be $\mathbf{0}$. It follows that λ_1 is an eigenvalue.

By Lemma 1, the factors of $f(A)$ may be rearranged in any order. Consequently, any factor $(A - \lambda_i I)$ of $f(A)$ will produce an eigenvalue λ_i, where the eigenvector is the product of the remaining factors applied to $\mathbf{u}$. ■

Example 1. *Find eigenvalues of*

$$A = \begin{pmatrix} 3 & 4 & 8 \\ 8 & 7 & 16 \\ -4 & -4 & -9 \end{pmatrix}.$$

We choose an arbitrary seed vector. There is no reason not to choose the simplest vector possible, so we pick $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.

Our theorem requires we find the least k for which

$$\mathcal{V}_k = \{\mathbf{u}, A\mathbf{u}, \ldots, A^k\mathbf{u}\}$$

is dependent. While we could find and check the independence of $\mathcal{V}_1, \mathcal{V}_2, \ldots$, in actual practice, we can proceed as follows: Since A acts on vectors in $\mathbb{R}^3$, we know that any set of four vectors is dependent, so we find $\mathcal{V}_3 = \{\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, A^3\mathbf{u}\}$. We compute

$$\mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 3 \\ 8 \\ -4 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 9 \\ 16 \\ -8 \end{pmatrix} \qquad A^3\mathbf{v} = \begin{pmatrix} 27 \\ 56 \\ -28 \end{pmatrix}$$

and seek nontrivial solutions to

$$x_1\mathbf{v} + x_2A\mathbf{v} + x_3A^2\mathbf{v} + x_4A^3\mathbf{v} = \mathbf{0}.$$

Setting $\mathbf{v}$, $A\mathbf{v}$, $A^2\mathbf{v}$, $A^3\mathbf{v}$ as column vectors and row reducing leads to

$$\begin{pmatrix} 1 & 3 & 9 & 27 \\ 0 & 8 & 16 & 56 \\ 0 & -4 & -8 & -28 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 3 & 6 \\ 0 & 1 & 2 & 7 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that if we let x_4 be a non-zero value, we will obtain a third degree polynomial. If $x_4 = 0$ but x_3 is nonzero, we will obtain a second degree polynomial. Since we are interested in a minimal polynomial, we will let $x_3 = 1$, $x_4 = 0$, which gives us $x_2 = -2$ and $x_1 = -3$ and corresponding vector equation $-3\mathbf{v} - 2A\mathbf{v} + A^2\mathbf{v} = \mathbf{0}$. Our minimal polynomial is $f(A) = A^2 - 2A - 3I$, and the roots $\lambda = 3$, $\lambda = -1$ are eigenvalues.

One advantage to this approach is clear from this example. If A is an $n \times n$ matrix, then the characteristic polynomial will always be of degree n. However, the minimal polynomial with respect to a given seed vector might have a lower degree. In fact, this will always occur, regardless of the seed vector, whenever any eigenvalue of A has a geometric multiplicity greater than 1. We leave the demonstration of this claim as an exercise for the reader.

Once we have the eigenvalues, how can we find the eigenvectors? There are two approaches. McWorter and Meyers use Theorem 3 to compute the eigenvectors directly. However, this is risky: if our eigenvalue has a geometric multiplicity greater than 1, computing the eigenvector using Theorem 3 will only give us one eigenvector, with no hint that there might be others.

Instead, we can proceed as we would if we had used the characteristic polynomial: once we know an eigenvalue, we can find and parameterize the nontrivial solutions to $(A - \lambda I)\mathbf{x} = \mathbf{0}$. In this case, we will find that $\lambda = 3$ corresponds to the eigenvector $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$, and $\lambda = -1$ corresponds to two linearly independent eigenvectors $\begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$

and $\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ Since we have found three linearly independent eigenvectors, there can be no others.

It should be clear that a randomly chosen seed vector might not give us all the eigenvalues. For example, if we chose $\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ as a seed vector, our minimal polynomial would be $f(A) = A - 3I$ (because this is in fact an eigenvector for $\lambda = 3$), and we would only find one eigenvalue.

So what happens if we find fewer eigenvectors than expected? The obvious solution is to pick a different seed vector. McWorter and Meyers proceed as follows: If k is the least value for which $\mathcal{V}_k = \{\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^k\mathbf{u}\}$ is dependent, then $\mathcal{V}_{k-1}$ is an independent set of vectors. Choose another vector $\mathbf{v}$ and find the least k' for which

$$\mathcal{V}_{k'} = \{\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^{k-1}\mathbf{u}, \mathbf{v}, A\mathbf{v}, \ldots, A^{k'}\mathbf{v}\}$$

is dependent. Consequently there is a nontrivial linear combination of these vectors that gives the zero vector.

Unfortunately, at this point the method requires the "thinking" that the student objected to, as recovering the eigenvalues and eigenvectors from the linear combination is nontrivial. In the following, we will present an alternative approach which will allow us to find all the eigenvectors. Moreover, our approach will provide a natural way to introduce (and find) generalized eigenvectors.

Generalized eigenvectors

Generalized eigenvalues and eigenvectors are ordinarily introduced in the context of finding an eigenbasis when you have a defective matrix. Their introduction usually goes something along the following lines.

Suppose A is an $n \times n$ matrix whose entries are from some field $\mathbb{F}$. The linearly independent eigenvectors of A form an eigenbasis, which spans a subspace of $\mathbb{F}^n$. If the eigenbasis is a basis for $\mathbb{F}^n$, then we say that A is a nondefective matrix.

What if the eigenbasis does not span all of $\mathbb{F}^n$? In that case, we can form a basis by including some additional vectors along with the linearly independent eigenvectors.

At this point, we usually introduce what appears to be an *ad hoc* requirement: the additional vectors $\mathbf{v}$ should satisfy $(A - \lambda I)^k\mathbf{v} = \mathbf{0}$, where λ is an eigenvalue and $(A - \lambda I)^p\mathbf{v} \neq \mathbf{0}$ for $p < k$. We say that $\mathbf{v}$ is a generalized eigenvector of rank k, and that

$$\mathbf{v}, (A - \lambda I)\mathbf{v}, (A - \lambda I)^2\mathbf{v}, \ldots, (A - \lambda I)^{k-1}\mathbf{v}$$

is a Jordan chain of generalized eigenvectors.

Except for having a certain pleasing symmetry, there is no obvious reason why we would flesh out an eigenbasis with generalized eigenvectors. Certainly, if asked to find additional vectors to form a basis, few would ever think to impose such requirements on what the additional vectors would look like.

In contrast, these expressions arise naturally from our approach.

Example 2. *Find eigenvalues and eigenvectors for* $A = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 2 & -2 \\ -8 & 5 & 1 \end{pmatrix}$.

Again, making the obvious choice for $\mathbf{u}$, we compute

$$\mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad A\mathbf{u} = \begin{pmatrix} 3 \\ -1 \\ -8 \end{pmatrix} \qquad A^2\mathbf{u} = \begin{pmatrix} 18 \\ 11 \\ -37 \end{pmatrix} \qquad A^3\mathbf{u} = \begin{pmatrix} 80 \\ 78 \\ -126 \end{pmatrix}.$$

Row reducing the column space leads to

$$\begin{pmatrix} 1 & 3 & 18 & 80 \\ 0 & -1 & 11 & 78 \\ 0 & -8 & -37 & -126 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 8 \\ 0 & 1 & 0 & -12 \\ 0 & 0 & 1 & 6 \end{pmatrix},$$

giving us solution

$$x_1 = -8, x_2 = 12, x_3 = -6, x_4 = 1,$$

and corresponding minimal polynomial

$$f(A) = A^3 - 6A^2 + 12A - 8 = (A - 2I)^3.$$

Because $f(A)$ is a minimal polynomial, we observe that $(A - 2I)^3\mathbf{u} = \mathbf{0}$, so the concept of a generalized eigenvector emerges naturally as a consequence of our approach.

Moreover, this gives us an direct method of finding the eigenvectors in a Jordan chain: they will be

$$\mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad (A - 2I)\mathbf{u} = \begin{pmatrix} 1 \\ -1 \\ -8 \end{pmatrix} \qquad (A - 2I)^2\mathbf{u} = \begin{pmatrix} 10 \\ 15 \\ -5 \end{pmatrix}.$$

Once this is established, it more naturally lends to the investigation of generalized eigenvectors. In particular, once the existence of a generalized eigenvector is established, it is easy to show that the vectors in a Jordan chain are independent. Indeed, the union of any number of Jordan chains for distinct eigenvalues are independent.

Seeds and seedlings

We are now in a position to find all eigenvectors and generalized eigenvectors for any $n \times n$ matrix. We proceed as follows:

- Choose an arbitrary seed vector and find the minimal polynomial.
- Use the roots of the minimal polynomial to find some of the eigenvalues.
- Use the eigenvalues to find eigenvectors and generalized eigenvectors.
- If we have found fewer than n eigenvectors and generalized eigenvectors, choose another seed vector and repeat.

However, there is an important requirement: Our second and subsequent seeds need to be independent of the known eigenvectors. So how can we choose such a vector?

Consider the subspace spanned by the eigenvectors and generalized eigenvectors we have already found. If we take these known eigenvectors as a partial basis for the whole space, then any vector in the space can be expressed as a linear combination of the known eigenvectors and generalized eigenvectors, plus a linear combination of some other unknown vectors.

Clearly:

Lemma 2. *Suppose $f(x)$ has roots $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, where λ_i has multiplicity r_i. Let $\mathcal{V}$ be a set of linearly independent eigenvectors and generalized eigenvectors corresponding to the eigenvalues in Λ, where the generalized eigenvectors have rank less than or equal to the corresponding r_i. If $\mathbf{x}$ is in the span of $\mathcal{V}$, then $f(A)\mathbf{x} = \mathbf{0}$.*

This follows from the commutativity of the factors of $f(A)$ and the linearity of matrix multiplication.

The significance of Lemma 2 is this: If $\mathbf{x}$ is a linear combination of the known eigenvectors and generalized eigenvectors, plus a linear combination of some as-yet-undetermined basis vectors, $f(A)$ applied to $\mathbf{x}$ will "zero out" the components corresponding to the known eigenvectors and generalized eigenvectors. Consequently $f(A)\mathbf{x}$ is a linear combination of the as-yet-undetermined basis vectors only.

This suggests the following: Our original seed vector $\mathbf{v}$ gave us a minimal polynomial $f(A)$. Choose any other seed vector $\mathbf{x}$. By Lemma 2, $\mathbf{v} = f(A)\mathbf{x}$ will be a vector that can be written entirely in terms of vectors we *do not* already know. In effect, we preprocess $\mathbf{x}$ to produce an improved seed vector $\mathbf{v} = f(A)\mathbf{x}$. We shall extend the botanical analogy and call $\mathbf{v}$ a "seedling" vector.

Using $\mathbf{v}$ as our seedling, we will be able to find a minimal polynomial $g(A)$ for which $g(A)\mathbf{v} = \mathbf{0}$, and Theorem 3 guarantees every root of $g(A)$ will be an eigenvalue. Moreover, growing the seed into a seedling further guarantees that $g(A)$ will have a degree less than n.

Consider our previous example with $A = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 2 & -2 \\ -8 & 5 & 1 \end{pmatrix}$, but suppose that instead of choosing the natural seed vector, which gave us an eigenbasis immediately, we chose the vector $\begin{pmatrix} 2 \\ 3 \\ -1 \end{pmatrix}$ as our initial seed. We would compute

$$\mathbf{v} = \begin{pmatrix} 2 \\ 3 \\ -1 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 4 \\ 6 \\ -2 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 8 \\ 12 \\ -4 \end{pmatrix} \qquad A^3\mathbf{v} = \begin{pmatrix} 16 \\ 24 \\ -8 \end{pmatrix}$$

then row reduce the matrix of column vectors

$$\begin{pmatrix} 2 & 4 & 8 & 16 \\ 3 & 6 & 12 & 24 \\ -1 & -2 & -4 & -8 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 4 & 8 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which gives us minimal polynomial $f(A) = A - 2I$ and eigenvalue $\lambda = 2$. The corresponding eigenvector is $\begin{pmatrix} 2 \\ 3 \\ -1 \end{pmatrix}$, with geometric multiplicity 1.

Since we should have three eigenvectors, we choose another seed vector. Again for illustrative purposes, we will avoid the obvious choice, and instead choose $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix}$. Since our minimal polynomial is $f(A) = A - 2I$, we will apply this to our seed vector to produce the seedling vector:

$$f(A)\mathbf{x} = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 0 & -2 \\ -8 & 5 & -1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ -8 \end{pmatrix}.$$

It is possible for our seed vector to "die" and give us the zero vector, not unlike last summer's attempt to grow heirloom tomatoes. In that case, we would have to choose another seed vector. We leave it as an exercise for the reader to produce an algorithm for generating "guaranteed-to-grow" seed vectors, though in practice simply choosing a random vector will usually work well enough.

Using our seedling vector we compute

$$\mathbf{v} = \begin{pmatrix} 4 \\ 4 \\ -8 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 16 \\ 20 \\ -20 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 48 \\ 64 \\ -48 \end{pmatrix} \qquad A^3\mathbf{v} = \begin{pmatrix} 128 \\ 176 \\ -112 \end{pmatrix}$$

(the alert reader will note we do not actually need $A^3\mathbf{v}$). Row reducing the column space produces

$$\begin{pmatrix} 4 & 16 & 48 & 128 \\ 4 & 20 & 64 & 176 \\ -8 & -20 & -48 & -112 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -4 & -16 \\ 0 & 1 & 4 & 12 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which gives us minimal polynomial

$$g(A) = A^2 - 4A + 4I = (A - 2I)^2.$$

Note that since $\mathbf{v} = (A - 2I)\mathbf{x}$, we have

$$g(A)\mathbf{v} = \mathbf{0},$$
$$(A - 2I)^2(A - 2I)\mathbf{x} = \mathbf{0}.$$

Consequently, our seed vector $\mathbf{x}$ is the first in a Jordan chain; we compute the others:

$$\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix} \qquad (A - 2I)\mathbf{x} = \begin{pmatrix} 4 \\ 4 \\ -8 \end{pmatrix} \qquad (A - 2I)^2\mathbf{x} = \begin{pmatrix} 8 \\ 12 \\ -4 \end{pmatrix}.$$

This leads to the following:

Lemma 3. *Let $\mathbf{v} = f(A)\mathbf{x}$, where $\mathbf{v} \neq \mathbf{0}$. Let $g(A)$ be the minimal polynomial for $\mathbf{v}$. If $g(A)$ and $f(A)$ have a common root λ, then λ corresponds to a Jordan chain of generalized eigenvectors.*

Proof. Let

$$f(A) = (A - \lambda I)^r f'(A) \quad \text{and} \quad g(A) = (A - \lambda I)^s g'(A),$$

where $f'(A)$ and $g'(A)$ have no factors of $A - \lambda I$. Since $g(A)f(A)\mathbf{x} = \mathbf{0}$ and the factors of f, g can be rearranged, we have

$$(A - \lambda I)^{r+s} g'(A) f'(A)\mathbf{x} = \mathbf{0}.$$

We claim that $g'(A)f'(A)\mathbf{x}$ is a generalized eigenvector of rank $r + s$.

First, we show that

$$g'(A)f'(A)\mathbf{x} \neq \mathbf{0}.$$

Suppose $f'(A)\mathbf{x} = \mathbf{0}$. Since

$$f(A) = (A - \lambda I)^r f'(A),$$

this would mean that $f(A)\mathbf{x} = \mathbf{0}$, which contradicts our assumption that $f(A)\mathbf{x} \neq \mathbf{0}$. Thus $f'(A)\mathbf{x} \neq \mathbf{0}$.

Now suppose

$$g'(A)\,f'(A)\mathbf{x} = \mathbf{0}.$$

Supplying the missing factors $(A - \lambda I)^r$ gives us

$$g'(A)(A - \lambda I)^r f'(A) = \mathbf{0}.$$

But

$$(A - I)^r f'(A) = f(A),$$

so we have $g'(A)\,f(A)\mathbf{x} = \mathbf{0}$. Consequently $g'(A)$ is a minimal polynomial of A with respect to $f(A)\mathbf{x} = \mathbf{v}$.

But

$$g(A) = (A - \lambda I)^s g'(A),$$

so $g'(A)$ would have a smaller degree than $g(A)$, which was assumed minimal. Consequently $g'(A)\,f'(A)\mathbf{x} \neq \mathbf{0}$. Thus $g'(A)\,f'(A)\mathbf{x}$ will be a generalized eigenvector of rank $r + s$, and λ will correspond to a Jordan chain of generalized eigenvectors. ■

Eigenvectors: now determinant free!

This leads to the following algorithm for finding all eigenvectors and generalized eigenvectors of an $n \times n$ matrix A:

- Pick any non-zero seed vector $\mathbf{v}_1$ and find the minimal polynomial $f_1(A)$ with respect to $\mathbf{v}_1$.
- Find the corresponding eigenvalues, eigenvectors, and generalized eigenvectors. If you have found n linearly independent eigenvectors and generalized eigenvectors, the algorithm terminates.
- If not, pick any non-zero seed vector $\mathbf{x}$ and grow it to seedling vector $\mathbf{v}_2 = f_1(A)\mathbf{x}$ (verifying $\mathbf{v}_2 \neq \mathbf{0}$: in the botanical analogy, it is possible that some of our seeds may fail to germinate!)
- Find the minimal polynomial $f_2(A)$ with respect to $\mathbf{v}_2$.
- Find the corresponding eigenvalues, eigenvectors, and generalized eigenvectors. If the eigenvalues are new, add the corresponding eigenvectors to the eigenbasis. If the eigenvalues are duplicates, replace the original Jordan chains with the new ones. If you have found n linearly independent eigenvectors and generalized eigenvectors, the algorithm terminates.
- If not, pick any non-zero seed vector $\mathbf{x}$ and grow it into non-zero seedling vector $\mathbf{v}_3 = f_2(A)\,f_1(A)\mathbf{x}$, etc.

McWorter and Meyers present a "Simon Legree" matrix as an illustration of the power of the determinant-free method. However, one of the referees for an earlier draft of this article noted that the standard properties of the determinant could be used to reduce the matrix to a more tractable one. For that reason, we present a more nightmarish problem.

Example 3. *Find all eigenvectors of*

$$A = \begin{pmatrix} 35 & 88 & 46 & 12 \\ -8 & -25 & -14 & -4 \\ -52 & -124 & -63 & -16 \\ 162 & 414 & 216 & 57 \end{pmatrix}.$$

The problem is solved too easily if we use our standard seed vector, as we would find all three eigenvalues in a single step. So, for illustrative purposes we will choose $\mathbf{v} = \begin{pmatrix} 0 & 4 & -10 & 9 \end{pmatrix}^T$ and compute

$$\mathbf{v} = \begin{pmatrix} 0 \\ 4 \\ -10 \\ 9 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 0 \\ 4 \\ -10 \\ 9 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 0 \\ 4 \\ -10 \\ 9 \end{pmatrix}$$

$$A^3\mathbf{v} = \begin{pmatrix} 0 \\ 4 \\ -10 \\ 9 \end{pmatrix} \qquad A^4\mathbf{v} = \begin{pmatrix} 0 \\ 4 \\ -10 \\ 9 \end{pmatrix}.$$

(Obviously, we could have stopped when we realized that $\{\mathbf{v}, A\mathbf{v}\}$ was a dependent set.)

Using these as column vectors and row reducing leads to

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 4 & 4 & 4 & 4 & 4 \\ -10 & -10 & -10 & -10 & -10 \\ 9 & 9 & 9 & 9 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

with corresponding minimal polynomial $f(A) = A - I$. Thus $\lambda = 1$ is an eigenvalue, and we find the corresponding set of linearly independent eigenvectors $\begin{pmatrix} 2 & -2 & 0 & 9 \end{pmatrix}^T$ and $\begin{pmatrix} 1 & -3 & 5 & 0 \end{pmatrix}^T$. Since A is a 4×4 matrix, we can find two more linearly independent eigenvectors and/or generalized eigenvectors.

To find them, we will pick another seed vector, then apply our minimal polynomial $f(A) = A - I$ to grow it into a seedling. This time we will use the standard choice, and compute:

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad (A - I)\mathbf{x} = \begin{pmatrix} 34 \\ -8 \\ -52 \\ 162 \end{pmatrix}.$$

Using $\mathbf{v} = (A - I)\mathbf{x}$ as our seedling, we compute as before:

$$\mathbf{v} = \begin{pmatrix} 34 \\ -8 \\ -52 \\ 162 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 38 \\ 8 \\ -92 \\ 198 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 178 \\ -8 \\ -340 \\ 882 \end{pmatrix}$$

$$A^3\mathbf{v} = \begin{pmatrix} 470 \\ 8 \\ -956 \\ 2358 \end{pmatrix} \qquad A^4\mathbf{v} = \begin{pmatrix} 1474 \\ -8 \\ -2932 \\ 7362 \end{pmatrix}.$$

We leave the reader to show that we do not actually need to compute all five vectors, since the minimal polynomial for $\mathbf{v}$ will have a degree of at most 2.

In any case, using these as column vectors and row reducing leads to,

$$\begin{pmatrix} 34 & 38 & 178 & 470 & 1474 \\ -8 & 8 & -8 & 8 & -8 \\ -52 & -92 & -340 & -956 & -2932 \\ 162 & 198 & 882 & 2358 & 7362 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 3 & 6 & 21 \\ 0 & 1 & 2 & 7 & 20 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

with corresponding minimal polynomial $f(A) = A^2 - 2A - 3I$. This gives us the remaining eigenvalues $\lambda = 3, -1$. These, in turn, give us the remaining eigenvectors.

What about generalized eigenvectors?

Example 4. *Find all eigenvectors and generalized eigenvectors for*

$$\begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \\ 0 & -1 & 1 \end{pmatrix}.$$

As above, the problem is solved too quickly if we use our obvious seed vector $\begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$, so instead we will choose seed vector $\mathbf{v} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$, and find that

$$\mathbf{v} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 4 \\ 1 \\ -2 \end{pmatrix} \qquad A^3\mathbf{v} = \begin{pmatrix} 6 \\ 1 \\ -3 \end{pmatrix}.$$

Row reducing our column space leads to

$$\begin{pmatrix} 0 & 2 & 4 & 6 \\ 1 & 1 & 1 & 1 \\ 0 & -1 & -2 & -3 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which has solution

$$x_1 = 1, x_2 = -2, x_3 = 1, x_4 = 0$$

and corresponding minimal polynomial $f(A) = A^2 - 2A + I = (A - I)^2$. Thus, $(A - I)^2\mathbf{v} = \mathbf{0}$, and our seed vector $\mathbf{v}$ is a rank 2 generalized eigenvector. The complete Jordan chain is

$$\mathbf{v} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad (A - I)\mathbf{v} = \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix},$$

which gives two eigenvectors.

We note that since $(A - I)^2\mathbf{v} = \mathbf{0}$, we have that $(A - I)\mathbf{v}$ is an eigenvector corresponding to eigenvalue $\lambda = 1$. We verify that $\lambda = 1$ has geometric multiplicity 1, so the corresponding eigenvector will be some scalar multiple of $(A - I)\mathbf{v}$. This means we have only found two out of three linearly independent eigenvectors and generalized eigenvectors.

Thus, we choose another seed vector. Again for illustrative purposes, we choose $\mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. First, we grow it into a seedling by applying $f(A) = (A - I)^2$:

$$\mathbf{u} = (A - I)^2\mathbf{x} = \begin{pmatrix} 4 \\ 0 \\ -2 \end{pmatrix}.$$

As above, the minimal polynomial with respect to $\mathbf{x}$ has degree 3 or less, and since $\mathbf{u} = (A - I)^2\mathbf{x}$, the minimal polynomial with respect to $\mathbf{u}$ will have degree 1. This means we only need to find

$$\mathbf{u} = \begin{pmatrix} 4 \\ 0 \\ -2 \end{pmatrix} \qquad A\mathbf{u} = \begin{pmatrix} 4 \\ 0 \\ -2 \end{pmatrix},$$

which gives us $(A - I)\mathbf{u} = \mathbf{0}$. Since $\mathbf{u} = (A - I)^2\mathbf{x}$, we have $\mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ as a generalized eigenvector of rank 3. This gives us the Jordan chain

$$\mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad (A - I)\mathbf{x} = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} \qquad (A - I)^2\mathbf{x} = \begin{pmatrix} 4 \\ 0 \\ -2 \end{pmatrix}.$$

We replace the original Jordan chain, giving us an eigenbasis corresponding to A.

In the preceding example, our eigenvalue had a geometric multiplicity equal to 1. What if it has a greater geometric multiplicity?

Example 5. *Find an eigenbasis corresponding to* $A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ -2 & -2 & -1 \end{pmatrix}$.

Using our standard seed vector, we find

$$\mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad A\mathbf{v} = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix} \qquad A^2\mathbf{v} = \begin{pmatrix} 3 \\ 2 \\ -4 \end{pmatrix} \qquad A^3\mathbf{v} = \begin{pmatrix} 4 \\ 3 \\ -6 \end{pmatrix}.$$

Row reducing the column space leads to

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \\ 0 & -2 & -4 & -6 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with solution $x_1 = 1, x_2 = -2, x_3 = 1, x_4 = 0$, giving us minimal polynomial $f(A) = A^2 - 2A + I$ with root $\lambda = 1$. We find $\lambda = 1$ corresponds to two linearly independent eigenvectors. We take these to be

$$\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

We *also* note that since our minimal polynomial was $A^2 - 2A + I$, our seed vector generated a Jordan chain of generalized eigenvectors, namely

$$\mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad (A - I)\mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}.$$

We have two linearly independent eigenvectors, and a Jordan chain of two vectors. However, our eigenbasis will have three vectors only, so we have too many vectors for a basis. The most straightforward way to find our eigenbasis is to row reduce the corresponding column space. Since the vectors in our Jordan chain are guaranteed to be independent, we will set these as our first two columns and row reduce to determine which of the remaining vectors is superfluous:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & -2 & -1 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 2 \end{pmatrix},$$

which tells us that the vector $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ can be expressed as linear combination of the others. Consequently, we may strike it from our set of vectors, leaving us with eigenbasis

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\}.$$

REFERENCES

[1] Axler, S. (1995). Down with determinants. *Amer. Math. Monthly*. 102(2): 139–154. doi.org/10.2307/2975348

[2] Axler, S. (2015). *Linear Algebra Done Right*, 3rd ed. New York: Springer-Verlag.

[3] McWorter Jr., W. A., Meyers, L. F. (1998). Computing eigenvalues and eigenvectors without determinants. *Math. Mag*. 71(1): 24–33. doi.org/10.1080/0025570X.1998.11996591

Summary. The standard approach to finding eigenvalues relies on solving the characteristic polynomial. But the characteristic polynomial for an $n \times n$ matrix will require computing a determinant with $n!$ terms and solving an nth degree polynomial equation, both of which are daunting tasks if $n \geq 3$. We present a determinant-free approach which often leads to lower-degree polynomial equations, and which provides a natural introduction to the concept of a generalized eigenvector.

JEFF SUZUKI an associate professor of mathematics at Brooklyn College, and more than twenty years after completing his dissertation, is still trying to figure out what he really wants to do. His publications cover such topics as the history of mathematics, the dynamics of revolutions, the mathematics of the U.S. Constitution, and patents based on mathematics known to any undergraduate math major. His current interests are linear algebra, mathematics education, and medieval dance.

Iterating the Locker Problem

REBECCA L. JAYNE
Hampden-Sydney College
Hampden-Sydney, VA 23943
rjayne@hsc.edu

ROBB T. KOETHER
Hampden-Sydney College
Hampden-Sydney, VA 23943
rkoether@hsc.edu

In a hallway there are 1000 closed lockers labeled 1 to 1000. There is also a group of 1000 students labeled 1 to 1000. Each of the 1000 students is sent down the hallway, one after the other, with the same instructions: for each k, student k is told to reverse the state of every kth locker door, starting with locker number k. That is, if the locker is closed, then open it, and if it is open, then close it. The question is, after all 1000 students have gone down the hallway, which lockers are left open?

The solution is straightforward and well known [**3**]. Each locker door is toggled once for every divisor of its number. If a locker has an even number of divisors, then it is toggled an even number of times and thus will be closed in the end. On the other hand, if a locker has an odd number of divisors, then it will be left open. So the question can be restated as follows: of the integers from 1 to 1000, which of them have an odd number of divisors? The answer is now clear. The perfect squares are the only integers with an odd number of divisors. (Every divisor of an integer n is a member of a pair of distinct integers with product n, except when $\sqrt{n}$ is itself an integer.)

Teachers use the locker problem with students from preteens to college-age to learn about factors and multiples, prime numbers, and problem solving skills. In addition, there have been some written works, for example, Dagal [**1**] and Torrance [**5**], that extend the idea of the locker problem. In this article, we think about some other variations of the problem.

To start thinking of such variations, we may begin by wondering: if there are 30 lockers and 30 students, what set of students will end up, after walking past the lockers and following their instructions, leaving open exactly the prime-numbered lockers? That is, which students should we send if our goal is to leave open precisely the following lockers?

$$\{2, 3, 5, 7, 11, 13, 17, 19, 23, 29\}.$$

The answer is that the set of students is

$$\{2, 3, 4, 5, 7, 9, 11, 12, 13, 17, 18, 19, 20, 23, 25, 28, 29, 30\}.$$

How do we determine that? One method is to consider the lockers in order, from locker 1 to locker 30. Do we want locker 1 to be open? No, so do not send student 1. Do we want locker 2 to be open? Yes, so send student 2, who will also open lockers 4, 6, 8, and so on. Then consider locker 3, and so on, until every locker has been properly set as open or closed. This method works, but it requires simulating the entire procedure.

Is there a better way to determine which students will leave open the prime numbered lockers? More generally, can we find, for any set of specified doors, a set of

Math. Mag. **93** (2020) 213–224. doi:10.1080/0025570X.2020.1736887
MSC: 00A00
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

students who will leave them open? The answer is, "Yes, although it is easier in certain cases than in others."

When we think about the problem this way, each locker door has two states: open and closed. We can think of this another way too. For some integer q, we can think of a locker door as having q different states: 0, 1, 2, ... $q-1$. Here we say that in state 0 a locker is closed, and in states $1, 2, \ldots, q-1$, it is open in varying degrees. We restrict our attention to the cases where q is prime and associate a value, say c_k, between 0 and $q-1$ for student k. When student k interacts with a door, he advances it by c_k modulo q. (So if a door is open in state $q-1$ and student k with value $c_k = 2$ interacts with it, the door will then be open in state 1.) We can ask similar questions as we did in the original ($q = 2$) case. Can we figure out which doors will be open, and by how much, if we send a particular group of students down the hall? If we have certain doors we want to be open certain amounts, can we find a set of students that makes this happen? Again, we answer with, "Yes, although it is easier in certain cases than in others."

What if we look at the locker problem iteratively? For arbitrary q, after sending students down the hallway, we record the state of each locker. Then we close all the lockers and send the students back down, the state of student k matching the state of locker k before we closed the lockers. What happens now? What happens if we keep doing this?

Computing locker states

The locker function Let N be the number of students and the number of lockers. The lockers are in any of q different states, for some prime q. (We require that q be a prime in order to facilitate later results.) We represent the states of students 1 to N as well as the states of lockers 1 to N with vectors in $\mathbb{Z}_q^N$. Now, define the *locker function* $f: \mathbb{Z}_q^N \to \mathbb{Z}_q^N$ to be the function that takes a vector of states of students and sends all the students down the hallway using our usual procedure. The image of this vector under f will be the vector that describes the states of all of the lockers once every student has gone down the hall.

For example, let us say $N = 10$, $q = 3$, and we have students with states

$$\mathbf{v} = (0, 2, 0, 1, 2, 1, 0, 0, 1, 1)$$

and we want to find $f(\mathbf{v})$. There is no need to send students 1, 3, 7, or 8 because these students are in state 0 and do not affect any doors. In Table 1, we show the status of the locker doors at each step, finding that

$$f(\mathbf{v}) = (0, 2, 0, 0, 2, 0, 0, 0, 1, 2).$$

	Locker number									
	1	2	3	4	5	6	7	8	9	10
Lockers begin closed	0	0	0	0	0	0	0	0	0	0
After student 2	0	2	0	2	0	2	0	2	0	2
After student 4	0	2	0	0	0	2	0	0	0	2
After student 5	0	2	0	0	2	2	0	0	0	1
After student 6	0	2	0	0	2	0	0	0	0	1
After student 9	0	2	0	0	2	0	0	0	1	1
After student 10	0	2	0	0	2	0	0	0	1	2

TABLE 1: Example of the function f.

We can establish that f is a bijection by induction on N. The case where $N = 1$ is trivial. Now, suppose that f is a bijection for some $N \geq 1$, and add in student and locker $N + 1$. This student has no effect on doors 1 through N, but he will be able to put locker $N + 1$ in any of q possible states. Thus, f is onto, and hence is a bijection. This tells us that we can obtain any configuration of lockers just by sending the appropriate students! It also tells us that there is only one set of students that gives us any particular configuration of lockers.

By the very nature of the locker function, it is clear that for any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_q^N$ and any $a \in \mathbb{Z}_q$, $f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{u})$ and $f(a\mathbf{v}) = af(\mathbf{v})$. Therefore, f is a linear transformation.

Iteration of the locker function Let us iterate f. After evaluating f once, we close all the locker doors and use the image vector as an input vector, as mentioned at the end of the previous section. In other words, we find $f(f(\mathbf{v})) = f^2(\mathbf{v})$. Returning to our recent example,

$$\mathbf{v} = (0, 2, 0, 1, 2, 1, 0, 0, 1, 1)$$

with $N = 10$ and $q = 3$. We find

$$\begin{aligned} f^2(\mathbf{v}) = f(f(\mathbf{v})) &= f((0, 2, 0, 0, 2, 0, 0, 0, 1, 2)) \\ &= (0, 2, 0, 2, 2, 2, 0, 2, 1, 0) \end{aligned}$$

by sending students 2, 5, 9, and 10 in their appropriate states.

What if we continue to do this, always sending down the hallway exactly those students whose lockers were left open from the previous iteration? What happens after n iterations?

Let us begin with

$$\mathbf{v}_0 = (1, 1, \ldots, 1) \in \mathbb{Z}_q^N,$$

that is, all students are in state 1. For $n \geq 1$, define $\mathbf{v}_n = f(\mathbf{v}_{n-1})$, and for $n \geq 0$, we denote the dth entry in $\mathbf{v}_n$ by $\mathbf{v}_n(d)$. Because we change a locker d only when we send a student that is a divisor of d, for $n \geq 0$, and all lockers d,

$$\mathbf{v}_n(d) \equiv \sum_{m|d} \mathbf{v}_{n-1}(m) \pmod{q}. \tag{1}$$

Our goal is to determine $\mathbf{v}_n(d)$ for any $n \geq 0$, $1 \leq d \leq N$. We begin by considering the simpler case where d is a power of a prime. Let $d = p^t$ for some prime p and for some integer $t \geq 0$.

For $q = 2$, we may plot the successive values of $\mathbf{v}_n(p^t)$ for $t \geq 0$ on the horizontal axis and $n \geq 0$ on the vertical axis. We draw a dark square for an open locker and a white square for a closed locker. See Figure 1, which is the same for any prime p. If the graph is rotated 135° clockwise, we see that it is Pascal's triangle modulo 2!

When we think of Pascal's triangle, we think about binomial coefficients, which leads us to the following theorem.

Theorem 1. *Let p be a prime. For all $n, t \geq 0$,*

$$\mathbf{v}_n(p^t) \equiv \binom{n+t}{t} \pmod{q}.$$

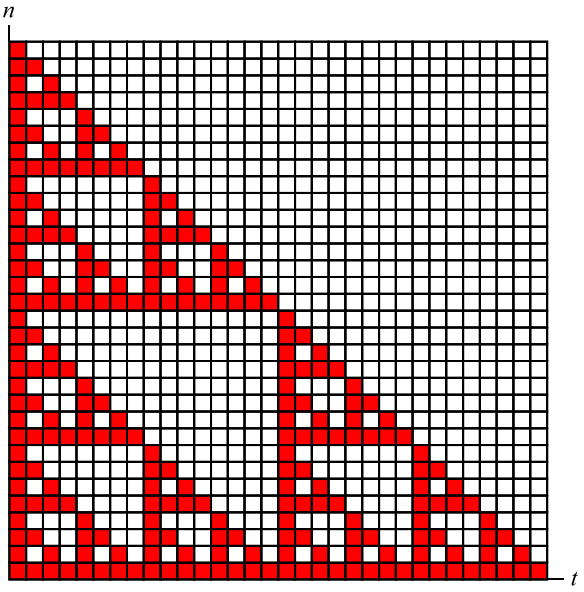


Figure 1 Values of $\mathbf{v}_n(p^t)$ for $0 \le n, t \le 31$

Proof. When $n = 0$ or $t = 0$, we have

$$\mathbf{v}_0(p^t) = 1 = \binom{0+t}{t},$$

$$\mathbf{v}_n(p^0) = \mathbf{v}_n(1) = 1 = \binom{n+0}{0}.$$

Now suppose that $n > 0$ and $t > 0$ and proceed by induction. The only divisors of d are the integers p^i where $0 \le i \le t$. Therefore,

$$\begin{aligned}
\mathbf{v}_n(p^t) &\equiv \sum_{i=0}^{t} \mathbf{v}_{n-1}(p^i) \pmod{q} \\
&\equiv \mathbf{v}_{n-1}(p^t) + \sum_{i=0}^{t-1} \mathbf{v}_{n-1}(p^i) \pmod{q} \\
&\equiv \mathbf{v}_{n-1}(p^t) + \mathbf{v}_n(p^{t-1}) \pmod{q} \\
&\equiv \binom{n-1+t}{t} + \binom{n+t-1}{t-1} \pmod{q} \\
&\equiv \binom{n+t}{t} \pmod{q}.
\end{aligned}$$

■

It will be convenient to use base q representations of integers. Let

$$a = a_k q^k + \cdots + a_1 q + a_0,$$

where $0 \le a_j \le q-1$. When necessary to avoid ambiguity, we will denote this by

$$a = [a_k] \cdots [a_1][a_0]_q.$$

The following theorem of Eduoard Lucas (see Niven and Zuckerman [4]) and the subsequent corollary (see Holte [2]) are instrumental to the development of our main results.

Theorem 2 (Lucas's theorem)**.** *Let q be a prime and let a and b be nonnegative integers with base q representations $[a_k] \cdots [a_1][a_0]_q$ and $[b_k] \cdots [b_1][b_0]_q$, respectively. Then*

$$\binom{a}{b} \equiv \binom{a_k}{b_k} \cdots \binom{a_1}{b_1}\binom{a_0}{b_0} \pmod{q}.$$

Corollary 1. *Using the notation of Lucas's theorem,*

$$\binom{a+b}{b} \equiv \binom{a_k+b_k}{b_k} \cdots \binom{a_1+b_1}{b_1}\binom{a_0+b_0}{b_0} \pmod{q}.$$

Let $n = [n_k] \cdots [n_1][n_0]_q$ and $t = [t_k] \cdots [t_1][t_0]_q$. Then, by Theorem 1 and Corollary 1,

$$\mathbf{v}_n(p^t) \equiv \binom{n_k+t_k}{t_k} \cdots \binom{n_1+t_1}{t_1}\binom{n_0+t_0}{t_0} \pmod{q}. \tag{2}$$

For any i, if $n_i + t_i \ge q$, then

$$\binom{n_i+t_i}{t_i} = \frac{(n_i+t_i)!}{n_i!t_i!} \equiv 0 \pmod{q}$$

because q is prime and $n_i, t_i < q$. Similarly,

$$\binom{n_i+t_i}{t_i} \not\equiv 0 \pmod{q}$$

when $n_i + t_i < q$. Therefore, $\mathbf{v}_n(p^t) \equiv 0 \pmod{q}$ if and only if $n_i + t_i \ge q$ for some i.

In other words, if the addition $n + t$ in base q results in at least one "carry," then door p^t is closed after n iterations. Define the operator $\oplus\colon \mathbb{Z}_q \times \mathbb{Z}_q \to \mathbb{Z}_q$ as

$$a \oplus b \equiv \binom{a+b}{b} \pmod{q}.$$

Then we may write

$$\mathbf{v}_n(p^t) = n \oplus t.$$

Equation (1) and the following well-known lemma [4] show that $\mathbf{v}_n$ is a multiplicative function. (A function f is *multiplicative* if $f(ab) = f(a)f(b)$ whenever a and b are relatively prime.)

Lemma 1. *If f is multiplicative and*

$$g(n) = \sum_{d|n} f(d),$$

then g is multiplicative.

Recall that $\mathbf{v}_0(d) = 1$ for all d, making $\mathbf{v}_0$ trivially multiplicative. That fact together with Lemma 1 and Equation (1) establish, by induction, that $\mathbf{v}_n$ is multiplicative.

Now we consider the general case where d is any locker between 1 and N. Let $d = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ be the prime factorization of d. By the previous lemma, we have

$$\mathbf{v}_n(d) = \mathbf{v}_n(p_1^{e_1})\mathbf{v}_n(p_2^{e_2}) \cdots \mathbf{v}_n(p_r^{e_r}) \equiv \prod_{i=1}^{r} n \oplus e_i \pmod{q}.$$

Inspired by the example of Torrence and Wagon [5], define the *signature* of d to be the *multiset* $\sigma(d) = \{e_1, e_2, \ldots, e_r\}$ (repetitions allowed) and as a special case define $\sigma(1) = \{0\}$. Extend the definition of $\oplus$ to these multisets as follows:

$$n \oplus \sigma(d) \equiv \prod_{i=1}^{r} n \oplus e_i \pmod{q}.$$

Then we have our main theorem:

Theorem 3. *For $n \geq 0$ and $d = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$,*

$$\mathbf{v}_n(d) = n \oplus \sigma(d) = \prod_{i=1}^{r} \prod_{j=1}^{k} \binom{n_j + e_{ij}}{e_{ij}},$$

where $e_i = [e_{ik}] \cdots [e_{i1}][e_{i0}]_q$.

Thus, locker d will be closed after n iterations if and only if, for at least one base-q digit n_j of n and at least one base-q digit e_{ij} of some exponent e_i, $n_j + e_{ij} \geq q$. With that in mind, we could streamline the computation by defining an integer m whose base-q digits are the maximums of the corresponding base-q digits of the exponents e_i to see whether any $n_j + e_{ij} \geq q$. Then compute $n \oplus m$ simply to determine whether the door is open or closed. However, if we also want to know the precise state of the door modulo q, then we must evaluate the full formula.

We include some examples.

Example 1. As in the original locker problem, let there be $N = 1000$ lockers, let $q = 3$, and consider locker number $d = 720$. After $n = 12$ iterations, will locker 720 be left open or closed?

To answer this, note that in base 3 we have $n = 110_3$. We factor $720 = 2^4 \cdot 3^2 \cdot 5^1$, and note that

$$\sigma(720) = \{1, 2, 4\} = \{001_3, 002_3, 011_3\}.$$

We see at a glance that $n \oplus e_i \neq 0$ for any i (no carries), so we know that the door is open. To find the state of the door (state 1 or state 2 modulo 3), we compute

$$\begin{aligned} \mathbf{v}_{12}(720) &= 12 \oplus \sigma(720) \\ &= 110_3 \oplus \{001_3, 002_3, 011_3\} \end{aligned}$$

$$= (110_3 \oplus 001_3)(110_3 \oplus 002_3)(110_3 \oplus 011_3)$$
$$= (1)(1)(2) = 2.$$

Therefore, locker 720 is open and in state 2 after 12 iterations.

Example 2. As a second example, consider the original question: Which lockers d are left open after $n = 1$ iteration (modulo 2)? In binary, $n = 00\ldots01_2$ so only those lockers whose signature contains only even numbers will be left open. For if any exponent e_i were odd, then its units digit in base 2 would be 1, thereby creating a carry. Such lockers are exactly the perfect squares.

Example 3. The previous example leads to a generalization of the original locker problem as follows. For the prime modulus q, which doors will be left open after $q - 1$ iterations? Because $q - 1 = [q - 1]_q$, it follows that

$$(q - 1) \oplus e \not\equiv 0 \pmod{q}$$

if and only if e is a multiple of q (base-q unit's digit 0). In other words, the doors left open after $q - 1$ iterations will be exactly the perfect qth powers.

What eventually happens?

As we continue to iterate the function f, what eventually happens? Because $\mathbb{Z}_q^N$ is a finite set, it follows that the vectors $\mathbf{v}_n$ will eventually repeat, after which they will fall into a cycle. Furthermore, because f is a bijection, the first repeated value must be the initial value. How long might that take? As it turns out, not long at all.

Let 2^K be the smallest power of 2 that is greater than N. Since 2 is the smallest prime, it follows that in the prime factorization $p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ of any $d \in \{1, 2, 3, \ldots, N\}$, every exponent will be less than K. The base-q representations of these exponents will require at most $M = \lfloor \log_q K \rfloor + 1$ digits.

Now consider what happens if we iterate $q^M - 1$ times. Because $q^M - 1$ is the largest base-q number expressible in M digits,

$$q^M - 1 = [q - 1] \cdots [q - 1][q - 1]_q.$$

When $d > 1$, at least one $e_i > 0$, and therefore $q - 1 + e_i \geq q$. In this case,

$$q^M - 1 \oplus \sigma(d) = 0$$

and the locker is closed. When $d = 1$, every $e_i = 0$. Therefore,

$$q^M - 1 \oplus \sigma(1) = 1$$

and locker 1 is open.

On the next iteration, we have $q^M = [1][0] \cdots [0][0]_q$. (The digit 1 is in position $M + 1$.) In this case, $0 + e_i < q$ in every position, so $q^M \oplus \sigma(d) = 1$ for every locker, meaning that every locker is left open and in state 1 after q^M iterations. That brings us back to the original situation on the next iteration. We have the following theorem.

Theorem 4. *Let N be the number of lockers, let $K = \lfloor \log_2 N \rfloor + 1$, and let $M = \lfloor \log_q K \rfloor + 1$. Then*

$$\mathbf{v}_{q^M} = \mathbf{v}_0.$$

For example, if $N = 100$ and $q = 3$, then $K = 7$ and $M = 2$, and the largest possible exponent would be 6. Any exponent less than or equal to 6 requires only 2 digits in base 3. Therefore, the sequence will repeat after at most $3^2 = 9$ iterations, and if not 9, then 3 iterations. In a more extreme example, if there were 10 million lockers and $q = 3$, then $K = 24$ and $M = 3$, so the sequence repeats after at most only $3^3 = 27$ iterations!

Although the locker problem begins with *all* students sent down the hallway, the previous observation suggests that it would be more natural to begin by sending only student 1 in state 1. That student will leave every locker open and in state 1, setting the stage for the standard locker problem. Since $\mathbf{v}_0 = (1, 1, 1, \ldots, 1)$, we could define

$$\mathbf{v}_{-1} = \mathbf{v}_{q^M-1} = (1, 0, 0, \ldots, 0).$$

The general case

Now we consider the more general question: what happens if we send an arbitrary set of students, each in an arbitrary state represented by the vector $(c_1, c_2, c_3, \ldots, c_N)$, on the first iteration? (Setting $c_i = 0$ is equivalent to not sending student i.) Based on the previous observations, it will be convenient to let this vector be $\mathbf{v}_{-1}$. Then, as usual, $\mathbf{v}_0 = f(\mathbf{v}_{-1})$. The question is, for $n \geq 0$, what states are the lockers in after n iterations. That is, what is the value of $\mathbf{v}_n$? The reverse question is, what should $\mathbf{v}_{q^M-2}$ be in order to obtain a desired configuration $\mathbf{v}_{-1}$? For example, for what configuration $\mathbf{v}_{q^M-2}$ will $\mathbf{v}_{-1}$ have every prime-numbered locker open in state 1 and the other lockers closed?

To see how to calculate $\mathbf{v}_n$, we begin with the simpler case where $\mathbf{v}_{-1}$ consists of only a single student s in state c. That is, $\mathbf{v}_{-1} = (0, \ldots, c, \ldots, 0)$. It is easy to see that only those lockers that are multiples of s will be affected. A locker d will be modified by student s as locker d/s was modified by student 1 if d is a multiple of s. If d is not a multiple of s, then locker d will not be affected. That is, it is the same as the original locker problem except that all locker numbers are scaled by the factor s, and their values are scaled by c.

From the fact that f is a linear transformation, our main theorem for an arbitrary $\mathbf{v}_{-1}$ follows immediately.

Theorem 5. *Let* $\mathbf{v}_{-1} = (c_1, c_2, c_3, \ldots, c_N)$. *Then for all* $n \geq 0$ *and for all lockers* d,

$$\mathbf{v}_n(d) \equiv \sum_{s|d} c_s \cdot n \oplus \sigma(d/s) \pmod{q}. \tag{3}$$

Is Theorem 5 consistent with Theorem 3? We see that it is by the following. If we begin with $\mathbf{v}_{-1} = (1, 0, 0, \ldots, 0)$, then

$$\begin{aligned} \mathbf{v}_n(d) &\equiv \sum_{s|d} c_s \cdot n \oplus \sigma(d/s) \pmod{q} \\ &\equiv 1 \cdot n \oplus \sigma(d/1) \pmod{q} \\ &\equiv n \oplus \sigma(d) \pmod{q}. \end{aligned}$$

Now we consider what eventually happens for arbitrary $\mathbf{v}_0$. In this case, the vectors $\mathbf{v}_n$ repeat in cycles for the same reasons as in the case where $\mathbf{v}_0 = (1, 1, \ldots, 1)$. Define M as before and consider $\mathbf{v}_{q^M-1}$. Recall that $(q^M - 1) \oplus \sigma(d)$ is 1 only when $d = 1$

and 0 otherwise. Thus, we have

$$\mathbf{v}_{q^M-1}(d) \equiv \sum_{s|d} c_s \cdot n \oplus \sigma(d/s) \pmod{q} = c_d$$

and so $\mathbf{v}_{q^M-1} = \mathbf{v}_{-1}$. It follows that that $\mathbf{v}_{q^M} = \mathbf{v}_0$.

Remark. Before presenting the next two examples, it will be helpful to note the following. For any exponent e,

$$(q^M - 2) \oplus e \equiv \begin{cases} 1 \pmod{q} & \text{if } e = 0, \\ -1 \pmod{q} & \text{if } e = 1, \\ 0 \pmod{q} & \text{if } e \geq 2. \end{cases}$$

This follows from the fact that $q^M - 2 = [q-1]\cdots[q-1][q-2]_q$.

We now give some examples.

Example 4. In the original locker problem, we saw that the lockers whose numbers were perfect squares were left open. We now ask a more general question: which set of students will leave open those lockers whose numbers are perfect kth powers, where $k \geq 2$? Let $\mathbf{v}_{-1}$ represent, in state 1, every student whose number is a kth power j^k for some j, and, in state 0, all other students. Let $d = m^k d'$ where d' is kth-power free. Then by Theorem 5 and the above remark,

$$\begin{aligned} \mathbf{v}_{q^M-2}(d) &\equiv \sum_{j^k|d} (q^M - 2) \oplus \sigma((m/j)^k d') \pmod{q} \\ &\equiv (q^M - 2) \oplus \sigma(d') \pmod{q}. \end{aligned} \tag{4}$$

This follows from the fact that for all $j^k < m^k$, we have

$$(q^M - 2) \oplus \sigma((m/j)^k d') = 0.$$

Let $d' = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ and let a be the number of e_i equal to 1 and b the number of e_i greater than 1. Then it follows that

$$\begin{aligned} \mathbf{v}_{(q^M-2)}(d') &\equiv (-1)^a \cdot 0^b \pmod{q} \\ &= \begin{cases} (-1)^a & \text{if } b = 0, \\ 0 & \text{if } b > 0. \end{cases} \end{aligned}$$

Thus, locker d will be closed if at least one exponent in the prime factorization of d is greater than 1 (mod k) and it will be open and in state $(-1)^b$ if all exponents are congruent to 0 or 1 (mod k), where b is the number of exponents that are congruent to 1 (mod k). In other words, we should select those students whose numbers are kth powers times a possibly empty product of distinct primes. In the original locker problem, where $k = 2$ and $q = 2$, this clearly includes all students.

Example 5. In this example, we answer the question of what set of students will leave open exactly the prime-numbered lockers? Let $\mathbf{v}_{-1}$ represent the lockers with each prime number open in state 1 and the others in state 0. Consider locker $d = p_1^{e_1} \cdots p_r^{e_r}$.

We have that

$$\begin{aligned}
\mathbf{v}_{q^M-2}(d) &\equiv \sum_{i=1}^{r} (q^M-2) \oplus \sigma((p_1^{e_1}\cdots p_r^{e_r})/p_i) \pmod q \\
&\equiv \sum_{i=1}^{r} (q^M-2) \oplus \sigma(p_1^{e_1}\cdots p_i^{e_i-1}\cdots p_r^{e_r}) \pmod q \\
&\equiv \sum_{i=1}^{r} (q^M-2) \oplus (\{e_1,\ldots,e_i-1,\ldots,e_r\}) \pmod q \\
&\equiv \sum_{i=1}^{r} [((q^M-2)\oplus e_1)\cdots((q^M-2)\oplus(e_i-1))\cdots((q^M-2)\oplus e_r)] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{mod } q) \\
&\equiv \begin{cases}
(-1)^{r-1}r \pmod q & \text{if all } e_i = 1, \\
(-1)^r \pmod q & \text{if one } e_i = 2, \text{ all other (if any) } e_i = 1, \\
0 \pmod q & \text{if one } e_i \geq 3, \text{ all other (if any) } e_i = 1, \\
0 \pmod q & \text{if more than one } e_i \geq 2.
\end{cases} \qquad (5)
\end{aligned}$$

Thus, the set of students that will leave open the prime numbered lockers consists of integers that are the product of r distinct primes, provided $r \not\equiv 0 \pmod q$, together with all integers that are the square of a prime times a possibly empty product of distinct other primes, with each student in the state described by congruence (5).

In the case where $N = 30$ and $q = 2$, that vector represents the set

$$\{2, 3, 4, 5, 7, 9, 11, 12, 13, 17, 18, 19, 20, 23, 25, 28, 29, 30\}$$

and when $N = 30$ and $q = 3$, the set is

$$\{2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 25, 26, 28, 29\}.$$

The locker function and Möbius inversion

As noted earlier, f is a linear transformation from the vector space $\mathbb{Z}_q^N$ to itself. Thus, f can be represented by an $N \times N$ matrix $\mathbf{T}$. By considering the action of f on the standard basis of $\mathbb{Z}_q^N$, we see that

$$\mathbf{T} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & \cdots & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1
\end{bmatrix},$$

in which column k contains a 1 in every kth position (the action of the kth student) and zeros elsewhere. Given any configuration $\mathbf{v}_n$, the subsequent configuration is $\mathbf{v}_{n+1}^\top = \mathbf{T}\mathbf{v}_n^\top$.

Note that column 1 represents the original locker problem with every student sent. If we let $\mathbf{v}_{-1} = (1, 0, 0, \ldots, 0)$, then $\mathbf{v}_n^\top = \mathbf{T}^{n+1}\mathbf{v}_{-1}^\top$, which implies that column 1 of $\mathbf{T}^{n+1}$ is congruent to $\mathbf{v}_n^\top$. It then follows from Theorem 5 that for any $n \geq 0$, $\mathbf{T}^{n+1}$ is congruent to the following matrix, modulo q:

$$\begin{bmatrix} n\oplus\sigma(1) & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ n\oplus\sigma(2) & n\oplus\sigma(1) & 0 & 0 & 0 & 0 & \cdots & 0 \\ n\oplus\sigma(3) & 0 & n\oplus\sigma(1) & 0 & 0 & 0 & \cdots & 0 \\ n\oplus\sigma(4) & n\oplus\sigma(2) & 0 & n\oplus\sigma(1) & 0 & 0 & \cdots & 0 \\ n\oplus\sigma(5) & 0 & 0 & 0 & n\oplus\sigma(1) & 0 & \cdots & 0 \\ n\oplus\sigma(6) & n\oplus\sigma(3) & n\oplus\sigma(2) & 0 & 0 & n\oplus\sigma(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n\oplus\sigma(N) & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & n\oplus\sigma(1) \end{bmatrix}.$$

It is clear that if we multiply $\mathbf{T}^{n+1}$ by $\mathbf{v}_{-1} = (c_1, c_2, c_3, \ldots, c_N)$, the result gives us equation (3):

$$\mathbf{v}_n(d) = \sum_{s|d} c_s \cdot n \oplus \sigma(d/s).$$

Before pursuing this any further, it will be beneficial to consider matrices of this general form and to relate them to the Dirichlet convolution product [4]. Let $g : \mathbb{N} \to \mathbb{Z}$ be an arithmetic function and define

$$m(g) = \begin{bmatrix} g(1) & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ g(2) & g(1) & 0 & 0 & 0 & 0 & \cdots & 0 \\ g(3) & 0 & g(1) & 0 & 0 & 0 & \cdots & 0 \\ g(4) & g(2) & 0 & g(1) & 0 & 0 & \cdots & 0 \\ g(5) & 0 & 0 & 0 & g(1) & 0 & \cdots & 0 \\ g(6) & g(3) & g(2) & 0 & 0 & g(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ g(N) & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & g(1) \end{bmatrix}.$$

If $h : \mathbb{N} \to \mathbb{Z}$ is another arithmetic function, then the (i, j)-element of the product $m(g)m(h)$ is

$$\sum_{d|n} g(d)h(n/d).$$

However, this is precisely the Dirichlet convolution product $g * h$ of g and h. Therefore,

$$m(g)m(h) = m(g * h).$$

Let us define $\varepsilon : \mathbb{N} \to \mathbb{Z}$ by $\varepsilon(1) = 1$ and $\varepsilon(n) = 0$ for all $n > 1$. Define $\mathbf{1} : \mathbb{N} \to \mathbb{Z}$ by $\mathbf{1}(n) = 1$ for all $n \in \mathbb{N}$. Then $m(\varepsilon) = \mathbf{I}$, $m(\mathbf{1}) = \mathbf{T}$, and

$$\mathbf{T}^n = m(\mathbf{1} * \mathbf{1} * \mathbf{1} * \cdots * \mathbf{1}) \quad (n \text{ copies of } \mathbf{1}).$$

It is well known that the set of arithmetic functions forms a commutative ring with unity ε under the operations of addition and the Dirichlet convolution product. The

units of this ring are those arithmetic functions g for which $g(1) \neq 0$. The inverse of $\mathbf{1}$ is the Möbius function μ, defined as

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1, \\ (-1)^r & \text{if } n \text{ is the product of } r \text{ distinct primes,} \\ 0 & \text{otherwise.} \end{cases}$$

From the equation $\mathbf{1} * \mu = \varepsilon$, it follows that

$$m(\mathbf{1})m(\mu) = m(\varepsilon) = \mathbf{I}$$

and therefore $\mathbf{T}^{-1} = m(\mu)$. The configuration that produces the primes after one iteration is given by

$$\mathbf{v}_{-1}(d) \equiv \sum_{\substack{p|d \\ p \text{ prime}}} \mu(d/p) \pmod{q}.$$

From this equation, one may quickly deduce equation (5). Similarly, one may easily obtain equation (4) from the equation

$$\mathbf{v}_{-1}(d) \equiv \sum_{j^k|d} \mu(d/j^k) \pmod{q}.$$

REFERENCES

[1] Dagal, K. (2013). Generalized locker problem. arxiv.org/pdf/1307.6455.pdf

[2] Holte, J. M. (1994). A Lucas-type theorem for Fibonomial-coefficient residues. *Fibonacci Q.* 32(1): 60–68. fq.math.ca/Scanned/32-1/holte.pdf

[3] Kimani, P. M., Olanoff, D., Masingila, J. O. (2016). The locker problem: an open and shut case. *Math. Teach. Middle School.* 22(3): 144–151. researchgate.net/publication/308921853_The_Locker_Problem_An_Open_and_Shut_Case

[4] Niven, I., Zuckerman, H. (1972). *An Introduction to the Theory of Numbers*. New York, NY: Wiley.

[5] Torrence, B., Wagon, S. (2007). The locker problem. *Crux Math*. 33(4): 232–236.

Summary. The well-known Locker Problem begins with N students assigned to N closed lockers. For each integer k from 1 to N, student k is instructed to reverse the state (either open or closed) of every kth locker door, beginning with locker k. The question is, once this is done, which lockers will be open? In this article, we extend the Locker Problem in two ways. Rather than the lockers being simply open or closed, we consider doors that are in any of q states, for some prime q, where only state 0 is closed and states $1, 2, \ldots, q-1$ are open in varying degrees. We also investigate what happens when we iterate the procedure, sending on each iteration every student whose own locker was left open in the previous iteration. We develop an explicit formula that describes the state of any locker after any number of iterations. Then we address the inverse question: given a desired configuration of lockers, each in a desired degree of openness, which students should be sent in order to leave the lockers in those states?

REBECCA L. JAYNE received her B.A. from McDaniel College and her M.S. and Ph.D. from North Carolina State University. She is now an Elliott Associate Professor of Mathematics at Hampden-Sydney College and lives nearby with her husband and two children.

ROBB T. KOETHER earned his Ph.D. in commutative algebra from the University of Oklahoma in 1978. He is currently retiring from Hampden-Sydney College in Virginia after teaching there for 39 years. In his spare time, he enjoys backpacking on the Appalachian Trail.

A Simple Proof of Kooi's Inequality

MARTIN LUKAREVSKI
University "Goce Delcev"
Stip, North Macedonia
martin.lukarevski@ugd.edu.mk

For a triangle with semiperimeter s, circumradius R, and inradius r, Kooi's inequality [1, Section 5.7], [2],

$$s^2 \leq \frac{R(4R+r)^2}{2(2R-r)}, \tag{1}$$

has important applications in the theory of triangle inequalities. We give a novel proof using only elementary algebra, based on the following well-known triangle inequality [1, Section 5.10]:

$$s^2 \leq 2R^2 + 10Rr - r^2 + 2(R-2r)\sqrt{R(R-2r)}. \tag{2}$$

By a simple algebraic manipulation, we observe that

$$\begin{aligned} &2(2R-r)(2R^2+10Rr-r^2) \\ &\qquad = R(4R+r)^2 - 4R(R-2r)^2 - (R-2r)(2R-r)^2. \end{aligned}$$

This identity and a skillful rearrangement yields

$$\begin{aligned} &2R^2+10Rr-r^2+2(R-2r)\sqrt{R(R-2r)} \\ &\quad = \frac{R(4R+r)^2}{2(2R-r)} - \left((R-2r)\sqrt{\frac{2R}{2R-r}} - \sqrt{\frac{(R-2r)(2R-r)}{2}}\right)^2. \end{aligned} \tag{3}$$

This expression is clearly smaller than

$$\frac{R(4R+r)^2}{2(2R-r)}.$$

Hence, by inequality (2) and equation (3), we obtain Kooi's inequality.

REFERENCES

[1] Bottema, O., Djordjevic, R. Z., Janic, R. R., Mitrinovic, D. S., Vasic, P. M. (1969). *Geometric Inequalities*. Groningen: Wolters-Noordhoff.

[2] Lukarevski, M., Marinescu, D. S. (2019). A refinement of the Kooi's inequality, Mittenpunkt and applications. *J. Inequal. Appl.* 13(3): 827–832. doi.org/10.7153/jmi-2019

Summary. We provide a simple, original proof of Kooi's inequality from Euclidean geometry.

MARTIN LUKAREVSKI earned his Ph.D. from Leibniz Universität Hannover, Germany, and currently he is an associate professor at University "Goce Delcev" (North Macedonia). He enjoys teaching mathematics, and his interests include classical geometry and geometric inequalities, analysis, and history of mathematics.

Math. Mag. **93** (2020) 225. doi:10.1080/0025570X.2020.1736875
MSC: 51M16

PROOFS WITHOUT WORDS

The Number of Bricks in a Ziggurat

BEN BLUMSON
National University of Singapore
Singapore 119077
benblumson@nus.edu.sg

JARINAH JABBAR
CAE Malaysia
47160 Puchong, Selangor
jarinah@cae.my

Theorem 1. *The number of square bricks in a hollow ziggurat n stories high and of base width n is $n^2 + (n-1)^2$.*

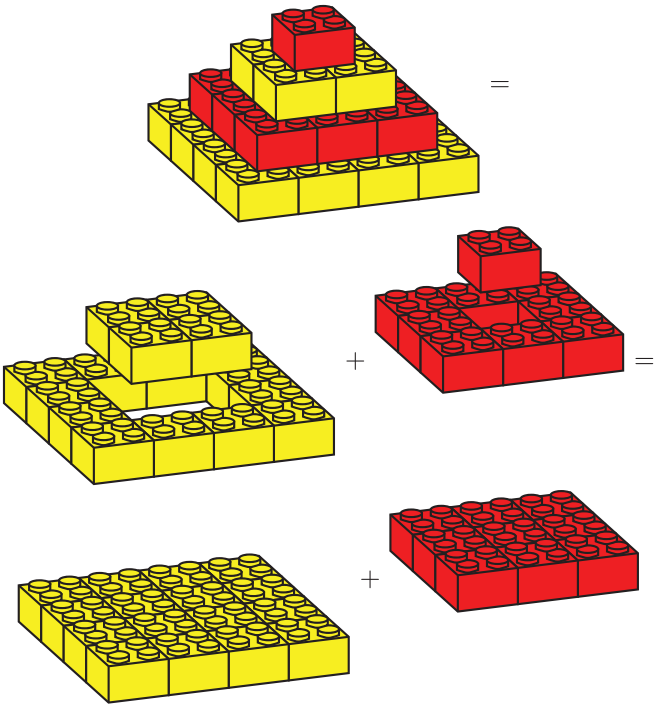
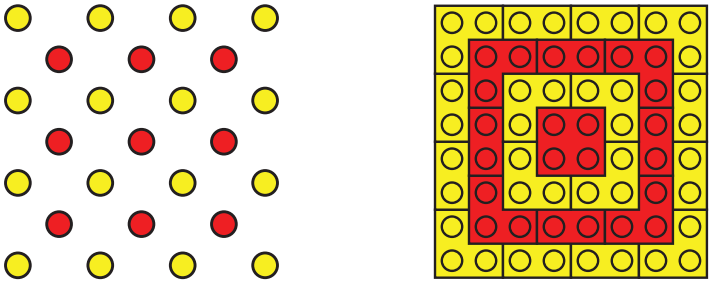


Figure 1

Remark. Recall that a *centered square* number is one that can be formed by placing one dot to serve as a center, and then by surrounding that center with square layers. Figure 2a is a well-known visual proof that such a centered square is the sum of

Math. Mag. **93** (2020) 226–227. doi:10.1080/0025570X.2020.1720495
MSC: 51E99
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

consecutive squares, as discussed in Conway and Guy [**1**, 41-42] and Deza and Deza [**2**, 54]. (See also sequence A001844 in the *Online Encyclopedia of Integer Sequences* [**3**]). Comparing it with Figure 2b of a ziggurat from above provides another proof of the theorem.



(a) (b)

Figure 2

Acknowledgments We would like to thank an anonymous referee, Jeremiah Joaquin, Michael Pelczar, and Weng-Hong Tang for comments on this paper. Figure 1 was produced using code adapted from github.com/cryingshadow/lego.

REFERENCES

[1] Conway, J. H., Guy, R. K. (1996). *The Book of Numbers*. New York: Springer.
[2] Deza, E., Deza, M. M. (2012). *Figurate Numbers*. Singapore: World Scientific.
[3] Sloane, N. J. A. (1964). Sequence A001844. *The On-Line Encyclopedia of Integer Sequences*.

Summary. The number of bricks in a ziggurat is a sum of consecutive squares.

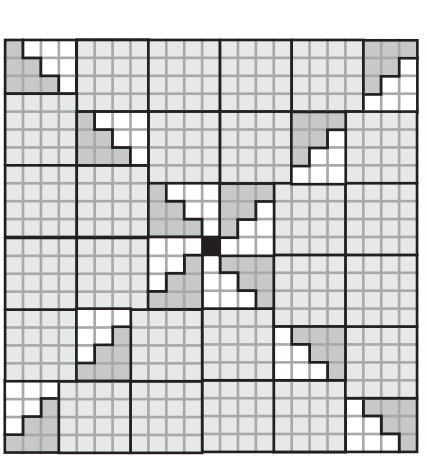
BEN BLUMSON is an Associate Professor in Philosophy at the National University of Singapore. He is the author of *Resemblance and Representation: An Essay in the Philosophy of Pictures* (Cambridge, Open Book Publishers, 2014).

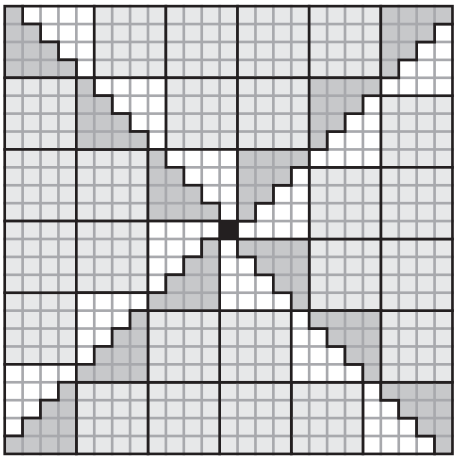
JARINAH JABBAR is Programme Director at CAE Malaysia in Kuala Lumpur. She is interested in special needs education and play therapy.

Squared Primes Modulo 24

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

Let p be a prime greater than or equal to 5. Then p equals $6n-1$ or $6n+1$ for some positive integer n. But the square of such a number is 1 more than a multiple of 24, as the following figures illustrate. Here, T_k denotes the triangular number $1+2+\cdots+k$ for $k \geq 1$ and $T_0 = 0$.





The left and right, respectively, illustrate the identities

$$(6n-1)^2 = 24n^2 + 24T_{n-1} + 1 \quad \text{and} \quad (6n+1)^2 = 24n^2 + 24T_n + 1.$$

Hence, we have the following theorem:

Theorem. *If $p \geq 5$ is prime, then $p^2 \equiv 1 \pmod{24}$.*

The theorem clearly also holds for composite p odd and not a multiple 3. That is, for $p \equiv \pm 1 \pmod{6}$.

Summary. We show visually that the square of a prime greater than or equal to 5 is congruent to 1 modulo 24.

ROGER B. NELSEN (MR Author ID 237909) is a professor emeritus at Lewis & Clark College, where he taught mathematics and statistics for 40 years.

Math. Mag. **93** (2020) 228. doi:10.1080/0025570X.2020.1736877
MSC: 97F60, 11A07, 11A41

PROBLEMS

LES REID, *Editor*
Missouri State University

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

RICHARD BELSHOFF, Missouri State University; *EYVINDUR ARI PALSSON*, Virginia Tech; *CODY PATTERSON*, Texas State University; *ROGELIO VALDEZ*, Centro de Investigación en Ciencias, UAEM, Mexico; *Assistant Editors*

Proposals

To be considered for publication, solutions should be received by November 1, 2020.

2096. *Proposed by H. A. ShahAli, Tehran, Iran.*

Any three distinct vertices of a polytope P form a triangle. How many of these triangles are isosceles if P is

(a) a regular n-gon?

(b) one of the Platonic solids?

(c) an n-dimensional cube?

2097. *Proposed by Omran Kouba, Higher Institute for Applied Sciences and Technology, Damascus, Syria.*

For a real number $x \notin \frac{1}{2} + \mathbb{Z}$, denote the nearest integer to x by $\langle x \rangle$. For any real number x, denote the largest integer smaller than or equal to x and the smallest integer larger than or equal to x by $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively. For a positive integer n let

$$a_n = \frac{2}{\langle \sqrt{n} \rangle} - \frac{1}{\lfloor \sqrt{n} \rfloor} - \frac{1}{\lceil \sqrt{n} \rceil}.$$

(a) Prove that the series $\sum_{n=1}^{\infty} a_n$ is convergent and find its sum L.

(b) Prove that the set

$$\left\{ \sqrt{n} \left(\sum_{k=1}^{n} a_k - L \right) : n \geq 1 \right\}$$

is dense in $[0, 1]$.

Math. Mag. **93** (2020) 229–238. doi:10.1080/0025570X.2020.1742543.

We invite readers to submit original problems appealing to students and teachers of advanced undergraduate mathematics. Proposals must always be accompanied by a solution and any relevant bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. Submitted problems should not be under consideration for publication elsewhere.

Proposals and solutions should be written in a style appropriate for this MAGAZINE.

Authors of proposals and solutions should send their contributions using the Magazine's submissions system hosted at http://mathematicsmagazine.submittable.com. More detailed instructions are available there. We encourage submissions in PDF format, ideally accompanied by LaTeX source. General inquiries to the editors should be sent to mathmagproblems@maa.org.

2098. *Proposed by Albert Natian, Los Angeles Valley College, Valley Glen, CA.*

Let $Z_0 = 0$, $Z_1 = 1$, and recursively define random variables $Z_2, Z_3, \ldots,$ taking values in $[0, 1]$ as follows: For each positive integer k, Z_{2k} is chosen uniformly in $[Z_{2k-2}, Z_{2k-1}]$ and Z_{2k+1} is chosen uniformly in $[Z_{2k}, Z_{2k-1}]$.

Prove that, with probability 1, the limit $Z^* = \lim_{n\to\infty} Z_n$ exists and find its distribution.

2099. *Proposed by Russ Gordon, Whitman College, Walla Walla, WA and George Stoica, Saint John, NB, Canada.*

Let r and s be distinct nonzero rational numbers. Find all functions $f : \mathbb{R} \to \mathbb{R}$ that satisfy

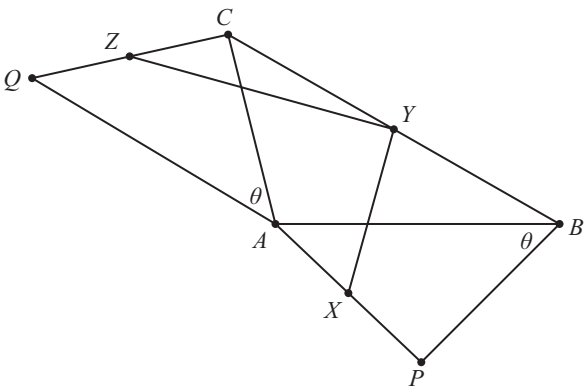
$$f\left(\frac{x+y}{r}\right) = \frac{f(x)+f(y)}{s}$$

for all real numbers x and y.

2100. *Proposed by Yevgenya Movshovich and John E. Wetzel, University of Illinois, Urbana, IL.*

Given $\triangle ABC$ and an angle θ, two congruent triangles $\triangle ABP$ and $\triangle QAC$ are constructed as follows: $AQ = AB$, $BP = AC$, $\mathrm{m}\angle ABP = \mathrm{m}\angle CAQ = \theta$, B and Q are on opposite sides of $\overleftrightarrow{AC}$, and C and P are on opposite sides of $\overleftrightarrow{AB}$, as shown in the figure. Let X, Y, and Z be the midpoints of segments AP, BC, and CQ, respectively.

Show that $\angle XYZ$ is a right angle.



Quickies

1101. *Proposed by Robert Calcaterra, University of Wisconsin, Platteville, WI.*

It is well known that it is impossible to square a circle using just a straightedge and compass. In other words, given a circle, it is impossible to construct a square having the same area as the given circle.

Given an arbitrary polygon, is it possible to construct a square with the same area using only straightedge and compass?

1102. *Proposed by Ovidiu Furdui and Alina Sîntămărian, Technical University of Cluj-Napoca, Cluj-Napoca, Romania.*

Let $n \geq 1$ be an integer. Calculate

$$I_n = \int_0^1 \frac{\ln(1-x) + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n}}{x^{n+1}}\, dx.$$

Solutions

Units in a familiar ring with unfamiliar multiplication — June 2019

2071. *Proposed by Ioan Băetu, Botoşani, Romania.*

Let $n > 1$ be an integer, and let $\mathbb{Z}_n$ be the ring of integers modulo n. For fixed $k \in \mathbb{Z}_n - \{0\}$, define a binary operation "$\circ$" on $\mathbb{Z}_n$ by $x \circ y = (x-k)(y-k) + k$ for all $x, y \in \mathbb{Z}_n$. Let U be the group of units of $\mathbb{Z}_n$ (under multiplication), and let U_k° be the set of elements of $\mathbb{Z}_n$ that are invertible under the operation $\circ$. Characterize those n with the property that $U \neq U_k^\circ$ for all $k \in \mathbb{Z}_n - \{0\}$.

Solution by Missouri State University Problem Solving Group, Missouri State University, Springfield, MO.

We claim that n has the desired property if and only if it is square-free.

First note that $x \circ (1+k) = x$ for all $x \in \mathbb{Z}_n$, so $1+k$ is the identity for $\circ$. If $u \in U$, then

$$(k+u) \circ (k+u^{-1}) = uu^{-1} + k = 1 + k,$$

so

$$k + U = \{k + u \mid u \in U\} \subseteq U_k^\circ.$$

Conversely, if $v \in U_k^\circ$, then there is a $w \in \mathbb{Z}_n$ with

$$1 + k = v \circ w = (v-k)(w-k) + k$$

and thus $1 = (v-k)(w-k)$. So $v - k \in U$ and hence $U_k^\circ = k + U$.

Suppose n is not square-free and let k be the product of all the distinct prime factors of n, so $k \in \mathbb{Z}_n - \{0\}$. If a is any integer, then $\gcd(a, n) = 1$ if and only if $\gcd(k + a, n) = 1$. It follows that $U = k + U = U_k^\circ$ and hence n does not have the desired property.

Suppose now that n is square-free. If $k \in \mathbb{Z}_n - \{0\}$ then at least one of the prime factors of n does not divide k; let q be the product of all the prime factors of n that do not divide k. Then $\gcd(q-k, n) = 1$ and hence $q - k \in U$. But then

$$q = k + (q-k) \in k + U = U_k^\circ,$$

and since $q \notin U$ then we have $U \neq U_k^\circ$ for all $k \in \mathbb{Z}_n - \{0\}$. So n has the desired property.

Also solved by Hafez Al-Assad (Syria), Anthony J. Bevelacqua, Elton Bojaxhiu (Germany) and Enkel Hysnelaj (Australia), Robert Calcaterra, Ali Deeb (Syria), Briana Foster-Greenwood, Tom Jager, Peter McPolin (Northern Ireland), and the proposer. There was one incomplete or incorrect solution.

Two initial value problems **June 2019**

2072. *Proposed by Julien Sorel, Piatra Neamt, PNI, Romania.*

(a) Show that the initial value problem

$$\begin{cases} y' = \sqrt{1 - y^2}, \\ y(0) = 1 \end{cases}$$

has infinitely many solutions defined on $\mathbb{R}$.

(b) By contrast, show that the initial value problem

$$\begin{cases} y' = \sqrt{x^2 - y^2}, \\ y(1) = 1 \end{cases}$$

has no solutions defined on an open interval containing $x = 1$.

Solution by Ali Deeb (student) and Hafez Al-Assad (student), Higher Institute for Applied Sciences and Technology, Damascus, Syria.

(a) First note that any solution with $y(b) = 1$ must have $y(x) = 1$ for all $x \geq b$. This follows from the fact that y is nondecreasing, since $y' \geq 0$, and that $-1 \leq y \leq 1$. Similarly, if $y(a) = -1$, then $y(x) = -1$ for all $x \leq a$.

For any $k \geq \pi/2$, let

$$y(x) = \begin{cases} -1 & x < -k - \frac{\pi}{2} \\ \sin(x + k) & -k - \frac{\pi}{2} \leq x < -k + \frac{\pi}{2} \\ 1 & x \geq -k + \frac{\pi}{2} \end{cases}$$

It is straightforward to check that y is continuous, differentiable, and satisfies the given differential equation.

(b) Suppose to the contrary that such a $y(x)$ exists. For x close to 1, $y(x)$ is also close to 1. In particular, they are both positive. The condition $y' = \sqrt{x^2 - y^2}$ forces $x \geq y(x) > 0$. Consider the function

$$g(x) = \frac{1 - y(x)}{1 - x}$$

defined for $x < 1$ and x close to 1. We have

$$\lim_{x \to 1^-} g(x) = y'(1) = \sqrt{1^2 - y(1)^2} = 0.$$

In particular, $g(x) < 1$ for x sufficiently close to 1. But this gives

$$\frac{1-y(x)}{1-x} < 1 \Rightarrow 1 - y(x) < 1 - x \Rightarrow y(x) > x$$

resulting in a contradiction.

Also solved by Yagub Aliyev (Azerbaijan), Michel Bataille (France), Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), David M. Bradley, Robert Calcaterra, Bruce E. Davis, John N. Fitch, Lixing Han, Eugene A. Herman, Tom Jager, Kee-Wai Lau (Hong Kong), Albert Natian, José Nieto (Venezuela), Northwestern University Math Problem Solving Group, Sonebi Omar (Morocco), Sung Hee Park (South Korea), Francisco Perdomo & Ángel Plaza (Spain), Edward Schmeichel, Nicholas C. Singer, Lawrence R. Weill, Xinyi Zhang (Canada), and the proposer. There were two incomplete or incorrect solutions.

Unambiguous factorial expansions **June 2019**

2073. *Proposed by Enrique Treviño, Lake Forest College, Lake Forest, IL.*

A *factorial expansion* is any formal expression of the form

$$\overline{a_k a_{k-1} \ldots a_2 a_1}$$

where $a_1, a_2, \ldots, a_k$ are k integers ($k \geq 1$) such that $0 \leq a_i \leq i$ for $i = 1, 2, \ldots, k$. The *value* of such a factorial expansion is

$$a_k \cdot k! + a_{k-1} \cdot (k-1)! + \cdots + a_2 \cdot 2! + a_1 \cdot 1!.$$

If the integers $a_1, \ldots, a_k$ are expressed in base 10 and their digits simply written together without separation, the value of the factorial expansion so written is often ambiguous. For instance, the expansion $\overline{10000000000}$ may be interpreted as having coefficients 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 and value $1 \times 11! + 0 \times (10! + 9! + \ldots + 1!) = 11!$, or having coefficients 10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 and value $10 \times 10! + 0 \times (9! + 8! + 7! + \cdots + 1!) = 10 \times 10!$. Such factorial expansions are called *ambiguous.* On the other hand, some factorial expansions are unambiguous: for example, the expansion $\overline{311}$ must have the value $3 \times 3! + 1 \times 2! + 1 \times 1! = 21$. Prove that there are only finitely many unambiguous factorial expansions, and find the one whose value is largest.

Solution by José Heber Nieto, Universidad del Zulia, Maracaibo, Venezuela.
Let us call the *length* of a factorial expansion $\overline{a_k a_{k-1} \ldots a_2 a_1}$ the total number of decimal digits in the a_i's. We claim that any factorial expansion with length greater than 99 is ambiguous. Indeed, let $d_n d_{n-1} \ldots d_1$ be the sequence of digits in such an expansion. Then it may be interpreted as

$$\overline{d_n d_{n-1} \ldots d_1} \text{ or as } \overline{(10d_n + d_{n-1}) d_{n-2} \ldots d_1},$$

and the first has greater value than the second. Therefore there are only finitely many unambiguous factorial expansions.

We claim that the unambiguous factorial expansion whose value is largest is

$$m = \overline{\underbrace{99\ldots99}_{91\text{ nines}}87654321},$$

whose value is

$$M = \sum_{k=1}^{8} k \cdot k! + 9\sum_{k=9}^{99} k!.$$

Indeed, the value of a factorial expansion $\overline{a_k a_{k-1}\ldots a_2 a_1}$ is

$$\sum_{j=1}^{k} a_k \cdot k! \le \sum_{j=1}^{k} k \cdot k! = (k+1)! - 1,$$

thus if $k < 99$ its value is less than 99! and less than M. If $k \ge 99$, to be unambiguous we must have $k = 99$ and all the a_i's must be digits. Then the greatest value is clearly attained with m.

Also solved by Hafez Al-Assad (Syria), Robert Calcaterra, Ali Deeb (Syria), Kelly Jahns, Vasile Teodorovici (Canada), and the proposer.

Zeta(2) in disguise **June 2019**

2074. *Proposed by Bao Do (student), Columbus State University, Columbus, GA.*

Evaluate

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k}\binom{n}{k} H_k,$$

where $H_k = \displaystyle\sum_{j=1}^{k} \frac{1}{j}$ is the kth harmonic sum.

Solution by Ulrich Abel and Vitaliy Kushnirevych, Technische Hochschule Mittelhessen, Friedberg, Germany.
Put

$$S_n := \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k}\binom{n}{k} H_k.$$

We have

$$\begin{aligned} S_n &= \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k}\left[\binom{n-1}{k} + \binom{n-1}{k-1}\right] H_k \\ &= S_{n-1} + \frac{1}{n}\sum_{k=1}^{n} (-1)^{k+1}\binom{n}{k} H_k \end{aligned}$$

and

$$\begin{aligned} T_n &:= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} H_k \\ &= \sum_{k=1}^{n} (-1)^{k+1} \left[\binom{n-1}{k} + \binom{n-1}{k-1} \right] H_k \\ &= T_{n-1} + \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} H_{k+1} \\ &= T_{n-1} - T_{n-1} + \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \frac{1}{k+1} \\ &= \frac{1}{n} \sum_{k=0}^{n-1} (-1)^k \binom{n}{k+1} \\ &= \frac{1}{n}. \end{aligned}$$

The recursive formula $S_n = S_{n-1} + n^{-2}$ and $S_1 = 1$ imply $S_n = \sum_{k=1}^{n} k^{-2}$, therefore $\lim_{n\to\infty} S_n = \pi^2/6$.

Also solved by Michel Bataille (France), Khristo Boyadzhiev, Brian Bradie, David Bradly, Hongwei Chen, Robert Doucette, GWstat Problem Solving Group, Lixing Han, Tom Jager, Walther Janous (Austria), Dixon Jones, Albert Natian, José Nieto (Venezuela), Chikanna Selvaraj, Nicholas Singer, Albert Stadler (Switzerland), Seán Stewart (Australia), Michael Vowe (Switzerland), and the proposer. There was one incomplete or incorrect solution.

A recursively defined sequence of tetrahedra **June 2019**

2075. *Proposed by Michael Goldenberg, The Ingenuity Project, Baltimore Polytechnic Institute, Baltimore, MD and Mark Kaplan, Towson University, Towson, MD.*

Consider the sequence $\{C_n\}$ defined recursively by $C_0 = 3$, $C_1 = 1$, $C_2 = 3$, and

$$C_n = C_{n-1} + C_{n-2} + C_{n-3} \qquad \text{for } n \geq 3.$$

Let $O = (0, 0, 0)$ be the origin of $\mathbb{R}^3$ and, for integer $n \geq 0$, let P_n be the point (C_n, C_{n+1}, C_{n+2}).

(a) Find the volume of the pyramid $O P_n P_{n+1} P_{n+2}$ in closed form.
(b) Show that the sequence $\{P_n\}$ asymptotically approaches a fixed line $\mathcal{L}$ through the origin of $\mathbb{R}^3$, and characterize this line.

Solution by Brandon Cho (student), The Nueva School, San Mateo, CA.

(a) If we record the nonzero coordinates of tetrahedron $O P_n P_{n+1} P_{n+2}$ in the rows of the matrix

$$T_n = \begin{pmatrix} C_n & C_{n+1} & C_{n+2} \\ C_{n+1} & C_{n+2} & C_{n+3} \\ C_{n+2} & C_{n+3} & C_{n+4} \end{pmatrix},$$

then the recursion

$$T_0 = \begin{pmatrix} 3 & 1 & 3 \\ 1 & 3 & 7 \\ 3 & 7 & 11 \end{pmatrix}, \quad T_{n+1} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} T_n, \quad n \geq 0,$$

enables us to generate the nonzero coordinates of successive tetrahedra. Since

$$\det T_{n+1} = \det \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \det T_n = \det T_n,$$

$|\det T_n| = |\det T_0| = 44$ by induction. The volume of a tetrahedron formed by three vectors (a_1, a_2, a_3), (b_1, b_2, b_3), and (c_1, c_2, c_3) that are coterminal at the origin is

$$\frac{1}{6} \left| \det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} \right|.$$

Therefore, the volume of tetrahedron $O P_n P_{n+1} P_{n+2}$ is $\frac{1}{6} |\det T_n| = \frac{44}{6} = \frac{22}{3}$ for all $n \geq 0$.

(b) The characteristic equation for $\{C_n\}$ is

$$r^3 - r^2 - r - 1 = 0,$$

which has one real root $t \approx 1.839$ (the exact value can be determined by the cubic formula, but is not needed here) and two complex roots β and $\overline{\beta}$, each of whose modulus is $1/\sqrt{t} < 1$. Thus, for some constants k_1, k_2, and k_3,

$$C_n = k_1 t^n + k_2 \beta^n + k_3 \overline{\beta}^n, \quad n \geq 0.$$

We claim that $k_1 = k_2 = k_3 = 1$. To see this, let

$$D_n = t^n + \beta^n + \overline{\beta}^n.$$

Note that $D_0 = 3$ and by Vieta's relations

$$\begin{aligned} D_1 &= t + \beta + \overline{\beta} = 1 \text{ and} \\ D_2 &= \left(t + \beta + \overline{\beta}\right)^2 - 2\left(t\beta + t\overline{\beta} + \beta\overline{\beta}\right) \\ &= 1^2 - 2(-1) = 3. \end{aligned}$$

The initial conditions are satisfied, so $C_n = D_n$ for all $n \geq 0$.

Since

$$(C_n, C_{n+1}, C_{n+2}) - C_n(1, t, t^2) =$$

$$\left(0, \beta^n(\beta - t) + \overline{\beta}^n\left(\overline{\beta} - t\right), \beta^n\left(\beta^2 - t^2\right) + \overline{\beta}^n\left(\overline{\beta}^2 - t^2\right)\right)$$

and

$$\lim_{n\to\infty} \beta^n = \lim_{n\to\infty} \overline{\beta}^n = 0,$$

the P_n asymptotically approach the line through the origin with direction vector $(1, t, t^2)$.

Also solved by Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Sarah Brickman (student), Robert Calcaterra, Robin Chapman (UK), George Washington University Problems Group, Eugene A. Herman, Tom Jager, Vitaliy Kushnirevych & Ulrich Abel (Germany), Harris Kwong, José H. Nieto (Venezuela), Jacob Siehler, Lawrence R. Weill, and the proposers. There were four incomplete or incorrect solutions.

Answers

Solutions to the Quickies from page 230.

A1101. The answer is yes.

Given a set

$$S = \{\alpha_1, \ldots, \alpha_n\} \subset \mathbb{R},$$

we say β is *S-constructible* if we can construct a segment of length β from segments having lengths $\alpha_1, \ldots, \alpha_n$ using only straightedge and compass. It is well known that if β_1 and β_2 are S-constructible, so are $\beta_1 + \beta_2$, $\beta_1\beta_2$, and β_1/n for any positive integer n. From these facts it is easy to see that $f(\alpha_1, \ldots, \alpha_n)$ is S-constructible for any polynomial function f with rational coefficients. Also, if β is S-constructible, so is $\sqrt{\beta}$.

Denote the area of a polygon P by $A(P)$. Given an arbitrary polygon P, we want to show that $\sqrt{A(P)}$ is S-constructible, where S is the set of all side lengths and diagonal lengths of P. It is well known that P can be triangulated using diagonals. This gives

$$A(P) = \sum A(T_i),$$

where the T_i are the triangles in the triangulation of P. By Heron's formula

$$A(T_i) = \frac{\sqrt{2a_i^2b_i^2 + 2b_i^2c_i^2 + 2a_i^2c_i^2 - a_i^4 - b_i^4 - c_i^4}}{4},$$

where a_i, b_i, c_i are the sidelengths of T_i. Now $\{a_i, b_i, c_i\} \subseteq S$, so the $A(T_i)$ are S-constructible. Therefore $A(P)$ is S-constructible and finally $\sqrt{A(P)}$ is S-constructible.

Editor's Note. (*i*) The negative answer to the problem of duplicating the cube shows that one cannot "cube" an arbitrary polyhedron.

(*ii*) Similar arguments to those above show that using only straightedge and compass, one can construct a hypercube with the same 4-content as an arbitrary polytope.

A1102. We have

$$\begin{aligned} I_n &= -\int_0^1 \frac{1}{x^{n+1}} \sum_{i=n+1}^{\infty} \frac{x^i}{i}\, dx \\ &= -\int_0^1 \sum_{i=n+1}^{\infty} \frac{x^{i-n-1}}{i}\, dx \\ &\overset{(*)}{=} -\sum_{i=n+1}^{\infty} \frac{1}{i(i-n)} \end{aligned}$$

$$= -\sum_{j=1}^{\infty} \frac{1}{j(j+n)}$$

$$= -\frac{1}{n}\sum_{j=1}^{\infty}\left(\frac{1}{j} - \frac{1}{j+n}\right)$$

$$= -\frac{1}{n}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)$$

$$= -\frac{H_n}{n}.$$

Step (*) is justified since the power series has nonnegative terms.

Correction to Solution 2062

There is a typographical error in the solution to Problem 2062 in the February 2020 issue of THE MAGAZINE. The inequality in the last sentence should be $n > 10^{100}$. We thank Stan Wagon for pointing this out. He also notes that the *On-Line Encyclopedia of Integer Sequences* gives all 13 solutions to the problem:

28263827, 35000000, 242463827, 500000000, 528263827, 535000000, 10000000000,

10028263827, 10035000000, 10242463827, 10500000000, 10528263827,

10535000000

See http://oeis.org/A101639.

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.

Su, Francis, *Mathematics for Human Flourishing*, Yale University Press, 2020; x+174 pp, $26. ISBN 978-0-300-23713-9.

This inspirational book is an expansion of Su's article of the same title (*American Mathematical Monthly* 124 (December 2017) 483–493) from his farewell address as President of the MAA. Su "grounds mathematics in what it means to be a human being and live a more fully human life." The book's chapters are organized around a dozen basic human desires . . .

> . . . whose fulfillment is a sign of human flourishing . . . I illustrate how the pursuit of mathematics can meet this desire, and I illuminate the virtues that are cultivated by engaging in math in this way. . . . When some people ask, 'When am I ever going to use this?' what they are really asking is 'When am I ever going to value this?' They're equating math's value with utility because they haven't seen that they can value anything more than its usefulness. A grander, more purposeful vision of mathematics would tap into the desires that can entice us to do mathematics as well as the virtues that mathematics can build.

Su returns us to a long-lost view of education—in mathematics or in any subject—as not solely job training but as *character building*. He enumerates more than 50 virtues that the pursuit of mathematics, by a "playful math explorer," can develop in association with human desires. "But does the proper practice of mathematics build *particular* virtues, like the ability to think clearly and to reason well? Unequivocally yes, and it may do so in a distinctive way." Included in the book are excerpts from Su's correspondence with a prison inmate with a thirst for mathematics. This is a book for everyone who wants to know where and how mathematics can fit into their lives. "Believe that you and every person in your life can flourish in mathematics."

Higham, Nicholas J., *Handbook of Writing for the Mathematical Sciences*, 3rd ed., SIAM, 2020; xxi+353 pp, $59, SIAM member $43.80. ISBN 978-1-611976-09-0.

The preceding edition of this book was in 1998, and much has happened since then to affect publication of mathematics. In addition to general principles of writing and their application to writing about mathematics, author Higham, like authors of other writing manuals, devotes chapters to usage, reference materials and style guides, English as a foreign language, and how to write a paper. Additional chapters treat TEX, LATEX, and BIBTEX; workflow; getting a paper published; giving a talk; preparing a poster; defending a thesis; writing a book; and even writing a blog. This is a rich resource for anyone trying to communicate mathematics.

Morris, Rebecca Lea, Motivated proofs: What they are, why they matter and how to write them, to appear in *The Review of Symbolic Logic*, arxiv.org/abs/2001.02657.

What are the qualities of a good proof? In addition to "explanatory power, depth, purity, beauty and fit," author Morris urges that they be motivated—that is, have no *deus ex machina* steps. The benefit of such a proof is greater understanding of the argument, plus the potential for further discoveries. Morris starts from discussions and examples by Pólya and identifies the key qualities of a motivated proof: The origin and intended tasks of each step are clear. She offers case studies of the Cauchy-Schwarz and general arithmetic-geometric mean inequalities, followed by suggestions on how to write a motivated proof.

Math. Mag. **93** (2020) 239–240. doi:10.1080/0025570X.2020.1742544

Larson, Don, Kristen Mazur, David White, and Carolyn Yarnall, The User's Guide Project: Looking back and looking forward, *Journal of Humanistic Mathematics* 10 (1) (January 2020) 411–430, scholarship.claremont.edu/jhm/vol10/iss1/23.

Authors of research papers are constrained to write concisely in a specific style. The authors of this paper recount the history, impact, and value of a project by algebraic topologists to write "user's guides" to accompany their research papers, to explain underlying intuitions and how to think about the content and results. Ideally, such a user's guide includes key insights, offers metaphors and imagery, recounts the story of development of the paper, and gives a colloquial summary. The project itself (2014–2017) died when the founder left academia; too bad! The profession would be better off if the value system of mathematical research rewarded such efforts, to the point that every research paper were accompanied by a user's guide.

Ouellette, Jennifer, Letting slower passengers board airplane first really is faster, study finds, arstechnica.com/science/2020/01/letting-slower-passengers-board-airplane-first-really-is-faster-study-finds/.

Erland, Sveinung, Jevgenijs Kaupužs, Vidar Frette, Rami Pugatch, and Eitan Bachmat. Lorentzian-geometry-based analysis of airplane boarding policies highlights "slow passengers first" as better. *Physical Review E* 100 (2019) 062313, journals.aps.org/pre/pdf/10.1103/PhysRevE.100.062313.

Previous research has shown that boarding airplane passengers in waves according to seat location (e.g., making passengers to be seated two rows apart be next to each other in the boarding line), or according to amount of luggage carried, can shorten boarding time. This new study shows that "it's better to get the slow people out of the way first and then let the fast people trickle in." The key to this counterintuitive conclusion is parallelism: More than one person sitting down at the same time. Desirable as it would be to airlines to minimize boarding time, would such plans work? First-class and frequent flyers expect to board first, there can be intense competition for overhead bin space, and "the fast people" watching the "slowpokes" are likely to get very impatient.

Devlin, Keith, What firefighting, military tactics, cancer treatment, and teenage parties can tell us about learning math, mathvalues.org/masterblog/what-firefighting-military-tactics-cancer-treatment-and-teenage-parties-can-tell-us-about-leaning-math [sic].

Campbell, Paul J., Does mathematics teach how to think?, in *The Best Writing on Mathematics 2019*, edited by Mircea Pitici, 27–42. Princeton, NJ: Princeton University Press, 2019.

Mathematician Devlin has the ultimate answer to those who reject algebra because they have never used it: It doesn't matter. "The chances of anyone who finds they need to make use of mathematics at some point in their life or career being able to use any specific high school curriculum method is close to zero." He asserts that mathematics education has to move from practicing algorithms to developing mathematical thinking, for which "*what* is taught is not . . . of any significance." Devlin, from his background in cognitive science, asserts that the necessary "deep learning" that mathematics should teach—flexible thinking required to tackle new problems—cannot be taught as a set of rules but can be acquired only through "prolonged engagement with *some* specific mathematics topics." Flexible thinking means being able to start from examples, recognize an underlying abstract structure, and then apply that structure to novel scenarios—without having to be primed to make the connection. You certainly have known people who cannot solve arithmetic problems in the abstract but can do so if a problem is couched in terms of money. Devlin gives examples of when, how, and why—and why not—people were able to move from real-world scenarios (military, medical, fire-fighting, age-checking) to realize an underlying structure and make the *transfer*—that is, apply the structure to a totally different scenario. Can Devlin's mission for mathematics be achieved in a society that regards all mathematics as just computational skills? Does (can) mathematics teach (flexible) thinking? For Francis Su, whose book I review above, the answer is an emphatic yes to the latter question. Does mathematics do so, as Su hints, "in a distinctive way"? Despite a paucity of evidence (see my own article cited above), I like to think (flexibly) that the answer is indeed yes.