LETTER FROM THE EDITOR

You have probably noticed something strange about this issue of *Mathematics Magazine*. It is twice as long as normal!

The magazine receives a large number of excellent submissions, and this has led to an extensive backlog of accepted articles awaiting publication. Consequently, some authors have had to wait an unacceptably long period of time to see their work in print. Since the flood of quality articles shows little sign of abating, we explored the only other solution—more pages! This and the December issue of the magazine will be full double issues. The first few issues of 2023 are likely to be longer than usual as well.

We are effectively publishing seven issues this year instead of the usual five, and this means a lot of extra work for everyone involved in the process. Let me thank Bonnie Ponce, Annie Petitt, and Amanda Gedney for all of their hard work. Let me also thank Taylor & Francis for their willingness to work with us to resolve this issue.

So let's get this party started!

Our lead article for this issue is a timely exploration of mathematics and epidemiology. Berit Nilsen Givens and Jennifer Switkes use a combination of graph theory and probability to examine the effectiveness of standard disease prevention measures. By considering the spread of infection through various sorts of networks, they show, among other things, that reducing the size of gatherings can dramatically slow the spread of disease.

Our co-lead article is a tour-de-force of applied trigonometry. Karen Bliss and Gregory Hartman consider the mathematics of compound miter saws. They work out the equations for finding the correct miter and bevel angles for crown molding and various sorts of boxes. Even if you have no interest in woodworking, you will find that their clever mathematics makes for engaging reading.

Fans of differential equations have two items to choose from. J. Alberto Conejero, Marina Murillo-Arcilla, Jesús M. Seoane, and Juan B. Seoane study the dynamics of car-following models. In addition to the mathematical interest of their results, they note the pedagogical value of these models in teaching students about dynamical systems and the physical interpretation of mathematical models. John E. Kampmeyer and Timothy J. McDevitt take as their starting point the deceptively simple equation $f(x) = f^{-1}(x)$, which was the subject of a classic problem in the *American Mathematical Monthly* in 1968. The ensuing investigation leads them to the golden ratio, Riemann surfaces, and the so-called "metallic numbers." Never heard of them? Well, neither had I before reading this fascinating and exceedingly clever article.

The number theorists can start with the amuse-bouche from Gary Reid Lawlor. He offers his contribution in the long-standing competition to find the shortest proof that nth roots of positive integers are either integers or irrational. Ye olde Basel problem receives a fresh treatment in Vilmos Komornik's article. Somphong Jitman and Ekkasit Sangwisut use group theory, hyperbolic trigonometry, and the perplex numbers to study a problem about Pythagorean triples.

If you are in the mood for algebra, you have come to the right place. Joseph A. Gallian and Shahriyar Roshan Zamir consider the venerable problem of working out the subgroup structure of the unit group modulo n. The magazine has published several articles on this topic over the years, and this is a welcome contribution to the genre. Amir Rastpour and Jacob Bourdeau-Marche consider a fascinating problem in matrix

Math. Mag. **95** (2022) 256–257. doi:10.1080/0025570X.2022.2103318 © Mathematical Association of America

algebra: under what conditions do matrices follow the same algebraic rules as scalars? If you prefer something more applied, then have a look at David Singer's article. He explores the mechanics of linear feedback shift registers, which play an important role in modern cryptography.

With Elias Abboud's article we come to Euclidean geometry. He starts with Marion's theorem: Trisect the sides of any triangle and connect the resulting points to the opposite vertices. The area of the resulting central hexagon is one-tenth the area of the original triangle. But what happens if we cut the sides in other ways? What other ratios can we obtain when we compare the hexagon to the original triangle? Gábor Gévay and Tomaž Pisanski study the famous six-circle theorem, which is one of several circle-incidence theorems attributed to Auguste Miquel. They show that the so-called "Miquel configuration" can be realized with circles of equal radius.

If you prefer your geometry differential as opposed to Euclidean, then have a look at the article by Hassan Boualem and Robert Brouzet. Everyone smiles the first time they notice that the standard $2\pi r$ formula for the circumference of a circle is the derivative of the standard area formula πr^2. But this relationship no longer holds if we use the diameter instead of the radius. This observation leads Boualem and Brouzet to a fascinating meditation on what a derivative really is, as well as to differential forms on manifolds.

Stephen M. Zemyan provides us with an elegant exercise in multivariable calculus, as he works out the polar moment of the solid Mylar balloon. Kenneth Levasseur considers the combinatorial game "Pass the Buck" and introduces readers to the stochastic abacus. George Stoica serves up an unusual characterization of anti-derivatives of real-valued functions. And Yves Nievergelt rounds out this issue's articles by considering the problem of round-off error in computer algebra systems.

Even more so than usual, we truly have something for everyone!

We also have original problems, reviews, proofs without words, and the presentation of the 2021 Allendoerfer Awards. Surely that will be enough to hold you until we do this all again in December.

Jason Rosenhouse, Editor

Spread of Infection in a Network

BERIT NILSEN GIVENS
California State Polytechnic
University, Pomona
Pomona, CA 91768
bngivens@cpp.edu

JENNIFER SWITKES
California State Polytechnic
University, Pomona
Pomona, CA 91768
jmswitkes@cpp.edu

Throughout the season of coronavirus, society has been studying, discussing, and debating the importance and efficacy of social distancing measures. These measures include wearing masks, isolating socially, and limiting the size of gatherings. We use the mathematical structure of graph theory, paired with combinatorial reasoning, to explore the impact of each of these measures on the spread of infection in a network.

Consider a graph consisting of vertices and edges. Each vertex will represent an individual. An edge between two vertices will mean that the two corresponding individuals interact. We will focus on complete graphs, which are graphs in which each vertex is adjacent to every other vertex. In Figure 1, we show the complete graph on six vertices; this could be a family of six people living together in their home.

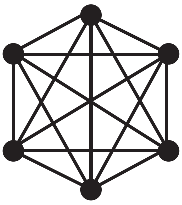


Figure 1 The complete graph on six vertices, K_6.

Larger complete graphs could represent the students in a classroom, the shoppers in a supermarket, or party-goers at an indoor party, and the list goes on.

What happens when infected individuals are included among the vertices of a complete graph?

Introduction to infection on K_n

Consider the complete graph on n vertices, K_n. Suppose that currently i vertices are infected, where $1 \leq i \leq n$. We will explore the dynamics of this network over time using a discrete time step. Suppose that in each step for each infected vertex and adjacent noninfected vertex the probability that the infected vertex spreads the infection to

Math. Mag. **95** (2022) 258–270. doi:10.1080/0025570X.2022.2092364 © Mathematical Association of America
MSC: Primary 60J20, Secondary 94C15

the noninfected vertex is p. The infection rate p represents the level of contagiousness. Less contagious illnesses would have a smaller value of p. The value of p could be lowered by wearing masks, maintaining distance from others, or vaccinations.

One-step probabilities Let $p_{i,j}$ be the probability, given that currently exactly i vertices are infected, that after exactly one step exactly j vertices are infected, with $0 \leq i \leq n$ and $0 \leq j \leq n$. For simplicity, we begin by assuming that individuals never recover from the infection. Later, we consider a modification that allows for recovery with no conferred immunity. Note that $p_{i,j} = 0$ if $j < i$, while for $j \geq i$,

$$p_{i,j} = \binom{n-i}{j-i} \left(1 - (1-p)^i\right)^{(j-i)} (1-p)^{i(n-j)}. \tag{1}$$

Here, the quantity $n - i$ is the number of previously noninfected vertices, and the quantity $j - i$ is the number of vertices to become newly infected in order to bring the number of infected vertices to j. Each of the $n - i$ noninfected vertices interacts with all i infected vertices and becomes infected if and only if at least one of the i already infected vertices manages to infect it.

We will use the term *total infection* for the state in which all n vertices are infected. Let t_i be the expected number of steps until total infection from a state in which i vertices are infected. Conditioning on the first step for $1 \leq i \leq n$,

$$t_i = p_{i,i}[t_i + 1] + p_{i,i+1}[t_{i+1} + 1] + \ldots + p_{i,n}[t_n + 1] \tag{2}$$

where $t_n = 0$. In the first step, we either remain with i infected vertices, or we end up with $i + 1, \ldots, n$ infected vertices. One step has already taken place, and we then look at the duration from there.

Thus, in matrix form, the system of equations for $t_1, t_2, \ldots, t_{n-1}$ is

$$\begin{bmatrix} 1 - p_{1,1} & -p_{1,2} & -p_{1,3} & \cdots & -p_{1,n-1} \\ 0 & 1 - p_{2,2} & -p_{2,3} & \cdots & -p_{2,n-1} \\ & & \ddots & & \\ 0 & 0 & \cdots & 0 & 1 - p_{n-1,n-1} \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_{n-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

where we have used the fact that

$$p_{i,i} + p_{i,i+1} + \ldots + p_{i,n} = 1.$$

Denoting the $(n-1) \times (n-1)$ coefficient matrix by T, the vector of expected durations by $\mathbf{t}$, and the vector of 1's by $\mathbf{1}$, we have $\mathbf{t} = T^{-1}\mathbf{1}$. Now define the matrix A to have entry $p_{i,j}$ in row i and column j, with $1 \leq i \leq n-1$ and $1 \leq j \leq n-1$. Note that $T = I - A$, where I is the $(n-1) \times (n-1)$ identity matrix. Thus,

$$\mathbf{t} = (I - A)^{-1}\mathbf{1} \tag{3}$$

gives the expected number of steps to total infection.

Multi-step probabilities For K_n, the probability of total infection after *at most* two steps, starting from one infected vertex, as a function of infection probability p, is given by

$$\sum_{k=1}^{n} p_{1,k} p_{k,n}.$$

In Figure 2, we plot the probability that K_{10}, K_{100}, and K_{1000} become totally infected after at most two steps, starting from one infected vertex, as a function of the infection

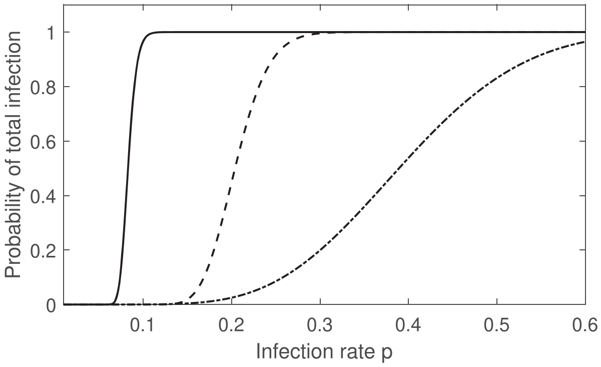


Figure 2 The probability that K_{10} (right), K_{100} (middle), and K_{1000} (left) become totally infected after at most two time steps, starting from one infected vertex, as a function of the infection rate p.

rate p. Total infection is more likely to happen quickly for larger complete graphs, hence the recommendations and mandates for social gatherings to be kept small.

The probability of total infection in *exactly* two steps, as a function of infection probability p, is given by

$$\sum_{k=1}^{n-1} p_{1,k} p_{k,n}.$$

This can be generalized to three, four, or an arbitrary number of steps. For example, the probability of total infection in exactly three steps is

$$\sum_{k=1}^{n-1} \sum_{\ell=k}^{n-1} p_{1,k} p_{k,\ell} p_{\ell,n}.$$

In Figure 3, we plot the probability that K_{1000} becomes fully infected after exactly three steps, starting from one infected vertex, as a function of the infection rate p. For small values of p, this probability is small since there has not been sufficient time yet for the infection to spread throughout the graph. For large values of p, this probability is small because the graph is likely to have become totally infected in no more than two steps. For intermediate values of p, the probability that K_{1000} becomes totally infected after exactly three steps peaks at close to probability 1. In large gatherings, a contagious disease can spread extremely quickly.

In general, the probability of total infection in exactly r steps is

$$\sum_{1 \le k_1 \le k_2 \le \cdots \le k_{r-1} \le n-1} p_{1,k_1} p_{k_1,k_2} \cdots p_{k_{r-1},n}. \tag{4}$$

In Figure 4, we look at K_{1000} and let $p = 0.01$. We plot the expected number of steps to total infection as a function of the number of already infected vertices. There are intervals along which the expected number of steps to total infection is approximately 3, and approximately 2. The curve approaches probability 1 as the number of already infected vertices approaches 999 since with 999 infected vertices it is almost certain that one (or more) of them will infect the final vertex in the next step. It is alarming to see that with $p = 0.01$ a network of 1000 individuals can become totally infected in a very small number of steps.

In Figure 5, we look at K_{1000} and let p range from 0.01 to 1, starting with one infected vertex. We plot the expected number of steps to total infection as a function

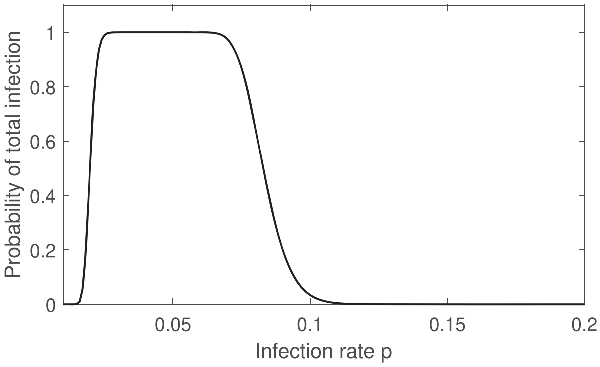


Figure 3 The probability that K_{1000} becomes fully infected after exactly three steps, starting from one infected vertex, as a function of the infection rate p.

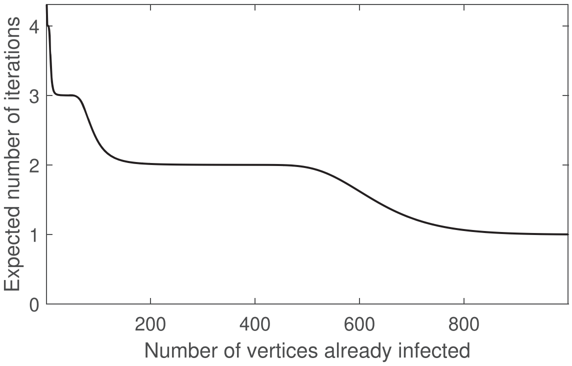


Figure 4 Expected number of steps to total infection in K_{1000} as a function of the number of already infected vertices; $p = 0.01$.

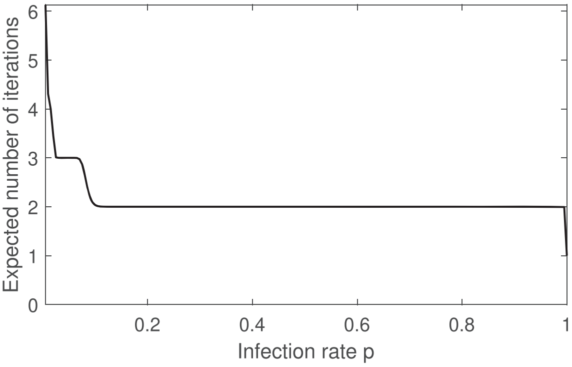


Figure 5 Expected number of steps to total infection starting with one infected vertex in K_{1000} as a function of p.

of the infection rate p. Again there are intervals along which the expected number of steps to total infection is approximately integral. Once again, the rapid spread of infection throughout a social network is dramatic.

In Figure 6, we plot the expected number of steps until total infection as a function of the number of vertices n for various values of the infection parameter p, starting with one infected individual. We see convergence to integer expected values as the size of the complete graph increases. For small infection rates, the expected number of steps to total infection is a decreasing function of the number of vertices. For large

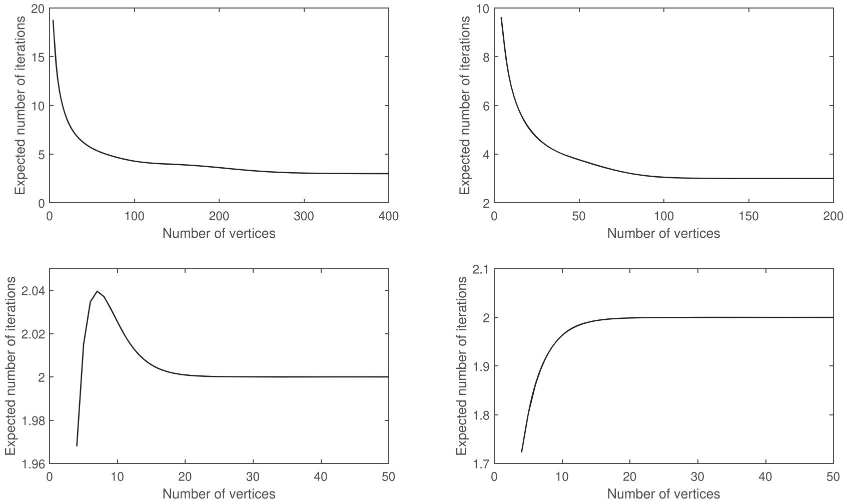


Figure 6 Expected number of steps until total infection as a function of the number of vertices n, for various values of the infection parameter p, starting with one infected individual. Upper left: $p = 0.05$. Upper right: $p = 0.1$. Lower left: $p = 0.6$. Lower right: $p = 0.7$.

infection rates, this function is increasing. For critical infection rates near $p = 0.6$, the expected number of steps to total infection increases and then decreases as the number of vertices increases. We highlight that, for small infection rates p, the disease spreads much more quickly on large social networks than on small social networks.

In the next section, we reformulate our model using the powerful mathematical tool of Markov chains.

Markov chain formulation

We now consider the model of infection on K_n as a Markov chain in which state i corresponds to the presence of exactly i infected vertices. We assume that initially at least one vertex is infected, so the state space is $\{1, 2, 3, \dots, n\}$. State n is absorbing.

One-step probabilities Define matrix P as the $n \times n$ matrix of one-step transition probabilities $p_{i,j}$, as defined previously in equation (1), with $1 \le i \le n$ and $1 \le j \le n$; see Mooney and Swift[1]. The matrix P can be decomposed into a block structure given by

$$P = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{5}$$

The $(n-1) \times (n-1)$ matrix A as defined previously is now observed to be the one-step transition probability matrix from non-absorbing states to non-absorbing states. Vector b is an $(n-1) \times 1$ column vector containing the probabilities $p_{1,n}$, $p_{2,n}$, $\dots$, $p_{n-1,n}$ of transitioning from less than total infection to total infection in one step. Vector $\mathbf{0}$ is a $1 \times (n-1)$ row vector with all entries zero, reflecting the fact that state n is absorbing and hence $p_{n,j} = 0$ for all $j \neq n$. The probability $p_{n,n} = 1$, again since state n is absorbing. For each row in matrix P, the entries sum to 1 as remarked earlier, and both matrix A and matrix P are upper triangular.

Multi-step probabilities Two-step transition probabilities are given by

$$P^2 = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} A & \mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} A^2 & (A+I)\mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix}.$$

Three-step transition probabilities are given by

$$P^3 = \begin{bmatrix} A^3 & (A^2 + A + I)\mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix}$$

and m-step transition probabilities are given by

$$P^m = \begin{bmatrix} A^m & (A^{m-1} + A^{m-2} + \ldots A^2 + A + I)\mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix}$$

for $m = 1, 2, \ldots$.

In the limit as $m \to \infty$, the infinite series

$$I + A + A^2 + A^3 + \ldots = (I - A)^{-1}.$$

Since matrix A is upper triangular, the eigenvalues of A are the diagonal elements of A, and since the diagonal elements $p_{i,i}$, for $1 \leq i \leq n-1$, are one-step transition probabilities of remaining in state i, the eigenvalues of A are all positive and less than 1. Therefore, the infinite series in powers of A does in fact converge to the result shown.

Thus, as $m \to \infty$,

$$P^m \to \begin{bmatrix} \mathbf{0} & (I - A)^{-1}\mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{0} & 1 \end{bmatrix}$$

where the (1, 1) block is the $(n-1) \times (n-1)$ zero matrix.

The entries $f_{i,j}$ of the *fundamental matrix*

$$F = (I - A)^{-1} \tag{6}$$

give the average number of times the process is in state j, given that it began in state i, where $1 \leq i \leq n-1$ and $1 \leq j \leq n-1$. To see this, expand back to $F = I + A + A^2 + \ldots$. Observe that entry $f_{i,j}$ gives the sum of the probabilities that the system is in state j after exactly zero steps, one step, two steps, …, given that the system started in state i. Thus, $f_{i,j}$ gives the expected number of times that the system will be in state j, given that the system started in state i. Summing over j, the sum of the entries of the ith row of the fundamental matrix F thus gives the expected number of times that a system initially in state i will be in states less than or equal to state $n-1$. That is, the sum of the entries in the ith row of F gives the expected number of steps for a process initially in state i to be absorbed, the same result we obtained as given by $\mathbf{t}$.

Starting with one infected individual, the probability that the system will be in state j after exactly r steps is given by the $1, j$ entry of P^r. Thus, the expected number of infected after exactly r steps is

$$\sum_{j=1}^{n} j \, (P^r)_{1,j} \, . \tag{7}$$

In Figure 7, we plot the expected number of infecteds as a function of number of steps for the complete graph K_{50} with infection rate $p = 0.02$, starting with one infected individual. Since there is no recovery, eventually everyone will be infected.

Next, we explore more complicated scenarios in which an infected individual interacts with several social networks.

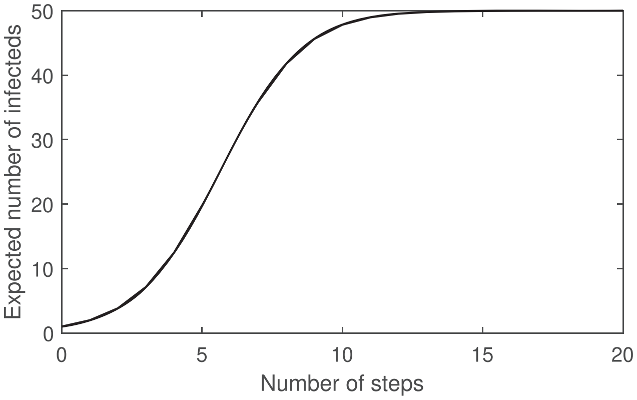


Figure 7 No recovery: expected number of infecteds as a function of number of steps, starting with one infected individual, K_{50}, $p = 0.02$.

Infection in star graphs

What happens if an infected college student interacts with several disjoint classrooms of students? What if an infected shopper runs several errands, interacting with shoppers in each store? We can model this scenario using the concept of star graphs.

Simple star graphs A simple star graph is a tree with one central node, with all other nodes appearing as leaves connected to the central node. The star graph S_m has one central node and $m - 1$ leaves, as seen in Figure 8.

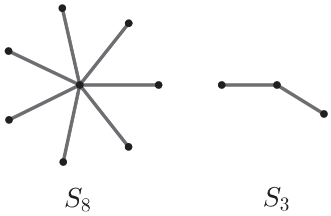


Figure 8 Two star graphs.

Suppose the central node is infected at time zero, with infection rate p. Then the probability of total infection of S_m after at most r time steps is

$$(1 - (1 - p)^r)^{m-1}.$$

Subtracting, the probability of total infection after exactly r time steps is

$$(1 - (1 - p)^r)^{m-1} - (1 - (1 - p)^{r-1})^{m-1}.$$

The expected time to total infection is therefore

$$\sum_{r=1}^{\infty} r \left[(1 - (1 - p)^r)^{m-1} - (1 - (1 - p)^{r-1})^{m-1}\right]. \tag{8}$$

An alternative approach is to use the fact that for any real numbers $X_1, X_2, \ldots, X_{m-1}$,

$$\max(X_1, \ldots, X_{m-1}) = \sum_i X_i - \sum_{i<j} \min(X_i, X_j)$$
$$+ \sum_{i<j<k} \min(X_i, X_j, X_k) - \cdots + (-1)^n \min(X_1, X_2, \ldots, X_{m-1}).$$

This identity can be established using an inclusion-exclusion type of argument. We can also note that the coefficient of X_i (when X_i is not the max value) in the expansion of the righthand side is an alternating sum of binomial coefficients, which is known to equal zero, while the maximum value is obtained only in the first sum and has a coefficient of 1.

Let X_i represent the time to infection for each leaf. Next, $E(X_i) = 1/p$ for each i. Given a pair X_i and X_j, the expected minimum value is the expected time until at least one of the pair is infected. The probability that at least one of an uninfected pair becomes infected is $1 - (1-p)^2$, so the expected time until at least one is infected is

$$E(\min(X_i, X_j)) = \frac{1}{1-(1-p)^2}.$$

Similarly,

$$E(\min(X_1, X_2, \ldots, X_j)) = \frac{1}{1-(1-p)^j},$$

for $j = 1, \ldots, m-1$.

Thus, the expected time until total infection is

$$E(\max(X_1, \ldots, X_{m-1})) = \sum_{j=1}^{m-1} \binom{m-1}{j} \frac{(-1)^{j+1}}{1-(1-p)^j}. \tag{9}$$

It is interesting to compare the two answers in equations (8) and (9). Letting $n = m-1$ and $q = 1-p$, we have shown that

$$\sum_{r=1}^{\infty} r\left[(1-q^r)^n - (1-q^{r-1})^n\right] = \sum_{j=1}^{n} \binom{n}{j} \frac{(-1)^{j+1}}{1-q^j} \tag{10}$$

Finally, if we instead assume that one of the leaves is infected, then the expected time to total infection is found by summing the expected time to infection of the central node (i.e., $1/p$) and the time to total infection of S_{m-1} after the central node is infected.

Composite star graph of rooted complete graphs We can extend the idea of star graphs further by imagining that leaves are replaced by complete graphs. There are two natural ways to imagine this, as seen in Figure 9.

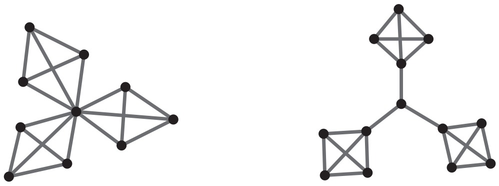


Figure 9 Two variations on a star graph with leaves replaced by complete graphs.

Suppose each time step is considered to be one day. Then we can imagine the graph on the right in Figure 9 as representing something like a store clerk as the central node, connected to the one member of each family of four that is doing daily shopping.

The graph on the left in Figure 9 can be thought of as representing a teacher who teaches in several disjoint classrooms of students. We will call this a *composite star graph of rooted complete graphs*, and we will explore it further.

Suppose that we have N complete graphs K_n, all rooted at a common vertex. Suppose further that this root vertex is infected. To compute the expected time to total infection, we follow the first approach used in computing the total infection of the simple star graphs.

For each complete graph K_n, the probability of total infection occurring in no more than r steps is $(P^r)_{1,n}$, and so the probability that all N complete graphs K_n have become totally infected in no more than r steps is

$$\left[(P^r)_{1,n}\right]^N.$$

Thus, the probability that it takes exactly r steps for all N complete graphs K_n to have become totally infected (some K_n's perhaps taking fewer than r steps, at least one K_n taking exactly r steps) is given by

$$\left[(P^r)_{1,n}\right]^N - \left[\left(P^{r-1}\right)_{1,n}\right]^N.$$

Finally, therefore, the expected number of steps for all N complete graphs K_n to have become totally infected (some K_n's perhaps taking fewer than r steps, at least one K_n taking exactly r steps) is given by

$$\sum_{r=1}^{\infty} r\left(\left[(P^r)_{1,n}\right]^N - \left[\left(P^{r-1}\right)_{1,n}\right]^N\right). \tag{11}$$

A simple star graph S_m can be thought of as $m-1$ complete graphs K_2 rooted at a common vertex. The transition matrix for K_2 is

$$P = \begin{bmatrix} 1-p & p \\ 0 & 1 \end{bmatrix}.$$

For K_2, it is straightforward to show that

$$(P^r)_{1,2} = 1-(1-p)^r.$$

In this case, we recover the result in equation (8).

In Table 1, we give the expected number of steps to total infection for a composite star graph consisting of N complete graphs K_n, all sharing a common root vertex that is infected.

In the upper table, $p = 0.01$. In the lower table, $p = 0.1$. As we would expect, as the number of complete graphs increases, the expected duration to total infection increases. As we have seen earlier, as the size of the complete graphs increases, the expected duration to total infection decreases, as infection spreads rapidly through large networks. As the infection rate increases, the expected duration to total infection decreases. The results suggest that it is much safer to congregate in many small groups than to gather in fewer but larger groups.

In Table 2 we explore this idea of restriction of the size of gatherings more systematically. We look at the expected number of steps to total infection for a composite rooted graph with one central vertex and 128 other vertices, by looking at (N, n) pairs

$p = 0.01$	$N = 10$	100	1000
$n = 10$	93.39	124.95	154.31
100	16.01	18.84	21.49
1000	4.98	5.01	5.06

$p = 0.1$	$N = 10$	100	1000
$n = 10$	10.19	13.20	16.00
100	3.42	4.00	4.00
1000	2.29	2.97	3.00

TABLE 1: Expected number of steps to total infection for a composite star graph consisting of N complete graphs K_n, all sharing a common root vertex that is infected. In the upper table, $p = 0.01$. In the lower table, $p = 0.1$.

N	1	2	4	8	16	32	64	128
n	129	65	33	17	9	5	3	2
$p = 0.01$	11.0	18.4	32.7	59.6	109.0	195.9	338.0	535.0
$p = 0.1$	3.0	3.7	4.9	7.2	11.6	19.6	33.0	52.1

TABLE 2: Expected number of steps to total infection for a composite star graph with one root vertex and 128 other vertices. As the number of groups increases (and the size of each group therefore decreases), the expected time to total infection increases.

for which $N(n-1) = 128$. As the number of groups increases (and the size of each group therefore decreases), the expected time to total infection increases dramatically. If the mathematics is to be believed (and we think it is), then small gatherings truly are the way to go when infection is a concern.

Throughout this article, we have assumed that individuals never recover from the infection. We conclude with a happier scenario in which individuals recover, though with no conferred immunity.

Recovery with no conferred immunity

Suppose now that when a vertex becomes infected, the vertex recovers in the next step, but receives no immunity upon recovery. The assumption that each infected individual recovers in one time step might seem very artificial. However, we can think of this as setting the time increment such that one time step represents the average length of time that the infection persists.

In this scenario, state 0 becomes the only absorbing state, and state n is unreachable since any infected vertices at the previous step are now recovered. Thus, we consider state space $\{1, 2, 3, \ldots, n-1, 0\}$. We will list the states in this order, and so the nth state corresponds to 0 infected vertices.

Recovery on a complete graph Define transition matrix P^* as

$$P^* = \begin{bmatrix} p_{1,2} & p_{1,3} & p_{1,4} & \cdots & p_{1,n-1} & p_{1,n} & p_{1,1} \\ p_{2,3} & p_{2,4} & p_{2,5} & \cdots & p_{2,n} & 0 & p_{2,2} \\ p_{3,4} & p_{3,5} & p_{3,6} & \cdots & 0 & 0 & p_{3,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{n-2,n-1} & p_{n-2,n} & 0 & \cdots & 0 & 0 & p_{n-2,n-2} \\ p_{n-1,n} & 0 & 0 & \cdots & 0 & 0 & p_{n-1,n-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}.$$

For $1 \leq i \leq n-1$ and $1 \leq j \leq n-i$, the entry i, j in matrix P^* is given by $p_{i,i+j}$ in the notation from equation (1). If one vertex is infected currently, that vertex might infect others and will recover, all in the next step. Thus, $p^*_{1,1} = p_{1,2}$ since in order for there to be one infected vertex at the next step, the currently infected vertex must have infected one vertex and then recovered. Similarly, $p^*_{1,2} = p_{1,3}$, since the currently infected vertex must have infected two vertices and then recovered. Similarly,

$$p^*_{n-1,1} = p_{n-1,n}$$

since the $n-1$ infected vertices must have infected the one remaining vertex, and then all the $n-1$ vertices recovered.

For $1 \leq i \leq n-1$, the entry i, n in matrix P^* is given by $p_{i,i}$. For example, $p^*_{1,n} = p_{1,1}$. Since the nth state corresponds to 0 infected vertices, the infected vertex must have infected no one and then recovered.

The matrix P^* has the form

$$P^* = \begin{bmatrix} A^* & \mathbf{b}^* \\ \mathbf{0} & 1 \end{bmatrix}.$$

In contrast to the earlier model for infection without recovery, in this model with recovery the matrix A^* is not triangular but rather anti-triangular.

Our model is a finite Markov chain with one absorbing state. In addition, the probability of moving from any state to the absorbing state is positive (for $p < 1$). In this type of Markov chain, we know that the probability of reaching the absorbing state approaches 1 as the number of steps goes to infinity. So theoretically, we know that in this model the disease will eventually be eradicated, no matter what the rate of infection is. However, the length of time to eradication is very much affected by the value of p.

In Figure 10, we plot the expected number of infecteds as a function of number of steps for the complete graph K_{50}, starting with one infected, for a variety of infection rates. If the infection rate was $p = 0$, the number of infecteds would drop to 0 in one step. With $p = 0.025$, the infection peaks and then is expected to die out over time. With $p = 0.05$, the infection appears to approach an endemic level (but will die out over an extended period of time). With infection rate $p = 0.1$, $p = 0.2$, $p = 0.4$, and $p = 0.8$, the infection appears to approach endemic levels through decaying oscillations (but again will die out over an extended period of time). If the infection rate was $p = 1$, the number of infecteds would forever alternate 1, 49, 1, 49, ... This idea of a quasi-endemic level of infection seems quite relevant as we think about the future.

We commented at the outset that wearing face masks can decrease the value of p. Here we see the potentially critical nature of this. What if, by wearing face masks, we can keep the infection rate p low enough that we move from a quasi-endemic scenario to a scenario in which the infection is eradicated promptly?

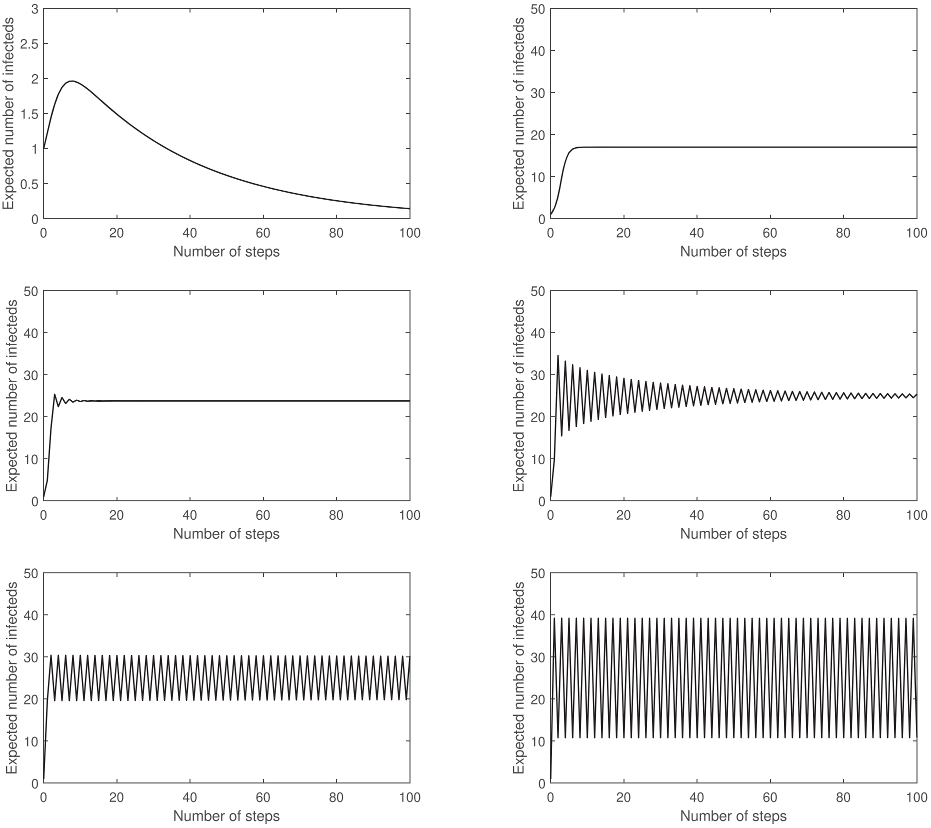


Figure 10 Recovery with no conferred immunity: Expected number of infecteds in K_{50}, starting with one infected, as a function of number of steps. Upper left: $p = 0.025$. Upper right: $p = 0.05$. Middle left: $p = 0.1$. Middle right: $p = 0.2$. Lower left: $p = 0.4$. Lower right: $p = 0.8$. In all but the upper left plot, the infection appears to approach or oscillate about an endemic level, but in fact will die out over an extended period of time.

Recovery on multiple complete graphs Here we are interested in looking at recovery on collections of complete graphs. With recovery, though, modeling of composite star graphs becomes much more complicated since the root vertex will recover and then perhaps get reinfected. Instead of exploring composite star graphs, therefore, we look at collections of disjoint complete graphs.

Suppose we have a community that is observing strict social distancing measures and thus consists of N disjoint complete graphs K_n, with each K_n currently having one infected individual. In Table 3, we look at the expected number of steps to total removal of infections for a community consisting of N disjoint complete graphs K_n, with each K_n beginning with one infected individual. As in Table 2, we use (N, n) pairs for which $N(n - 1) = 128$. In each pair, there is a total of 128 susceptible individuals, gathered in N disjoint groups. Here, we use $p = 0.01$ and $p = 0.025$. As the number of groups increases (and the size of each group therefore decreases), the expected time to total eradication decreases. It is remarkable that for $p = 0.01$ in moving from one group of 128 susceptible individuals to two groups of 64 susceptible individuals, the number of time steps to total eradication of infection drops by a factor of 100. We see a similar, and even more dramatic, trend with $p = 0.025$. Sometimes people wonder if the permitted gathering size has truly mattered. Our results suggest a resounding yes. In addition, our results suggest that the importance of keeping the value of p low, for example by wearing masks, almost cannot be overstated.

N	1	2	4	8	16	32	64	128
n	129	65	33	17	9	5	3	2
$p = 0.01$	307.53	2.93	2.17	1.93	1.83	1.77	1.75	1.74
$p = 0.025$	NA	1.6×10^4	5.69	3.16	2.53	2.25	2.11	2.04

TABLE 3: Expected number of steps to total eradication of infections for a community consisting of N disjoint complete graphs K_n, with each K_n beginning with one infected individual.

Final thoughts

We set out to explore questions regarding the importance and efficacy of many of the social distancing measures used in the past couple of years, such as wearing masks, isolating socially, and limiting the size of gatherings. Our results suggest that each of these three types of measures can play a very significant role in reducing the spread of infection.

The results are even more dramatic than we would have guessed, including tremendous improvements as communities self-isolated into smaller groups, and game-changing effects as the infection rate dropped. Future work will explore the impact of vaccination among a subset of the population (and corresponding lack of vaccination among the complement).

If this mathematical modeling can contribute to the work of surviving a pandemic and staying healthy post-pandemic, we will be most pleased.

REFERENCES

[1] Mooney, D. D., Swift, R. J. (1999). *A Course in Mathematical Modeling*. Washington D. C.: Mathematical Association of America.

Summary. The spread of an infection across a network depends on the degree of interconnectedness of the network and on the level of contagiousness of the infection. In this article, we explore what happens as networks grow or shrink in size, along with how the infection rate p affects the spread. We first consider a complete graph as a model of a "bubble" of people who interact freely. Then we explore a variant of star graphs, which could model a teacher interacting with several disjoint classes of students. Using graph theory and combinatorial reasoning, we show that limiting the size of gatherings and reducing the value of p can indeed dramatically slow the spread of a highly contagious disease like COVID-19.

BERIT NILSEN GIVENS (MR Author ID: 709024) received her PhD from the University of Wisconsin, Madison, in 2003. She is a professor at Cal Poly Pomona, where she enjoys teaching a great variety of courses for mathematics majors. Her research interests include combinatorics, algebra, and geometry. In her spare time, she is an avid knitter.

JENNIFER SWITKES (MR Author ID: 675737) received her PhD from Claremont Graduate University in 2000. She is a professor at Cal Poly Pomona, with research interests focused on mathematical modeling. She very much enjoyed this collaboration between a pure mathematician and an applied mathematician. In her spare time, she loves hiking.

The Mathematics of Compound Miter Saws

KAREN BLISS
Virginia Military Institute
Lexington, VA 24450
blisskm@vmi.edu

GREGORY HARTMAN
Virginia Military Institute
Lexington, VA 24450
hartmangn@vmi.edu

The authors were intrigued by an image and mathematical graph shown in the user's manual for a *double bevel compound miter saw*. This paper investigates the properties of the functions drawn in that graph, then generalizes them.

Miter saw basics

A miter saw is designed to make angled *crosscuts*, that is, cuts across the grain of a board. Consider Figure 1(a); a board is placed flat on the *table* and held against the *fence* at back. When a miter saw is oriented in space as shown in Figure 1(b), a *mitered cut* (or, simply, a miter) can be made by rotating the saw blade about the z-axis, typically a maximum of about 50° left/right. A *compound miter saw* allows the blade to be tilted (or *beveled*) out of the x-z plane, typically toward the negative y-axis (though a double-bevel compound miter saw can tilt toward both the positive and negative y-axes). Figure 5(a) shows the saw after it has been mitered and beveled. Miter saws are an essential tool for carpenters, who commonly use them to cut window and door trim, baseboards, and crown molding.

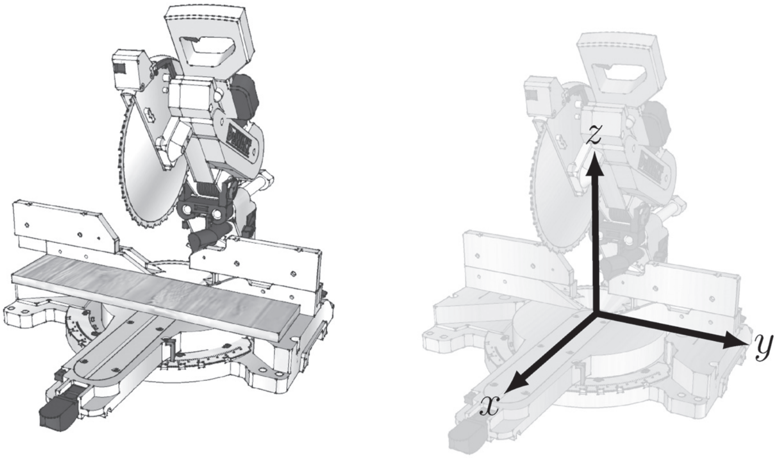
The first author purchased a DeWalt DWS779 compound miter saw and, math nerd that he is, immediately began reading the manual [1]. Page 11 contains the image shown in Figure 2, along with the following text:

> "A compound miter is a cut made using a miter angle and a bevel angle at the same time. This is the type of cut used to make frames or boxes with slanting sides, like the one shown in Figure 15 ... The chart at the end of this manual (Table 1) will assist you in selecting the proper bevel and miter settings for common compound miter cuts."

The chart referenced as "Table 1" in the quote is shown here in Figure 3. To use it, we first decide the number of sides needed for our box. If we choose four, look along the curve that is labeled "Square Box" (a title inconsistent with the labels of the other curves). Then we determine the exterior slope angle of our box, in degrees, referenced as "ANGLE A" in Figure 2. If we choose A= 70°, we look along the Square Box curve to the dot marked "70". The "x"-coordinate of this dot is approximately 41.5°, and this is the bevel angle. The "y"-coordinate of the dot is approximately 19°, and this is the miter angle.

The authors were fascinated by this chart. The curves, which the authors named "D-curves" in honor of the DeWalt company, appear to be quarter-circles, with the "Square Box" curve having a radius of 45°. It looks as though one might determine the

Math. Mag. **95** (2022) 271–284. doi:10.1080/0025570X.2022.2093547 © Mathematical Association of America
MSC: 00A69



(a) A miter saw

(b) Orienting axes on the saw

Figure 1 Basic miter saw images (blade guard of saw removed for clarity). (*All miter saw figures based on [**2**].*)

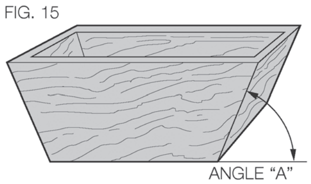


Figure 2 The box with sloped sides shown in the DeWalt DWS779 manual [**1**].

bevel and miter angles as 45 cos A and 45 sin A, respectively. At the same time, there are hints that these formulas are not correct. For one, the three points marked "70" do not appear to be collinear.

The rest of this paper will determine parametric equations for the D-curves, highlight a few of their interesting properties, and then generalize them.

Solving for miter and bevel angle functions

Throughout, we use the term "edge" as a woodworker would, meaning "one end of a board." We are not referring to the "edge" of a polyhedron. Since woodworking angles are measured in degrees, we measure all angles in degrees. We use α to represent the manual's angle A and let φ be the exterior angle of the box's corner. We will assume that our goal is to make boxes with corner angles of equal measure. Thus, for a box with n sides, $\varphi = 360°/n$. (Later, we will let φ represent the exterior angle of any corner, not just those associated with an n-sided, equal-angled box.)

As shown in Figure 1(b), we orient the saw by placing the miter/bevel pivot point at the origin, the saw blade (in its standard position) in the x-z plane, the table in the x-y plane, and the fence in the y-z plane, placing the line of intersection of the table and fence along the y-axis. With this orientation, it is clear that the unit vector $\vec{s}_1 = \langle 0, 1, 0\rangle$ is normal to the plane of the saw blade.

We want to position the saw blade so that it will properly cut one edge of a box. In Figure 4(a), a box with exterior slope angle $\alpha = 70°$ is shown; in Figures 4(b) and (c) one side, then one edge, are isolated. We will focus our attention on cutting that edge. In this "upright" position, we denote the plane containing this edge as p_e, and let its unit normal be $\vec{e}_1$.

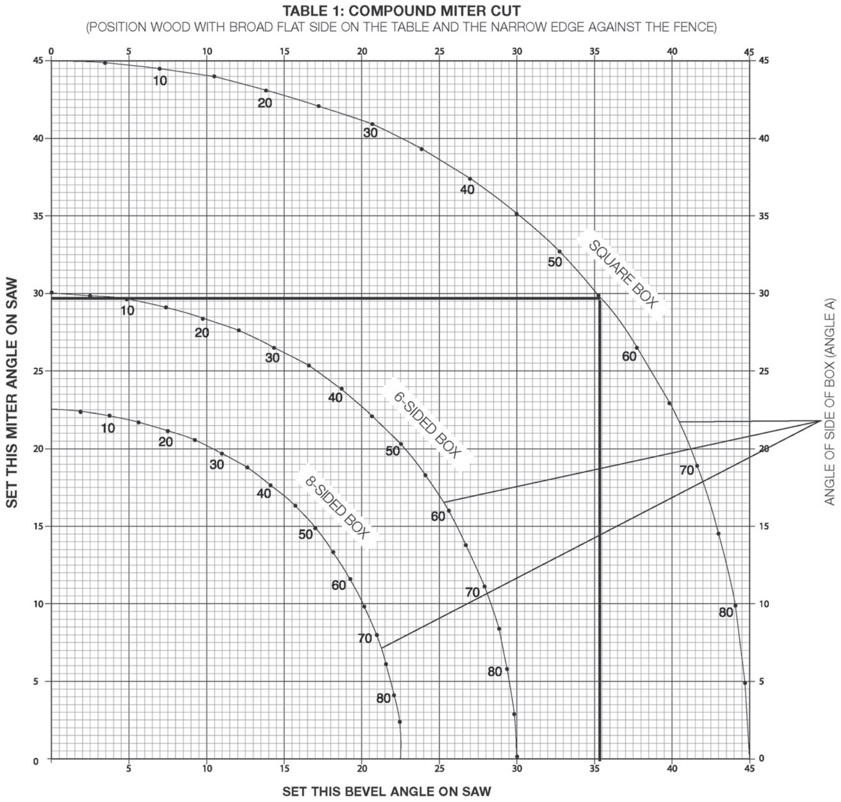
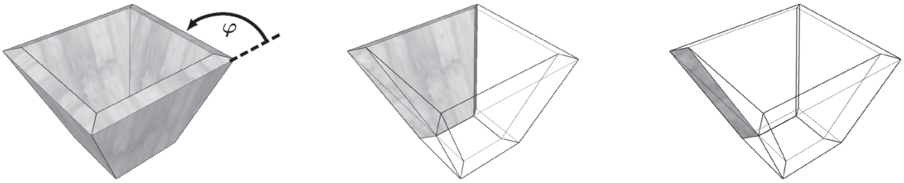


Figure 3 The chart given in the DeWalt DWS779 manual [1] for determining the miter and bevel angles needed to create the sloped-sided box shown in Figure 2.



(a) Viewing the whole box

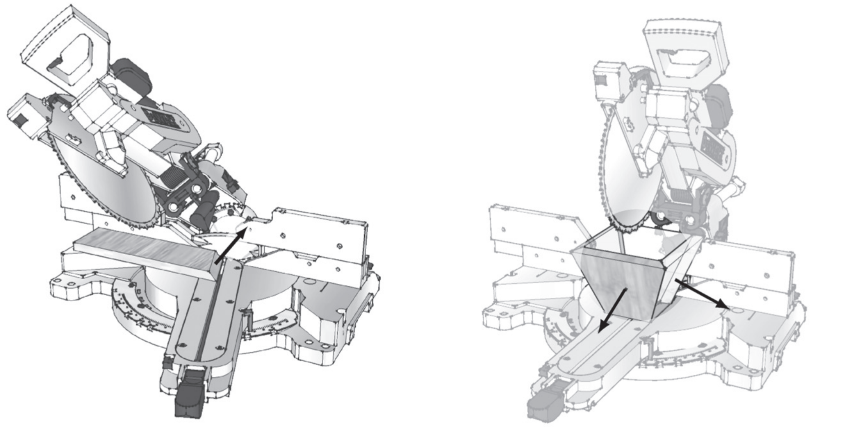
(b) One side

(c) One edge

Figure 4 Understanding the components of a box.

In Figure 5(a), we see the saw having cut a board on its table. The cut edge is the isolated edge of Figure 4(c). As positioned in Figure 5(a), the cut edge lies in a plane with unit normal vector $\vec{e}_2$, where $\vec{e}_2$ is chosen to have a positive z-coordinate. Once we know the components of $\vec{e}_2$, we can accomplish the first goal of this paper, which is to find the proper miter and bevel angles, m and b, respectively, so that $\vec{s}_1$ is rotated to be equal to $\vec{e}_2$. That is, we want to know how we can rotate the blade so that the blade is in the same plane as the edge we want to cut.

To determine $\vec{e}_2$, first consider our box as oriented in Figure 5(b). The isolated side, which we will call "box side 1," is parallel to the y-axis, and the box rests on the x-y plane. Let p_1 be the plane containing the exterior face of the isolated side with unit normal $\vec{n}_1$, let p_2 be the plane containing the exterior face of the adjacent, transparent, side (which we will call "box side 2") with unit normal $\vec{n}_2$, and let p_e (which we referenced two paragraphs earlier) be the plane containing the face of the intersection of these two box sides with unit normal $\vec{e}_1$. Once we know $\vec{e}_1$, we will show how to easily find $\vec{e}_2$.



(a) Mitering and beveling to cut a board, with vector in the direction of $\vec{e}_2$ shown.

(b) Illustrating vectors in the direction of $\vec{n}_1$ and $\vec{n}_2$.

Figure 5 Understanding the compound cut.

The experience of woodworkers makes determining $\vec{e}_1$ relatively straightforward. It is known that if our box is placed on the saw table as shown in Figure 5(b), such that the plane p_e contains the origin (i.e., such that the plane containing the edge between box sides 1 and 2 contains the miter/bevel swivel point), then that edge could be cut with just a miter and no bevel. That is, if we could hold a board parallel to the fence and at an angle of α with the table, then we would only need to miter the saw blade and not bevel it as well. Moreover, the miter angle is rather intuitive: for a 4-sided box, we miter at 45°; for a 6-sided box, we miter at 30°; when the exterior angle of the corner is φ, we miter at $\varphi/2$. The angle α does not influence the miter angle with this setup. (While this concept is straightforward, in practice, this is hard to do. It is difficult to hold a board steady while cutting it with a whirling blade!) Thus, $\vec{e}_1$ is just the rotation of the saw blade's normal vector, $\vec{s}_1 = \langle 0, 1, 0 \rangle$, about the z-axis by an angle of $\varphi/2$.

We will need to rotate vectors about each of the coordinate axes. We use the standard rotation matrices, given by the equations in equation (1), to rotate 3D space counter-clockwise an angle of θ about a coordinate axis, and we use subscripts to indicate the axis.

$$M_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \qquad M_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$$
$$M_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Thus $\vec{e}_1 = M_z(\varphi/2)\vec{s}_1 = \langle -\sin(\varphi/2), \cos(\varphi/2), 0 \rangle$.

Later, it will be useful to have a method of finding $\vec{e}_1$ that is more generalizable, so we will develop one here. Again, let $\vec{n}_1$ be the unit normal vector to "box side 1" and let $\vec{n}_2$ be the unit normal vector to "box side 2," as described above. Though we gloss over a few of the finer details, the vector $\langle \cos(\alpha), 0, \sin(\alpha) \rangle$ has an angle of elevation of α with the x-axis and is parallel to box side 1, hence $\vec{n}_1 = \langle \sin(\alpha), 0, -\cos(\alpha) \rangle$ is a unit normal vector to this side (with a negative z-component, as indicated in Figure 5(b)). To determine $\vec{n}_2$, we can rotate $\vec{n}_1$ counter-clockwise about the z-axis by an angle of φ:

$$\vec{n}_2 = M_z(\varphi)\vec{n}_1 = \langle \sin(\alpha)\cos(\varphi), \sin(\alpha)\sin(\varphi), -\cos(\alpha) \rangle.$$

Since these vectors are of the same length, the plane p_e (again, that is the intersecting edge of the two box sides) is parallel to $\vec{n}_1 + \vec{n}_2$ and is orthogonal to $\vec{n}_2 - \vec{n}_1$.

This latter vector is not a unit vector. Dividing by its magnitude, we discover $\vec{e}_1$:

$$\begin{aligned}\vec{e}_1 &= \frac{\vec{n}_2 - \vec{n}_1}{\|\vec{n}_2 - \vec{n}_1\|}\\ &= \frac{\big\langle \sin(\alpha)\big(\cos(\varphi) - 1\big), \sin(\alpha)\sin(\varphi), 0\big\rangle}{\sqrt{\sin^2(\alpha)\big(\cos(\varphi) - 1\big)^2 + \sin^2(\alpha)\sin^2(\varphi)}}.\end{aligned}$$

Assuming $0° < \alpha < 180°$ and $0° < \varphi \leq 180°$, we apply several trigonometric identities and simplify to obtain

$$\vec{e}_1 = \langle -\sin(\varphi/2), \cos(\varphi/2), 0\rangle \tag{2}$$

as determined before.

We have found $\vec{e}_1$, the unit normal to the edge when in its upright position. We need $\vec{e}_2$, the edge's unit normal when it is laid down on the table. To obtain $\vec{e}_2$ from $\vec{e}_1$, we rotate $\vec{e}_1$ counter-clockwise about the y-axis an angle of α.

With $\vec{e}_2 = M_y(\alpha)\vec{e}_1$, we have

$$\vec{e}_2 = \big\langle -\cos(\alpha)\sin(\varphi/2), \cos(\varphi/2), \sin(\alpha)\sin(\varphi/2)\big\rangle. \tag{3}$$

Letting b represent the angle by which the saw is beveled, and letting m represent the angle by which the saw is mitered, we seek the proper bevel/miter combination such that the unit normal to the saw blade, $\vec{s}_1 = \langle 0, 1, 0\rangle$, is equal under rotation to $\vec{e}_2$.

Adjusting the miter angle of the saw is performed by rotating the saw's blade about the z-axis. In our setup, we are interested in counter-clockwise rotations, so we can multiply by $M_z(m)$ to perform the rotation. Adjusting the bevel angle is performed by rotating the saw's blade about the x-axis. Again, our setup requires a counter-clockwise rotation, achieved by multiplying by $M_x(b)$.

We can multiply these matrices together to perform both a miter and a bevel, but we must multiply in the correct order. When one adjusts the miter angle of a miter saw, the pivot point of the beveling mechanism also rotates, so that beveling no longer occurs as a rotation around the x-axis. Therefore, the product $M_x(b)M_z(m)$ represents rotating the saw blade about the z-axis, followed by a rotation about the x-axis, which is not how real saws move.

Instead, if we bevel first, adjusting the miter angle is still a rotation about the z-axis. Hence, we can represent a compound miter cut with the rotation matrix $M_z(m)M_x(b)$.

Applying this rotation to $\vec{s}_1 = \langle 0, 1, 0\rangle$, we get a unit normal to the saw's blade after miter and bevel adjustment of angles m and b, respectively:

$$\vec{s}_2 = M_z(m)M_x(b)\vec{s}_1 = \big\langle -\cos(b)\sin(m), \cos(b)\cos(m), \sin(b)\big\rangle. \tag{4}$$

We seek the bevel angle b and miter angle m so that $\vec{s}_2 = \vec{e}_2$. Equating the third components of each vector, as seen in equations (3) and (4), we can solve for b:

$$b(\alpha, \varphi) = \sin^{-1}\big(\sin(\alpha)\sin(\varphi/2)\big). \tag{5}$$

To solve for m, equate the proportions of the first component of each vector divided by the second component.

From $\vec{e}_2$ in equation (3):

$$\frac{-\cos(\alpha)\sin(\varphi/2)}{\cos(\varphi/2)} = -\cos(\alpha)\tan(\varphi/2). \tag{6}$$

From $\vec{s}_2$ in equation (4):

$$\frac{-\cos(b)\sin(m)}{\cos(b)\cos(m)} = -\tan(m). \tag{7}$$

From equations (6) and (7):

$$m(\alpha, \varphi) = \tan^{-1}\big(\cos(\alpha)\tan(\varphi/2)\big). \tag{8}$$

(Equations (5) and (8) give the proper bevel and miter angles to cut the *left* edge of a board, as shown in Figure 4.

We leave it to the reader to confirm that in order to cut the *right* edge, one needs to bevel and miter with angles $-b(\alpha, \varphi)$ and $-m(\alpha, \varphi)$, respectively. That is, one bevels and miters the same amount, but in the opposite direction.)

By fixing a value for φ, the parametric equations

$$x(\alpha) = b(\alpha, \varphi), \quad y(\alpha) = m(\alpha, \varphi), \quad 0^\circ \le \alpha \le 360^\circ, \quad 0^\circ \le \varphi < 180^\circ \tag{9}$$

determine a D-curve, denoted by D_φ. The curves D_{90}, D_{60} and D_{45} are shown in the DeWalt saw manual in Figure 3. (Note how in equation (9) we remove some of the restrictions on α and φ established before equation (2), as the equations for the bevel and miter angles admit a greater range of values.) We will use $D_\varphi(\alpha)$ to describe the point on D_φ that gives the bevel and miter angles needed to make a box with sides sloped an angle of α. We will use D'_α to denote the curve parametrized in a manner similar to equation (9), where instead α is fixed and φ varies. Each point on this curve describes a corner, with exterior angle φ, where the sides have slope angle α.

The equations for generating D-curves are well known, appearing in many books, magazines, and websites, though their properties seem not to have been explored.

We were unable to find the earliest reference to these formulas in the literature. Vautaw investigates cutting four-sided picture frames with sloped sides on a table saw [**3**]. Since the blade of a table saw is always parallel to its fence, the formulas given there are similar, though different, to the ones derived here.

Properties of D-curves

Graphing In Figure 6, the curves D_{90}, D_{60}, and D_{45} are shown, corresponding to the exterior angles needed for regular 4-, 6-, and 8-sided boxes, respectively. Also shown, with dashed lines, are quarter circles with radii of 45, 30, and 22.5. In the introduction, it was stated that D-curves appeared to be quarter circles; the figure and the equations of the D-curve confirm that they are not. (Note, though, the similarity between D_{45} and its quarter circle.)

Also shown in the figure is the curve D'_{70}. Every point on this curve determines the bevel/miter angles needed to make a corner where the sides are sloped at 70°. What is key to note here is that this curve is not linear, as alluded to in the introduction (the line $y = \cot(70^\circ)x$ is drawn with a dashed line for comparison).

Symmetry D-curves are symmetric about the line $y = x$, though not in the "expected" way. Revisit the vectors $\vec{e}_2$ and $\vec{s}_2$ from equations (3) and (4), respectively, which we equated to determine equations for b and m. Equating their second components gives

$$\cos(b)\cos(m) = \cos(\varphi/2). \tag{10}$$

When b, m, and φ satisfy this equation, we know that the ordered pair (b, m) lies on D_φ. But as the equation is symmetric in terms of b and m, it also means that the ordered pair (m, b) also lies on D_φ, giving us the symmetry of D_φ about the line $y = x$.

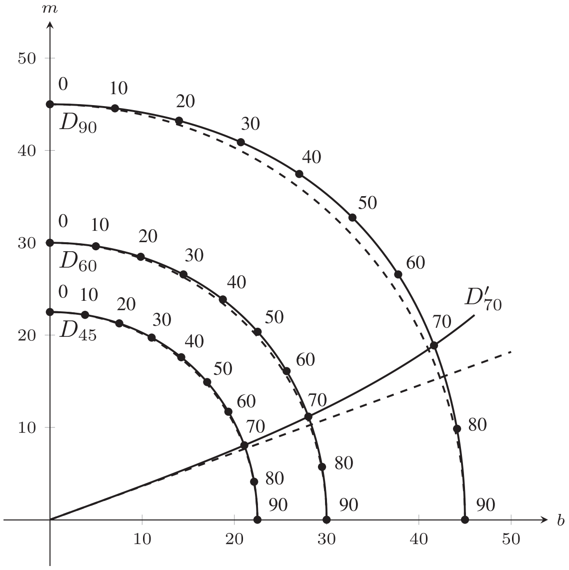


Figure 6 Drawing D_{90}, D_{60}, D_{45}, and D'_{70}, along with related quarter-circles and line with a dashed pen.

The unexpected nature of the symmetry is this: the image of $D_\varphi(\alpha)$ under reflection about $y = x$ is not, in general, $D_\varphi(90^\circ - \alpha)$. For example, $D_{90}(70^\circ) \approx (41.6^\circ, 18.9^\circ)$. While the (approximate) point $(18.9^\circ, 41.6^\circ)$ does lie on D_{90}, it is not $D_{90}(20^\circ)$, which is approximately $(14.0^\circ, 43.2^\circ)$. One can also "eyeball" the odd nature of this symmetry by noticing that $D_{90}(0^\circ)$ and $D_{90}(10^\circ)$ are much closer together than $D_{90}(80^\circ)$ and $D_{90}(90^\circ)$.

Limiting curves If we fix α and consider what happens as $\varphi \to 0^\circ$ (i.e., as we design boxes with increasing numbers of sides), both $b(\alpha, \varphi)$ and $m(\alpha, \varphi)$ approach 0° for all α. However, as illustrated in Figure 6, the D-curves they describe approach a circle of radius $\varphi/2$.

To confirm this, consider:

$$\lim_{\varphi\to 0^\circ} \frac{b(\alpha,\varphi)}{\varphi/2} = \lim_{\varphi\to 0^\circ} \frac{\sin^{-1}\big(\sin(\alpha)\sin(\varphi/2)\big)}{\varphi/2} = \sin(\alpha),$$

where the limit is evaluated with L'Hopital's rule. It is similarly straightforward to confirm that as $\varphi \to 0^\circ$, $m(\alpha,\varphi)/(\varphi/2) \to \cos(\alpha)$.

Putting these two concepts together, for small φ, the curve D_φ is well-approximated by the parameterization $x = \frac{\varphi}{2}\sin(\alpha)$, $y = \frac{\varphi}{2}\cos(\alpha)$, which describe a circle of radius $\varphi/2$. That is, for small φ, $b(\alpha,\varphi) \approx \frac{\varphi}{2}\sin(\alpha)$ and $m(\alpha,\varphi) \approx \frac{\varphi}{2}\cos(\alpha)$.

We can also investigate curves that are not drawn in the DEWALT manual. Suppose, for example, we fix α between 0° and 90°.

As $\varphi \to 180^-$, it is straightforward to see that $b(\alpha,\varphi) \to \alpha$. Also, as $\varphi \to 180^{\circ -}$, consider

$$m(\alpha,\varphi) = \tan^{-1}\big(\cos(\alpha)\tan(\varphi/2)\big).$$

The argument of arctangent approaches infinity, showing that $m(\alpha,\varphi) \to 90^\circ$. (We encourage the reader to consider the development of the "two-sided box" constructed as $\varphi \to 180^\circ$.)

Letting α again vary between 0° and 360° and using the principles of the previous paragraph, we let the reader confirm that as $\varphi \to 180^-$, the D-curves approach a square, centered at the origin, with opposite corners $(-90^\circ, -90^\circ)$ and $(90^\circ, 90^\circ)$, as illustrated in Figure 7.

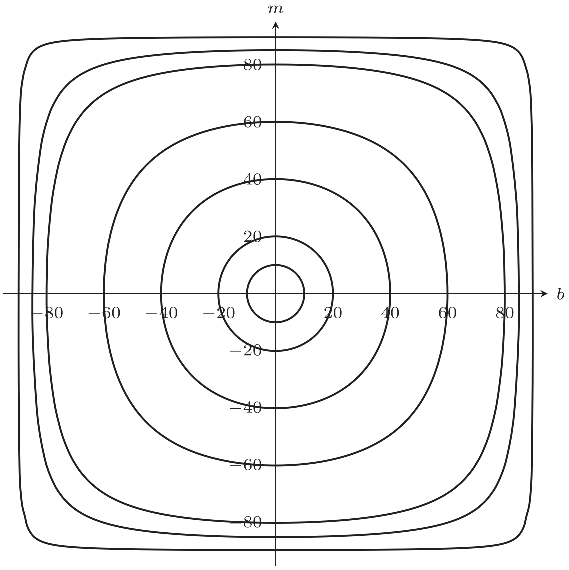


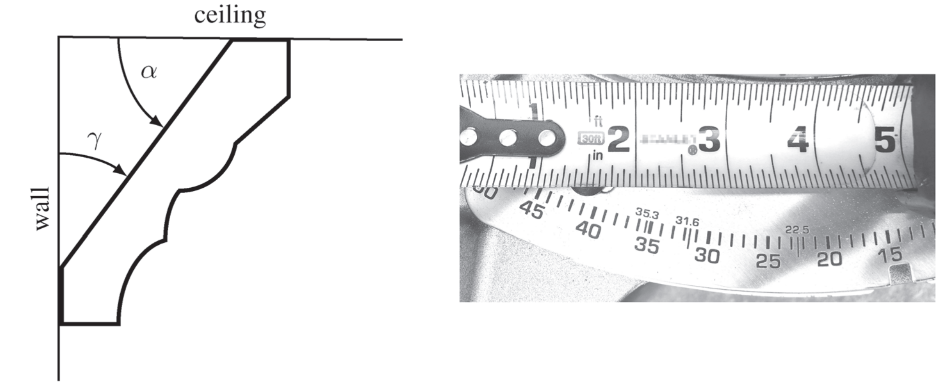
Figure 7 Drawing D_φ for $\varphi = 20^\circ, 40^\circ, 80^\circ, 120^\circ, 160^\circ, 170^\circ$, and 179°. For small values of φ, the curve approximates a circle. For values near 180°, the curve approaches a square.

Every bevel/miter combination is unique (almost) Consider only the first quadrant of the b-m plane, where $0^\circ \le b, m < 90^\circ$, as illustrated in Figure 6. For each point (b, m) in this region, there exists exactly one point (α, φ), $0^\circ \le \alpha \le 90^\circ$, $0^\circ \le \varphi < 180^\circ$, such that $(b, m) = D_\varphi(\alpha) = D'_\alpha(\varphi)$. That is, if one cuts the left edge of a board with a random bevel/miter combination (b, m), along with the right edge of a board with the combination $(-b, -m)$, then those boards can be paired to form a unique corner with exterior angle φ and common slope angle α. We can verify this claim using equation (10); given b and m, there is only one $0^\circ \le \varphi \le 180^\circ$ such that $\cos(b)\cos(m) = \cos(\varphi/2)$. Of course, there is probability 0 that this random combination is the basis for a regular n-sided box; at the same time, one could find select points along D'_α to construct an *irregular* n-sided box wherein all sides have a slope angle of α.

This combination is unique if we constrain ourselves to both boards having the same slope angle. In the next section we'll look at boxes where boards with different slope angles meet at a corner. When considering that construction, each bevel/miter combination can be used to construct an infinite array of corners.

Application to crown molding Making a box with sloped sides is directly related to the cutting of crown molding. Figure 8(a) shows a common cross-sectional view of how crown molding is placed in the corner between a wall and ceiling. Viewing this piece of molding as an upside-down side of a box with sloped sides, we can see that the slope angle of the "box" is α. Therefore, one can use the bevel and miter formulas in this paper to properly cut crown molding.

There are two common varieties of crown molding, determined by the angles α and γ as shown in the figure, where γ is often referred to as the "spring angle." When $\alpha = \gamma = 45^\circ$, the trim is referred to as "45/45°crown," though most crown molding is "52/38°crown," where $\gamma = 38^\circ$. (It is not hard to determine if a given piece of



(a) Illustrating typical crown molding with spring angle γ.

(b) Part of the DeWalt DWS779 angle gauge with tape measure for size reference.

Figure 8 Figures related to crown molding.

trim is 45/45°crown or not; so common are these types that if a given piece is not 45/45°crown, one generally just assumes it is 52/38°crown.)

If we are cutting 52/38°crown for a normal 90° room corner, we would set the bevel to $b(52°, 90°) = 33.86°$ and the miter to $m(52°, 90°) = 31.62°$. So important are these angles that most miter saws have them specially marked on their angle gauges; some saws even have "stops" at these locations, wherein the saw naturally locks into these positions. In Figure 8(b), the miter-angle gauge of the author's saw is shown below a measuring tape. One can see how the miter angle of 31.6° is specially marked; though not shown, the bevel angle of 33.9° is also marked on the bevel-angle gauge. We leave it to the reader to discover the importance of a 35.3° angle.

Other miter saw manuals It is not uncommon for the corners of a room to *not* meet at an exact right angle. Due to imprecision in construction, the angle may be only *near* 90°. Special features in a room, such as a bay window, may require an angle other than 90°. Because of this, most miter saw manuals contain tables of the bevel and miter angles needed to cut crown molding given the interior angle of a corner, i.e., $180° - \varphi$, and the spring angle $\gamma = 90° - \alpha$.

We checked the miter saw manuals from six other major tool manufacturers. All of the manuals, except the one from DeWalt, had such a table. With the formulas of this paper (inspired by the DeWalt manual), one can recreate these tables and compute the proper bevel/miter angles for a corner of, say, 80°. However, one would find it difficult to use just the graph given in the DeWalt manual, as the proper D-curve is not shown.

We also noted two interesting facts about every table in the manuals we checked. First, they all listed bevel/miter angles with two places after the decimal. This level of accuracy is interesting because the bevel and miter angles on every saw are set by hand, where angles are determined by printed guides on the saw. Accuracy to the hundredth of a degree is not feasible; again refer to Figure 8(b) to see the typical size of such guides. Even the digital angle gauges one can buy from hardware stores generally only display one digit after the decimal (and their manuals often include the disclaimer "accurate to ±0.2°.") The given tables seem to give a high degree of accuracy, the applicability of which is lost to most woodworkers.

Second, each table, in every manual, contained errors. Most of the errors in the tables come in the hundredth decimal place, meaning these mistakes will incur no harm to the typical user. What the authors find interesting about these errors is that they are not the result of truncation (i.e., an angle of 32.177° being truncated to 32.17°), but are likely either simple typographical errors or the results of rounding somewhere else in the calculation process. Either way, it seems these tables *were not* generated via a spreadsheet, which would seem to be the natural way to create a table.

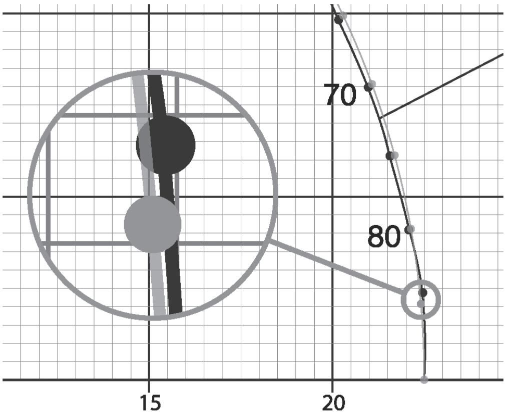


Figure 9 Illustrating an error between the actual bevel/miter settings and the manual.

The DeWalt graphic is also not without error. The authors were able to accurately draw D-curves on top of a .pdf version of the graphic, and while most of the author's points match the manual's points, there are some discrepancies. Perhaps the most significant is the location of the point $D_{45}(85°)$, as seen in Figure 9. The space between grid lines in the graphic represents 0.5°; one can see that the difference in y-values between the manual's point in black and the authors' point in gray is about 0.3°. (One can check that $m(85°, 45°) \approx 2.1°$.) Again, this difference is not large when one considers the manual placement of the bevel angle on the saw. There are enough places for one to make a mistake in building an eight-sided box with sides sloped at 85° that one would be hard pressed to pin any inaccuracies in building on this chart. What is interesting are these questions: how did the writer of the manual generate this graph? In the process of using a computer to create such a beautiful image, how is it that some points are off?

Box sides with different slope angles

Thus far we have investigated finding the proper bevel and miter angles for joining two boards (sides of a box or pieces of crown molding) with the same slope angle α. What if these two boards had different slope angles, such as shown in Figure 10? What bevel and miter angle settings would properly join such boards?

Revisiting Figure 5(b), let the highlighted box side ("box side 1") have a slope angle of α, let the adjacent, transparent box side ("box side 2") have a slope angle of β, and let φ be the exterior angle of the corner. We again seek a unit vector $\vec{e}_1$ normal to the plane containing the face where the two boards meet. We will then use this vector to obtain $\vec{e}_2$, the unit vector normal to the cut edges illustrated in Figure 5(a).

As indicated by the Figure, position box side 1 so that it is parallel to the y-axis. As before,

$$\vec{n}_1 = \langle \sin(\alpha),\ 0,\ -\cos(\alpha) \rangle$$

is a unit normal to the plane containing the exterior face of box side 1. To determine $\vec{n}_2$, a unit normal vector to the plane containing the exterior face of box side 2, we again refer to our previous work determining $\vec{n}_2$, replacing α with β:

$$\vec{n}_2 = \langle \sin(\beta)\cos(\varphi),\ \sin(\beta)\sin(\varphi),\ -\cos(\beta) \rangle.$$

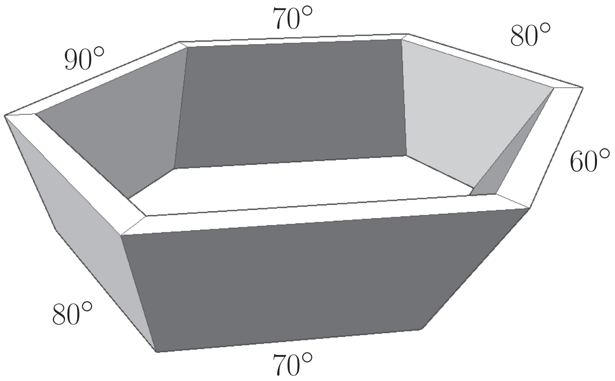


Figure 10 A six-sided box ($\varphi = 60°$) with sides sloped with angles as indicated.

While we cannot rely on "woodworker's wisdom" to determine $\vec{e}_1$, we can use the second method shown previously to find $\vec{e}_1$. Nothing in the second method was dependent on the two sides of the box having the same slope angle. So again, $\vec{n}_2 - \vec{n}_1$ is normal to the plane we seek; the unit vector in this direction is

$$\begin{aligned}\vec{e}_1 &= \frac{\vec{n}_2 - \vec{n}_1}{\|\vec{n}_2 - \vec{n}_1\|}\\ &= \frac{\langle \cos(\varphi)\sin(\beta) - \sin(\alpha), \sin(\varphi)\sin(\beta), \cos(\alpha) - \cos(\beta)\rangle}{\sqrt{\big(\cos(\varphi)\sin(\beta) - \sin(\alpha)\big)^2 + \sin^2(\varphi)\sin^2(\beta) + \big(\cos(\alpha) - \cos(\beta)\big)^2}}\\ &= \frac{\langle \cos(\varphi)\sin(\beta) - \sin(\alpha), \sin(\varphi)\sin(\beta), \cos(\alpha) - \cos(\beta)\rangle}{\sqrt{2\big(1 - \cos(\varphi)\sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta)\big)}}.\end{aligned}$$

As before, $\vec{e}_2$ is the counter-clockwise rotation of $\vec{e}_1$ about the y-axis by an angle of α:

$$\begin{aligned}\vec{e}_2 &= M_y(\alpha)\vec{e}_1\\ &= \Big\langle \cos(\varphi)\cos(\alpha)\sin(\beta) - \sin(\alpha)\cos(\beta), \sin(\varphi)\sin(\beta),\\ &\qquad 1 - \cos(\varphi)\sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta)\Big\rangle\Big/ X, \end{aligned}\tag{11}$$

where

$$X = \sqrt{2\big(1 - \cos(\varphi)\sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta)\big)}.$$

Recalling from equation (4) that

$$\vec{s}_2 = \big\langle -\cos(b)\sin(m), \cos(b)\cos(m), \sin(b)\big\rangle,$$

we can solve for the bevel and miter angles, b and m, respectively, so that $\vec{s}_2 = \vec{e}_2$ as we did before in equations (5) and (8).

To find a formula for the bevel angle, we equate the third components of $\vec{s}_2$ and $\vec{e}_2$, giving

$$b(\alpha, \beta, \varphi) = \sin^{-1}\left(\frac{1}{\sqrt{2}}\sqrt{1 - \cos(\varphi)\sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta)}\right); \tag{12}$$

using identities this can be rewritten as

$$= \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\alpha-\beta}{2}\right)+\sin^2\left(\varphi/2\right)\sin(\alpha)\sin(\beta)}\right). \tag{13}$$

(Equation (11) dictates we restrict α and β from being simultaneously 0°, though we can relax that restriction in our formulation of $b(\alpha, \beta, \varphi)$ in equations (12) and (13) as when both sides have a slope angle of 0°, we use a bevel angle of 0°, as the formulas show.)

Note that these equations are symmetric in terms of α and β; the bevel angle remains the same if box side 1 has slope angle β and box side 2 has slope angle α. We can state this more colloquially: the same bevel angle is used to cut both boards that meet in a corner. An advantage to equation (13) over (12) is that when $\alpha = \beta$, one can more readily see how this equation reduces to equation (5).

To find an equation for the miter angle, we again equate the proportions of the first component of each vector divided by the second component.

From $\vec{s}_2$:

$$\frac{-\cos(b)\sin(m)}{\cos(b)\cos(m)} = -\tan(m). \tag{14}$$

From $\vec{e}_2$ in equation (11):

$$\frac{\cos(\varphi)\cos(\alpha)\sin(\beta)-\sin(\alpha)\cos(\beta)}{\sin(\varphi)\sin(\beta)}. \tag{15}$$

Equating the expressions in equations (14) and (15) leads to

$$m(\alpha, \beta, \varphi) = \tan^{-1}\left(\frac{\sin(\alpha)\cos(\beta)-\cos(\varphi)\cos(\alpha)\sin(\beta)}{\sin(\varphi)\sin(\beta)}\right). \tag{16}$$

Using trigonometric identities, this can be rewritten as

$$\tan^{-1}\left(\frac{\sin(\alpha-\beta)}{\sin(\varphi)\sin(\beta)}+\cos(\alpha)\tan(\varphi/2)\right), \tag{17}$$

where $0° < \beta, \varphi < 180°$.

Note that these equations are not symmetric with respect to α and β; at a corner, each board is mitered differently (unless $\alpha = \beta$). An advantage of equation (17) over equation (16) is that again, when $\alpha = \beta$, one can more readily see how this equation reduces to equation (8).

Equations (13) and (17) give the proper bevel and miter angles for cutting the *left* edge of *box side 1*. It is natural to wonder what bevel and miter angles are needed to cut the matching *right* edge of *box side 2*. We again leave it to the reader to confirm that this right edge can be cut with bevel $-b(\beta, \alpha, \varphi)$ and miter $-m(\beta, \alpha, \varphi)$; not only is there a sign change, the roles of α and β are switched.

This leads to the following understanding of the bevel and miter equations: let α be the slope angle of the box side you are cutting, and let β be the slope angle of the other side of the corner. If one is cutting the left edge, use bevel and miter angles $b(\alpha, \beta, \varphi)$ and $m(\alpha, \beta, \varphi)$, respectively; if cutting the right edge, use the opposite of these angles.

For example, consider the corner shown at the top of Figure 10, where the 90°- and 70°-sloped sides meet (and $\varphi = 60°$). The left edge of the 70° board is cut with

$$\text{bevel: } b(70°, 90°, 60°) \approx 30.99°$$

$$\text{miter: } m(70°, 90°, 60°) \approx -11.17°.$$

When cutting the 90° board, we have $\alpha = 90°$ and $\beta = 70°$; since we are cutting the right edge, we use the opposite of the b and m functions:

$$\text{bevel:}\ -b(90°, 70°, 60°) \approx -30.99°$$

$$\text{miter:}\ -m(90°, 70°, 60°) \approx -22.80°.$$

One item we find interesting is that the base of the 70°-sloped side is wider than its top. In the other figures of boxes with sloped sides, the base of each side has always been narrower than the top. The negative miter angle found when cutting the *left* edge means we miter "the opposite of the normal direction," giving a wider base.

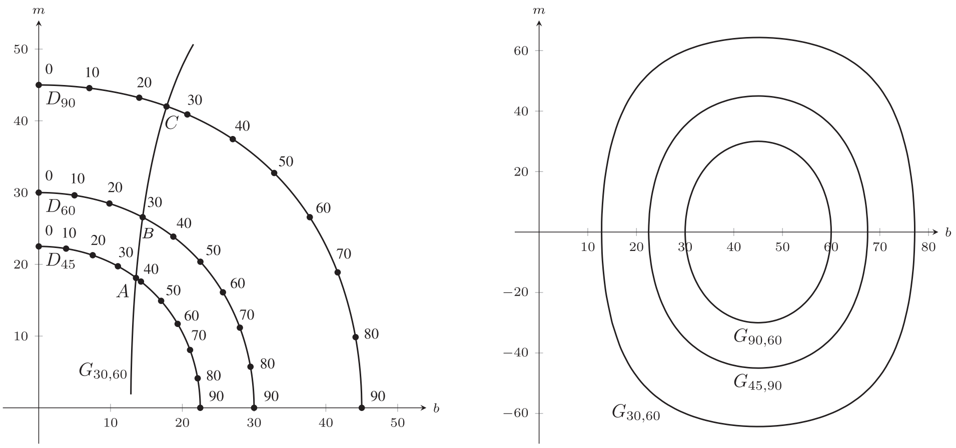
One of us made a four-sided box where each side has a different slope angle using these formulas to verify their correctness.

The box shown in Figure 10 is a 3-D model made using SketchUp®. Each edge was beveled/mitered separately according to the formulas, then the sides were placed together.

Generalized *D*-curves We can generate curves from equations (12) and (16) by fixing two of the variables and letting a third vary. Letting α vary and fixing β and φ, we say the parametric equations

$$x(\alpha) = b(\alpha, \beta, \varphi), \quad y(\alpha) = m(\alpha, \beta, \varphi), \quad 0° \leq \alpha \leq 360°, \quad 0° < \beta, \varphi < 180°$$

determine a G-curve (for *g*eneralized D-curve), denoted $G_{\beta,\varphi}$.



(a) Drawing $G_{30,60}$ along with D_{90}, D_{60} and D_{45}.

(b) Drawing $G_{30,60}$, $G_{90,60}$ and $G_{45,90}$.

Figure 11 Illustrating G-curves.

Figure 11(a) shows a portion of $G_{30,60}$. This curve shows the bevel and miter angles needed to cut the left edge of a box side, where the exterior angle of the corner is $\varphi = 60°$ and the other box side has a slope angle of 30°. This G-curve is plotted along with three D-curves; note their intersections at points A, B, and C.

The intersection point B is "expected." It lies on D_{60}, giving the bevel and miter angles to create a 6-sided box where both sides have a slope angle of 30°. That point also lies on $G_{30,60}$, where $\alpha = \beta = 30°$.

The intersection points A and C are less expected. At A, we find a bevel and miter angle combination that serves two purposes. First, lying on D_{45}, the combination allows one to build an eight-sided box with slope angle of $\approx 37.8°$. Second, as A lies on $G_{30,60}$, this bevel/miter combination allows one to build the corner of a box with an exterior angle of 60°, where the "left-hand" side of the corner has a slope angle of 30° and the "right-hand" side has a slope angle of $\approx 25.1°$. A similar statement can be made about the point C: cutting a board with that bevel/miter combination is the start of making a 4-sided box with all sides sloped at $\approx 25.7°$ or the start of making a particular six-sided box.

In Figure 11(b), three G-curves are drawn for $0° \leq \alpha \leq 360°$. We make two observations about this figure, each leading to follow-up opportunities.

First, these curves are centered around the point $(45°, 0°)$. We wonder why, both mathematically and intuitively. How do the equations create the observed symmetry? How could one have anticipated this by simply thinking about the construction of box corners?

Second, as β varies from 30° to 90°, the G-curve $G_{30,60}$ continuously deforms to $G_{90,60}$. In the process of that deformation, one expects intersections of $G_{\beta,60}$ with the drawn curve $G_{45,90}$ for various values of β. However, it seems they intersect for only one value of β, namely

$$\hat{\beta} = \cos^{-1}\left(1/\sqrt{3}\right) \approx 54.74°;$$

that is, it seems that the curve $G_{45,90}$ is identical to the curve $G_{\hat{\beta},60}$. A first challenge is to show this is true, a second is to understand why.

What's next?

While the equations generating D-curves are well known, the equations of miter and bevel angles for corners with different slope angles do not seem to appear in the literature. As such, we wonder if the equations given are "the best." That is, the equations for the bevel angle in equations (12) and (13) have a certain beauty and simplicity to them; the equations for the miter angle in equations (16) and (17) do not seem as "nice." Using trigonometric identities, is there are better way to express these formulas?

Finally, there is much to discover about G-curves. We are especially interested in understanding when $G_{\beta_1,\varphi_1} = G_{\beta_2,\varphi_2}$ for $\beta_1 \neq \beta_2$, $\varphi_1 \neq \varphi_2$.

REFERENCES

[1] DeWalt DWS 779 Manual. (2015). DeWalt Industrial Tool Co.

[2] Inav, DW718_12_Double Bevel Sliding Compound Miter Saw. (2017). 3D Warehouse, Trimble Inc.

[3] Vautaw, W. R. (2008). Two problems with table saws. *College Math. J.* 39(2): 121–128. doi.org/10.1080/07468342.2008.11922285

Summary. Compound miter saws are used to cut crown molding and to make boxes with equally-sloped sides. Given the slope of the molding or the side of the box, we derive the equations for the proper miter and bevel angles to form the correct corner. The equations are generalized to form boxes where the sides have different slopes.

KAREN BLISS is an associate professor of applied mathematics at the Virginia Military Institute. She likes home improvement projects, including minor electrical, plumbing, flooring, etc. One of her favorite classes to teach is Calculus III, which is one of the reasons she found this project so enjoyable.

GREG HARTMAN is a professor of applied mathematics at Virginia Military Institute. He enjoys woodworking and home renovation projects. One of his favorite classes to teach is Calculus III, which is one of the reasons he found this project so enjoyable.

The Group of Primitive Pythagorean Triples and Perplex Numbers

SOMPHONG JITMAN
Department of Mathematics
Faculty of Science
Silpakorn University
Nakhon Pathom 73000, Thailand
sjitman@gmail.com

EKKASIT SANGWISUT
Department of Mathematics and Statistics
Faculty of Science
Thaksin University
Phattalung 93110, Thailand
ekkasit@tsu.ac.th

A *Pythagorean triple* (a, b, c) is an ordered triple of integers satisfying the equation $a^2 + b^2 = c^2$. Alternatively, such a triple can be represented by a right triangle with legs of lengths a and b and hypotenuse of length c.

In the case where $\gcd(a, b, c) = 1$, such a Pythagorean triple is called primitive. Otherwise, (a, b, c) is called non-primitive. It is not difficult to see that every non-primitive Pythagorean triple can be written as a multiple of a primitive one. Precisely, a Pythagorean triple is of the form $(ka, kb, kc) = k(a, b, c)$ for some primitive Pythagorean triple (a, b, c) and positive integer k. Therefore, the primitive Pythagorean triple (a, b, c) will be used to represent the Pythagorean triples (ka, kb, kc) for all positive integers k. It is well-known that if (a, b, c) is a primitive Pythagorean triple, then exactly one of the integers a or b is even, and c is always odd. In the rest of this paper, for each primitive Pythagorean triple (a, b, c), we assume that a and c are odd positive integers and b is an even integer (which can be zero or negative).

Let $\mathcal{P}$ be the set of all primitive Pythagorean triples and let (a, b, c) and (d, e, f) be elements in $\mathcal{P}$. Then $a^2 + b^2 = c^2$ and $d^2 + e^2 = f^2$. Equivalently, $a^2 = c^2 - b^2$ and $d^2 = f^2 - e^2$. The following identity ensures that the product of two differences of squares is again a difference of squares:

$$a^2d^2 = (c^2 - b^2)(f^2 - e^2) = (be + cf)^2 - (bf + ce)^2. \tag{1}$$

By equation (1), a new Pythagorean triple of the form $(ad, bf + ce, be + cf)$ can be obtained from the primitive Pythagorean triples (a, b, c) and (d, e, f). It follows that the primitive representation of $(ad, bf + ce, be + cf)$ is an element in $\mathcal{P}$.

Let $\oplus$ be a binary operation on $\mathcal{P}$ defined by

$$(a, b, c) \oplus (d, e, f) = (u, v, w), \tag{2}$$

where (u, v, w) is the primitive representation of $(ad, bf + ce, be + cf)$. For example, by equation (1), the primitive Pythagorean triples $(3, 4, 5)$ and $(3, -4, 5)$ produce the Pythagorean triple $(9, 0, 9)$, whose primitive representation is $(1, 0, 1)$. Hence, $(3, 4, 5) \oplus (3, -4, 5) = (1, 0, 1)$. Some further examples of the addition operation $\oplus$ defined in equation (2) are given as follows:

$$(3, 4, 5) \oplus (3, 4, 5) = (9, 40, 41),$$

Math. Mag. **95** (2022) 285–293. doi:10.1080/0025570X.2022.2092383 © Mathematical Association of America
MSC: 11A99

$$(3,4,5) \oplus (1,0,1) = (3,4,5),$$
$$(3,4,5) \oplus (5,12,13) = (15,112,113).$$

In 1962, W. Sierpinski [7] posed the question "How many primitive Pythagorean triples have the same hypotenuse?" The question was answered by E. J. Eckert in 1984 [2]. We shall focus on a parallel question about the odd legs of Pythagorean triples. Precisely, we consider two questions "How many primitive Pythagorean triples have the same odd leg?" and "How can we construct such primitive Pythagorean triples?" We first show that $(\mathcal{P}, \oplus)$ is a free abelian group with identity $(1,0,1)$ and in which the inverse of (a,b,c) is $(a,-b,c)$. Based on the group structure of $(\mathcal{P}, \oplus)$, we show how to construct all primitive Pythagorean triples having the same fixed odd leg.

Perplex numbers

In this section, we present the definition and basic properties of perplex numbers (or hyperbolic numbers). For more details about perplex numbers, the reader is referred to the works of Fjelstad [3], Harkin [4] Poodiak [6], or Sobczyk [8].

A *perplex number* is a number of the form $c + bh$, where b and c are real numbers and h is an indeterminate such that $h^2 = 1$. For a given perplex number $z = c + bh$, c is called the *real part* of z and b is called the *hyperbolic part* of z. The addition and multiplication of perplex numbers $c + bh$ and $f + eh$ are defined as usual by

$$(c + bh) + (f + eh) = (c + f) + (b + e)h$$
$$(c + bh)(f + eh) = (cf + be) + (ce + bf)h.$$

The conjugate of $z = c + bh$ is defined to be $\bar{z} = c - bh$. The multiplication of a perplex number and its conjugate is a real number of the form

$$z \cdot \bar{z} = (c + bh)(c - bh) = c^2 - b^2,$$

which can be negative, zero, or positive. In particular, $z \cdot \bar{z} = 0$ if $c = b$, $z \cdot \bar{z} > 0$ if $c > b$, and $z \cdot \bar{z} < 0$ otherwise. The magnitude of z is defined by

$$|z| = \sqrt{|z \cdot \bar{z}|} = \sqrt{|c^2 - b^2|}. \tag{3}$$

Let $z = c + bh$ and $z' = f + eh$ be perplex numbers such that $|z| = a$ and $|z'| = d$. Then

$$|(ce + bf)^2 - (cf + be)^2| = |zz'|^2 = |z|^2|z'|^2 = |c^2 - b^2||e^2 - f^2|.$$

If $c > b$ and $f > e$, then

$$(cd + bf)^2 - (cf + be)^2 = (c^2 - b^2)(e^2 - f^2),$$

which is relevant to equation (1). For a given perplex number $z = c + bh$ such that $c > b$ and $|z| = a$, z can be identified with the point (c, b) on the hyperbola $x^2 - y^2 = a^2$. The line from the origin to the point (c, b) cuts the unit hyperbola $x^2 - y^2 = 1$ at the points $(\frac{c}{a}, \frac{b}{a})$ and $(-\frac{c}{a}, -\frac{b}{a})$ as shown in Figure 1.

The points $(\frac{c}{a}, \frac{b}{a})$ and $(-\frac{c}{a}, -\frac{b}{a})$ can be viewed in terms of the hyperbolic functions sinh and cosh. Precisely,

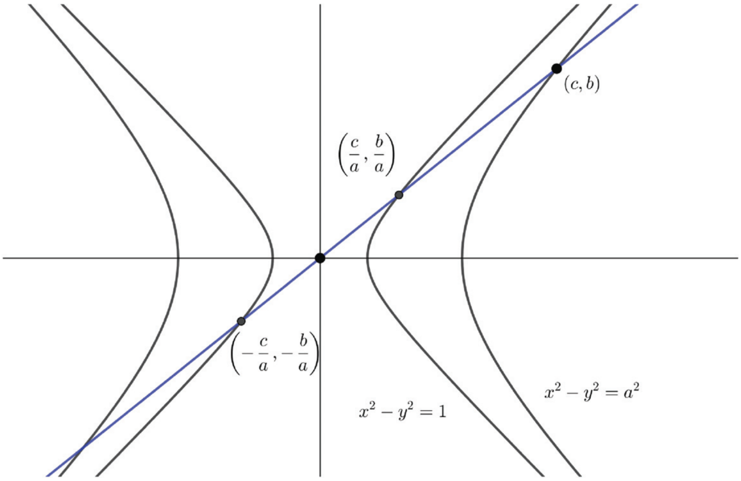

(c, b)
$\left(\frac{c}{a}, \frac{b}{a}\right)$
$\left(-\frac{c}{a}, -\frac{b}{a}\right)$
$x^2 - y^2 = 1$
$x^2 - y^2 = a^2$

Figure 1 Perplex numbers can be identified with points on a hyperbola.

$$\left(\frac{c}{a}, \frac{b}{a}\right) = (\cosh\alpha, \sinh\alpha)$$
$$\left(-\frac{c}{a}, -\frac{b}{a}\right) = (-\cosh\alpha, -\sinh\alpha)\,,$$

where $\alpha/2$ is the shaded area in Figure 2.

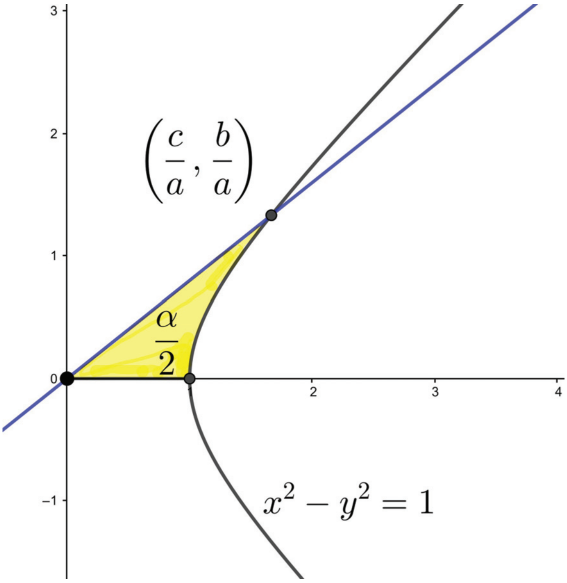


Figure 2 The geometrical interpretation of the quantity α.

Primitive Pythagorean triples and perplex numbers

We will now establish a connection between primitive Pythagorean triples and perplex numbers.

Let (a, b, c) be a primitive Pythagorean triple. Then $a^2 = c^2 - b^2$. Based on equation (3), the triple (a, b, c) is identified with the perplex number $z = c + bh$ of magnitude a. By dividing z by its magnitude a, we get that $(z/a) = (c/a) + (b/a)h$, and $\left(\frac{c}{a}, \frac{b}{a}\right)$ is a point on the unit hyperbola $x^2 - y^2 = 1$.

For given primitive Pythagorean triples (a, b, c) and (d, e, f),

$$(a, b, c) \text{ is represented by } \left(\frac{c}{a}, \frac{b}{a}\right) = (\cosh\alpha, \sinh\alpha)$$

and

$$(d, e, f) \text{ is represented by } \left(\frac{f}{d}, \frac{e}{d}\right) = (\cosh\beta, \sinh\beta),$$

for some α and β.

From equation (2), recall that $(a, b, c) \oplus (d, e, f) = (ad, bf + ce, be + cf)$, which is represented by

$$\left(\frac{be + cf}{ad}, \frac{bf + cd}{ad}\right) = (\cosh\gamma, \sinh\gamma)\,,$$

where $\cosh\gamma$ and $\sinh\gamma$ are determined by the hyperbolic addition formulas,

$$\begin{aligned}\cosh(\alpha+\beta) &= \cosh\alpha\cosh\beta + \sinh\alpha\sinh\beta \\ &= \left(\frac{c}{a}\right)\left(\frac{f}{d}\right) + \left(\frac{b}{a}\right)\left(\frac{e}{d}\right) = \frac{cf + be}{ad} = \cosh\gamma \\ \sinh(\alpha+\beta) &= \sinh\alpha\cosh\beta + \cosh\alpha\sinh\beta \\ &= \left(\frac{b}{a}\right)\left(\frac{f}{d}\right) + \left(\frac{c}{a}\right)\left(\frac{e}{d}\right) = \frac{bf + ce}{ad} = \sinh\gamma.\end{aligned}$$

The group $(\mathcal{P}, \oplus)$ and primitive Pythagorean triples with odd leg

We now show that $(P, \oplus)$ is a free abelian group. Subsequently, the enumeration and construction of the primitive Pythagorean triples on a fixed odd leg are given.

First, we recall Euclid's formula for generating primitive Pythagorean triples.

Theorem 1. *([1, Theorem 12.1]). Primitive Pythagorean triples* (a, b, c) *are uniquely determined by the formulas* $a = m^2 - n^2,\ b = \pm 2mn,\ c = m^2 + n^2$ *for integers* $m > n > 0$ *such that* $\gcd(m, n) = 1$ *and* $m \not\equiv n \pmod 2$.

In the case where the odd leg of a primitive Pythagorean triple is a prime power, the next corollary can be obtained as an immediate consequence of Theorem 1.

Corollary 1. *Let* p^k *be an odd prime power. Then the primitive Pythagorean triples of odd leg* p^k *are of the forms*

$$\left(p^k,\ \pm\frac{p^{2k} - 1}{2},\ \frac{p^{2k} + 1}{2}\right).$$

Proof. By Theorem 1, we have $p^k = m^2 - n^2 = (m + n)(m - n)$ for some integers m and n. Then $m - n = p^\ell$ and $m + n = p^{k-\ell}$ for some $k > 2\ell$. It follows that

$$m = \frac{p^{k-\ell} + p^\ell}{2} \qquad \text{and} \qquad n = \frac{p^{k-\ell} - p^\ell}{2}.$$

If $\ell > 0$, then $\gcd(m, n) \neq 1$, implying that (a, b, c) is not primitive. Thus, ℓ must be 0, implying that $m = (p^k + 1)/2$ and $n = (p^k - 1)/2$. Again by Theorem 1, we have

$$b = \pm 2mn = \frac{p^{2k} - 1}{2} \qquad \text{and} \qquad c = m^2 + n^2 = \frac{p^{2k} + 1}{2},$$

as desired. ■

Next, we show that $(P, \oplus)$ is an abelian group.

Theorem 2. *The set $\mathcal{P}$ of primitive Pythagorean triples together with the binary operation $\oplus$ defined in equation (2) forms an abelian group.*

Proof. Let (a, b, c) and (d, e, f) be elements in $\mathcal{P}$. Then

$$(ad)^2 + (bf + ce)^2 = (be + cf)^2$$

by equation (1), which implies that the primitive representation $(a, b, c) \oplus (d, e, f)$ of $(ad, bf + ce, be + cf)$ is an element in $\mathcal{P}$. Hence, the set $\mathcal{P}$ is closed under $\oplus$. The proofs of associativity and commutativity are straightforward, and we omit the details.

Let (a, b, c) be an element in $\mathcal{P}$. Then

$$(a, b, c) \oplus (1, 0, 1) = (a, b + 0, 0 + c) = (a, b, c)$$

which implies that $(1, 0, 1)$ is the identity in $\mathcal{P}$. Since

$$\begin{aligned}(a, b, c) \oplus (a, -b, c) &= (a^2, bc + (-cb), -b^2 + c^2)\\ &= (a^2, 0, a^2) = (1, 0, 1),\end{aligned}$$

we have that $(a, -b, c)$ is the inverse of (a, b, c). ■

The following lemma plays a crucial role in our proof of the main result in Theorem 3.

Lemma 1. *Let b, c, e, f be integers, and let p be an odd positive integer such that $bcef \not\equiv 0 \pmod{p^2}$ and $c^2 - b^2 \equiv f^2 - e^2 \equiv 0 \pmod{p^2}$. Then exactly one of the following statements holds.*

1. $bf + ce \equiv be + cf \equiv 0 \pmod{p^2}$.
2. $bf - ce \equiv cf - be \equiv 0 \pmod{p^2}$.

Proof. It is not difficult to see that

$$\begin{aligned}(bf + ce)(bf - ce) &= b^2 f^2 - c^2 e^2\\ &= b^2(f^2 - e^2) - e^2(c^2 - b^2) \equiv 0 \pmod{p^2}.\end{aligned}$$

We show that $p \mid (bf + ce)$ or $p \mid (bf - ce)$, but not both. Assume that $p \mid (bf + ce)$ and $p \mid (bf - ce)$. Then $bf + ce = pm$ and $bf - ce = p\ell$ for some integers m and ℓ. Hence, $p \mid 2bf$ and $p \mid 2ce$, which implies that $p^2 \mid bcef$, a contradiction. We now consider the following two cases.

Case One. $bf + ce \equiv 0 \pmod{p^2}$. Then $(bf + ce)^2 \equiv 0 \pmod{p^4}$ and we have

$$-(bf + ce)^2 + (be + cf)^2 = (f^2 - e^2)(c^2 - b^2) \equiv 0 \pmod{p^4}.$$

Hence, $(be + cf)^2 \equiv 0 \pmod{p^4}$. As desired, we have $be + cf \equiv 0 \pmod{p^2}$.

Case Two. $bf - ce \equiv 0 \pmod{p^2}$. Then $(bf - ce)^2 \equiv 0 \pmod{p^4}$, which implies that

$$-(bf - ce)^2 + (cf - be)^2 = (f^2 - e^2)(c^2 - b^2) \equiv 0 \pmod{p^4}.$$

Hence, $(cf - be)^2 \equiv 0 \pmod{p^4}$, implying that $cf - be \equiv 0 \pmod{p^2}$. ■

Note that p^k is odd for all odd positive integers p and positive integers k which implies that Lemma 1 holds true for all integers p^k.

By applying Corollary 1, Theorem 2, and Lemma 1, we obtain a formula for the primitive Pythagorean triples on a fixed odd leg.

Theorem 3. *$(\mathcal{P}, \oplus)$ is a free abelian group generated by the set of primitive Pythagorean triples of odd prime leg. Moreover, each primitive Pythagorean triple (a, b, c) can be written in the form*

$$(a,b,c) = r_1\left(p_1, \pm\frac{p_1^2-1}{2}, \frac{p_1^2+1}{2}\right) \oplus \cdots \oplus r_k\left(p_k, \pm\frac{p_k^2-1}{2}, \frac{p_k^2+1}{2}\right)$$

where the prime factorization of a is $a = p_1^{r_1} p_2^{r_2} \dots p_k^{r_k}$, k and r_i are positive integers, the numbers $p_1, p_2, \dots, p_k$ are distinct odd primes, and the signs $\pm$ are all independent of each other.

Proof. Let $(a, b, c) \in \mathcal{P}$ be such that $(a, b, c) \neq (1, 0, 1)$. Write the prime factorization of a.

$$a = p_1^{r_1} p_2^{r_2} \dots p_k^{r_k}.$$

We prove the statement by induction on k. For the basis step, assume that $a = p_1^{r_1}$. By Corollary 1, we then have

$$\begin{aligned}(p_1{}^{r_1}, b, c) &= \left(p_1{}^{r_1},\ \pm\frac{p_1{}^{2r_1}-1}{2},\ \frac{p_1{}^{2r_1}+1}{2}\right)\\ &= r_1\left(p_1,\ \pm\frac{p_1^2-1}{2},\ \frac{p_1^2+1}{2}\right).\end{aligned}$$

For the inductive step, let $k \geq 2$ be an integer, and assume that the statement holds true for all positive integers less than k. Write $a = p_1^{r_1} q$, where $q = p_2^{r_2} \dots p_k^{r_k}$. Since $p_1^{r_1}$ is an odd prime power, the primitive Pythagorean triples having odd leg $p_1^{r_1}$ are

$$\left(p_1^{r_1},\ \pm\frac{p_1^{2r_1}-1}{2},\ \frac{p_1^{2r_1}+1}{2}\right),$$

by Corollary 1. For convenience, write $e = (p_1^{2r_1} - 1)/2$ and $f = (p_1^{2r_1} + 1)/2$. Then $bcef \not\equiv 0 \pmod{p_1^{2r_1}}$ and $c^2 - b^2 \equiv f^2 - e^2 \equiv 0 \pmod{p_1^{2r_1}}$. Since p^{r_1} is an odd positive integer, the conditions in Lemma 1 are satisfied. This leads to two distinct cases.

Case One. $bf + ce \equiv be + cf \equiv 0 \pmod{p^{2r_1}}$. It follows that $(bf + ce)/p_1^{2r_1}$ and $(be + cf)/p_1^{2r_1}$ are integers. Suppose that

$$p_i \mid \gcd\left(q,\ \frac{bf+ce}{p_1^{2r_1}},\ \frac{be+cf}{p_1^{2r_1}}\right)$$

for some $2 \leq i \leq k$. Then

$$c = \frac{c(e^2-f^2)}{p_1^{2r_1}} = e\frac{bf+ce}{p_1^{2r_1}} - f\frac{be+cf}{p_1^{2r_1}}$$

is divisible by p_i, which is contrary to our assumption that $\gcd(q, c) = 1$. Consequently, we have

$$\gcd\left(q, \frac{bf+ce}{p_1^{2r_1}}, \frac{be+cf}{p_1^{2r_1}}\right) = 1.$$

From equation (2), it follows that

$$\left(q, \frac{bf+ce}{p_1^{2r_1}}, \frac{be+cf}{p_1^{2r_1}}\right) = (p_1^{r_1}q, b, c) \oplus (p_1^{r_1}, e, f) \in \mathcal{P}$$

is the primitive representation of $\left(p_1^{2r_1}q, bf+ce, be+cf\right)$. As $\mathcal{P}$ is an abelian group, we have

$$(p_1^{r_1}q, b, c) = (p_1^{r_1}, -e, f) \oplus \left(q, \frac{bf+ce}{p_1^{2r_1}}, \frac{be+cf}{p_1^{2r_1}}\right). \tag{4}$$

Since $q = p_2^{r_2} \dots p_k^{r_k}$, by the induction hypothesis, we have

$$\begin{aligned}\left(q, \frac{bf+ce}{p_1^{2r_1}}, \frac{be+cf}{p_1^{2r_1}}\right) &= r_2\left(p_2, \pm\frac{p_2^2-1}{2}, \frac{p_2^2+1}{2}\right)\\ &\qquad \oplus \cdots \oplus r_k\left(p_k, \pm\frac{p_k^2-1}{2}, \frac{p_k^2+1}{2}\right),\end{aligned}$$

where the signs ± are all independent of each other. From (4), it follows that

$$\begin{aligned}(a, b, c) &= (p_1^{r_1}q, b, c)\\ &= (p_1^{r_1}, -e, f) \oplus \left(q, \frac{bf+ce}{p_1^{2r_1}}, \frac{be+cf}{p_1^{2r_1}}\right)\\ &= \left(p_1^{r_1}, -\frac{p_1^{2r_1}-1}{2}, \frac{p_1^{2r_1}+1}{2}\right) \oplus \left(q, \frac{bf+ce}{p_1^{2r_1}}, \frac{be+cf}{p_1^{2r_1}}\right)\\ &= r_1\left(p_1, -\frac{p_1^2-1}{2}, \frac{p_1^2+1}{2}\right) \oplus r_2\left(p_2, \pm\frac{p_2^2-1}{2}, \frac{p_2^2+1}{2}\right)\\ &\qquad \oplus \cdots \oplus r_k\left(p_k, \pm\frac{p_k^2-1}{2}, \frac{p_k^2+1}{2}\right).\end{aligned}$$

Case Two. $bf - ce \equiv cf - be \equiv 0 \pmod{p^{2r_1}}$. It follows that $(bf-ce)/p_1^{2r_1}$ and $(-be+cf)/p_1^{2r_1}$ are integers. Suppose that

$$p_i \mid \gcd\left(q, \frac{bf-ce}{p_1^{2r_1}}, \frac{-be+cf}{p_1^{2r_1}}\right)$$

for some $2 \le i \le k$. Then

$$c = \frac{c(e^2-f^2)}{p_1^{2r_1}} = -e\frac{bf-ce}{p_1^{2r_1}} - f\frac{-be+cf}{p_1^{2r_1}}$$

is divisible by p_i, which is contrary to our assumption that $\gcd(c, q) = 1$. We therefore have

$$\gcd\left(q, \frac{bf-ce}{p_1^{2r_1}}, \frac{-be+cf}{p_1^{2r_1}}\right) = 1.$$

From equation (2), it follows that

$$\left(q,\ \frac{bf-ce}{p_1^{2r_1}},\ \frac{-be+cf}{p_1^{2r_1}}\right) = (p_1^{r_1}q,\ b,\ c) \oplus (p_1^{r_1},\ -e,\ f) \in \mathcal{P}$$

is the primitive representation of $\left(p_1^{2r_1}q,\ bf-ce,\ -be+cf\right)$. Since $\mathcal{P}$ is an abelian group, we have

$$(p_1^{r_1}q,\ b,\ c) = (p_1^{r_1},\ e,\ f) \oplus \left(q,\ \frac{bf-ce}{p_1^{2r_1}},\ \frac{-be+cf}{p_1^{2r_1}}\right). \tag{5}$$

Since $q = p_2^{r_2} \dots p_k^{r_k}$, by the induction hypothesis, we have

$$\left(q,\ \frac{bf-ce}{p_1^{2r_1}},\ \frac{-be+cf}{p_1^{2r_1}}\right) = r_2\left(p_2,\ \pm\frac{p_2^2-1}{2},\ \frac{p_2^2+1}{2}\right) \oplus \dots \oplus r_k\left(p_k,\ \pm\frac{p_k^2-1}{2},\ \frac{p_k^2+1}{2}\right),$$

where the signs ± are all independent of each other. From equation (5), it follows that

$$\begin{aligned}(a,b,c) &= (p_1^{r_1}q, b, c)\\ &= (p_1^{r_1}, e, f) \oplus \left(q, \frac{bf-ce}{p_1^{2r_1}}, \frac{-be+cf}{p_1^{2r_1}}\right)\\ &= \left(p_1^{r_1}, \frac{p_1^{2r_1}-1}{2}, \frac{p_1^{2r_1}+1}{2}\right) \oplus \left(q, \frac{bf-ce}{p_1^{2r_1}}, \frac{-be+cf}{p_1^{2r_1}}\right)\\ &= r_1\left(p_1, \frac{p_1^2-1}{2}, \frac{p_1^2+1}{2}\right) \oplus r_2\left(p_2, \pm\frac{p_2^2-1}{2}, \frac{p_2^2+1}{2}\right)\\ &\qquad \oplus \dots \oplus r_k\left(p_k, \pm\frac{p_k^2-1}{2}, \frac{p_k^2+1}{2}\right).\end{aligned}$$

From the two cases, it can be concluded that

$$(a,b,c) = r_1\left(p_1, \pm\frac{p_1^2-1}{2}, \frac{p_1^2+1}{2}\right) \oplus \dots \oplus r_k\left(p_k, \pm\frac{p_k^2-1}{2}, \frac{p_k^2+1}{2}\right),$$

where the signs ± are all independent of each other. ■

From Theorem 3, the number of primitive Pythagorean triples with the same odd leg $a = p_1^{r_1}p_2^{r_2}\dots p_k^{r_k}$ is 2^k, and their explicit representations are of the forms

$$(a,b,c) = r_1\left(p_1, \pm\frac{p_1^2-1}{2}, \frac{p_1^2+1}{2}\right) \oplus \dots \oplus r_k\left(p_k, \pm\frac{p_k^2-1}{2}, \frac{p_k^2+1}{2}\right),$$

where the signs ± are all independent of each other.

Here is an illustrative example:

Example. Let $a = 4725 = 3^3 \cdot 5^2 \cdot 7$. Then there are $2^3 = 8$ primitive Pythagorean triples with leg 4725 of the forms

$$\begin{aligned}(4725, b, c) &= 3(3, \pm 4, 5) \oplus 2(5, \pm 12, 13) \oplus (7, \pm 24, 25)\\ &= (27, \pm 364, 365) \oplus (25, \pm 312, 313) \oplus (7, \pm 24, 25).\end{aligned}$$

Therefore, the primitive Pythagorean triples with leg 4725 are

- $(27, 364, 365) \oplus (25, 312, 313) \oplus (7, 24, 25) = (4725, 11162812, 11162813)$,
- $(27, -364, 365) \oplus (25, 312, 313) \oplus (7, 24, 25) = (4725, 14948, 15677)$,
- $(27, 364, 365) \oplus (25, -312, 313) \oplus (7, 24, 25) = (4725, 17548, 18173)$,
- $(27, 364, 365) \oplus (25, 312, 313) \oplus (7, -24, 25) = (4725, 227788, 227837)$,
- $(27, -364, 365) \oplus (25, -312, 313) \oplus (7, 24, 25) = (4725, -227788, 227837)$,
- $(27, 364, 365) \oplus (25, -312, 313) \oplus (7, -24, 25) = (4725, -14948, 15677)$,
- $(27, -364, 365) \oplus (25, 312, 313) \oplus (7, -24, 25) = (4725, -17548, 18173)$,
- $(27, -364, 365) \oplus (25, -312, 313) \oplus (7, -24, 25) = (4725, -11162812, x)$, where $x = 11162813$.

Acknowledgments The authors would like to thank the anonymous referees for helpful comments and suggestions.

REFERENCES

[1] Burton, D. M. (2007). *Elementary Number Theory.* 6nd. ed. New York: McGraw-Hill.
[2] Eckert, E. J. (1984). The group of primitive Pythagorean triangles. *Math. Mag*. 57(1): 22–27. doi.org/10.1080/0025570X.1984.11977070
[3] Fjelstad, P. (1986). Extending special relativity via the perplex numbers. *Amer. J. Phys*. 54(5): 416–422. doi.org/10.1119/1.14605
[4] Harkin, A. A., Harkin J. B. (2004). Geometry of generalized complex numbers. *Math. Mag*. 77(2): 118–129. doi.org/10.1080/0025570X.2004.11953236
[5] Sporn, H. (2017). Pythagorean triples, complex numbers, and perplex numbers. *College Math. J.* 48(2): 115–122. https://doi.org/10.4169/college.math.j.48.2.115
[6] Podiack, R. D., LeClair, K. J. (2009). Fundamental theorems of algebra for the perplexes. *College Math. J.* 40(5): 322–335. doi.org/10.4169/074683409X475643
[7] Sierpinski, W. (1962). *Pythagorean Triangles*. The Scripta Mathematia Studies, no 9. New York: Yeshia Univ.
[8] Sobczyk, G. (1995). The hyperbolic number plane. *College Math. J.* 26(4): 268–280. doi.org/10.1080/07468342.1995.11973712

Summary. We study a group structure on primitive Pythagorean triples, as well as its applications to the construction and enumeration of primitive Pythagorean triples on a fixed odd leg. We determine the number of such triples and show how to list them.

SOMPHONG JITMAN received his Ph.D. in mathematics from Chulalongkorn University in 2011. Since 2013, he has been a lecturer in the Department of Mathematics, Faculty of Science, Silpakorn University, Thailand. Prior to that, he was a Research Fellow with the Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. His research interests include algebra, number theory, and their applications to classical and quantum coding theory.

EKKASIT SANGWISUT received his Ph.D. in mathematics from Chulalongkorn University in 2014. He has been a lecturer with the Department of Mathematics and Statistics, Faculty of Science, Thaksin University, Thailand since 2015. His research interests are in commutative ring theory, polynomials, and their applications in algebraic coding theory.

Routh's Theorem, Morgan's Story, and New Patterns of Area Ratios

ELIAS ABBOUD
Beit Berl College
Doar Beit Berl 44905, Israel
eabboud@beitberl.ac.il

Ryan Morgan, who was a high school student in 1994, discovered the theorem which states: If each side of a triangle is divided into n equal parts, where n is odd, then the area of the hexagon formed by connecting the cevians from the vertices to the two central division points on the opposite sides equals $8/(9n^2 - 1)$ times the area of the triangle [3].

Morgan's theorem is a generalization of Marion's theorem that concerns the case $n = 3$ [6]. The hexagon resulting from trisecting the sides has one-tenth the area of the original triangle.

Morgan's theorem and Marion's theorem are classified as theorems of *affine* geometry, which deals with affine transformations that preserve properties such as collinearity of points, parallelism of lines, the ratios of lengths of line segments, and the ratios of areas [2].

The proofs of both theorems depend on Routh's theorem from affine geometry.

Theorem 1. *(Routh's Theorem) If the sides BC, CA, AB of a triangle ABC are divided at L, M, N in the respective ratios $\lambda : 1$, $\mu : 1$, $\nu : 1$, then the cevians AL, BM, CN form a triangle whose area is*

$$\frac{(\lambda\mu\nu - 1)^2}{(\lambda\mu + \lambda + 1)(\mu\nu + \mu + 1)(\nu\lambda + \nu + 1)}$$

times that of ABC.

In his classic geometry textbook, Coxeter emphasized that this result was discovered by Steiner, but simultaneously cited two references [2, p. 211]. The first was Steiner's original work [5, pp. 163–168] and the second was Routh's work [4, p. 82]. Later in his book, he referred to the result as "Routh's theorem" [2, p. 219].

Let us return to the story of Ryan Morgan. According to an article in *The Baltimore Sun* newspaper, Morgan formulated the conjecture (now theorem) when he was 15 [8]. The proof of the theorem was published two years later by Watanabe, Hanson, and Nowosielski [7] by applying Routh's theorem several times.

In recent work by the present author, the case of dividing the sides of a triangle in the respective ratios $1 : \lambda : 1$, $\lambda > 0$, was explored, and nice expressions of area ratios were obtained [1]. In particular, the theorems of Morgan and Marion were derived as special cases.

In this article, we use the same method to explore the case of dividing the sides of a triangle in the respective ratios $1 : \lambda : 2$, $\lambda > 0$. We show that the area ratio of the resulting hexagon to the original triangle is given by a rational function with cubic polynomials in both the numerator and denominator. We then study some special values of λ and obtain a geometric interpretation of the corresponding values of the rational function.

Math. Mag. **95** (2022) 294–301. doi:10.1080/0025570X.2022.2093575 © Mathematical Association of America
MSC: 51N10

Barycentric coordinates

We begin with a brief summary of barycentric coordinates. For more details we refer the reader to Coxeter [2, pp. 218–220]. Coxeter gave a general proof of Theorem 1, attributed to Möbius, using barycentric coordinates. These are *homogeneous* coordinates (t_1, t_2, t_3), where t_1, t_2, t_3 are masses at the vertices of a reference triangle $A_1A_2A_3$. In particular $(1, 0, 0)$ is A_1, $(0, 1, 0)$ is A_2, $(0, 0, 1)$ is A_3, and (t_1, t_2, t_3) corresponds to a point P such that the areas of the triangles PA_2A_3, PA_3A_1, PA_1A_2 are proportional to the barycentric coordinates t_1, t_2, t_3 of P, respectively (see Figure 1). If $t_1 + t_2 + t_3 = 1$, then the normalized barycentric coordinates (t_1, t_2, t_3) are called *areal* coordinates. In this case, the areas of the triangles PA_2A_3, PA_3A_1, PA_1A_2 are t_1, t_2, t_3 times the area of the whole triangle $A_1A_2A_3$, respectively.

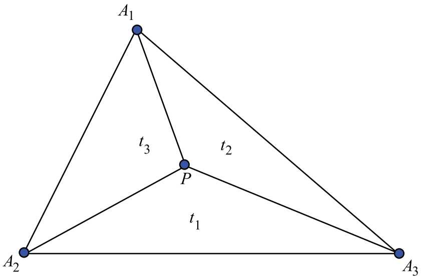


Figure 1 The areas of the triangles $PA_2A_3, PA_3A_1, PA_1A_2$ are proportional to the barycentric coordinates t_1, t_2, t_3, respectively.

In barycentric coordinates, a line has a linear homogeneous equation. We can find the equation for the line joining two given points A and B with barycentric coordinates (r_1, r_2, r_3) and (s_1, s_2, s_3), respectively, by calculating the determinant:

$$\begin{vmatrix} r_1 & r_2 & r_3 \\ s_1 & s_2 & s_3 \\ t_1 & t_2 & t_3 \end{vmatrix} = 0.$$

Moreover, if C is a point dividing the segment AB in the ratio $\alpha : \beta$, then the homogeneous barycentric coordinates of C are

$$\alpha(r_1, r_2, r_3) + \beta(s_1, s_2, s_3) = (\alpha r_1 + \beta s_1, \alpha r_2 + \beta s_2 + \alpha r_3 + \beta s_3).$$

Finally, in terms of areal coordinates, with the reference triangle as unit, the area of the triangle with vertices (q_1, q_2, q_3), (r_1, r_2, r_3), and (s_1, s_2, s_3) is given by the determinant

$$\begin{vmatrix} q_1 & q_2 & q_3 \\ r_1 & r_2 & r_3 \\ s_1 & s_2 & s_3 \end{vmatrix}.$$

These ideas will be applied in the next section.

New patterns of area ratios

Suppose now that the sides of a triangle $A_2A_3A_1$ are divided at $A_{i,1}$, $A_{i,2}$, $1 \leq i \leq 3$ in the respective ratios $1 : \lambda : 2$, $\lambda > 0$, as shown in Figure 2.

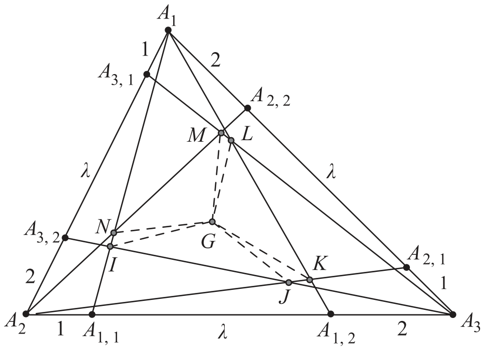


Figure 2 Dividing the sides of the triangle in the ratios $1 : \lambda : 2$.

Let I, J, K, L, M, N be the points of intersection of the corresponding cevians, as shown in Table 1.

TABLE 1: Cevians and points of intersection.

point	cevians
I	$A_1A_{1,1} \cap A_3A_{3,2}$
J	$A_2A_{2,1} \cap A_3A_{3,2}$
K	$A_2A_{2,1} \cap A_1A_{1,2}$
L	$A_3A_{3,1} \cap A_1A_{1,2}$
M	$A_2A_{2,2} \cap A_3A_{3,1}$
N	$A_1A_{1,1} \cap A_2A_{2,2}$

Let G be the center of gravity of $A_1A_2A_3$. Since the barycentric coordinates of A_1, A_2, A_3 are $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, respectively, the barycentric coordinates of G are $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. By the symmetry of the division, the area of hexagon $IJKLMN$ is three times the area of the quadrilateral $GLMN$, which is the union of the triangles GMN and GLM. Readers are encouraged to check out this fact for themselves, by imitating the steps below.

Therefore, it is sufficient to compute the barycentric coordinates of L, M, and N. To find the area ratio of the hexagon $IJKLMN$ to the triangle $A_2A_3A_1$, we should follow the following steps.

STEP ONE. Find the barycentric coordinates of $A_{i,1}$ and $A_{i,2}$, $1 \leq i \leq 3$.

The points $A_{i,1}$, $1 \leq i \leq 3$ divide the sides A_2A_3, A_3A_1, A_1A_2 by the ratio $1 : \lambda + 2$, respectively, and the points $A_{i,2}$, $1 \leq i \leq 3$, divide the sides A_2A_3, A_3A_1, A_1A_2 by the ratio $1 + \lambda : 2$, respectively. Table 2 shows the respective barycentric coordinates.

TABLE 2: The respective barycentric coordinates.

point	barycentric coordinates
$A_{1,1}$	$(0, \lambda + 2, 1)$
$A_{2,1}$	$(1, 0, \lambda + 2)$
$A_{3,1}$	$(\lambda + 2, 1, 0)$
$A_{1,2}$	$(0, 2, \lambda + 1)$
$A_{2,2}$	$(\lambda + 1, 0, 2)$
$A_{3,2}$	$(2, \lambda + 1, 0)$

Step Two. Compute the equations of the cevians $A_1A_{1,1}$, $A_1A_{1,2}$, $A_2A_{2,2}$, and $A_3A_{3,1}$.

The cevian $A_1A_{1,1}$ has equation

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & \lambda+2 & 1 \\ t_1 & t_2 & t_3 \end{vmatrix} = 0.$$

Computing this determinant we get $-t_2 + (\lambda + 2)t_3 = 0$. The cevian $A_1A_{1,2}$ has the equation

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 2 & \lambda+1 \\ t_1 & t_2 & t_3 \end{vmatrix} = 0.$$

Equivalently, $-(\lambda + 1)t_2 + 2t_3 = 0$. The cevian $A_2A_{2,2}$ has the equation

$$\begin{vmatrix} 0 & 1 & 0 \\ \lambda+1 & 0 & 2 \\ t_1 & t_2 & t_3 \end{vmatrix} = 0.$$

After computing the determinant, we get the equation $-2t_1 + (\lambda + 1)t_3 = 0$. Finally, the cevian $A_3A_{3,1}$ has the equation

$$\begin{vmatrix} 0 & 0 & 1 \\ \lambda+2 & 1 & 0 \\ t_1 & t_2 & t_3 \end{vmatrix} = 0,$$

which simplifies into the equation $-t_1 + (\lambda + 2)t_2 = 0$.

Step Three. Find the barycentric coordinates of the points L, M, and N.

Since L is the intersection of the cevians $A_3A_{3,1} \cap A_1A_{1,2}$, we have to solve the system

$$\begin{cases} -t_1 + (\lambda + 2)t_2 = 0 \\ -(\lambda + 1)t_2 + 2t_3 = 0. \end{cases}$$

Substituting $t_2 = 2$, we get $t_1 = 2(\lambda + 2)$ and $t_3 = \lambda + 1$. Hence, $L = (2(\lambda + 2), 2, \lambda + 1)$.

Likewise, M is the intersection of the cevians $A_2A_{2,2} \cap A_3A_{3,1}$. Therefore, we have to solve the following system of two equations:

$$\begin{cases} -2t_1 + (\lambda + 1)t_3 = 0 \\ -t_1 + (\lambda + 2)t_2 = 0. \end{cases}$$

Substituting $t_1 = (\lambda + 1)(\lambda + 2)$ we get $t_3 = 2(\lambda + 2)$ and $t_2 = \lambda + 1$. Hence, $M = ((\lambda + 1)(\lambda + 2), \lambda + 1, 2(\lambda + 2))$.

Similarly, N is the intersection of the cevians $A_1A_{1,1} \cap A_2A_{2,2}$. Therefore, we have to solve the following system of two equations:

$$\begin{cases} -t_2 + (\lambda + 2)t_3 = 0 \\ -2t_1 + (\lambda + 1)t_3 = 0. \end{cases}$$

Substituting $t_3 = 2$ we get $t_2 = 2(\lambda + 2)$ and $t_1 = (\lambda + 1)$. Hence, $N = (\lambda + 1, 2(\lambda + 2), 2)$.

STEP FOUR. Compute the areas of triangles GMN and GLM, and normalize by dividing each determinant by the product of the sums of the rows.

The areas of GMN and GLM are proportional to the respective determinants

$$\begin{vmatrix} 1/3 & 1/3 & 1/3 \\ (\lambda+1)(\lambda+2) & \lambda+1 & 2(\lambda+2) \\ \lambda+1 & 2(\lambda+2) & 2 \end{vmatrix} = \frac{2}{3}\lambda^3 + \frac{5}{3}\lambda^2 - \frac{7}{3}$$

and

$$\begin{vmatrix} 1/3 & 1/3 & 1/3 \\ 2(\lambda+2) & 2 & \lambda+1 \\ (\lambda+1)(\lambda+2) & \lambda+1 & 2(\lambda+2) \end{vmatrix} = \frac{1}{3}\lambda^3 - \frac{1}{3}\lambda^2 - 3\lambda - \frac{7}{3}.$$

To find the ratio of the area of the hexagon $IJKLMN$ to the area of the triangle $A_2A_3A_1$ we have to normalize the barycentric coordinates by dividing each one of the determinants by the product of the sums of the rows. Hence, the area of GMN is

$$h(\lambda) = \frac{\frac{2}{3}\lambda^3 + \frac{5}{3}\lambda^2 - \frac{7}{3}}{(\lambda^2 + 6\lambda + 7)(3\lambda + 7)},$$

times the area of the triangle $A_2A_3A_1$. Besides, the area of GLM is

$$g(\lambda) = \frac{\frac{1}{3}\lambda^3 - \frac{1}{3}\lambda^2 - 3\lambda - \frac{7}{3}}{(\lambda^2 + 6\lambda + 7)(3\lambda + 7)},$$

times the area of the triangle $A_2A_3A_1$.

Therefore, the area of the hexagon $IGKLMN$ is

$$f(\lambda) = 3h(\lambda) + 3g(\lambda) = \frac{3\lambda^3 + 4\lambda^2 - 9\lambda - 14}{(\lambda^2 + 6\lambda + 7)(3\lambda + 7)}$$

times the area of the triangle $A_2A_3A_1$.

Notice that $g(\lambda) = 0$ implies that

$$\frac{1}{3}\lambda^3 - \frac{1}{3}\lambda^2 - 3\lambda - \frac{7}{3} = \frac{1}{3}(\lambda + 1)\left(\lambda^2 - 2\lambda - 7\right) = 0.$$

Hence, the polynomial has one positive root at $\lambda = 2\sqrt{2} + 1 \approx 3.828$. In this case, the areas of the triangles GLM, GNI, and GJK equal 0 and the hexagon degenerates to a triangle whose area is $f(2\sqrt{2} + 1) = \frac{10}{7} - \frac{6}{7}\sqrt{2} \approx 0.216$, times the area of $A_2A_3A_1$ (see Figure 3).

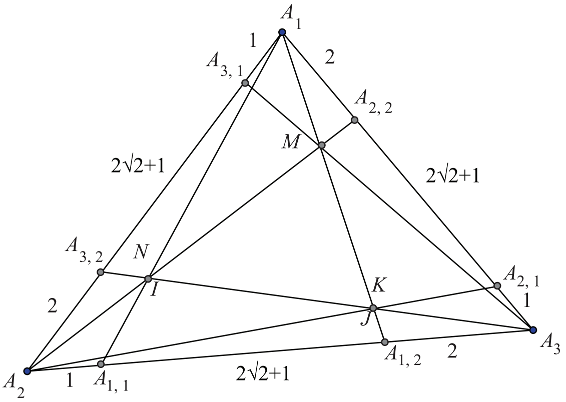


Figure 3 If $\lambda = 2\sqrt{2} + 1$ then the hexagon degenerates to a triangle.

Summarizing the above results, we have the following theorem:

Theorem 2. *Suppose that the sides of a triangle $A_2A_3A_1$ are divided at $A_{i,1}$, $A_{i,2}$, $1 \leq i \leq 3$ in the respective ratios $1 : \lambda : 2$. If the division points $A_{i,1}$, $A_{i,2}$, $1 \leq i \leq 3$, are connected to the opposite vertices then for $\lambda \geq 2\sqrt{2} + 1$ the ratio of the area of the resulting hexagon to the area of the triangle is*

$$f(\lambda) = \frac{3\lambda^3 + 4\lambda^2 - 9\lambda - 14}{(\lambda^2 + 6\lambda + 7)(3\lambda + 7)}.$$

If $\lambda = 2\sqrt{2} + 1$, then the hexagon degenerates to a triangle whose area is $\frac{10}{7} - \frac{6}{7}\sqrt{2}$ times the area of $A_2A_3A_1$.

Other values of *f* What happens in Theorem 2 if $\lambda < 2\sqrt{2} + 1$? We aim to explain the results geometrically. In this case, the triangles GMN and GLM in Figure 2 overlap. Indeed, reducing the values of λ causes the points L and M to move closer until they meet at $\lambda = 2\sqrt{2} + 1$. Then, these two points begin to move apart from each other and the triangle GLM changes its direction. The interchange of two rows in the corresponding determinant of GLM (step 4) will change its sign, thus explaining the negative values of the function g for $0 < \lambda < 2\sqrt{2} + 1$ (see Figure 4, which has been cut down to focus on the positive roots of the functions f, g and h).

As mentioned previously, $g(\lambda) = 0$ implies $\lambda = 2\sqrt{2} + 1 \approx 3.828$ and the hexagon degenerates to a triangle. On the other hand, $h(\lambda) = 0$ implies $\frac{2}{3}\lambda^3 + \frac{5}{3}\lambda^2 - \frac{7}{3} = \frac{1}{3}(\lambda - 1)\left(7\lambda + 2\lambda^2 + 7\right) = 0$. Hence, h has one real root at $\lambda = 1$. In this case the triangles GMN, GIJ, and GKL in Figure 2, degenerate and the hexagon takes the shape of three triangles with a common vertex at G (see Figure 5).

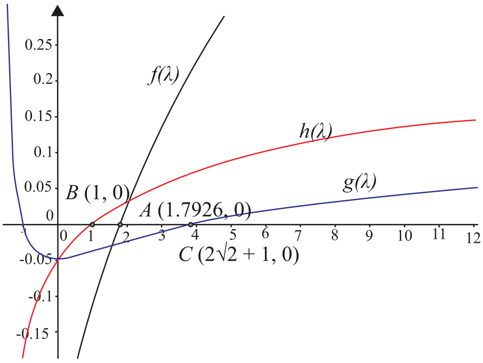


Figure 4 The graphs of the functions f, g, and h.

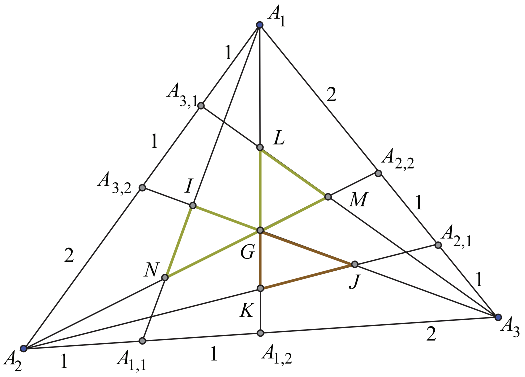


Figure 5 Dividing the sides of the triangle by the ratios 1 : 1 : 2

Finally, $f(\lambda) = 0$ implies that $3\lambda^3 + 4\lambda^2 - 9\lambda - 14 = 0$. This equation has one real root at $\lambda_0 \approx 1.7926$. The geometric interpretation of the fact that the function f vanishes at $\lambda_0 \approx 1.7926$ leads to the following observation: For $\lambda_0 < \lambda < 2\sqrt{2} + 1$, the values of f express the ratio of three times the difference $S_{GMN} - S_{GML}$, and the area of the whole triangle (S_{GMN} and S_{GML} denotes the areas of GMN and GML, respectively). If $\lambda = \lambda_0 \approx 1.7926$, then the function f vanishes at λ_0 and hence $S_{GMN} = S_{GLM}$ (see Figure 6).

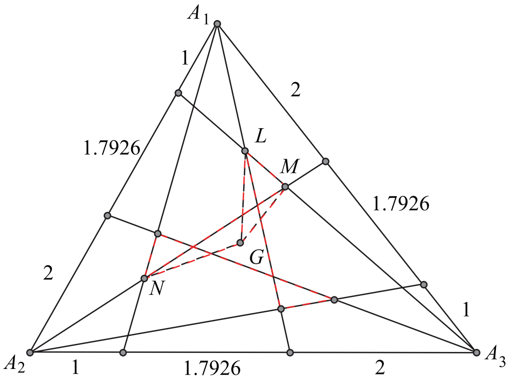


Figure 6 The triangles GMN and GML have equal area.

Thus, we have the following result.

Corollary 1. *If the sides of a triangle are divided in the respective ratios* $1 : \lambda_0 : 2$, *where* $\lambda_0 \approx 1.7926$ *is the real root of the polynomial* $3\lambda^3 + 4\lambda^2 - 9\lambda - 14$, *then the triangles* GMN *and* GML *in Figure 6 have equal area.*

Concluding remarks

The readers are encouraged to draw a dynamic figure similar to Figure 2, where the sides of the triangle are divided in the ratios $1 : \lambda : 2$. This can help with exploring the above results. The readers are also encouraged to give a geometric interpretation, in view of the above discussion, to the negative values of the function f in the interval $0 < \lambda < \lambda_0 \approx 1.7926$.

Steps 1–4 in the previous section allow us to generalize the pattern by dividing the sides of the triangle in the ratios $1 : \lambda : \alpha$, for any positive real numbers λ and α. The following expression generalizes the corresponding formula in Theorem 2:

$$\frac{(\alpha+1)\lambda^3 + 2\alpha\lambda^2 - 3\left(\alpha^3 - \alpha^2 - \alpha + 1\right)\lambda - 2\left(\alpha^4 - \alpha^3 - \alpha + 1\right)}{(\lambda^2 + 2(\alpha+1)\lambda + \alpha^2 + \alpha + 1)((\alpha+1)\lambda + \alpha^2 + \alpha + 1)}.$$

The details are left to the reader.

Acknowledgments The author is indebted to the referee for his valuable comments that improved the exposition of the paper and for providing the link of Morgan's story. Special thanks are due to the Editor for his constructive comments and suggestions.

REFERENCES

[1] Abboud, E. (2015). On the Routh-Steiner theorem and some generalisations. *Math. Gaz.* 99(544): 45–53. doi:10.1017/mag.2014.6

[2] Coxeter, H. S. M. (1969). *Introduction to Geometry*, 2nd ed. New York: Wiley.

[3] Morgan, R. (1994). No restriction needed. *Math. Teacher* 87(9): 726.

[4] Routh, E. J. (1896). *A Treatise on Analytical Statics with Numerous Examples*, Vol. 1, 2nd. ed. Cambridge: Cambridge Univ. Press.

[5] Steiner, J. (1882). *Gesammelte Werke*, Vol. 1. Berlin: Reimer.

[6] Walter, M. (1993). Reader reflections: Marion's theorem. *The Mathematics Teacher* 86(8): 619.

[7] Watanabe, T., Hanson, R., Nowosielski, F. D. (1996). Morgan's theorem. *Math. Teacher* 89(5): 420–423.

[8] Maushard, M. (1994). Something new in a triangle. *The Baltimore Sun*. Available at: https://www.baltimoresun.com/news/bs-xpm-1994-12-20-1994354105-story.html. Accessed June 2022.

Summary. We explore new patterns of area ratios discovered from dividing the sides of a triangle in the ratios $1 : \lambda : 2$. We obtain a rational function with cubic polynomials in both the numerator and denominator and give a geometric interpretation of some if its unique values.

ELIAS ABBOUD (MR Author ID: 249090) received his D.Sc. from the Technion, Israel. Since 1992, he has taught mathematics at Beit Berl College. Between the years 2010–2017 he served as the Mathematics Chair in the Arab Academic Institution within the Faculty of Education of Beit Berl College. Since 2001, he has worked partially at the Academic Arab College of Education-Haifa.

When Does Chaos Appear While Driving? Learning Dynamical Systems Via Car-Following Models

J. ALBERTO CONEJERO
Instituto Universitario de Matemática Pura y Aplicada
Universitat Politècnica de València
46022, València, Spain
aconejero@upv.es

MARINA MURILLO-ARCILA
Instituto Universitario de Matemática Pura y Aplicada
Universitat Politècnica de València
46022, València, Spain
mamuar1@upv.es

JESÚS M. SEOANE
Nonlinear Dynamics, Chaos and Complex Systems Group
Departamento de Física, Universidad Rey Juan Carlos
Tulipán s/n, 28933 Móstoles, Madrid, Spain
jesus.seoane@urjc.es

JUAN B. SEOANE-SEPÚLVEDA
Instituto de Matemática Interdisciplinar (IMI)
Plaza de las Ciencias 3, 28040 Madrid, Spain
jseoane@mat.ucm.es

To the loving memory of Dr. D. José Manuel Seoane Capote (1942–2016)

In mathematics, a *dynamical system* is a system in which a function describes the time dependence of a point in a geometrical space. There are many examples of this, including the very famous mathematical model describing the swinging of a clock pendulum, which started with Galileo's research in 1602, and the famous three-dimensional Lorenz attractor, which provided the earliest example of chaos in a dynamical system in the early 1960s.

Teaching dynamical systems is not a simple task. At the moment, there are too many potential examples and models that one could use to present this notion. Not all of them are accessible to all due either to their technicality or to their complexity and level of abstraction. The most common way to introduce this notion at an undergraduate level is by means of the famous Lotka-Volterra system. The classical Lotka-Volterra system is a two-dimensional system in which the concept of stability can be easily presented by means of the typical example of a predator/prey situation. For instance, a typical example is seen in

$$\text{(LV)} \qquad \begin{cases} x'(t) = a_1 x(t) - a_2 x(t) y(t) \\ y'(t) = -b_1 y(t) + b_2 x(t) y(t), \end{cases}$$

with $a_1, a_2, b_1, b_2 > 0$, and Figure 1 shows the typical trajectories of such a system.

In this paper, we are interested in innovating by employing a different approach to teaching dynamical systems. Towards that end, we propose the use of the "not that typical" car-following models.

Math. Mag. **95** (2022) 302–313. doi:10.1080/0025570X.2022.2092382 © Mathematical Association of America
MSC: Primary 34-01, Secondary 93A30

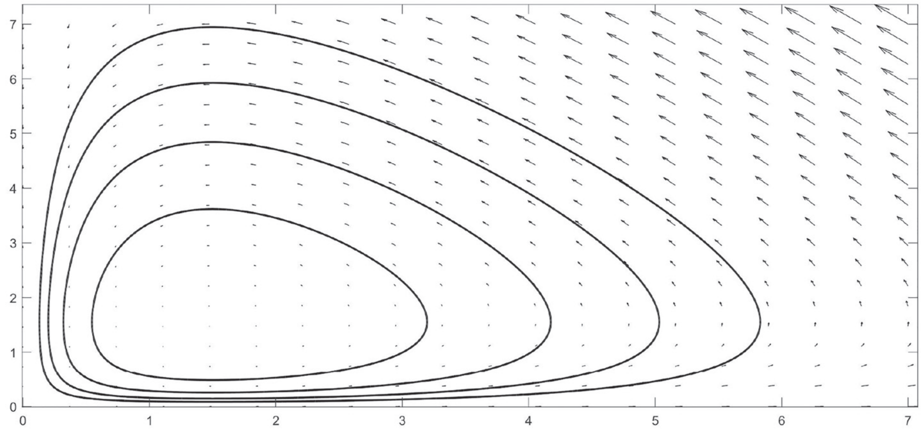


Figure 1 Trajectories for a the classical Lotka-Volterra system such as system (LV) for values $a_1 = 1.4$, $a_2 = 0.9$, $b_1 = 1.8$, and $b_2 = 1.2$.

Car-following models were introduced with the intention of describing a driver's reaction to the changes in speed of the car in front of him on a single lane. Modeling this behavior is necessary for the development of traffic flow theory. The first car-following models were due to Greenshields [15, 16] in the 1930s. During the 50s and 60s, car-following models were refined by taking into account considerations involved in driving a motor vehicle on a lane [9]. These considerations include the difference between the velocities of a car and the car in front of it, the distance of a car with respect to the preceding one, and the driver's reaction time. See, for instance, Forbes [11] and Pipes [23]. Chandler et al. [6] and Herman et al. [18] proposed a mathematical model that assumes the acceleration of the following car in each two-vehicle unit is linearly proportional to the cars' relative velocities at some earlier time, with a fixed time lag of transmission of the driver-vehicle system. This model is well-known as the *Quick-Thinking Driver* QTD model.

In practice, the acceleration of a car depends not only on the velocity of the car in front, but also slightly on the velocity of a car two ahead, as it is considered in the nearest and *Next-Nearest* (NN) model. It can also be modeled taking into account the speeds of the cars that go in front of and behind it, as is considered in the *Forward and Backward Control* FBC model. An interested reader can find a discussion of the historical evolution of these models in Brackstone and McDonald [5] and Hoogendoorn and Bovy [17].

Chaos is closely linked with car-following models. Even in a simple model like QTD, it is possible to find chaos relating its dynamics to certain solutions of the logistic equation [20, 21]. Such a model is a particular case of a more general nonlinear car-following model developed by Gazis, Herman, and Rothery (GHR) for General Motors [13, 24]. The discontinuous behavior of some of its solutions suggested the existence of chaos for a certain range of input parameters. Other authors studied the existence of chaos for this model under various assumptions: Disbro and Frame [10] showed chaos for the (GHR) model without taking into account signals, bottlenecks, intersections, etc. or with a coordinated signal network. In Addison and Low [1], and in Addison et al. [2], chaos was observed for a platoon of vehicles described by the (GHR) model when adding a nonlinear inter-car separation dependent term.

More recently, Barrachina et al. [4] and Conejero, Murillo-Arcila, and Seoane-Sepúlveda [7] used techniques from semigroup theory to study the existence of chaos in different car-following models for an infinite number of cars driving on a road. It is worth mentioning that chaos can also be found when studying traffic models at a macroscopic level, as is the case in the Lighthill-Whitham-Richards equation [8].

Our concern here is to study the dynamics of the continuous dynamical system that represents the behavior of cars driving on a road when considering some classical car-following models, such as the QTD and NN models. More concretely, we determine their equilibria and stability in terms of the parameters involved in the models. Moreover, we illustrate the outcome with numerical solutions.

The models we propose can be used in teaching for many different applications such as the study of dynamical systems and differential equations and the improvement of computing skills with mathematical software such as Maple, Matlab, R, or Mathematica. It is also useful as a proposal for developing skills in model formulation, solution, and interpretation.

Preliminaries

We first introduce the models that we are going to study. In the basic formulation of any of the car-following models, there is a relation between the acceleration of a car and the difference between its velocity and the velocity of the car in front of it. In a basic formulation, the driver of a car adjusts her speed according to the relative velocity between her car and the one in front. That is,

$$x_1''(t+t_1) = \lambda_1(x_2'(t) - x_1'(t)), \tag{1}$$

where $x_2(t)$ denotes the position of the car which goes in front of car 1 at time t and whose position is given by $x_1(t)$, t_1 denotes the reaction time of driver 1, and the positive number λ_1 is a sensitivity coefficient that measures how strongly driver 1 responds to the acceleration of the car in front of her. Usually, λ_1 lies between 0.3 and $0.4\,\mathrm{s}^{-1}$ [5]. Under the assumption that all drivers react "very quickly," one can take $t_1 = 0$. This is known as the *Quick-Thinking-Driver* (QTD) model.

$$x_1''(t) = \lambda_1(x_2'(t) - x_1'(t)), \tag{2}$$

This model can be reformulated in terms of velocities $u_1(t) = x_1'(t)$ and $u_2(t) = x_2'(t)$, and we have

$$u_1'(t+t_1) = \lambda_1(u_2(t) - u_1(t)), \tag{3}$$

It can also be improved by taking into account that the reaction time depends on the speed of the car, as is done in McCartney [20] and in McCartney and Gibson [21, p. 92]. This leads us to formulate a modified version of it:

$$u_1'(t) = \gamma_1 u_1(t)(u_2(t) - u_1(t)), \tag{4}$$

As is indicated by McCartney and Gibson [21], the models in equation (3) and equation (4) should provide the same acceleration for the same relative velocity. For instance, in the case of a car moving at 45 km/h, about 13m/s, we take $13\gamma_1 = \lambda_1$ in order to ensure that models (3) and (4) predict the same acceleration, and typical values of γ_1 will be in the range of 0.02–0.03 s^{-1}. For further details on driving simulations, we refer readers to the excellent handbook edited by Fisher et al. [12, Chapters 5, 7, and 12].

It is also interesting to investigate the effect of a control that involves the car two ahead in addition to the car in front. This model is known as the *nearest and next-nearest* (NN) model, and it is given by

$$u_1'(t) = \lambda_{1,1}(u_2(t) - u_1(t)) + \lambda_{1,2}(u_3(t) - u_1(t)), \tag{5}$$

in which $\lambda_{1,1}$ stands for the sensitivity coefficient in relation to the car ahead and $\lambda_{1,2}$ the one with the car two ahead, so that $\lambda_{1,1} + \lambda_{1,2}$ plays the role of λ_1.

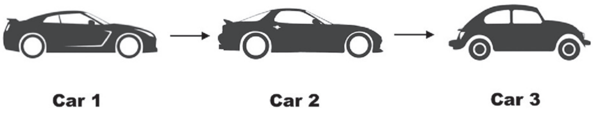


Figure 2 The case of three cars, with the speed of the leading car constant and equal to v, and the speed of the others behind, $u_2(t)$ and $u_1(t)$, respectively. Designed by Freepik.

Analogously, one can also improve it in the same way as (4). This yields

$$u_1'(t) = \gamma_{1,1} u_1(t)(u_2(t) - u_1(t)) + \gamma_{1,2} u_1(t)(u_3(t) - u_1(t)), \tag{6}$$

with $\gamma_{1,1} + \gamma_{1,2}$ instead of γ_1.

Some results related to the stability of dynamical systems will be needed. Let us consider a dynamical system on $\mathbb{R}^n_{+,0}$ of the form $x' = f(x)$, $x \in \mathbb{R}^n_{0,+}$, and let f be a differentiable function. We recall that an equilibrium point x_0 is called *hyperbolic* if all the eigenvalues of the Jacobian matrix $J(x_0)$ have nonzero real part. Such a point is called a *sink* if all the eigenvalues of $J(x_0)$ have negative real part. It is said to be a *source* if all the eigenvalues have positive real part, and it is a *saddle point* if it is a hyperbolic point and has at least one eigenvalue with positive real part and one with negative real part. In terms of stability, sinks correspond to asymptotically stable equilibria points. Hyperbolic equilibrium points are *unstable* if and only if they are saddles or sources. The stability of nonhyperbolic equilibrium points is more difficult to determine, and it is typically necessary to use the famous Lyapunov functions. In the next section, we study the equilibria of the Quick-Thinking-Driver model with three cars and we also analyze its trajectories and the ones of the perturbed model by adding an oscillation term.

We then study the situation of three cars with the leading car moving at a fixed speed, and with two cars following behind in a line. Such a model can be perturbed in order to provide a cyclic orbit to which the speeds converge.

The next case adds an additional car, and we compare the QTD and NN models. We present two situations: In the first, the cars start with different speeds, but as time goes by, their speeds tend to the one of the leading car. In the second, we perturbate the speed of the leading car, and we analyze the propagation of that perturbation along the cars on the lane.

We close by proposing some possible extensions and class activities.

The Quick-Thinking-Driver model with three cars

First, we study the stability of the QTD model with three cars, with the leading car driving at constant speed. Here, the model describing the speed of the cars behind the leading car can be described by the following two equations:

$$\begin{cases} u_1'(t) = \gamma_1 u_1(t)(u_2(t) - u_1(t)) \\ u_2'(t) = \gamma_2 u_2(t)(v - u_2(t)). \end{cases} \tag{7}$$

Solving the system

$$\begin{cases} 0 = \gamma_1 u_1(u_2 - u_1) \\ 0 = \gamma_2 u_2(v - u_2), \end{cases} \tag{8}$$

we obtain the following three equilibrium points: $P_1 = (0, 0)$, $P_2 = (0, v)$, $P_3 = (v, v)$, which can be seen by looking at the nullclines, see Figure 3. Looking at the phase plane, we can appreciate that P_3 is an attractor. To confirm this, we now calculate the eigenvalues associated to the equilibrium points P_1, P_2, and P_3. The Jacobian of the system is given by

$$J(u_1, u_2) = \begin{pmatrix} \gamma_1(u_2 - 2u_1) & \gamma_1 u_1 \\ 0 & \gamma_2(v - 2u_2) \end{pmatrix}. \tag{9}$$

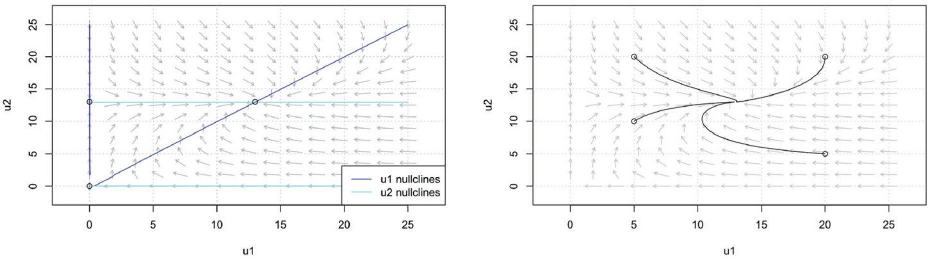


Figure 3 On the left, we see the phase plane, equilibrium points, and nullclines for the solution of system (7), where $\gamma_1 = \gamma_2 = 0.03$, and $v = 13$ m/s. On the right, we see the trajectories for the initial conditions (5, 10), (5, 20), (20, 5), and (20, 20) after $t = 15$ s.

We next obtain the eigenvalues associated to $J(P_i)$, $i = 1, \ldots, 3$. $J(P_1)$ has a null and a positive eigenvalue $\gamma_1 v$, but $J(P_2)$ has a negative $(-\gamma_2 v)$ and a positive $(\gamma_2 v)$ eigenvalue. Anyway, both P_1 and P_2 are unstable. In contrast, P_3 has both eigenvalues, $-\gamma_1 v$ and $-\gamma_2 v$, with negative real part, and it is a stable equilibrium point for the system. This can be seen also by looking at the trajectories depicted in Figure 3.

If we perturb the speed of the leading car by a term $\sin(t)$, then we get

$$\begin{cases} u_1'(t) = \gamma_1 u_1(t)(u_2(t) - u_1(t)) \\ u_2'(t) = \gamma_2 u_2(t)(v + \sin(t) - u_2(t)). \end{cases} \tag{10}$$

the speed falls into a cycle around the point (13, 13), see Figure 4, which is similar to what happens with the Lotka-Volterra trajectories, but it is obtained in this case through a nonautonomous dynamical system. This can be related to the existence of pullback attractors, see Harraga and Yebdri [19], for example. More information about the topological connections of both systems in the discrete case can be found in Balibrea [3]. From a physical point of view, these kinds of perturbations are very typical in modeling dynamical systems in the presence of external perturbations as it occurs, for instance, with the pendulum [14]. In this specific case, this perturbation can be seen in situations in which the driver accelerates or decelerates as a consequence of a multi-car collision which occurs when the velocities are equal.

Stability of traffic models with three cars

Let us analyze the stability of the dynamical systems that model how three cars will progress in time, following a car at fixed speed, when we consider the QTD and the NN models. We assume that the leading car goes at constant speed v, which is followed by cars 3, 2, and 1.

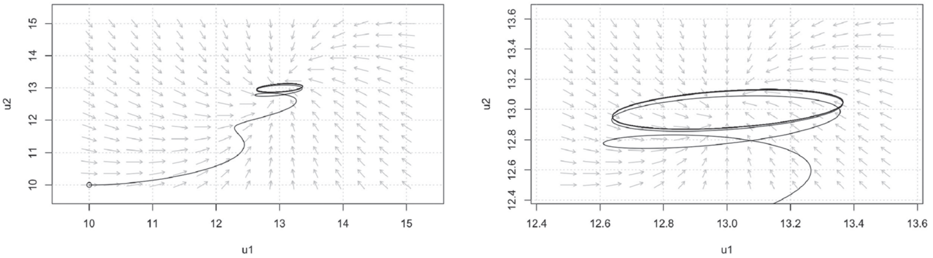


Figure 4 On the left, the trajectory for the initial condition (10, 10) after 100 s. On the right, an amplified view of the converging cycle.

Stability of the QTD model First, we concentrate on the study of the dynamics of these three cars when assuming that they follow the leading one according to QTD model.

$$\begin{cases} u_1'(t) = \gamma_1 u_1(t)(u_2(t) - u_1(t)) \\ u_2'(t) = \gamma_2 u_2(t)(u_3(t) - u_2(t)) \\ u_3'(t) = \gamma_3 u_3(t)(v - u_3(t)). \end{cases} \tag{11}$$

Solving the system

$$\begin{cases} 0 = \gamma_1 u_1(u_2 - u_1) \\ 0 = \gamma_2 u_2(u_3 - u_2) \\ 0 = \gamma_3 u_3(v - u_3), \end{cases} \tag{12}$$

we obtain the equilibrium points: $P_1 = (0, 0, 0)$, $P_2 = (0, 0, v)$, $P_3 = (0, v, v)$, and $P_4 = (v, v, v)$.

We now calculate the eigenvalues associated to the equilibrium points P_1, P_2, P_3, and P_4. The Jacobian of the system is given by

$$J(u_1, u_2, u_3) = \begin{pmatrix} \gamma_1(u_2 - 2u_1) & \gamma_1 u_1 & 0 \\ 0 & \gamma_2(u_3 - 2u_2) & \gamma_2 u_2 \\ 0 & 0 & \gamma_3(v - 2u_3) \end{pmatrix}. \tag{13}$$

We next obtain the eigenvalues associated to $J(P_i)$, $i = 1, \ldots, 4$. $J(P_1)$ has $\gamma_3 v$ as a positive eigenvalue, and therefore P_1 is not stable. Points P_2 and P_3 are unstable equilibrium points because they both have positive and negative eigenvalues, and P_4 is a stable equilibrium point for the system since all its eigenvalues have negative real part. It is important to point out that an equilibrium for system (11) corresponds to a stationary solution, specifically, a solution for which each of the three cars has constant velocity.

Since our model is nonlinear, it may present chaotic motions. To check whether or not it is chaotic, we analyze both the bifurcation diagrams and the Lyapunov exponent of the system in order to see the global behavior of our model for different values of the parameters.

On the one hand, *bifurcation diagrams* provide graphical information on the changes in the dynamics of the system in terms of one of its parameters [25]. They are very useful for checking, from a qualitative point of view, if the system is chaotic or not. To plot it, we take an arbitrary initial condition, and we compute the final state of the system versus a chosen parameter of it. If the system is periodic of period n, then n points appear in a vertical line in our bifurcation diagram. Otherwise, if the system is chaotic, a continuous vertical line is depicted in the diagram.

On the other hand, the *Lyapunov exponent* is the most common tool to have a quantitative indicator to observe chaotic motions [22]. These tools are the most useful indicators for characterizing possible chaotic regimes in a dynamical system, but they are mostly unfamiliar to undergraduate students in mathematics and science. The Lyapunov exponent, namely λ, indicates the divergence between two trajectories of the system which start their motion with very similar initial conditions. It measures this distance as a function of the time, and it can be calculated as follows:

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \ln \frac{||\delta x(t_i)||}{||\delta x_0||}, \tag{14}$$

where $||\delta x(t_i)||$ denotes the distance between the trajectories after the time $t = t_i$, and $||\delta x_0||$ denotes the distance between the trajectories at the initial time $t = 0$. If the system diverges, then the Lyapunov exponent is positive, and therefore our system exhibits chaotic motions. This is due to the nonlinear nature of our equations, and therefore it satisfies the necessary condition for that. We can observe that these two nearby trajectories are sensitive to the initial conditions, and their distance increases with respect to time in an exponential manner and therefore, if the equations are nonlinear, they become chaotic.

Otherwise, our system is stable and periodic motions take place. Figure 5 illustrates this. It shows both the bifurcation diagram and the Lyapunov exponent of the QTD model by taking as a parameter the sensitivity coefficient $\gamma = \gamma_i$, $i = 1, 2, 3$, representing the drivers' reaction times in a realistic situation. In Figure 5, we consider speeds around 13 m/s and values of γ around 0.3 s^{-1}. To compute both, we have taken as initial condition $(u_1(0), u_2(0), u_3(0)) = (20, 13, 10)$, a physical situation in which collisions can take place since the first and the second car are at rest and the last one has positive velocity.

Notice that to compute numerically the Lyapunov exponents, we take high values of the integration times according to the experiment we have carried out. In our case, we have taken 500 $t.u$ which is very large in comparison with the times we use in the computations of the trajectories. In that sense, the time, in these practical situations, can be treated as infinite, and therefore the estimation of the Lyapunov exponents is very accurate.

In both plots, we clearly see the final stabilization of the dynamics of our system, and therefore the nonexistence of chaotic motions. This result is interesting since the equations are nonlinear, but, in a practical case, there are no vehicle collisions and the cars finish in a stable situation.

Stability of the NN model We now suppose that the three cars following the leader at constant speed behave according to an NN model. Now, the acceleration of each car depends not only on the velocity of the car just in front, but also on the two cars ahead of it. For the car that only has one car ahead of them, we assume they follow the QTD model.

In this case, the system describing the model can be written as follows:

$$\begin{cases} u_1'(t) = \gamma_{1,1} u_1(t)(u_2(t) - u_1(t)) + \gamma_{1,2} u_1(t)(u_3(t) - u_1(t)) \\ u_2'(t) = \gamma_{2,1} u_2(t)(u_3(t) - u_2(t)) + \gamma_{2,2} u_2(t)(v - u_2(t)) \\ u_3'(t) = \gamma_3 u_3(t)(v - u_3(t)). \end{cases} \tag{15}$$

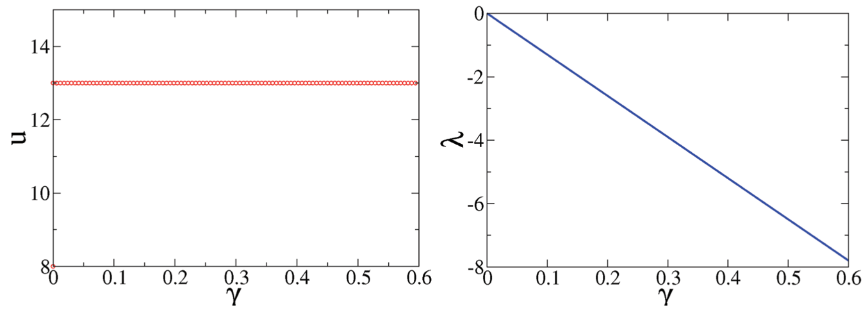


Figure 5 Plot of both the bifurcation diagram, by plotting u versus γ, and the Lyapunov exponent versus γ. We can observe that the motions are not chaotic for any parameter value, and the system is always stable. The initial condition is given by $(u_1(0), u_2(0), u_3(0)) = (20, 13, 10)$. Observe that in this case the value of velocity $v = 13$ m/s as an asymptotic fixed point and therefore there is not presence of chaos. The corresponding Lyapunov exponent distribution corroborates it properly.

We now obtain the equilibrium points of the system

$$\begin{cases} 0 = \gamma_{1,1}u_1(u_2 - u_1) + \gamma_{1,2}u_1(u_3 - u_1) \\ 0 = \gamma_{2,1}u_2(u_3 - u_2) + \gamma_{2,2}u_2(v - u_2) = 0 \\ 0 = \gamma_3 u_3(v - u_3) = 0. \end{cases} \tag{16}$$

Again, we also have

$$P_1 = (0, 0, 0), \quad P_2 = (0, 0, v), \quad P_3 = (0, v, v), \quad P_4 = (v, v, v),$$

as in the previous case, and

$$P_5 = \left(0, \frac{v\gamma_{2,1}}{\gamma_{2,1} + \gamma_{2,2}}, 0\right), \qquad P_6 = \left(\frac{v\gamma_{1,2}}{\gamma_{1,1} + \gamma_{1,2}}, 0, v\right),$$
$$P_7 = \left(\frac{v\gamma_{1,1}}{(\gamma_{1,1} + \gamma_{1,2})(\gamma_{2,1} + \gamma_{2,2})}, \frac{v\gamma_{2,1}}{\gamma_{2,1} + \gamma_{2,2}}, 0\right).$$

We note that points P_5, P_6, and P_7 will not be considered when analyzing the system in the car-following context since they do not represent realistic situations.

In this case, the Jacobian matrix $J(u_1, u_2, u_3)$ is given by:

$$\begin{pmatrix} \gamma_{1,1}u_2 - (2\gamma_{1,1} + 2\gamma_{1,2})u_1 + \gamma_{1,2}u_3 & \gamma_{1,1}u_1 & \gamma_{1,2}u_1 \\ 0 & X & \gamma_{2,1}u_2 \\ 0 & 0 & \gamma_3 v - 2\gamma_3 u_3 \end{pmatrix}, \tag{17}$$

where

$$X = \gamma_{2,2}u_3 - (2\gamma_{2,1} + 2\gamma_{2,2})u_2 + \gamma_{2,2}v.$$

Substituting the equilibrium points, we get that the P_i, for all $1 \le i \le 7$ with $i \ne 4$, are unstable equilibrium points. For the case of P_4, all the eigenvalues of $J(P_4)$ are negative, and thus it is a stable point. So all models agree in the fact that the unique stable solution is obtained when all the cars approach the speed of the leading car.

Stability of perturbed traffic models

In realistic situations, when a car approaches, there is a variation in the motion like the one given by a periodic force acting on the cars. When trying to maintain a certain distance, cars sometimes get a little closer, and they sometimes get a little more separated. This is to prevent collisions of one car with the two others when they are too close and they circulate in the same direction. This effect can be modeled by adding a new term to the equations, namely $\alpha \sin(t)$, with $\alpha > 0$.

We discuss the stability of the previous models for the aforementioned case in which we introduce a small perturbation given by $\alpha \sin(t)$, $\alpha > 0$ in the speed of the leading car. Then, the QTD model with three cars following a leader at constant speed v is given by

$$\begin{cases} u_1'(t) = \gamma_1 u_1(t)(u_2(t) - u_1(t)) \\ u_2'(t) = \gamma_2 u_2(t)(u_3(t) - u_2(t)) \\ u_3'(t) = \gamma_3 u_3(t)(v + \alpha \sin(t) - u_3(t)). \end{cases} \tag{18}$$

In the case of the NN model, the analogous perturbed model is described as follows:

$$\begin{cases} u_1'(t) = \gamma_{1,1} u_1(t)(u_2(t) - u_1(t)) + \gamma_{1,2} u_1(t)(u_3(t) - u_1(t)) \\ u_2'(t) = \gamma_{2,1} u_2(t)(u_3(t) - u_2(t)) + \gamma_{2,2} u_2(t)(v - u_2(t)) \\ u_3'(t) = \gamma_3 u_3(t)(v + \alpha \sin(t) - u_3(t)). \end{cases} \tag{19}$$

Even in this case, chaotic motions cannot be found for any value of γ, as we show in Figure 6, in which both bifurcation diagrams and the Lyapunov exponent are calculated for the QTD model. As in Figure 5, we have taken as initial condition $(u_1(0), u_2(0), u_3(0)) = (20, 13, 10)$, and $v = 13$, a physical situation in which collisions can take place. If $v = 13$, the speeds of each car will converge to a cycle around (v, v, v), and the speeds can be assumed to be greater than one, in order to ensure that all accelerations will be positive. These tools can be used easily for other systems to see the existence of irregular motions, and they are very intuitive concepts for undergraduate students in mathematics or science.

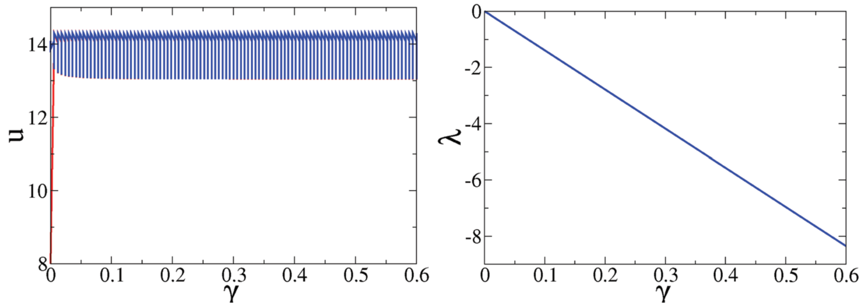


Figure 6 Plot of the bifurcation diagram, by plotting u versus γ, and the Lyapunov exponent versus γ. We can observe that the motions are not chaotic for any parameter value, and the system is always stable even under the existence of an external force. The initial condition is given by $(u_1(0), u_2(0), u_3(0)) = (20, 13, 10)$. We can see, on the left side, that the velocity v changes periodically with period equal one. On the other hand, and in the right panel, the Lyapunov exponent is zero or negative. Therefore, the non-chaotic behavior is corroborated.

In Figure 7, we plot the trajectories of non-perturbed and perturbed models and we see that the perturbation is transmitted to the cars following the leading one both for the QTD and the NN models. Moreover, we observe that the amplitude of these perturbations tends to zero for the NN model, in contrast to what happens for the QTD.

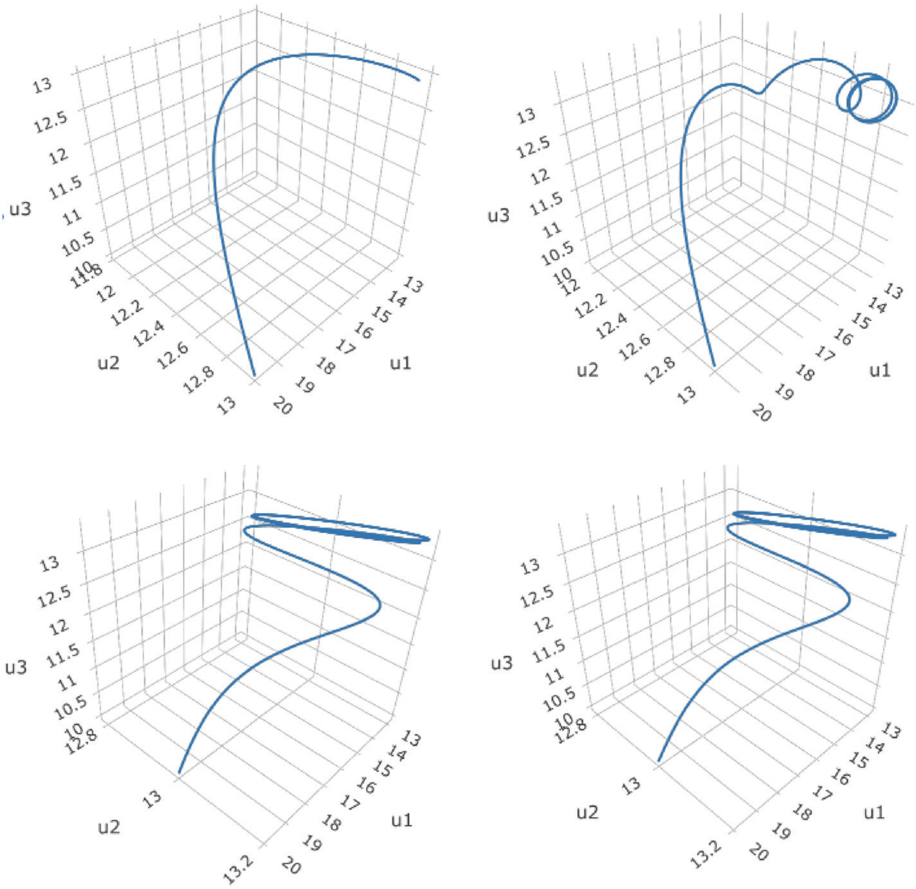


Figure 7 Top left: The trajectory following the QTD model, see (11), where the velocity of car 1 is represented on the x-axis, the velocity of car 2 on the y-axis, and the velocity of car 3 on the z-axis. Top right: The trajectory following the perturbed QTD model, see (18). Bottom left: The trajectory following the NN model (15). Bottom right: The trajectory following the perturbed NN model, see (19). The initial condition in all cases is $(u_1(0), u_2(0), u_3(0)) = (20, 13, 10)$. The parameters γ_i, $i = 1, 2, 3$ are equal to 0.03. The parameters $\gamma_{i,j}$, $i, j = 1, 2$ are equal to 0.015. The parameter $\alpha = 1$.

Possible extensions and class activities

Once the QTD and NN models have been studied, and the students are familiarized with them, the most natural activities to propose are possible extensions and modifications of the models. One possible activity is to generalize and study the previous models for a finite given number of vehicles. That is, to study the systems

$$\begin{cases} u_i'(t) = \gamma_i u_i(t)(u_{i+1}(t) - u_i(t)) & \text{for } 1 \leq i \leq k-1 \\ u_k'(t) = \gamma_k u_k(t)(v - u_k(t)), \end{cases}$$

and

$$\begin{cases} u_i'(t) = \gamma_{i,1}u_i(t)(u_{i+1}(t) - u_i(t)) + \gamma_{i,2}u_i(t)(u_{i+2}(t) - u_i(t)), 1 \le i \le k-2 \\ u_{k-1}'(t) = \gamma_{k-1,1}u_{k-1}(t)(u_k(t) - u_{k-1}(t)) + \gamma_{k-1,2}u_{k-1}(t)(v - u_{k-1}(t)) \\ u_k'(t) = \gamma_k u_k(t)(v - u_k(t)). \end{cases}$$

Another possible activity is to propose the same analysis, but now for a new model that takes into account the speeds of the cars that drive in front and behind the main driver, known in its linear version as the Forward and Backward Control model. This model was developed by Herman et al. [18] for General Motors and is studied by Barrachina et al. [4], Students can investigate if this new consideration can lead to chaotic situations in contrast to the other models:

$$u_1'(t) = -\gamma_1 u_1(t) + \gamma_2 u_2(t)(u_2(t) - u_1(t)),$$
$$u_2'(t) = \gamma_1(u_1(t) - u_2(t)) + \gamma_2 u_2(u_3(t) - u_2(t)),$$

with control constants $\gamma_1, \gamma_2 > 0, \gamma_1 < \gamma_2$. We have considered the original model, but adding a nonlinearity assumption that the control parameter γ_2 is proportional to $u_2(t)$.

Closing remarks Although classical mathematical models, such as population models, are really useful for teaching dynamical systems, our objective in this paper is to provide a not-so-common model to emphasize the numerous applications that dynamical systems present. Moreover, the apogee of autonomous cars encourages the study of car-following models in this area. The car-following models considered in this paper can also serve to improve students' skills for modeling dynamical systems using software. Moreover, it presents an opportunity to introduce the students to a simple, but important, class of traffic models which are widely used in engineering, and they can help them to give a physical interpretation of mathematical models.

Acknowledgments J. Alberto Conejero and Marina Murillo-Arcila were supported by MEC, grant MTM2016-75963-P. Jesús M. Seoane was supported by the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) under Project No. PID2019-105554GB-I00. Juan B. Seoane-Sepúlveda was supported by grant PGC2018-097286-B-I00.

REFERENCES

[1] Addison, P. S., Low, D. J. (1998). A novel nonlinear car-following model. *Chaos: Interdiscip. J. Nonlinear Sci.* 8(4): 791–799. doi.org/10.1063/1.166364

[2] Addison, P. S., McCann, J. M., Low, D. J., Currie, J. I. (1996). Order and chaos in the dynamics of vehicle platoons. *Traffic Eng. Control*. 37(7–8): 456–459.

[3] Balibrea. F. (2016). On problems of topological dynamics in non-autonomous discrete systems. *Appl. Math. Nonlinear Sci.* 1(2) : 391–404. doi.org/10.21042/AMNS.2016.2.00034

[4] Barrachina, X., Conejero, J. A., Murillo-Arcila, M., Seoane-Sepúlveda, J. B. (2015). Distributional chaos for the forward and backward control traffic model. *Linear Algebra Appl.* 479 (15 Aug 2015) : 202–215. doi.org/10.1016/j.laa.2015.04.010

[5] Brackstone, M., McDonald, M. (1999). Car-following: a historical review. *Transp. Res. Part F Traffic Psychol. Behav*. 2(4): 181–196. doi.org/10.1016/S1369-8478(00)00005-X

[6] Chandler, R. E., Herman, R., Montroll, E. W. (1958). Traffic dynamics: studies in car following. *Operations Res.* 6(2): 165–184. doi.org/10.1287/opre.6.2.165

[7] Conejero, J. A., Murillo-Arcila, M., Seoane-Sepúlveda, J. B. (2016). Linear chaos for the Quick-Thinking-Driver model. *Semigroup Forum*. 92(2): 486–493. doi.org/10.1007/s00233-015-9704-6

[8] Conejero, J. A., Martínez-Giménez, F., Peris, A., Ródenas, F. (2016). Chaotic asymptotic behaviour of the solutions of the Lighthill-Whitham-Richards equation. *Nonlinear Dyn.* 84(1): 127–133. doi.org/10.1007/s11071-015-2245-4

[9] Cumming, R. W. (1964) The analysis of skills in driving. *J. Aust. Road Res. Board*. 1(9): 4–14.

[10] Disbro, J. E., Frame, M. (1989). Traffic flow theory and chaotic behavior. *Transp. Res. Rec*. 1225: 109–115.

[11] Forbes, T. W. (1963). Human factor considerations in traffic flow theory. *Highw. Res. Board Rec*. 15: 60–66.

[12] Fisher, D. L., Rizzo, M., Caird, J, Lee, J. D. (2011). *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. Boca Raton, FL: CRC Press.

[13] Gazis, D. C., Herman, R., Rothery, R. W. (1961). Nonlinear follow-the-leader models of traffic flow. *Oper. Res.* 9(4): 545–567. doi.org/10.1287/opre.9.4.545

[14] Gitterman, M. (2010). *The Chaotic Pendulum*. Singapore: World Scientific.

[15] Greenshields, B. D. (1934). The photographic method of studying traffic behavior. *Proceedings of the 13th Annual Meeting of the Highway Research Board*, pp. 382–399.

[16] Greenshields, B. D. (1935). A study of traffic capacity. *Proceedings of the 14th Annual Meeting of the Highway Research Board*, pp. 448–477.

[17] Hoogendoorn, S. P., Bovy, P. H. L. (2001). State-of-the-art of vehicular traffic flow modeling. *Proc. Inst. Mech. Eng. Part I: J. Syst. Control Eng*. 215(4): 283–303. doi.org/10.1243.0959651011541120

[18] Herman, R., Montroll, E. W., Potts, R. B., Rothery, R. W. (1959). Traffic dynamics: analysis of stability in car following. *Oper. Res.* 7(1): 86–106. doi.org/10.1287/opre.7.1.86

[19] Harraga, H., Yebdri, M. (2018). Attractors for a nonautonomous reaction-diffusion equation with delay. *Appl. Math. Nonlinear Sci*. 3(1): 127–150. doi.org/10.21042/AMNS.2018.1.00010

[20] McCartney, M. (2009). A discrete time car following model and the bi-parameter logistic map. *Commun. Nonlinear Sci. Numer. Simul.* 14(1): 233–243. doi.org/10.1016/j.cnsns.2007.06.012

[21] McCartney, M., Gibson, S. (2004). On the road to chaos. *Teaching Mathematics and its Applications*. 23(2): 89–96. doi.org/10.1093/teamat/23.2.89

[22] Nusse, H. E., Yorke, J. A. (2012). *Dynamics: Numerical Explorations*. New York: Springer.

[23] Pipes, L. A. (1953). An operational analysis of traffic dynamics. *J. Appl. Phys*. 24(3): 274–281. doi.org/10.1063/1.1721265

[24] Rothery, R. W. (1992). Car following models. In: *Revised Monograph on Traffic Flow Theory*, Federal Highway Administration, Oak Ridge National Laboratory. Update and expansion of the transportation research board (trb) special report 165, “Traffic flow theory,” 1975 ed.

[25] Strogatz, S. H. (2019). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Boca Raton, FL: CRC Press.

Summary. We analyze the dynamics of a system of several cars in a traffic lane, each car following the ones in front of it. The effect of small perturbations in the speed of a certain car is propagated to the cars behind it in the lane. Nevertheless, these perturbations tend to dissipate along the lane. The results can be used as an activity for undergraduate students to improve their knowledge of dynamical systems, modeling, and the physical interpretation of mathematical models.

J. ALBERTO CONEJERO (MR Author ID: 684956) is a full professor in the Applied Mathematics Department of Universitat Politècnica de València. He is also responsible for the M.Sc. program in Mathematics Research at UPV. His research interests include dynamical systems, partial differential equations, network science, data analysis, and applications of mathematics to computer science, engineering, and biotechnology.

MARINA MURILLO-ARCILA (MR Author ID: 998995) is an associate professor in the Department of Mathematics at Universitat Jaume I. Her research interests include dynamical systems, fractional calculus, and partial differential equations and their applications.

JESÚS M. SEOANE (MR Author ID: 790243) is a full professor of physics at Universidad Rey Juan Carlos (URJC) in Madrid, Spain. His research is in the fields of nonlinear dynamics, chaos theory, and complex systems. More particularly, his research interests are in cancer physics, fractal structures, Hamiltonian systems, chaotic scattering, synchronization, and chaos control, among others.

JUAN B. SEOANE-SEPÚLVEDA (MR Author ID: 680972) received his first Ph.D. at the Universidad de Cádiz (Spain), jointly with Universität Karlsruhe (Germany) in 2005. His second Ph.D. was earned at Kent State University (Kent, Ohio, USA) in 2006. His main interests include real analysis, set theory, Banach space geometry, and lineability. He has authored two books and over 150 research papers. He is currently a full professor at Universidad Complutense de Madrid (Spain), and is the Editor of both the *Real Analysis Exchange* and the *Banach Journal of Mathematical Analysis*.

In Search of Lost Time (Coordinate)

HASSAN BOUALEM
Université de Montpellier
Montpellier, France
hassan.boualem@umontpellier.fr

ROBERT BROUZET
Université Perpignan Via Domitia
Perpignan, France
robert.brouzet@univ-perp.fr

This paper starts with the amazing relationship between the perimeter of a circle and the area of the disk it bounds: the first is the derivative of the second. This subject has been studied in numerous, very interesting papers [2–6,8,9]. Our approach is quite different because we adopt a differential geometry point of view that is motivated by the following considerations.

To better understand this phenomenon, it seems useful to do a preliminary review of the meaning of the derivative with respect to a variable. In fact, the sentence " the derivative of the area of a disk is equal to its perimeter" makes no sense, or, at least, it contains some implicit assumptions. Why? Because this assertion is true: $\frac{d}{dr}(\pi r^2) = 2\pi r$, and also because it is wrong: $\frac{d}{dD}(\pi D^2/4) \neq \pi D$! So, depending on whether one thinks in terms of the radius or the diameter, we are either right or wrong. Is the diameter therefore less natural than the radius? We would like to say no, and yet . . . The *functions* perimeter and area are well-defined and have an intrinsic meaning, but there are many ways to write them according to the choice of parameter or variable.

By way of comparison, if you give a price in dollars or euros, a distance in kilometers or miles, you only change units, and this also changes the numbers representing what you want to quantify. However, the price or the distance remains intrinsically the same. This is a similar phenomenon to the problem that concerns us here. Let us go on in this direction: Take two arbitrary functions; you will always manage to say that one of them is the derivative of the other! For instance, it is possible to get the monomial x^2 as the derivative of the polynomial $x^7 + 3$ or, more striking, the exponential function as the derivative of the tangent function!

How do we achieve this crazy trick? It is very easy. Let us consider this problem with two general functions before coming back to the previously mentioned examples. Let f and g be two real-valued functions defined on the same interval I. We will assume that these two functions are differentiable and, even, of class $\mathcal{C}^1$. We denote by x the variable of these functions. What we have just said is that there is a new variable $t = t(x)$, $x = x(t)$ such that $\frac{df}{dt} = g$, in the precise sense that $\frac{d}{dt} f(x(t)) = g(x(t))$. Indeed, let us write that $\frac{df}{dt} = \frac{df}{dx}\frac{dx}{dt} = g$; if $\frac{df}{dt} = g$, then $\frac{df}{dx}\frac{dx}{dt} = g$ and so

$$\frac{dt}{dx} = \frac{f'(x)}{g(x)} \qquad \text{or} \qquad t(x) = \int \frac{f'(x)}{g(x)}\, \mathrm{d}x.$$

Of course, we must assume that g is never zero on I; we even need more: we will assume that we have $g > 0$ and $f' > 0$ on I. With these assumptions, the well-defined function $x \mapsto t(x)$ is strictly increasing and so a $\mathcal{C}^1$-diffeomorphism from I to its range, and therefore it is an admissible change of coordinates. We have, of course, $\frac{df}{dt} = g$ as required.

Math. Mag. **95** (2022) 314–328. doi:10.1080/0025570X.2022.2093584 © Mathematical Association of America
MSC: Primary M3A04, Secondary 97G30

Now let us look at our two examples. starting with the polynomials. On the interval $I = \mathbb{R}_+^*$, we immediately get by calculating $\int \frac{f'(x)}{g(x)}\,\mathrm{d}x$ that $t(x) = \frac{7}{5}x^5$, implying that we should set $x(t) = \sqrt[5]{\frac{5}{7}t}$. For the example of the exponential and tangent functions, namely for the case where $f(x) = \tan x$ and $g(x) = e^x$, both considered as defined on $I = (-\pi/2, \pi/2)$, all the assumptions of the previous general calculation are satisfied, and the variable t such that $\frac{d}{dt}(e^t) = \tan t$ is given by

$$t : (-\pi/2, \pi/2) \to \mathbb{R},\ x \mapsto t(x) = \int_0^x \frac{1 + \tan^2 s}{e^s}\,\mathrm{d}s.$$

Unfortunately, in this case we cannot go further because of the impossibility of writing this integral in terms of elementary functions. It would be a more delicate matter if we were to consider the same problem for figures with several degrees of freedom, as we will see later.

Now let us return to our initial problem of the relationship between area and perimeter for the circle. Let us study why the radius is the parameter that makes the miracle possible. In other words, what makes the radius natural? For a family of squares parameterized by the side c, the area function is $A(c) = c^2$, the perimeter function is $P(c) = 4c$, and the equality $A'(c) = P(c)$ is not satisfied. But another choice—the radius r of the inscribed circle—gives $A(r) = 4r^2$, $P(r) = 8r$, and so $A'(r) = P(r)$. We remark that if s is the chosen parameter, then the calculation of the derivative of the area function

$$A'(s) = \lim_{\epsilon \to 0} \frac{A(s+\epsilon) - A(s)}{\epsilon},$$

leads us to consider *deformations* of squares. Figure 1 illustrates such deformations in the case $s = c$, then $s = r$.

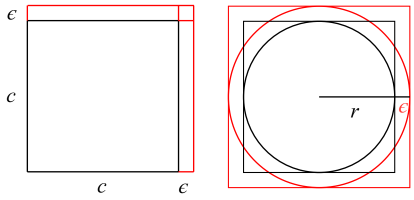


Figure 1 Two possible deformations of a square corresponding respectively to $s = c$ and $s = r$.

We have the same phenomenon that we have already noted in the case of the circle with the radius versus the diameter. We can see that a deformation with respect to the diameter leads to an increase in the area of the disk that is half as much as a deformation with respect to its radius. This is because we must divide the ϵ into two parts, so the variation of the area is not large enough in this case (see Figure 2).

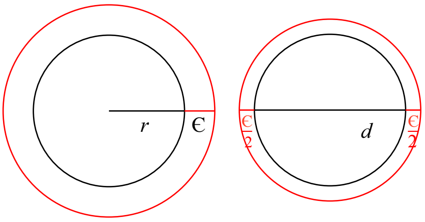


Figure 2 Two possible deformations of a circle: radius *vs* diameter.

This geometrical idea of deformation will be very important from the point of view we develop; this notion appears in Krenicky and Rychtář [6] under the guise of *similarity*. The reason is the following: We can see, through the basic examples of circles and squares, that we are dealing with a one-parameter family of curves, and that the choice of the "right" parameter is critical. But, if we want to consider the initial question in a more general way, we will need to work with figures that are described using several parameters and not just one, like, for example, rectangles or triangles. Of course, inside the family of all rectangles, or the family of all triangles, we can only consider the particular family of regular polygons. They can be parameterized by the radius of the incircle, which is also known as the apothem, and, in this case, the studied relation makes sense and is true! But for the complete family of all the figures of a given type (rectangles, triangles etc.) we will need to work in the framework of *differential geometry* and to consider *manifolds of figures* on which will be defined differentiable functions *the perimeter, area, and volume*. The initial problem will then be recast in terms of *directional derivatives* and therefore of deformations of figures.

In light of the points previously outlined, our text will be divided into two parts. The first will deal with the case of one-parameter families of geometrical figures in a Euclidean space. After a brief survey of classical results, we will state a new one, presented in the general framework of codimension 1 submanifolds of a Riemannian manifold, which explains why the derivative of the n-dimensional volume of a round ball with respect to its radius is equal to the $(n-1)$-volume of its boundary (a sphere). The second part is dedicated to the general case of d-dimensional manifolds of n-dimensional geometrical figures. We will show that one can always find a vector field X such that the Lie derivative of the n-dimensional volume with respect to this field is equal to the $(n-1)$-dimensional volume of its boundary. We will illustrate both the notion of manifolds of geometrical figures and this result by giving several examples. In each case, we will study the deformations of figures by the flow of the vector field X.

One-parameter family of geometrical figures

Ball and sphere in three-dimensional Euclidean space Let us begin with some recaps consisting of simple calculations about them in the usual space $\mathbb{R}^3$. Because length, area, and volume are invariant by translation, we only consider a ball with center at the origin, and because they are changed by multiplying by r, r^2, and r^3, respectively, when we pass from radius 1 to radius r, we only look at the case where the radius is 1. Let us denote by V the volume of such a ball and by A the area of the corresponding sphere. The ideas described below to calculate V and A will be the same for the general calculation in the next section. The main idea is to consider a ball as an onion! You can cut it by slices with a knife or peel it layer by layer. In the first case you see the ball as a union of parallel disks with radii between 0 and 1, and in the second case as a union of concentric spheres with radii between 0 and 1. The first way leads to:

$$\begin{aligned} V &= \int\int\int_{\{(x,y,z)\in\mathbb{R}^3,\ x^2+y^2+z^2\le 1\}} dx\,dy\,dz \\ &= \int_{-1}^{1}\left[\int\int_{\{(x,y)\in\mathbb{R}^2,\ x^2+y^2\le 1-z^2\}} dx\,dy\right] dz. \end{aligned}$$

The double integral in the brackets is just the area of a disk with radius equal to $\sqrt{1-z^2}$, and so it is equal to $\pi(1-z^2)$. One can conclude from this remark that

$$V = \pi\int_{-1}^{1}(1-z^2)\,dz = 2\pi\int_{0}^{1}(1-z^2)\,dz = 2\pi\left(1-\frac{1}{3}\right) = \frac{4\pi}{3}.$$

The second way leads to $V = \int_0^1 A(r)\mathrm{d}r$ where $A(r)$ is the area of the sphere with radius r. But by the previous argument about homogeneity, we have $A(r) = r^2 A$ and so

$$V = A\int_0^2 r^2\, dr = \frac{1}{3}A.$$

Finally, we can conclude that $V = \frac{4}{3}\pi$ and $A = 4\pi$. Therefore, a ball with radius r has a volume equal to $\frac{4}{3}\pi r^3$, and the area of the corresponding sphere is equal to $4\pi r^2$, as is well-known by all students.

Balls and spheres in a general Euclidean space Let us recall some well-known facts about the notion of volume and, in particular, let us deal with the special case of balls and spheres in n-dimensional Euclidean space. To deal with volume, you first need to choose some reference unit. If you work in some n-dimensional vector space (not necessarily endowed with a Euclidean structure), then you can choose some basis $\mathcal{B} = (e_1, \dots, e_n)$ and decide that the polytope based on it is your unit. Now, the volume of a polytope constructed on vectors $v_1, \dots, v_n$ is just the determinant of this family of vectors in the basis $\mathcal{B}$. In the case of an oriented Euclidean space, you have a natural choice for the initial basis: any directed orthonormal basis. If you have a domain of E more complicated than a simple polytope, then you must use integral calculus. The classical language for this is that of differential forms. The space of differential forms with degree n is a line. and any generator of this line is called a volume-form. Let us choose such a form ω_n that satisfies the condition $\omega_n(e_1, \dots, e_n) = 1$, where $(e_1, \dots, e_n)$ is the canonical basis of $\mathbb{R}^n$. If K is a compact set of $\mathbb{R}^n$, its n-volume is then defined by

$$\mathrm{Vol}_n(K) := \int_K \omega_n.$$

If we assume that the boundary ∂K of K is a smooth hypersurface, and N is a field of unit vectors normal to ∂K, then we can define the $(n-1)$-volume of ∂K as

$$\mathrm{Vol}_{n-1}(\partial K) := \int_{\partial K} i_N\omega_n,$$

where $i_N\omega_n$ is the interior product of ω_n by N, i.e., the $(n-1)$-differential form that you get by putting N into the first argument.

Calculating the volume of a ball or a sphere in the usual Euclidean space $\mathbb{R}^n$ is a classical problem. There are several methods available. We briefly describe one of them. Let n be a positive integer and r a positive real number. Let us define

$$B_n(r) := \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n,\ \sqrt{\sum_{i=1}^n x_i^2} \le r \right\},$$

$$\mathbb{S}^{n-1}(r) := \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n,\ \sqrt{\sum_{i=1}^n x_i^2} = r \right\},$$

to be, respectively, the Euclidean ball of $\mathbb{R}^n$ with radius r and center at the origin, and the sphere which is its boundary. For $r = 1$, we simply denote them by B_n and $\mathbb{S}^{n-1}$.

Then we have

$$V_n(r) := \mathrm{Vol}_n(B_n(r)) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} r^n$$

and

$$A_{n-1}(r) := \mathrm{Vol}_{n-1}(\mathbb{S}^{n-1}(r)) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} n r^{n-1},$$

where Γ is the Euler Gamma function defined for $x \in (0, +\infty)$ by the formula

$$\Gamma(x) := \int_0^{+\infty} e^{-t} t^{x-1}\, dt.$$

One can find this result in Mneimné and Testard [7] or, for a generalization to polytopes, in Emert and Nelson [4]. The idea is the following: we start with the classical value of the Gaussian integral

$$\int_{-\infty}^{+\infty} e^{-x^2}\, dx = \sqrt{\pi}.$$

Using the Fubini-Tonelli theorem we get that

$$J_n := \int_{\mathbb{R}^n} e^{-x_1^2-\cdots-x_n^2}\, dx_1 \cdots dx_n = \pi^{\frac{n}{2}}.$$

Now let us consider the change of variables to spherical coordinates, that is, the $\mathcal{C}^1$-diffeomorphism

$$\mathbb{R}^n \setminus \{0\} \to \mathbb{S}^{n-1} \times (0, +\infty),\ x \mapsto \left(\frac{x}{\|x\|}, \|x\|\right).$$

Then we have

$$J_n = \int_{\mathbb{S}^{n-1}\times(0,+\infty)} e^{-r^2} r^{n-1}\, dr\, d\sigma = \int_{\mathbb{S}^{n-1}} d\sigma \int_0^{+\infty} e^{-r^2} r^{n-1}\, dr,$$

where the last equality uses the Fubini-Tonelli theorem. Now the change of variable $t = r^2$ transforms the simple integral which appears in the last formula to

$$\int_0^{+\infty} e^{-r^2} r^{n-1}\, dr = \int_0^{\infty} e^{-t} t^{\frac{n-1}{2}} \frac{dt}{2\sqrt{t}} = \frac{1}{2}\Gamma\left(\frac{n}{2}\right).$$

So,

$$J_n = \frac{1}{2}\Gamma\left(\frac{n}{2}\right) A_{n-1}(1).$$

Using spherical coordinates, we also get

$$V_n(1) = A_{n-1}(1) \int_0^1 r^{n-1}\, dr = \frac{A_{n-1}(1)}{n}.$$

Finally,

$$V_n(r) = r^n V_n(1) = r^n \frac{A_{n-1}(1)}{n} = r^n \frac{\pi^{\frac{n}{2}}}{\frac{n}{2}\Gamma\left(\frac{n}{2}\right)} = r^n \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n+2}{2}\right)},$$

using the classic functional equation satisfied by the Gamma function—for all positive real numbers x, $\Gamma(x+1) = x\Gamma(x)$.

Now,

$$A_{n-1}(r) = \int_{\mathbb{S}^{n-1}(r)} d\sigma = r^{n-1} \int_{\mathbb{S}^{n-1}} d\sigma = r^{n-1} A_{n-1}(1).$$

But we have seen the relation between $A_{n-1}(1)$ and the volume of the unit ball: $V_n(1) = \frac{A_{n-1}(1)}{n}$, or $A_{n-1}(1) = nV_n(1)$. So,

$$A_{n-1}(r) = nr^{n-1} \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n+2}{2}\right)} = \frac{dV_n(r)}{dr}.$$

Why is the radius the natural coordinate? This relation between the n-volume of a ball of radius r and the $(n-1)$-volume of the sphere of radius r is a consequence of two facts: the $(n-1)$-sphere is the *boundary* of the n-ball and the gradient of the "distance to the center" function which defines the sphere and the ball is a *unitary* vector field. We can state a result pointing out these two facts in a very general context involving submanifolds of codimension 1 in a Riemannian manifold. But before that, let us be a bit more explicit in the case of the Euclidean balls, and let us try to show the main ideas. A derivative is a limit of a ratio involving the difference between two values of the function. Here, this difference makes integrals appear on *different domains*, namely

$$\frac{dV_n}{dr}_{|r=r_0} = \lim_{r\to r_0} \frac{\int_{B_n(r)} \omega_n - \int_{B_n(r_0)} \omega_n}{r - r_0}.$$

The first thing to do is to work in the *same domain*. It is possible using two tools: we can transform one ball into the other by a convenient homothety ($B_n(r)$ is the image of $B_n(r_0)$ by the transformation $v \mapsto \frac{r}{r_0}v$, and this is called a homothety) , and use the change of variables theorem. When we get the same domain, we can swap the limit and the integral, and thus actually the derivative with the integral, namely

$$\frac{dV_n}{dr}_{|r=r_0} = \int_{B_n(r_0)} \lim_{r\to r_0} \frac{1}{r_0^n} \frac{r^n - r_0^n}{r - r_0} \omega_n.$$

In the case of the balls, we have finished because we recognize the derivative of the function $r \mapsto r^n$ at r_0 and so

$$\frac{dV_n}{dr}_{|r=r_0} = \int_{B_n(r_0)} \frac{n}{r_0}\omega_n = \frac{n}{r_0} V_n(r_0),$$

and we know that the right hand side is also equal to $A_{n-1}(r_0)$, as already seen. Actually, this case is too simple, and so we can miss the deep purpose of the calculation. Indeed, let us write $\omega_n = dx_1 \wedge \cdots \wedge dx_n$ and introduce the $(n-1)$-form given by

$$\alpha := \sum_{i=1}^{n} (-1)^{i+1} \frac{x_i}{r_0}\, dx_1 \wedge \cdots \widehat{dx_i} \wedge \cdots \wedge dx_n,$$

where the symbol $\widehat{dx_i}$ denotes the fact that the term with index i does not appear in the sum. Then $d\alpha = \frac{n}{r_0}\omega_n$. So we can rewrite

$$\frac{dV_n}{dr}_{|r=r_0} = \int_{B_n(r_0)} \frac{n}{r_0}\omega_n = \int_{B_n(r_0)} d\alpha,$$

and using Stokes's theorem,

$$\frac{dV_n}{dr}_{|r=r_0} = \int_{\partial B_n(r_0)} \alpha.$$

But α is exactly the form $i_{\nabla f}\omega_n$ so

$$\frac{dV_n}{dr}_{|r=r_0} = \int_{\partial B_n(r_0)} i_{\nabla f}\omega_n,$$

which is precisely equal to $A_{n-1}(r_0)$ because the gradient is a unitary vector field!

Let us state a general result for which the previous one is a particular case.

Theorem 1. *Let M be a Riemannian manifold of dimension n, and let f be a proper onto map from M to $\mathbb{R}_+$ which is a submersion of class $\mathcal{C}^\infty$ on the set $M \setminus f^{-1}(0)$.*

For all $r > 0$, let us write

$$K_r := \{x \in M,\ f(x) \le r\} = f^{-1}([0, r])$$

and $H_r = \partial K_r$. Moreover, let us write $v_n(r)$ and $a_{n-1}(r)$ for their respective volumes.

If the gradient of f is unitary on $M \setminus f^{-1}(0)$, then we get the relation:

$$\forall r > 0,\ \frac{dv_n}{dr} = a_{n-1}(r).$$

What is playing the role of the previous homothety? It is the flow φ_t of the vector field ∇f. It is easy to verify that for all $r > 0$ and t sufficiently near to 0, we have $\varphi_t(K_r) = K_{r+t}$ and $\varphi_t(H_r) = H_{r+t}$. Thus, if we fix $r_0 > 0$ and choose a volume-form Ω on M, using the change of variables theorem, we get, for r sufficiently near to r_0, that

$$v_n(r) = \int_{K_{r_0}} \varphi^*_{r-r_0}\Omega.$$

So we can work with two integrals on the same domain. The derivative of v_n at r_0 is given by

$$\frac{dv_n}{dr}_{|r=r_0} = \int_{K_{r_0}} \frac{d}{dt}\varphi_t^*\Omega_{|t=0},$$

where $\varphi_t^*\Omega$ is the pullback of the volume form Ω by the application φ_t. So this formula gives us the commutativity of integral and derivative as in the case of the balls. This derivative is just the directional derivative (or Lie derivative) of Ω in the direction of the gradient:

$$\frac{d}{dt}\varphi_t^*\Omega_{|t=0} = \mathcal{L}_{\nabla f}\Omega.$$

A general formula of differential geometry, called Cartan's formula, gives a relation between the Lie derivative, the inner product, and the exterior derivative of forms: $L_X = di_X + i_X d$. So we have,

$$\frac{dv_n}{dr}_{|r=r_0} = \int_{K_{r_0}} i_{\nabla f}d\Omega + di_{\nabla f}\Omega.$$

But $d\Omega = 0$, so

$$\frac{dv_n}{dr}_{|r=r_0} = \int_{K_{r_0}} di_{\nabla f}\Omega.$$

Using Stokes's theorem, we get

$$\frac{dv_n}{dr}_{|r=r_0} = \int_{H_{r_0}} i_{\nabla f}\Omega,$$

and because ∇f is a unitary vector field, we can conclude that

$$\frac{dv_n}{dr}_{|r=r_0} = a_{n-1}(r_0).$$

We can see that this general proof perfectly follows the previous one given for balls. It makes apparent the deep reasons for which we have equality between the derivative of the n-volume and the $(n-1)$-volume of the boundary in the special case of Euclidean balls.

Convex compact sets contained in balls The example of the family of squares mentioned in the introduction illustrates that, in general, we cannot hope to have equality between the derivative of the area and the length, or their multidimensional generalizations. All the natural examples with which we deal are convex, compact sets. So a natural question to ask is: "If we consider a one-parameter family $(K_r)_{r\geq 0}$ of convex, compact sets of the usual Euclidean space $\mathbb{R}^n$ which satisfy the relationships studied in this paper, is it necessary that each K_r be the ball $B_n(r)$?"

Of course not! As we have already seen, the family of squares parameterized by the apothem satisfy this relation, and they are not circles. So is there a natural assumption on the K_r which forces them to be balls? Before beginning any calculations, let us provide a precise framework.

Let $(K_r)_{r\geq 0}$ be a family of convex compact sets of the Euclidean space $\mathbb{R}^n$ so that for each of the sets we could define its n-volume $v_n(r)$ and the $(n-1)$-volume $a_{n-1}(r)$ of its boundary. Let us denote by $B_n(r)$ the closed ball with center O and radius r, and by $\mathbb{S}^{n-1}(r) = \partial B_n(r)$ the sphere with center O and radius r.

Let us assume that:

(i) the functions $r \mapsto v_n(r)$ and $r \mapsto a_{n-1}(r)$ are of class $\mathcal{C}^1$ on $[0, +\infty)$;

(ii) for all $r > 0$, the interior of K_r is nonempty.

Continuing to let $V_n(r)$ and $A_{n-1}(r)$, respectively, to be the volume of a Euclidean ball of $\mathbb{R}^n$ with radius r and the volume of its boundary, then, using the famous isoperimetric inequality, (see Berger [1] for example), we get the following relation between V_n, v_n, A_{n-1}, and $a_{n-1}(r)$:

$$\frac{a_{n-1}(r)^n}{v_n(r)^{n-1}} \geq \frac{A_{n-1}(r)^n}{V_n(r)^{n-1}},$$

or, similarly,

$$a_{n-1}(r) \geq \frac{A_{n-1}(r)}{V_n(r)^{1-\frac{1}{n}}} v_n(r)^{1-\frac{1}{n}}.$$

The relation $\frac{A_{n-1}(r)}{V_n(r)^{1-\frac{1}{n}}} = nV_n(1)^{\frac{1}{n}}$, valid for the balls, leads then to

$$a_{n-1}(r) \geq nV_n(1)^{\frac{1}{n}} v_n(r)^{1-\frac{1}{n}}.$$

So, if we suppose now that $\frac{dv_n}{dr} = a_{n-1}$, we get the differential inequality

$$\frac{dv_n}{dr} \geq nV_n(1)^{\frac{1}{n}} v_n(r)^{1-\frac{1}{n}}.$$

Let us define the auxiliary function

$$\varphi_n(r) = v_n(r)^{1/n} - V_n(1)^{\frac{1}{n}} r, \quad r \in [0, +\infty).$$

This function has a derivative on $(0, +\infty)$ because $v_n(r) > 0$ for $r > 0$. This derivative is given by

$$\varphi_n'(r) = \frac{v_n'(r)}{nv_n(r)^{1-\frac{1}{n}}} - V_n(1)^{\frac{1}{n}} = \frac{v_n'(r) - nV_n(1)^{\frac{1}{n}} v_n(r)^{1-\frac{1}{n}}}{nv_n^{1-\frac{1}{n}}}.$$

So $\varphi_n'(r) \geq 0$, and φ_n is an increasing function and therefore nonnegative since $\varphi_n(0) = 0$.

We have obtained the following inequality:

$$\forall r \geq 0, \ v_n(r) \geq V_n(1) r^n = V_n(r).$$

In other words: The volume of K_r is necessarily greater than the volume of the ball with radius r. This is an inequality that we can easily observe with the example of squares parameterized by the apothem.

We can therefore force K_r to be equal to the ball $B_n(r)$ if we assume that $K_r \subset B_n(r)$! Indeed, in this case we have the converse inequality and so, finally, the equality $v_n(r) = V_n(r)$. We can then conclude that for all $r \geq 0$, we have that $K_r = B_n(r)$ and $\partial K_r = \mathbb{S}^{n-1}(r)$ (the case of equality in the isoperimetric inequality).

Finally, because we have already proved that balls satisfy the derivative relation, we can state the following proposition which is finally proved:

Proposition 2. *Let $(K_r)_{r\geq 0}$ be a family of convex, compact sets in the usual Euclidean space $\mathbb{R}^n$. For each set of the family we can define its n-volume $v_n(r)$ and the $(n-1)$-volume $a_{n-1}(r)$ of its boundary. Denote by $B_n(r)$ the closed ball with center O and radius r, and by $\mathbb{S}^{n-1}(r) = \partial B_n(r)$ the sphere with center O and radius r.*

Let us assume that:

(i) the functions $r \mapsto v_n(r)$ and $r \mapsto a_{n-1}(r)$ are of class $\mathcal{C}^1$ on $[0, +\infty)$;

(ii) for all $r > 0$, the interior of K_r is nonempty;

(iii) for all $r \geq 0$, $K_r \subset B_n(r)$.

Then, $\frac{dv_n}{dr} = a_{n-1}$ if and only if for all $r \geq 0$, $K_r = B_n(r)$ (and so $\partial K_r = \mathbb{S}^{n-1}(r)$).

Remark 1.

1. Without hypothesis (iii) this equivalence is false. Indeed, it suffices to consider a family of regular polygons in the plane, parameterized by the apothem.
2. Actually, the previous proposition, which concerns a one-parameter family of convex, compact sets of a Euclidean space, shows that only balls satisfy the property studied in this paper, at least if each of these sets is contained in the corresponding ball. In this result, we can see the balls playing an extremal role relative to the studied relationship in the same way as in the isoperimetric inequality; it is not surprising since this famous inequality is the main tool of the proof.
3. We can illustrate this result in the case, previously and widely mentioned, of a family of squares of the Euclidean plane. For $r > 0$, let us denote K_r the square with center at the origin and side of length $r/\sqrt{2}$. The family $(K_r)_{r>0}$ satisfies all the assumptions of the proposition. Because clearly $K_r \subsetneqq B_2(r)$, we confirm the fact that the studied relationship is not verified in this case.

Manifolds of figures

Presentation of the problem In the first section, we only considered the case of one-parameter families of geometrical figures, and we took the derivative with respect to this parameter (typically the radius in the case of the balls). However, when we consider a general family of geometrical figures, the number of degrees of freedom is greater than one. That is, the set forms what is called, in the learned language of differential geometry, a *differential manifold* of dimension $d \geq 1$. The undergraduate reader must not be afraid of such a word; it is not really necessary to know precisely what a differential manifold is to understand our purpose. It suffices to have in mind that a manifold describes a world where objects are localized unequivocally using a d-tuple of coordinates (corresponding to the number of degrees of freedom mentioned above).

To begin, let us consider the family $\mathcal{P}$ of parallelograms on an affine plane. An element of $\mathcal{P}$ is completely determined by knowledge of three of its vertices, so by six real numbers when working in some affine frame. In other words, $\mathcal{P}$ is a six-dimensional manifold. Let us continue with some other examples. For instance, consider the family $\mathcal{R}$ of rectangles in the Euclidean plane $\mathbb{R}^2$. Because they are particular parallelograms, six real numbers are sufficient to determine each of them. But we can reduce this number because the orthogonality of the sides is a constraint which links the six previous numbers, and we finally have five independent real numbers describing the rectangle, implying that the set $\mathcal{R}$ is a five-dimensional manifold. But, actually, since we are only interested in questions concerning length and area, notions which are invariant under the group of displacements, we are not dealing with the whole set of rectangles, but only with the set—that we will denote by $\mathcal{R}_0$—of these rectangles modulo the displacements. In particular, by first applying some suitable translation, we can assume that one of the vertices is at the origin. The number of real numbers describing such a new rectangle is thereby reduced to 3, but we can also apply a convenient rotation to orient the larger side along the positive x-axis. So, finally, the number of remaining numbers describing the rectangle is 2, the length of its two sides. Such an element of $\mathcal{R}_0$ is therefore unambiguously identified by a pair (x, y) with x (its length) strictly greater than y (its width). Consequently, the "manifold of rectangles modulo displacements" is just the open set of $\mathbb{R}^2$ defined by

$$\mathcal{R}_0 = \{(x, y) \in \mathbb{R}^2,\ 0 < y < x\},$$

and so it is a two-dimensional manifold.

As we wrote in the introduction, a very important notion used in this paper to deal with general geometrical figures is that of *deformation*. Like the concept of a manifold, the notion of a deformation belongs also to the area of differential geometry. But as with the manifolds, the non-initiated reader must not be afraid of this new technical word because it describes a very intuitive idea, easy to keep in mind. Let us explain what it is for the example of rectangles.

A continuous (or differentiable, or of class $\mathcal{C}^r$) deformation is a one-parameter family of rectangles $(R_t)_{t\in I}$ that is represented by a parameterized curve $t \mapsto (x(t), y(t))$, that is continuous (or differentiable, or of class $\mathcal{C}^r$), defined on I, and drawn in $\mathcal{R}_0$. We will come back to this notion later when we speak about its relation with the derivative with respect to a vector field.

Let us take the example of the set $\mathcal{T}$ of triangles. A triangle is defined by three points, and so by six independent real numbers—$\mathcal{T}$ is a six-dimensional manifold. The work done above for rectangles can also be done in this case, and therefore we can reduce the manifold $\mathcal{T}$ to a three-dimensional manifold* $\mathcal{T}_0 = (0, +\infty) \times (0, +\infty) \times (0, \pi)$. In this type of situation, the functions *area* and *perimeter*, denoted respectively by A and P, are differentiable functions on the reduced manifold, and so depend on n variables, where n is the dimension of the manifold. At any rate, we are in a situation where as soon as $n \geq 2$, the problem of the derivative with respect to a variable is not well-defined. In particular, the relation $\frac{dA}{dr} = P$ no longer makes sense. So what is there to say in this context about this question?

Directional derivative and deformations of geometrical figures The key result of this section is the following theorem, though we will defer the proof to the appendix.

Theorem 3. *Let $\mathcal{V}$ be a manifold of plane geometrical figures. Let us denote by A and P the differentiable functions defined on $\mathcal{V}$ with values in $(0, +\infty)$ that respectively assign its area and its perimeter to an element of $\mathcal{V}$. If A is not singular, that is, $dA \neq 0$ on $\mathcal{V}$, then there exists a vector field X such that the derivative of A in the direction of X is equal to P. In other words,*

$$X.A = L_X A = dA(X) = P.$$

We will only verify it with many examples: circles, squares, rectangles, parallelograms, and triangles. In each case, we will give an explicit vector field X that satisfies the condition of the theorem, and we will describe the associated deformation. However, first it is necessary to provide some clarifications about vector fields and their link with the notion of deformation.

If

$$f : (x_1, \ldots, x_d) \mapsto f(x_1, \ldots, x_d)$$

*A triangle is so described by the length of two of its sides and the angle between them.

is a function defined on coordinates $(x_1, \dots, x_d)$, we write $\frac{\partial f}{\partial x_i}$ for the partial derivative of f with respect to the variable x_i. We recall that the differential df of f is given by the formula $df := \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} dx_i$, where dx_i denotes the differential of the function $(x_1, \dots, x_d) \mapsto x_i$.

Now, a vector field on this manifold is a vector-valued function

$$X : (x_1, \dots, x_d) \mapsto \sum_{i=1}^{d} X_i e_i,$$

where the X_i are numerical functions of the variables $x_1, \dots, x_d$ with some regularity according to the context (continuous, differentiable, etc.), and the e_i are the vectors of the canonical basis for $\mathbb{R}^d$. If X is such a vector field, and if f is a differentiable real function defined on the manifold, then we define the derivative of f along X, or with respect to X, or the Lie derivative of f with respect to X, to be the function, denoted by $X.f$, or $L_X f$, defined by

$$X.f = L_X f := df(X) = \sum_{i=1}^{d} X_i \frac{\partial f}{\partial x_i}.$$

One writes $X.f = \left(\sum_{i=1}^{d} X_i \frac{\partial}{\partial x_i} \right).f$, and so the vectors e_i are denoted by $\frac{\partial}{\partial x_i}$.

Now, let m_0 be a point of the manifold corresponding to coordinates $(a_1, \dots, a_d)$, and let $X = \sum_{i=1}^{d} X_i \frac{\partial}{\partial x_i}$ be a vector field. The deformation of m_0 at the time t with respect to the field X is just the unique solution $t \mapsto (x_1(t), \dots, x_d(t))$ of the differential system with initial conditions given by $\forall i,\ \dot{x}_i = X_i,\ x_i(0) = a_i$.

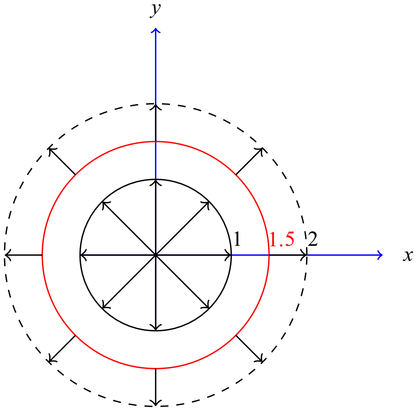


Figure 3 Deformations of a circle for different values of the time.

Figure 3 illustrates the notion of deformation. We can see the deformation after a time $t = 1/2$ of the red circle[†] with radius $3/2$ along the unitary radial vector field into the dashed circle of radius 2. (Note that the smaller circle is the unitary one indicating the unitary vector field).

Readers will be able to convince themselves that this is indeed the generalization of what we have named "deformation" in the introduction to the examples of one-parameter families.

[†] The online version of this article has color diagrams.

Circles, squares and equilateral triangles We have already answered the question for these three very simple cases corresponding to 1-dimensional manifolds. Indeed, for the circles, which were the starting point of our study, we have $\mathcal{V} = (0, +\infty)$ and, with the radius r as variable, $X = \frac{\partial}{\partial r}$. For the squares, we have the same manifold and the same field if the variable is the radius of the inscribed circle or what we have called the apothem. As for the equilateral triangles, we take the radius of the inscribed circle as the variable as well. Then we have $A(r) = 3\sqrt{3}r^2$, $P(r) = 6\sqrt{3}r$, $X = \frac{\partial}{\partial r}$, and the corresponding deformation is illustrated by Figure 4.

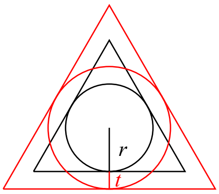


Figure 4 Deformation of an equilateral triangle along the apothem.

Rectangles In the case of the rectangles, the manifold is two-dimensional and given by

$$\mathcal{V} = \{(x, y) \in \mathbb{R}^2,\ 0 < y < x\}.$$

Then we have $A(x, y) = xy$ and $P(x, y) = 2(x + y)$. Now, because $dA = y\,dx + x\,dy$, in order to satisfy the condition $dA(X) = P$, it suffices to take

$$X = 2\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right).$$

Then the associated deformation is shown in Figure 5.

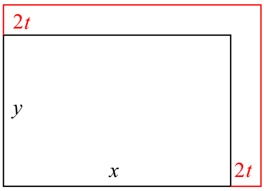


Figure 5 Deformation of a rectangle.

Parallelograms The manifold of parallelograms is three-dimensional; it is the product $(0, \pi) \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ where a parallelogram is described by the angle between two adjacent sides and their lengths. The functions A and P are given by $A(\theta, x, y) = xy \sin\theta$, $P(\theta, x, y) = 2(x + y)$.

The vector field $X = \dfrac{2}{\sin\theta}\left(\dfrac{\partial}{\partial x} + \dfrac{\partial}{\partial y}\right)$ defined above satisfies the relation $dA(X) = P$. The associated differential system is given by:

$$\dot{\theta} = 0, \quad \dot{x} = \frac{2}{\sin\theta}, \quad \dot{y} = \frac{2}{\sin\theta},$$

and its integration leads to the trajectories

$$\theta(t) = \theta_0, \quad x(t) = \frac{2t}{\sin\theta_0} + x_0, \quad y(t) = \frac{2t}{\sin\theta_0} + y_0.$$

This gives the deformation shown in Figure 6.

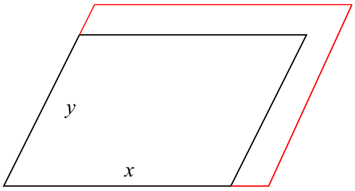


Figure 6 Deformation of a parallelogram.

Triangles Now let us study the case of *triangles*. We have already studied the case of equilateral triangles, whose manifold is one-dimensional. Equilateral triangles have a lot of symmetry properties, and that is why this dimension is so low. Now, if we give up some symmetries, this dimension becomes larger. For example, isosceles triangles need two coordinates. One way to proceed is to choose the length r of equal sides and the angle θ between them.

In these coordinates, the area and perimeter functions are given by

$$A(r,\theta)=\frac{1}{2}r^2\sin(\theta),\;\; P(r,\theta)=2r+2r\sin\left(\frac{\theta}{2}\right).$$

Thus, to get $dA(X)=P$, the field

$$X=\frac{2\left(1+\sin\left(\frac{\theta}{2}\right)\right)}{\sin\theta}\frac{\partial}{\partial r}$$

is suitable. The corresponding deformation is given by the flow of the differential system

$$\dot{\theta}=0,\;\; \dot{r}=\frac{2\left(1+\sin\left(\frac{\theta}{2}\right)\right)}{\sin\theta},$$

which is easy to integrate:

$$\theta(t)=\theta_0,\;\; r(t)=\frac{2\left(1+\sin\left(\frac{\theta_0}{2}\right)\right)}{\sin\theta_0}t+r_0.$$

This last example is illustrated by Figure 7.

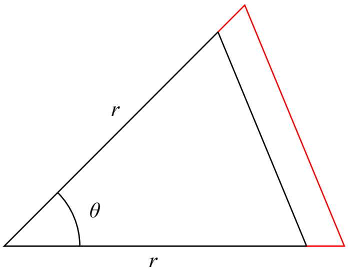


Figure 7 Deformation of an isosceles triangle.

We have previously explained how to determine the manifold of general triangles seen as the set $(0,+\infty)^2\times(0,\pi)$. Unfortunately, with this choice of coordinates the calculation of the perimeter and the area is not easy. So we prefer to see this manifold as the set

$$\mathcal{V}=(0,+\infty)^2\times(0,\pi/2)\ni(r,b,\theta),$$

where the meaning of r, b and θ can be seen in Figure 8.

The area and the perimeter of a triangle of $\mathcal{V}$ are then given by:

$$A(r,b,\theta)=\frac{1}{2}rb$$

$$P(r,b,\theta)=b+\frac{r}{\cos\theta}+\sqrt{r^2+(b-r\tan\theta)^2}.$$

An example of a vector field X satisfying $dA(X) = P$ is

$$X = \frac{2}{\cos\theta}\frac{\partial}{\partial b} + \left(2 + \frac{2}{b}\sqrt{r^2 + (b - r\tan\theta)^2}\right)\frac{\partial}{\partial r}.$$

The determination of its flow leads to the following differential system:

$$\begin{cases} \dot{\theta} = 0 \\ \dot{b} = 2/\cos\theta \\ \dot{r} = 2 + \frac{2}{b}\sqrt{r^2 + (b - r\tan\theta)^2} \end{cases}$$

The first two equations are easy to integrate, leading, respectively, to

$$\theta(t) = \theta_0 \qquad \text{and} \qquad b(t) = \frac{2}{\cos\theta_0}t + b_0.$$

On the other hand, the third equation is terrible! But it is not necessary if we want to visualize the deformation that consists in remaining at a constant angle and in pushing the side perpendicular to this fixed direction, as illustrated in Figure 8.

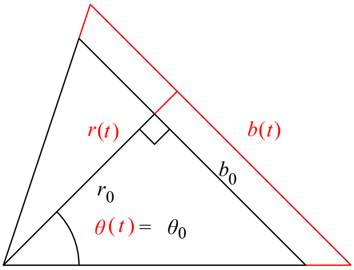


Figure 8 Deformation of a general triangle.

In the particular case of *right-angled triangles*, the previous manifold loses one dimension because the length b is completely determined by r. Moreover, the explicit calculation of the flow is easy. So $\mathcal{V} = (0, +\infty) \times (0, \pi/2) \ni (r, \theta)$. Indeed, in this case the area and perimeter functions are, respectively, given by the formulas:

$$A(r, \theta) = \frac{r^2}{\sin 2\theta}$$

$$P(r, \theta) = \frac{r}{\sin\theta} + \frac{r}{\cos\theta} + \frac{2r}{\sin 2\theta},$$

and the vector field $X = (1 + \cos\theta + \sin\theta)\frac{\partial}{\partial r}$ satisfies the relation $dA(X) = P$. Here, the associated differential system is

$$\dot{r} = 1 + \cos\theta + \sin\theta, \quad \dot{\theta} = 0$$

and this is easy to integrate, leading to $r(t) = (1 + \cos\theta_0 + \sin\theta_0)t + r_0, \ \theta(t) = \theta_0$.

Appendix: Two proofs of Theorem 3

We will give two proofs of the theorem.

First, let g be a Riemannian metric on $\mathcal{V}$. Denote by ∇A the gradient of A with respect to the metric g. It is defined by $\forall Y \in \mathcal{X}(\mathcal{V}), \ dA(Y) = g(\nabla A, Y)$. Since $dA \neq 0$, the vector field ∇A is nowhere equal to zero. Thus, we can define

$$X := \frac{P}{g(\nabla A, \nabla A)}\nabla A.$$

So we have

$$dA(X) = dA\left(\frac{P}{g(\nabla A, \nabla A)}\nabla A\right) = \frac{P}{g(\nabla A, \nabla A)}g(\nabla A, \nabla A) = P,$$

showing that the vector field X is suitable.

Now for the second proof. Since A is not singular, i.e., dA is nowhere zero on the manifold, we can find an atlas $((U, \varphi_U))_{U \in \mathcal{O}}$ of the manifold such that the area function A may be taken as the first coordinate on each open set U. The local existence of a vector field is certain since if $(x_1^U, \dots, x_n^U)$ are coordinates on U, then the field defined on U by $P\frac{\partial}{\partial x_1^U}$ is suitable on U. Now, let us choose a smooth partition of unity subordinate to the open cover $\mathcal{O}$, i.e., a family of smooth functions $(\rho_U)_U$ on the manifold, locally finite, nonnegative and satisfying $\sum_U \rho_U = 1$. If $m \in \mathcal{V}$, let us define

$$X(m) := \sum_U \rho_U(m)P(m)\frac{\partial}{\partial x_1^U}.$$

This field, well-defined on the whole manifold $\mathcal{V}$, satisfies the relation $X.A = P$.

Acknowledgments We thank the two anonymous referees for their remarks, advice, and points of view, which allowed us to produce a more readable and understandable text. We thank also Timothy Neal and Christelle Quéru for their proofreading and for greatly improving the English language of our paper!

REFERENCES

[1] Berger, M. (2009). *Geometry I and II*. Berlin: Springer-Verlag.

[2] Dorff, M., Hall, L. (2003). Solids in $\mathbb{R}^n$ whose area is the derivative of the volume. *College Math. J.* 34(5): 350–358. doi.org/10.1080/07468342.2003.11922029

[3] Dorff, M., Marichal, J-L. (2007). Derivative relationships between volume and surface area of compact regions in $\mathbb{R}^n$. *Rocky Mountain J. Math.* 37(2): 551–571. doi.org/10.1216/rmjm/1181068766.

[4] Emert, J., Nelson, R. (1997). Volume and surface area for polyhedra and polytopes. *Math. Mag.* 70(5): 365–371. doi.org/10.1080/0025570X.1997.11996576

[5] Fjelstad, P., Ginchev, I. (2003). Volume, surface area, and the harmonic mean, *Math. Mag.* 76(2): 126–129. doi.org/10.1080/0025570X.2003.11953164

[6] Krenicky, J. N., Rychtář, J. (2010). On the relationship between volume and surface area. *Involve: A Jour. of Math.* 3(1): 1–8. doi.org/10.2140/involve.2010.3.1

[7] Mneimné, R., Testard, F. (1986). *Introduction à la théorie des groupes de Lie classiques*. Paris: Hermann.

[8] Struss, K. A. (1990). Exploring the volume-surface area relationship. *College Math. J.* 21(1): 40-43. doi.org/10.1080/07468342.1990.11973282

[9] Tong, J. (1997). Area and perimeter, volume and surface area. *College Math. J.* 28(1): 57. doi.org/10.1080/07468342.1997.11973833

Summary. As is shown by the large literature on this subject, we are not the first to be surprised by the fact that, for the circle, the derivative of the area is equal to its perimeter. Our point of view to better understand this seemingly miraculous relationship, comes from differential geometry, leading us to necessary thoughts on what a derivative is and the major role played by changes of coordinates. Moreover, this use of differential geometry seems unavoidable when it comes to studying the case of figures depending on several parameters. Our approach to these questions will take place in the framework of manifolds of figures and will use the notions of directional derivatives and deformations of figures.

HASSAN BOUALEM AND ROBERT BROUZET are assistant professors at Montpellier and Perpignan universities (France), respectively. Their main research area is differential geometry. They started to work together at the end of the nineties. Since then, they have written more than 10 common research papers. They have also been widely involved in teacher training, and they have participated in the writing of three books on mathematical teaching. They are deeply committed to spreading mathematical culture, especially for secondary and undergraduate students.

The Polar Moment of Inertia of the Solid Mylar Balloon

STEPHEN M. ZEMYAN
Penn State Mont Alto
Mont Alto, PA 17237
smz3@psu.edu

The polar moment of inertia I of a solid body about a fixed axis of rotation measures the resistance of the body to changes in its rate of rotation. It is naturally related to several other quantities in the theory of rotational dynamics. If ω denotes the angular velocity of the rotating body and α denotes its angular acceleration, then the angular momentum L is given by $L = I\omega$, the torque τ is given by $\tau = I\alpha$, and the rotational kinetic energy K is given by $K = \frac{1}{2}I\omega^2$.

The simplest definition of the polar moment of inertia of a solid body is given by the integral

$$I = \int r^2 \, dm,$$

where r is the perpendicular distance of the point mass dm to the axis of rotation. We shall assume that all solid bodies under discussion are homogeneous with unit density, so that their masses and volumes are numerically equal.

The polar moment of inertia has been calculated for many solid bodies with circular symmetry and constant density. The following formulas can be found in any standard calculus text that covers applications of multivariable integrals.

- The polar moment of inertia I_C of a solid circular cylinder of radius a and height h is given by $I_C = \frac{1}{2}ma^2$, where $m = \pi a^2 h$.
- The polar moment of inertia I_S of a solid sphere of radius a is given by $I_S = \frac{2}{5}ma^2$, where $m = \frac{4}{3}\pi a^3$.
- The polar moment of inertia I_E of a solid ellipsoid of revolution, whose surface is given by the equation
$$\frac{x^2}{a^2} + \frac{y^2}{a^2} + \frac{z^2}{c^2} = 1,$$
is given by $I_E = \frac{2}{5}ma^2$, where $m = \frac{4}{3}\pi a^2 c$.

The Mylar balloon is physically constructed by gluing two circular sheets of Mylar together along their common edges and then inflating it to full capacity. Paulsen [2, p. 956] showed that the thickness of the inflated Mylar balloon is given by

$$T_M = \frac{2a\sqrt{2}\,\pi^{3/2}}{\left(\Gamma\left(\frac{1}{4}\right)\right)^2},$$

and that its volume is given by

Math. Mag. **95** (2022) 329–331. doi:10.1080/0025570X.2022.2093593 © Mathematical Association of America
MSC: 70E17

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

$$V_M = \frac{a^3\sqrt{2\pi}}{12}\left(\Gamma\left(\frac{1}{4}\right)\right)^2, \tag{1}$$

where a is the radius of the inflated balloon. Here, we assume that the Mylar balloon-shaped object is solid and homogeneous of unit density, so that $V_M = m$.

The Mylar balloon and an ellipsoid of the same inflated radius and thickness are similar in shape. Figure 1 compares a portion of their profile curves. The profile curve for the Mylar balloon is above the profile curve for the ellipsoid.

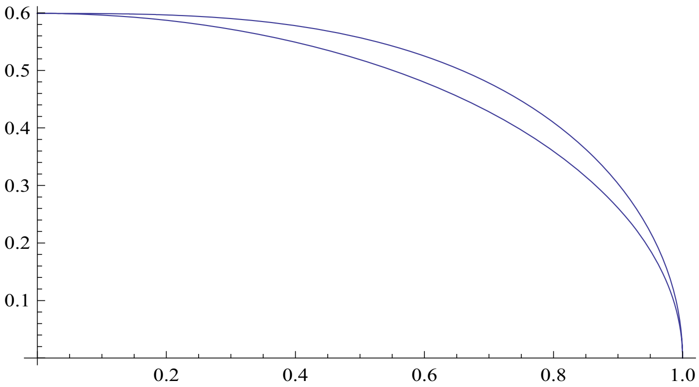


Figure 1 A comparison of the profile curves of the Mylar balloon and an ellipsoid of revolution, generated with *Mathematica* software.

We may now state our main result.

Theorem 1. *The polar moment of inertia I_M of the solid Mylar balloon M is given by $I_M = kma^2$, where m is the mass, a is the radius of the solid balloon, and*

$$k = \frac{9(\Gamma(\frac{3}{4}))^4}{5\pi^2} = \frac{36\pi^2}{5(\Gamma(\frac{1}{4}))^4} \approx 0.4112519\ldots$$

Proof. In order to prove the theorem, we require a parametrization of the profile curve for the Mylar balloon. Several have been presented in the literature, but we prefer the one given by Paulsen [2, p. 955] for its ease of use. By using a classical argument from the calculus of variations, Paulsen showed that the profile curve for the Mylar balloon is given by

$$f(r) = \int_r^a \frac{t^2}{\sqrt{a^4 - t^4}}\,dt,$$

where a is the radius of the inflated balloon, with $0 \le r \le a$. For objects with circular symmetry, the use of cylindrical coordinates is most advantageous. In this setting, the formula for the polar moment of inertia of the Mylar balloon M becomes

$$\begin{aligned}
I_M &= \int\int\int_M r^2\,dV \\
&= 2\int_{\theta=0}^{2\pi}\int_{r=0}^{a}\int_{z=0}^{f(r)} r^2 r\,dz\,dr\,d\theta \\
&= 4\pi\int_{r=0}^{a} r^3 f(r)\,dr = 4\pi\int_{r=0}^{a}\int_{t=r}^{a}\frac{r^3t^2}{\sqrt{a^4-t^4}}\,dt\,dr
\end{aligned}$$

$$= 4\pi \int_{t=0}^{a} \int_{r=0}^{t} \frac{r^3 t^2}{\sqrt{a^4 - t^4}}\, dr\, dt$$

$$= \pi \int_{t=0}^{a} \frac{t^6}{\sqrt{a^4 - t^4}}\, dt.$$

By utilizing the substitution $a^4 w = t^4$, this last integral becomes

$$I_M = \frac{1}{4}\pi a^5 \int_0^1 w^{3/4}(1-w)^{-1/2}\, dw,$$

which may be easily evaluated in terms of the classical beta function [1, p. 24] as

$$I_M = \frac{1}{4}\pi a^5 \frac{\Gamma\left(\frac{7}{4}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{9}{4}\right)} = \frac{3\sqrt{2\pi}}{10}\left(\Gamma\left(\frac{3}{4}\right)\right)^2 a^5.$$

Since

$$a^3 = \frac{12}{\sqrt{2\pi}\left(\Gamma\left(\frac{1}{4}\right)\right)^2}\, m$$

from equation 1, and $\Gamma\left(\frac{1}{4}\right)\Gamma\left(\frac{3}{4}\right) = \pi\sqrt{2}$, the theorem is now proven. ■

In comparison, note that $I_S = I_E < I_M < I_C$.

The following generalization to higher moments can be established in a similar manner. The proof is left as an exercise for the reader.

Theorem 2. *The n^{th} moment of the solid Mylar balloon M is given by*

$$M_n = \int_M r^n\, dm = k_n m a^n,$$

where

$$k_n = \frac{6(n+1)}{(n+2)(n+3)} \frac{\Gamma\left(\frac{n+1}{4}\right)\Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{n+3}{4}\right)\Gamma\left(\frac{1}{4}\right)}.$$

Note that $M_0 = m$ and that $M_2 = I_M$. Note also that the constant k_n can be simplified depending upon the value of $n \pmod 4$. For example, if $n \equiv 0 \pmod 4$, then k_n is rational.

REFERENCES

[1] Bell, W. W. (2004). *Special Functions for Scientists and Engineers*. Mineola: Dover.

[2] Paulsen, W. H. (1994). What is the shape of the Mylar balloon? *Amer. Math. Monthly*. 101(10): 953–958. doi.org/10.2307/2975161

Summary. In this paper, we determine the polar moment of inertia I_M of a solid Mylar balloon-shaped object with unit density by evaluating a triple integral in cylindrical coordinates. This integral is resolved in terms of the classical beta function. The same technique can be employed to determine all of its higher moments as well.

STEPHEN M. ZEMYAN is an emeritus professor of mathematics at Penn State Mont Alto. He pursued his graduate studies at the University of Maryland, College Park under the direction of Brit Kirwan, specializing in complex analysis. His textbook, *The Classical Theory of Integral Equations*, was published in 2012. Other mathematical interests include number theory, difference equations, differential geometry, classical analysis, and conformal dynamics. In his spare time, he enjoys playing the piano, traveling, and studying foreign languages.

A Simple Proof that n^{th} Roots are Always Integers or Irrational

GARY REID LAWLOR
Brigham Young University
Provo, UT 84602
lawlor@math.byu.edu

In the June 2017 issue of this MAGAZINE, Jeffrey Bergen posed a challenge to find the easiest possible proof of the irrationality of those roots of positive integers that are not themselves integers [1]. Here is our attempt to meet Bergen's challenge.

Theorem 1. *If $m, n \in \mathbb{N}$ then $\sqrt[n]{m}$ is either an integer or irrational.*

Proof. Suppose that $\alpha = \sqrt[n]{m} = \frac{p}{q}$ and $r < \alpha < r + 1$ with $m, n, p, q, r \in \mathbb{N}$. For any $k \in \mathbb{N}$, if we expand $(\alpha - r)^k$ with the binomial theorem, then we can use the identity $\alpha^n = m$ to reduce all powers of α to at most $n - 1$, and then combine terms to write $(\alpha - r)^k$ as a fraction with denominator q^{n-1}. But then the set

$$\{\alpha - r,\ (\alpha - r)^2,\ (\alpha - r)^3, \ldots\}$$

contains infinitely many distinct fractions between 0 and 1 that can all be written with denominator q^{n-1}, which is impossible. ■

In addition to Jeffrey Bergen, other worthy entries in the present challenge include those by David M. Bloom [2] and T. Estermann [3] (square roots only) and David Gilat [4] (all roots of monic, integer polynomials).

REFERENCES

[1] Bergen, J. (2017). Is this the easiest proof that *n*th roots are always integers or irrational? *Math. Mag.* 90: 225. doi.org/10.4169/math.mag.90.3.225

[2] Bloom, D. M. (1995). A one-sentence proof that $\sqrt{2}$ is irrational. *Math. Mag.* 68(4): 286. doi.org/10.1080.0025570X.1995.11996338

[3] Estermann, T. (1975). The irrationality of $\sqrt{2}$. *Math. Gazette* 59: 110. doi.org/10.2307/3616647

[4] Gilat, D. (2012). Gauss's Lemma and the irrationality of roots, revisited. *Math. Mag.* 85: 114–116. doi.org/10.4169/math.mag.85.2.114

Summary. We present an especially succinct proof that if m and n are natural numbers, then $\sqrt[n]{m}$ is either an integer or irrational.

GARY REID LAWLOR researches area minimization at Brigham Young University, where he has taught since 1991. Born in Alberta, Canada, Lawlor enjoys hiking, spending time with family, and searching for simpler proofs of classical theorems.

Math. Mag. **95** (2022) 332. doi:10.1080/0025570X.2022.2094628 © Mathematical Association of America
MSC: 11A05

A “Luminous” Solution to the Basel Problem

VILMOS KOMORNIK
Université de Strasbourg,
67084 Strasbourg Cedex, France
vilmos.komornik@math.unistra.fr

Dedicated to the memory of Dominique Dumont.

The Gregory–Leibniz equality

$$1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots=\frac{\pi}{4}$$

may be easily justified (today) by integrating on $(0, 1)$ the identity

$$\frac{1}{1+x^2}=1-x^2+x^4-\cdots;$$

see, for example, Courant and Robbins [2, pp. 441–442] or Komornik and Schäfke [7]. Euler’s equivalent equations

$$1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots=\frac{\pi^2}{6}$$

and

$$1+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\cdots=\frac{\pi^2}{8}$$

have many different proofs, but none of them is as simple; see, for example, Harper [4] Hofbauer [5], or Kalman [6] and their references. The equation

$$1+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\cdots=2\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots\right)^2 \tag{1}$$

allows us to reduce Euler’s theorem to that of Gregory–Leibniz. A tricky direct proof of this equation was outlined by Borwein and Borwein [1, p. 381].

Dumont [3] gave a transparent, but heuristic, proof as follows. He considered the lattice formed by the points of the plane with odd positive integer coordinates and the corresponding function

$$f(2p-1,2q-1):=\frac{(-1)^{p+q}}{(2p-1)(2q-1)}.$$

He introduced, for $n = 1, 2, \ldots$, the “light ray” R_n reflected by the angle 45° at the points $(0, 2n)$ and $(2n, 0)$; see Figure 1.

Denoting by r_n the sum of the values of f at the lattice points lying on R_n, and observing that each lattice point belongs to two rays, except for those on the diagonal D_0 with $p = q$ that belong only to one ray, he obtained the formal equality

$$\sum_{p=1}^{\infty}\frac{1}{(2p-1)^2}+\sum_{n=1}^{\infty}r_n=2\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{(-1)^{p+q}}{(2p-1)(2q-1)}=2\left(\sum_{p=1}^{\infty}\frac{(-1)^{p+1}}{2p-1}\right)^2.$$

Math. Mag. **95** (2022) 333–337. doi:10.1080/0025570X.2022.2094138 © Mathematical Association of America
MSC: 40A25

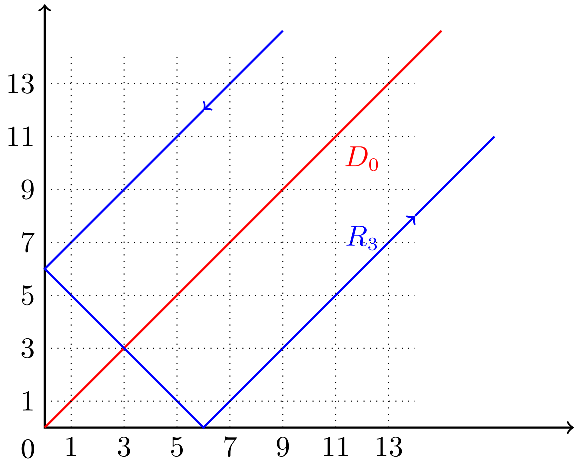


Figure 1 The ray R_3 and the diagonal D_0.

He concluded by checking that all subsums r_n vanish. The computation is not justified because $\sum r_n$ and the double series are not absolutely convergent. In the next section we make his idea rigorous.

Proof of equation (1)

Fix a large positive integer N, and consider within the square $(0, 2N) \times (0, 2N)$ the lattice points $A = (x_A, y_A)$ with odd integer coordinates. If we denote by S_n and D_n the sets of lattice points that lie on the lines of equation $y = 2n - x$ and $y = 2n + x$, respectively, then each lattice point belongs to a unique set among $S_1, \dots, S_{2N-1}$, and to a unique set among $D_{1-N}, \dots, D_{N-1}$; see Figure 2. Therefore, denoting by s_i and d_j the sum of the values of f at the points of S_i and D_j, respectively, we have the following equation:

$$\sum_{i=1}^{2N-1} s_i + \sum_{j=1-N}^{N-1} d_j = 2\sum_{p=1}^{N}\sum_{q=1}^{N} f(2p-1, 2q-1) = 2\left(\sum_{p=1}^{N} \frac{(-1)^p}{2p-1}\right)^2.$$

Since

$$d_0 = \sum_{p=1}^{N} \frac{1}{(2p-1)^2},$$

equation (1) will follow from the limit relation

$$s_N + \sum_{n=1}^{N-1} (s_n + s_{2N-n} + d_n + d_{-n}) \to 0 \quad \text{as} \quad N \to +\infty.$$

We can also write it in the form

$$r_1 + \cdots + r_{N-1} + s_N \to 0 \quad \text{as} \quad N \to +\infty, \tag{2}$$

where $r_n := s_n + s_{2N-n} + d_n + d_{-n}$ is the sum of the values of f at the lattice points of the rectangle R_n described by the light ray R_n in the square $[0, 2N] \times [0, 2N]$, starting from the point $(2n, 0)$ by the angle 45°, and reflected by the four sides of the square;

see Figure 3. Note that R_N is a degenerate rectangle, coinciding with a diagonal of the square, and $r_N = 2s_N$.

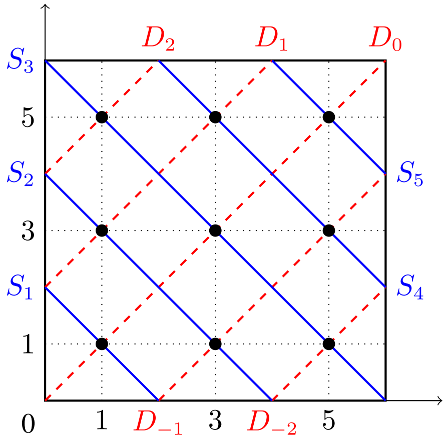


Figure 2 The lattice points and the lines S_n, D_n within the square for $N = 3$.

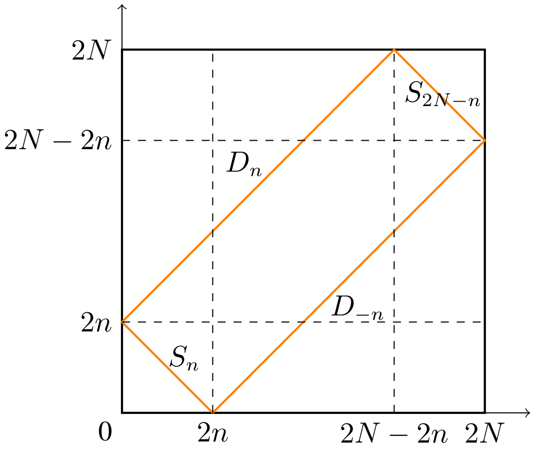


Figure 3 A reflected light ray R_n.

Let us compute s_n, d_n, and r_n. Using the identities

$$\frac{1}{xy} = \frac{1}{x+y}\left(\frac{1}{x}+\frac{1}{y}\right) \quad \text{and} \quad \frac{1}{xy} = \frac{1}{y-x}\left(\frac{1}{x}-\frac{1}{y}\right),$$

we have

$$s_n = (-1)^{n+1} \sum_{A \in S_n} \frac{1}{x_A y_A} = \frac{(-1)^{n+1}}{2n} \sum_{A \in S_n} \left(\frac{1}{x_A}+\frac{1}{y_A}\right) \tag{3}$$

for $n = 1, \ldots, 2N-1$, and

$$d_n = (-1)^n \sum_{A \in D_n} \frac{1}{x_A y_A} = \frac{(-1)^n}{2n} \sum_{A \in D_n} \left(\frac{1}{x_A}-\frac{1}{y_A}\right) \tag{4}$$

for $n = 1, \dots, N$. The exponents of -1 come from the observation that if $x_A = 2p - 1$ and $y_A = 2q - 1$, then on S_n we have

$$p + q = \frac{x_A + y_A + 2}{2} = n + 1, \quad \text{and thus} \quad (-1)^{p+q} = (-1)^{n+1},$$

while on D_n we have

$$q - p = \frac{y_A - x_A}{2} = n, \quad \text{so that} \quad (-1)^{p+q} = (-1)^{q-p} = (-1)^n.$$

From equation (4), we deduce that:

$$d_n = \frac{(-1)^n}{2n} \left(\sum_{A \in S_n} \frac{1}{x_A} - \sum_{A \in S_{2N-n}} \frac{1}{y_A} \right). \tag{5}$$

For this, first we note that if A runs over D_n, then the denominators x_A and y_A in the last sum of equation (4) run over the odd integers in the intervals $[0, 2N - 2n]$ and $[2n, 2N]$, respectively; see Figure 3. The fractions with $x_A, y_A \in [2n, 2N - 2n]$ eliminate each other, and the sum reduces to

$$\sum_{\substack{A \in D_n \\ x_A < 2n}} \frac{1}{x_A} - \sum_{\substack{A \in D_n \\ y_A > 2N - 2n}} \frac{1}{y_A}.$$

Equation (5) now follows by observing that the last expression is equal to

$$\sum_{A \in S_n} \frac{1}{x_A} - \sum_{A \in S_{2N-n}} \frac{1}{y_A}.$$

Indeed, in both expressions, x_A and y_A run over the odd integers in the intervals $[0, 2n]$ and $[2N - 2n, 2N]$, respectively; see Figure 3 again.

By symmetry, we get from equation (5) the equation

$$d_{-n} = \frac{(-1)^n}{2n} \left(\sum_{A \in S_n} \frac{1}{y_A} - \sum_{A \in S_{2N-n}} \frac{1}{x_A} \right), \tag{6}$$

and then we infer the following from equations (3), (5), and (6):

$$\begin{aligned} r_n &= s_n + d_n + d_{-n} + s_{2N-n} \\ &= \frac{(-1)^{n+1}}{2n} \sum_{A \in S_{2N-n}} \left(\frac{1}{x_A} + \frac{1}{y_A} \right) + \frac{(-1)^{2N-n+1}}{4N - 2n} \sum_{A \in S_{2N-n}} \left(\frac{1}{x_A} + \frac{1}{y_A} \right) \\ &= (-1)^{n+1} \frac{4N}{2n(4N - 2n)} \sum_{A \in S_{2N-n}} \left(\frac{1}{x_A} + \frac{1}{y_A} \right) \\ &= (-1)^{n+1} \frac{2N}{2N - n} \left(\frac{1}{n} \sum_{p=N-n+1}^{N} \frac{1}{2p - 1} \right). \end{aligned}$$

If $1 \le n \le N - 1$, then

$$\frac{2N}{2N - n} < \frac{2N}{2N - n - 1},$$

and

$$\frac{1}{n}\sum_{p=N-n+1}^{N}\frac{1}{2p-1}<\frac{1}{n+1}\sum_{p=N-n}^{N}\frac{1}{2p-1}$$

because the arithmetic mean increases when we add a larger number to the set. Therefore, we have

$$|r_1|<|r_2|<\cdots<|r_{N-1}|<|r_N|\,.$$

Since $r_1,\ldots,r_{N-1}$ have alternating signs and $r_N=2s_N$, it follows that

$$\begin{aligned}|r_1+\cdots+r_{N-1}+s_N|&\le|r_1+\cdots+r_{N-1}|+|s_N|\\&\le|r_{N-1}|+|s_N|<3\,|s_N|\,.\end{aligned}$$

This implies equation (2) because if $N\to+\infty$, then

$$\begin{aligned}|s_N|=\frac{1}{N}\sum_{p=1}^{N}\frac{1}{2p-1}&\le\frac{1}{N}\left(1+\int_1^N\frac{1}{2x-1}\,dx\right)\\&=\frac{1}{N}+\frac{\ln(2N-1)}{2N}\to 0.\end{aligned}$$

Acknowledgment The author thanks the referee for her/his comments on the first version of this note.

REFERENCES

[1] Borwein, J. M., Borwein, P. (1987). *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity.* New York: Wiley.

[2] Courant, R., Robbins, H. (1941). *What Is Mathematics? An Elementary Approach to Ideas and Methods.* Oxford: Oxford Univ. Press.

[3] Dumont, D. (1992). Une preuve lumineuse de la relation $1+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\cdots=2\big(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots\big)^2$. *L'Ouvert.* 69: 17–20.

[4] Harper, J. D. (2003). Another simple proof of $1+\frac{1}{2^2}+\frac{1}{3^2}+\cdots=\frac{\pi^2}{6}$. *Amer. Math. Monthly.* 110(6): 540–541. doi.org/10.1080/00029890.2003.11919994

[5] Hofbauer, J. (2002). A simple proof of $\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$ and related identities. *Amer. Math. Monthly.* 109(2): 196–200. doi.org/10.80.0029890.2002.11919855

[6] Kalman, D. (1993). Six ways to sum a series. *College Math. J.* 24(5): 402–421. doi.org/10.1080/07468342.1993.11973562

[7] Komornik, V., Schäfke, R. (2021). Leibniz, Newton, and Cauchy: a complex relationship. *Amer. Math. Monthly.* 128(4):367–369. doi.org/10.1080/00029890.2021.1867465.

Summary. Years ago, D. Dumont gave a beautiful heuristic proof relating two famous equalities of Gregory–Leibniz and Euler. The purpose of this note is to make his proof rigorous.

VILMOS KOMORNIK (MR Author ID: 104490) studied in Budapest and has taught in Hungary and France. He had collaborated with J.-L. Lions on control theory and with P. Erdős on combinatorial number theory. He likes the artistic aspects of mathematics, and he has collected many elegant results and proofs in his textbooks. He is an external member of the Hungarian Academy of Sciences.

Heavy Metal ODEs

JOHN E. KAMPMEYER, III
Elizabethtown College
Elizabethtown, PA 17022
jkampmeyer3@gmail.com

TIMOTHY J. MCDEVITT
Elizabethtown College
Elizabethtown, PA 17022
McDevittT@etown.edu

Consider the deceptively simple differential equation

$$f'(x) = f^{-1}(x). \tag{1}$$

Because of the presence of the inverse, none of the solution techniques that students typically study in differential equations courses can be applied to equation (1). We cannot even construct a direction field or apply any standard numerical methods to see the nature of the solutions. To make matters worse, the inverse may restrict the domain of the solution in an unexpected way.

This problem dates back to at least 1968, where it was first posed as an Elementary Problem in the *American Mathematical Monthly*. In the published solution [2], the author proved that the only real-valued solution for $x > 0$ is

$$f(x) = \varphi\left(\frac{x}{\varphi}\right)^{\varphi}, \tag{2}$$

where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio. A graph of that solution can be seen in Figure 1. This solution is remarkable for a couple of reasons. First, the appearance

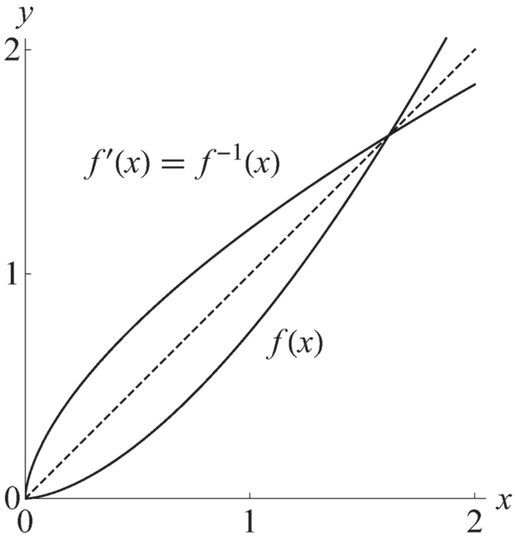


Figure 1 Graph of $f(x) = \varphi\left(\frac{x}{\varphi}\right)^{\varphi}$ and $f'(x) = f^{-1}(x)$ for $x > 0$.

Math. Mag. **95** (2022) 338–348. doi:10.1080/0025570X.2022.2094168 © Mathematical Association of America
MSC: Primary 34A99, Secondary 30F99

of the golden ratio is unexpected, and second, one usually finds unique solutions to differential equations only when they are augmented with an initial condition.

In this paper, we find additional solutions to equation (1) using techniques that involve an interesting combination of differential equations and elementary complex variable theory.

Some special ratios

As we saw above, the unique, real-valued solution of equation (1) for $x > 0$ involves the golden ratio $\varphi = (1+\sqrt{5})/2 \approx 1.618$, which is the positive solution of the quadratic equation $b^2 - b - 1 = 0$. The second solution is $1 - \varphi \approx -0.618$. Along with π and e, the golden ratio is one of the most well-known mathematical constants. For a geometric interpretation of φ, consider a rectangle with dimensions $(\varphi + 1) \times \varphi$, as shown in Figure 2. If we carve out a $\varphi \times \varphi$ square from one end of the rectangle, then the remaining $\varphi \times 1$ rectangle is similar to the original $(\varphi + 1) \times \varphi$ rectangle.

Throughout this paper, we will make liberal use of some algebraic identities involving the golden ratio such as $\varphi^2 = \varphi + 1$ and $\varphi - 1 = 1/\varphi$.

A related number is the golden *angle* ψ. Referring to Figure 3, if a circle is broken into two disjoint arcs, with the length of the longer arc being φ times larger than the length of the smaller arc, then the golden angle ψ is the interior angle subtended by the smaller arc. Since $\varphi\psi + \psi = 2\pi$, we have that

$$\psi = \frac{2\pi}{1+\varphi} \approx 2.400.$$

We will see below that the golden angle naturally appears in complex-valued solutions of equation (1).

It is worth mentioning that the golden ratio is the first number in a more general sequence called the *metallic numbers* [3]. For a positive integer n, the nth metallic number M_n is the positive solution of $b^2 - nb - 1 = 0$, or

$$M_n = \frac{n + \sqrt{n^2+4}}{2}, \quad n \in \mathbb{N}. \tag{3}$$

The first few values in the sequence $\{M_n\}_{n=1}^{\infty}$ are $\{1.618, 2.414, 3.303, 4.236, \ldots\}$, and for large n, $M_n \approx n$. For a geometric interpretation of M_n, consider a rectangle with dimensions $(nM_n + 1) \times M_n$. If we remove n adjacent $M_n \times M_n$ squares from one side, then the remaining $M_n \times 1$ rectangle is similar to the original large rectangle. See Figure 4 for an example with $n = 5$. We will see below that the metallic numbers appear in the solutions of a generalization of equation (1),

$$f^{(n)}(x) = f^{-1}(x), \quad n \in \mathbb{N}. \tag{4}$$

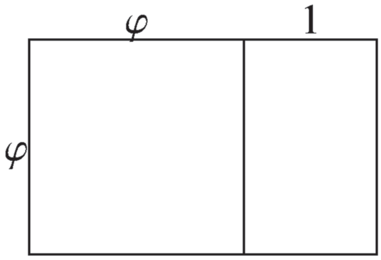


Figure 2 The aspect ratio of this rectangle is $1/\varphi$, where φ is the golden ratio.

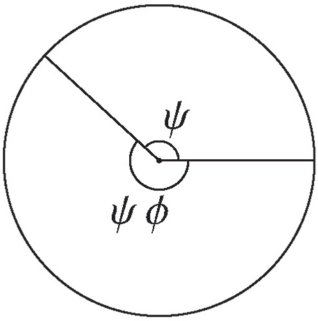


Figure 3 The ratio of the larger central angle to the smaller central angle is ψ, the golden angle.

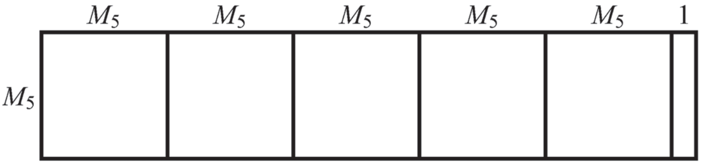


Figure 4 If M_5 is the fifth metallic number, then the large $(5M_5 + 1) \times M_5$ rectangle is similar to the small $M_5 \times 1$ rectangle at the right end.

Background

A video on a popular YouTube channel [4] considers solutions of equation (1) in the form of power functions, specifically $f(x) = ax^b$, where a and b are constants. The creator of the video claims that since

$$f^{-1}(x) = \left(\frac{x}{a}\right)^{1/b} = \frac{x^{1/b}}{a^{1/b}} = abx^{b-1} = f'(x), \tag{5}$$

we have that $a = b^{-b/(b+1)}$ and $b^2 - b - 1 = 0$, implying that b is either the golden ratio φ or its companion value $1 - \varphi$. In addition to equation (2), this appears to produce a second solution,

$$f(x) = (-\varphi)^{-\varphi} x^{1-\varphi}, \tag{6}$$

but, since the solution in equation (2) is unique for $x > 0$, equation (6) cannot be correct. So what went wrong? The first hint is that $(-\varphi)^{-\varphi} \approx -0.166 + 0.428i$ is not real, and familiar algebraic rules often do not apply for complex variables. For example, if a or b is a positive real number, then $(ab)^c = a^c b^c$, but a misapplication of that rule can lead to obviously incorrect results such as

$$-1 = i^2 = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)(-1)} = \sqrt{1} = 1. \tag{7}$$

The same problem occurs in equation (5) since

$$\left(\frac{x}{a}\right)^{1/b} \neq \frac{x^{1/b}}{a^{1/b}}$$

in general, and this leads to the spurious solution in equation (6).

Let us review some familiar rules from algebra in the context of complex numbers. First, we recall the notion of the *principal value* of a complex number z. If we write a nonzero* z in exponential form as $Ze^{i\zeta}$, where $Z > 0$, then we require ζ to be in

*The number 0 is an exception because Arg(0) is undefined.

$(-\pi, \pi]$. This choice of ζ is called the *principal argument* of z and is denoted by $\operatorname{Arg}(z)$. For nonzero complex numbers z and w, $\operatorname{Log} z = \ln|z| + i \operatorname{Arg}(z)$, and we take the principal value of z^w to be defined as $z^w = e^{w \operatorname{Log} z}$. For example,

$$\begin{aligned} i^{3+2i} &= e^{(3+2i)\operatorname{Log} i} = e^{(3+2i)(\ln|i| + i\operatorname{Arg}(i))} \\ &= e^{(3+2i)(\ln 1 + i\pi/2)} = e^{-\pi} e^{i3\pi/2} = e^{-\pi} e^{-i\pi/2}. \end{aligned} \tag{8}$$

Note that $e^{i3\pi/2}$ and $e^{-i\pi/2}$ have the same value $(-i)$, but we choose the latter because $-\pi/2$ is in $(-\pi, \pi]$, and this follows our convention of using the principal argument. In addition, even though, for example, $i = e^{i5\pi/2} = e^{i\pi/2}$, if we were to use $\operatorname{Arg}(i) = 5\pi/2$ instead of $\operatorname{Arg}(i) = \pi/2$ at the beginning of the calculation in equation (8), then we would erroneously compute the modulus of i^{3+2i} to be $e^{-5\pi}$. Therefore, it is very important to remember to keep arguments of complex numbers in the interval $(-\pi, \pi]$.

As a result, if a, b, and c are nonzero complex numbers, then

$$(ab)^c = a^c b^c e^{-2\pi i p c}, \tag{9}$$

where p is an integer such that $-\pi < \operatorname{Arg}(a) + \operatorname{Arg}(b) - 2\pi p \leq \pi$. Note that the introduction of $-2\pi p$ guarantees that the argument of the product ab is in $(-\pi, \pi]$. We conclude that $(ab)^c = a^c b^c$ only if pc is an integer, and this resolves the false proof that $-1 = 1$ in equation (7). Recomputing equation (7), if we let $a = b = -1 = e^{i\pi}$ and $c = 1/2$, then $p = 1$ in equation (9) and

$$1 = \sqrt{(-1)(-1)} = \sqrt{-1}\sqrt{-1}e^{-2\pi i(1)(1/2)} = (i)(i)(-1) = 1$$

instead of -1. It is also useful to point out that we recover the familiar rule from high school algebra that $(ab)^c = a^c b^c$ if either a or b is a positive real number. If, for example, $a > 0$, then $\operatorname{Arg}(a) = 0$, $p = 0$, and equation (9) reduces to

$$(ab)^c = a^c b^c, \qquad a > 0. \tag{10}$$

In a similar way,

$$\left(a^b\right)^c = a^{bc} e^{-2\pi i q c}, \tag{11}$$

where q is an integer such that $-\pi < \operatorname{Re}(b)\operatorname{Arg}(a) + \operatorname{Im}(b)\ln|a| - 2\pi q \leq \pi$. In this case, the introduction of $-2\pi q$ keeps the argument of a^b in $(-\pi, \pi]$, and we also have that $\left(a^b\right)^c = a^{bc}$ only if qc is an integer. To see the importance of this in a numerical example, consider

$$\left(i^3\right)^i = (-i)^i = e^{i\operatorname{Log}(-i)} = e^{\pi/2} \approx 4.810,$$

whereas

$$i^{3i} = e^{3i\operatorname{Log}(i)} = e^{3i(i\pi/2)} = e^{-3\pi/2} \approx 0.00898.$$

This apparent discrepancy is resolved by choosing $a = c = i$, $b = 3$, and $q = 1$ in equation (11) so that

$$\left(i^3\right)^i = i^{3i} e^{-2\pi i(1)(i)} = e^{-3\pi/2} e^{2\pi} = e^{\pi/2}.$$

The rule in equation (11) also resolves a second bogus proof that $-1 = 1$:

$$-1 = (-1)^{1/3} = (-1)^{2/6} = \left[(-1)^2\right]^{1/6} = 1^{1/6} = 1.$$

According to equation (11) with $a = -1$, $b = 2$, and $c = 1/6$, q must be 1 so that

$$\left[(-1)^2\right]^{1/6} = (-1)^{2(1/6)} e^{-2\pi i(1)(1/6)} = (-1)^{1/3} e^{-\pi i/3} = e^{\pi i/3} e^{-\pi i/3} = 1$$

since the principal cube root of -1 is $e^{\pi i/3}$.

What is f^{-1}?

Before we proceed to find complex-valued solutions to equation (1), we need to clarify the meaning of $f^{-1}(x)$ because the inverse of a complex-valued function can be multivalued. As mentioned above, to eliminate that ambiguity, we always take the principal value of any complex number z.

Using principal values of complex numbers can lead to domain restrictions. For example, if $g(z) = z^2$ and $g^{-1}(z) = z^{1/2}$, then

$$g^{-1}\left(g\left(e^{i\pi/4}\right)\right) = \left[\left(e^{i\pi/4}\right)^2\right]^{1/2} = \left(e^{i\pi/2}\right)^{1/2} = e^{i\pi/4},$$

but

$$\begin{aligned} g^{-1}\left(g\left(e^{3i\pi/4}\right)\right) &= \left[\left(e^{3i\pi/4}\right)^2\right]^{1/2} \\ &= \left(e^{-i\pi/2}\right)^{1/2} = e^{-i\pi/4} \neq e^{3i\pi/4}. \end{aligned}$$

Let us extend this analysis to all $z \in \mathbb{C}$ by letting $z = Ze^{i\zeta}$. Then, by equation (11), we have

$$\begin{aligned} g^{-1}\left(g(z)\right) &= \left[\left(Ze^{i\zeta}\right)^2\right]^{1/2} \\ &= \left[Z^2e^{2i\zeta}\right]^{1/2} = Ze^{i(2\zeta - q\pi)/2} = ze^{-iq\pi/2}, \end{aligned} \tag{12}$$

where q is an integer such that

$$-\pi < 2\zeta - q\pi \leq \pi. \tag{13}$$

From equation (12), we see that $g^{-1}\left(g(z)\right) = z$ only if $q = 0$, and equation (13) subsequently implies that $-\pi/2 < \zeta \leq \pi/2$. On the other hand, if we compose g and g^{-1} in the opposite order, then

$$g\left(g^{-1}(z)\right) = \left[\left(Ze^{i\zeta}\right)^{1/2}\right]^2 = \left[Z^{1/2}e^{i\zeta/2}\right]^2 = Ze^{i\zeta} = z$$

for all $z \in \mathbb{C}$. So $g^{-1}(z) = z^{1/2}$ is an inverse of g for both composition orders ($g \circ g^{-1}$ and $g^{-1} \circ g$) provided that $-\pi/2 < \text{Arg}(z) \leq \pi/2$.

We therefore reformulate equation (1) as

$$f\left(f'(z)\right) = z \quad \text{and} \quad f'\left(f(z)\right) = z \tag{14}$$

for $z \in \mathbb{C}$, using principal values. That is, the derivative of f must be a bona fide inverse regardless of the order of composition.

Solving $f'(z) = f^{-1}(z)$

We now solve equation (1), reformulated as equation (14), by seeking power function solutions of the form $f(z) = az^b$ for complex a, b, and z. If we write $a = Ae^{i\alpha}$, $b = Be^{i\beta}$, and $z = Ze^{i\zeta}$, then

$$f\left(f'(z)\right) = a(abz^{b-1})^b = a\left(ABZ^{b-1}e^{i(\alpha+\beta+(b-1)\zeta)}\right)^b.$$

Since AB is real and positive, equation (10) implies that we can pull it outside of the parentheses without introducing any additional exponential factors. We also use equations (9) and (11) to pull the Z^{b-1} term out of the parentheses and combine the exponents $b - 1$ and b. In all, using equations (9)–(11),

$$f\left(f'(z)\right) = a(AB)^b\left(Z^{b-1}e^{i(\alpha+\beta+(b-1)\zeta)}\right)^b, \qquad \text{by (10)},$$

$$= a(AB)^b \left(Z^{b-1}\right)^b \left(e^{i(\alpha+\beta+(b-1)\zeta)}\right)^b e^{-2\pi i p b}, \qquad \text{by (9)},$$

$$= a(AB)^b Z^{b^2-b} \left(e^{i(\alpha+\beta+(b-1)\zeta)}\right)^b e^{-2\pi i p b} e^{-2\pi i q b}, \qquad \text{by (11)},$$

for some integers p and q satisfying

$$-\pi < \operatorname{Arg}\left(Z^{b-1}\right) + \operatorname{Arg}\left(e^{i(\alpha+\beta+(b-1)\zeta)}\right) - 2\pi p \le \pi \tag{15}$$

and

$$-\pi < \operatorname{Re}(b-1)\operatorname{Arg}(Z) + \operatorname{Im}(b-1)\ln Z - 2\pi q \le \pi. \tag{16}$$

For $f(f'(z)) = z = Ze^{i\zeta}$, matching powers of Z implies that $b^2 - b = 1$, which in turn implies that $b = \varphi$ or $b = 1-\varphi$. In either case, since b is real and Z is positive, we can conclude that $p = 0$ in equation (15), and $q = 0$ in equation (16).

To summarize our progress so far, we have

$$f\left(f'(z)\right) = a(AB)^b Z \left[e^{i(\alpha+\beta+(b-1)\zeta)}\right]^b,$$

which can be further reduced using equation (11) to

$$f\left(f'(z)\right) = a(AB)^b z e^{ib(\alpha+\beta)} e^{-2\pi i m b}, \tag{17}$$

where m is an integer such that $-\pi < \alpha + (b-1)\zeta - 2m\pi \le \pi$ and b is either φ or $1-\varphi$. Now, we consider the two possible values for b.

The $b = \varphi$ case First, if $b = \varphi$, then $B = \varphi$, $\beta = 0$, and equation (17) becomes

$$f(f'(z)) = A^{1+\varphi}\varphi^{\varphi} e^{i[(1+\varphi)\alpha - 2m\pi\varphi]} z, \tag{18}$$

where m is an integer such that

$$-\pi < \alpha + (\varphi-1)\zeta - 2m\pi \le \pi. \tag{19}$$

For equation (18) to reduce to $f(f'(z)) = z$, we need $A^{1+\varphi}\varphi^{\varphi} = 1$ and

$$(1+\varphi)\alpha - 2m\pi\varphi = 2k\pi$$

for some integer k, which gives

$$A = \varphi^{1-\varphi} \quad \text{and} \quad \alpha = \psi(k + m\varphi), \tag{20}$$

where ψ is the golden angle. Substituting our new expression for α into equation (19) gives

$$-\pi < \psi(k-m) + (\varphi-1)\zeta \le \pi. \tag{21}$$

We now turn our attention to determining which (m, k) pairs satisfy equation (21) and $-\pi < \alpha \le \pi$.

Since $-\pi < \alpha \le \pi$, equation (20) implies that

$$-\frac{1+\varphi}{2} < k + m\varphi \le \frac{1+\varphi}{2}, \tag{22}$$

and since $-\pi < \zeta \le \pi$, (21) implies that

$$-\pi\varphi \le -\pi - (\varphi-1)\zeta < \psi(k-m) \le \pi - (\varphi-1)\zeta < \pi\varphi,$$

which is equivalent to

$$-\frac{2\varphi+1}{2} < k - m < \frac{2\varphi+1}{2}. \tag{23}$$

The only possible values for (m, k) that satisfy the inequalities in equations (22) and (23) can readily be seen in Figure 5, but we can also derive those solutions algebraically. Multiplying equation (23) by φ and adding it to equation (22) gives

$$-\left(\frac{3}{2} + \frac{1}{2\varphi}\right) < k \le \frac{3}{2} + \frac{1}{2\varphi} \approx 1.81,$$

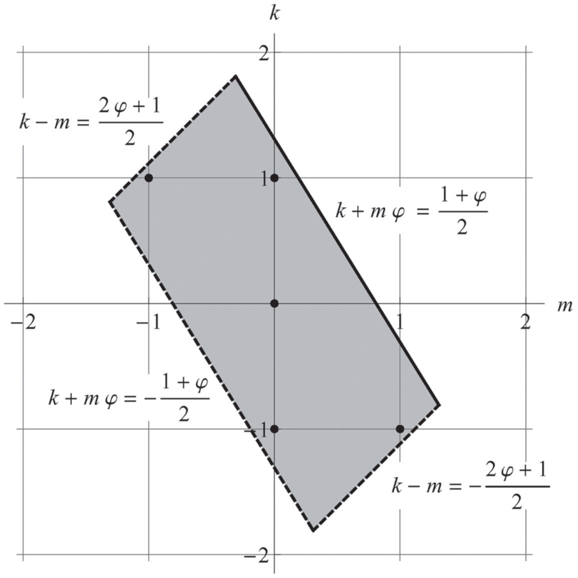


Figure 5 The shaded area is a graphical representation of the constraints in equations (22) and (23). Note that there are only five points with integer coordinates that satisfy equations (22) and (23): $(m, k) = (0, 0)$, $(0, 1)$, $(0, -1)$, $(1, -1)$, and $(-1, 1)$.

so $k \in \{-1, 0, 1\}$. Equation (22) then implies that there are five (m, k) pairs that satisfy both equations (22) and (23):

$$(0, 0), \quad (0, 1), \quad (0, -1), \quad (1, -1), \quad (-1, 1).$$

If we now return to equation (21) and solve for ζ, then we have

$$-\psi\left[\frac{1}{2} + \varphi(1 - m + k)\right] < \zeta \le \psi\left[\frac{1}{2} + \varphi(1 + m - k)\right].$$

However, we also must have $-\pi < \zeta \le \pi$, so

$$-\min\left\{\pi, \psi\left[\frac{1}{2} + \varphi(1 - m + k)\right]\right\} < \zeta$$
$$\le \min\left\{\pi, \psi\left[\frac{1}{2} + \varphi(1 + m - k)\right]\right\}. \qquad (24)$$

The five (m, k) pairs shown in Figure 5 give five solutions of $f(f'(z)) = z$,

$$f(z) = \varphi\left(\frac{z}{\varphi}\right)^{\varphi}, \quad \varphi e^{i\psi}\left(\frac{z}{\varphi}\right)^{\varphi}, \quad \varphi e^{-i\psi}\left(\frac{z}{\varphi}\right)^{\varphi},$$
$$\varphi e^{i\psi/\varphi}\left(\frac{z}{\varphi}\right)^{\varphi}, \quad \varphi e^{-i\psi/\varphi}\left(\frac{z}{\varphi}\right)^{\varphi},$$

with corresponding domains specified by (24),

$$(-\pi, \pi], \quad \left(-\pi, \frac{\psi}{2}\right], \quad \left(-\frac{\psi}{2}, \pi\right],$$
$$\left(\psi\left(\varphi - \frac{1}{2}\right), \pi\right], \quad \left(-\pi, -\psi\left(\varphi - \frac{1}{2}\right)\right],$$

respectively.

TABLE 1: Solutions of equation (14) of the form az^{φ}.

$f(z)$	Interval for $\zeta = \mathrm{Arg}(z)$	
$\varphi\left(\frac{z}{\varphi}\right)^{\varphi}$	$\left(-\frac{\pi}{\varphi}, \frac{\pi}{\varphi}\right]$	
$\varphi e^{i\psi}\left(\frac{z}{\varphi}\right)^{\varphi}$	$\left(\frac{\psi}{2\varphi^2}, \frac{\psi}{2}\right]$	
$\varphi e^{-i\psi}\left(\frac{z}{\varphi}\right)^{\varphi}$	$\left(-\frac{\psi}{2}, -\frac{\psi}{2\varphi^2}\right]$	

Now we need to determine which solutions of $f(f'(z)) = z$ also satisfy $f'(f(z)) = z$. For the first solution, $f(z) = \varphi\left(\frac{z}{\varphi}\right)^{\varphi}$, our rule for powers of powers in (11) implies that

$$f'(f(z)) = \varphi\left[\left(\frac{z}{\varphi}\right)^{\varphi}\right]^{\varphi-1} = (z^{\varphi})^{\varphi-1} = ze^{-2\pi iq(\varphi-1)},$$

for an integer q such that

$$-\pi < \varphi\zeta - 2\pi q \leq \pi.$$

For $f'(f(z))$ to reduce to z, we need $q = 0$ and $-\pi/\varphi < \zeta \leq \pi/\varphi$, which restricts the domain of the solution. Therefore, $f(f'(z)) = z$ and $f'(f(z)) = z$ for the first solution provided that $-\pi/\varphi < \zeta \leq \pi/\varphi$.

For the second solution, $f(z) = \varphi e^{i\psi}\left(\frac{z}{\varphi}\right)^{\varphi}$, a similar analysis shows that

$$\begin{aligned} f'(f(z)) &= \varphi e^{i\psi}\left[e^{i\psi}\left(\frac{z}{\varphi}\right)^{\varphi}\right]^{\varphi-1} \\ &= e^{i\psi}ze^{i(\varphi-1)(\psi-2\pi)} = ze^{i2\pi} = z, \end{aligned}$$

provided that $\psi/(2\varphi^2) < \zeta \leq \pi$. The intersection of $(-\pi, \psi/2]$ and $(\psi/(2\varphi^2), \pi]$ gives $(\psi/(2\varphi^2), \psi/2]$ as the domain of the second solution of (14). The analysis for the third solution is almost identical to the second, and the results are summarized in Table 1.

For the fourth solution of $f(f'(z)) = z$, $f(z) = \varphi e^{i\psi/\varphi}\left(\frac{z}{\varphi}\right)^{\varphi}$,

$$\begin{aligned} f'(f(z)) &= \varphi e^{i\psi/\varphi}\left[e^{i\psi/\varphi}\left(\frac{z}{\varphi}\right)^{\varphi}\right]^{\varphi-1} \\ &= e^{i\psi/\varphi}ze^{i(\varphi-1)(\psi/\varphi+2\pi)} = ze^{i4\pi} = z, \end{aligned}$$

provided that $-\pi < \zeta \leq \psi(2-\varphi/2)$. However, the intersection of this interval with $(\psi(\varphi - 1/2), \pi]$ is empty, so this solution of equation (14) is not valid. A similar analysis also reveals that the fifth solution is not valid.

The first solution in Table 1 is a complex version of the original solution in equation (2). We also see from Table 1 that there is only one for (real-valued) $z > 0$, which is consistent with the analysis in Hindmarsh [2]. Figure 6[†] shows the Riemann surface for the first solution in Table 1. Specifically, the real part of $f(z)$ is graphed as the third dimension and the shading indicates the imaginary part of $f(z)$. The part of the surface that is displayed in a checkerboard fashion is actually excluded by the domain

[†] The online version of this paper has color diagrams.

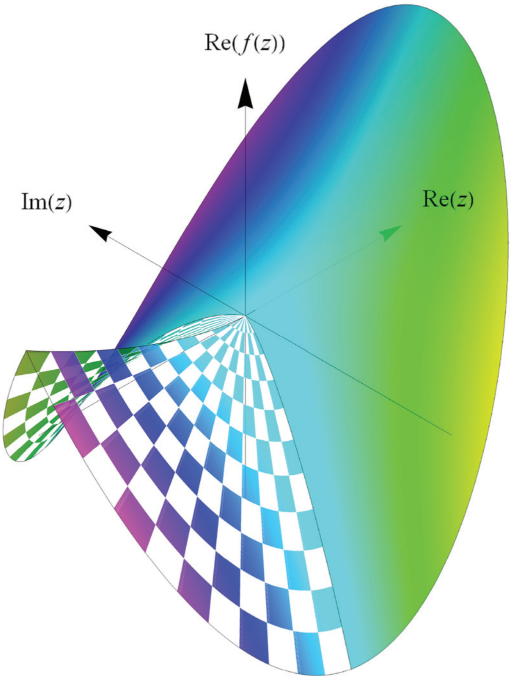


Figure 6 Riemann surface for $f(z) = \varphi\left(\frac{z}{\varphi}\right)^{\varphi}$ for $-\pi < \zeta \leq \pi$. The solid part of the surface corresponds to $-\pi/\varphi < \zeta \leq \pi/\varphi$ and the shading corresponds to Im$(f(z))$.

restriction $-\pi/\varphi < \zeta \leq \pi/\varphi$, so the solution really only consists of the solid part of the surface.

Figure 7 shows the Riemann surface for $f'(z)$ along with an attempt to visually graph the inverse of the solution in Figure 6. The graph of the inverse of a real-valued function is its reflection in the identity. In a similar way, the second graph in Figure 7 switches the roles of z and $f(z)$ from Figure 6; however, this makes Re(z) take on two values if ζ is not in $(-\pi/\varphi, \pi/\varphi]$, and this is indicated with a checkerboard style. For graphical equality of $f'(z)$ and $f^{-1}(z)$ in Figure 7, the domain of ζ must be restricted to $(-\pi/\varphi, \pi/\varphi]$.

The $b = 1 - \varphi$ case Now we return to equation (17) and consider the second value of $b = 1 - \varphi$. Since $1 - \varphi < 0$, $B = \varphi - 1$, $\beta = \pi$, and equation (17) reduces to

$$f\left(f'(z)\right) = A^{2-\varphi}\varphi^{1/\varphi}e^{i[(2-\varphi)\alpha+(2m-1)(\varphi-1)\pi]}z, \tag{25}$$

where m is an integer such that

$$-\pi < \alpha - \varphi\zeta - (2m-1)\pi \leq \pi. \tag{26}$$

For $f(f'(z)) = z$, we must have

$$A = \varphi^{-\varphi} \qquad \text{and} \qquad \alpha = \varphi\pi\left[2(k\varphi - m) + 1\right],$$

where k is an integer. Like we did for the $b = \varphi$ case, we need to complete our analysis by finding (m, k) pairs that simultaneously satisfy equation (26) and the fact that $-\pi < \alpha \leq \pi$.

Plugging the expression for α into equation (26) and using the fact that $-\pi < \zeta \leq \pi$, we find that $-1 < k - m \leq 0$, but since m and k are integers, they must be equal. Using $m = k$ in $-\pi < \alpha \leq \pi$, we find that $m = k = -1$, and that implies, from equation (26), that

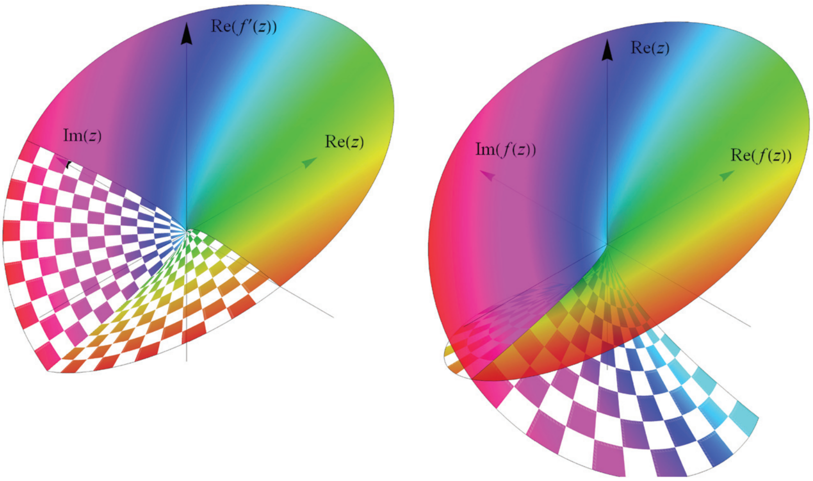


Figure 7 Riemann surfaces for $f'(z) = \varphi\left(\frac{z}{\varphi}\right)^{\varphi-1}$ for $-\pi < \zeta \leq \pi$ (left) and an "inversion" of Figure 6 (right). The solid part of the surfaces correspond to $-\pi/\varphi < \zeta \leq \pi/\varphi$ and the shading corresponds to Im($f'(z)$) (left) and Im(z) (right).

$$\pi \leq \zeta < \pi + \frac{2\pi}{\varphi},$$

from which we conclude that $\zeta = \pi$. This means that by looking for *complex* solutions, we have found a second *real* solution that solves both $f(f'(x)) = x$ and $f'(f(x)) = x$,

$$f(x) = -\frac{1}{\varphi^{\varphi}|x|^{\varphi-1}} = (1-\varphi)\left(\frac{x}{1-\varphi}\right)^{1-\varphi}, \qquad x < 0, \tag{27}$$

which can be readily identified as the original solution to equation (2) after replacing φ with its complementary value $1-\varphi$! Note that the existence of a second solution does not contradict the result in Hindmarsh [**2**], which established that the solution in equation (2) is unique for *positive* x.

Heavy metal addendum

We close by briefly considering the generalized problem in equation (4), reinterpreted as solving

$$f\left(f^{(n)}(z)\right) = z \qquad \text{and} \qquad f^{(n)}(f(z)) = z \tag{28}$$

simultaneously for complex z. Finding functions of a complex variable that satisfy equation (28) requires careful use of equations (9) and (11) and a lot of patience! For the sake of brevity, we only consider real-valued solutions of a real variable. We leave it as an exercise to the interested reader to work out the details for the general complex variable case.

Taking $f(x) = ax^b$ and differentiating n times yields

$$f^{(n)}(x) = ab(b-1)\cdots(b-n+1)x^{b-n} = a\frac{\Gamma(b+1)}{\Gamma(b-n+1)}x^{b-n},$$

where Γ denotes the gamma function. Composing f and $f^{(n)}$ in either order and equating the result to z implies that $b^2 - nb - 1 = 0$, which has solutions involving the metallic numbers. In particular, $b = M_n$ and $b = n - M_n$.

Using the primary root, $b = M_n$, we find

$$f(x) = \left[\frac{\Gamma(M_n - n + 1)}{\Gamma(M_n + 1)}\right]^{M_n/(M_n+1)} x^{M_n}, \quad x > 0, \tag{29}$$

for all integers $n \geq 1$ and

$$f(x) = -\left|\frac{\Gamma(M_n - n + 1)}{\Gamma(M_n + 1)}\right|^{M_n/(M_n+1)} (-x)^{M_n}, \quad x < 0, \tag{30}$$

for even integers $n \geq 2$. The absolute value in equation (30) is not strictly necessary because $\Gamma(M_n - n + 1)$ and $\Gamma(M_n + 1)$ are both positive, but including the absolute value allows us to formally obtain the other real solutions by replacing M_n with its complementary value $n - M_n$ in (29) and (30), respectively. Specifically, the complementary solutions are

$$f(x) = \left[\frac{\Gamma(n - M_n + 1)}{\Gamma(1 - M_n)}\right]^{1/(M_n-1)} x^{n-M_n}, \quad x > 0, \tag{31}$$

for all even integers $n \geq 2$ and

$$f(x) = -\left|\frac{\Gamma(n - M_n + 1)}{\Gamma(1 - M_n)}\right|^{1/(M_n-1)} (-x)^{n-M_n}, \quad x < 0. \tag{32}$$

for all integers $n \geq 1$. The absolute value in equation (32) is needed because $\Gamma(1 - M_n)$ is positive for even n but negative for odd n, whereas $\Gamma(n - M_n + 1) > 0$. Finally, we conclude by observing that if $n = 1$, then $M_1 = \varphi$ and equations (29) and (32) reduce to the solutions in equations (2) and (27).

REFERENCES

[1] Euler, L. (1738). De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt. *Comment. acad. scientiarum Petropolitanae*. 36–57.

[2] Hindmarsh, A. C. (1969). Solution of Elementary Problem 2105. *Amer. Math. Monthly*. 76(6): 690–701. doi.org/10.1080/00029890.1969.12000310

[3] de Spinadel, V. W. (1998). The metallic means and design. *Nexus II: Arquitecture&Mathematics* 5: 141–157.

[4] Derivative equals inverse. Available at: https://www.youtube.com/watch?v=0IlWyIaMXqI&t=492s&index=2&list=WL. (accessed June 2022).

Summary. We solve the differential equation $f'(z) = f^{-1}(z)$, where z is complex, and show that the solutions involve the golden ratio φ. As an addendum, we briefly consider solutions of $f^{(n)}(z) = f^{-1}(z)$, n a natural number, which involve the family of metallic numbers M_n, defined as the solutions to the equations $M_n^2 - nM_n - 1 = 0$.

JOHN E. KAMPMEYER, III earned his B.S. in mathematics, with a concentration in pure mathematics, from Elizabethtown College in 2019. He currently works as a systems engineer at Lockheed Martin Valley Forge in King of Prussia, PA. One of John's many hobbies is finding and evaluating integrals that *Mathematica* cannot compute.

TIMOTHY J. MCDEVITT is a professor of mathematics at Elizabethtown College, where he has taught since 2005. He does a lot of interdisciplinary research and he especially enjoys doing research with undergraduates.

"Pass the Buck" on a Complete Binary Tree

KENNETH LEVASSEUR
University of Massachusetts, Lowell
Lowell, MA 01854
kenneth_levasseur@uml.edu

In the 1970s, Engel [2] devised the stochastic abacus as a way to solve certain discrete probability problems with minimal numerical computation. More recently, Torrence used the same technique to determine winning probabilities for players in the game "Pass the Buck" [6] for a variety of families of graphs. The stochastic abacus has found more widespread exposure due to a recent article by Propp in *Math Horizons* [4]. In this note, the game is analyzed for complete binary trees where the root is the start vertex, and we derive winning probabilities for nodes at different levels. We show that the limiting probability for the root to win as the number of levels goes to infinity is $\sqrt{2} - 1$. We also observe that the derivation easily generalizes to complete ℓ-ary trees.

Pass the Buck

The game "Pass the Buck" is played on a connected, undirected graph with a distinguished "start vertex." The game proceeds in steps starting with the start vertex holding a prize (the "buck"). At each stage in the game, a vertex is selected randomly and uniformly from among the vertex currently holding the buck and its neighboring vertices. If the current vertex is selected, then the game ends with that vertex winning. If a neighboring vertex is selected, then the buck is passed there and the process is repeated. More precisely, if the degree of the vertex that holds the buck is k, then the buck moves to any of the neighbors with probability $\frac{1}{k+1}$, and the game ends with the player at the current vertex winning with probability $\frac{1}{k+1}$.

The stochastic abacus

For different graphs, the probabilities of any vertex winning can be derived in a variety of ways. For example, we can develop a system of equations that determines the probabilities. The game can also be modeled as a Markov chain, and our desired probabilities can be computed using well-known techniques. See Kemeny and Snell [3] for a general introduction to Markov chains and Snell [5] for a discussion of the connection between Markov chains and the Stochastic Abacus. Alternatively, the Stochastic Abacus method (also known as Engel's Algorithm) uses only elementary transition rules to compute winning probabilities. We will illustrate all three methods for the case of a complete binary tree up to level 1 (see Figure 1), which will serve as a basis for computing the probabilities in larger trees.

Math. Mag. **95** (2022) 349–353. doi:10.1080/0025570X.2022.2094165 © Mathematical Association of America
MSC: 60J22

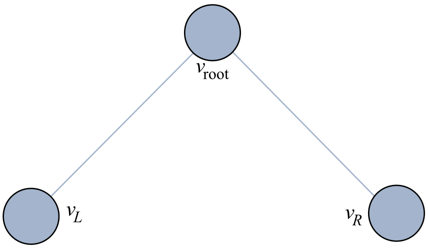


Figure 1 A complete binary tree to one level

In this form of the game, we have three players (root, L, and R), and the buck starts at the root. We can easily derive the probabilities for each vertex by observing that

$$p_{root} = \frac{1}{3} + \frac{2}{3}\left(\frac{1}{2}p_{root}\right),$$

implying that $p_{root} = 1/2$. By symmetry, $p_L = p_R = \frac{1}{4}$. This is the shortest of our derivations, but it does not scale so easily.

Next, we will derive the probabilities using Markov chain theory. The states v_x corresponding to each of the three vertices are states for which the game is in progress, and they are non-absorbing states for the process. We add three absorbing states t_x to represent winning outcomes for each player. An absorbing state is one in which the process, having arrived at that state, never leaves, thus representing the end of the game in our case. The transition matrix for the process with ordering of states v_L, v_{root}, v_R, t_L, t_{root}, (t_R) is

$$T = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

By listing absorbing states last, the form of the transition matrix of an absorbing Markov chain with k absorbing states is

$$\left[\begin{array}{c|c} Q & R \\ \hline \mathbf{0} & I_k \end{array}\right].$$

In our case, $k = 3$. If there are a total of m states, then Q is an $(m-k) \times (m-k)$ matrix of transition probabilities between the non-absorbing states. The transition probabilities from non-absorbing states to absorbing states is contained within R. The matrix NR, where $N = (I - Q)^{-1}$ is the matrix of probabilities into the different absorbing states. The j^{th} row of NR contains the probabilities of ending in the absorbing states, assuming the process starts in state j.

In our example,

$$NR = \left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 \end{pmatrix}\right)^{-1} \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{5}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{4} & \frac{5}{8} \end{pmatrix}. \quad (1)$$

We are mostly concerned with the middle row, which gives probabilities that are consistent with the previous derivation. For this simple case, the graph we have considered is also a path graph. The probabilities for the cases where the starting position is either v_L or v_R are consistent with the general case of “pass the buck” on a path graph starting at an end vertex, as analyzed by Torrence [6].

The same probabilities we have observed twice will now be arrived at using a stochastic abacus. The abacus is constructed by first considering the binary tree to be directed, with each undirected edge becoming a pair of directed edges. The vertices of the tree correspond to the non-absorbing states of the Markov chain, but we will refer to them here as “internal vertices.” Then we add absorbing vertices to the graph, one for each internal vertex, and an edge leading into each absorbing vertex, as in Figure 2. The root is designated as the starting position. Initially, we deposit chips into the internal vertices, the number of chips being one less than the outdegree of each vertex in the abacus, as indicated in each vertex of Figure 2. At this point, the system is “critically loaded.” The process then consists of, sequentially, adding a chip to the start vertex v_{root}, and then repeatedly “firing” chips whenever the content of an internal vertex is greater than or equal to the outdegree of that vertex. This involves distributing a chip along each outgoing edge of the “loaded vertex” to neighboring vertices. This process of adding chips and firing as long as possible continues until the chip content of the internal vertices returns to the critical loading state.

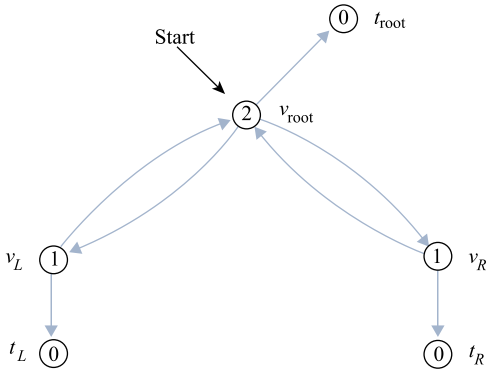


Figure 2 Stochastic Abacus for a complete level one binary tree

The remarkable fact is that after we have returned to the original critical loading of internal vertices, the probability that any vertex wins the game is equal to the number of chips in its corresponding absorbing vertex divided by the total number of chips in all absorbing vertices. See Snell [5] for a proof.

Table 1 provides a step by step account of how the process plays out in our example.

It may not be obvious, but if two vertices can fire, as is the case before steps 4 and 5, it does not matter in what order they are fired. See Bjöner [1] for a proof. After step 9, the three interior vertices are back to being critically loaded, and the process ends. The total number of chips in the absorbing vertices is 4 and the root had 2, so its probability of winning is $\frac{1}{2}$, consistent with our previous derivation. The other two vertices again have winning probability $\frac{1}{4}$.

Pass the Buck on a complete binary tree

Consider the game of Pass the Buck on a complete binary tree with n full levels, $n \geq 0$, where the buck starts at the root. We have seen three approaches that can be used in

TABLE 1: Applying the stochastic abacus to a level 1 binary tree.

Step	Comment	v_{root}	v_L	v_R	t_{root}	t_L	t_R
1	Critically loaded	2	1	1	0	0	0
2	Add 1 to v_{root}	3	1	1	0	0	0
3	v_{root} fires	0	2	2	1	0	0
4	v_L fires	1	0	2	1	1	0
5	v_R fires	2	0	0	1	1	1
6	Add 1 to root	3	0	0	1	1	1
7	v_{root} fires	0	1	1	2	1	1
8	add 1 to v_{root}	1	1	1	2	1	1
9	add 1 to v_{root}	2	1	1	2	1	1

this general case. However, Engel's method is the only one that easily scales to reach the conclusions that follow. We derive formulae for the probabilities that any vertex at level k of the tree, with $0 \leq k \leq n$, will win the game. Our derivation is based on the observation that the chips needed at different levels is recursive, with a second order recurrence.

Theorem 1. *The number of chips in the absorbing vertex of the root at the end of the stochastic abacus process, denoted by $a(n)$, follows the recursion $a(n) = 4a(n-1) - 2a(n-2)$, for $n \geq 2$.*

Proof. Suppose a complete binary tree is critically loaded to n levels and augmented with absorbing vertices. Then in order to return the two subtrees starting at level 1 to critical loading status, each vertex at level 1, which are roots of binary trees of level $n-1$, must fire $a(n-1)$ times. Every time these vertices fire, they need four chips. The initial loading of three chips to each of the vertices at level 1 are used in the first firing, but then to return to critical loading, three other chips are needed. This suggests the root must fire $4a(n-1)$ times, but there is one other source of chips to each vertex at level 1. Those are the vertices at level 2. Each time they fire, they return a chip back up to level 1. Therefore, the root must fire $a(n) = 4a(n-1) - 2a(n-2)$ times to complete the process. This is then the number of chips that are deposited into the absorbing vertex of the root. ■

We will apply this recursion, together with the two base cases $a(0) = 1$ and $a(1) = 2$, to compute our probabilities.

Corollary 1.1. *At the end of the stochastic abacus process on a complete binary tree of level n, the number of chips deposited into each absorbing vertex at level k, for $0 \leq k \leq n$, is $a(n-k)$. The total number of chips that are in all absorbing vertices is then $\sum_{k=0}^{n} 2^k a(n-k)$.*

Corollary 1.2. *With the initial conditions $a(0) = 1$ and $a(1) = 2$, we have $a(n) = \frac{1}{2}((2-\sqrt{2})^n + (2+\sqrt{2})^n)$, and the total number of chips in all absorbing vertices at the end of the stochastic abacus process is*

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

If we select one of the 2^k vertices at level k of the tree, then let $p(n,k)$ be the probability that it wins the level n game. The probability that the root is the winner of the level n game is

$$p(n,0) = \frac{a(n)}{t(n)} = \frac{\sqrt{2}((2-\sqrt{2})^n + (2+\sqrt{2})^2)}{(2+\sqrt{2})^{n+1} - (2-\sqrt{2})^{n+1}}.$$

Interestingly, $\lim_{n\to\infty} p(n,0) = \sqrt{2} - 1$. More generally,

$$p(n,k) = \frac{a(n-k)}{t(n)} \quad \text{and} \quad \lim_{n\to\infty} p(n,k) = \frac{\sqrt{2}}{(2+\sqrt{2})^{k+1}}.$$

Pass the buck on ℓ-ary trees

We can generalize our argument from binary trees to ℓ-ary trees, for $\ell \geq 2$. A complete ℓ-ary tree to level n, $n > 0$, will have a root and ℓ subtrees, each an ℓ-ary tree to level $n-1$. An ℓ-ary tree up to level 0 is a single vertex, which is the form of a leaf for larger trees. By the same logic as the binary case, if we critically load the augmented directed graph for a complete ℓ-ary tree up to level n, $n \geq 2$, then the number of firings of the start vertex that are needed to return to the critically loaded state, $a(\ell, n)$, satisfies the recursion $a(\ell, n) = (\ell+2)a(\ell, n-1) - \ell a(\ell, n-2)$. The base cases are $a(\ell, 0) = 1$ and $a(\ell, 1) = 2$.

REFERENCES

[1] Bjöner, A., Lovasz, L., Shor, P. (1991). Chip-firing games on graphs. *Eur. J. Comb.* 12(4): 283–291. doi.org/10.1016/s0195-6698(13)80111-4.

[2] Engel, A. (1976), Why does the probablistic abacus work? *Educ. Stud. Math.* 7: 59–69. doi.org/10.1007/BF00144359

[3] Kemeny, J. G., Snell, J. L. (1976). *Finite Markov Chains*. New York: Springer-Verlag.

[4] Propp, J. (2018). Prof. Engel's marvelously improbable machines. *Math Horizons*. 26(2): 5–9. doi.org/10.1080/10724117.2018.1518840.

[5] Snell, J. L. *The Engel algorithm for absorbing Markov chains*, Available at: https://arxiv.org/abs/0904.1413v1.

[6] Torrence, B. Passing the buck and firing Fibonacci: Adventures with the stochastic abacus. *Amer. Math. Monthly*. 126(5): 387–399. doi.org/10.1080/00029890.2019.1577089.

Summary. We employ the stochastic abacus to compute winning probabilities at each level of the game "Pass the Buck" on a complete binary tree with the starting vertex being the root of the tree. The derivation is also generalized to play on a complete ℓ-ary trees.

KENNETH LEVASSEUR is a professor of mathematics at the University of Massachusetts, Lowell.

Isometric Miquel Configurations of Points and Circles

GÁBOR GÉVAY
Bolyai Institute
University of Szeged
Szeged, Hungary
gevay@math.u-szeged.hu

TOMAŽ PISANSKI
University of Primorska
Koper, Slovenia
and Institute of Mathematics,
Physics, and Mechanics
University of Ljubljana
Ljubljana, Slovenia
Tomaz.Pisanski@upr.si

There are several classical incidence theorems involving circles which are attributed to Auguste Miquel since they occur in the third part of a series of his 1838 papers published in Liouville's journal [15]. One of them is the following, usually called the "Six Circle Theorem."

Theorem 1. *Consider four circles C_1, C_2, C_3, C_4 in the plane. Suppose that C_1 and C_2 intersect at P_1 and Q_1, that C_2 and C_3 intersect at P_2 and Q_2, that C_3 and C_4 intersect at P_3 and Q_3, and that C_4 and C_1 intersect at P_4 and Q_4. Then the points P_1, P_2, P_3, P_4 are concyclic if and only if Q_1, Q_2, Q_3, Q_4 are concyclic.*

The various proofs of this theorem use different techniques. For example: oriented angles (Richter-Gebert [19]), the complex cross ratio (Hahn [11]), cubic curves (Coolidge [1]), and bracket algebra (Richter-Gebert [19]), among others. In the particular case of the arrangement of the six circles shown on the left in Figure 1, there is a remarkably simple proof that could well be called a "Proof Without Words" [7, 18] (see the image on the right of Figure 1).

It deserves a brief digression to see why the proof is clear from direct observation. Consider the six quadrilaterals on the right-hand side. One can choose five of them, say, the one in the center of the figure and the four others adjacent to it. By assumption, they are all cyclic quadrilaterals. Hence, the opposite angles add up to π in each of them. This implies that so do the angles in the sixth, outermost quadrilateral. Thus, it is cyclic as well, and hence the theorem is proved.

Note that each of the cyclic quadrilaterals on the right in Figure 1 are convex.* This makes it possible to arrange the circumscribed circles on the left in Figure 1 in such a way that one can clearly distinguish an outer and an inner circle, as well as a closed chain formed by the remaining four circles with their points of intersection, as specified by the theorem. In a more general setting, like the one in Figure 2, some of the quadrilaterals are not convex. Hence, the structure of the circle configuration cannot be seen so easily, and the proof is less obvious than before.

Note also that each of the right-hand images in Figures 1 and 2 can be considered as a graph G associated to the given configuration in the following way: G has the configuration points as vertices. Two vertices of G are adjacent if they belong to two

Math. Mag. **95** (2022) 354–361. doi:10.1080/0025570X.2022.2094164 © Mathematical Association of America
MSC: Primary 05B30, Secondary 51B99

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

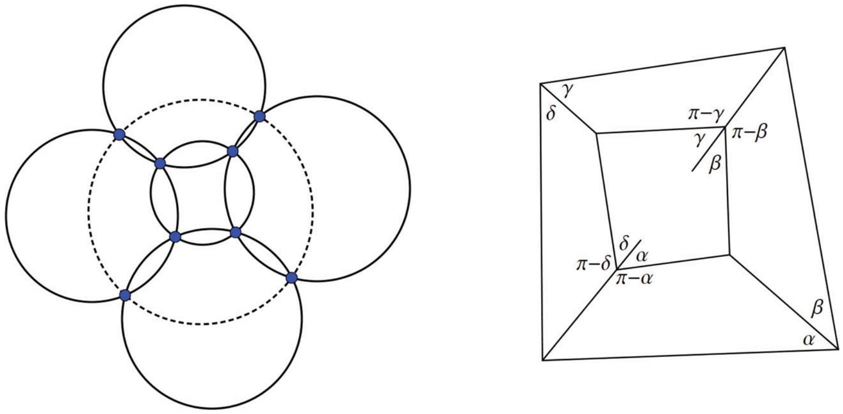


Figure 1 One arrangement of the six circles in the Miquel configuration (left), and a visual proof of the theorem for this particular type of arrangement (right). The points $P_1, \dots, P_4$ in Miquel's theorem correspond to the points on the small inner circle, while the points $Q_1, \dots, Q_4$ correspond to the points on the outer, dotted circle.

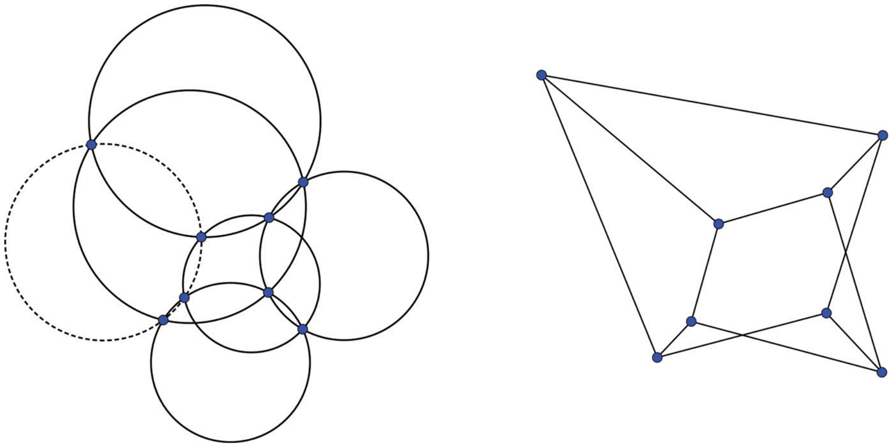


Figure 2 A general representation of the Miquel configuration (left), and the graph formed by the cyclic quadrilaterals inscribed in its circles (right).

configuration circles (they represent the intersection of two circles). It follows that each 4-cycle of G can then be viewed as a cyclic quadrilateral.

Miquel's six circle theorem plays a key role in the axiomatization of the inversive plane by van der Waerden and Smid [20]. Specifically, if the theorem holds, then the associated coordinate skew field is commutative. Thus, its status is similar to that of Pappus' theorem in projective geometry.

There is another context where the theorem's relevance can be compared to that of Pappus' theorem. It is one of the first incidence theorems that lead to configurations of points and circles; its type is $(8_3, 6_4)$. In this sense, the relationship with Pappus' theorem, which, in turn, leads to a configuration of points and lines (of type (9_3)), is clear.

Recall that a configuration $\mathcal{C}$ is said to be of type (p_q, n_k) if it consists of p points and k blocks, such that each point is incident with precisely q of the blocks, while each block is incident with precisely n of the points. In a geometric configuration, the set of blocks may consist of lines, circles, or anything else. If $p = n$, then clearly $q = k$. In this case, we use the more concise notation (n_k), and, following Grünbaum, we use the term *balanced configuration* (for other details on configurations, see the recent monographs by Grünbaum [10], and by Pisanski and Servatius [17]). We emphasize that, historically, the first geometric configurations originated from incidence theorems. The classical examples for configurations of points and lines is Pappus' (9_3) and Desargues' (10_3) configuration [3, 13].

Peculiarly, during the long period since the discovery of Miquel's six circle theorem, a simple question seems to have escaped the attention of authors: *Can the Miquel configuration be realized with circles of equal radius?*

We note that a configuration of points and circles in which all circles are of equal radius is called *isometric* [9]. The nonappearance of this question in the literature is even more peculiar in light of a result by Ziegenbein [21]. He proved as long ago as 1940 that all the members of the infinite series of point-circle configurations originating from Clifford's famous "chain of theorems" are isometric. (Clifford discovered his chain about one and a half centuries after Miquel's theorem! [4]).

Here we show that the Miquel configuration can be realized isometrically.

Theorem 2. *The Miquel configuration can be represented as an isometric configuration. Moreover, if five of its circles have unit radius, then the sixth circle must also have unit radius.*

In our proof, we shall use the notion of a Levi graph. Let $\mathcal{C}$ be a configuration. Following Coxeter [3], a graph, denoted by $L(\mathcal{C})$, is called the *Levi graph* of $\mathcal{C}$ if

- it is bipartite,
- its bipartition classes are in one-to-one correspondence with the set of vertices and the set of blocks of $\mathcal{C}$, respectively, and
- two of its vertices are connected by an edge if and only if the corresponding point and block in $\mathcal{C}$ are incident.

For example, Figure 3 shows the Pappus configuration and its Levi graph (called the *Pappus graph*) [17].

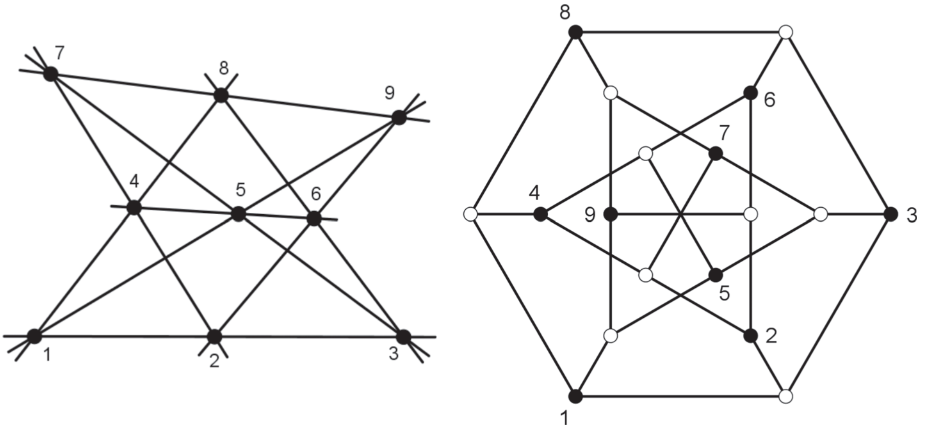


Figure 3 The Pappus configuration (left), and its Levi graph (right).

The *combinatorial structure* of any configuration (p_q, n_k) is completely determined by a (q, k)-regular bipartite graph with a given black-and-white vertex coloring, where black vertices correspond to points, and white vertices correspond to blocks. Such a graph is called a *colored Levi graph*.

Consider a configuration $\mathcal{C}$ in which each block is incident with three points, and apply the following construction on its Levi graph $L(\mathcal{C})$: For each white vertex v, draw a circle on the black vertices adjacent to v (here we assume that we use a representation of $L(\mathcal{C})$ in which no such triple of black vertices contains collinear points). We call this procedure a *V-construction* [9]. It provides a simple way to obtain a drawing of a point-circle configuration from its Levi graph. Moreover, given any other type of configuration in which the blocks are not circles, we obtain in this way its *point-circle representation*. For example, applying it to the Pappus graph, we obtain a point-circle representation of the Pappus configuration, as shown in Figure 4.

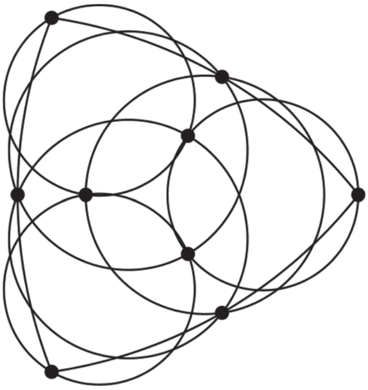


Figure 4 A point-circle representation of the Pappus configuration (circles of large size are shown only by their arcs).

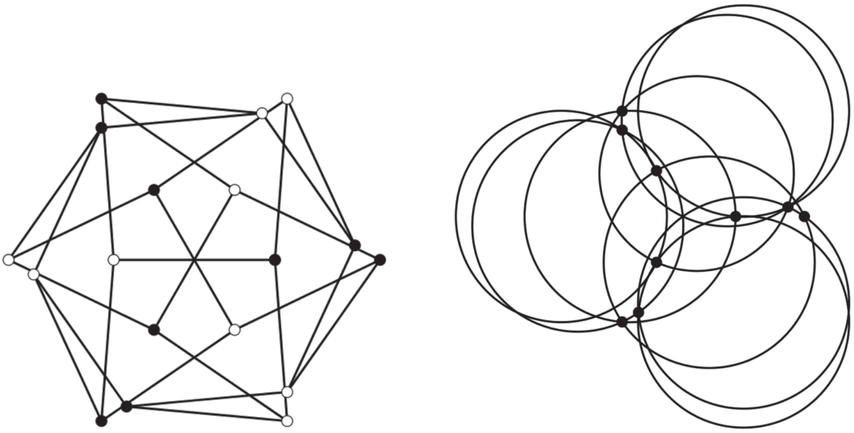


Figure 5 Gerbracht's representation of the Pappus graph (left), and the isometric point-circle Pappus configuration derived from it (right).

Assume now that the Levi graph of a configuration $\mathcal{C}$ can be drawn in such a way that all of its edges are of the same length. A graph with this property is called a *unit-distance graph* [9, 14]. It is clear that our V-construction applied to a unit-distance graph yields an isometric point-circle configuration. In this case, the centers of the circles coincide precisely with the white vertices of the Levi graph. Again taking the Pappus configuration for an example, the first unit-distance representation of its Levi graph is due to Gerbracht [8]. From this graph, we can easily derive the isometric point-circle representation of the Pappus configuration (see Figure 5).

Consider now the Levi graph of the Miquel configuration. It is isomorphic to the skeleton of a *rhombic dodecahedron*. Indeed, the Miquel configuration can be conceived of as a spatial configuration whose circles are circumscribed around the square faces of an ordinary cube. This can easily be seen by comparing the two parts of our Figure 1. The left image in Figure 6 shows how to derive a rhombic dodecahedron from an ordinary cube. The right image in Figure 6 shows that the skeleton of the rhombic dodecahedron is a bipartite graph which has eight black vertices of valence three, and six white vertices of valence four. (Thus, these vertices correspond to the points, respectively the circles, of the configuration).

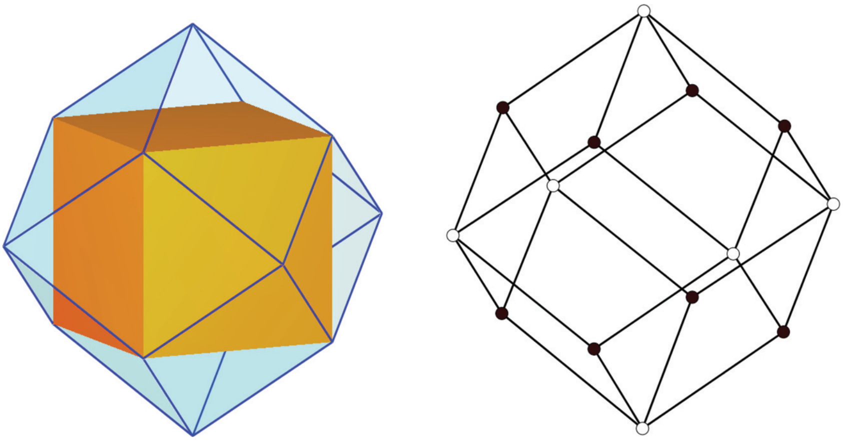


Figure 6 On the left, we construct the rhombic dodecahedron from an ordinary cube. On the right we see its skeleton, which is isomorphic to the Levi graph of the $(8_3, 6_4)$ Miquel configuration.

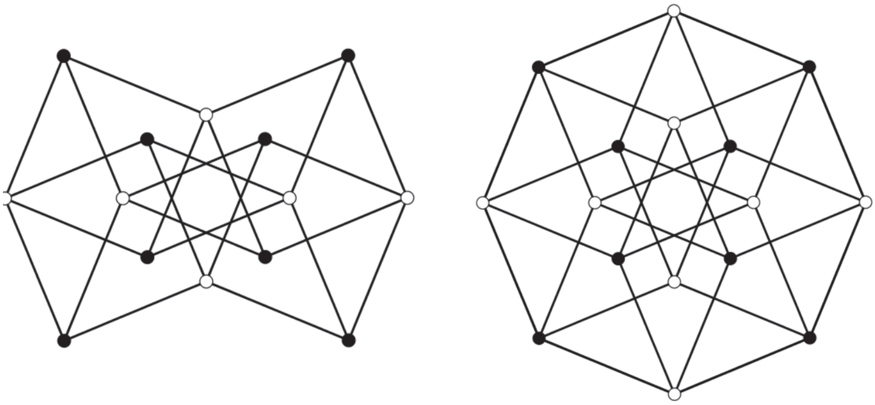


Figure 7 The skeleton of the rhombic dododecahedron (left) is a subgraph of the 4-cube graph (right).

Furthermore, the skeleton of the rhombic dodecahedron is a subgraph of the graph of a four-dimensional cube. (In what follows, we denote these graphs by RD and Q_4, respectively). Indeed, as it can be seen directly from Figure 7, RD can be obtained from Q_4 by removing two antipodal vertices together with the edges incident with them. (Here we use a regular octagonal representation of the four-cube graph, which is known, for example, from the paper by Coxeter cited above, see [3, Fig. 5]; see also the cover illustration of the journal *Discrete Applied Mathematics*).

We are now prepared to prove our main theorem.

Proof. Let K_2 denote the graph consisting of two vertices and a single edge. It is well known that the n-cube graph Q_n (i.e., the one-dimensional skeleton of an

n-dimensional cube) is the Cartesian product of n copies of K_2 (see Erdös, Harary, and Tutte [6], for example). It was shown by Horvat and Pisanski [14] that the Cartesian product of unit-distance graphs is a unit-distance graph. (Note that here we only make use of this result, hence we may omit the definition of the Cartesian product of graphs. For the definition, and other related details, see Horvat and Pisanski [14].) Since K_2 is itself a unit-distance graph, it follows that Q_n is as well. We have seen above that the graph RD is a subgraph of Q_4. Hence, RD is also a unit-distance graph. Thus, applying the V-construction on RD yields an isometric representation of the Miquel configuration.

If the radii of five of the circles are equal, then the position of all eight of the configuration points is already determined by intersections of these five circles. Moreover, the four unsaturated points are endpoints of radii starting from a common vertex. These edges are of unit length since they form the radii of four of the five unit circles. Hence, the points in question have a unit circumcircle, as desired (Figures 8 and 9). ■

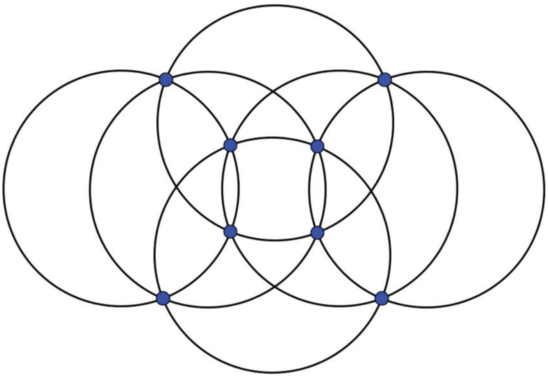


Figure 8 An isometric representation of the Miquel configuration obtained from the drawing of the rhombic dodecahedron graph on the left of Figure 7 by the V-construction.

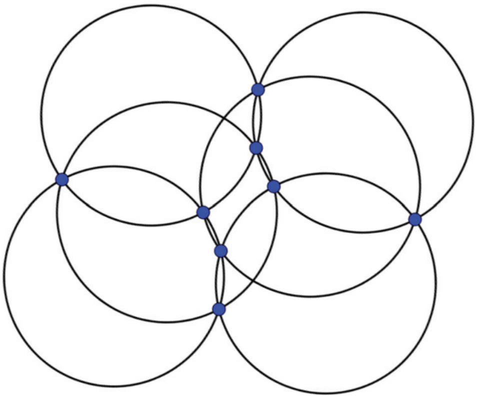


Figure 9 An isometric representation of the Miquel configuration, general setting.

To conclude, we mention that the V-construction can also be used to derive the *dual* of the original configuration. Recall that two configurations $\mathcal{C}$ and $\mathcal{C}^*$ are called dual of each other if there is a one-to-one correspondence between the points of $\mathcal{C}$ and the blocks of $\mathcal{C}^*$, and vice versa, such that incidences are preserved [10,17]. In our case the blocks are circles. Hence, the dual of the Miquel configuration is a configuration consisting of 6 points and 8 circles. Note that the duality relation in this case corresponds

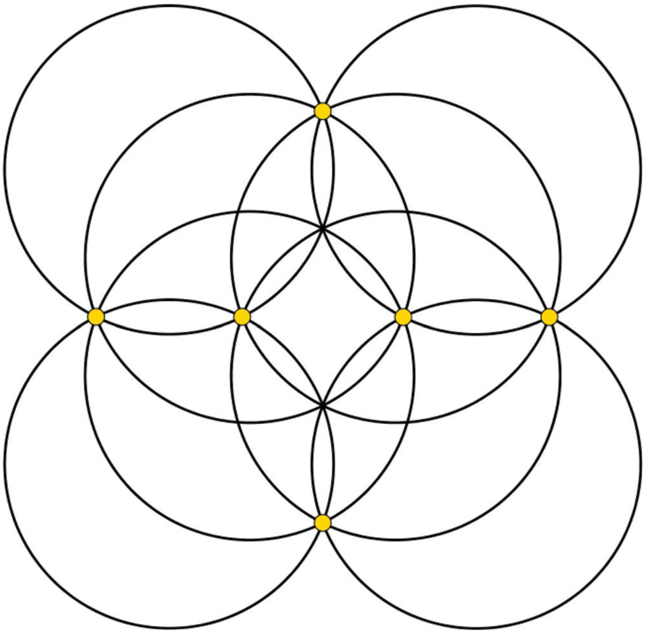


Figure 10 An isometric representation of the dual Miquel $(6_4, 8_3)$ configuration. The six configuration points carry light color. By adding the two "missing" points, one obtains an isometric Clifford (8_4) configuration depicted in Figure 6(b) of [9].

to duality between the cube and the octahedron: the circles of the Miquel configuration can be considered as the circumcircles of the faces of a cube, while the circles of the dual configuration can be considered as circumcircles of the faces of an octahedron. Considering the Levi graph, this means that the role of the black and white vertices is interchanged. Hence, applying the V-construction in the reverse way on RD, we obtain the dual configuration of type $(6_4, 8_3)$ such that it will also be isometric (see Figure 10). We note that this dual configuration occurs in a paper by Dorwart [5] (but with no regard to the size of the circles).

Acknowledgments The authors acknowledge support by the ARRS grant N1-0032, Slovenia and by the OTKA grant NN-114614, Hungary. The work of the second author was supported in part by the ARRS grants P1-0294 and J1-9187.

REFERENCES

[1] Coolidge, J. L. (1923). Some unsolved problems in solid geometry. *Amer. Math. Monthly*. 30(4): 174–180. doi.org/10.1080/00029890.1923.11986227

[2] Coxeter, H. S. M. (1948). *Regular Polytopes*. London: Methuen.

[3] ——— (1950). Self-dual configurations and regular graphs. *Bull. Amer. Math. Soc*. 56(5): 413–455. doi.org/10.1090/S0002-9904-1950-09407-5 Reprinted in: H. S. M. Coxeter, *Twelve Geometric Essays*, Carbondale: South. Ill. Univ. Press, Carbondale, 1968.

[4] Coxeter, H. S. M. (1961). *Introduction to Geometry*. New York: Wiley.

[5] Dorwart, H. R. (1988). Point and circle configurations: a new theorem. *Math. Magazine*. 61(4): 253–259. doi.org/10.2307/2689362

[6] Erdös, P., Harary, F., Tutte, W. T. (1965). On the dimension of a graph. *Mathematika*. 12(2): 118–122. doi.org/10.1112/S0025579300005222

[7] French, D. (2004). *Teaching and Learning Geometry*. London/New York: Continuum.

[8] Gerbracht, E. H. (2008). *On the Unit Distance Embeddability of Connected Cubic Symmetric Graphs*. *Kolloquium über Kombinatorik*. Magdeburg, Germany. Nov. 15, 2008.

[9] Gévay, G., Pisanski, T. (2014). Kronecker covers, V-construction, unit-distance graphs and isometric point-circle configurations. *Ars Math. Contemp*. 7(2): 317–336. doi.org/10.26493/1855-3974.359.8eb

[10] Grünbaum, B. (2009). *Configurations of Points and Lines*. Providence: Amer. Math. Soc.

[11] Hahn, L. (1994). *Complex Numbers and Geometry*, Washington D. C.: Math. Ass. Amer.

[12] Hammack, R., Imrich, W., Klavžar, S. (2011). *Handbook of Product Graphs*, 2nd Ed. Boca Raton: CRC Press.

[13] Hilbert, D., Cohn-Vossen, S. (1932). *Anschauliche Geometrie* Berlin: Springer. Hungarian translation: *Szemléletes geometria*. (1982). Budapest: Gondolat.

[14] Horvat, B., Pisanski, T. (2010). Products of unit distance graphs. *Discrete Math.* 310(12): 1783–1792. doi.org/10.1016/j.disc.2009.11.035

[15] Miquel, A. (1838). Théorèmes sur les intersections des cercles et des sphères. *Journal de mathématiques pures et appliquées* 1^{re} *série* 3: 517–522.

[16] Miquel, A. (1844). Mémoire de géometrie. *Journal de mathématiques pures et appliquées* 1^{re} *série*. 9: 20–27.

[17] Pisanski, T., Servatius, B. (2013). *Configurations from a Graphical Viewpoint*. Boston: Birkhäuser.

[18] Pritchard, C. (2003). *The Changing Shape of Geometry*. Cambridge: Cambridge Univ. Press.

[19] Richter-Gebert, J. (2011). *Perspectives on Projective Geometry*. New York: Springer.

[20] van der Waerden, B. L., Smid, J. L. (1935). Eine Axiomatik der Kreisgeometrie und der Laguerregeometrie, *Math. Ann*. 110(1): 753–776. doi.org/10.1007/BF01448057

[21] Ziegenbein, P. (1940). Konfigurationen in der Kreisgeometrie. *J. Reine Angew. Math.* 1941(183): 9–24. doi.org/10.1515/crll.1941.183.9

Summary. The classical six circle theorem of Miquel gives rise to a configuration consisting of eight points and six circles. We prove that this configuration can be realized by circles of equal size. Moreover, if five of the circles have unit radius, then the sixth circle must also have unit radius. In the proof, we use the Levi graph of this configuration, which is isomorphic to the skeleton of the rhombic dodecahedron.

GÁBOR GÉVAY (MR Author ID: 326668) received his Ph.D. in mathematics at Bolyai Institute, University of Szeged, Hungary, where he has been teaching since 1991. His main mathematical interests are in combinatorial, convex, and discrete geometry. His current interest in geometric configurations stems, among other things, from a visual appeal to these structures.

TOMAŽ (TOMO) PISANSKI (MR Author ID: 139995) is a senior Slovenian mathematician working mainly in discrete mathematics and graph theory. He is founding co-editor of two research journals: *Ars Mathematica Contemporanea* and *The Art of Discrete and Applied Mathematics*.

A Characterization of Antiderivatives

GEORGE STOICA
Saint John, New Brunswick, Canada
gstoica2015@gmail.com

Consider the following functions: $f, g : [-\pi/2, \pi/2] \to \mathbb{R}$ given by $f(x) = e^x$ and $g(x) = \cos x$. Obviously, an antiderivative of f on $[-\pi/2, \pi/2]$ is $F(x) = e^x$, and a simple integration by parts shows that

$$\int_{-\pi/2}^{\pi/2} \left[f(x)g(x) + F(x)g'(x) \right] dx = \int_{-\pi/2}^{\pi/2} \left(e^x \cos x - e^x \sin x \right) dx = 0.$$

We shall prove in the sequel that the above calculation remains true if we replace the function g by *any* continuous function with continuous derivative, null at the endpoints of its domain. This will lead us to an interesting characterization of antiderivatives for real valued functions that are continuous on a closed interval. Specifically, we have the following

Proposition. *Let $f, F : [a, b] \to \mathbb{R}$ be continuous functions. The following statements are equivalent:*

(i) If $g : [a, b] \to \mathbb{R}$ is a continuous function with a continuous derivative such that $g(a) = g(b) = 0$, then we have that

$$\int_a^b \left[f(x)g(x) + F(x)g'(x) \right] dx = 0.$$

(ii) F is differentiable and $F'(x) = f(x)$ for all $x \in [a, b]$.

Proof. For the implication $(ii) \Rightarrow (i)$ simply note that

$$\begin{aligned} \int_a^b \left[f(x)g(x) + F(x)g'(x) \right] dx &= \int_a^b \left[F'(x)g(x) + F(x)g'(x) \right] dx \\ &= \left[F(x)g(x) \right]_{x=a}^{x=b} = 0. \end{aligned}$$

For the implication $(i) \Rightarrow (ii)$, let us consider the function

$$A(x) := \int_a^x f(t)\, dt,$$

for all $x \in [a, b]$. This function is continuous and differentiable on $[a, b]$. Integrating by parts, we find that

$$\int_a^b A'(x)g(x)\, dx = \left[A(x)g(x) \right]_{x=a}^{x=b} - \int_a^b A(x)g'(x)\, dx = -\int_a^b A(x)g'(x)\, dx.$$

Thus, using that $A'(x) = f(x)$ on $[a, b]$, the condition in (i) can be re-written as

$$\int_a^b \left[-A(x) + F(x) \right] g'(x)\, dx = \int_a^b \left[f(x)g(x) + F(x)g'(x) \right] dx = 0.$$

Math. Mag. **95** (2022) 362–363. doi:10.1080/0025570X.2022.2094170 © Mathematical Association of America
MSC: 26A06

Let c be the constant defined by the condition

$$\int_a^b [-A(x) + F(x) - c]\ dx = 0,$$

and let

$$h(x) := \int_a^x [-A(t) + F(t) - c]\ dt,$$

for all $x \in [a, b]$, so that h is continuous with continuous derivative, and such that $h(a) = h(b) = 0$.

Then, on the one hand, using the condition in (i) with g replaced by h, we obtain that

$$\int_a^b [-A(x) + F(x) - c]\, h'(x)\, dx$$
$$= \int_a^b [-A(x) + F(x)]\, h'(x) dx - c[h(b) - h(a)] = 0,$$

while on the other hand

$$\int_a^b [-A(x) + F(x) - c]\, h'(x)\, dx = \int_a^b [-A(x) + F(x) - c]^2\ dx.$$

Comparing the last two equations, we conclude that $-A(x) + F(x) = c$ for all $x \in [a, b]$.

Since A is differentiable on $[a, b]$, the latter formula shows that F is differentiable on $[a, b]$. Taking derivatives in the same formula leads to: $-A'(x) + F'(x) = 0$, or $-f(x) + F'(x) = 0$ for all $x \in [a, b]$. ■

Remark. This result serves as a potential answer to a question going back to W.H. Young, formulated and partially treated in Freiling [3, p. 806]. It was inspired by, and complements, the study by Appell and Reinwand [1, 2]. It can also be used as a good classroom exercise. We let the readers check that one can replace "continuous derivative" in (i) by "integrable derivative."

REFERENCES

[1] Appell, J., Reinwand, S. (2018). Functions with antiderivative. I: Characterizations and properties. *Math. Semesterber.* 65(2): 195–210. doi.org/10.1007/s00591-017-0215-2

[2] Appell, J., Reinwand, S. (2019). Functions with antiderivative. II: Products and compositions. *Math. Semesterber.* 66(1): 49–72. doi.org/10.1007/s00591-017-0216-1

[3] Freiling, C. (1998). On the problem of characterizing derivatives. *Real Analysis Exchange.* 23(2): 805–812. doi.org/10.2307/44154004

Summary. We present an interesting characterization of antiderivatives for real-valued functions that are continuous on a closed interval.

GEORGE STOICA has a Ph.D. in Mathematics and another one in Statistics; completed a postdoc in Finance and another one in Biology, and had a 36-year career in academia, government, and the private sector, in Canada, the United States, France, and the United Kingdom. He has written four textbooks and more than 150 research papers. He has participated in some 100 projects in mathematics and statistics, both pure and with applications to: medicine, health, psychology, sociology, mechanics, astronomy, finance, economics, management and education. He has coordinated: two postdoc, six doctoral, and 14 Master's students. He speaks six languages (including Latin!) and has traveled (still does!) to 40 countries all around the world.

A Unified Treatment for the Inverses of M-Matrices and Scalars

AMIR RASTPOUR
Ontario Tech University
Oshawa, Ontario, Canada
amir.rastpour@ontariotechu.ca

JACOB BOURDEAU-MARCHE
Ontario Tech University
Oshawa, Ontario, Canada
jacob.bourdeaumarche@ontariotechu.net

Sometimes, when we are stuck in the complexities of matrix algebra, we wish we could deal with matrices as if they were scalars. For example, it would make our life easier if the inverse of the sum of two real matrices A and B could be expressed as a function of A and B, or if we could say $A \leq B$ implies $A^{-1} \geq B^{-1}$—the ordering symbols are understood to be element-wise. Although matrix algebra does not generally work this way, there are special cases that make these wishes come true. For example, if both A and $A + B$ are nonsingular and B has rank one, then $(A + B)^{-1}$ can be expressed as a function of A and B [5]. On the other hand, there is a class of matrices, known as M-matrices, for which the second wish comes true. That is, if A and B are nonsingular M-matrices, then $A \leq B$ implies $A^{-1} \geq B^{-1}$ [4].

We provide an alternative proof for this property of M-matrices using the representation of $(A + B)^{-1}$ in terms of A and B, and the fact that $A + X \geq A$, for any element-wise nonnegative matrix X. Our proof is more elementary than the original proof provided by Fiedler and Pták [4], which employs advanced theorems related to the eigenvalues of A and B.

Let A be a real $n \times n$ matrix with nonpositive off-diagonal elements. That is,

$$A = \begin{pmatrix} a_{1,1} & -a_{1,2} & \dots & -a_{1,n} \\ -a_{2,1} & a_{2,2} & \dots & -a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n,1} & -a_{n,2} & \dots & a_{n,n} \end{pmatrix},$$

where $a_{i,j} \geq 0, i \neq j, \forall i, j \in \{1, \dots, n\}$. A submatrix of A that is obtained by removing one or more of its rows and their corresponding columns is called a *principal submatrix* of A, and its determinant is called a *principal minor* of A. The matrix A is an M-matrix if all of its principal minors are nonnegative, and A is a nonsingular M-matrix if all of its principal minors are positive [2].

To the best of our knowledge, there has been no systematic survey of the history of M-matrices. However, Robert J. Plemmons, in a short note on the history of this matrix family published in 1976 [11], mentions that the name M-matrix was first used by Alexander Ostrowski [10] in 1937. Ostrowski was honoring Hermann Minkowski, whose work in the early 1900s proved that if all row sums of a real square matrix with nonpositive off-diagonal elements are positive, then the determinant of that matrix is positive [11]. Because of their structural properties, M-matrices have found use in a wide range of applications. Building upon the influential work of Alexander Ostrowski, the theory of M-matrices was developed by two streams of researchers: mathematicians and economists. The applications of M-matrices in mathematics and

Math. Mag. **95** (2022) 364–367. doi:10.1080/0025570X.2022.2094169 © Mathematical Association of America

economics included (but was not limited to) developing efficient methods to solve large sparse systems of linear equations and analyzing the equilibrium stability of economic systems, respectively. In 1962, Fiedler and Pták made one of the first attempts to collect various properties of nonsingular M-matrices from different fields of the science into one document [4].

More than a century after the publication of Hermann Minkowski's seminal work [6, 10], which formed the basis for the development of M-matrices, this matrix family is still widely used and is finding its way into new fields of science. For example, properties of the M-matrix are used to study a class of Markov chains named "quasi-birth-and-death (QBD) processes," which generalize the more familiar birth-and-death processes [9]. The infinitesimal generator of a QBD process takes the form of a block-tridiagonal matrix, with diagonal matrix blocks that are nonsingular M-matrices. Because of this special structure, the theory of M-matrices is extensively used to calculate and perform sensitivity analyses on the stationary distribution of QBDs; see Dayar [3] and Rastpour [12], for example.

Another recent field of application for M-matrices is at the intersection of two branches of mathematics, namely matrix theory and graph theory, where matrices are used to represent and analyze graphs. For example, a simple undirected graph can be represented by the Laplacian matrix, which is defined as the difference between the graph's degree matrix (a diagonal matrix with the number of edges connected to each vertex on the diagonals) and its adjacency matrix (a symmetric square matrix with 1s at coordinates that correspond to connected vertices and 0s elsewhere). For a connected graph, it turns out that the Laplacian matrix is a singular M-matrix, and all principal submatrices of the Laplacian matrix are nonsingular M-matrices [8]. The theory of M-matrices provide useful insights about the Laplacian matrix and the structure of the associated graph; see Barik and Pati [1] and Molitierno [8], for example.

Element-wise ordering of the inverse of two M-matrices

We use 0 and 1 to denote the all-zero and all-one matrices of appropriate size, respectively, and we use $\mathrm{tr}(X)$ to denote the sum of elements on the main diagonal of a square matrix X, which is known as the *trace* of X.

Berman and Plemmons thoroughly discuss the definition and properties of M-matrices. They list 50 necessary and sufficient conditions for the statement "A is a nonsingular M-matrix" [2, chapter 6]. We use two of these necessary and sufficient conditions:

Lemma 1. If A is a square matrix with nonpositive off-diagonal elements, then:

(a) A is a nonsingular M-matrix if and only if there exists a column vector $a > 0$ such that $Aa > 0$.
(b) A is a nonsingular M-matrix if and only if A is inverse-positive; that is, A^{-1} exists, $A^{-1} \geq 0$, and $A^{-1} \neq 0$.

Proof. This follows from properties $I27$ and $N38$ in [2, chapter 6]. ■

We also use a formula for the inverse of the sum of two matrices [5]:

Lemma 2. Let G and $G + E$ be nonsingular matrices where E is a matrix of rank 1. Let $g = \mathrm{tr}(EG^{-1})$. Then $g \neq -1$ and

$$(G + E)^{-1} = G^{-1} - \frac{1}{1+g} G^{-1} E G^{-1}.$$

Proof. This follows from Miller [5, p. 68]. ■

Theorem 1. *If A and B are nonsingular M-matrices and $A \leq B$, then $A^{-1} \geq B^{-1} \geq 0$.*

Proof. Lemma 1b implies that $A^{-1} \geq 0$ and $B^{-1} \geq 0$. It remains to prove that $A^{-1} \geq B^{-1}$. Define

$$Y = B - A \geq 0.$$

If $Y = 0$, then the theorem holds. We proceed by showing that the theorem holds if $Y \geq 0$ as well. Assume Y has r nonzero columns and denote them as y_i, for $i = 1, \dots, r$. Let $q(i)$ be the column index of y_i in Y and $d_{q(i)}$ be the column vector with a 1 in the $q(i)$th coordinate and 0s elsewhere. Define

$$E_i = y_i d_{q(i)}^T,$$

which is a rank-one matrix whose column $q(i)$ equals column $q(i)$ of Y, and whose other elements are 0. The definition of E_i implies that $E_i \geq 0$ and

$$Y = E_1 + \cdots + E_r.$$

Define

$$Z_i = Z_{i-1} + E_i,$$

for $i = 1, \dots, r$, with $Z_0 = A$. The definition of Z_i implies

$$Z_i - A = \sum_{j=1}^{i} E_j \geq 0 \qquad \text{and} \qquad Z_r = B.$$

We first show that Z_i is a nonsingular M-matrix, and then we obtain a formula to calculate Z_i^{-1} as a function of Z_{i-1}^{-1}.

By assumption, A is a nonsingular M-matrix. Therefore, following Lemma 1a, there exists a column vector $a > 0$ such that $Aa > 0$. We multiply both sides of the inequality $Z_i - A \geq 0$ by the positive vector a and obtain $Z_i a \geq Aa > 0$, which implies that Z_i is also a nonsingular M-matrix (by Lemma 1a).

Given that Z_i is a nonsingular M-matrix, and E_i is a rank-one nonnegative matrix for $i = 1, \dots, r$, we iteratively apply Lemma 2 to $Z_i = Z_{i-1} + E_i$ and calculate Z_i^{-1} as a function of Z_{i-1}^{-1}. That is,

$$Z_i^{-1} = (Z_{i-1} + E_i)^{-1} \tag{1}$$

$$= Z_{i-1}^{-1} - \frac{1}{1 + \operatorname{tr}\left(E_i Z_{i-1}^{-1}\right)} Z_{i-1}^{-1} E_i Z_{i-1}^{-1}, \quad i = 1, \dots, r. \tag{2}$$

Since $Z_i^{-1} \geq 0$ and $Z_{i-1}^{-1} \geq 0$ (by Lemma 1b), and $E_i \geq 0$ (by definition), the following inequalities also hold:

$$Z_{i-1}^{-1} E_i Z_{i-1}^{-1} \geq 0 \qquad \text{and} \qquad \left(1 + \operatorname{tr}\left(E_i Z_{i-1}^{-1}\right)\right) \geq 0.$$

These inequalities imply that equation (2) can be presented as

$$Z_i^{-1} = Z_{i-1}^{-1} - X,$$

where $X \geq 0$. Therefore,

$$Z_0^{-1} \geq Z_1^{-1} \geq \cdots \geq Z_r^{-1}.$$

Note that $Z_0 = A$ and $Z_r = B$ and the proof is complete. ■

REFERENCES

[1] Barik, S., Pati, S. (2005). On algebraic connectivity and spectral integral variations of graphs. *Linear Algebra Appl.* 397: 209–222. doi.org/10.1016/j.laa.2004.10.015

[2] Berman, A., Plemmons, R. J. (1994). *Nonnegative Matrices in the Mathematical Sciences*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

[3] Dayar, T., Sandmann, W., Spieler, D., Wolf, V. (2011). Infinite level-dependent QBD processes and matrix-analytic solutions for stochastic chemical kinetics. *Adv. Appl. Probab*. 43(4): 1005–1026. doi.org/10.1239/aap/1324045696

[4] Fiedler, M., Pták, V. (1962). On matrices with nonpositive off-diagonal elements and positive principal minors. *Czechoslovak Math. J.* 12: 382–400. doi.org/10.21136/CMJ.1962.100526

[5] Miller, K. S. (1981). On the inverse of the sum of matrices, *Math. Mag*. 54: 67–72. doi.org/10.1080/0025570X.1981.11976898

[6] Minkowski, H. (1900). Zur Theorie der Einheiten in den algebraischen Zahlkörpern. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*. 1: 90–93. Available at: eudml.org/doc/58467

[7] Minkowski, H. (1907). *Diophantische approximationen: eine einfü hrung in die zahlentheorie*. Leipzig: Teubner.

[8] Molitierno, J. J. (2012). *Applications of Combinatorial Matrix Theory to Laplacian Matrices of Graphs.* Boca Raton, FL: CRC Press.

[9] Neuts, M. F. (1981). *Matrix-geometric Solutions in Stochastic Models–an algorithmic approach.* Baltimore: Johns Hopkins University Press.

[10] Ostrowski, A. M. (1937). Über die Determinanten mit überwiegender Hauptdiagonale. *Comment. Math. Helvetici*. 10: 69–96. doi.org/10.1007/BF01214284

[11] Plemmons, R. J. (1976). *A Survey of M-Matrix Characterizations. I. Nonsingular M-Matrices.* No. MRC-TSR-1651. U. Wisconsin, Madison, Math. Res. Center.

[12] Rastpour, A., Ingolfsson, A., Sandıkçı B. (2020). *Algorithms for Queueing Systems with Reneging and Priorities Modeled as Quasi-Birth-Death Processes. Working paper*. Univ. Ontario Inst. Tech., Canada.

Summary. It is interesting to find conditions under which matrices follow the same algebraic rules as scalars. For example, conditions that guarantee if two nonsingular matrices A and B are ordered as $A \leq B$, then their inverses satisfy $A^{-1} \geq B^{-1}$, where the ordering is understood to be element-wise. In this article, we provide an elementary proof for such a theorem.

AMIR RASTPOUR is a professor of Operations Management in the Faculty of Business and Information Technology at the Ontario Tech University. He received his PhD from the University of Alberta School of Business in 2015.

JACOB BOURDEAU-MARCHE is a recent graduate of the Ontario Tech University's Commerce program with a major in finance.

Secrets of Linear Feedback Shift Registers

DAVID SINGER
Case Western Reserve University
Cleveland, OH 44106
david.singer@case.edu

Modern cryptography comes in two flavors: *Private key*, or classical, crypto uses complicated substitution and transposition techniques to obscure a message and relies on the receiver sharing a secret key with the sender. *Public key* cryptography, an essential tool in internet security, uses powerful mathematical ideas to allow secure communication between parties who do not have a shared key.

Courses in the mathematics of cryptography attract at least two different groups of students: mathematics majors, many with strong backgrounds in algebra or number theory, and engineering students, with strong backgrounds in computer science and computer engineering.

For public-key cryptography, an understanding of finite fields is essential, and mathematics majors are likely to have the necessary background in field theory and linear algebra.

Engineering students often learn about Linear Feedback Shift Registers (LFSRs), which are used in communications and in generating pseudorandom sequences; students may even know how to physically build them. LFSRs can also be used to create extremely efficient private-key cryptosystems, although in their straightforward implementation they are not cryptographically secure. The students are taught that so-called maximal-length LFSRs employ *primitive polynomials*, which can be found in look-up tables, but the students rarely know how a primitive polynomial works, or what happens if they are using one that is not primitive. The explanation for this requires an understanding of finite fields.

So, it is useful for both math majors and CS majors to learn the theory of finite fields and then apply this to the theory of LFSRs. The connection between the two topics is rather subtle. In fact, the question of what periodicity properties a not-necessarily-maximal n-bit LFSR may have does not seem to be addressed in the literature. In the last part of this article, we illustrate the solution in the case of $n = 6$, which is just large enough to make the result interesting, but small enough to allow for a complete solution.

Linear feedback shift registers

A Linear Feedback Shift Register (LFSR) is a device that can generate a long, seemingly random, sequence of ones and zeroes. They are used in computer simulations of random processes, error-correcting codes, and other engineering applications. The ease with which shift registers can produce such sequences make them an attractive topic in an introductory course in the mathematics of cryptography.

A first course in cryptography inevitably explores the notion of the *One-Time Pad*. This system, introduced by G. S. Vernam in 1917, is a "perfectly secure cryptosystem," that is, the cipher text does not leak any information about the message (See Beutelspacher [**1**, p. 53] or Trappe [**5**, p. 336]). It relies on generating long random sequences of letters or numbers.

Math. Mag. **95** (2022) 368–375. doi:10.1080/0025570X.2022.2094166 © Mathematical Association of America
MSC: 12E20, Secondary 11T71

Suppose the message M consists of a sequence $m_1 m_2 \dots m_\ell$ of ℓ letters taken from the usual 26-letter English alphabet. The venerable *Caesar cipher* works by shifting each letter of the alphabet by some fixed amount; the Vernam cipher shifts the letters by a different amount at each position in the message. To encrypt M, we generate a sequence $k_1 k_2 \dots k_\ell$ of ℓ random letters of the alphabet, each letter chosen randomly and independently with uniform probability 1/26. The ciphertext is then the sequence $c_1 c_2 \dots c_\ell$ with c_i determined by adding m_i to k_i, where we treat the letters as the integers from 0 to 25 mod 26.

Mod 26 arithmetic is somewhat inconvenient; mod 2 arithmetic is more natural in digital computers. So, we assume that a message has been encoded in some standard way as a string $m_1 m_2 \dots m_\ell$ of "bits", i.e., zeroes and ones. The key is then a random binary string $k_1 k_2 \dots k_\ell$ and we compute the ciphertext by $c_i = m_i \oplus k_i$, where the operation $\oplus$ is addition mod 2. To decrypt, we use the simple formula $m_i = c_i \oplus k_i$. To do this, of course, the recipient must have the key string. Since this string is completely arbitrary, it is theoretically impossible to recover the message without the key since every possible message of length ℓ can be encrypted to any ciphertext of length ℓ.

There are some major practical difficulties with this scheme. First, the recipient must have previously received the key, which is as large as the message! Second, the key must be chosen completely at random. To overcome these difficulties in practice, cryptographers try to come up with a device or algorithm for generating a long, seemingly random, binary string of bits using only a small random string S (called the "seed"). Then the sender and receiver only need to agree on the seed, which they can exchange using public-key cryptography.

By definition, this long string is *not* random since we generate it algorithmically, but perhaps it simulates a random string in the sense of being unpredictable for someone who does not possess the seed. Such a sequence is called "pseudorandom." This sequence should have statistical properties that true random sequences have. For example, 0 and 1 should appear with roughly the same frequency. Likewise, the four strings of length two should each appear with roughly the same frequency, and so on. This is necessary, but by no means sufficient, for a secure cryptosystem. We need a *semantically secure* algorithm, so an adversary can not recover partial information about a message in a reasonable amount of time. As you might guess, this last item is one of the most challenging problems of modern cryptography.

One simple and elegant (but definitely *not* cryptographically secure) algorithm, or *machine*, for generating a pseudorandom string is the LFSR, shown in Figure 1. It consists of n cells, each capable of storing one bit, either a 0 or 1. The device is controlled by a clock. At each time step it transfers the content of each cell to the next cell. The last cell outputs its bit to the stream. To get the new content of the leftmost cell, we feed back the mod 2 sum of the contents of certain specified cells. Mathematically, this is expressed using *connection coefficients* c_i, one for each cell, each of which is 0 or 1. When $c_i = 1$, the content of the ith cell is added in to the feedback. The machine is exactly determined by these coefficients, as in equation 1 We will always assume $c_n = 1$ since otherwise we would get the same output by deleting the last cell. For each choice of the connection coefficients we get a machine, and since there are 2^{n-1} possibilities (assuming $c_n = 1$), there are 2^{n-1} different LFSRs with n cells.

If $X_i(t)$ denotes the content of the ith cell at time step t, then the rules for the cells are

$$\begin{aligned} X_i(t+1) &= X_{i-1}(t) \qquad (2 \le i \le n) \\ X_1(t+1) &= c_1 X_1(t) \oplus c_2 X_2(t) \oplus \cdots \oplus c_n X_n(t). \end{aligned} \tag{1}$$

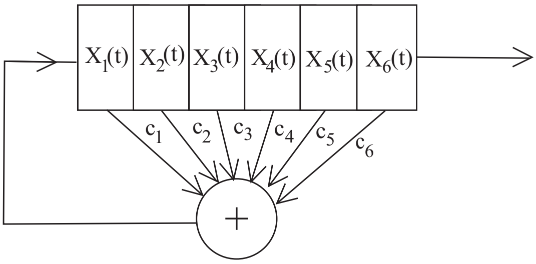


Figure 1 A 6-cell LFSR.

The state of the system at time t is given by the *column* vector

$$x(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \\ \dots \\ X_n(t) \end{pmatrix}.$$

The initial configuration is

$$x(0) = \begin{pmatrix} X_1(0) \\ X_2(0) \\ \dots \\ X_n(0) \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \dots \\ s_n \end{pmatrix} = s,$$

where the initial state of the system is given by the seed s.

Since there are only 2^n possible states for the machine, it is obvious that whatever initial state is specified, the machine must eventually repeat. It is not quite as obvious that it must return to its initial state, and if we did not assume $c_n = 1$ this would not be true! For instance, the initial state $(0, 0, \dots, 1)$ would drop into the zero state. We will soon see that if $c_n = 1$, then the machine will always return to its initial state. If the initial state is the zero vector, then the machine will remain in that state forever, so we exclude that from consideration. Consequently, the maximum number of steps before the machine returns to its initial state is $2^n - 1$. Given a seed s, the *period* of s is the number of steps it takes to return to s; the period is the smallest positive r such that $x(r) = s = x(0)$. The period of the machine is the maximum period achieved for any seed. If the period of s is $2^n - 1$, then the machine must visit every nonzero state, and so the period for any seed must be $2^n - 1$. Call such a machine a *maximal machine*. A fundamental fact in the theory of LFSRs is that for every n, a maximal machine exists. Since the goal is to achieve a long string from a small seed, this is the preferred result.

This is where the subject becomes mysterious. Of the 2^{n-1} possible machines, which choices correspond to ones that are maximal machines? What happens in the cases where the machine is not maximal?

Associate with the machine the *connection polynomial*

$$C(x) = x^n - c_1 x^{n-1} - \dots - c_{n-1}x - c_n,$$

whose coefficients are the connection coefficients. Typically, a cryptography text will say that the condition for the machine to be maximal is that the polynomial $C(x)$ is *primitive* (see, e.g., Welsh [7, p. 130], or Menezes [4, p. 197]). But what does that mean? More generally, what kind of periodicity can occur for an LFSR? This is the question we propose to explore.

As a teaser, consider the following question: which of the following integers *cannot* be the period of a 6-cell machine: 6, 7, 8, 9, 10, 11, 12, 21, 30, or 31? The full answer is at the end of the paper. As Beutelspacher wisely suggests, it is instructive "to construct—without any theory—your own shift registers of maximum period" [1, p. 59].

Using linear algebra

As the name suggests, a Linear Feedback Shift Register can be viewed through the lens of linear algebra. The relation between $x(t+1)$ and $x(t)$ is given by the equation $x(t+1) = Ax(t)$, where A is the matrix

$$A = \begin{pmatrix} c_1 & c_2 & c_3 & \cdots & c_{n-1} & c_n \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{2}$$

A is nonsingular exactly when $c_n \neq 0$. To see that, one can either observe that in that case the *rows* are obviously independent or note that the (mod 2) determinant is c_n. Note that we are doing linear algebra over $\mathbb{F}_p$, the field of integers modulo a prime number p, which is an important tool in cryptography. The matrix is invertible, which implies that the output is periodic for any initial seed. We have $x(t) = A^t x(0)$. The group $GL(n, \mathbb{F}_2)$ of invertible $n \times n$ matrices with entries in $\mathbb{F}_2$ has finite order, and therefore the matrix A has finite order k. That is, $A^k = I$, where I is the $n \times n$ identity matrix. Thus, the period of any seed must be a divisor of k. What are the possible values of k?

Using the idea that a matrix is invertible if and only if the columns (or rows) are linearly independent, we can count the number of invertible matrices. This leads to the following well-known result:

Proposition 1. *The number N of elements of $GL(n, \mathbb{F}_2)$ is given by the formula*

$$\begin{aligned} N &= (2^n - 1)(2^n - 2)(2^n - 2^2)\cdots(2^n - 2^{n-1}) \\ &= 2^{\frac{n^2-n}{2}}(2^n - 1)(2^{n-1} - 1)\cdots(2^2 - 1) \end{aligned}$$

Proof. To construct an invertible matrix, we must choose the n rows(or columns) to be linearly independent vectors. The top row can be anything except the zero vector. The second row can be any other nonzero row, so there are $2^n - 2$ choices. These two vectors span a two-dimensional subspace. The third row must avoid the four vectors in it, leaving $2^n - 4$ choices. Continuing along these lines, the $k + 1$st row will have $2^n - 2^k$ possible vectors available. ∎

Since the order of an element of a group divides the order of the group, k must be a divisor of this number. For example, if $n = 6$, then k must divide $N = 2^{15} \times 3^4 \times 5 \times 7^2 \times 31$. In particular $k \neq 11$, giving a partial answer to the teaser. Furthermore, k must be no larger than $2^n - 1$. We will demonstrate this here and again in the next section. The key fact we need is the following algebraic lemma (see Lidl [3, p. 77]).

Lemma 1. *Let $P(x) = a_0 + a_1 x + \cdots + x^n$ be a polynomial with coefficients in $\mathbb{F}_p$, p a prime, $a_0 \neq 0$. Then for some $k < p^n$, $P(x)$ is a factor of $x^k - 1$.*

Proof. The ring $R = \mathbb{F}_p[x]/(P(x))$ consists of congruence classes of polynomials with $\mathbb{F}_p$ coefficients, where two polynomials are in the same congruence class if they differ by a multiple of $P(x)$. Using division with remainder, we see that every polynomial is equivalent to exactly one polynomial of degree less than n, and therefore there are exactly $p^n - 1$ nonzero equivalence classes in R. Since there are p^n monomials in the set $\{x^i : 0 \leq i < p^n\}$ there must exist numbers $0 \leq i < j < p^n$ with $x^i \equiv x^j \mod P(x)$. Then $P(x)$ divides

$$x^j - x^i = x^i(x^{j-i} - 1).$$

Since $P(0) \neq 0$, $P(x)$ must divide $x^{j-i} - 1$. ■

Now, a basic fact from linear algebra is that every $n \times n$ matrix satisfies a polynomial relation of degree n, namely its characteristic equation. In the case of the matrix A, the polynomial is the *connection polynomial* $C(x) = x^n - c_1x^{n-1} - c_2x^{n-2} - \cdots - c_n$, so

$$C(A) = A^n - c_1A^{n-1} - c_2A^{n-2} - \cdots - c_nI = 0.$$

Any polynomial $P(x)$ having $C(x)$ as a factor must also satisfy $P(A) = 0$. By Lemma 1, $C(x)$ is a factor of $x^k - 1$ for some $k < 2^n$. So $A^k - I = 0$.

So far, we see that the order of A must divide the order of $GL(n, \mathbb{F}_2)$ and be no larger than $2^n - 1$. These conditions, while necessary, are not sufficient. For example, no matrix in $GL(6, \mathbb{F}_2)$ has order exactly 35 or 49. The precise set of conditions is quite complicated; see, for example, Golumb [2, pp. 41–43], for the detailed result. In the next section, we indicate a method for finding matrices of given orders.

We conclude this section with an important definition.

Definition. A polynomial $C(x)$ of degree n in $\mathbb{F}_2[x]$ is *primitive* if it is irreducible and divides $x^{2^n-1} - 1$ but does not divide $x^k - 1$ for $k < 2^n - 1$.

A basic result in the theory of polynomials is that primitive polynomials of degree n exist for all $n \geq 1$. This implies the existence of maximal machines; a machine is maximal if and only if the connection polynomial is primitive.

Using finite fields

The basic facts we need about finite fields are the following, all of which are elementary:

1. If p is a prime, then the field $\mathbb{F}_p$ of integers mod p is a finite field with p elements.
2. If $\mathbb{F}_q$ is a finite field with q elements, then the ring of polynomials $\mathbb{F}_q[x]$ has unique factorization and in fact has a Euclidean algorithm.
3. If $P(x)$ is an irreducible polynomial of degree n, then the quotient ring $R = \mathbb{F}_p[x]/(P(x))$ is a field with q^n elements. If $P(x)$ is not irreducible, then the resulting quotient ring has zero-divisors.
4. Every finite field has a subfield $\mathbb{F}_p$ of prime order p. It can be viewed as a vector space over $\mathbb{F}_p$, and therefore has order $q = p^k$ for some prime p and positive integer k.

A less elementary fact is that up to isomorphism there is exactly one field of order p^k. Such a field can be constructed by finding an irreducible polynomial of degree k over $\mathbb{F}_p$ and forming the quotient field. Since there are generally many such polynomials, the uniqueness is certainly not obvious. The standard proof, from the observation that $\mathbb{F}_q$ is the splitting field of $x^q - x$, is generally not accessible to students in a first cryptography course.

Suppose now that $P(x)$ is irreducible over $\mathbb{F}_p$, of degree n, and consider the field $\mathbb{F} = \mathbb{F}_p[x]/(P(x))$. We will always use the symbol α to denote the element of $\mathbb{F}$ representing the congruence class of x. So, in $\mathbb{F}$ we have $P(\alpha) = 0$. Another way of thinking about this is that $\mathbb{F}$ is obtained by enlarging the field $\mathbb{F}_p$ by throwing in a root of the polynomial $P(x)$.

An element of $\mathbb{F}$ is the congruence class of a polynomial, which can be uniquely taken to be of degree less than n. Therefore, the elements are uniquely expressible in the form

$$b = b_{n-1}\alpha^{n-1} + \cdots + b_2\alpha^2 + b_1\alpha + b_0 \qquad b_i \in \mathbb{F}_p$$

Now, represent b by the *row* vector $(b_{n-1} \ \ldots \ b_2 \ b_1 \ b_0)$. That is, $\mathbb{F}$ will be viewed as the n-dimensional vector space of row vectors over $\mathbb{F}_p$. Then multiplication by α is a linear map and can be represented by a matrix A; $b\alpha$ is represented by $(b_{n-1} \ \ldots b_2 \ b_1 \ b_0)A$. If the polynomial is the connection polynomial

$$C(x) = x^n - c_1x^{n-1} - \cdots - c_{n-1}x - c_n,$$

then the matrix is nothing more than the same matrix A defined earlier in equation 2!

It follows from this that the order of the matrix A is the order of the element α. It is a fact about finite fields that the multiplicative group of nonzero elements is a cyclic group. A generator of this group is called a primitive element. The polynomial $C(x)$ is primitive if and only if α is a primitive element of the field $\mathbb{F}$. The existence of primitive roots mod p is an important fact from number theory, and it comes up naturally in a cryptography course. In general α need not be a primitive element, although it is elementary (from group theory) that $\alpha^{p^n-1} = 1$ since the nonzero elements of a field form a group.

Remark. Once we have found a matrix A as in equation 2 of order $p^n - 1$, we can use it to give a lovely description of the finite field $\mathbb{F}_{p^n}$, namely as the powers of the matrix A together with the zero matrix. See Wardlaw [6] for a discussion of this point.

But suppose $C(x)$ is not irreducible. Then the quotient ring R is not a field, and the nonzero elements do not form a group. If we look instead at the *invertible* elements of R, they do form a group. This is the secret key that unlocks the mystery of the LFSR! Although the argument below generalizes, we will restrict to the case $p = 2$. In this case addition is the same as subtraction, so we do not have to worry about signs.

The matrix A represents multiplication by α in the ring R. Since $c_n = 1$, we can write

$$1 = (\alpha^{n-1} + c_1\alpha^{n-2} + \cdots + c_{n-1})\alpha$$

Therefore, α is invertible, and its order (which is the order of the matrix A) divides the number of invertible elements. This helps explain why the period of an LFSR may be a number not dividing $2^n - 1$. If the polynomial is irreducible, however, α must have order dividing $2^n - 1$.

To determine the possible periods of LFSRs, then, we need to consider the ways in which a polynomial of degree n can be constructed from irreducible polynomials. Here are the basic rules; details may be found in Lidl and Niederreiter [3, chapter 3].) Write $C(x) = g_1g_2 \ldots g_r$, where $g_1, \ldots, g_r$ are pairwise relatively prime. Call the *order* of a polynomial f the order of the element α; note that this is potentially confusing terminology! Then the order of f is the least common multiple of the orders of g_i. If $g_i = h_i^b$, with h_i irreducible, and the order of h_i is e, then the order of g_i is $2^t e$, where t is the smallest integer with $2^t \geq b$. Finally, if h has degree k, then its order is a divisor of $2^k - 1$.

Example: LFSRs with six cells

Suppose a polynomial $C(x)$ has degree 6. Then, since the sum of the degrees of the irreducible factors is 6, $C(x)$ determines a set D of positive integers adding up to 6. We use the notation m^i to represent a factor of degree m repeated i times. For example, $(4, 1^2)$ represents a product of a polynomial of degree 4 and the square of a linear polynomial. The order of an irreducible polynomial of degree 4 must divide $2^4 - 1 = 15$. In fact, such a polynomial will have order 15 or order 5. There is only one linear polynomial, and its square has order 2. Therefore, a polynomial with this pattern must have order 30 or 10. Since $2^2 - 1 = 3$, $2^3 - 1 = 7$, and $2^5 - 1 = 31$ are all primes, irreducible polynomials of degrees 2, 3, and 5 are automatically primitive. The order of a sixth degree polynomial is either 9, 21, or 63. Table 1 includes all possible cases. Each of the 32 possible polynomials fits one of these patterns.

TABLE 1: Representative connection polynomials.

D	$(c_1c_2c_3c_4c_5c_6)$	*Factors*	O
6	(000011)	$z^6 + z + 1$	63
6	(010111)	$z^6 + z^4 + z^2 + z + 1$	21
6	(001001)	$z^3 + z + 1$	9
5, 1	(101111)	$(z + 1)(z^5 + z^2 + 1)$	31
$4, 1^2$	(011111)	$(z^4 + z + 1)(z + 1)^2$	30
$4, 1^2$	(100011)	$(z^4 + z^3 + z^2 + z^1)(z + 1)^2$	10
4, 2	(111001)	$(z^4 + z + 1)(z^2 + z + 1)$	15
$3, 1^3$	(001011)	$(1 + z)^3(1 + z^2 + z^3)$	28
3, 2, 1	(010011)	$(1 + z)(1 + z + z^2)(1 + z + z^3)$	21
3^2	(010001)	$(1 + z^2 + z^3)^2$	14
3, 3	(111111)	$(z^3 + z + 1)(z^3 + z^2 + 1)$	7
2^3	(101011)	$(z^2 + z + 1)^3$	12
$2^2, 1^2$	(000001)	$(z + 1)^2(z^2 + z + 1)^2$	6
$2, 1^4$	(110111)	$(z + 1)^4(z^2 + z + 1)$	12
1^6	(010101)	$(z + 1)^6$	8

So the possible orders of the matrix are 6, 7, 8, 9, 10, 12, 14, 15, 21, 28, 30, 31, and 63. These numbers represent the *largest* periods of seeds in the corresponding machines. It is easy to show that the initial seed (0, 0, 0, 0, 0, 1) will always achieve the largest period. Of course, the seed (0, 0, 0, 0, 0, 0) always achieves the shortest period, namely 1. Other periods are also possible, although they must be divisors of the largest period. For example, for the machine with connection polynomial

$$z^6 + z^4 + z^3 + z^2 + z + 1,$$

depending on the initial nonzero seed, the period will be 30, 15, 2, or 1. So perhaps we have not yet uncovered all the secrets of the LFSR.

REFERENCES

[1] Beutelspacher, A. (1994). *Cryptology*. Translation by J. C. Fisher Washington D. C.: Math. Assoc. Amer.

[2] Golumb, S. (1967). *Shift Register Sequences*. San Francisco, CA: Holden-Day.

[3] Lidl, R., Niederreiter, H. (2002). *Introduction to Finite Fields and their Applications*. Cambridge: Cambridge Univ. Press.

[4] Menezes, A., van Oorschot, P., Vanstone, S. (1996). *Handbook of Applied Cryptography*. Boca Raton, FL: CRC Press. Available at http://cacr.uwaterloo.ca/hac/.

[5] Trappe, W., Washington, L. C. (2006). *Introduction to Cryptography with Coding Theory*, 2nd ed. Upper Saddle River, NJ: Pearson.

[6] Wardlaw, W. P. (1994). Matrix representation of finite fields. *Math. Magazine*. 67(4): 289–293. doi.org/10.1080/0025570X.1994.11996233

[7] Welsh, D. (1988). *Codes and Cryptography*. New York: Oxford Univ. Press.

Summary. A Linear Feedback Shift Register (LFSR) is a device that can generate a long, seemingly random, sequence of ones and zeroes. This is important in cryptography. We consider the sometimes unexpected periodic properties of LFSRs, how to understand them using linear algebra, and how to relate them to finite fields, another important topic in cryptography. Along the way, we resolve the puzzle of what it means for a polynomial to be primitive.

DAVID SINGER learned mathematics and how to teach from his mother and at the University of Pennsylvania, receiving his Ph.D. under Herman Gluck in 1970. After a National Science Foundation postdoctoral fellowship at Princeton and a stint teaching at Cornell University, he came to Case Western Reserve University in Cleveland in 1975. He has pursued research interests in differential geometry and Hamiltonian and other dynamical systems. He is grateful for receiving the MAA's Ohio Section Award for Distinguished College or University Teaching of Mathematics in 2005 and is very proud of his four wonderful grandchildren.

Factoring Subgroups and Factor Groups of the Unit Group Modulo *n*

JOSEPH A. GALLIAN
University of Minnesota, Duluth
Duluth, MN 55812
jgallian@d.umn.edu

SHAHRIYAR ROSHAN ZAMIR
University of Nebraska-Lincoln
Lincoln, NE 68588
sroshanzamir2@huskers.unl.edu

Early in a course on abstract algebra, one encounters the multiplicative group $U(n)$ of integers modulo n, consisting of the set of integers less than or equal to n and relatively prime to n. The order of these groups is $|U(n)| = \phi(n)$, where ϕ is the Euler phi function. This group was introduced by Euler in 1761 and investigated in detail by Gauss in 1801 in his famous book on number theory *Disquisitiones Arithmeticae*. Gauss elucidated its structure as a direct product of groups of the form Z_m, the group of integers modulo m under addition.

In his classic book on algebra *Lehrbuch der Algebra*, Heinrich Weber gave an extensive treatment of the groups $U(n)$ and described them as the most important examples of finite Abelian groups. One of their striking properties, proved later in this paper, is that every finite Abelian group is isomorphic to a subgroup of $U(n)$ for infinitely many n. The textbook by Gallian [4] uses the groups $U(n)$ and their subgroups to illustrate, in a concrete way, the concepts of cyclic and noncyclic groups, isomorphisms, homomorphisms, internal and external direct products, cosets, Lagrange's Theorem, factor groups, and the fundamental theorem of finite Abelian groups. These connections will be evident in this paper as well.

The groups $U(n)$ arise naturally in algebra, number theory, cryptography, and computer science. They have been studied in this MAGAZINE by Cheng [2], Devries [3], Gallian and Rusin [5], and Guichard [6]). Moreover, Allan, Dunne, Jack, Lynd, and Ellingsen [1] provide the classification of the group of units of the ring of Gaussian integers modulo n.

In Gallian [4] and Gallian and Rusin [5], it is shown how to express $U(n)$ and certain subgroups of $U(n)$ as a direct product of subgroups of $U(n)$ and as a direct product of groups of the form Z_m. We provide similar results about the structures of some subgroups and factor (quotient) groups of the groups $U(n)$.

Central to our discussion is the following theorem of Gauss:

$$U(p^n) \approx Z_{p^n - p^{n-1}} \text{ for an odd prime } p.$$

$$U(2^n) \approx Z_2 \oplus Z_{2^{n-2}} \text{ for } n \geq 3.$$

$$U(4) \approx Z_2 \text{ and } U(2) \approx U(1) \approx Z_1 \approx \{0\}.$$

Also recall the standard result that if $n_1, n_2, \dots n_r$ are pairwise relatively prime natural numbers with $n = n_1 n_2 \dots n_r$, then we have

$$U(n) \approx U(n_1) \oplus U(n_2) \oplus \cdots \oplus U(n_r).$$

Math. Mag. **95** (2022) 376–388. doi:10.1080/0025570X.2022.2094167 © Joseph A. Gallian and Shahriyar Roshan Zamir 2022. All Rights Reserved
MSC: 20E07

By combining these results, we can easily write every U-group as a direct product of groups of the form Z_m. For example,

$$U(1400) = U(2^3 \cdot 5^2 \cdot 7) \approx U(2^3) \oplus U(5^2) \oplus U(7) \approx Z_2 \oplus Z_2 \oplus Z_{20} \oplus Z_6.$$

This example raises the question of how can we do similar things for certain subgroups and factor groups of $U(n)$. In the next section we show how for $n \geq 1$ and an integer k, the subgroup of $U(n)$ defined by

$$U_k(n) = \{x \in U(n) \mid x \equiv kt + 1 \text{ for } t \in \mathbb{Z}\},$$

and the factor group $U(n)/U_k(n)$ can both be expressed as a direct product of groups of the form Z_m. In later sections, we do the same for two generalizations of $U_k(n)$ and for subgroups of $U(n)$ of the form

$$U(n)^{(k)} = \{x^k \mid x \in U(n)\}.$$

Results related to $U_k(n)$

In Gallian [4] and Gallian and Rusin [5], $U_k(n)$ is defined only for positive divisors k of n. Although our definition does not include that requirement, our first theorem shows that for questions about the structure of groups of the form $U_k(n)$, we may assume that k is a positive divisor of n.

Theorem 1. *Let n and k be positive integers. Then $U_k(n) = U_{\gcd(n,k)}(n)$.*

Proof. Let $\gcd(n, k) = d$, $k = dh$, and $x \in U_k(n)$. Then $x \equiv kt + 1 \pmod{n}$ implies $x \equiv d(ht) + 1 \pmod{n}$, which is in $U_d(n)$. For $x \in U_d(n)$ we have $x \equiv dt' + 1 \pmod{n}$. We know there exists integers s and t such that $sk + tn = d$. Hence, $x = (sk + tn)t' + 1 \equiv k(st') + 1 \pmod{n}$, and therefore $x \in U_k(n)$. ■

Corollary 1. *For any positive integer n and an arbitrary odd integer h, we have $U_{2h}(n) = U_h(n)$.*

Proof. If n is odd, then $\gcd(2h, n) = \gcd(h, n)$ and by Theorem 1 we get

$$U_{2h}(n) = U_{\gcd(2h,n)}(n) = U_{\gcd(h,n)}(n) = U_h(n).$$

Now suppose n is even. If $2h$ does not divide n, then again by Theorem 1 we get $U_{2h}(n) = U_{\gcd(2h,n)}(n)$. Since n is even, the greatest common divisor of $2h$ and n must equal $2h'$ for some odd h'. Hence, we may assume $2h$ divides n. It follows, by definition, that $U_{2h}(n) \subseteq U_h(n)$. Let $x \in U_h(n)$. Since h divides n, we have that $x = hk + 1$, where x is smaller than n. If k is odd, then x is even and hence not relatively prime to n, so k has to be even. Let $k = 2t$. Then $x = 2ht + 1$ and therefore $x \in U_{2h}(n)$. ■

Theorem 1 shows that any factor of k in $U_k(n)$ that is relatively prime to n can be "canceled." Corollary 1 shows that if k has exactly one factor of 2, then it can be "canceled" as well. For example:

$$U_{24}(30) = \{1, 19, 13, 7\} = \{1, 7, 13, 19\} = U_6(30) = U_3(30)$$

$$U_{15}(70) = \{1, 31, 61, 51, 11, 41\} = \{1, 11, 31, 41, 51, 61\} = U_5(70) = U_{10}(70)$$

The above examples illustrate the utility of using cancelation. For $U_{15}(70)$, we generate the set by starting with 1. We then successively add 15 to an element to get the next one. This results in terms that exceed the modulus and elements that are not in increasing order. In contrast, for $U_5(70)$ or $U_{10}(70)$ we generate the elements without using modular arithmetic, and the elements are in increasing order.

Noting that $U_5(70) = U_{10}(70)$ demonstrates the interesting fact that Corollary 1 is useful in opposite ways depending on the parity of n. When n is odd, canceling the 2 offers the same advantages as Theorem 1. When n is even, it is more efficient not to cancel the 2 because in examining the elements of the form $ht + 1$, we find that every other element is even and therefore is not in $U_h(n) = U_{2h}(n)$. So, one needs only to examine half as many integers for $U_{2h}(n)$. This observation is often overlooked by students.

We next classify the structure of subgroups of the form $U_k(n)$ and their respective factor groups when n is a power of a prime. After that, we shift our attention to the general case of arbitrary positive integers n and k. For the proof of Lemma 1 and Proposition 1, we only need to find the order of $U_k(n)$ and use the fact that every subgroup and every factor group of a cyclic group is cyclic.

Lemma 1. *For an odd prime p and $1 \leq k \leq m$, we have that $U_{p^k}(p^m) \approx Z_{p^{m-k}}$.*

Proof. Note that

$$U_{p^k}(p^m) = \{1, p^k + 1, 2p^k + 1, 3p^k + 1 \ldots, (p^{m-k} - 1)p^k + 1\}.$$

Since $U(p^m)$ is cyclic, the result follows. ■

Let us consider an example. For $p^m = 11^5$ and $p^k = 11^3$, Lemma 1 gives

$$U_{11^3}(11^5) \approx Z_{11^{5-3}} \approx Z_{121}.$$

Note that for $k = m$ we get the subgroup consisting of the identity only. That is,

$$U_{11^5}(11^5) \approx \{1\} \approx Z_1.$$

It is worth mentioning that for an odd prime p, Lemma 1 and the formula

$$|U(p^m)| = (p - 1)p^{m-1}$$

give us the attractive result that the Sylow p-subgroup of $U(p^m)$ is $U_p(p^m)$.

Proposition 1. *For an odd prime p and $1 \leq k \leq m$, we have*

$$\frac{U(p^m)}{U_{p^k}(p^m)} \approx Z_{p^{k-1}(p-1)}.$$

Proof. Since $U(p^m)$ is cyclic, we only need to find the order of $U(p^m)/U_{p^k}(p^m)$, which is

$$|U(p^m)/U_{p^k}(p^m)| = \frac{p^{m-1}(p-1)}{p^{m-k}} = p^{k-1}(p-1).$$

■

Suppose that in the previous example we had wanted to find the structure of $U(11^5)/U_{11^3}(11^5)$. By Proposition 1, we have

$$U(11^5)/U_{11^3}(11^5) \approx Z_{11^{3-1}(11-1)} \approx Z_{1210}.$$

Notice how much faster this was compared to having to do the calculations by hand. Also note that the structure of the factor group depends only on k.

Lemma 2. *Let $n \geq 1$ and $2 \leq i \leq n$. Then we have that $U_2(2^n) = U(2^n)$ and $U_{2^i}(2^n) \approx Z_{2^{n-i}}$.*

Proof. The first assertion follows by the definition. For the second part, observe that $|U_{2^i}(2^n)| = 2^{n-i}$ because

$$U_{2^i}(2^n) = \{1, 2^i + 1, \ldots, (2^{n-i} - 1)2^i + 1\}.$$

Since $U_{2^i}(2^n)$ is a subgroup of

$$U(2^n) \approx Z_2 \oplus Z_{2^{n-2}},$$

we know that $U_{2^i}(2^n)$ is isomorphic to either $Z_{2^{n-i}}$ or $Z_2 \oplus Z_{2^{n-i-1}}$, where $2 \leq i$. This implies that the subgroup $U_{2^i}(2^n)$ has either one or three elements of order 2, respectively. We will show it has one. Note that the group $U(2^n)$ has exactly three elements of order 2, namely $2^n - 1$ and $2^{n-1} \pm 1$. If $2^n - 1 \in U_{2^i}(2^n)$ then $2^n - 1 = k \cdot 2^i + 1$ for some integer k. This is a contradiction since the left-hand side is -1 mod 2^i, and the right-hand side is 1 mod 2^i. So $U_{2^i}(2^n)$ has only one element of order 2, and therefore is isomorphic to $Z_{2^{n-i}}$. ■

The following result about factor groups of finite Abelian groups will be helpful for our results about factor groups of U-groups.

Proposition 2. *Let*

$$G \approx Z_{p_1^{n_1}} \oplus \cdots \oplus Z_{p_k^{n_k}},$$

and let H be a subgroup of G such that

$$|H| = p_1^{n_1 - m_1} \ldots p_k^{n_k - m_k},$$

where p_i is prime and $0 \leq m_i \leq n_i$ for all i. Then

$$\frac{G}{H} \approx Z_{p_1^{m_1}} \oplus \cdots \oplus Z_{p_k^{m_k}}.$$

Porposition 2 follows from the fact that if G is k-generated, then G/H is generated by the canonical image of the generators of G. Thus, the number of components in a cyclic decomposition of a factor group is less than or equal to the number of components in the cyclic decomposition of the original group.

Proposition 3. *For $n = i$ we have $U(2^n)/U_{2^i}(2^n) \approx Z_1$. For $n = 2$ and $i = 1$ we have $U(4)/U_2(4) \approx Z_1$. For $2 \leq i < n$ we have*

$$\frac{U(2^n)}{U_{2^i}(2^n)} \approx Z_2 \oplus Z_{2^{i-2}}.$$

Proof. The first two assertions are obvious, so we assume that $2 \leq i < n$. Observe that

$$\left| \frac{U(2^n)}{U_{2^i}(2^n)} \right| = 2^{i-1}.$$

By Proposition 2, $U(2^n)/U_{2^i}(2^n)$ is isomorphic to either $Z_{2^{i-1}}$ or $Z_2 \oplus Z_{2^{i-2}}$ and therefore it has either one or three elements of order 2. We will show the latter is the case by exhibiting two elements of order 2. Let $H = U_{2^i}(2^n)$. Since

$$((2^n - 1)H)^2 = (-1H)^2 = H,$$

we know that $|(2^n - 1)H| = 1$ or 2. If $|(2^n - 1)H| = 1$, then $2^n - 1 \in H$, and therefore $2^n - 1 = 2^i \cdot k + 1$. But that is impossible since the left side is $-1 \pmod{2^i}$ and the right side is $1 \pmod{2^i}$. Similarly, we can show that $|(2^{n-1} - 1)H| = 2$. ■

Theorem 2. *Let $p_1, \dots, p_k$ be distinct primes. For $1 \le m_i$, $0 \le j_i \le m_i$, and $1 \le i \le k$, we have that*

$$U_{p_1^{j_1} \dots p_k^{j_k}}(p_1^{m_1} \dots p_k^{m_k}) \approx U_{p_1^{j_1}}(p_1^{m_1}) \oplus \cdots \oplus U_{p_k^{j_k}}(p_k^{m_k}).$$

Proof. We know that $U(p_1^{m_1} \dots p_k^{m_k})$ is isomorphic to

$$U(p_1^{m_1}) \oplus \cdots \oplus U(p_k^{m_k})$$

under the mapping

$$\gamma(x) = \left(x \pmod{p_1^{m_1}}, \dots, x \pmod{p_k^{m_k}}\right).$$

We will show the same mapping is the required isomorphism. For convenience, let

$$a = p_1^{m_1} \dots p_k^{m_k} \qquad \text{and} \qquad b = p_1^{j_1} \dots p_k^{j_k}.$$

If b is divisible by 2 but not 4, then by Corollary 1, we can ignore that factor of 2 in b. So, we may assume that if b is even, then b is divisible by 4. The restriction of the domain of γ from $U(a)$ to $U_b(a)$ is a well-defined, one-to-one, and operation preserving mapping from $U_b(a)$ to

$$U_{p_1^{j_1}}(p_1^{m_1}) \oplus \cdots \oplus U_{p_k^{j_k}}(p_k^{m_k})$$

because γ is an isomorphism. We need only show that this mapping is onto. Since γ is a one-to-one mapping, it suffices to show that $\gamma(U_b(a))$ is into

$$U_{p_1^{j_1}}(p_1^{m_1}) \oplus \cdots \oplus U_{p_k^{j_k}}(p_k^{m_k})$$

and that they have the same order. It follows from the definition and from Corollary 1, that the order of $U_b(a)$ is

$$\frac{a}{b} = p_1^{m_1 - j_1} \dots p_k^{m_k - j_k}.$$

(In the case of b even, the previous assumption of b being divisible by 4 was necessary for this and the next claim.)

By definition, we have:

$$|U_{p_1^{j_1}}(p_1^{m_1})| = p_1^{m_1 - j_1} \dots |U_{p_k^{j_k}}(p_k^{m_k})| = p_k^{m_k - j_k}.$$

Therefore, the order of

$$U_{p_1^{j_1}}(p_1^{m_1}) \oplus \cdots \oplus U_{p_k^{j_k}}(p_k^{m_k}) = p_1^{m_1 - j_1} \dots p_k^{m_k - j_k}.$$

To show the into part, let $x \in U_b(a)$. Note that $\gcd(p_i, x) = 1$ for all i since $\gcd(a, x) = 1$. Moreover,

$$\gamma(x) = (x \pmod{p_1^{m_1}}, \dots, x \pmod{p_k^{m_k}}).$$

To show $\gamma(x)$ is in

$$U_{p_1^{j_1}}(p_1^{m_1}) \oplus \cdots \oplus U_{p_k^{j_k}}(p_k^{m_k}),$$

we mod the ith component by $p_i^{j_i}$. This yields:

$$([x \pmod{p_1^{m_1}}] \pmod{p_1^{j_1}}, \ldots, [x \pmod{p_k^{m_k}}] \pmod{p_k^{j_k}})$$
$$= (x \pmod{p_1^{j_1}}), \ldots, x \pmod{p_k^{j_k}}) = (1, \ldots, 1)$$

since $x \equiv 1 (\bmod\ b))$. ■

The following corollaries are direct consequences of Theorem 2, Lemma 1, and Lemma 2.

Corollary 2. *Let k, k' be positive integers such that*

$$U_{\gcd(k,n)}(n) = U_{\gcd(k',n)}(n).$$

Then $\gcd(k, n) = \gcd(k', n)$.

Corollary 3. *If $|U_k(n)| = p^m$ for an odd prime p and $1 \leq m$ then $U_k(n) \approx Z_{p^m}$.*

Corollary 4. *Let $p_1, \ldots, p_k$ be distinct odd primes. For $1 \leq j_i \leq m_i$ and $1 \leq i \leq k$, we have*

$$U_{p_1^{j_1} \ldots p_k^{j_k}}(p_1^{m_1} \ldots p_k^{m_k}) \approx Z_{p_1^{m_1-j_1} \ldots p_k^{m_k-j_k}}$$

and

$$U_{p_1^{j_1} \ldots p_k^{j_k}}(p_1^{m_1} \ldots p_k^{m_k}) = \langle p_1^{j_1} \ldots p_k^{j_k} + 1 \rangle.$$

Proof. The first part follows directly from Theorem 2 and Lemma 1. To see that $p_1^{j_1} \ldots p_k^{j_k} + 1$ is a generator for

$$U_{p_1^{j_1} \ldots p_k^{j_k}}(p_1^{m_1} \ldots p_k^{m_k}),$$

observe that the isomorphism from

$$U_{p_1^{j_1} \ldots p_k^{j_k}}(p_1^{m_1} \ldots p_k^{m_k}) \quad \text{to} \quad Z_{p_1^{m_1-j_1} \ldots p_k^{m_k-j_k}}$$

is given by

$$\gamma(x) = x \pmod{p_1^{j_1} \ldots p_k^{j_k}}$$

maps $p_1^{j_1} \ldots p_k^{j_k} + 1$ to a generator of $Z_{p_1^{m_1-j_1} \ldots p_k^{m_k-j_k}}$. ■

Corollary 5. *For $1 \leq k \leq n$, we have $U_k(n) = U(n)$ if and only if $\gcd(n, k) = 1$ or 2.*

Proof. If $\gcd(k, n) = 1$, then by Theorem 1 we get

$$U_k(n) = U_{\gcd(k,n)}(n) = U_1(n) = U(n).$$

If $\gcd(k, n) = 2$, then it follows from Theorem 1 and Corollary 1 that

$$U_k(n) = U_{\gcd(k,n)}(n) = U_2(n) = U(n).$$

Now suppose $U_k(n) = U(n)$. From Theorem 1 and Corollary 1 we get:

$$U_k(n) = U_{\gcd(k,n)}(n) = U(n) = U_2(n) = U_{\gcd(2,n)}(n).$$

It follows from Corollary 2 that $\gcd(k, n) = \gcd(2, n)$. This implies that $\gcd(k, n) = 1$ or 2. ■

Example 1. To demonstrate how we can use Theorem 2, suppose we want to express $U_{140}(1800)$ as a direct product of groups of the form Z_m. We know

$$U_{140}(1800) = U_{20}(1800) = U_{20}(8 \cdot 9 \cdot 25),$$

and by Theorem 2, Lemma 1, and Lemma 2, we get

$$U_{20}(1800) \approx U_4(8) \oplus U(9) \oplus U_5(25) \approx Z_2 \oplus Z_6 \oplus Z_5.$$

Theorem 3. *Let $n > 1$ be odd and k a divisor of n. Then $U(n)/U_k(n) \approx U(k)$.*

Proof. Let $n = p_1^{n_1} \dots p_j^{n_j}$ and $k = p_1^{m_1} \dots p_j^{m_j}$. Consider the homomorphism $\gamma : U(n) \to U(k)$ given by $\gamma(x) = x \pmod{k})$. By definition, $\operatorname{Ker}(\gamma) = U_k(n)$. Therefore, the first group isomorphism theorem gives us

$$\frac{U(n)}{U_k(n)} \approx \gamma(U(n)).$$

Moreover, $\gamma(U(n))$ is a subgroup of $U(k)$. We will show $|\gamma(U(n))| = |U(k)|$. We know that

$$|U(k)| = \phi(k) = p_1^{m_1-1}(p_1 - 1) \dots p_j^{m_j-1}(p_j - 1).$$

By Theorem 2 and Lemma 1, we get

$$|\gamma(U(n))| = |\frac{U(n)}{U_k(n)}| = \frac{|U(n)|}{|U_k(n)|} = \frac{p_1^{n_1-1}(p_1 - 1) \dots p_j^{n_j-1}(p_j - 1)}{p^{n_1-m_1} \dots p^{n_j-m_j}}$$
$$= p_1^{m_1-1}(p_1 - 1) \dots p_j^{m_j-1}(p_j - 1).$$

■

Theorem 4. *If n is even and k is divisible by 4, then $U(n)/U_k(n) \approx U(k)$.*

Proof. The argument is identical to the proof of Theorem 3. To find the order of $U(n)/U_k(n)$, we use Theorem 2, Lemma 1, and Lemma 2. ■

Theorem 5. *If n is even and $k = 2h$, with h odd, then $U(n)/U_k(n) \approx U(h)$.*

Proof. We know from Corollary 1 that $U_k(n) = U_h(n)$. We change the mapping in Theorem 3 to $\gamma : U(n) \to U(h)$, where $\gamma(x) = x \pmod{h})$. The rest of the proof is an order argument identical to the one in the proof of Theorem 3. ■

Generalizations to $U_{\pm k}(n)$ and $U_{k,H}(n)$

Does every subgroup of $U(n)$ have the form $U_k(n)$, where k is a divisor of n? The answer is no. For example, $U(36)$, which is isomorphic to $Z_2 \oplus Z_6$, has a subgroup isomorphic to $Z_2 \oplus Z_2$. But looking at cases reveals that for no divisor k of 36 do we get $U_k(36) \approx Z_2 \oplus Z_2$. This motivates our next theorem. It will allow us to give a description of the elements of $U(36)$ that form the subgroup isomorphic to $Z_2 \oplus Z_2$.

Theorem 6. *For $n \geq 1$ and a positive integer k, the set*

$$U_{\pm k}(n) = \{x \in U(n) \mid x \equiv kt \pm 1 \pmod{n} \text{ for } t \in \mathbb{Z}\}$$

is a subgroup of $U(n)$.

Proof. It suffices to show that $U_{\pm k}(n)$ is closed (see Gallian [4, Theorem 3.3]). If $a, b \in U_{\pm k}(n)$, then

$$ab \pmod n \equiv (a \pmod n)(b \pmod n) \equiv (\pm 1)(\pm 1) \equiv \pm 1.$$

■

Here are a few examples of $U_{\pm k}(n)$:

$$\begin{aligned} U_{\pm 9}(36) &= \{1, 17, 19, 35\} \\ U_{\pm 11}(33) &= \{1, 10, 23, 32\} \\ U_{\pm 5}(45) &= \{1, 4, 11, 14, 16, 19, 26, 29, 31, 34, 41, 44\}. \end{aligned}$$

The first example answers our question about a noncyclic subgroup of order four in $U(36)$ since

$$U_{\pm 9}(36) \approx Z_2 \oplus Z_2.$$

As was the case with $U_k(n)$, we do not need to know all of the elements of $U(n)$ to find the elements of $U_{\pm k}(n)$. The algorithm is similar. Add ± 1 to all nonnegative integer multiples of k, mod by n, and check to see if the result is relatively prime to n. Continue in this fashion until you reach 1. For example, $(1 \cdot 9) \pm 1 \pmod{36}$ are not relatively prime to 36 so we discard them, and $(4 \cdot 9) \pm 1 \equiv 1, 35 \pmod{36}$, so we are done.

Theorem 7. *Let $n = st$ with $n \geq 3$ and $\gcd(s, t) = 1$. Then*

$$U_{\pm s}(n) \approx U_s(n) \times \{1, n-1\} \approx U(t) \oplus Z_2.$$

Proof. We know that if $G = H \times K$, the internal direct product of H and K, then $G = H \oplus K$. By inspection, $U_{\pm s}(n) = U_s(n) \times \{1, -1\}$, where $-1 \equiv n - 1 \pmod n$. (A detailed and more general proof of why the two subgroups $U_{\pm s}(n)$ and $U_s(n) \times \{1, -1\}$ are equal is given in Theorem 8.) Since $\gcd(s, t) = 1$, it follows that

$$U_s(n) = U_s(st) \approx U(t),$$

where the last isomorphism is a result from Gallian and Rusin [5]. (See also Gallian [4, p. 160].) Moreover, $\{1, -1\} \approx Z_2$. Therefore $U_{\pm s}(n) \approx U(t) \oplus Z_2$. ■

One might wonder if $U_{\pm k}(n) = U_k(n) \times \{1, -1\}$ for all $1 \leq k \leq n$. The answer is yes in all non-trivial cases. In Corollary 5, we proved that $U_k(n) = U(n)$ if and only if $\gcd(k, n) = 1$ or 2. So, we now ignore this case.

Theorem 8. *For $1 \leq k \leq n$, and $U_k(n) \neq U(n)$, we have*

$$U_{\pm k}(n) = U_k(n) \times \{1, -1\} \approx U_{\gcd(n,k)}(n) \oplus Z_2.$$

Proof. Suppose $U_k(n) \neq U(n)$. It suffices to show

$$U_{\pm k}(n) = U_k(n) \times \{1, -1\}.$$

(The rest follows from Theorem 1.) Let $A = U_{\pm k}(n)$, and let $B = U_k(n) \times \{1, -1\}$. The assumption that $U_k(n) \neq U(n)$ allows for set B to exist. Otherwise, the notion of internal direct product would not make sense. Because both A and B are subgroups of $U(n)$, it suffices to show A and B are subsets of each other. For $x \in A$, if $x = kt + 1$, then we are done. If $x = kt - 1$ then $x = -(k(-t) + 1)$, which is an element of B.

Now let $x \in B$. If $x = (kt + 1)(1)$, then we are done. If $x = (kt + 1)(-1)$ then $x = k(-t) - 1$, which is an element of A. Finally, we have

$$U_{\pm k}(n) = U_k(n) \times \{1, -1\} \approx U_{\gcd(n,k)}(n) \oplus Z_2.$$

■

The following example demonstrates how the above results, taken together, easily dispatch problems that appear to be intimidating.

Example 2. Consider the case

$$n = 2^3 \cdot 3^3 \cdot 5^2 \cdot 11 = 59400 \qquad \text{and} \qquad k = 2^2 \cdot 3 \cdot 5 \cdot 13 = 780.$$

We will find the structure of $U_{\pm 780}(59400)$. Note that the fact that $\gcd(59400, 780) = 60$, implies

$$\begin{aligned} U_{\pm 780}(59400) &\approx U_{780}(59400) \oplus Z_2 \approx U_{60}(59400) \oplus Z_2 \\ &\approx U_{4\cdot 3\cdot 5}(8 \cdot 27 \cdot 25 \cdot 11) \oplus Z_2 \\ &\approx U_4(8) \oplus U_3(27) \oplus U_5(25) \oplus U(11) \oplus Z_2 \\ &\approx Z_2 \oplus Z_9 \oplus Z_5 \oplus Z_{10} \oplus Z_2 \approx Z_{45} \oplus Z_{10} \oplus Z_2 \oplus Z_2. \end{aligned}$$

We finish this section with a generalization of $U_{\pm k}(n)$. The following are alternate definitions for the subgroups $U_k(n)$ and $U_{\pm k}(n)$:

$$\begin{aligned} U_k(n) &= \{x \in U(n) \mid x \pmod{k} \in \{1\}\} \\ U_{\pm k}(n) &= \{x \in U(n) \mid x \pmod{k} \in \{1, -1\}\}. \end{aligned}$$

We generalize these by replacing $\{1\}$ or $\{1, -1\}$ with any other subgroup H of $U(n)$.

Theorem 9. *For $n > 1$, let k be a positive divisor of n and H be a subgroup of $U(n)$. The set $U_{k,H}(n) = \{x \in U(n) \mid |x \pmod{k} \in H\}$ is a subgroup of $U(n)$.*

Proof. The proof follows from the closure of H. ■

The advantage of using these subgroups is that by picking certain positive divisors k of n and a subgroup H of $U(n)$, we are able to construct a new subgroup of $U(n)$ by changing the divisor k or the subgroup H (or both).

Example 3. Let $n = 80$, $k = 10$ and $H = \{1, 9\}$. Then we have

$$U_{10,\{1,9\}}(80) = \{x \in U(80) \mid x = 10t + 1 \text{ or } x = 10t + 9,\ t \in \mathbb{Z}\}.$$

For $t = 0$, we get H. For $t = 1$, we get 11 and 19. For $t = 2$ we get 21 and 29, and so on. Notice that we need only check up to $t = 8$. Finally, we have

$$U_{10,\{1,9\}}(80) = \{1, 9, 11, 19, 21, 29, 31, 39, 41, 49, 51, 59, 61, 69, 71, 79\},$$

which is indeed a subgroup of $U(80)$.

Our results about when $U_{\pm k}(n) = U_k(n) \times \{1, -1\}$ raise the question of when $U_{k,H}(n) = U_k(n) \times H$.

Theorem 10. *Let $n > 1$, k a positive divisor of n, and H a subgroup of $U(n)$. Then $U_{k,H}(n) = U_k(n) \times H$ if and only if $U_k(n) \cap H = \{1\}$.*

Proof. Suppose $U_k(n) \cap H = \{1\}$. It suffices to show $U_{k,H}(n) = U_k(n)H$ since the rest follows from the definition of internal direct product. For $x \in U_k(n)H$, we have that for some $h \in H$,

$$x \equiv (kt+1)h \pmod{n} \equiv k(th) + h \pmod{n}),$$

which is an element of $U_{k,H}(n)$. For $x \in U_{k,H}(n)$ we have $x = kt + h$ for some $h \in H$. The following chain of equalities shows $x \in U_k(n)H$:

$$x = (kt+h)1 = (kt+h)h^{-1}h = (k(th^{-1})+1)h.$$

Thus x has the desired form. If $U_{k,H}(n) = U_k(n) \times H$, then by the definition of internal direct product, we get $U_k(n) \cap H = \{1\}$. ■

Results about $U(n)^{(k)}$ and a general result about $U(n)$ groups

We now ask the following question: Is every subgroup of $U(n)$ expressible in the form $U_{\pm k}(n)$ or $U_k(n)$? The answer is again no For instance, for

$$U(252) \approx Z_2 \oplus Z_6 \oplus Z_6,$$

there is no divisor k of 252 such that $U_k(252)$ or $U_{\pm k}(252)$ yields the subgroup of $U(252)$ isomorphic to $Z_2 \oplus Z_2 \oplus Z_2$. This question motivates another way of producing subgroups in a $U(n)$ group.

Definition 1. *Let $n > 1$, and let k be any integer. We define*

$$U(n)^{(k)} = \{x^k \mid x \in U(n)\}.$$

That $U(n)^{(k)}$ is a subgroup of $U(n)$ follows from the closure of $U(n)^{(k)}$. If k does not divide $|U(n)| = \phi(n)$, then this subgroup can be viewed as the image of the automorphism given by $\gamma(x) = x^k$. If k is a divisor of $\phi(n)$, then γ defines a homomorphism from $U(n)$ to itself with kernel:

$$\text{Ker}(\gamma) = \{x \in U(n) \mid x^k = e\}.$$

Consequently, by the first isomorphism theorem for groups, we have

$$\frac{U(n)}{\text{Ker}(\gamma)} \approx U(n)^{(k)}.$$

Example 4. Consider $U(13) = \{1, 5, 7, 12\}$ and $k = 2$. Then squaring each element gives us $\{1, 12, 12, 1\}$, implying that $U(13)^{(2)} = \{1, 12\}$.

Our next result is the counterpart of $U_k(n) = U_{\gcd(n,k)}(n)$.

Proposition 4. *For $n > 1$ and any integer k,*

$$U(n)^{(k)} = U(n)^{(\gcd(\phi(n),k)}.$$

Proof. Let $\gcd(\phi(n), k) = d$, and let $k = dh$. Since $U(n)^{(d)}$ and $U(n)^{(k)}$ are both subgroups of $U(n)$, we only need to show they are subsets of each other. Clearly $U(n)^{(k)} \subseteq U(n)^{(d)}$ since $x^{hd} = (x^h)^d$. Now let $x^d \in U(n)^{(d)}$. We know $d = t_1k + t_2\phi(n)$, which implies

$$x^d = x^{t_1k+t_2\phi(n)} = x^{t_1k} \cdot x^{t_2\phi(n)} = (x^{t_1})^k \in U(n)^{(k)}.$$

■

Proposition 4 allows us to assume the superscript k is always a divisor of $\phi(n)$.

Example 5. Consider

$$U(252) = U(4 \cdot 9 \cdot 7) \approx U(4) \oplus U(9) \oplus U(7) \approx Z_2 \oplus Z_6 \oplus Z_6.$$

Direct calculations show that

$$U(252)^{(3)} = \{1, 55, 71, 125, 127, 181, 197, 251\}$$

and every nonidentity element has order two. Thus, we have

$$U(252)^{(3)} \approx Z_2 \oplus Z_2 \oplus Z_2.$$

Notice that in the previous example, raising the elements of $U(252)$ to the third power is equivalent to multiplying the elements of $Z_2 \oplus Z_6 \oplus Z_6$ by 3. So, in order to find the structure of the latter, all we have to do is trace the generator of each component, namely 1, after being multiplied by 3. In the first component, which is Z_2, 3 (mod 2) is 1, and therefore we get Z_2. In the next two Z_6 components, 1 goes to 3 which yields a Z_2. To summarize, finding the structure of $U(n)^{(k)}$ is equivalent to tracing 1 in each term in the cyclic group decomposition of $U(n)$. This is the main idea of Theorem 11.

Theorem 11. *Let $n = p_1^{m_1} \dots p_j^{m_j}$ for distinct odd primes p_i and positive integers m_i. Then*

$$U(p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx Z_{d_1} \oplus \cdots \oplus Z_{d_j}$$

where

$$d_i = \frac{\phi(p_i^{m_i})}{\gcd(\phi(p_i^{m_i}), k)}$$

for all $1 \leq i \leq j$.

Proof. We know that

$$U(n) \approx Z_{\phi(p_1^{m_1})} \oplus \cdots \oplus Z_{\phi(p_j^{m_j})}.$$

Raising every element in $U(n)$ to the kth power is equivalent to multiplying all the elements of

$$Z_{\phi(p_1^{m_1})} \oplus \cdots \oplus Z_{\phi(p_j^{m_j})}$$

by k. This is a mapping of cyclic groups to themselves. So, one needs only to trace where the generator, 1, of each cyclic component is mapped. Observe that 1 goes to k for each term. Hence $Z_{\phi(p_i^{m_i})}$ is mapped to Z_{d_i} where $d_i = \phi(p_i^{m_i})/\gcd(\phi(p_i^{m_i}), k)$. ■

Corollary 6. *Let $n = 2^b p_1^{m_1} \dots p_j^{m_j}$ for distinct odd primes p_i and positive integers b and m_i for all i. Define $d_i = \phi(p_i^{m_i})/\gcd(\phi(p_i^{m_i}), k)$ for all $1 \leq i \leq j$. Then*

1. $U(2 \cdot p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx U(p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx Z_{d_1} \oplus \cdots \oplus Z_{d_j}$.
2. $U(2^b p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx Z_2 \oplus Z_{d_1} \oplus \cdots \oplus Z_{d_j}$ *if $b = 2$ and k is odd.*
3. $U(2^b p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx Z_{d_1} \oplus \cdots \oplus Z_{d_j}$ *if $b = 2$ and k is even.*
4. $U(2^b p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx Z_2 \oplus Z_{2^{b-2}} \oplus Z_{d_1} \oplus \cdots \oplus Z_{d_j}$ *if $b \geq 3$ and k is odd.*
5. $U(2^b p_1^{m_1} \dots p_j^{m_j})^{(k)} \approx Z_{\frac{2^{b-2}}{\gcd(2^{b-2},k)}} \oplus Z_{d_1} \oplus \cdots \oplus Z_{d_j}$ *if $b \geq 3$ and k is even.*

Proof. For $b = 1$ observe that

$$U(2 \cdot p_1^{m_1} \dots p_j^{m_j}) \approx U(p_1^{m_1} \dots p_j^{m_j}).$$

If $b = 2$, then we have

$$U(n) \approx Z_2 \oplus Z_{\phi(p_1^{m_1})} \oplus \cdots \oplus Z_{\phi(p_j^{m_j})}.$$

If k is odd, then the additional Z_2 term does not change and the rest is identical to Theorem 11. If k is even, the first Z_2 is gone because we are mapping 1 to $k \pmod 2$ which yields zero. For $b \geq 3$ we get

$$U(n) \approx Z_2 \oplus Z_{2^{b-2}} \oplus Z_{\phi(p_1^{m_1})} \oplus \cdots \oplus Z_{\phi(p_j^{m_j})}.$$

For odd k, the term $Z_2 \oplus Z_{2^{b-2}}$ stays the same. For even k, the first Z_2 is gone. We need to find the order of $1 \cdot k = k$ in the $Z_{2^{b-2}}$ term to find the structure of the first component of the direct product. But that order is exactly $2^{b-2}/\gcd(2^{b-2}, k)$. Since every subgroup of a cyclic group is cyclic, the result follows. ■

The previous theorem and its corollary help us find the explicit group elements of various subgroups with desired structures, including p-Sylow subgroups.

Example 6. Let $n = 3^3 \cdot 7 \cdot 19$. The cyclic group decomposition of $U(n)$ is $Z_6 \oplus Z_{18} \oplus Z_{18}$. By Theorem 11 we have

$$U(n)^{(9)} \approx Z_{\frac{6}{\gcd(6,9)}} \oplus Z_{\frac{18}{\gcd(18,9)}} \oplus Z_{\frac{18}{\gcd(18,9)}} \approx Z_2 \oplus Z_2 \oplus Z_2.$$

Therefore, after raising every element of $U(n)$ to the 9th power, the elements that are left form the Sylow 2-subgroup of $U(n)$. Define $\gamma : U(n) \to U(n)$ by $\gamma(x) = x^9$. We claim Ker(γ) is the Sylow 3-subgroup of $U(n)$. By the first isomorphism theorem, we observe that

$$U(n)/\text{Ker}(\gamma) \approx (U(n))^{(9)} \approx Z_2 \oplus Z_2 \oplus Z_2,$$

which implies that Ker(γ) is the set of all elements of $U(n)$ whose orders divide 9 and is isomorphic to $Z_3 \oplus Z_9 \oplus Z_9$, which is the structure of the 3-Sylow subgroup of $U(n)$. Hence, Ker$(\gamma) \approx Z_3 \oplus Z_9 \oplus Z_9$. The group $U(n)^{(2)}$ is another way to obtain the Sylow 3-subgroup of $U(n)$. Observe, by Theorem 11, that

$$U(n)^{(2)} \approx Z_{\frac{6}{\gcd(6,2)}} \oplus Z_{\frac{18}{\gcd(18,2)}} \oplus Z_{\frac{18}{\gcd(18,2)}} \approx Z_3 \oplus Z_9 \oplus Z_9.$$

Example 7. Suppose in the previous example we wanted to produce the elements of $U(n)$ that form a subgroup isomorphic to $Z_6 \oplus Z_2$. To this end let

$$H = U_{7 \cdot 19}(3^3 \cdot 7 \cdot 19)^{(9)} \qquad \text{and} \qquad K = U_{3^3 \cdot 19}(3^3 \cdot 7 \cdot 19).$$

Using Theorem 2 it is clear that $U_{7 \cdot 19}(3^3 \cdot 7 \cdot 19) \approx Z_{18}$ and by Theorem 11 we have $H = U_{7 \cdot 19}(3^3 \cdot 7 \cdot 19)^{(9)} \approx Z_2$. Noting that $K \approx Z_6$, we let $L = H \times K$. Since $H \cap K = \{1\}$, we have

$$L \approx H \oplus K \approx Z_2 \oplus Z_6.$$

For completeness, we conclude this paper by proving that every finite Abelian group is a subgroup of a U-group, thereby offering support for Weber's assertion in the introduction that the U-groups are the most important examples of finite Abelian groups. We know of no proof of the fact that does not use number theory in an essential way.

Indeed, we will use Dirichlet's theorem, also called Dirichlet's prime number theorem, which states that for any two relatively prime integers a and b, there are infinitely many primes of the form $q = an + b$, where n is a nonnegative integer. (See Shanks [7].)

Theorem 12. *Every finite Abelian group is isomorphic to a subgroup of a U-group.*

Proof. Let G be a finite Abelian group. By the fundamental theorem of finite Abelian groups, we have that

$$G \approx Z_{p_1^{a_1}} \oplus \cdots \oplus Z_{p_1^{a_i}} \oplus \cdots \oplus Z_{p_s^{r_1}} \oplus \cdots \oplus Z_{p_s^{r_h}},$$

where the p_i's are distinct primes, and we have arranged the subscripts such that a_1 is the largest exponent of p_1, and r_1 is the largest exponent of p_s. Let $a = p_1^{a_1}$ and $b = 1$ in the statement of Dirichlet's theorem. Then there are infinitely many primes of the form $q = p_1^{a_1}n + 1$, which implies that $p_1^{a_1}$ divides $\phi(q)$. Therefore, $U(q)$ has a subgroup of order $p_1^{a_1}$. We can find i distinct primes, $q_1, \ldots, q_i$ of the form $p_1^{a_1}n + 1$, each of which will have a subgroup of order $p_1^{a_1}$. Since that was the largest power of p_1, we can get every subgroup of a smaller power of p_1. Repeating this process for each prime up to p_s and multiplying all these primes, we obtain the desired n. ■

Acknowledgments We would like to thank the referees and the Editor of *Mathematics Magazine* for their constructive comments and suggestions.

REFERENCES

[1] Allan, A. A., Dunne, M. J., Jack, J. R., Lynd, J. C., Ellingsen Jr, H. W. (2008). Classification of the group of units in Gaussian integers modulo n. *Pi Mu Eps. J.* 12(9): 513–519.

[2] Cheng, Y. (1989). Decompositions of U-groups. *Math. Mag.* 62(4): 271–273. doi.org/10.1080/0025570X.1989.11977454

[3] Devries, D. J. (1989). The group of units in Z_m. *Math. Mag.* 62(5): 340–342. doi.org/10.1080/0025570X.1989.11977467

[4] Gallian, J. A. (2017). *Contemporary Abstract Algebra*, 9th Ed. Boston: Cengage.

[5] Gallian, J. A., Rusin, D. (1980). Factoring groups of integers modulo n. *Math. Mag.*. 53(1): 33–36. doi.org/10.1080/0025570X.1980.11976823

[6] Guichard, D. R. (1999). When is $U(n)$ cyclic? An algebraic approach. *Math. Mag.* 72(2): 139–142. doi.org/10.1080/0025570X.1999.11996716

[7] Shanks, D. (1978) *Solved and Unsolved Problems in Number Theory*, 2nd ed. New York: Chelsea.

Summary. We introduce several ways of producing subgroups of $U(n)$, the group of units of Z_n. Our results find the structure of these various subgroups in terms of external direct product of cyclic groups. We then use our classifications to give descriptions of elements of $U(n)$ that form a subgroup of $U(n)$ with a desired cyclic group decomposition. This includes a description of elements that form a Sylow p-subgroup.

JOE GALLIAN received a PhD from Notre Dame in 1971. He has had the good fortune of being at the University of Minnesota, Duluth since 1972. He is proud to be a joint author with his student and friend Shah. He enjoys spending time nearly everyday with his ten year old grandson Joey.

SHAHRIYAR ROSHAN ZAMIR obtained his B.A. in Pure Mathematics from Georgia Gwinnett College and his M.S. in Mathematics from the University of Minnesota, Duluth, where he had the honor of having Joe Gallian as his Master's thesis adviser. He considers it a great privilege to call Joe his friend. He is currently a Ph.D student at the University of Nebraska-Lincoln.

To Fix Those Plots, Use Limits!

YVES NIEVERGELT
Eastern Washington University
Cheney, WA 99004
ynievergelt@ewu.edu

A study of electrons in electromagnetic fields involves a dimensionless function f defined by a formula recently published in *Science* [7, p. 190, equation (3)]:

$$f(z) = \frac{2}{z^2}\left[\frac{1}{\left(1-z^2\right)^{-1/2}-1}+\delta\right]^{-1}. \tag{1}$$

Figure 1 shows my plots of f for $\delta = 0$ and z near 0. Our boss says: fix those plots!

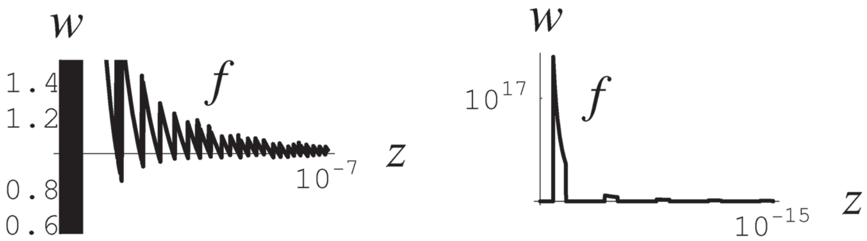


Figure 1 Plots of f for $\delta = 0$ with *Mathematica*™ 5 on an iMac12,1 with Intel Core i7 under OS 10.7.2, with $1 - z^2$ replaced by $(1 - z)(1 + z)$ in Equation (1).

How would *you* improve those plots? Techniques for calculating limits may help.

In his article "Where are Limits Needed in Calculus?", R. Michael Range notes that students raised on graphing calculators sometimes find it difficult to grasp the need for limits [11]. A need for limits may arise in computing arc lengths, or in solving isoperimetric problems and differential equations, as pointed out several decades earlier in this MAGAZINE by Judith V. Grabiner [1]. However, introducing limits in any one such context would involve introducing at least two new concepts at once, such as limits and arc lengths.

In contrast, this article demonstrates by examples how the intermediate logical, algebraic, and analytical steps used in the calculation of limits assist in computing or plotting functions, including algebraic functions, with simple graphing calculators or fancy professional computing systems. Specifically, where hardware and software produce erroneous graphs off the mark by several orders of magnitude, steps in the derivation of limits can be used to lead calculators and computers to correct sketches of curves. The examples given here could be used in courses ranging from algebra to multivariable calculus. They become more interesting in classes where students use different computing and plotting hardware, software, and options within software, as well as "equivalent" algebraic formulae, which can produce different graphs for the same function.

Math. Mag. **95** (2022) 389–396. doi:10.1080/0025570X.2022.2094180 © Mathematical Association of America
MSC: 26-04

Sketching functions of one variable using algebra

Our first example demonstrates how to use algebra to improve the plots in Figure 1.

Example 1. As defined, the function f may be difficult to plot near the origin because equation (1) can exacerbate rounding errors in calculators and computers. Indeed, if $\delta = 0$, then equation (1) with $z = 0$ gives the indeterminate form $0/0$, where different rounding errors on the numerator and denominator lead to the wild plots in Figure 1. Nevertheless, the same ideas used in calculus to find limits apply here, such as combining fractions and multiplying numerators and denominators by a conjugate quantity. In other words, one way to fix Figure 1 consists of "simplifying" equation (1) to discover a function g that is continuous at the origin and such that $f(z) = (z^2/z^2) \cdot g(z)$ for $0 < |z| < 1$. The difficulty is not in doing the algebra, but in deciding what steps to ask a human or symbolic manipulation software to do.

For instance, finding a common denominator and multiplying by a conjugate quantity leads to

$$\frac{1}{\sqrt{1-z^2}} - 1 = \frac{1-\sqrt{1-z^2}}{\sqrt{1-z^2}} \cdot \frac{1+\sqrt{1-z^2}}{1+\sqrt{1-z^2}} = \frac{z^2}{\sqrt{1-z^2}+1-z^2}.$$

Taking reciprocals, adding δ, and reducing the sum to a common denominator gives

$$\frac{1}{\left(1-z^2\right)^{-1/2}-1} + \delta = \frac{\sqrt{1-z^2}+1-z^2}{z^2} + \delta = \frac{\sqrt{1-z^2}+1-(1-\delta)z^2}{z^2}.$$

Multiplying the reciprocal by $2/z^2$ yields a relation with another formula, which we denote by g:

$$f(z) = \frac{2}{\sqrt{1-z^2}+1-(1-\delta)z^2} = g(z). \tag{2}$$

For $0 \leq \delta < 1$, if $0 < z^2 < 1$, and if z^2 increases, then $1-z^2$, $\sqrt{1-z^2} = \sqrt{(1-z)(1+z)}$, and $1-(1-\delta)z^2$ all decrease, which implies that $f(z)$ increases. Also,

$$f(z) = g(z) = \frac{2}{\sqrt{1-z^2}+1-(1-\delta)z^2} > \frac{2}{1+1} = 1, \tag{3}$$

in contrast to the left-hand panel in Figure 1. Moreover, if $\delta = 0$ and $0 < z < 0.0001$, then $0.999\,999\,990 < 1-z^2 < 1$ and $0.999\,999\,995 < \sqrt{1-z^2} < 1$, implying that

$$1 < f(z) = \frac{2}{\sqrt{1-z^2}+1-z^2} \tag{4}$$

$$< \frac{2}{1.999\,999\,985} = 1.000\,000\,007\,500\ldots \tag{5}$$

Thus, for $\delta = 0$ and $0 < z < 10^{-4}$, on a plot that also shows the origin, the increase in $f(z)$ is imperceptible. Consequently, the graph of f must appear visually as a horizontal line at height 1, as correctly shown in the plot of g from equation (2) in Figure 2.*

To show how various algebraically equivalent formulae and various versions of the same brands of hardware and software produce incorrect plots of f for $0 < z < 10^{-7}$, Figure 2 also shows a plot of f from equation (1), while the left-hand panel in Figure 1 shows a plot of f with $1-z^2$ replaced by $(1-z)(1+z)$ in equation (1).

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

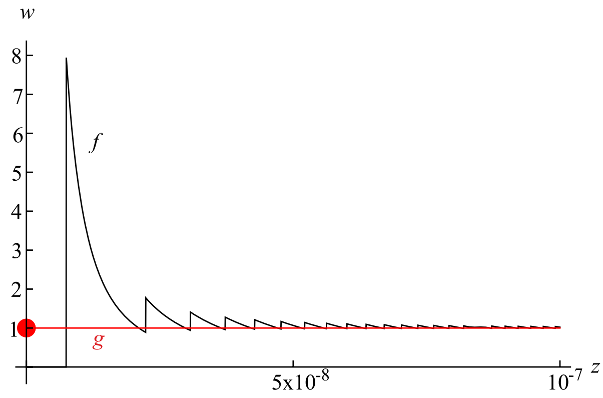


Figure 2 For $\delta = 0$, plots of f from equation (1), and g with $g(0)$ from equation (2), by *Mathematica* 10 on an iMac17,1 with Intel Core i5 under OS 10.11.6.

Notice that the activity that really assists in plotting the function f is the determination of a function g that is continuous at 0 and such that $f(z) = g(z)$ for z away from 0. Thus, this example nicely demonstrates why one might need to apply the techniques used to compute a limit by hand in an alternative setting, without necessarily using limits.

The question arises, why use equation (1), which is less accurate and requires more arithmetic operations than equation (2)? Equation (1) was chosen, among formulae with the same values away from the origin, to single out the role of the parameter δ: In private correspondence, Frank H. L. Koppens, one of the authors of the *Science* article we cited, explained that they chose the most elegant notation for their function, and also wanted it to be clear that δ was an additional, small term.

The lesson is that communication and computation may require different formulae.

Computing the harmonic mean of two numbers

We now show that professional software can deliver erroneous values of the harmonic mean outside the range of the data, and we show how to use the algebraic derivation of the limit to guide the same software toward accurate values of the harmonic mean. The *harmonic mean*, H, of two positive numbers x and y, is defined by

$$H(x, y) = \frac{2}{\frac{1}{x} + \frac{1}{y}}. \tag{6}$$

Equation (6) is undefined if $x = 0$ or $y = 0$, even though these values can arise in practice. Indeed, in physics it is common to use the function

$$R(x, y) = \frac{1}{\frac{1}{x} + \frac{1}{y}}, \tag{7}$$

which is half of the harmonic mean. This formula represents the so-called *reduced mass* of two positive masses x and y, as well as the *resistance* of two parallel resistors with positive resistances x and y. (See Kittel, Knight, and Ruderman [6, p. 290], and Purcell [10, p. 132]). If either $x = 0$ or $y = 0$, which happens if either resistor shorts out, then the resulting resistance is exactly zero, but then equation (7) is undefined. To get a valid formula for this case, denote the larger and smaller of x and y by

$\min(x, y)$ and $\max(x, y)$, respectively, and multiply the top and bottom of equation (7) by $\min(x, y)$, to get

$$R(x, y) = \frac{1}{\frac{1}{x} + \frac{1}{y}} = \frac{\min(x, y)}{\frac{\min(x,y)}{x} + \frac{\min(x,y)}{y}} = \frac{\min(x, y)}{1 + \frac{\min(x,y)}{\max(x,y)}}. \tag{8}$$

Note that if $\min(x, y) = x$, then $\max(x, y) = y$. Hence,

$$\frac{\min(x, y)}{x} = 1 \qquad \text{and} \qquad \frac{\min(x, y)}{y} = \frac{\min(x, y)}{\max(x, y)},$$

with a comparable result if $\min(x, y) = y$. Therefore, $0 \le \min(x, y)/\max(x, y) \le 1$, implying that $1 \le 1 + (\min(x, y)/\max(x, y)) \le 2$.

Equation (8) yields

$$\frac{\min(x, y)}{1 + 1} \le R(x, y) = \frac{\min(x, y)}{1 + \frac{\min(x,y)}{\max(x,y)}} \le \frac{\min(x, y)}{1 + 0} = \min(x, y). \tag{9}$$

The lower bound in (9) will be invoked in Example 3. The upper bound (9) reveals that the reduced mass or equivalent resistance does not exceed the smaller of the two masses or resistances. Would the resistance vanish if both parallel resistors short out? Intuition based on experience with circuits suggests that if both short out, then their equivalent resistance is also zero. Does algebra support this intuition? Equation (9) is valid for all positive values of x and y. It extends to nonnegative values of x and y as follows: If $x = y > 0$, then equation (9) gives $R(x, x) = x/2 = \min(x, y)/2$. If either $x = 0$ or $y = 0$, but not both, then it gives $R(x, x) = 0 = \min(x, y)/2$. However, if $x = y = 0$, then $x/2 = \min(x, y)/2 = 0$, too. Thus,

$$R(x, y) = \begin{cases} x/2 & \text{if } x = y \ge 0 \\[2ex] \dfrac{\min(x, y)}{1 + \frac{\min(x,y)}{\max(x,y)}} & \text{if } \max(x, y) > \min(x, y) \ge 0, \end{cases} \tag{10}$$

yields the correct resistance for *all* nonnegative x and y, including $R(0, 0) = 0$.

Returning to the harmonic mean, multiplying inequalities (9) by 2 gives the bounds $\min(x, y) \le H(x, y) \le 2\min(x, y)$. Equation (6) then leads to the bounds

$$\min(x, y) \le H(x, y) = \frac{2}{\frac{1}{\min(x,y)} + \frac{1}{\max(x,y)}} \le \max(x, y). \tag{11}$$

Inequalities (11) are useful for work with calculators and computer hardware that come "out of the box" wired to compute only with a finite set of numbers, for instance, in scientific notation, with a fixed number of digits and a bounded range of exponents. Thus, if x and y are two positive numbers in your computing system, then their harmonic mean, $H(x, y)$, can neither exceed the largest, nor fall below the smallest, positive number in your computing system. Yet equation (11) cannot be used to compute $H(x, y)$ because the reciprocal of the smallest number $1/\min(x, y)$ can exceed the largest positive number, or the reciprocal of the largest number, $1/\max(x, y)$, can fall below the smallest positive number. Instead, compute $H(x, y)$ with equation (12):

$$H(x, y) = \begin{cases} x & \text{if } x = y \ge 0 \\[2ex] \dfrac{\min(x, y)}{\frac{1}{2}\left[1 + \frac{\min(x,y)}{\max(x,y)}\right]} & \text{if } \max(x, y) > \min(x, y) \ge 0. \end{cases} \tag{12}$$

To avoid falling out of range, instead of doubling equation (10), use equation (12). For instance, if x is the smallest positive number in your computing system, and y is the next larger one, then $R(x, y) < x$, and equation (10) may round $R(x, y)$ to zero, whereas equation (12) can round $H(x, y)$ to $x = \min(x, y)$. Example 2 shows how equation (12) beats professional computing systems that give results out of range.

Example 2. Set x and y to the largest floating-point number on your machine; the specifics depend on your system. For instance, *Mathematica* 10 finds that largest number with `$MaxMachineNumber`, which returns `1.797693134862316`$\times$`10`^{308} on an iMac17,1. *Mathematica* 10's harmonic mean is `HarmonicMean`. Exceptionally accurate, `HarmonicMean[{$MaxMachineNumber,$MaxMachineNumber}]` from *Mathematica* 10 delivers the correct value, `1.797693134862316`$\times$`10`^{308}.

MATLAB® R2019a does not fare as well: its `realmax` function fetches the largest floating-point number on your machine, but its harmonic mean `harmmean` returns an infinity:

```
X = realmax;      Z = [X, X];      harmmean(Z)      Inf
```

However, using equation (12) with the same `Z`, MATLAB does yield the correct value:

$$\mathtt{H} = \frac{\mathtt{min(Z)}}{\frac{1}{2}(1 + (\mathtt{min(Z)/max(Z)}))} \qquad \mathtt{1.797693134862316e+308}.$$

With Julia, Python, or Sage, you may first have to import two packages, with commands such as `import sys`, to get a command to find the largest floating-point number and `import stats` to get the harmonic mean, `hmean`, which gives an infinity:

```
X = sys.float_info.max;      Z = [X, X];      stats.hmean(Z)      inf
```

So do `Hmean` and `harmonic.mean` from `R` with `X=as.numeric(.Machine[4])`.

Example 3. Instead of repeating Example 2 near (0, 0), Figure 3 displays plots of harmonic means, computed in two different ways, along with straight lines through the origin. If $y = c \cdot x$ with $0 \leq c \leq 1$, then $\min(x, y) = y = c \cdot x$. Equation (12) gives

$$H(x, c\cdot x) = \frac{\min(x, y)}{\frac{1}{2}\cdot\left[1 + \frac{\min(x,y)}{\max(x,y)}\right]} = \frac{c\cdot x}{\frac{1}{2}\cdot\left[1+\frac{c\cdot x}{x}\right]} = \frac{2\cdot c}{1+c}\cdot x. \tag{13}$$

Equation (13) shows that the graph of $z = H(x, c \cdot x)$ is a straight line with slope $2 \cdot c/(1 + c)$. Figure 3(A) shows that MATLAB's `harmmean` rounds to 0 there, contrary to the lower bounds from equations (9) and (11). In contrast, the steps from equation (7) to equation (9) are what give more accurate values in Figure 3(B), which uses equation (12).

In a multivariable calculus class, inequalities (11) give $\lim_{(x,y)\to(0,0)} H(x, y) = 0$. However, this limit only extends the graph of H to one point, where $H(0, 0) = 0$.

Do values of x and y of the order of 10^{308} or 2^{-1023} arise in the real world? Such values are near the largest and smallest positive numbers available out of the box in computers that conform to the IEEE Standard 754-1985 [3]. They were chosen so that you may try Examples 2 and 3 on your own computer. However, there are other computing systems. For instance, in the IEEE Standard 754-2019 half-precision binary floating-point format [4], the largest available positive number is $(2 - 2^{-10}) \cdot 2^{15} = 65\,504$, which does not accommodate the mass of Pluto (10^{22} kg), or the mass of any other planet [5, Table E.8, p. 476], while the smallest available positive number is $2^{-24} \approx 6 \cdot 10^{-8}$ [9]. In comparison, the electron, proton, and neutron have respective

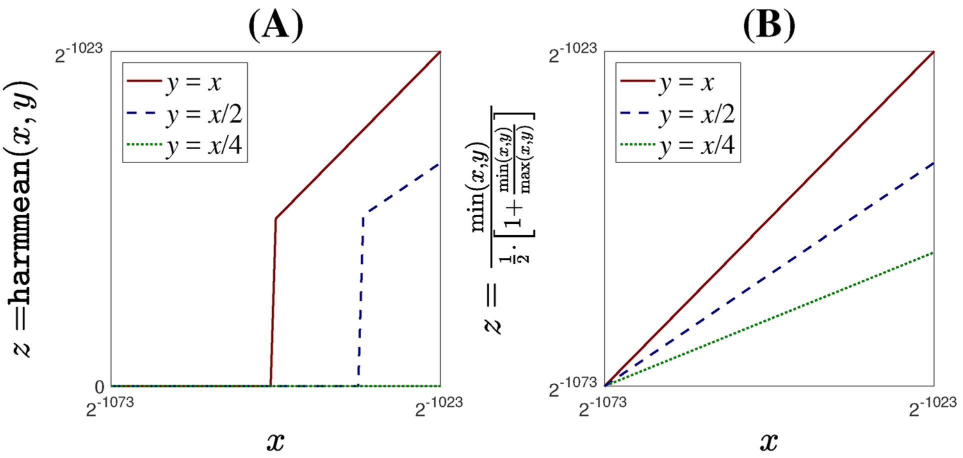


Figure 3 Plots of computed harmonic means near the origin: **(A)** from MATLAB's `harmmean`, **(B)** from equation (12), withMATLAB R2019a on an iMac17,1.

masses listed as $x = 9.035 \cdot 10^{-31}$ kg, $p = 1.67239 \cdot 10^{-27}$ kg, $n = 1.67470 \cdot 10^{-27}$ kg [**8**, p. 868]. The reduced mass of an electron, x, and a nucleus of tritium, $y = p + 2n \approx 5 \cdot 10^{-27}$, is about $R(x, y) \approx 9 \cdot 10^{-31}$ kg. Thus, there exist physical quantities with positive values that are larger than the largest positive number, or are smaller than the smallest positive number, available on some computing systems.

Analyzing and plotting transcendental function

We get similar results when we plot transcendental functions, as our final example shows.

Example 4. The rate of some chemical reactions, which is the number of moles reacting per unit of time, is proportional to a dimensionless function η of a dimensionless parameter ω called the *Thiele modulus* [**2**, p. 533]:

$$\eta(\omega) = \frac{3}{\omega}\left[\frac{1}{\tanh(\omega)} - \frac{1}{\omega}\right]. \tag{14}$$

Equation (14) fails for $\omega = 0$. We note that L'Hospital's rule or *Mathematica* gives $\lim_{\omega\to 0} \eta(\omega) = 1$, but this limit tells us nothing about $\eta(\omega)$ for any particular value $\omega \neq 0$. Figure 4 shows the limit point disconnected from a computed plot of η for $0 < \omega < 10^{-7}$.

Rearranging and substituting a Maclaurin series in equation (14) leads to:

$$\begin{aligned}\eta(\omega) &= 3 \cdot \frac{\omega \cdot \cosh(\omega) - \sinh(\omega)}{\omega^2 \cdot \sinh(\omega)}\\ &= 3 \cdot \frac{\sum_{k=0}^{\infty} \frac{\omega^{2k+1}}{(2k)!} - \sum_{k=0}^{\infty} \frac{\omega^{2k+1}}{(2k+1)!}}{\omega^2 \cdot \sum_{k=0}^{\infty} \frac{\omega^{2k+1}}{(2k+1)!}}. \end{aligned} \tag{15}$$

Substitute

$$\varphi(\omega) = \sum_{\ell=1}^{\infty}\left(\frac{3}{2\ell+3}\right)\left(\frac{\omega^{2\ell}}{(2\ell+1)!}\right) \quad \text{and} \quad \psi(\omega) = \sum_{k=1}^{\infty}\frac{\omega^{2k}}{(2k+1)!}$$

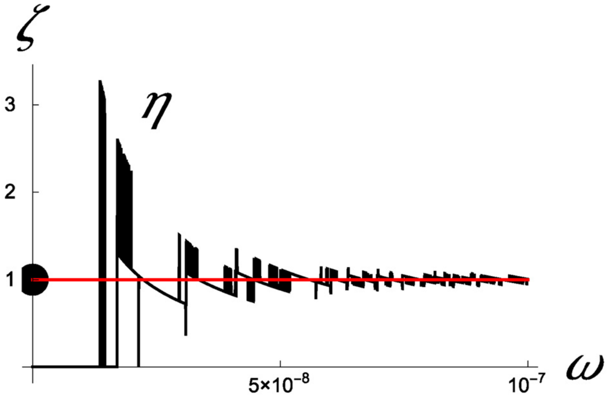


Figure 4 The limit $\lim_{\omega\to 0}\eta(\omega)=1$ tells us nothing about $\eta(\omega)$ for $\omega\neq 0$. Plots of η from equation (14) and $\eta\approx 1$ from inequalities (18), by *Mathematica* 10 on an iMac17,1.

to define a function α:

$$\eta(\omega)=\frac{1+\sum_{\ell=1}^{\infty}\frac{3}{2\ell+3}\frac{\omega^{2\ell}}{(2\ell+1)!}}{1+\sum_{k=1}^{\infty}\frac{\omega^{2k}}{(2k+1)!}}=\frac{1+\varphi(\omega)}{1+\psi(\omega)}=\alpha(\omega)\leq 1. \tag{16}$$

Thus, φ, ψ, and α are even and differentiable, with $\alpha(0)=1$, but

$$0=\alpha'(0)=\varphi'(0)=\psi'(0)=\varphi(0)=\psi(0).$$

The upper bound (16) is strict for $\omega\neq 0$ because each Maclaurin coefficient of ψ exceeds the corresponding coefficient of φ, which also shows that $\varphi''(\omega)<\psi''(\omega)$.

Substituting these results into the quotient rule yields $\alpha''(0)<0$. Thus, α has a strict global maximum at the origin. The first term of the power series for φ shows that $\varphi(\omega)\geq\frac{3}{5}\cdot\frac{\omega^2}{6}$. Using $(2k+1)!\geq 6^k$ for $k\geq 1$ in the power series for ψ gives

$$\psi(\omega)\leq\sum_{k=1}^{\infty}\left(\frac{\omega^2}{6}\right)^k=\frac{\omega^2}{6-\omega^2},$$

from the geometric series with ratio $\omega^2/6$ for $\omega^2<6$.

The inequality $1/(1+\psi)\geq 1-\psi$ then leads to

$$1\geq\eta(\omega)\geq\frac{1+\frac{3}{5}\cdot\frac{\omega^2}{6}}{1+\psi(\omega)}\geq\left(1+\frac{3}{5}\cdot\frac{\omega^2}{6}\right)\cdot\left[1-\frac{\omega^2}{6-\omega^2}\right]\geq 1-\frac{\omega^2}{10}, \tag{17}$$

where the lower bound (17) holds for $\omega^2\leq 2/3$. In particular, for $0<\omega<10^{-7}$,

$$0.999999999999999=1-10^{-15}=1-\frac{(10^{-7})^2}{10}<1-\frac{\omega^2}{10}<\eta(\omega)<1. \tag{18}$$

Thus, in Figure 4, the horizontal straight line at height 1 is a more accurate graph of η.

Conclusions

Computers can plot, do algebra, and find limits, but we must still tell them what to do. By default, computers use a fixed number of digits to evaluate formulae and plot them. In some situations, rounding errors can overwhelm computations, producing plots that are erroneous by orders of magnitude, with features that do not match the formulae.

To produce significantly more accurate plots, we may have to provide computers with other formulae, which are equivalent over some domain, but which avoid indeterminate forms at specific points. Merely instructing computers to “simplify” formulae does not necessarily work. To get better formulae, we may have to discover on our own specific sequences of logical, algebraic, or analytic steps, the same type of steps used to find limits. Software may be able to do such steps, but we still have to specify the steps.

Acknowledgments I thank the referees for suggestions that greatly improved the manuscript. This work was supported in part by a professional leave from Eastern Washington University.

REFERENCES

[1] Grabiner, J. V. (1983). The changing concept of change: the derivative from Fermat to Weierstrass. *Math. Mag.* 56(4): 195–206. doi.org/10.2307/2689807

[2] Helfferich, F. (1995). *Ion Exchange*. New York: Dover.

[3] Institute of Electrical and Electronic Engineers. (1985). *IEEE Standard for Binary Floating-Point Arithmetic (ANSI/IEEE Standard 754-1985)*.

[4] Institute of Electrical and Electronic Engineers. (2019). *IEEE Standard for Floating-Point Arithmetic (IEEE Standard 754-2019)*. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8766229. Last accessed June 2022.

[5] Karttunen, H., Kröger, P., Oja, H., Poutanen, M., Karl Johan Donner, K. J., eds. (1994). *Fundamental Astronomy*, 2nd ed. New York: Springer-Verlag.

[6] Kittel, C., Knight, W. D., Ruderman, M. A., Helmholz, A. C., Moyer, B. J. (1973). *Mechanics*. 2nd ed. Vol. 1 of *Berkeley Physics Course*. New York: McGraw-Hill. Available at: https://b-ok.org/book/1199550/31aa18/. Last accessed June 2022.

[7] Lundeberg, M. B., Gao, Y., Asgari, R., Tan, C., Van Duppen, B., Autore, M., Alonso-Gonzáles, P., Woessner, A., Watanabe, K., Taniguchi, T., et. al (2017). Tuning quantum nonlocal effects in graphene plasmonics. *Science*. 357(6385): 187–190. doi.org/10.1126/science.aan2735

[8] Pauling, L. (1988). *General Chemistry*. New York: Dover.

[9] Pranesh, S. (2019). Low precision floating-point formats: The wild west of computer arithmetic. *SIAM News*. 52(9): 12.

[10] Purcell, E. M. (1965). *Electricity and Magnetism*. 3rd ed. Vol. 2 of *Berkeley Physics Course*. New York: McGraw-Hill. Available at: http://www.sicyon.com/resources/library/pdf/Electricity_and_Magnetism.pdf. Last accessed June 2022.

[11] Range, R. M. (2011). Where are limits needed in calculus? *Amer. Math. Monthly*. 118(5): 404–417. doi: 10.4169/amer.math.monthly.118.05.404

Summary. Professional mathematical software may produce plots of elementary functions that are in error by orders of magnitude or show erroneous features. The same software may give a correct limit near an endpoint, but such a limit yields at most a single point on the graph. In contrast, proofs or derivations of the same limit provide information to guide the software toward a sketch that shows correct trends and stays within correct bounds.

YVES NIEVERGELT completed his undergraduate degree in mathematics from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in December of 1976, skipped the graduation ceremony in 1977 for compulsory Swiss military service, but increasingly appreciates the quality of the education he got at the EPFL. In 1984, he earned a Ph.D. in several complex variables under the guidance of James R. King at the University of Washington in Seattle, where he also benefited from Caspar R. Curjel’s mentoring in teaching. Since 1985, he has been teaching complex and numerical analysis as well as other topics at Eastern Washington University.

Carl B. Allendoerfer Awards

The Carl B. Allendoerfer Awards, established in 1976, are presented to authors of articles of expository excellence published in MATHEMATICS MAGAZINE. The Award is named for Carl B. Allendoerfer, a distinguished mathematician at the University of Washington and president of the Mathematical Association of America, 1959–1960.

David J. Hunter and Chisondi Warioba

"Segregation Surfaces" MATHEMATICS MAGAZINE, Volume 94, Number 3, June 2021, pages 163–172.

Measuring segregation on a city map is not simple, and such measurements are not defined in just one way. The article "Segregation Surfaces" shares with us several such approaches that have been developed by social scientists, with most of these ideas using concepts from multivariable calculus. Throughout, the authors make an excellent case for applying mathematical techniques with care and caution, recognizing that there is no single correct approach and no quick fix.

Simultaneously, the article does a marvelous job highlighting how undergraduate data analysis and mathematical techniques can lend insight into how we quantify segregation patterns.

The approaches to measuring segregation are illustrated via formulas, full-color maps, and mathematical explanations. A theme across these measurements is that they occur on two-dimensional maps displaying data representing varying concentrations of groups of people. The maps show contours identifying levels of concentration, as well as directions of greatest change, also known as gradients. The mathematical tools used quickly demonstrate not only that this article is showing us mechanisms for indexing segregation, but also that we can study this topic by drawing many of our ideas from a typical course in multivariable calculus.

Early in the article, its authors discuss the conversion of map data into a surface. This process begins with probability density functions describing how two groups, A and B, are distributed, then uses kernel density estimation to determine the surface. The estimation involves parameters that may be chosen in different ways. One system for this process leads to a choropleth map, which shades the map according to the proportion of white residents and draws contour lines showing the probability that a resident is white. Another visualization on a 2D map shows segregation gradients, with direction indicating the greatest increase in proportion of white residents, and length showing how rapidly the proportion changes.

These ideas come together to form two different indices of segregation. One index computes the average gradient length along the entire 50% contour, that is, along the contour showing 50% white residents. Depending on the data and segregation patterns in a city, such a contour may not be defined, so another index computes the average gradient length across the entire region. The first index makes use of gradients and a contour integral; the second uses gradients and a double integral across a region's area.

Math. Mag. **95** (2022) 397–400. doi:10.1080/0025570X.2022.2094679 © Mathematical Association of America
MSC: Primary 01A70

These formulas bring together calculus concepts, while showing how these ideas can be used in the context of a meaningful data set.

The authors end by providing readers with several articles where we can learn more, presenting several related problems to try, and linking to their code and data. Throughout, "Segregation Surfaces" makes an excellent case for applying mathematical techniques with care and caution, recognizing that there is no single correct approach and no quick fix. Simultaneously, the article does a marvelous job highlighting how undergraduate data analysis and mathematical techniques can lend insight into how we quantify segregation patterns.

Response from the Authors

David J. Hunter

We are honored that our paper on segregation measures and visualization has been selected for a Carl B. Allendoerfer award. We would like to thank the editorial staff of *Mathematics Magazine* and the careful work of anonymous reviewers whose insightful comments improved the paper. We are also grateful for powerful open source tools and open data practices that can support undergraduate research in a range of disciplines. Our hope is that this work will inspire other mathematical investigations into topics that address important questions.

Chisondi Warioba

Math has always fascinated me as a language. It is a universal way to communicate. As someone who speaks English as a second language, the ability to describe the world we live in with such universal descriptors will always take my breath away. It is an honor to contribute to this field and an even greater honor to be recognized as a recipient of this year's Allendoerfer Award.

DAVID J. HUNTER (MR Author ID: 633493) received his Ph.D. from the University of Virginia, Charlottesville in 1997 and now teaches mathematics and computer science at Westmont College, Santa Barbara, CA. As a transplanted Chicagoan living in Santa Barbara, he loves to walk around cities and hike in the mountains.

CHISONDI WARIOBA (ORCID: 0000-0002-4266-2673) is originally from Tanzania, East Africa. He graduated from Westmont College in 2021 with a Bachelor of Science in Chemistry, Physics, and Biology. He is currently a second-year Ph.D. student in Medical Physics at the University of Chicago, studying resting state functional connectivity in MR-imaged stroke models. He is using his love for math in the statistical analysis required for the study. He is passionate about equity in higher education and anything music-related.

Kaity Parsons Peter Tingley, and Emma Zajdela

"When to Hold 'Em" Mathematics Magazine Volume 94, Number 3, June 2021, pages 201–212.

"So you want to win at poker?" Thus begins this exciting article. The authors work through strategies based on the hands that are dealt, their probabilities of winning, random behavior of players, bluffing, and slow-play. They begin with just two players and only six possible hands. In an expert teaching move, they use this simple case to build reader intuition. From here, they progress to a discussion of infinitely many poker hands and widely expanded ways for players to bet. Their explanations are approachable, and their lively writing style welcomes us to continue reading and thinking.

To simplify things, the article focuses on a game of poker involving only two players. To further set rules that allow the authors to analyze the game's outcomes, they require both players to write a computer program specifying how they will play. Player 1 writes their program first, and Player 2 gets to see Player 1's program while deciding on their own program! Though this may seem less than fair, the authors make a case that poker players who have known each other a long while are each likely to know the other's typical strategies. Thus, seeing another player's computer program is not so far from the reality of competing against each other. As a fascinating outcome, if Player 1 proceeds in a totally straightforward way, betting only on hands that are likely to win, then the best outcome for Player 1 is to break even. However, by bluffing, Player 1 can win money from Player 2.

From here, the article takes the logical next step of assuming that Player 1 also knows the entire strategy for Player 2. Therefore, each player knows what to expect from the other, which is similar to what might happen between two people who have played poker together many times before. Both players can then determine their best system of play, based on full information about the other player's decision-making strategies, which can lead to a Nash equilibrium. In developing the many possible outcomes they consider, the authors use a tried-and-true teaching technique: they begin with a simplified set of rules through which they build intuition with their readers, and then they progress to wide-ranging and much more abstract ideas. In particular, the authors initially allow only six possible poker hands, determined randomly by the roll of a die. With only six hands, the table of outcomes fits nicely onto a journal page. Once readers understand these outcomes, the authors introduce the idea of infinitely many outcomes, having all possible probabilities from 0 to 1, and this concept appears completely natural and quickly understandable. The betting possibilities also expand dramatically throughout the article.

"When to Hold 'Em" is lively and conversational throughout. In this inviting format, the authors carefully and fully develop several approaches to evaluating poker play. Their explanations are approachable, and their writing style welcomes us to continue reading and thinking. They have put together a wonderfully readable examination of mathematical ideas.

Response from the Authors

Kaity Parsons

I am both honored and surprised to receive this award. This paper was the greatest achievement of my undergraduate career at Loyola University, Chicago. When Dr. Tingley first approached me with this project, I was thrilled at the concept. I grew up playing card games, poker included. Though I have yet to win millions in Vegas, the theories outlined in this paper have served me well. However, the experience of this project was the most vital in developing my love for playing with numbers. The many hours spent sitting with a simple question—How can you win, or rather, lose the least, at poker?—was pivotal in my math journey. Now, I try to do the same for my own students. Math has a bad reputation and I, like Dr. Tingley and this project did for me, am determined to make math fun.

Peter Tingley

It is a great honor and pleasure to accept this award, thank you! The writing of "When to Hold–Em" played out over several years and involved many people. It really began in 2012, when I saw an amazing and inspiring talk on poker and math by Yan X. Zhang. It continued when Nick Barron convinced me to teach game theory, which was not at all my field, but which I greatly enjoyed. Kaity Parsons and Emma Zajdela were both students in that class, and both did research projects on poker with me, leading to the paper's first draft. We kept in touch after they graduated, and the paper slowly evolved. It has been used in my game theory classes ever since, and many students have commented on and improved it—most notably Emily Danning He who in 2017 gave it a super thorough proofreading. The whole process has been a wonderful experience, which I am grateful to have shared with these amazing students. It was its own reward, but winning an actual award is certainly a nice addition!

Emma Zajdela

I am honored and delighted to receive the Carl B. Allendoerfer award for our paper on game theory and poker. Historically, people have become interested in mathematical problems through games of chance (think of Pascal in the 17th century), and this holds true today—workshops based on this paper have been successful in several outreach programs designed to introduce high-school students to mathematical research. It also sparked a fascination for me with the idea that we can use math to understand human behavior. This research was the impetus for me to pursue a Ph.D. in applied math, with a focus on modeling complex social systems, for which I received the NSF Graduate Research Fellowship.

KAITY PARSONS (MR Author ID: 1345241) received a B.S. in mathematics from Loyola University Chicago in 2017. Since graduating, Kaity has been working as a center director at the Town and Country Mathnasium in St. Louis Missouri. She and her partner are currently planning a one-way cross-country road trip with their dog, Angel and their cat, Salem, to seek what they will find.

PETER TINGLEY (MR Author ID: 679482) received a Ph.D. in Mathematics from the University of California, Berkeley in 2008. He spent short periods at the University of Melbourne (Australia) and MIT, and since 2012 has been at Loyola University Chicago. He also helps run the Chicago Math Teachers' circle, which he co-founded in 2015, and is the current blue division racquetball champion at the Evanston YMCA.

EMMA ZAJDELA (MR Author ID: 1345242) received a B.S. in math and physics from Loyola University Chicago in 2016 and an M.S. in math from the University of Illinois Chicago in 2018. She is currently an NSF fellow and Ph.D. student in applied mathematics at Northwestern University. Since 2016 she has served as Assistant to the President of the Malta Conferences Foundation, a nonprofit that uses science as a bridge to peace in the Middle East. She also recently received her yellow belt in judo.

PROOFS WITHOUT WORDS

Euler Bricks

TOM EDGAR
Pacific Lutheran University
Tacoma, WA 98447
edgartj@plu.edu

An *Euler brick*, also called a rational cuboid, is a rectangular prism with the property that the length, the width, the height, and *all* the diagonals of the rectangular faces are integers. We provide two visual proofs implying the existence of infinitely many Euler bricks using Saunderson's parametrization [5]. For ease of notation, we use (a, b, c) to denote a Pythagorean triple of the form $a^2 + b^2 = c^2$. Given two positive integers m and n, it is clear that

$$(|m^2 - n^2|, 2mn, m^2 + n^2)$$

is a Pythagorean triple (see Houston [3] for a visual proof). Using this type of triple, we substitute and scale to get Pythagorean triples of the form

$$(x|4y^2 - z^2|, 4xyz, x(4y^2 + z^2))$$

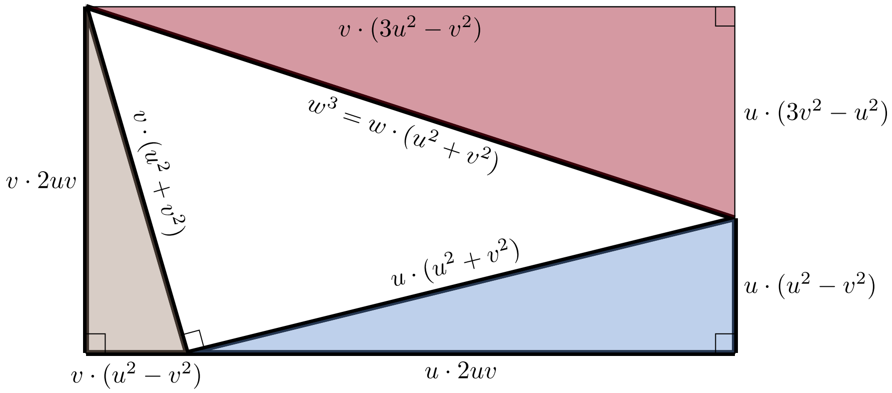
for any positive integers x, y, and z. We then apply a visual technique known informally as Garfield's trapezoid (see Alsina and Nelsen [1]) to obtain another Pythagorean triple via the following lemma.

Lemma. *Let u, v and w all be positive integers. If (u, v, w) is a Pythagorean triple then*

$$(v|3u^2 - v^2|, u|3v^2 - u^2|, w^3)$$

is also a Pythagorean triple.

Proof. We may assume that $u > v$. The following diagram demonstrates the proof when $3v^2 > u^2$. For the case $3v^2 < u^2$, we interchange the respective labels of the left (brown) and right (blue) triangles, so the top (red) triangle has height $u \cdot (u^2 - 3v^2)$.*



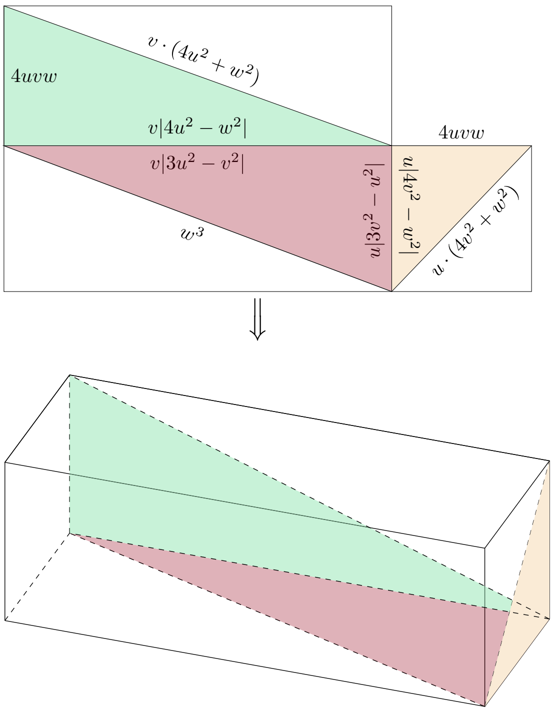
■

Math. Mag. **95** (2022) 401–402. doi:10.1080/0025570X.2022.2094178 © Mathematical Association of America

*The online version of this article has color diagrams.

Theorem. *There are infinitely many Euler bricks: if* (u, v, w) *is a Pythagorean triple, then there is an Euler brick with side lengths* $u|3v^2 - u^2|$, $v|3u^2 - v^2|$, *and* $4uvw$.

Proof.



■

The Euler bricks constructed here are known to not have an integer-valued interior diagonal [6]. It is still an open question to determine if there exists a *perfect cuboid*, which is an Euler brick that also has an integer-valued interior diagonal [2, 4].

REFERENCES

[1] Alsina. C., Nelsen, R. B. (2011). *Icons of Mathematics: An Exploration of Twenty Key Images*. Washington D. C.: Mathematical Association of America.

[2] Guy, R. K. (2004). *Unsolved Problems in Number Theory*, 3rd ed. New York: Springer-Verlag.

[3] Houston, D. (1994). Proof without words: Pythagorean triples via double angle formulas. *Mathematics Magazine* 67(3): 187. doi.org/10.1080/0025570X.1994.11996211

[4] Leech, J. (1977). The rational cuboid revisted. *Amer. Math. Monthly* 84(7): 518–533. doi.org/10.1080/00029890.1977.11994405

[5] Saunderson, N. (1740). *The Elements of Algebra*, Vol. 2. Cambridge: Cambridge University Press.

[6] Spohn, W. G. (1972). On the integral cuboid. *Amer. Math. Monthly*. 79(1): 57–59. doi.org/10.1080/00029890.1972.11992984

Summary. We provide a visual argument that there are infinitely many Euler bricks, also known as rational cuboids.

TOM EDGAR (MR Author ID: 821633) is a professor of mathematics at Pacific Lutheran University and the editor of *Math Horizons*. His colleague, Jessica K. Sklar, inspired this note by finding a reference to Euler bricks in the movie Escape Room (2019).

Nesbitt's Inequality

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

Nesbitt's inequality, a staple of mathematics competitions, states that for all positive numbers a, b, and c, we have that

$$\frac{a}{b+c}+\frac{b}{c+a}+\frac{c}{a+b}\geq\frac{3}{2}$$

with equality if and only if $a=b=c$ (see Nesbitt [1] or Steele [2, Exercise 5.6]). It has been proved many times in a variety of ways. In this note, we establish two lemmas visually, from which the inequality immediately follows.

Lemma 1. *If* $a, b, c > 0$, *then*

$$(a+b+c)\left(\frac{1}{a+b}+\frac{1}{b+c}+\frac{1}{c+a}\right)=\frac{a}{b+c}+\frac{b}{c+a}+\frac{c}{a+b}+3.$$

Proof. See Figure 1. ■

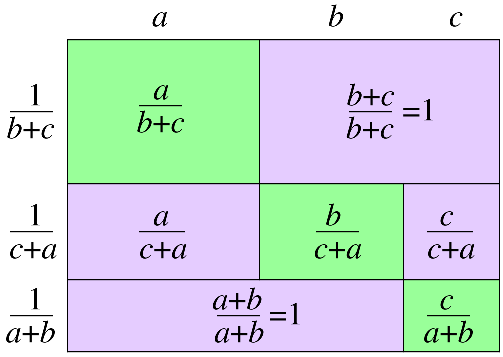


Figure 1 The proof of Lemma 1.

Lemma 2. *If* $x, y, z > 0$, *then*

$$(x+y+z)\left(\frac{1}{x}+\frac{1}{y}+\frac{1}{z}\right)\geq 9.$$

Proof. See Figure 2. ■

PROOF OF NESBITT'S INEQUALITY:

$$\frac{a}{b+c}+\frac{b}{c+a}+\frac{c}{a+b}=(a+b+c)\left(\frac{1}{a+b}+\frac{1}{b+c}+\frac{1}{c+a}\right)-3$$

Math. Mag. **95** (2022) 403–404. doi:10.1080/0025570X.2022.2094179 © Mathematical Association of America

$$= \frac{1}{2}[(a+b)+(b+c)+(c+a)]\left(\frac{1}{a+b}+\frac{1}{b+c}+\frac{1}{c+a}\right)-3$$

$$\geq \frac{9}{2}-3=\frac{3}{2}.$$

■

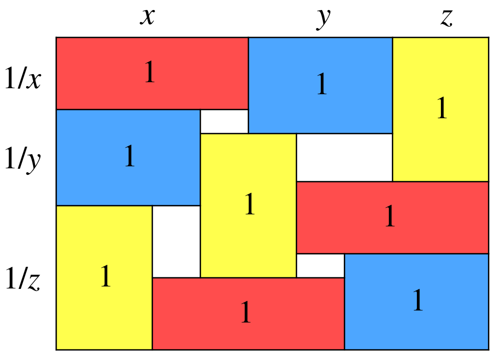


Figure 2 The proof of Lemma 2.

NOTE. When a, b, and c are the side lengths of a triangle, Nesbitt's inequality becomes

$$\frac{3}{2} \leq \frac{a}{b+c}+\frac{b}{c+a}+\frac{c}{a+b} < 2.$$

To establish the upper bound, note that the triangle law implies that the denominators are each greater than the semi-perimeter $s=(a+b+c)/2$. Hence,

$$\frac{a}{b+c}+\frac{b}{c+a}+\frac{c}{a+b} < \frac{a+b+c}{s} = 2.$$

To show that the upper bound is best possible; consider triangles with sides n, n, and ϵ.

REFERENCES

[1] Nesbitt A. M. (1903). Problem 15114. *Educational Times*. 2: 37–38

[2] Steele, J. M. (2004). *The Cauchy-Schwarz Master Class.* Washington, DC: Mathematical Association of America; and Cambridge, UK: Cambridge University Press.

Summary. We prove two lemmas visually to establish Nesbitt's inequality.

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

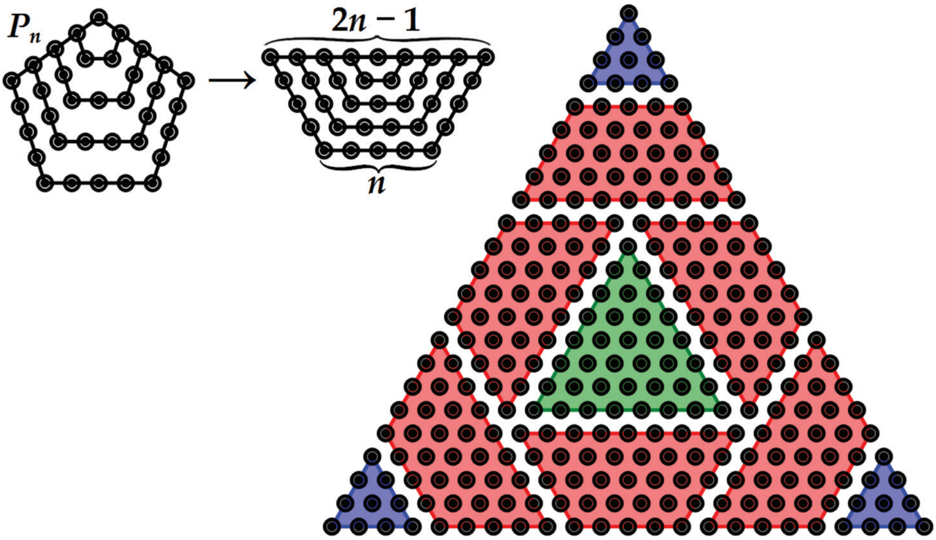
An Identity Relating Triangular and Pentagonal Numbers

GUNHAN CAGLAYAN
New Jersey City University
Jersey City, NJ 07305
gcaglayan@ncju.edu

Theorem 1. *Let P_n and T_n represent the nth pentagonal and the nth triangular number, respectively, for $n \in \mathbb{N}$. Then we have the following identity:*

$$T_{5n-2} = T_{2n-2} + 6P_n + 3T_{n-1}.$$

We illustrate the proof for $n = 5$.



Acknowledgment The author wishes to thank the *Mathematics Magazine* editorial board and anonymous reviewers for their helpful comments on earlier drafts.

Summary. We give a visual proof for an identity relating triangular and pentagonal numbers.

GUNHAN CAGLAYAN (MR Author ID: 1116420) teaches mathematics at New Jersey City University. His main interests are visual mathematics and student learning through modeling and visualization.

Math. Mag. **95** (2022) 405. doi:10.1080/0025570X.2022.2104560 © Mathematical Association of America

PROBLEMS

Proposals

To be considered for publication, solutions should be received by March 1, 2023.

2151. *Proposed by Tran Quang Hung, Hanoi, Vietnam.*

Let $ABCD$ and $XYZT$ be two directly similar squares such that A and Y lie on the lines XT and CD, respectively. Let M be the intersection of lines XZ and AC, and let N be the intersection of lines XY and BC. Prove that circumcenter of $\triangle XAC$ lies on the line MN.

2152. *Proposed by Paul Bracken, University of Texas Rio Grande Valley, Edinburg, TX.*

Evaluate

$$\int_0^1 \int_0^1 \frac{dy\,dx}{\sqrt{1-x^2}\sqrt{1-y^2}\,(1+xy)}.$$

2153. *Proposed by Rex H. Wu, New York, NY.*

Let F_n and L_n be the Fibonacci and Lucas numbers, respectively. Evaluate the following for $k \geq 0$.

(a) $\displaystyle\sum_{n=0}^{\infty} \arctan \frac{F_{2k}}{F_{2n+1}}$

(b) $\displaystyle\sum_{n=0}^{\infty} \arctan \frac{L_{2k+1}}{L_{2n}}$

Math. Mag. **95** (2022) 406–414. doi:10.1080/0025570X.2022.2103317 © Mathematical Association of America

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

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

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

2154. *Proposed by the Columbus State University Problem Solving Group, Columbus, GA.*

Let $f(n)$ denote the number of ordered partitions of a positive integer n such that all of the parts are odd. For example, $f(5) = 5$, since 5 can be written as $5, 3 + 1 + 1, 1 + 3 + 1, 3 + 1 + 1$, and $1 + 1 + 1 + 1 + 1$. Determine $f(n)$.

2155. *Proposed by Ioan Băetu, Botoşani, Romania.*

Let R be a ring with identity and U a subset of the units of R with $|U| = p$, where p is an odd prime. Suppose that for all $a \in R$, there is a $u \in U$ and a $k \in \mathbb{Z}^+$ such that $ua^k = a^{k+1}$. Show that

(a) For all $a \in R$, there is a $u \in U$ such that $ua = a^2$.

(b) The ring R is commutative.

Quickies

1123. *Proposed by George Stoica, Saint John, NB, Canada.*

Given a function $f : \mathbb{R}^\ell \to \mathbb{R}$, we say that f satisfies condition P_n if

$$f\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right) = \frac{1}{n}\sum_{i=1}^{n} f(A_i)$$

for all $A_1, \ldots, A_n \in \mathbb{R}^\ell$. Show that for all $m, n \geq 2$, conditions P_m and P_n are equivalent.

1124. *Proposed by A. Berele and T. Kyle Petersen, DePaul University, Chicago, IL.*

Let $\{F_n\}_{n=1}^\infty$ be the Fibonacci sequence $1, 1, 2, 3, 5, 8, 13, \ldots$. Does there exist an infinite subsequence $\{F_{n_i}\}_{i=1}^\infty$, the sum of whose reciprocals converges to 1?

Solutions

Minimize the length of the tangent segment **October 2021**

2126. *Proposed by M. V. Channakeshava, Bengaluru, India.*

A tangent line to the ellipse

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

meets the x-axis and y-axis at the points A and B, respectively. Find the minimum value of AB.

Solution by Kangrae Park (student), Seoul National University, Seoul, Korea.
We may assume that $a, b > 0$ and that the point of tangency $P = (\alpha, \beta)$ lies in the first quadrant. One readily verifies that the tangent line to the ellipse at P is

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

Therefore, A and B are $(a^2/\alpha, 0)$ and $(0, b^2/\beta)$, respectively. Note that

$$\frac{\alpha^2}{a^2} + \frac{\beta^2}{b^2} = 1$$

since the point P is on the ellipse. Applying the Cauchy-Schwarz inequality with

$$\mathbf{u} = \left(\frac{a^2}{\alpha}, \frac{b^2}{\beta}\right) \quad \text{and} \quad \mathbf{v} = \left(\frac{\alpha}{a}, \frac{\beta}{b}\right),$$

we obtain

$$\frac{a^4}{\alpha^2} + \frac{b^4}{\beta^2} = \left(\frac{a^4}{\alpha^2} + \frac{b^4}{\beta^2}\right)\left(\frac{\alpha^2}{a^2} + \frac{\beta^2}{b^2}\right) = (\mathbf{u}\cdot\mathbf{u})(\mathbf{v}\cdot\mathbf{v}) \geq (\mathbf{u}\cdot\mathbf{v})^2 = (a+b)^2.$$

It follows that

$$AB = \sqrt{\frac{a^4}{\alpha^2} + \frac{b^4}{\beta^2}} \geq a + b.$$

This lower bound is attained if and only if $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent. A straightforward calculation shows that this occurs if and only if

$$\alpha^2 = \frac{a^3}{a+b} \quad \text{and} \quad \beta^2 = \frac{b^3}{a+b}.$$

This gives the esthetically pleasing result that when AB attains its minimum value of $a + b$, we have $PB = a$ and $PA = b$.

Also solved by Ulrich Abel & Vitaliy Kushnirevych (Germany), Yagub Aliyev (Azerbaijan), Michel Bataille (France), Bejmanin Bittner, Khristo Boyadzhiev, Paul Bracken, Brian Bradie, Robert Calcaterra, Hongwei Chen, Joowon Chung (South Korea), Robert Doucette, Rob Downes, Eagle Problem Solvers (Georgia Southern University), Habib Y. Far, John Fitch, Dmitry Fleischman, Noah Garson (Canada), Kyle Gatesman, Subhankar Gayen (India), Jan Grzesik, Emmett Hart, Eugene A. Herman, David Huckaby, Tom Jager, Walther Janous (Austria), Mark Kaplan & Michael Goldenberg, Kee-Wai Lau (Hong Kong), Lucas Perry & Alexander Perry, Didier Pinchon (France), Ivan Retamoso, Celia Schacht, Randy Schwartz, Ioannis Sfikas (Greece), Vishwesh Ravi Shrimali (India), Albert Stadler (Switzerland), Seán M. Stewart (Saudi Arabia), David Stone & John Hawkins, Nora Thornber, R. S. Tiberio, Michael Vowe (Switzerland), Lienhard Wimmer (Germany), and the proposer. There were seventeen incomplete or incorrect solutions.

Two idempotent matrices October 2021

2127. *Proposed by Jeff Stuart, Pacific Lutheran University, Tacoma, WA and Roger Horn, Tampa, FL.*

Suppose that $A, B \in M_{n\times n}(\mathbb{C})$ is such that $AB = A$ and $BA = B$. Show that

(a) A and B are idempotent and have the same null space.

(b) If $1 \leq \operatorname{rank} A < n$, then there are infinitely many choices of B that satisfy the hypotheses.

(c) $A = B$ if and only if $A - I$ and $B - I$ have the same null space.

Solution by Michel Bataille, Rouen, France.
(a) The fact that $A^2 = A$ and $B^2 = B$ follows from:

$$A^2 = (AB)A = A(BA) = AB = A, \qquad B^2 = (BA)B = B(AB) = BA = B.$$

In addition, if X is a column vector and $AX = 0$, then $BAX = 0$, that is, $BX = 0$. Thus, $\ker A \subseteq \ker B$. Similarly, if $BX = 0$, then $ABX = 0$. Hence $AX = 0$ so that $\ker B \subseteq \ker A$. We conclude that $\ker A = \ker B$.

(b) Let $r = \text{rank}(A)$. Since A is idempotent, we have $\text{range}(A) \oplus \ker A = \mathbb{C}^n$. Since $AX = X$ if $X \in \text{range}(A)$ and $\dim(\text{range}(A)) = r$, it follows that $A = PJ_rP^{-1}$ for some invertible $n \times n$ matrix P and

$$J_r = \left(\begin{array}{c|c} I_r & O \\ \hline O & O \end{array}\right),$$

where I_r denotes the $r \times r$ unit matrix and O a null matrix of the appropriate size. Consider the matrices $B = PB'P^{-1}$ with

$$B' = \left(\begin{array}{c|c} I_r & O \\ \hline C & O \end{array}\right),$$

where C is an arbitrary $(n - r) \times r$ matrix with complex entries. There are infinitely many such matrices B, and we calculate

$$AB = PJ_rP^{-1}PB'P^{-1} = PJ_rB'P^{-1} = PJ_rP^{-1} = A,$$

and

$$BA = PB'P^{-1}PJ_rP^{-1} = PB'J_rP^{-1} = PB'P^{-1} = B.$$

(c) Clearly, $A - I$ and $B - I$ have the same null space if $A = B$. Conversely, suppose that $\ker(A - I) = \ker(B - I)$. Let X be a column vector. Since $(A - I)A = O$, the vector AX is in $\ker(A - I)$, hence is in $\ker(B - I)$. This means that $(B - I)AX = 0$, that is, $BX = AX$ (since $BA = B$). Since X is arbitrary, we can conclude that $A = B$.

Also solved by Paul Budney, Robert Calcaterra, Hongwei Chen, Robert Doucette, Dmitry Fleischman, Kyle Gatesman, Eugene A. Herman, Tom Jager, Rachel McMullan, Thoriq Muhammad (Indonesia), Didier Pinchon (France), Michael Reid, Randy Schwartz, Omar Sonebi (Morroco), and the proposer. There was one incomplete or incorrect solution.

Two exponential inequalities October 2021

2128. *Proposed by George Stoica, Saint John, NB, Canada.*

Let $0 < a < b < 1$ and $\epsilon > 0$ be given. Prove the existence of positive integers m and n such that $(1 - b^m)^n < \epsilon$ and $(1 - a^m)^n > 1 - \epsilon$.

Solution by Robert Doucette, McNeese State University, Lake Charles, LA.
It is well known that

$$\lim_{x\to 0}(1 - x)^{1/x} = e^{-1}.$$

Suppose $0 < \alpha < 1$. Then, since $\alpha^x \to 0^+$ as $x \to \infty$,

$$\lim_{x\to\infty}(1 - \alpha^x)^{\alpha^{-x}} = e^{-1}.$$

Hence,

$$\lim_{x\to\infty}(1 - \alpha^x)^{\beta^{-x}} = \lim_{x\to\infty}\left[(1 - \alpha^x)^{\alpha^{-x}}\right]^{(\beta/\alpha)^{-x}} = \begin{cases} 0, & \text{if } 0 < \beta < \alpha < 1 \\ 1, & \text{if } 0 < \alpha < \beta < 1 \end{cases}.$$

Choose c and d such that $0 < a < c < d < b < 1$. Note that $c^{-x} - d^{-x} \to \infty$ as $x \to \infty$.

By the limits established above, there exists a positive integer m such that

$$(1 - b^m)^{d^{-m}} < \epsilon, (1 - a^m)^{c^{-m}} > 1 - \epsilon, \text{ and } c^{-m} - d^{-m} > 1.$$

There also exists a positive integer n such that $d^{-m} < n < c^{-m}$. Therefore,

$$(1 - b^m)^n < (1 - b^m)^{d^{-m}} < \epsilon \text{ and } (1 - a^m)^n > (1 - a^m)^{c^{-m}} > 1 - \epsilon.$$

Also solved by Levent Batakci, Michel Bataille (France), Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Bruce Burdick, Michael Cohen, Dmitry Fleischman, Kyle Gatesman, Michael Goldenberg & Mark Kaplan, Eugene Herman, Miguel Lerma, Reiner Martin (Germany), Raymond Mortini (France), Michael Nathanson, Moubinool Omajee (France), Didier Pinchon (France), Albert Stadler (Switzerland), Omar Sonebi (Morroco), and the proposer.

Two improper integrals October 2021

2129. *Proposed by Vincent Coll and Daniel Conus, Lehigh University, Bethlehem, PA and Lee Whitt, San Diego, CA.*

Determine whether the following improper integrals are convergent or divergent.

(a) $\displaystyle\int_0^1 \exp\left(\sum_{k=0}^{\infty} x^{2^k}\right) dx$

(b) $\displaystyle\int_0^1 \exp\left(\sum_{k=0}^{\infty} x^{3^k}\right) dx$

Solution by Gerald A. Edgar, Denver, CO.
(a) The integral diverges. For $0 < x < 1$ we have

$$\begin{aligned}\log\frac{1}{1-x} &= \sum_{n=1}^{\infty}\frac{1}{n}x^n = \sum_{k=0}^{\infty}\left(\sum_{n=2^k}^{2^{k+1}-1}\frac{1}{n}x^n\right)\\ &\le \sum_{k=0}^{\infty}\left(\sum_{n=2^k}^{2^{k+1}-1}\frac{1}{2^k}x^{2^k}\right) = \sum_{k=0}^{\infty}\left(\frac{2^k}{2^k}x^{2^k}\right) = \sum_{k=0}^{\infty}x^{2^k}.\end{aligned}$$

Therefore,

$$\exp\left(\sum_{k=0}^{\infty}x^{2^k}\right) \ge \frac{1}{1-x}.$$

The integral (a) diverges by comparison with the divergent integral $\int_0^1 dx/(1-x)$.

(b) The integral converges. We will need an estimate for a harmonic sum. The function $1/x$ is decreasing, so for $k \ge 1$

$$\sum_{n=3^{k-1}}^{3^k-1}\frac{1}{n} > \int_{3^{k-1}}^{3^k}\frac{dx}{x} = \log 3.$$

Now, for $0 < x < 1$ we have

$$\log \frac{1}{1-x} = \sum_{n=1}^{\infty} \frac{1}{n} x^n = \sum_{k=1}^{\infty} \left(\sum_{n=3^{k-1}}^{3^k - 1} \frac{1}{n} x^n \right)$$

$$> \sum_{k=1}^{\infty} \left(\sum_{n=3^{k-1}}^{3^k - 1} \frac{1}{n} \right) x^{3^k} > \sum_{k=1}^{\infty} (\log 3) x^{3^k}.$$

Let $r = 1/\log 3$, so that $0 < r < 1$. Then

$$r \log \frac{1}{1-x} > \sum_{k=1}^{\infty} x^{3^k},$$

$$\log \frac{1}{(1-x)^r} + 1 > \sum_{k=0}^{\infty} x^{3^k},$$

$$\frac{e}{(1-x)^r} > \exp\left(\sum_{k=0}^{\infty} x^{3^k} \right).$$

The integral (b) converges by comparison with the convergent integral

$$\int_0^1 \frac{e}{(1-x)^r}\, dx.$$

Editor's Note. A more detailed analysis shows that

$$\int_0^1 \exp\left(\sum_{k=0}^{\infty} x^{\alpha^k} \right) dx$$

converges if $\alpha > e$ and diverges if $1 \le \alpha \le e$.

Also solved by Michael Bataille (France), Robert Calcaterra, Dmitry Fleischman, Eugene A. Herman, Walther Janous (Austria), Albert Natian, Moubinool Omarjee (France), Didier Pinchon (France), Albert Stadler (Switzerland), and the proposers. There was one incomplete or incorrect solution.

When does the circumcenter lie on the incircle? October 2021

2130. *Proposed by Florin Stanescu, Şerban Cioculescu School, Găeşti, Romania.*

Given the acute $\triangle ABC$, let D, E, and F be the feet of the altitudes from A, B, and C, respectively. Choose $P, R \in \overleftrightarrow{AB}$, $S, T \in \overleftrightarrow{BC}$, $Q, U \in \overleftrightarrow{AC}$ so that

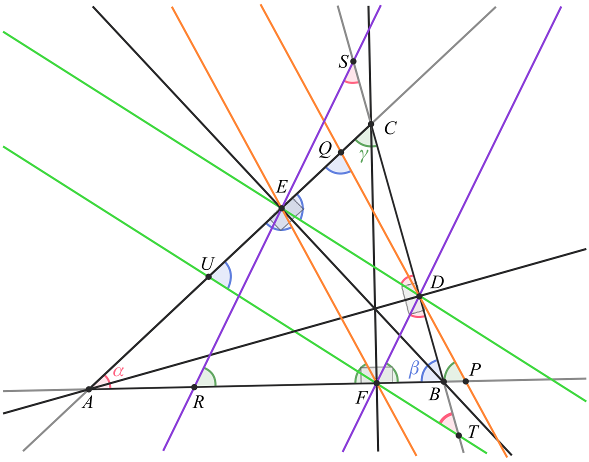
$$D \in \overleftrightarrow{PQ},\ E \in \overleftrightarrow{RS},\ F \in \overleftrightarrow{TU} \text{ and } \overleftrightarrow{PQ} \parallel \overleftrightarrow{EF},\ \overleftrightarrow{RS} \parallel \overleftrightarrow{DF},\ \overleftrightarrow{TU} \parallel \overleftrightarrow{DE}.$$

Show that

$$\frac{PQ + RS - TU}{AB} + \frac{RS + TU - PQ}{BC} + \frac{TU + PQ - RS}{AC} = 2\sqrt{2}$$

if and only if the circumcenter of $\triangle ABC$ lies on the incircle of $\triangle ABC$.

Solution by the Fejéntaláltuka Szeged Problem Solving Group, University of Szeged, Szeged, Hungary.



Let O and I be the circumcenter and the incenter of $\triangle ABC$. Then Euler's theorem states that $OI^2 = R(R-2r)$, where R and r are the circumradius and the inradius of the triangle, respectively. Now O lies on the incircle if and only if $R(R-2r) = r^2$, which is equivalent to $\left(\frac{r}{R}\right)^2 + 2\frac{r}{R} - 1 = 0$. Therefore, $\frac{r}{R} = \sqrt{2} - 1$ since $\frac{r}{R} > 0$. Since $\cos\alpha + \cos\beta + \cos\gamma = 1 + \frac{r}{R}$ in any triangle, we can reduce the original condition to $\cos\alpha + \cos\beta + \cos\gamma = \sqrt{2}$ where α, β and γ are the angles of $\triangle ABC$.

We have

$$\begin{aligned} DE^2 &\overset{(1)}{=} CD^2 + CE^2 - 2CD \cdot CE \cos\gamma \\ &\overset{(2)}{=} (CA\cos\gamma)^2 + (BC\cos\gamma)^2 - 2(CA\cos\gamma)(BC\cos\gamma)\cos\gamma \\ &= (CA^2 + BC^2 - 2CA \cdot BC\cos\gamma)\cos^2\gamma \overset{(3)}{=} AB^2\cos^2\gamma, \end{aligned}$$

where (1) and (3) are the result of the law of cosines applied to $\triangle CDE$ and $\triangle ABC$, respectively, and (2) follows from the fact that CD and CE are altitudes. Since $\triangle ABC$ is acute, $\cos\alpha > 0$, so

$$DE = AB\cos\gamma, \text{ and similarly } EF = BC\cos\alpha \text{ and } FD = CA\cos\beta. \qquad (1)$$

Because $\angle BFC$ and $\angle BEC$ are right angles, E and F lie on the circle with diameter BC, thus $BCEF$ is a cyclic quadrilateral. Hence, $m\angle EFA = 180^\circ - m\angle BFE = m\angle ECB = \gamma$ and $m\angle AEF = 180^\circ - m\angle FEC = m\angle CBF = \beta$. We can similarly see that $m\angle FDB = m\angle CDE = \alpha$, $m\angle DEC = \beta$ and $m\angle BFD = \gamma$. Since $PQ \parallel EF$, $RS \parallel FD$ and $TU \parallel DE$ we have

$$\begin{gathered} m\angle RSB = m\angle FDB = \alpha = m\angle CDE = m\angle CTU, \\ m\angle AQP = m\angle AEF = \beta = m\angle DEC = m\angle TUC, \\ m\angle BRS = m\angle BFD = \gamma = m\angle EFA = m\angle QPA. \end{gathered}$$

Therefore, the following triangles are all isosceles (because they all have two congruent angles): $\triangle DQE$, $\triangle EDS$, $\triangle ERF$, $\triangle FEU$, $\triangle FTD$, and $\triangle DFP$. Therefore,

$$DQ = DE = ES, RE = EF = FU, \text{ and } TF = FD = PD,$$

which (by (1)) leads to

$$\begin{aligned} PQ &= PD + DQ = FD + DE = CA\cos\beta + AB\cos\gamma, \\ RS &= RE + ES = EF + DE = BC\cos\alpha + AB\cos\gamma, \\ TU &= TF + FU = FD + EF = CA\cos\beta + BC\cos\alpha. \end{aligned}$$

Substituting these into our original statement, we get that

$$\frac{PQ+RS-TU}{AB} + \frac{RS+TU-PQ}{BC} + \frac{TU+PQ-RS}{CA} = 2\left(\cos\gamma + \cos\alpha + \cos\beta\right).$$

In the first paragraph, we showed that the right side of the last equation equals $2\sqrt{2}$ if and only if the circumcenter lies on the incircle, which is exactly what we wanted to prove.

Also solved by Michel Bataille (France), Kyle Gatesman, Volkhard Schindler (Germany), Albert Stadler (Switzerland), and the proposer.

Answers

Solutions to the Quickies from page 407.

A1123. We will need the fact that if f satisfies P_2, then

$$f\left(\frac{n}{n+1}A_1 + \frac{1}{n+1}A_2\right) = \frac{n}{n+1}f(A_1) + \frac{1}{n+1}(A_2). \tag{1}$$

We proceed by induction. When $n = 1$ this is just condition P_2. Let

$$X = \frac{n+1}{n+2}A_1 + \frac{1}{n+2}A_2 \quad \text{and} \quad Y = \frac{1}{n+2}A_1 + \frac{n+1}{n+2}A_2.$$

We have

$$X = \frac{n}{n+1}A_1 + \frac{1}{n+1}Y \quad \text{and} \quad Y = \frac{1}{n+1}X + \frac{n}{n+1}A_2,$$

so, by the induction hypothesis,

$$f(X) = \frac{n}{n+1}f(A_1) + \frac{1}{n+1}f(Y) \quad \text{and} \quad f(Y) = \frac{1}{n+1}f(X) + \frac{n}{n+1}f(A_2).$$

Eliminating $f(Y)$ gives the desired result.

We will now use induction to show that $P_2 \Rightarrow P_n$ for all $n \geq 2$, the case $n = 2$ being immediate. Let

$$G = \frac{1}{n+1}\sum_{i=1}^{n+1} A_i \quad \text{and} \quad G' = \frac{1}{n}\sum_{i=1}^{n} A_i.$$

Hence,

$$G = \frac{n}{n+1}G' + \frac{1}{n+1}A_{n+1}.$$

Therefore,

$$f(G) = \frac{n}{n+1}f(G') + \frac{1}{n+1}f(A_{n+1}) \text{ (by (1))}$$

$$= \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n} f(A_i)\right) + \frac{1}{n+1} f(A_{n+1}) \text{ (by induction)}$$

$$= \frac{1}{n+1}\sum_{i=1}^{n+1} f(A_i),$$

as desired.

To show that $P_n \Rightarrow P_2$, let $M = (A_1 + A_2)/2$. Then,

$$f\left(\frac{1}{n}\left(M + M + \sum_{i=3}^{n} A_i\right)\right) = f\left(\frac{1}{n}\left(A_1 + A_2 + \sum_{i=3}^{n} A_i\right)\right)$$

$$\frac{1}{n}\left(2f(M) + \sum_{i=3}^{n} f(A_i)\right) = \frac{1}{n}\left(f(A_1) + f(A_2) + \sum_{i=3}^{n} f(A_i)\right) \text{ (by } P_n\text{)},$$

so $f(M) = (f(A_1) + f(A_2))/2$ as we wished to show.

A1124. The answer is yes. Note that if $1/F_n < x \le 1/F_{n-1}$ with $n \ge 3$, then

$$0 < x - \frac{1}{F_n} \le \frac{1}{F_{n-1}} - \frac{1}{F_n} \le \frac{2}{F_n} - \frac{1}{F_n} = \frac{1}{F_n}.$$

For $y \le 1$, let $g(y)$ denote the unique positive integer m such that

$$\frac{1}{F_m} < y \le \frac{1}{F_{m-1}}.$$

The relation above shows that $g(x - 1/F_n) > n$. Now take $x_1 = 1, n_1 = 3$ and recursively define

$$x_{k+1} = x_k - \frac{1}{F_{n_k}} \quad \text{and} \quad n_{k+1} = g(x_{k+1}).$$

This gives

$$1 = \frac{1}{F_3} + \frac{1}{F_4} + \frac{1}{F_6} + \frac{1}{F_9} + \frac{1}{F_{11}} + \frac{1}{F_{21}} + \frac{1}{F_{23}} + \cdots.$$

Note that the analogous result holds for any a such that

$$0 < a \le \sum_{n=1}^{\infty} \frac{1}{F_n} = 3.35988\ldots.$$

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.

Winston, Wayne L., Scott Nestler, and Konstantinos Pelechrinis, *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports*, 2nd ed., Princeton University Press, 2022; xxi + 584 pp, $24.95(P). ISBN 978-0-691-17762-5.

This book is a huge and highly interesting compendium of "sports analytics," the application of mathematical and statistical techniques to evaluating, rating, predicting performance, and betting on sports teams and players. An associated website contains datasets and programs. The coverage is astonishingly broad and thorough, encompassing linear regression (and ridge regression), Monte Carlo simulation, Poisson events, conditional probability, game theory, Bayesian statistics, and numerous models. So too are the applications, to baseball (when does it pay to bunt?), football (go for 1 point or 2 points after a touchdown?), and basketball (is there a "hot hand" phenomenon?). Soccer, volleyball, hockey, and golf get some coverage, too. This second edition contains 17 new chapters plus newer data than the first edition in 2012.

Williams, G. Arnell, *Algebra the Beautiful: An Ode to Math's Least-Loved Subject*, Basic Books, 2022; xiii + 395 pp, $32(P). ISBN 978-1-5416-0068-3.

At last! After all the attacks on the need for students to study algebra, the subject finally has the benefit of a gifted defender. The book's title is an apt summary of the author's enthusiasm, which he communicates through humanistic, aesthetic, and conceptual approaches. Algebra's rules store ideas, its word problems are "numerical symphonies," and it teaches how to transform relationships between quantities into symbolic relationships that can be "maneuvered" and analyzed. "[T]his book aims to inform, bolster, and inspire your mathematical soul." What could be better!?

Bressoud, David, Decades later, problematic role of calculus as gatekeeper to opportunity persists, https://www.utdanacenter.org/blog/decades-later-problematic-role-calculus-gatekeeper-opportunity-persists.

De Loera, Jesus A., and Francis Su, Calculus isn't the only option. Let's broaden and update the current math curriculum, https://www.sacbee.com/opinion/op-ed/article260529232.html.

Bressoud points out that of those who take calculus in high school, 30% retake Calculus I in college, only 20% skip past Calculus I in college, and—"[m]ost disturbing of all"—30–35% are placed into precalculus/college algebra. In addition, White students are much more likely to take the AP Calculus exam, and also to pass it, than Black and Latino students. Meanwhile, calculus is not available to many students in majority-minority high schools. Given the statistics noted, that might seem like a blessing, except that students taking calculus for the first time in college are competing against others who have already had it. In part, it is the "competing" that makes calculus a gatekeeper. De Loera and Su react to controversy over proposed revisions to the California state initiative on school curriculum. They urge "advanced electives"—beyond geometry and two years of algebra—in statistics, data science, probability, discrete mathematics—courses "aligned with [students'] aspirations ... to thrive as scholars and professionals." That sounds grand, but many students do not know themselves or their talents and potential well enough to know what to aspire to (and few aspire to be scholars). De Loera and Su do not mention that many schools do not have teachers who can do those courses, nor summer programs (as supported by the NSF years ago) to prepare them to do so.

Math. Mag. **95** (2022) 415–416. doi:10.1080/0025570X.2022.2094680. © Mathematical Association of America

Barrow-Green, June, Jeremy Gray, and Robin Wilson, *The History of Mathematics: A Source-Based Approach*, vol. 2, MAA Press, 2022; xiv + 687 pp, $89 ($66.75 AMS or MAA member), $89(E) ($66.75 AMS or MAA member). ISBN 978-1-4704-4382-5, 978-1-4704-5693-1.

This volume picks up where the authors left off in their first volume (2019), from about 1650 to the start of the 20th century. The intent of these books is to ask: Who did the mathematics, and why? How was the work disseminated (or not)? How did it emerge from the culture of the time, and why is it still relevant today? The approach is to use extensive quotations from original sources as the best way to answer some of those questions. The book concludes uniquely with dozens of suggested essay exercises that are "firmly historical, rather than primarily mathematical"; some call for supporting or contesting claims about mathematical discoveries. The two volumes were designed for a year-long course, and they provide excellent material for a senior-level course to help students survey the mathematics that they have learned and put it into cultural and scientific context.

Clayton, Aubrey, Damned lies: Eugenics and the myth of objectivity in statistics, *Nautilus* online issue 092, chapter 2; print issue 33, 19–35; https://nautil.us/issue/92/frontiers/how-eugenics-shaped-statistics.

"[P]urging statistics of the ghosts of its eugenicist past is not a straightforward proposition," because so much of statistics arose from Galton, Pearson, and Fisher. This article details their eugenicist legacy and goes further to suggest that current controversies in statistics, such as over significance testing, trace back to the biases of those three. "[S]tatistics needs to free itself from the ideal of being perfectly objective" and strive for "moral objectivity."

Wolfram, Stephen, How inevitable is the concept of numbers?, https://writings.stephenwolfram.com/2021/05/how-inevitable-is-the-concept-of-numbers/.

"Why do we use numbers so much? Is it something about the world? Or is it more something about us? ... Are numbers even inevitable in mathematics?" Author Wolfram philosophizes, imagining aliens arriving on a starship having views of mathematics "incoherently different from our own" without "any of the familiar features of our typical view of mathematics, like numbers." That would indeed be strange and perhaps incomprehensible to us, since Wolfram feels that "numbers seem to be inextricably connected to core aspects of our existence."

Kendig, Keith, *A Gateway to Number Theory: Applying the Power of Algebraic Curves*, MAA Press, 2021; xv + 207 pp, $59(P) ($44.25 MAA or AMS member). ISBN 978-1-4704-5622-1.

This book solves homogeneous diophantine equations in three variables (e.g., $a^2 + b^2 = 2c^2$). The technique is to set $x = a/c$ and $y = b/c$ to arrive at a polynomial $p(x, y) = 0$ and then investigate rational points on the corresponding curve. The main investigation is of equations of degree 3, which correspond to elliptic curves. This pleasant and accessible journey continues into curves over $\mathbb{C}$ and the topology of algebraic curves. There are plenty of concrete examples, plus code (in GeoGebra, Maple, and Mathematica) for creating animations and solving the equations.

Long, Mark, Liberal arts, meet computation: A Wolfram Community introduction, https://blog.wolfram.com/2022/05/20/liberal-arts-meet-computation-a-wolfram-community-introduction/.

Does computation offer anything to non-science liberal arts students? Author Long provides examples of "Wolfram Community posts that exemplify classical liberal arts subjects," as manifested in the categories of the trivium (grammar, logic, rhetoric): an implementation of Wordle, ternary logic tables, text-image analysis; and of the quadrivium (arithmetic, geometry, music and astronomy): invention of imaginary numbers, computational art contest, playing with musical scales and composition, and watching the path of an asteroid.

Rosenhouse, Jason, The failures of mathematical anti-evolutionism, *Skeptical Inquirer* 46(3) (May/June 2022) 41–44, https://skepticalinquirer.org/2022/05/the-failures-of-mathematical-anti-evolutionism/.

In recent years, creationists have manufactured "mathematical" arguments against the possibility of evolution, from probability, information theory, and combinatorial search. Author Rosenhouse explains the arguments and why they fail. [Disclosure: I picked this article for the Reviews column before noticing the name of the author, who is the editor of This Magazine.]