LETTER FROM THE EDITOR

In the first article of this issue, Kenneth Williams recalls some history of Liouville's work on quadratic forms. Then, Williams uses some results of Gauss to show that Liouville's work can be proved from a result of Jacobi's.

Gauss makes an appearance in the second article of this issue, too. Easter this year is on April 21. In 1800, Gauss wrote about calculating the Easter date and, in this issue, Donald Teets summarizes and explains Gauss' calculation of the Easter date, as well as Gauss' 1816 correction of an error in his calculation. Somewhere in there is a joke on Easter bonnets and the Gauss–Bonnet theorem. Also, this month's cover art is motivated by Teets' article and contains a grid of possible Easter dates. Read more about it on the inside front cover.

If you only have candles to heat your bedroom, how should you position the candles to warm the room as efficiently as possible? What if your room is a torus? Surprisingly, in their article, Florian Pausinger and Stefan Steinerberger use number theory to answer the question.

The Gion shrine problem is a geometry problem from eighteenth-century Japan. In their article, David Clark, J. Arias de Reyna, and Noam Elkies provide a new solution to the problem that, like the classical Japanese solution, involves solving a degree-ten polynomial. However, their polynomial is easier to write down. In their analysis, they use a result of Fermat to show that no rational solution to the Gion shrine problem exists.

I have always enjoyed the combinatorics of determining the number of nonnegative integer solutions to equations like $x_1 + \cdots + x_n = k$ with inequality constraints on the x_i. Hideo Hirose determines the number of solutions to $x_1 + \cdots + x_n \equiv k \pmod{p}$ using probability theory.

In the next article, Edwin O'Shea examines divisibility tests, the algorithms used to determine if one integer is divisible by another. He uses only basic divisibility properties to unify previously examined tests based on trimming and summing.

Weak induction and strong induction are two well-known proof techniques for statements indexed by the natural numbers. What about real induction? In the final article, Pete Clark provides an instructor's guide to real induction, a proof technique, that is, applicable to statements indexed by an interval on the real line. Clark applies real induction to prove basic results in real analysis and topology.

Throughout this issue are three proofs without words. In the first, Günhan Caglayan provides an identity on the difference between triangular numbers. In the second, Rex Wu considers arctangent identities involving 2 and the golden ratio. In the third, Steve Butler shows a relationship between independent sets in grid graphs and the tiling of Aztec diamonds. There are also two Math Bites in this issue: Tristen Pankake-Sieminski and Raymond Viglione provide a math bite on averages of averages and Konstantinos Gaitanas uses the discriminant of a quadratic equation to prove the root-mean square–arithmetic mean inequality.

As with every issue of 2019, David Nacin provides a TRIBUS puzzle. And, as with every issue of the MAGAZINE, the Problems section includes new problems to get you thinking and the Reviews section provides summaries of some recent articles and books.

Michael A. Jones, Editor

Some Formulas of Liouville in the Spirit of Gauss

KENNETH S. WILLIAMS
Carleton University
Ottawa, ON, Canada
kennethwilliams@cunet.carleton.ca

The French mathematician Joseph Liouville (1809–1882) made major contributions to a great many different areas of mathematics, including complex analysis, differential geometry, mathematical physics and the theory of integration in finite terms. At least six theorems are named after him [**13**, p. vii]. He is particularly known for Liouville's theorem (a bounded entire function is a constant) and for Liouville numbers (the first known examples of transcendental numbers).

At the age of 47, Liouville began his research in the theory of numbers which was to occupy him for the rest of his life. Liouville's objective was to give some basic elementary principles from which arithmetic formulas proved by his colleagues Jacobi, Hermite, Kronecker and others would follow. There is no doubt that he discovered such underlying principles from which came numerous results. Unfortunately, probably because of the pressure on his time from his administrative responsibilities, Liouville published his results, almost always without proof, in a series of eighteen articles in his journal *Journal de Mathématiques Pures et Appliquées* and the application of them to quadratic forms in a series of ninety notes in the same journal (as one of the referees of this article put it succinctly in his/her report "without editorial interference"). Later, other mathematicians proved Liouville's arithmetic formulas though not always in the way Liouville had in mind. For example Pepin [**14**] used trigonometric identities to prove Liouville-type arithmetic identities from which he deduced many of Liouville's formulas involving quadratic forms. Even today we cannot be absolutely sure what Liouville's arithmetic methods were.

Liouville's secrecy concerning his methods resulted in him not receiving the recognition he deserved for his work on quadratic forms, even though he was a pioneer in this area. Liouville's interest was in determining the representation numbers of quadratic forms $c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_4^2$, where c_1, c_2, c_3, c_4 are positive integers; that is, he sought formulas for

$$N(c_1, c_2, c_3, c_4; n) = \tag{1}$$
$$|\{(x_1,x_2, x_3, x_4) \in \mathbb{Z}^4 \mid n = c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_4^2\}|,$$

valid for every $n \in \mathbb{N}$. When $n = 0$ the only solution to $0 = c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_4^2$ is $x_1 = x_2 = x_3 = x_4 = 0$ because c_1, c_2, c_3, c_4 are all positive, so $N(c_1, c_2, c_3, c_4; 0) = 1$. Perhaps Liouville was motivated by one of the consequences of Jacobi's monumental work [**7**] of 1829 on elliptic and theta functions, namely that the number of representations of a positive integer n as a sum of four squares of integers is 8 times the sum of the positive divisors of n which are not multiples of 4, see [**5**, **6**], that is

$$N(1, 1, 1, 1; n) = 8a_4(n), \text{ where } a_k(n) = \sum_{\substack{d|n \\ k\nmid d}} d, \text{ for } k \in \mathbb{N}. \tag{2}$$

Math. Mag. **92** (2019) 83–90. doi:10.1080/0025570X.2019.1543989
MSC: Primary 11E25, Secondary 11D85; 11E20

If m is a positive rational number but not a positive integer we define $a_k(m) = 0$, so that for example $a_4(n/8) = 0$ for every $n \in \mathbb{N}$ which is not a multiple of 8. If $k, l, n \in \mathbb{N}$, by splitting the sum $a_{kl}(n)$ into two sums according to whether a divisor d of n is a multiple of k or not, we deduce that

$$a_{kl}(n) = ka_l(n/k) + a_k(n),$$

so that we have the useful formula

$$a_l(n/k) = \frac{1}{k}a_{kl}(n) - \frac{1}{k}a_k(n). \tag{3}$$

The purely arithmetical nature of Jacobi's formula in equation (2) perhaps suggested to Liouville that he look for other quadratic forms $c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_4^2$ $(c_1, c_2, c_3, c_4 \in \mathbb{N})$ for which the representation numbers defined in equation (1) can be expressed in a purely arithmetic way and to give proofs of them based on arithmetic principles. Today the theory of modular forms explains when this is possible and when it is not.

How did Jacobi's arithmetic formula of equation (2) for $N(1, 1, 1, 1; n)$ $(n \in \mathbb{N})$ follow from Jacobi's analytic work? What Jacobi proved was the identity

$$(1 + 2x + 2x^4 + 2x^9 + \cdots)^4 = 1 + 8\left(\frac{x}{1-x} + \frac{2x^2}{1+x^2} + \frac{3x^3}{1-x^3} + \cdots\right). \tag{4}$$

The reader will find this identity (with x replaced by q) in Jacobi's *Gesammelte Werke* [**8**, Vol. 1, p. 239]. The two series in equation (4) converge for all complex numbers x satisfying $|x| < 1$, as do all the series considered in this article. The left-hand side of equation (4) is

$$\left(\sum_{m=-\infty}^{\infty} x^{m^2}\right)^4 = \sum_{(m_1,m_2,m_3,m_4)\in\mathbb{Z}^4} x^{m_1^2+m_2^2+m_3^2+m_4^2} = \sum_{n=0}^{\infty} \sum_{\substack{(m_1,m_2,m_3,m_4)\in\mathbb{Z}^4 \\ m_1^2+m_2^2+m_3^2+m_4^2=n}} x^n,$$

so that

$$\left(\sum_{m=-\infty}^{\infty} x^{m^2}\right)^4 = \sum_{n=0}^{\infty} N(1, 1, 1, 1; n)x^n, \tag{5}$$

proving that $\left(\sum_{m=-\infty}^{\infty} x^{m^2}\right)^4$ is the generating function of the $N(1, 1, 1, 1; n)$. The right-hand side of equation (4) can be shown to be

$$1 + 8\sum_{n=1}^{\infty} \sum_{\substack{d|n \\ 4\nmid d}} dx^n = 1 + 8\sum_{n=1}^{\infty} a_4(n)x^n$$

using the well-known geometric series

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

The reader can find the details of this calculation in [**3**, p. 61]. Equating coefficients of x^n for $n \in \mathbb{N}$, we obtain the formula in equation (2).

It is perhaps not so well-known that there are other forms $c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_4^2$ whose representation numbers can be expressed in terms of arithmetic sums of the type $a_k(n)$ for certain values of $k \in \mathbb{N}$. We give five such forms for which this is the case. Formulas for the representation numbers of these forms were given by Liouville [9–12] in the 1860s. Many proofs of these formulas are known, see [1] for some references. The nature of these proofs varies from proofs using elementary methods to proofs using theta function identities to modern proofs using modular forms. Our contribution will be to show that they can all be proved in a simple systematic way from Jacobi's theorem, expressed in the form with equations (2) and (5) combined,

$$\left(\sum_{m=-\infty}^{\infty} x^{m^2}\right)^4 = 1 + \sum_{n=1}^{\infty} 8a_4(n)x^n, \tag{6}$$

by using some results of Gauss.

Liouville's formulas

For $n \in \mathbb{N}$,

$$\begin{aligned}
N(1,1,2,2;n) &= 2a_2(n) - 2a_4(n) + 4a_8(n),\\
N(1,2,2,4;n) &= a_2(n) - a_8(n) + 2a_{16}(n),\\
N(1,1,1,4;n) &= 2\alpha(n)a_2(n) + 5a_4(n) - 3a_8(n) + 2a_{16}(n),\\
N(1,1,4,4;n) &= (2\alpha(n)+1)a_2(n) - a_8(n) + 2a_{16}(n),\\
N(1,4,4,4;n) &= (\alpha(n)+3/2)a_2(n) - (5/2)a_4(n) + 2a_{16}(n),
\end{aligned}$$

where

$$\alpha(n) = \begin{cases} 0 & \text{if } n \equiv 0 \pmod 2,\\ (-1)^{(n-1)/2} & \text{if } n \equiv 1 \pmod 2. \end{cases} \tag{7}$$

We now describe some results of Gauss that we shall use. Gauss [4, p. 465] considered the infinite series

$$P(x) = 1 + 2x + 2x^4 + \cdots \quad \text{and} \quad Q(x) = 1 - 2x + 2x^4 - \cdots, \tag{8}$$

and found some of their properties. Clearly $Q(-x) = P(x)$. The functions P and Q are now called theta functions and the variable x is often written as q but we will use Gauss' notation.

Gauss' relationships involving $P(x)$ and $Q(x)$

Gauss proved among others the following relationships, see [4, formulas (13), (16), (19), (20, (21); pp. 466–467]. For $|x| < 1$,

$$P(x) + Q(x) = 2P(x^4), \tag{9}$$

$$P^2(x) + Q^2(x) = 2P^2(x^2), \tag{10}$$

$$P(x)Q(x) = Q^2(x^2), \tag{11}$$

$$P(x) + iQ(x) = (1+i)Q(ix), \tag{12}$$

$$P(x) - iQ(x) = (1-i)P(ix). \tag{13}$$

Formulas (9), (12) and (13) are easily proved by calculating $P(x) + Q(x)$, $P(x) + iQ(x)$ and $P(x) - iQ(x)$ from the series in equation (8). For completeness we prove (10) and (11). First we prove (10). We set, for $n \in \mathbb{N} \cup \{0\}$,

$$s(n) = |\{(x, y) \in \mathbb{Z}^2 \mid n = x^2 + y^2\}|,$$

so that $s(0) = 1$. Then

$$P^2(x) = \left(\sum_{m=-\infty}^{\infty} x^{m^2}\right)^2 = \sum_{(m_1,m_2)\in\mathbb{Z}^2} x^{m_1^2+m_2^2} = \sum_{n=0}^{\infty} \sum_{\substack{(m_1,m_2)\in\mathbb{Z}^2 \\ m_1^2+m_2^2=n}} x^n = \sum_{n=0}^{\infty} s(n)x^n.$$

If a and b are integers such that $a^2 + b^2 = n$, then $(a + b)^2 + (a - b)^2 = 2n$. Conversely, if c and d are integers such that $c^2 + d^2 = 2n$, then c and d have the same parity, so we can define integers a and b by $a = (c + d)/2$ and $b = (c - d)/2$, so that

$$a^2 + b^2 = \left(\frac{c+d}{2}\right)^2 + \left(\frac{c-d}{2}\right)^2 = \frac{c^2+d^2}{2} = n.$$

Thus there is a bijection between $(a, b) \in \mathbb{Z}^2$ with $a^2 + b^2 = n$ and $(c, d) \in \mathbb{Z}^2$ with $c^2 + d^2 = 2n$. Hence $s(n) = s(2n)$. Then

$$\begin{aligned} P^2(x) + Q^2(x) = P^2(x) + P^2(-x) &= \sum_{n=0}^{\infty} s(n)x^n + \sum_{n=0}^{\infty} s(n)(-x)^n \\ &= 2\sum_{n=0}^{\infty} s(2n)x^{2n} = 2\sum_{n=0}^{\infty} s(n)x^{2n} = 2P^2(x^2), \end{aligned}$$

which is equation (10). We now prove equation (11). By appealing to equations (9) and (10), we have

$$\begin{aligned} 4P^2(x^4) = (P(x) + Q(x))^2 &= P^2(x) + Q^2(x) + 2P(x)Q(x) \\ &= 2P^2(x^2) + 2P(x)Q(x) = 4P^2(x^4) - 2Q^2(x^2) + 2P(x)Q(x), \end{aligned}$$

which gives equation (11).

By exactly the same kind of calculation as the one we did to show that $P^4(x)$ is the generating function of $N(1, 1, 1, 1; n)$ for $n \in \mathbb{N} \cup \{0\}$, we find that the generating function of $N(c_1, c_2, c_3, c_4; n)$ is $P(x^{c_1})P(x^{c_2})P(x^{c_3})P(x^{c_4})$. Thus in order to find the generating functions of $N(1, 1, 2, 2; n)$, $N(1, 2, 2, 4; n)$, $N(1, 1, 1, 4; n)$, $N(1, 1, 4, 4; n)$ and $N(1, 4, 4, 4; n)$ we must find the power series expansions in powers of x of $P^2(x)P^2(x^2)$, $P(x)P^2(x^2)P(x^4)$, $P^3(x)P(x^4)$, $P^2(x)P^2(x^4)$ and $P(x)P^3(x^4)$, respectively. We begin by using Gauss' formulas to replace each occurrence of $P^2(x^2)$ by $\frac{1}{2}P^2(x) + \frac{1}{2}Q^2(x)$ and each occurrence of $P(x^4)$ by $\frac{1}{2}P(x) + \frac{1}{2}Q(x)$ in each of these five products to show that each of them can be expressed as a rational linear combination of $P^4(x)$, $P^3(x)Q(x)$, $P^2(x)Q^2(x)$, $P(x)Q^3(x)$ and $Q^4(x)$. Although $Q^4(x)$ does not actually appear in these formulas it is convenient to treat it along with the others as it will be used later in this article. We obtain

$$\begin{aligned} P^2(x)P^2(x^2) &= \frac{1}{2}P^4(x) + \frac{1}{2}P^2(x)Q^2(x), \\ P(x)P^2(x^2)P(x^4) &= \frac{1}{4}P^4(x) + \frac{1}{4}P^3(x)Q(x) + \frac{1}{4}P^2(x)Q^2(x) + \frac{1}{4}P(x)Q^3(x), \end{aligned}$$

$$P^3(x)P(x^4) = \frac{1}{2}P^4(x) + \frac{1}{2}P^3(x)Q(x),$$

$$P^2(x)P^2(x^4) = \frac{1}{4}P^4(x) + \frac{1}{2}P^3(x)Q(x) + \frac{1}{4}P^2(x)Q^2(x), \text{ and}$$

$$P(x)P^3(x^4) = \frac{1}{8}P^4(x) + \frac{3}{8}P^3(x)Q(x) + \frac{3}{8}P^2(x)Q^2(x) + \frac{1}{8}P(x)Q^3(x).$$

To find the power series expansions of $P^2(x)P^2(x^2)$, $P(x)P^2(x^2)P(x^4)$, $P^3(x)P(x^4)$, $P^2(x)P^2(x^4)$, and $P(x)P^3(x^4)$, we must therefore determine the power series expansions of $P^{4-l}(x)Q^l(x)$ for $l = 0, 1, 2, 3, 4$. These are given in the following theorem. The first of these is Jacobi's theorem and we deduce the remaining formulas from it.

Theorem.

$$P^4(x) = 1 + \sum_{n=1}^{\infty} 8a_4(n)x^n, \tag{14}$$

$$P^3(x)Q(x) = 1 + \sum_{n=1}^{\infty} (4a_{16}(n) - 6a_8(n) + 2a_4(n) + 4\alpha(n)a_2(n))x^n, \tag{15}$$

$$P^2(x)Q^2(x) = 1 + \sum_{n=1}^{\infty} (8a_8(n) - 12a_4(n) + 4a_2(n))x^n, \tag{16}$$

$$P(x)Q^3(x) = 1 + \sum_{n=1}^{\infty} (4a_{16}(n) - 6a_8(n) + 2a_4(n) - 4\alpha(n)a_2(n))x^n, \tag{17}$$

$$Q^4(x) = 1 + \sum_{n=1}^{\infty} (16a_4(n) - 24a_2(n))x^n, \tag{18}$$

where $a_k(n)$ is defined in equation (2) and $\alpha(n)$ in equation (7).

Proof. Equation (14) is Jacobi's theorem from equation (6). It is stated as part of this theorem for convenience and completeness.

Now we prove equation (18). Replacing x by $-x$ in equation (14), we obtain

$$Q^4(x) = P^4(-x) = 1 + \sum_{n=1}^{\infty} 8(-1)^n a_4(n)x^n.$$

We now show that

$$(-1)^n a_4(n) = 2a_4(n) - 3a_2(n) \tag{19}$$

to complete the proof. If n is odd then $a_4(n) = a_2(n)$ and equation (19) follows. If n is even, from (3) with $(k, l) = (2, 2)$, we have

$$a_2(n/2) = \frac{1}{2}a_4(n) - \frac{1}{2}a_2(n). \tag{20}$$

As n is even we have $a_2(n/2) = a_2(n)$ so $a_4(n) = 3a_2(n)$ and equation (19) follows.

Next we prove equation (16). By equations (11) and (18) we have

$$P^2(x)Q^2(x) = Q^4(x^2) = 1 + \sum_{n=1}^{\infty} (16a_4(n/2) - 24a_2(n/2))x^n. \tag{21}$$

By (3) with $(k, l) = (2, 4)$ we have

$$a_4(n/2) = \frac{1}{2}a_8(n) - \frac{1}{2}a_2(n). \tag{22}$$

Using equations (20) and (22) in equation (21) we obtain equation (16).

Finally, we prove equations (15) and (17). First we determine the power series expansion of $\frac{1}{2}(P^3(x)Q(x) + P(x)Q^3(x))$. Appealing to Gauss' formulas in equations (10) and (11), we deduce

$$\begin{aligned}\frac{1}{2}(P^3(x)Q(x) + P(x)Q^3(x)) &= \frac{1}{2}(P^2(x) + Q^2(x))P(x)Q(x)\\ &= P^2(x^2)Q^2(x^2) = Q^4(x^4).\end{aligned}$$

From equation (18) of the theorem, with x replaced by x^4, we obtain

$$Q^4(x^4) = 1 + \sum_{n=1}^{\infty}(16a_4(n/4) - 24a_2(n/4))x^n. \tag{23}$$

Taking $(k, l) = (4, 2)$ and $(4, 4)$ in (3), we obtain

$$a_2(n/4) = \frac{1}{4}a_8(n) - \frac{1}{4}a_4(n), \quad a_4(n/4) = \frac{1}{4}a_{16}(n) - \frac{1}{4}a_4(n). \tag{24}$$

Using equation (24) in equation (23), we deduce

$$\frac{1}{2}(P^3(x)Q(x) + P(x)Q^3(x)) = 1 + \sum_{n=1}^{\infty}(4a_{16}(n) - 6a_8(n) + 2a_4(n))x^n. \tag{25}$$

Next we determine the power series expansion of $\frac{1}{2}(P^3(x)Q(x) - P(x)Q^3(x))$. Factoring we have

$$P^3(x)Q(x) - P(x)Q^3(x) = (P(x)Q(x))(P(x) + Q(x))(P(x) - Q(x)). \tag{26}$$

We express each of the bracketed factors on the right-hand side of equation (26) in terms of $P(ix)$ and $Q(ix)$. By equations (10) and (11), we have

$$P(x)Q(x) = Q^2(x^2) = P^2(-x^2) = \frac{1}{2}(P^2(ix) + Q^2(ix)).$$

From equation (9) we see that

$$P(x) + Q(x) = 2P(x^4) = P(ix) + Q(ix).$$

From equations (12) and (13) we have

$$P(x) - Q(x) = -i(P(ix) - Q(ix)).$$

Thus the right-hand side of equation (26) is

$$\frac{-i}{2}(P^4(ix) - Q^4(ix)) = \frac{1}{2i}(P^4(ix) - P^4(-ix)) = \sum_{n=1}^{\infty} 8\alpha(n)a_4(n)x^n,$$

where we appealed to equation (14) for the power series expansions of $P^4(ix)$ and $P^4(-ix)$. Thus

$$\frac{1}{2}(P^3(x)Q(x) - P(x)Q^3(x)) = \sum_{n=1}^{\infty} 4\alpha(n)a_4(n)x^n. \tag{27}$$

Adding and subtracting equations (25) and (27), as $\alpha(n)a_4(n) = \alpha(n)a_2(n)$, we obtain equations (15) and (17). ■

Proof of Liouville's formulas

Liouville's formulas follow by putting the power series expansions of $P^{4-j}(x)Q^j(x)$ $(j = 0, 1, 2, 3, 4)$ given in the theorem into the formulas expressing the five products $P^2(x)P^2(x^2), \ldots, P(x)P^3(x^4)$ as linear combinations of $P^{4-j}(x)Q^j(x)$ $(j = 0, 1, 2, 3, 4)$, and then equating the coefficients of x^n $(n \in \mathbb{N})$. We just give the details for $N(1, 1, 1, 4; n)$. We have

$$\begin{aligned}\sum_{n=0}^{\infty} N(1,1,1,4;n)x^n &= P^3(x)P(x^4) = \frac{1}{2}P^4(x) + \frac{1}{2}P^3(x)Q(x)\\ &= \frac{1}{2}\left(1 + \sum_{n=1}^{\infty} 8a_4(n)x^n\right)\\ &\quad + \frac{1}{2}\left(1 + \sum_{n=1}^{\infty}(4a_{16}(n) - 6a_8(n) + 2a_4(n) + 4\alpha(n)a_2(n))x^n\right)\\ &= 1 + \sum_{n=1}^{\infty}(2\alpha(n)a_2(n) + 5a_4(n) - 3a_8(n) + 2a_{16}(n))x^n,\end{aligned}$$

and equating coefficients of x^n for $n \in \mathbb{N}$, we obtain the third of Liouville's five formulas. The remaining four formulas follow similarly.

The ideas of this article can be extended to express the representation numbers of the ten forms $x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_4^2$ for $(c_2, c_3, c_4) = (1, 1, 2)$, $(1, 1, 8)$, $(1, 2, 4)$, $(1, 4, 8)$, $(2, 2, 2)$, $(2, 2, 8)$, $(2, 4, 4)$, $(2, 8, 8)$, $(4, 4, 8)$ and $(8, 8, 8)$ in terms of the arithmetic sums $\sum_{\substack{d|n \\ d\equiv j \pmod 8}} d$ for $j = 1, 3, 5, 7$. The details are more complicated than for the forms treated in this article. Proofs of Liouville's formulas for the representation numbers of these ten forms can be found in [**2**, Theorems 5.1–5.10].

Acknowledgments. The author thanks the referees for their very valuable suggestions which enabled him to improve the presentation of this article.

REFERENCES

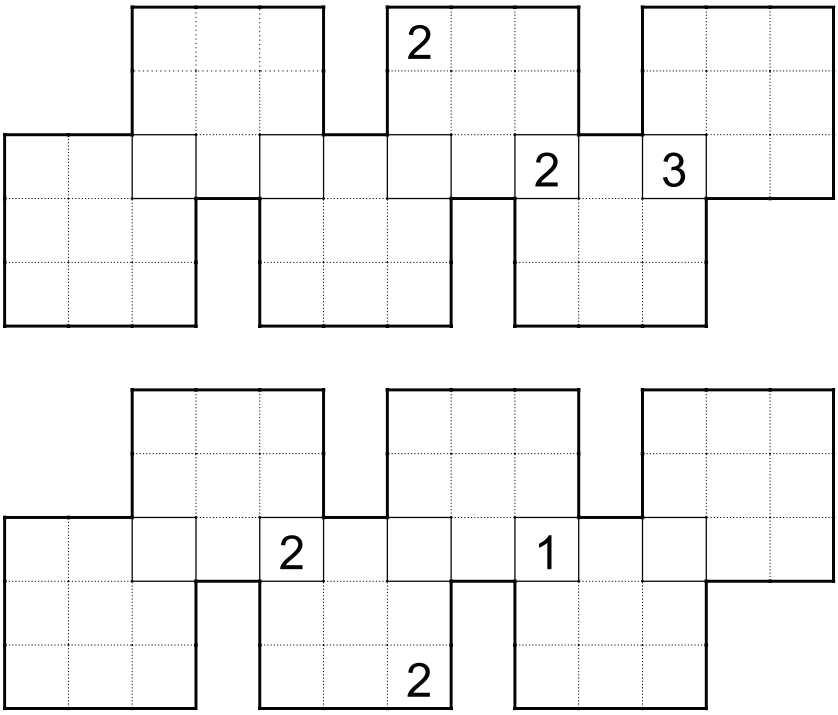
[1] Alaca, A., Alaca, Ş., Lemire, M. F., Williams, K. S. (2007). Nineteen quaternary quadratic forms. *Acta Arith.* 130: 277–310.
[2] Alaca, A., Alaca, Ş., Lemire, M. F., Williams, K. S. (2009). The number of representations of a positive integer by certain quaternary quadratic forms. *Int. J. Number Theory* 5: 13–40.
[3] Berndt, B. C. (2006). *Number Theory in the Spirit of Ramanujan*. Student Mathematics Library, Vol. 34, Providence, RI: American Math. Society.
[4] Gauss, C. F. (1900). *Werke* VIII. Leipzig: B. G. Teubner.
[5] Jacobi, C. G. J. (1828). Lettre à Legendre, le 9 septembre 1828; *Werke* I, pp. 422–425.

[6] Jacobi, C. G. J. (1828). Note sur la décompostion d'un nombre donné en quatre carrés. *J. reine angew. Math.* 3: 191; *Werke* I, p. 247.
[7] Jacobi, C. G. J. (1829). *Fundamenta Nova Theoriae Functionum Ellipticarum.* Regiomonti: Bornträger; 1829; *Werke* I, pp. 49–239.
[8] Jacobi, C. G. J. (1881). *Gesammelte Werke* I, Berlin; republished by Chelsea Publ. Co., New York, 1969.
[9] Liouville, J. (1860). Sur la forme $x^2 + y^2 + 2(z^2 + t^2)$. *J. Math. Pures Appl.* (2) 5: 269–272.
[10] Liouville, J. (1860). Sur la forme $x^2 + y^2 + 4(z^2 + t^2)$. *J. Math. Pures Appl.* (2) 5: 305–308.
[11] Liouville, J. (1861). Sur les deux formes $X^2 + Y^2 + Z^2 + 4T^2$, $X^2 + 4Y^2 + 4Z^2 + 4T^2$. *J. Math. Pures Appl.* (2) 6: 440–448.
[12] Liouville, J. (1862). Sur la forme $X^2 + 2Y^2 + 2Z^2 + 4T^2$. *J. Math. Pures Appl.* (2) 7: 1–4.
[13] Lützen, J. (1990). *Joseph Liouville 1809–1882: Master of Pure and Applied Mathematics*. New York: Springer-Verlag.
[14] Pepin, T. (1890). Sur quelques formes quadratiques quaternaires. *J. Math. Pures Appl.* (4) 6: 5–67.

Summary. We show how Liouville's formulas for the number of representations of a positive integer by the forms $x_1^2 + x_2^2 + 2x_3^2 + 2x_4^2$, $x_1^2 + 2x_2^2 + 2x_3^2 + 4x_4^2$, $x_1^2 + x_2^2 + x_3^2 + 4x_4^2$, $x_1^2 + x_2^2 + 4x_3^2 + 4x_4^2$, and $x_1^2 + 4x_2^2 + 4x_3^2 + 4x_4^2$ follow in a simple systematic way from a beautiful identity of Jacobi using some elementary relationships between the infinite series $P(x) = 1 + 2x + 2x^4 + 2x^9 + \cdots$ and $Q(x) = 1 - 2x + 2x^4 - 2x^9 + \cdots$ given by Gauss.

KENNETH S. WILLIAMS (MR Author ID: 183165) is Professor Emeritus and Distinguished Research Professor at Carleton University, Ottawa, Canada. He has been a member of the Mathematical Association of America since 1963 and has been an avid reader of *The American Mathematical Monthly* and *Mathematics Magazine* since 1959 when he entered university as a freshman.

Tribus Puzzle



How to play. Fill each of the three-by-three squares with either a 1, 2, or 3 so that each number appears exactly once in each column and row. Some cells apply to more than one square, as the squares overlap. Each of the three-by-three squares must be distinct. The solution can be found on page 98.

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

Gauss's Computation of the Easter Date

DONALD TEETS
South Dakota School of Mines and Technology
Rapid City, SD 57701
donald.teets@sdsmt.edu

In the article "Berechnung des Osterfestes (Calculation of the Easter Date)" [4] that appeared in August, 1800, Carl Friedrich Gauss offers an extraordinarily simple set of arithmetic rules for calculating the Easter date in any given year. He illustrates it with an example, concluding that "so it is, for example, for the year 4763 Easter is . . . the 7th of April . . . " This article will take a close look at Gauss's algorithm and will offer a glimpse at the history surrounding it.

There is a vast literature on the subject of Easter and the Easter date (see the references in [1], for example), so why another paper on this topic? First and foremost, this paper is *pure Gauss*: *his* method and *his* comments, offering readers a glimpse into his original work. Second, Gauss's algorithm, and variations thereof, are widely reproduced in the literature of the subject just as Gauss presented it; that is, with little or no explanation whatsoever. (For example, see the Easter date algorithm in the U. S. Naval Observatory's *Explanatory Supplement to the Astronomical Almanac* [9].) The ten modular arithmetic formulas appearing in the algorithm are simple to implement, yet so inscrutable that they beg further explanation: where do they come from, and why does the algorithm work? Third, frequent reference is made to an error in Gauss's algorithm, suggesting that his method should be discounted entirely. We shall see that Gauss is undeserving of much of the criticism surrounding this error.

All in all, despite the immense number of papers on the subject of the Easter date that have appeared through the centuries, the elegance and simplicity of Gauss's algorithm are not widely understood. And surprisingly, this very popular topic seems never to have been addressed in a century of MAA journals.

Gauss was 23 years old and struggling to find an appointment to support himself when "Berechnung des Osterfestes" was published. His biographer G. W. Dunnington [3] writes that "according to his own story his mother could not tell him the exact day on which he was born; she only knew that it was a Wednesday, eight days before Ascension Day. That started him on his search for the formula." (Note: Ascension Day is the Thursday 40 days after Easter, counting the latter as day one. In Gauss's birth year 1777, Easter fell on March 30, Ascension Day was Thursday, May 8, so Gauss calculated his birthday as April 30.)

Gauss achieved fame in the fields of number theory and astronomy, and the Easter date computation is a mix of the two. But each of these appears in a trivial way; the real complexity of the problem stems from the byzantine rules upon which the Easter date is based. In Gauss's own words, "the intention of this article is not to discuss the standard method of computing the Easter date, which one can find in any manual of mathematical chronology, and which is easy enough, once one knows the meaning and practice of the usual vocabulary of the profession, *golden number, epact, Easter limit, solar cycle, and Dominical letter*, and has the necessary assistance tables available; rather than this task, to give a means of help, independent and clear, a pure analytical solution based on the simplest calculations and operations."

Math. Mag. **92** (2019) 91–98. doi:10.1080/0025570X.2019.1549889
MSC: Primary 01A55

Definition of Easter

A popular definition of Easter is "the Sunday after the first full moon on or after the vernal equinox." But this definition comes with certain difficulties; for example, a full moon may occur just before midnight in one time zone and just after in another, thus on different days. In the Catholic Encyclopedia [7], we find the following: "Seeing, therefore, that astronomical accuracy must at some point give way to convenience . . . , the Church has drawn up a lunar calendar which maintains as close a relation with the astronomical moons as is practicable . . . ."

Quite simply, the Church uses an idealized, formulaic full moon date rather than a true astronomical full moon to determine when Easter falls. Likewise, the idealized vernal equinox March 21 is used instead of the true equinox, which may or may not occur on that date. The first of these formulaic full moons occurring on or after March 21 is known as the *paschal full moon*, and Easter is properly defined as the Sunday immediately *after* the paschal full moon (PFM). Thus, determining the date of the PFM in a given year is at the heart of the Easter algorithm. In order to make use of his original symbols, it will be convenient to introduce Gauss's algorithm before we address the PFM problem.

Gauss's Easter algorithm

Here is Gauss's Easter algorithm in his own words and symbols [4]:

Complete general rules for the calculation of the Easter date for the Julian, as well as the Gregorian Calendar.

If the result of the division of	by	is the remainder
the year number	19	a
the year number	4	b
the year number	7	c
the number $19a + M$	30	d
the number $2b + 4c + 6d + N$	7	e

Then Easter falls on the $(22 + d + e)$th of March
or the $(d + e - 9)$th of April.

The symbols M and N are described (in Gauss's own words) as follows:

> M and N are numbers that have unchanging values for all time in the Julian calendar, and always throughout at least 100 years in the Gregorian calendar; in the former, $M = 15$ and $N = 6$ . . .
>
> In general, one can find the values for M and N in the Gregorian calendar for any given century from $100k$ to $100k + 99$ through the following rule:
>
> Suppose that k divided by $\begin{Bmatrix} 3 \\ 4 \end{Bmatrix}$ gives the (entire) quotients $\begin{Bmatrix} p \\ q \end{Bmatrix}$ where no consideration is given to the remainders. Then $\begin{Bmatrix} M \\ N \end{Bmatrix}$ is the remainder one obtains, when one divides $\begin{Bmatrix} 15 + k - p - q \\ 4 + k - q \end{Bmatrix}$ by $\begin{Bmatrix} 30 \\ 7 \end{Bmatrix}$.

With these descriptions of a, b, c, d, e, k, p, q, M, and N, Gauss's algorithm is complete, but far from easily understood. We shall now proceed with our examination of the formulas for these ten values that lie at the heart of the algorithm.

The Metonic cycle and the PFM date

The Julian calendar, in use until the late sixteenth century (and still later in England and its colonies), consists of ordinary years of 365 days and leap years of 366 days, achieved by the familiar rule of inserting February 29th into years divisible by four. Thus the average Julian year is exactly 365.25 days. (The Gregorian calendar, to be discussed later, modifies this plan slightly.)

By the fourth century, it was known that 19 (average) Julian years span almost exactly the same length of time as 235 lunar months (new moon to new moon). This equivalence gives an average lunar month of $19 \times 365.25/235 = 29.53085$ days, whereas the true lunar month is approximately 29.53059 days. This close agreement forms the basis of the so-called *Metonic cycle,* a tabulation of new moons formulated roughly as follows. Starting with an observed new moon December 24, 322 AD, new moons are inserted into the table in intervals alternating between 30 and 29 days: January 23, February 21, March 23, ..., December 13 in the first year, then January 12, February 10, ... in the second year, etc. This pattern will, of course, yield an average lunar month of 29.5 days, which is slightly too short. Thus, six "leap months" of 30 days and one of 29 days were inserted in the 19 year cycle at intervals selected to keep the Metonic new moons as close to the true new moons as possible. The Metonic cycle asserts that this pattern repeats *exactly* every 19 years. More information about the Metonic cycle can be found in [2] and [7].

Though the construction of the entire table of Metonic new moons is a bit of a puzzle, only those corresponding to the first full moon on or after March 21 are critical to the Easter date. Over the entire 19 year cycle, these new moons occur (with M for March and A for April) M23, M12, M31, M20, M9, M28, M17, A5 (=M36), M25, M14, A2 (=M33), M22, M11, M30, M19, M8, M27, M16, A4 (=M35). Using Gauss's symbol A for the year and the familiar *mod* notation for the remainder in integer division, we see that his first computation $a = A \mod 19$ simply determines where a given year falls in the ever-repeating 19 year Metonic cycle, and thus which of the 19 new moon dates listed above is relevant. One might also observe that each new moon date in the list can be obtained from the preceding one by subtracting 11 days or adding 19 days, operations that are closely related in modulo 30 arithmetic. In fact it is not difficult to verify that this entire list of March and April new moon dates can be generated as March $8 + d$, where $d = (19a + M) \mod 30$, $M = 15$, and $a = 0, 1, 2, \ldots, 18$. Of course, the nth day of March with $n > 31$ must be interpreted as the $(n - 31)$st day of April; for example, when $a = 7$, we get $d = 28$, and the new moon of "March 36" actually falls on April 5. We will assume this convention for the remainder of the article.

Finally, we observe that the full moon date is always taken as the fourteenth day of the lunar cycle; that is, 13 days after the new moon. Thus, *the paschal full moon (PFM) date in year A is March* $21 + d$ in Gauss's formulation.

The Sunday formula

In the list of Metonic cycle new moon dates in the previous section, the earliest and latest are March 8 and April 5, respectively, with corresponding full moons March 21 and April 18. Easter falls at least one day, but at most seven days after the PFM date, making March 22 the earliest and April 25 the latest possible Easter dates. Gauss sets the Easter date as March $22 + d + e$, where $e \in \{0, 1, 2, 3, 4, 5, 6\}$ is chosen so that Easter falls on a Sunday (the first Sunday after the PFM). Finding e is a straightforward problem that amounts to counting elapsed days from a particular, known Sunday to March $22 + d$ in the given year, then determining what e must be added to make

the count divisible by 7. Curiously enough, though this is perhaps the most easily understood part of the whole process, it is the part that Gauss explains most thoroughly. His plan for finding e is based on the Gregorian calendar, which we shall address later; for now, we present Gauss's plan with minor modifications to complete our work for the Julian calendar.

The number of days from Sunday, March 20, 1580, to March $22 + d + e$ in year A is $2 + d + e + i + 365(A - 1580)$, where i counts February 29ths occurring in this interval. (Here $A \leq 1582$; in years before 1580 our count of elapsed days will be negative.) Letting $b = A \mod 4$, it is not hard to see that $i = \frac{1}{4}(A - b - 1580)$. Thus we want to choose e to make

$$2 + d + e + \frac{1}{4}(A - b - 1580) + 365(A - 1580)$$

divisible by 7. Adding $\frac{7}{4}(A - b - 1580)$ will not affect divisibility by 7, and produces the simpler form

$$2 + d + e + 367(A - 1580) - 2b. \tag{1}$$

With simple reductions and the introduction of the symbols $c = A \mod 7$ and $N = 6$ one finds that the quantity in equation (1) is divisible by 7 exactly when

$$e = (2b + 4c + 6d + N) \mod 7.$$

Gauss's formulation of the Easter date in year A in the Julian Calendar is complete: Fix $M = 15$ and $N = 6$. Set $a = A \mod 19$ to determine the year's position in the Metonic cycle. Let $d = (19a + M) \mod 30$ so that March $21 + d$ gives the PFM date for year a in the Metonic cycle. Let $b = A \mod 4$, $c = A \mod 7$, and $e = (2b + 4c + 6d + N) \mod 7$ to find the next Sunday. Then Easter Sunday is March $22 + d + e$ or April $d + e - 9$, as appropriate.

The Gregorian calendar reform

Because the Julian calendar slightly overestimates the length of the year, Easter dates based on March 21 shifted further and further from the true vernal equinox. In 1582 Pope Gregory instituted corrective changes resulting in the so-called *Gregorian calendar* still in widespread use today. First, to correct the drift in the equinox date that had already occurred after several centuries under the Julian calendar, the day after October 4, 1582 was declared to be October 15, 1582. Second, to prevent the same problem from arising in the future, the length of the average year was reduced by declaring that century years would be leap years only if divisible by 400. Thus 1600 was a leap year, 1700, 1800, 1900 were common years, 2000 was a leap year, etc.

Clearly these two corrections affect the dates upon which Sundays fall, a change that can be addressed by modifying the value of N. (This change is broadly described in the literature as the "solar equation.") Once again our work closely mimics Gauss's presentation, adding a few details that he chose to omit.

Gauss counts the number of days from Sunday, March 21, 1700 to March $22 + d + e$ in year A as $1 + d + e + i + 365(A - 1700)$, where once again i counts February 29ths in this interval. (Here $A \geq 1583$.) By setting $k = \lfloor A/100 \rfloor$ (the integer part of the quotient) and $q = \lfloor k/4 \rfloor$ and recalling that $b = A \mod 4$, we have $i = \frac{1}{4}(A - b - 1700) - (k - 17) + (q - 4)$. (Here we are discarding February 29ths in the century years and adding them back in for the years divisible by 400.) As before, we want to choose e to make

$$1 + d + e + \frac{1}{4}(A - b - 1700) - (k - 17) + (q - 4) + 365(A - 1700)$$

divisible by 7. By adding $\frac{7}{4}(A - b - 1700)$ and simplifying it is not hard to show that the resulting expression is divisible by 7 exactly when

$$e = (2b + 4c + 6d + 4 + k - q) \mod 7.$$

Now the choice of $N = (4 + k - q) \mod 7$ in Gauss's algorithm is clear.

The Gregorian calendar reform contained a third correction that does not influence the civil calendar, but does affect the Easter date. The reader may recall that the Metonic cycle produces a lunar month of 29.53085 days, compared to the true lunar month averaging 29.53059 days. Just as the incorrect length of the Julian year slowly accumulated to a noticeable error, the same was true for this incorrect length of the lunar month. The accumulated error is approximately 1 day in 312.5 years, or very nearly 8 days every 2500 years. Thus the Gregorian calendar reform included a one-day correction in the PFM date every 300 years, seven consecutive times, followed by a one-day correction after another 400 years, repeated indefinitely. The so-called "lunar equation" affects the PFM date by changing M (and thus d) according to this plan.

The solar equation, which resulted from removing February 29ths from the calendar, affects the PFM dates just as it affects the Sunday calculation. Thus the value of M must change by $k - q$, as did the value of N. All that remains is to incorporate the lunar equation into M, but here, unfortunately, the extraordinarily dependable Dr. Gauss has a minor failure. For the lunar equation, Gauss defines $p = \lfloor k/3 \rfloor$, then builds

$$M = (15 + k - p - q) \mod 30. \tag{2}$$

His error is clear: this M properly accounts for the solar equation with $k - q$, but applies the one-day correction dictated by the lunar equation *every* 300 years, rather than the prescribed sequence of seven consecutive 300-year intervals followed by an eighth interval of 400 years. Only mathematicians could be troubled by an error in the Easter date that won't surface for another two thousand years, but since we have the means to correct it, let us do so!

Gauss's error

One of the most widely quoted sources on the subject of the Easter date is J. M. Oudin's 1940 paper "Étude sur la date de Paques" [8]. Indeed, it is Oudin's Easter algorithm that appears in the U.S. Naval Observatory's almanac cited above. And it is common for those who reproduce Oudin's algorithm to do so after enthusiastically describing Gauss' algorithm, then "pulling the rug out from under it" by mentioning the error described above. (Or simply declaring that Gauss's method is wrong without understanding the nature of the error at all.)

Oudin writes, "Gauss, having forgotten to take into account these delays of the lunar equation in his formula for M, the latter ... works only until 4199 inclusively and is found thus devoid of the character of generality that its author thought to have given it." Oudin is correct, but he writes without benefit of the ability to conduct a simple internet search. With this modern tool, it is not difficult to find the following piece missing from his puzzle.

In the January/February 1816 edition of the *Zeitschrift fur Astronomie und verwandte Wissenschaften (Journal for Astronomy and Related Sciences)* we find a contribution from Gauss [5]. It is titled "Berichtigung zu dem Aufsatze: Berechnung des Osterfestes (Correction to the Essay: Computation of the Easter Date)," and gives publication information for the original August, 1800 paper. Gauss writes

> ...my establishment of that rule [for p] took no notice of the circumstance that the lunar equation, so-called by the originators of the Gregorian calendar, which is reasonable every 300 years ... must actually become reasonable once every $312\frac{1}{2}$ years. Without my engaging the question here of whether this achieves the intended purpose, I only remark that my rule, with the arrangement of the Gregorian Calendar as it is, can easily be brought into complete agreement when one does not accept for the number p the quotient in the division of the number k by 3, as stipulated in the citation above, rather one accepts the quotient in the division of the number $13 + 8k$ by 25.

And of course, when Gauss's corrected value $p = \lfloor \frac{13+8k}{25} \rfloor$ is incorporated into equation (2) for M, it exactly produces the necessary one-day changes every 300 years, seven consecutive times, followed by a one-day change after 400 years. Oudin's criticism was 124 years too late!

We close this section by noting that in the Gauss *Werke*, "Berechnung des Osterfestes" closes with "Handschriftliche Bemerkung" or "Handwritten remarks." Among these is the correct formula for p. Presumably this was Gauss's handwritten remark, added to the manuscript sometime after his note of 1816.

The exceptional cases

Though it would appear that the Easter date algorithm is complete, there are two special cases, not at all obvious, that must be addressed. The reader will gain a much better appreciation of the challenges inherent in examining Gauss's work by considering these exceptions in his own words:

> From the above rules, one finds unique and alone in the *Gregorian calendar* the following two exceptional cases:
>
> I. If the calculation gives Easter on the 26th of April, then one *always* takes the 19th of April. (e.g., 1609, 1989).
>
> One easily sees that this case can only occur where the calculation gives $d = 29$ and $e = 6$; d can only obtain the value 29 when $11M + 11$ divided by 30 gives a remainder that is *smaller* than 19; to this end, M must have one of the following 19 values:
>
> $$0, 2, 3, 5, 6, 8, 10, 11, 13, 14, 16, 17, 19, 21, 22, 24, 25, 27, 29.$$
>
> II. If the calculation gives $d = 28$, $e = 6$, and meets the requirement that $11M + 11$ divided by 30 gives a remainder that is smaller than 19, then Easter does not fall, as follows from the calculation, on the 25th, rather on the 18th of April. One can easily convince oneself that this case can only occur in those centuries in which M has one of the following eight values:
>
> $$2, 5, 10, 13, 16, 21, 24, 29.$$
>
> *With these two exceptional cases accounted for, the above rules are completely general.*

When translating Gauss, it is always useful to have a list of German synonyms for "easily!"

The first exception is straightforward, though Gauss's added remarks make it less so. We noted previously that under the Julian calendar and the original Metonic cycle,

Easter dates always fall between March 22 and April 25, inclusive. This rule was continued under the Gregorian calendar. If $d = 29$ and $e = 6$, however, Gauss's formula produces a PFM date of April 19 and an Easter date of April 26, in violation of the rule. The Church's Easter tables simply shifted the PFM date to one day earlier under this circumstance (from Sunday, April 19 to Saturday, April 18), effectively replacing $d = 29$ by $d = 28$. This has no effect at all *unless* $e = 6$, that is, *unless* April 19 is a Sunday; in that case it forces Easter to occur a week earlier, on April 19.

The comments that Gauss supplies along with the first exception are puzzling until one observes that by fixing $d = 29$ in the PFM formula $d = (19a + M) \mod 30$, one can solve to find $a = (11M + 11) \mod 30$ and $M = (29 + 11a) \mod 30$. The first of these equations reduces Gauss's claim of "a remainder that is *smaller* than 19" to the simple observation that a takes on the values $0, 1, 2, \ldots, 18$. Now Gauss's list of M values can be generated by successively substituting these a values into the second equation. One should also note that Gauss's choice of 1989 as an example of the first exception is erroneous; he corrects this without comment in an article on the Easter date that appeared in *Brunswick Magazine* [**6**] in 1807, replacing 1989 with 1981.

The second exceptional case is a result of the change described in the first. The Gregorian calendar reformers wished to preserve a basic feature of the original Metonic full moons: the PFM date is never duplicated within one 19-year cycle. But now this is sure to happen when $d = 28$ and $d = 29$ occur within the same 19-year cycle, because the $d = 29$ PFM (April 19) has been shifted to the $d = 28$ PFM date (April 18). Gauss's peculiar way of identifying these cycles in which both $d = 28$ and $d = 29$ occur needs a closer look.

Suppose that for a given M value, there is an $a \in \{0, 1, \ldots, 18\}$ for which $d = (19a + M) \mod 30$ produces $d = 28$; suppose also that "$11M + 11$ divided by 30 gives a remainder that is smaller than 19." The latter requirement means that there is an $a \in \{0, 1, \ldots, 18\}$ for which $a = (11M + 11) \mod 30$. As in the first exception, this can be rearranged to obtain $29 = (19a + M) \mod 30$; that is, for the given M value $d = 29$ is also achieved.

So how do we avoid the duplication of PFM dates when a given M value produces both $d = 28$ and $d = 29$ in a 19-year cycle? Simply shift the $d = 28$ PFM date (April 18) to that of $d = 27$ (April 17)! The apparent cascade of shifts that might be produced by this plan does not occur, because the 19-year cycles containing both $d = 28$ and $d = 29$ never contain $d = 27$. (Probably the simplest way to see this is by brute force: construct a table of $d = (19a + M) \mod 30$ values for $a = 0, 1, 2, \ldots, 18$ and $M = 0, 1, 2, \ldots, 29$.) Since $e = 6$ in Gauss's second exceptional case, the PFM associated with $d = 28$ falls on *Sunday*, April 18, so reducing d to 27 places the PFM on Saturday, April 17 and Easter on Sunday, April 18, just as Gauss's second exception dictates.

Finally, to find the eight values of M supplied by Gauss in the second exceptional case, it is probably simplest to fix $d = 28$ and consider "the requirement that $11M + 11$ divided by 30 gives a remainder that is smaller than 19" by listing the triples

$$(a, M, (11M + 11) \mod 30)$$

for $a = 0, 1, 2, \ldots, 18$. Those meeting the requirement are $(11, 29, 0)$, $(12, 10, 1)$, $(13, 21, 2)$, $(14, 2, 3)$, $(15, 13, 4)$, $(16, 24, 5)$, $(17, 5, 6)$, and $(18, 16, 7)$, from which we can read off the middle entries as Gauss's eight M values.

Gauss's method is complete!

Conclusion

The mathematics in Gauss's Easter algorithm is trivial. But his ability to transform a desperately arcane set of rules and tables into a simple arithmetic process illustrates

once again his uncanny ability to see deeply into complicated matters. If only he would give mere mortals a little more guidance through his work . . . ah, but perhaps not. For then the joy of completing the many puzzles he has left behind would be lost to the rest of us!

REFERENCES

[1] Bien, R. (2004). Gauss and beyond: The making of Easter algorithms, *Arch. Hist. Exact Sci.* 58: 439–452.

[2] Bond, J. J. (1875). *Handy-book of rules and tables for verifying dates with the Christian Era, &c.* London: George Bell and Sons, available at http://books.google.com/books/about/Handy_book_of_Rules_and_Tables_for_Verif.html?id=liq4tEW3dc8C.

[3] Dunnington, G. W. (1955). *Carl Friedrich Gauss: Titan of Science*. New York: Hafner.

[4] Gauss, C. F. (1800). Berechnung des Osterfestes, *Werke*. 6: 73–79.

[5] Gauss, C. F. (1816). Berichtigung zu dem Aufsatze: Berechnung des Osterfestes. *Werke*. 11(1): 201.

[6] Gauss, C. F. (1807). Noch etwas über die Bestimmung des Osterfestes. *Werke*. 6: 82–86.

[7] Kennedy, T. (1909). Epact, *The Catholic Encyclopedia*, v. 5. New York: Robert Appleton Company, available at http://www.newadvent.org/cathen/05480b.htm.

[8] Oudin, J. M. (1940). Étude sur la date de Paques. *Bulletin Astronomique* 12: 391–410.

[9] Richards, E. G. (2012). Calendars. In: P. K. Seidelman, S. E. Urban, eds. *Explanatory Supplement to the Astronomical Almanac*. 3rd ed. Mill Valley, CA: University Science Books, pp. 600–601.

Summary. This article summarizes and explains the formula for the Easter date provided by Gauss in his paper "Berechnung des Osterfestes" written in 1800. It addresses an error that was corrected by Gauss in 1816.

DONALD TEETS (MR Author ID: 311701) received a D.A. from Idaho State University in 1988 and has taught at the South Dakota School of Mines and Technology since then. He received the Carl B. Allendoerfer writing award from the MAA in 2000 for an article on the astronomical work of Gauss, and the Burton W. Jones Distinguished Teaching Award from the Rocky Mountain Section of the MAA in 2004.

TRIBUS Solution

		3	1	2		2	3	1		1	2	3
		1	2	3		1	2	3		2	3	1
3	1	2	3	1	2	3	1	2	1	3	1	2
2	3	1		3	1	2		1	3	2		
1	2	3		2	3	1		3	2	1		

		3	2	1		1	3	2		3	2	1
		2	1	3		2	1	3		1	3	2
2	3	1	3	2	1	3	2	1	3	2	1	3
3	1	2		3	2	1		3	2	1		
1	2	3		1	3	2		2	1	3		

Heating a Torus with Number Theory

FLORIAN PAUSINGER
Queen's University
Belfast BT7 1NN
f.pausinger@qub.ac.uk

STEFAN STEINERBERGER
Yale University
New Haven, CT 06520
stefan.steinerberger@yale.edu

Suppose you live on a torus and want to heat it as efficiently as possible with a finite number of identical heat sources. What configuration is optimal and how much better is the optimal solution compared with, say, a random placement of the sources? The answer to this question is surprisingly simple and involves a small detour into some elementary number theory. Interestingly, the solution to this problem has connections to more advanced questions in numerical integration.

Our space of interest is the two-dimensional torus $\mathbb{T}^2$ represented as the square $[0, 2\pi]^2$ in which opposite boundary faces are glued together; see Figure 2. The aim is to find ways of effectively heating the torus $\mathbb{T}^2$. If we are given N identical radiators, how are we supposed to place them to guarantee that the temperature in $\mathbb{T}^2$ becomes everywhere nice and cozy as quickly as possible? We start by describing the heat given off by a radiator as a function $\phi : \mathbb{T}^2 \to \mathbb{R}$. Such a function could, for example, have the form

$$\phi(x, y) = e^{-36(x^2+y^2)},$$

which describes an initial temperature maximum at $(0, 0)$ that decays exponentially with growing distance; see Figure 1.

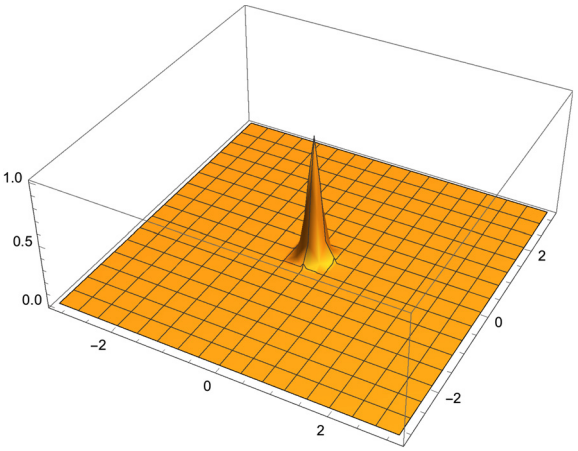


Figure 1 Plot of the function $\phi(x, y) = e^{-36(x^2+y^2)}$.

Math. Mag. **92** (2019) 99–106. doi:10.1080/0025570X.2018.1546507
MSC: Primary 35K05, Secondary 11L03
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

We assume the N radiators are placed at the points $(x_n, y_n)_{i=1}^N$, turned on for a little while and then turned off at time $t = 0$. We then watch as the heat spreads throughout the room. Each radiator will have contributed a particular temperature distribution ϕ and, since all radiators are assumed to be identical, the temperature in (x, y) at time $t = 0$, denoted by $u(0, x, y)$, will be of the form

$$u(0, x, y) = \sum_{n=1}^{N} \phi(x - x_n, y - y_n). \tag{1}$$

Denoting the temperature in $(x, y) \in \mathbb{T}^2$ at time t by $u(t, x, y)$, the governing physical law is the *free heat equation* $u_t = \Delta u$ (which we won't use directly). Physical intuition tells us that the heat is going to spread: points adjacent to warm points will heat up while, conversely, warm points surrounded by colder points will cool down. We expect the heat to be spread more and more evenly and that temperature will eventually converge to a constant (which can be explicitly computed because the total amount of heat in a closed system such as $\mathbb{T}^2$ has to stay constant):

$$\text{for all } (x, y) \in \mathbb{T}^2, \quad \lim_{t\to\infty} u(t, x, y) = \frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} u(0, x, y)dxdy.$$

It turns out to be useful to rewrite $u(0, x, y)$ as a Fourier series

$$u(0, x, y) = \sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} e^{i(kx+my)}.$$

This representation immediately implies that

$$\begin{aligned}
\frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} u(0, x, y)dxdy &= \frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} \left(\sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} e^{i(kx+my)} \right) dxdy \\
&= \frac{1}{\text{area}(\mathbb{T}^2)} \left(\sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} \int_{\mathbb{T}^2} e^{i(kx+my)} \right) dxdy \\
&= c_{0,0},
\end{aligned}$$

where $c_{0,0}$ is the constant term of the Fourier series of $u(0, x, y)$ because all the other integrals vanish. How should we pick the points $\{(x_n, y_n)_{n=1}^N\}$ to ensure that the initial condition $u(0, x, y)$ converges to the constant temperature as quickly as possible? We do not want to make any special assumptions on the form of $\phi : \mathbb{T}^2 \to \mathbb{R}$. In particular, our placement rule will be valid for *all* smooth functions ϕ—even for heat distributions ϕ that are not radially symmetric or even physically meaningful. The key is to utilize elementary number theory in the form of particular permutations of finite fields.

Result

Our argument will immediately show that *every* heat distribution converges to a constant with speed at least e^{-t}. Moreover, we present a general and particularly nice placement of N points (with N prime) that uses number theory in an essential way to get a much faster convergence speed of *at least* $e^{-(N/4+\varepsilon)t}$ with $\varepsilon > 0$: for a prime

number N, an integer p satisfying $\sqrt{N}/2 < p \leq \sqrt{N}$ and an arbitrary $q \in \mathbb{N}$, we define the point set $\{(x_n, y_n)_{1 \leq n \leq N}\}$ with

$$x_n = 2\pi \frac{n}{N} \qquad \text{and} \qquad y_n = 2\pi \frac{(pn+q) \mod N}{N};$$

Figure 2 illustrates the construction for $N = 7$, $p = 2$ and $q = 3$. Number theoretical constructions of point sets of this kind are called *lattice rules* and have a long history in the field of numerical integration (see, for example, the book by Sloan and Joe [**6**]). It seems that the connection to the placement of radiators has not been observed before.

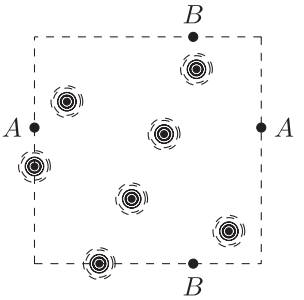


Figure 2 Placing seven (identical) heat distributions (radiators, candles, ...) at these points leads to convergence towards the constant room temperature with a speed of at least e^{-4t}. The points A and B illustrate how boundary faces are identified. In other words, exiting to the left (up) makes you appear on the right (down).

Theorem. *This set of points has the following property: for every smooth heat distribution $\phi : \mathbb{T}^2 \to \mathbb{R}$, the initial distribution $u(0, x, y)$ given by*

$$u(0, x, y) = \sum_{n=1}^{N} \phi(x - x_n, y - y_n)$$

converges to the equilibrium with speed at least

$$\max_{(x,y) \in \mathbb{T}^2} \left| u(t, x, y) - c_{0,0} \right| \leq c e^{-\alpha t},$$

where $c_{0,0}$ is the constant term of the Fourier series of $u(0, x, y)$, c is a constant independent of N and t, and $\alpha = \lfloor \sqrt{N}/2 \rfloor^2 + 2\lfloor \sqrt{N}/2 \rfloor + 1 \geq N/4$.

How this works

The argument comes in two parts. First, we use a neat formula for the heat equation to derive a condition on the points $\{(x_n, y_n)_{1 \leq n \leq N}\}$ that ensures that the heat equation converges quickly to its equilibrium state. In the second part of the argument, we verify that the condition is valid for our proposed set of points.

Step 1. It will be practical to use an explicit formula for the solution of the heat equation in terms of Fourier series. More precisely, if

$$u(0, x, y) = \sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} e^{i(kx+my)},$$

then (see, e.g., [3]) the solution of the heat equation is given by

$$u(t, x, y) = \sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} e^{-(k^2+m^2)t} e^{i(kx+my)},$$

which can be easily verified by explicit computation since

$$\frac{\partial}{\partial t} u(t, x, y) = \sum_{(k,m)\in\mathbb{Z}^2} c_{k,m}(-k^2 - m^2) e^{-(k^2+m^2)t} e^{i(kx+my)} = \Delta u(t, x, y).$$

This already shows that

$$u(t, x, y) - \frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} u(0, x, y) dx dy = \sum_{\substack{(k,m)\in\mathbb{Z}^2 \\ (k,m)\neq(0,0)}} c_{k,m} e^{-(k^2+m^2)t} e^{i(kx+my)}$$

can be written as the sum of exponential functions each of which decays at least as quickly as e^{-t} independently of everything else.

We now write the heat distribution of a single heat source (radiator, candle, ...) as

$$\phi(x, y) = \sum_{(k,m)\in\mathbb{Z}^2} a_{k,m} e^{i(kx+my)}.$$

Plugging this into equation (1) and exchanging the order of summation leads to

$$\begin{aligned}
\sum_{n=1}^{N} \phi(x - x_n, y - y_n) &= \sum_{n=1}^{N} \sum_{(k,m)\in\mathbb{Z}^2} a_{k,m} e^{i(k(x-x_n)+m(y-y_n))} \\
&= \sum_{n=1}^{N} \sum_{(k,m)\in\mathbb{Z}^2} a_{k,m} e^{-ikx_n} e^{-imy_n} e^{i(kx+my)} \\
&= \sum_{(k,m)\in\mathbb{Z}^2} \underbrace{a_{k,m} \left(\sum_{n=1}^{N} e^{-ikx_n} e^{-imy_n} \right)}_{:= c_{k,m}} e^{i(kx+my)}.
\end{aligned}$$

The formula for the heat equation tells us that we would like to pick the points $\{(x_n, y_n)_{n=1}^N\}$ such that the expression in brackets vanishes for as many small values of k, m as possible. More precisely, we obtain the following lemma.

Lemma 1. *If the set of points $(x_n, y_n)_{1\leq n\leq N}$ has the property that*

$$\sum_{n=1}^{N} e^{ikx_n} e^{imy_n} = 0 \tag{2}$$

for all $(k, m) \in \mathbb{Z}^2$ *with* $(k, m) \neq (0, 0)$, $|k|, |m| \leq \ell$ *and* $\ell \in \mathbb{Z}$, *then for all smooth* $\phi : \mathbb{T}^2 \to \mathbb{R}$ *and corresponding* $u(0, x, y)$ *as in equation* (1) *we have, for some constant* $c > 0$,

$$\max_{(x,y)\in\mathbb{T}^2} \left| u(t, x, y) - \frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} u(0, x, y) dx dy \right| \leq c e^{-(\ell^2+2\ell+1)t}.$$

Proof. We have the explicit solution to the heat equation as

$$u(t, x, y) = \sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} e^{-(k^2+m^2)t} e^{i(kx+my)},$$

where $c_{k,m} = 0$ for all $|k|, |m| \leq \ell$, $\ell \in \mathbb{Z}$. This means that the smallest sum of squares $k^2 + m^2$ of a pair (k, m) for which $c_{k,m} \neq 0$ is at least $(\ell + 1)^2 + 0^2 = \ell^2 + 2\ell + 1$ for the pair $(\ell + 1, 0)$. Hence, the first nonzero exponential term decays at least as quickly as $e^{-(\ell^2+2\ell+1)t}$. As for nonzero coefficients, we can use the fact that ϕ is smooth to conclude that

$$|a_{k,m}| = \left| \frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} \phi(x, y) e^{-i(kx+my)} dx dy \right| \leq \max_{(x,y)\in\mathbb{T}^2} |\phi(x, y)|$$

and classical results (see, e.g., [3]) on the decay of Fourier coefficients of smooth functions to ensure that everything is summable. ∎

Step 2. We will now show that our point set has the property from Lemma 1.

Lemma 2. *For a prime* N, *let* $\{(x_n, y_n) : 1 \leq n \leq N\}$ *be as defined above. Then for* $(0, 0) \neq (k, m) \in \mathbb{Z}^2$ *with* $|k|, |m| \leq \sqrt{N}/2$

$$\sum_{n=1}^{N} e^{ikx_n} e^{imy_n} = \sum_{n=1}^{N} e^{2\pi i \frac{kn+m(pn+q)}{N}} = \sum_{n=1}^{N} e^{2\pi i \frac{(k+mp)n+mq}{N}} = 0. \tag{3}$$

Proof. This is a fairly intricate expression. We want to relate it to the fact that the Nth roots of unities sum to 0 because they form a geometric progression, that is,

$$\sum_{n=1}^{N} e^{2\pi i \frac{n}{N}} = \sum_{n=1}^{N} \left(e^{2\pi i \frac{1}{N}} \right)^n = \frac{e^{2\pi i \frac{N+1}{N}} - 1}{e^{2\pi i \frac{1}{N}} - 1} - 1 = 0. \tag{4}$$

Therefore, *if* $n \to (k + mp)n + mq$ is a bijection on $\mathbb{Z}_N = \{0, 1, \ldots, N - 1\}$, then equation (4) would imply equation (3) since we are still summing over the roots of unity (just in a different order); see Figure 3.

For which values of k and m is this the case? First, it is clear that the map $n \to (k + mp)n + mq$ is a bijection on $\mathbb{Z}_N$ if and only if $n \to (k + mp)n$ is a bijection on $\mathbb{Z}_N$. Now, since N is prime, the map $n \to (k + mp)n$ is a bijection if and only if N does not divide evenly into $mp + k$, which is true for k, m and p, as assumed. Indeed, since

$$|k|, |m| \leq \frac{1}{2}\sqrt{N} < p \leq \sqrt{N},$$

we have

$$|mp + k| \leq |mp| + |k| \leq \frac{1}{2}N + \frac{1}{2}\sqrt{N} < N.$$

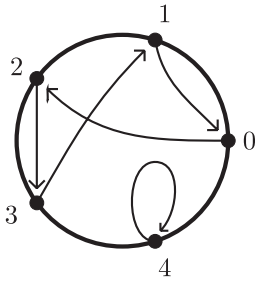


Figure 3 The map $n \to 3n + 2$ is a bijection on $\mathbb{Z}_5$ and induces a reordering of the 5th roots of unity.

This means that N does divide evenly into $mp + k$ only if $mp + k = 0$. However, $|k| < p$ implies $-p < k < p$ and so

$$(m-1)p < mp + k < (m+1)p.$$

Thus $mp + k = 0$ implies $m - 1 < 0 < m + 1$, which gives $m = 0$ and thus $k = 0$. But this contradicts the assumption that $(k, m) \neq (0, 0)$, thus we must have that N does not divide evenly into $mp + k$, as claimed. ■

Since $k, m \in \mathbb{Z}$ and $|k|, |m| \leq \sqrt{N}/2$, we set $\ell = \lfloor \sqrt{N}/2 \rfloor$ in Lemma 1 to obtain the result stated in the theorem.

Example. To illustrate our result we place 7 radiators as shown in Figure 2 and use the heat distribution ϕ from the introduction. It turns out that ϕ has a particularly nice Fourier series, i.e.,

$$\phi(x, y) = e^{-36(x^2+y^2)} = \sum_{(k,m)\in\mathbb{Z}^2} a_{k,m} e^{i(kx+my)},$$

with

$$a_{k,m} = \frac{1}{4 \cdot 36\pi} e^{-\frac{k^2+m^2}{4\cdot 36}}.$$

The coefficients $a_{k,m}$ are all we need to build the functions $u(0, x, y)$ and $u(t, x, y)$. We observe that the average value of one heat source is $a_{0,0}$ and thus, by summation, $c_{0,0} = N a_{0,0}$. This shows that the average temperature is

$$\frac{1}{\text{area}(\mathbb{T}^2)} \int_{\mathbb{T}^2} u(0, x, y) dx dy = 7a_{0,0} = \frac{7}{144\pi} \approx 0.01547\ldots$$

Thus, we can numerically investigate how fast $\max_{(x,y)} u(t, x, y)$ converges to $7a_{0,0}$ as $t \to \infty$ for different point sets. According to Lemma 2 all coefficients $c_{k,m}$ with $|k|, |m| \leq \sqrt{7}/2$ vanish. Hence, setting $\ell = \lfloor \sqrt{7}/2 \rfloor$ in Lemma 1, we obtain a convergence of at least e^{-4t}, which can be also observed from our numerical results in Figure 4. This can be easily compared to random points: for a set of N random points, we expect a speed of convergence of $(c/\sqrt{N})e^{-t}$, where c is a constant depending only on ϕ. This shows that there is quite a bit of decay for t small but, as t becomes large, the result is much worse than our construction.

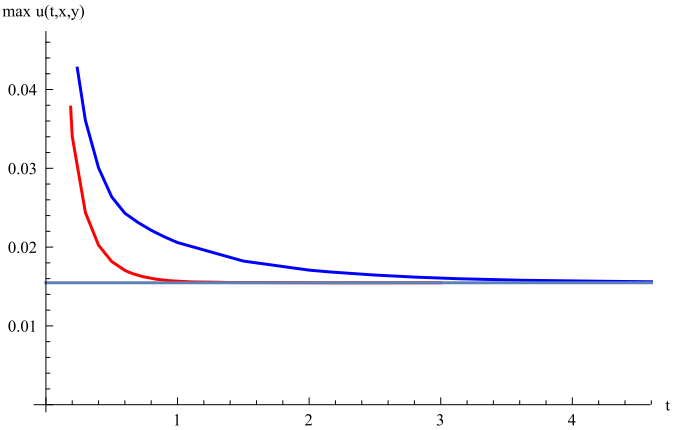


Figure 4 Comparison of convergence of $\max_{(x,y)} u(t, x, y)$ for different point sets. The red graph shows the convergence for our point set (with convergence speed e^{-4t}), the blue graph corresponds to a set of 7 random points, which almost surely will always only converge with speed e^{-t}.

The big picture

To illustrate the generality of the ideas from the example, we outline the appearance of these number theoretical constructions in two more situations.

Fourier analysis. A result of Montgomery [4, 5] (proven for a quite different purpose) implies that our construction is optimal up to a constant: N heat sources can never decay faster than $e^{-\alpha t}$ where $\alpha \sim N$. A result [7] of the second author implies that this is even true if we allow the heat sources to have different (positive weights). Statements of this type are probably true on very general domains, the best-known result in that direction is [1].

Cubature formulas. A set of N points $\{(x_n, y_n)_{n=1}^N\}$ is said to be an *exact cubature formula of degree* ℓ on $\mathbb{T}^2$ if we can compute the integral of every function of the form

$$f(x, y) = \sum_{\substack{(k,m)\in\mathbb{Z}^2 \\ |k|+|m|\leq \ell}} a_{k,m} e^{i(kx+my)},$$

exactly by taking the average value at the N points, i.e., if for all such f, then

$$\frac{1}{(2\pi)^2} \int_{\mathbb{T}^2} f(x, y) dx dy = a_{0,0} = \frac{1}{N} \sum_{n=1}^{N} f(x_n, y_n).$$

Plugging in and exchanging the order of summation gives

$$\sum_{n=1}^{N} f(x_n, y_n) = \sum_{\substack{(k,m)\in\mathbb{Z}^2 \\ |k|+|m|\leq \ell}} a_{k,m} \left(\sum_{n=1}^{N} e^{ikx_n} e^{imy_n} \right)$$

and, hence the similarity. The underlying expression is the same and the goal is to select points that make it vanish for all $|k| + |m| \leq \ell$ with $(k, m) \neq (0, 0)$. Interestingly, Cools and Sloan [2] have discovered cubature formulas with the minimal

possible number of points for a given ℓ that are *not* lattice rules. It would be interesting to see similar results in our context.

Wave equation. The radiator problem also appears in other physical problems. Suppose we have a $\mathbb{T}^2$-swimming pool and want to create waves by throwing in N stones simultaneously. The goal is to pick the points in such a way that we get no low-frequency waves at all. For simplicity, we can model this with the wave equation

$$u_{tt} = \Delta u \qquad \text{and} \quad u_t\big|_{t=0} = 0$$

for $u(0, x, y)$ as in equation (1). Using again the theory of Fourier series, we follow our line of thought from Step 1. We solve the wave equation explicitly with the formula

$$u(t, x, y) = \sum_{(k,m)\in\mathbb{Z}^2} c_{k,m} e^{i(k^2+m^2)t} e^{i(kx+my)},$$

and see by the same subsequent computation, that, in order not to generate low-frequency waves, it is required that equation (2) holds for as many small $(k, m) \in \mathbb{Z}^2$ as possible. These point sets and the same arguments even apply to more complicated equations such as the Schrödinger equation $iu_t = \Delta u$.

REFERENCES

[1] Bilyk, D., Dai, F., Steinerberger, S. (2018). General and refined Montgomery lemmata, arXiv:1801.07701
[2] Cools, R., Sloan, I. (1996). Minimal cubature formulae of trigonometric degree. *Math. Comp.* 65 (216): 1583–1600.
[3] Evans, L. (1998). *Partial Differential Equations*. Graduate Studies in Mathematics, Vol. 19. Providence, RI: American Mathematical Society. ISBN: 0-8218-0772-2
[4] Montgomery, H. (1989). Irregularities of distribution by means of power sums. Proceedings of the Congress on Number Theory (Spanish) (Zarauz, 1984), 11–27.
[5] Montgomery, H. (1994). Ten lectures at the interface of harmonic analysis and number theory. CBMS Regional Conference Series in Mathematics, Vol. 84. Providence, RI: American Mathematical Society.
[6] Sloan, I., Joe, S. (1994). *Lattice Methods for Multiple Integration.* New York: Clarendon Press, Oxford University Press. ISBN: 0-19-853472-8
[7] Steinerberger, S. (2017). Spectral limitations of quadrature rules and generalized spherical designs, arXiv:1708.08736

Summary. We discuss an amusing application of number theory: suppose you find yourself on the two-dimensional torus $\mathbb{T}^2$ equipped with N candles and want to position the candles in such a way that they heat up the room as efficiently as possible. To achieve this goal, we give a construction that uses number theory as the main ingredient and explain related results.

FLORIAN PAUSINGER (MR Author ID: 920909) is a lecturer at Queen's University Belfast. He studied pure mathematics at the University of Salzburg, Austria, before he moved to Vienna where he received his PhD from IST Austria under the direction of Herbert Edelsbrunner. In his research he is mainly interested in the very basic question of how to distribute points in a uniform manner in squares, on spheres or on more complicated shapes. Moreover, he is interested in waves—both in classroom teaching fluid mechanics as well as on the Irish beaches while surfing.

STEFAN STEINERBERGER (MR Author ID: 869041) received his undergraduate education in Austria and a PhD in Mathematical Analysis from Bonn, Germany. He then moved to Yale where he is an Assistant Professor of Mathematics. He spends most of his days trying to figure out why things vibrate the way they do (not so easy!) and drinking coffee.

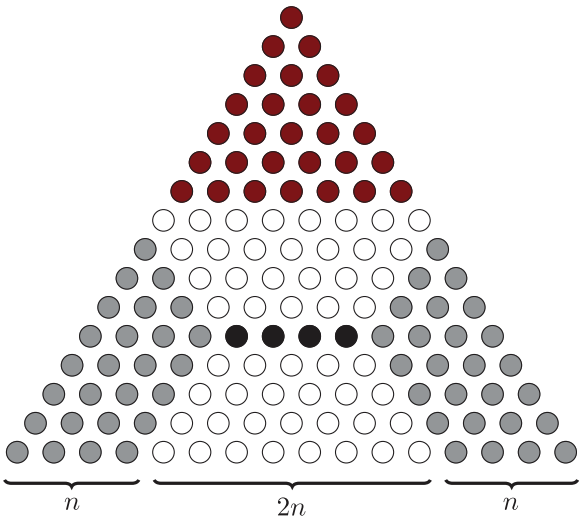
Proof Without Words: An Identity on Difference Between Triangular Numbers

GÜNHAN CAGLAYAN
New Jersey City University
Jersey City, NJ 07305
gcaglayan@njcu.edu

Proposition. *For $n \in \mathbb{N}$, the following identity holds:*

$$T_{4n} = 4(T_{2n} - T_n) + n + T_{2n-1}.$$

Proof. The proof is demonstrated for $n = 4$.



■

Summary. This proof without words demonstrates an identity on difference between triangular numbers.

GÜNHAN CAGLAYAN (MR Author ID: 1116420) teaches mathematics at New Jersey City University. His main interests are mathematical visualization and student learning through modeling and visualization.

Math. Mag. **92** (2019) 107–107. doi:10.1080/0025570X.2019.1562297
MSC: Primary 05A19
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

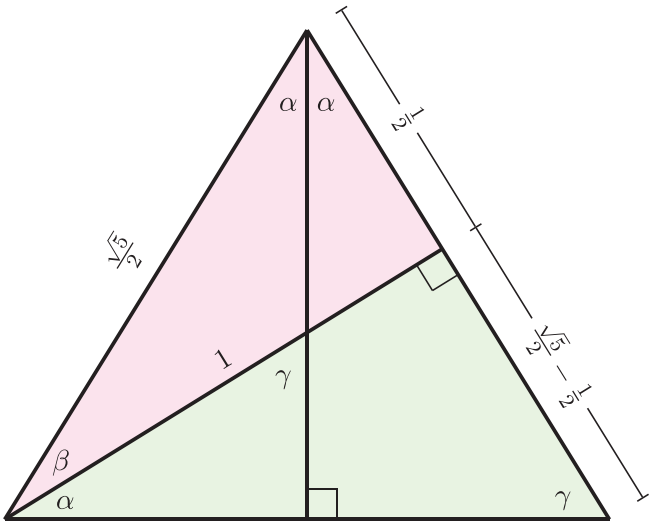
Proof Without Words: Some Arctangent Identities Involving 2, the Golden Ratio, and Their Reciprocals

REX H. WU
New York Presbyterian Lower Manhattan Hospital
New York, NY 10038
rexhwu@yahoo.com

Theorem. *The golden ratio* $\varphi = \dfrac{\sqrt{5}+1}{2}$ *satisfies* $\dfrac{1}{\varphi} = \dfrac{\sqrt{5}-1}{2}$. *The following relationships hold:*

(1) $\arctan\varphi = \arctan(1/2) + (1/2)\arctan 2$,
(2) $\arctan\varphi - \arctan(1/\varphi) = \arctan(1/2)$,
(3) $2\arctan(1/\varphi) = \arctan 2$,
(4) $\arctan\varphi + (1/2)\arctan 2 = \pi/2 = 2\arctan(1/\varphi) + \arctan(1/2)$, *or,*
$\arctan\varphi - 2\arctan(1/\varphi) = \arctan(1/2) - (1/2)\arctan 2$, *and*
(5) $\arctan\varphi = \pi/4 + (1/2)\arctan(1/2)$.

The following image can be used to show (1)–(5) above.



$$\alpha = \arctan(1/\varphi),\ \beta = \arctan(1/2),\ \gamma = \alpha + \beta = \operatorname{arccot}(1/\varphi) = \arctan\varphi,$$
$$2\alpha = \arctan 2,\ 2\alpha + \beta = \pi/2.$$

Math. Mag. **92** (2019) 108–109. doi:10.1080/0025570X.2019.1562296
MSC: Primary 51M04
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

Beyond the proof without words, a linear combination of any two of the first four identities would give a general form of all the identities. For example, using (2) and (3), for $x, y \in \mathbb{R}$, we have $x \arctan \varphi + (2y - x) \arctan(1/\varphi) = x \arctan(1/2) + y \arctan 2$. All five identities are special cases of this general identity.

Challenge. Relabel the above figure and use $2\varphi - 3 = (1/\varphi^3)$ and $2\varphi - 1 = \sqrt{5}$ to prove the following:

(1) $\arctan \varphi^3 = (1/2) \arctan(1/2) + \arctan 2$,
(2) $\arctan \varphi^3 - \arctan(1/\varphi^3) = \arctan 2$,
(3) $2 \arctan(1/\varphi^3) = \arctan(1/2)$,
(4) $\arctan \varphi^3 + (1/2) \arctan(1/2) = (\pi/2) = 2 \arctan(1/\varphi^3) + \arctan 2$, and
(5) $\arctan \varphi^3 = \pi/4 + (1/2) \arctan 2$.

For another proof without words that relates the arctangent of 2 to the arctangent of the reciprocal of the golden ratio, see [**1**].

REFERENCE

[1] Plaza, Á. (2017). Proof without words: Arctangent of two and the golden ratio. *Math. Mag.* 90(3): 179.

Summary. Using an isosceles triangle with its heights from the vertex and a base angle, we illustrate some arctangent relationships among 2, the golden ratio and their reciprocals.

REX H. WU (MR Author ID: 1293646) is interested in the visualization of mathematical identities and concepts. He is also interested in the history of trigonometry.

A Modern Solution to the Gion Shrine Problem

J. ARIAS DE REYNA
Universidad de Sevilla
Sevilla, Spain
arias@us.es

DAVID CLARK
Randolph-Macon College
Ashland, Virginia, 23005
davidclark@rmc.edu

NOAM D. ELKIES
Harvard University
Cambridge, Massachusetts, 02138
elkies@math.harvard.edu

If you had visited Kyoto's Gion shrine around the middle of the eighteenth century, you might have noticed a wooden tablet, inscribed with geometric figures, hanging from one of the eaves. This may not have been a great surprise, since such *sangaku* (literally "mathematical tablets") were not uncommon in Japanese temples and shrines at the time. However, this particular tablet happened to hold a problem that would rise to great fame among a generation of Japanese mathematicians.

Problem. We have a segment of a circle. The line segment m bisects the arc and chord AB. As shown, we draw a square with side s and an inscribed circle of diameter d. Let the length $AB = a$. Then, if

$$p = a + m + s + d \quad \text{and} \quad q = \frac{m}{a} + \frac{d}{m} + \frac{s}{d}$$

and p and q are given, find a, m, s, and d.

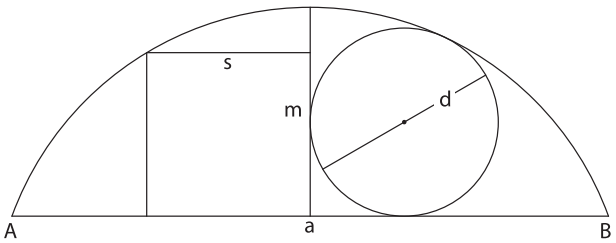


Figure 1 Gion shrine problem.

The algebraic particulars of this challenge might strike modern mathematicians as odd, but it has roots in a fascinating niche of mathematical history.

In this paper, after taking a brief look at Japan's past, we return to the present and offer a new solution to the Gion shrine problem. We also address questions of existence and uniqueness, which curiously lead us back to a result of Pierre de Fermat.

Math. Mag. **92** (2019) 110–122. doi:10.1080/0025570X.2019.1557007
MSC: Primary 01A27, Secondary 51M04; 11D41
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

Background

Eighteenth-century Japan, unified under the Tokugawa shogunate, was a relatively peaceful place where artistic recreation flourished. Those with leisure time indulged in the emerging arts of kabuki and bunraku theater, haiku poetry, and ukiyo-e woodblock printing—and a homegrown style of mathematics called *wasan* (*wa-* = "Japanese" + *-san* = "mathematics"). Practitioners of *wasan* tended to gravitate toward the aesthetics of geometry, and proved wonderful (though esoteric) results about packings of circles, polygons, and ellipses, as well as analogous problems in three dimensions. When a collection of theorems was deemed especially beautiful, it would be inscribed on a *sangaku* and hung in a Buddhist temple or Shinto shrine—as an offering to the gods, a challenge to other worshippers, and an advertisement for the school producing the work.

At the same time, the shogunate's policy of *sakoku* ("closed country") kept Japan intellectually distant from the scientific revolution of the West. The result was an insulated discipline that relied heavily on two sources of established knowledge: the planar geometry results of the Greeks and the rich body of mathematics imported from China, both of which had long been present in Japanese mathematics. *Sangi* computing rods, a notable Chinese technology, allowed for the numerical computation of roots of polynomials, and were used extensively in Japan. For more about traditional Japanese mathematics, see [**9**], [**10**], and [**14**].

The Gion shrine problem exhibits both geometrical aesthetics and an opportunity to harness the computational power of *sangi*. Tsuda Nobuhisa solved the problem first by deriving a polynomial of degree $1024 = 2^{10}$ from whose roots one could derive the result; his solution appeared on a *sangaku* hung from the Gion shrine in 1749. Subsequent progress was made by a mathematician named Nakata, who was able to reduce the necessary polynomial degree to 46. However, a celebrated breakthrough was made by Ajima Naonobu, who in a 1774 handwritten manuscript entitled *Kyoto Gion Gaku Toujyutsu* (literally "The Solution to the Kyoto Gion *Sangaku*") presented a degree ten polynomial solution. Ajima's derivation was first published in 1966 [**1**], and has since received a modern analysis [**8**].

Figure 2 Yasaka (formerly Gion) shrine in Kyoto, Japan.

Strikingly, Ajima's approach uses no geometric techniques more sophisticated than the Pythagorean theorem. With a great deal of algebraic persistence, he is able to manipulate a few basic geometric relations into a system of high degree equations in a and d. A clever substitution yields four cubic equations in a single variable X whose coefficients are given in terms of a, p, and q. This can be viewed as a homogeneous linear system with nontrivial solution $(X^3, X^2, X, 1)$; any such system must have determinant zero. Ajima then uses a technique equivalent to Laplace's method of cofactor expansion (c. 1776) to arrive at a polynomial equation of degree 10 in a, which requires nearly a full page to write out completely. It should be noted that, because of *sakoku*, Ajima (1732–1798) may not have even heard of Laplace (1749–1827), and likely was unaware of his results.

The solution given in this paper also has the form of a tenth degree polynomial. In contrast to Ajima's, ours makes extensive use of trigonometric functions, though it should be noted that eighteenth-century Japanese mathematicians had inherited a basic understanding of trigonometry from the Chinese. Indeed, trigonometric tables and a version of the Law of Cosines have been found in Tokugawa period (1603–1868) documents; see [10] and [11]. An advantage of this approach is that it allows for greater geometrical insight—and yields a polynomial that can comfortably be written in two lines, thanks in large part to a choice of variables more convenient than Ajima's. We also show existence and uniqueness of solutions (an issue Ajima does not seem to have addressed), and that, in general, for rational p and q, the numbers a, m, s and d are contained in an extension of $\mathbb{Q}$ of degree 20. We then consider the problem of realizing the Gion configuration with rational lengths, and prove that this is impossible using a result of Fermat pivotal to the history of (Western) number theory.

Solution

In contrast to Ajima's solution, given in the form of a polynomial in a, ours uses a new variable t. While t arises somewhat mysteriously from a series of ad-hoc substitutions, we shall ultimately see that in fact $t = d/a$.

Solution 1. We start by fixing the constants

$$q_0 = -3 + \frac{3\sqrt{5}}{2} + \tfrac{1}{2}\sqrt{\tfrac{1}{2}\,(125 - 41\sqrt{5})} \approx 2.394972$$

and

$$t_0 = \tfrac{1}{2}\left(1 - \sqrt{5} + \sqrt{2(5 - \sqrt{5})}\right) \approx 0.557537.$$

Given p and q, with $2 < q \le q_0$, we first find the unique solution $t \in (0, t_0]$ of the equation

$$8t^{10} + (16q - 33)t^8 + 16t^7 + (8q^2 - 49q + 56)t^6 + (16q - 33)t^5$$
$$- (16q^2 - 55q + 39)t^4 - (16q - 22)t^3 + (8q^2 - 23q + 18)t^2 - t + q - 2 = 0.$$

Then, using t, we compute the quantities

$$m' = 16t^2, \quad d' = 16t^2(1 - t^2), \quad a' = 16t(1 - t^2),$$
$$s' = -1 + 6t^2 - t^4 + \sqrt{1 + 20t^2 - 26t^4 + 20t^6 + t^8},$$

and define $p' = a' + m' + s' + d'$. Finally, the desired quantities will be

$$a = \frac{p}{p'}a', \quad m = \frac{p}{p'}m', \quad s = \frac{p}{p'}s', \quad d = \frac{p}{p'}d'.$$

Proof. From the definitions of p and q, one immediately sees that scaling all lengths by λ changes p to λp but leaves q invariant. As such, the problem is, for all practical purposes, independent of the overall scale. For convenience, we first seek a solution in which the radius of the circular arc is 1.

Observe that the angle φ (see Figure 3) determines the circular segment and, as we shall see, the solution. But for the problem to have a solution, the angle φ is limited to

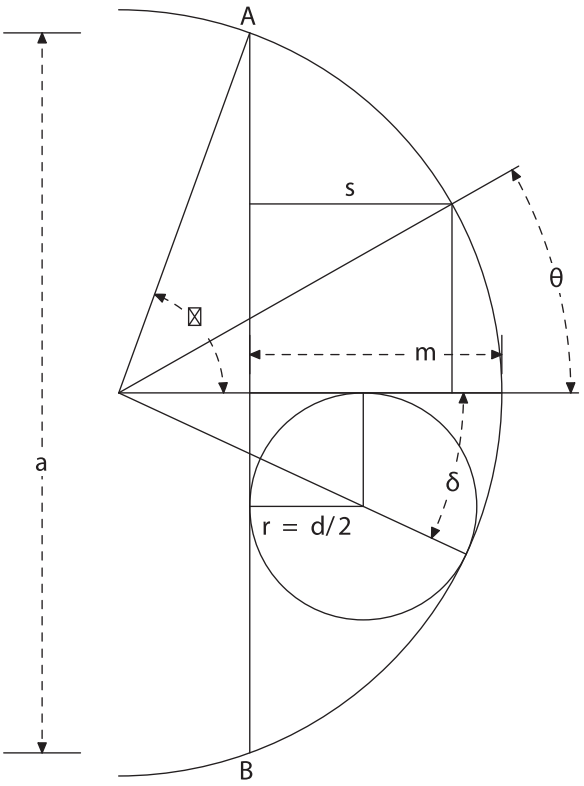


Figure 3 Construction for the solution. We first assume that the radius of the circular arc with center C is 1.

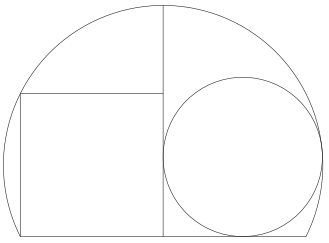


Figure 4 The extreme case.

the interval $0 < \varphi \le \varphi_0 = \frac{\pi}{2} + \arctan 1/2 \approx 117°$. If $\varphi > \varphi_0$, the square will fail to fit inside the segment. Figure 4 shows the limiting case in which $\varphi = \varphi_0$. Not surprisingly, the given value of q completely determines φ. We shall see this correspondence shortly (Figure 5).

Our first goal will be to express each of the quantities d, m, a, and s in terms of a single variable; the quantity $r = d/2$, the radius of the small circle, happens to be a convenient choice. From Figure 3 we have

$$a = 2 \sin \varphi, \tag{1}$$

$$s = \cos \theta - \cos \varphi = \sin \theta. \tag{2}$$

From (2), we find that $\cos\theta = \cos\varphi + \sin\theta$. Thus

$$\cos^2\theta = \cos^2\varphi + 2\cos\varphi\sin\theta + \sin^2\theta,$$

and finally

$$2\sin^2\theta + 2\cos\varphi\sin\theta - \sin^2\varphi = 0.$$

This equation, quadratic in $\sin\theta$, has two solutions whose product is negative. Since in our case $\sin\theta > 0$, we must take the greater of the two solutions,

$$\sin\theta = -\tfrac{1}{2}\cos\varphi + \tfrac{1}{4}\sqrt{4\cos^2\varphi + 8\sin^2\varphi},$$

from which it follows that

$$\sin\theta = \tfrac{1}{4}\sqrt{4 + 4\sin^2\varphi} - \tfrac{1}{2}\cos\varphi = \tfrac{1}{4}\sqrt{8 - 4\cos^2\varphi} - \tfrac{1}{2}\cos\varphi. \tag{3}$$

The angle $\delta > 0$ in Figure 3 can be described by the equations

$$(1-r)\cos\delta - r = 1 - m \quad \text{and} \quad \sin\delta = \frac{r}{1-r},$$

where $0 < r < 1/2$. It follows that

$$(1-r)\sqrt{1 - \frac{r^2}{(1-r)^2}} - r = 1 - m,$$

whence

$$m = 1 + r - \sqrt{1-2r}. \tag{4}$$

We also have

$$m = 1 - \cos\varphi,$$

from which it follows that

$$\cos\varphi = -r + \sqrt{1-2r}. \tag{5}$$

Taking a brief digression, we can use (1), (2), (3), and (5) to write q in terms of φ. A plot of the implicit function $\varphi(q)$ is given in Figure 5.

Since $0 < \varphi \le \varphi_0$ we have $1 > \cos\varphi \ge -1/\sqrt{5} = \cos\varphi_0$. Then (5) implies that

$$0 < r \le r_0 = -1 + \frac{1}{\sqrt{5}} + \sqrt{2 - \frac{2}{\sqrt{5}}}.$$

Equation (5) also tells us that

$$\begin{aligned} \sin^2\varphi &= 1 - \left((1-2r) + r^2 - 2r\sqrt{1-2r}\right) \\ &= 2r - r^2 + 2r\sqrt{1-2r}, \end{aligned} \tag{6}$$

which, when combined with (3) and (5), yields

$$\sin\theta = \tfrac{1}{2}\left(r - \sqrt{1-2r} + \sqrt{1 + 2r - r^2 + 2r\sqrt{1-2r}}\right). \tag{7}$$

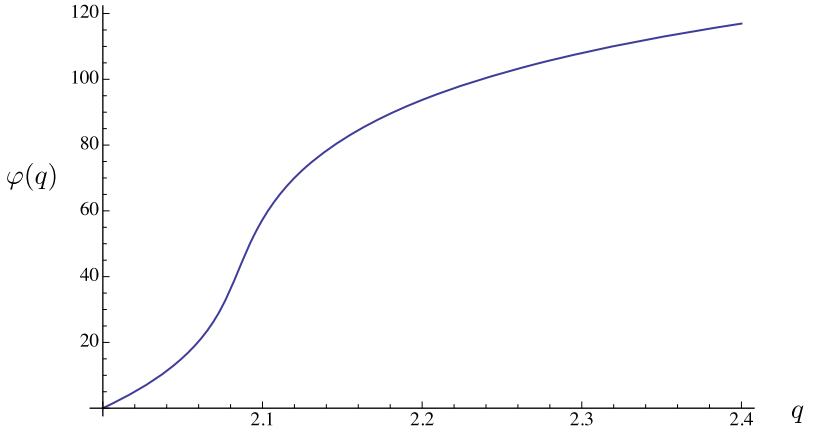


Figure 5 A given q determines the angle φ, in degrees.

Bringing together our results from (1), (2), (4), (6), and (7), we now have the problem's four desired quantities in terms of r:

$$\begin{aligned} d &= 2r, \\ m &= 1 + r - \sqrt{1-2r}, \\ a &= 2\sqrt{2r - r^2 + 2r\sqrt{1-2r}}, \\ s &= \tfrac{1}{2}\left(r - \sqrt{1-2r} + \sqrt{1 + 2r - r^2 + 2r\sqrt{1-2r}}\right). \end{aligned} \tag{8}$$

At this stage, we could write down a polynomial equation relating q and r, plug in the given value of q and solve for $r \in (0, r_0]$ (using *sangi*, or *Mathematica*), and use this value to recover the quantities d, m, a, and s. But such an equation would have an unnecessarily large degree due to the numerous radicals in (8). Thus we first eliminate some of these radicals with a sequence of changes of variable. First, define $x > 1$ by $2r = x(2-x)$. The inequality $0 < r \le r_0 < 1/2$ guarantees that $2 > x \ge x_0 = 1 + \sqrt{1-2r_0} \approx 1.051462$. We can write the above equations (8) in terms of x:

$$\begin{aligned} d &= x(2-x) \\ m &= \tfrac{1}{2}(4 - x^2) \\ a &= x\sqrt{4-x^2} \\ s &= \tfrac{1}{4}\left(2 - x^2 + \sqrt{4 + 4x^2 - x^4}\right). \end{aligned} \tag{9}$$

We next remove the radical $\sqrt{4-x^2}$ using the following rational parametrization of the circle $x^2 + y^2 = 4$ by a new variable $t = (2-x)/y$:

$$x = 2\frac{1-t^2}{1+t^2}, \qquad y = \sqrt{4-x^2} = \frac{4t}{1+t^2}.$$

As an aside, these x and y are obtained by doubling the coordinates of the well-known rational parametrization of the unit circle $x^2 + y^2 = 1$; a further consequence is that the sides of any right triangle are proportional to (t^2-1), $2t$, and (t^2+1), for some rational number t. See for example [**4**, §6.1, pp. 58–60] for an overview.

The equations for x and y give us $t^2 = (2-x)/(2+x)$, and it is easy to see that when $x \in [x_0, 2)$ we get $t^2 \in (0, t_0^2)$. Thus we must restrict t so that

$$0 < t \le t_0 = \tfrac{1}{2}\left(1 - \sqrt{5} + \sqrt{2(5-\sqrt{5})}\right) \approx 0.557537. \tag{10}$$

In terms of the new variable t equations (9) become

$$\begin{aligned} d &= \frac{8t^2(1-t^2)}{(1+t^2)^2}, \quad m = \frac{8t^2}{(1+t^2)^2}, \quad a = \frac{8t(1-t^2)}{(1+t^2)^2}, \\ s &= \frac{-1+6t^2-t^4+\sqrt{1+20t^2-26t^4+20t^6+t^8}}{2(1+t^2)^2}. \end{aligned} \tag{11}$$

We simplify these expressions further by changing the overall scale of the problem, choosing the radius of the circular arc to be $2(1+t^2)^2$ instead of 1. Our four quantities d, m, a, s are then given by

$$\begin{aligned} d &= 16t^2(1-t^2), \quad m = 16t^2, \quad a = 16t(1-t^2), \\ s &= -1+6t^2-t^4+\sqrt{1+20t^2-26t^4+20t^6+t^8}. \end{aligned} \tag{12}$$

Observe that the new variable t equals $2r/a$.

Putting everything together, we have

$$\begin{aligned} p &= a+m+s+d \\ &= -1+16t+38t^2-16t^3-17t^4+\sqrt{1+20t^2-26t^4+20t^6+t^8}, \\ q &= \frac{m}{a}+\frac{d}{m}+\frac{s}{d} \\ &= \frac{-1+22t^2+16t^3-33t^4+16t^6+\sqrt{1+20t^2-26t^4+20t^6+t^8}}{16t^2(1-t^2)}. \end{aligned} \tag{13}$$

Equation (13), relating q and t, can be rewritten as

$$\begin{aligned} &\left(16t^2(-1+t^2)q + (-1+22t^2+16t^3-33t^4+16t^6)\right)^2 \\ &\qquad = 1+20t^2-26t^4+20t^6+t^8, \end{aligned}$$

and further expanded to $32t^2P(t,q) = 0$ where

$$\begin{aligned} P(t,q) &= 8t^{10} + (16q-33)t^8 + 16t^7 + (8q^2-49q+56)t^6 + (16q-33)t^5 \\ &\quad - (16q^2-55q+39)t^4 - (16q-22)t^3 + (8q^2-23q+18)t^2 - t + q - 2 = 0. \end{aligned}$$

Suppose then that we are given p and q. We first solve for t in $P(t,q) = 0$, choosing a root $t \in (0, t_0]$. Then equations (12) recover quantities a', m', s', and d' that correspond to q. To get the correct p we need only rescale the solution by the reciprocal of $p' = a' + m' + s' + d'$. This is precisely the procedure indicated in Solution 1. ■

Not all values of p and q are allowed. We can plot q in terms of t as given by equation (13); see Figure 6. We see that q varies on the interval $2 < q \le q_0$ where

$$q_0 = q(t_0) = -3 + \frac{3\sqrt{5}}{2} + \tfrac{1}{2}\sqrt{\tfrac{1}{2}(125-41\sqrt{5})} \approx 2.394972. \tag{14}$$

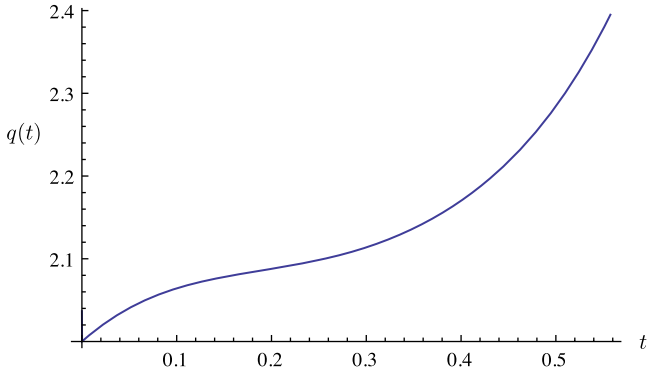


Figure 6 Plot of $q(t)$.

If q is not in the interval $(2, q_0]$, it cannot correspond to an allowable value of t; in this case there is no solution to the problem.

On the other hand, $q(t)$ is an increasing function on the interval $[0, t_0]$. To show this, we compute and rationalize $q'(t)$. The inequality $q'(t) > 0$ is equivalent to one of the form $u(t) > 0$ for a polynomial $u(t)$; the latter inequality is easy to check. Therefore, for each q in the interval $(2, q_0]$ there is a unique positive $t \leq t_0$ with $P(q, t) = 0$.

To summarize, given a pair of positive real numbers p and q:

- if $q \leq 2$ or $q > q_0$, there is no solution to the Gion shrine problem;
- if $2 < q \leq q_0$, there is exactly one solution.

For many values of $q \in \mathbb{Q}$, the polynomial $P(q, t)$ is irreducible. For example, this happens with $q = 9/4$. In such a case, the corresponding t generates a field $\mathbb{Q}(t)$ that is a degree 10 extension of $\mathbb{Q}$. Here the numbers a', m', s' and d' are contained in $\mathbb{Q}(t)$ or an extension of degree 2 of $\mathbb{Q}(t)$. If we choose p to be rational, the same occurs with the final numbers a, m, s, and d. These numbers might be contained in a proper subfield of $\mathbb{Q}(t)$; but since $t = 2r/a$, this subfield can be only $\mathbb{Q}(t)$ or a degree-2 extension of $\mathbb{Q}(t)$. Thus, the solutions generate a field extension of degree 20 or possibly 10.

Rational solutions

In *yosan* (*yo-* = "Western"), following a tradition that goes at least as far back as Diophantus and remains productive in modern number theory (which also incorporates contributions from modern Japan and elsewhere), one is often interested in the existence of integral or rational solutions to geometrical problems. So we ask: *Is there any rational solution of the Gion shrine problem, that is, a solution for which each of the quantities p, q, a, m, s, d is a rational number?* We shall show that there is no such solution, and that by a happy coincidence this fact is obtained from a theorem of Fermat that forms a key juncture in the history of *yosan* number theory.

Given our analysis in the previous section, we soon see that a rational solution of the Gion shrine problem is tantamount to a rational value of t in our range $0 < t \leq t_0$ for which

$$s' + 1 - 6t^2 + t^4 = \sqrt{1 + 20t^2 - 26t^4 + 20t^6 + t^8}$$

is also rational; that is, a rational point on the algebraic curve

$$C : u^2 = 1 + 20t^2 - 26t^4 + 20t^6 + t^8 \tag{15}$$

with $0 < t \leq t_0$. We shall show the following proposition.

Proposition. *The only solutions in rational numbers to equation in (15) are the pair* $(t, u) = (0, \pm 1)$ *and the quadruple* $(t, u) = (\pm 1, \pm 4)$.

In particular it will follow there are none with $0 < t \leq t_0$.

Proof. If $t = 0$ then clearly $u = \pm 1$. Assume, then, that (t, u) is a rational solution of equation (15) with $t \neq 0$. Divide both sides by t^4 to obtain

$$\left(u/t^2\right)^2 = t^{-4} + 20t^{-2} - 26 + 20t^2 + t^4 = \left(t - \frac{1}{t}\right)^4 + 24\left(t - \frac{1}{t}\right)^2 + 16.$$

Thus $(t - (1/t),\ u/t^2)$ is a rational solution of the equation

$$U^2 = T^4 + 24T^2 + 16, \tag{16}$$

with $T = 0$ if and only if $t = \pm 1$. We shall show that there are no rational solutions of equation (16) other than $(T, U) = (0, \pm 4)$, which will prove our proposition.

We next eliminate the T^4 term from equation (16) by introducing a new variable $\delta = T^2 - U$. Thus $U = T^2 - \delta$ and $T^4 + 24T^2 + 16 = (T^2 - \delta)^2$, whence

$$(24 + 2\delta)T^2 = \delta^2 - 16 = (\delta - 4)(\delta + 4), \tag{17}$$

so $(24 + 2\delta)(\delta - 4)(\delta + 4)$ is a rational square. The further change of variable $\delta = 8X - 4$ makes $(24 + 2\delta)(\delta - 4)(\delta + 4) = (16X + 16)(8X - 8)(8X) = 2^{10}(X + 1)(X - 1)X$, so we obtain a rational solution of

$$Y^2 = (X + 1)(X - 1)X = X^3 - X. \tag{18}$$

But Fermat already proved that the only solutions of equation (18) with both variables rational are the three with $Y = 0$. Hence $\delta \in \{4, -4, -12\}$. But $\delta = -12$ is not possible in equation (17), and each of $\delta = \pm 4$ makes $T = 0$ as claimed. This completes the proof of our proposition. ∎

Fermat's theorem occupies a pivotal point in the history of number theory: it settled a centuries-old problem going back at least to Fibonacci; it yields the exponent 4 case of Fermat's last theorem, which is the only case for which we have a proof by Fermat himself; it is also the only example we have of a proof by Fermat using his method of descent; and this method remains the basis of our only technique for describing the group of rational points on an elliptic curve (see [**13**, Chapters VIII and X, especially X.1, X.4, and X.6]). The result is sometimes given in one of the two following equivalent forms.

1. *The area of a right triangle whose sides are integers is not a square.* This is how Fermat recorded the result in the margin of his copy of Diophantus' *Arithmetica* [**12**]. We noted already the sides of a right triangle are proportional to $t^2 - 1, 2t, t^2 + 1$ for some rational t; the sides are positive when $t > 1$. The area is thus $\alpha^2(t^3 - t)$ for some rational α, so the fact that this cannot be a square is a consequence of the fact that $Y^2 = X^3 - X$ has no rational solutions with $X > 1$. Note that this shows only one implication, but for the other direction we use the fact that $Y^2 > 0$ implies either $X > 1$ or $-1 < X < 0$, and in the latter case $(-1/X, Y/X^2)$ satisfies the same equation with $-1/X > 1$.

2. *There is no nonconstant three-term arithmetic progression of squares whose common difference is a square.* Such a progression would have the form $x^2 - y^2$, x^2, $x^2 + y^2$ for integers x, y with $0 < y < x$, and then $X = x^2/y^2$ would satisfy $X^3 - X = (zxz'/y^3)^2$ where $x^2 - y^2 = z^2$ and $x^2 + y^2 = z'^2$. Again we have proven only the implication that obtains the new formulation of the result from the impossibility of $Y^2 = X^3 - X$ in nonzero $X, Y \in \mathbb{Q}$. To go the other way, we may note that if $X^3 - X$ is a nonzero square then $(X^2 - 2X - 1)^2$, $(X^2 + 1)^2$, $(X^2 + 2X - 1)^2$ is a three-term arithmetic progression of squares whose common difference is the square $4(X^3 - X)$.

The common difference of a three-term arithmetic progression of squares is said to be "congruent" (because the common difference of an arithmetic progression was called its "congruum"); so this form of Fermat's theorem was stated as *no [nonzero] square is a congruent number*.

We can now appreciate Dickson's words on Fermat's results at the start of chapter XXII ("Equations of degree four" on pp. 615 ff.) of his *History of the Theory of Numbers, Vol. II: Diophantine Analysis* [3]. Departing from his usual telegraphic presentation, Dickson devotes more than a page to Fermat's original proof and an introduction, which we quote next:

> Leonardo Pisano [=Fibonacci] recognized the fact, but gave an incomplete proof, that no square is a congruent number (i. e., $x^2 + y^2$ and $x^2 - y^2$ are not both squares) [...] Four centuries later, Fermat[2] stated and proved the result thus implied by Leonardo: *no right triangle with rational sides equals a square with a rational side.* The occasion was the twentieth of Bachet's problems inserted at the end of book VI of Diophantus: to find a right triangle whose area is a given number A. [...]
>
> Fermat's proof is of especial interest as it illustrates in detail his method of infinite descent and as it presents the only instance of a detailed proof left by him. [...]

Footnote 2 refers to "Fermat's marginal notes in his copy of Bachet's edition of Diophantus' Arithmetica; Oeuvres de Fermat, Paris, 1, 1891, 340; 3, 1896, 271." See [7]. The ensuing reproduction of Fermat's proof concludes with the sentence "The margin is too narrow for the complete demonstration and all its developments" which echoes the famous marginal note recording what we now know as Fermat's last theorem.

The case $n = 4$ of Fermat's assertion on $x^n + y^n = z^n$ quickly follows from his result on $Y^2 = X^3 - X$. Indeed for rational y, z it is not possible for $z^4 - y^4$ to be a nonzero square, let alone a fourth power, because then we could set $X = (z/y)^2$, and $X^3 - X = (z/y^3)^2(z^4 - y^4)$ would be a nonzero square. To complete this circle of allusions we note that it was the paragraph on Pythagorean triangles in Diophantus (whose parametrization we used in our solution of the Gion shrine problem) in whose margin Fermat recorded that such triangles cannot have square area.

We conclude with some general remarks on Diophantine equations such as equation (15), to give more context to the particular curve C and our determination of all its rational points.

In general if $D(t)$ is a nonconstant polynomial without repeated roots then the equation $u^2 = D(t)$ defines a hyperelliptic curve of genus $g = \lceil \frac{1}{2} \deg D \rceil - 1$. (Note that any polynomial D can be factored uniquely as $F^2 D_1$, where F is monic and D_1 has no repeated roots, and then $D(t)$ is a square if and only if $D_1(t)$ is a square or $F(t) = 0$; thus the assumption that D has distinct roots loses no real generality. Further, we use the expansive sense of "hyperelliptic curve" that allows genus 1 (elliptic curve) and 0 (conic curve).) Hence our C has $g = \lceil 8/2 \rceil - 1 = 3$.

Since $g > 1$, Faltings' theorem (proof of the Mordell conjecture) [2] guarantees that there are only finitely many solutions. Unfortunately both of Faltings' proofs [5, 6], and all later proofs based on them, are fundamentally "ineffective": they prove that the list of rational points is finite, but provide no general algorithm guaranteed to compute this list. In some cases one can prove that a curve has no rational solutions using elementary inequalities or congruences; for example, $-1 - t^4$ is never a square because it is negative for all $t \in \mathbb{Q}$, nor is $-1 + 3t^4$ by reduction modulo powers of 3 (or of 2). But these methods cannot apply to our C, because it does have a few rational points outside the range $0 < t \le t_0$, such as $(t, u) = (1, \pm 1)$. Once we find a few rational points on C it is often very difficult to prove that there are no others.

Fortunately our C was still tractable, thanks in part to having more automorphisms than the "hyperelliptic involution" $h : (t, u) \leftrightarrow (t, -u)$ shared by all hyperelliptic curves. Because $D(t)$ is an even polynomial, C also has the involution $i_1 : (t, u) \leftrightarrow (-t, u)$; because D is "palindromic" (the coefficients read the same in each direction), it satisfies $D(t) = t^{\deg D} D(1/t)$, and since $\deg D$ is even this yields a further involution $i_2 : (t, u) \leftrightarrow (1/t, u/t^4)$. (This involution fixes the points $(\pm 1, \pm 4)$, but pairs the points $(0, \pm 1)$ with "points of infinity" of C when C is regarded as a projective curve.) Even these do not generate the full group of automorphisms of C, which includes also the somewhat unexpected involution

$$i_3 : (t, u) \leftrightarrow \left(\frac{t+1}{t-1}, \frac{4u}{(t-1)^4} \right); \tag{19}$$

which switches the $t = 0$ and $t = 1$ points (and switches the $t = -1$ points with the pair of points at infinity).

In general a curve $u^2 = D(t)$ with $\deg(D) = 8$ has an action of i_1 and i_2 if and only if

$$D(t) = \alpha t^8 + \beta t^6 + \gamma t^4 + \beta t^2 + \alpha$$

for some α, β, γ. The fractional linear transformation $t \mapsto (t + 1)/(t - 1)$ conjugates each of $t \mapsto -t$ and $t \mapsto 1/t$ to the other, and thus normalizes the group generated by these two involutions. Hence $t \mapsto (t + 1)/(t - 1)$ (which is itself an involution) transforms a hyperelliptic curve $u^2 = \alpha t^8 + \beta t^6 + \gamma t^4 + \beta t^2 + \alpha$ into another curve of the same form (as can also be seen by direct calculation). We thus tried this substitution on our curve C, which has $(\alpha, \beta, \gamma) = (1, 20, -26)$, hoping to find a simpler equation. To our surprise we instead found C itself, and thus produced an extra automorphism of C.

It can be shown that h, i_1, i_2, i_3 do generate $\mathrm{Aut}(C)$; but we did not need this fact, nor the involution (19) itself, to prove the proposition.

There are several choices of subgroups H of $\mathrm{Aut}(C)$ for which the quotient curve $E = C/H$ has genus 1. Unlike curves of genus > 1, a genus-1 curve may have either finitely or infinitely many rational points; but when there are finitely many points we can often list them all even when the list is nonempty. (In the remaining case of genus zero, a smooth curve with a single rational point has a rational parametrization and therefore has infinitely many rational points; so we cannot use genus-zero quotients for our purpose.) Fortunately this happened for our C with $H = \{1, i_1 i_2\}$. Since the quotient map $C \to E$ takes rational points to rational points, we could then find all the rational points on C by computing the preimages of each rational point of E. (There are four rational points when we include points at infinity; of the four, three are finite on the models (17, 18), and two on (16), with one of the other two mapping to the point with $\delta = -12$ in (17).) Indeed this is quite similar to the proof of the exponent 4

case of Fermat: the curve $x^4 + y^4 = z^4$ defines a projective plane curve F_4 of genus 3 (though not a hyperelliptic one), with several involutions including $i : (x : y : z) \leftrightarrow (-x : y : z)$, and the quotient curve $C/\{1, i\}$ has genus 1 and admits a nonconstant map to $Y^2 = X^3 - X$. One can thus use Fermat's result on this curve to find all rational solutions of $x^4 + y^4 = z^4$.

Naturally some relevant details are different between our approach and Fermat's exponent 4 case. It so happens that both curves have "good reduction outside 2", remaining smooth mod p for all odd primes p. For instance, for C this follows from the fact that the discriminant of $1 + 20t^2 - 26t^4 + 20t^6 + t^8$ is a power of 2 (namely 2^{64}). Thus the same is true for any quotient curve. But for the Fermat quartic, all the quotient curves of genus 1 have only finitely many points; for example, if we took the quotient by $(x : y : z) \leftrightarrow (x : y : -z)$ we would reach $Y^2 = X^3 + X$, and this curve has only two rational points, one at infinity and the other at $(0, 0)$, which is also proved by Fermat's method of descent. For our curve C, we had to be careful to avoid quotients such as $C/\{1, i_1\}$ and $C/\{1, i_2\}$ which have genus 1 but infinitely many rational points.

Acknowledgment The authors thank Tony Rothman, for his interest and helpful comments, and Hidetoshi Fukagawa, for sharing with us a small piece of his life's work on *sangaku* mathematics.

REFERENCES

[1] Ajima, N., Hirayama, A., Matsuoka, M. (1966). *Naonobu Ajima's Complete Works*. Publication Committee of Naonobu Ajima's Complete Works. Tokyo: Fuji Tanki Daigaku Shuppanbu, Shōwa 41.
[2] Bloch, S. (1984). The proof of the Mordell conjecture. *Math. Intelligencer*. 6(2): 41–47.
[3] Dickson, L. E. (1934). *History of the Theory of Numbers, Vol. II: Diophantine Analysis*. Mineola, NY: Stechert.
[4] Elkies, N. D. (2007). The ABC's of number theory. *Harvard College Mathematics Review.* 1(1): 58–60.
[5] Faltings, G. (1983). Endlichkeitssätze für abelsche Varietäten über Zahlkörpern. *Inventiones. Mathematicae* 73: 349–366.
[6] —. (1991). Diophantine approximation on Abelian varieties. *Annals of Mathematics.* 133(2): 549–576.
[7] de Fermat, P., Tannery, P., Henry, C., de Billy, J., Wallis, J., France. Ministère de l'éducation nationale. (1896). *Oeuvres de Fermat*. Oeuvres de Fermat, Gauthier-Villars et fils.
[8] Fukagawa, H. (1983). *Zoku zoku sangaku no kenkyū*. Nagoya-shi: Narumi Dofūkai.
[9] Fukagawa, H., Rothman, T. (2008). *Sacred Mathematics: Japanese Temple Geometry*. Princeton, NJ: Princeton University Press.
[10] Horiuchi, A. (2010). *Japanese Mathematics in the Edo Period (1600–1868)*. Translated from the 1994 French original by Silke Wimmer-Zagier. Science Network Historical Studies, Vol. 40, Basel: Birkhäuser Verlag.
[11] Hosking, R.J. (2016). Sangaku: A mathematical, artistic, religious, and diagrammatic examination. PhD dissertation. University of Canterbury, New Zealand.
[12] Kato, K., Kurokawa, N., Saito, T. (2000). *Number Theory 1: Fermat's Dream*. Translated from the 1996 Japanese original by Masato Kuwata. Translations of Mathematical Monographs. Vol. 186. Providence, RI: American Mathematical Society.
[13] Silverman, J. H. (1986). *The Arithmetic of Elliptic Curves*. New York: Springer.
[14] Smith, D. E., Mikami, Y. (2004). *A History of Japanese Mathematics*. Reprint of the 1914 original. Mineola, NY: Dover.

Summary. We give a new solution to the famous Gion shrine geometry problem from eighteenth-century Japan. Like the classical Japanese solution, ours is given in the form of a degree-ten equation. However, our polynomial has the advantage of being much easier to write down. We also provide some additional analysis, including a discussion of existence and uniqueness. Finally, we prove that the Gion shrine problem has no rational solutions.

J. ARIAS DE REYNA (MR Author ID: 27005) learned mathematics from books starting at age thirteen; at the time, even books were difficult to get in dictator Franco's Spain. He has published a book about Carleson's proof on the convergence of Fourier series, defining the largest known rearrangement invariant space of functions with almost everywhere convergent series. He also obtained good bounds for the Riemann–Siegel expansion.

DAVID CLARK (MR Author ID: 867065) was trained as a quantum topologist, but has recently become interested in the history of Japanese mathematics, and more broadly the exchange of mathematical ideas across cultural barriers. He has twice taken students to Japan to learn about *sangaku* tablets, contemplate Zen gardens, and watch sumo wrestling.

NOAM D. ELKIES (MR Author ID: 229330) is a number theorist, much of whose work concerns Diophantine geometry and computational number theory. He was granted tenure at Harvard at age 26, the youngest in the University's history. Outside of math, Elkies' main interests are music—mainly classical piano and composition—and chess, where he specializes in composing and solving problems.

Math Bite: When an Average of Averages is the Average

If you want to find the mean of a data set, you would not, say, split the data in half, find the average of each half, and average those results. The average of a data set is generally not obtained by averaging the averages of some of its subsets, unless one considers *all* possible subsets of a *fixed* size!

In particular, suppose the data is given by $\{a_1, ..., a_n\}$. Given $1 \leq k \leq n$, there are $\binom{n}{k}$ subsets of size k, and any a_i lies in $\binom{n-1}{k-1}$ of them. Thus, averaging the averages of all k-subsets produces

$$\frac{\frac{\binom{n-1}{k-1}(a_1+\cdots+a_n)}{k}}{\binom{n}{k}} = \frac{a_1+\cdots+a_n}{n},$$

via an application of the well-known identity $n\binom{n-1}{k-1} = k\binom{n}{k}$.

—Contributed by Tristen Pankake-Sieminski (MR Author ID: 1297564) Aturian LLC, Lake Forest, IL and Raymond Viglione (MR Author ID: 704098) School of Mathematical Sciences, Kean University, Union, NJ.

A Probability Perspective to a Combinatorics Problem

HIDEO HIROSE
Data Science Research Center
Hiroshima Institute of Technology
h.hirose.tk@it-hiroshima.ac.jp

This paper gives a simple proof that the number of solutions to

$$x_1 + x_2 + \cdots + x_n \equiv k \pmod{p} \tag{1}$$

is p^{n-1} for each $k \in \{0, 1, \ldots, p-1\} = [p-1]$, where $x_i \in [p-1]$ for $i = 1, \ldots, n$.

For example, when $p = 2$, and $n = 2$, then the solutions to $x_1 + x_2 \equiv 0 \pmod 2$ are $(x_1, x_2) = (0, 0)$ and $(1, 1)$, while the solutions to $x_1 + x_2 \equiv 1 \pmod 2$ are $(x_1, x_2) = (0, 1)$ and $(1, 0)$. The number of solutions is $2^{2-1} = 2$ in each case. As another example, when $p = 2$, and $n = 3$, then the solutions to $x_1 + x_2 + x_3 \equiv 0 \pmod 2$ are $(x_1, x_2, x_3) = (0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$, and $(1, 1, 0)$; and, the solutions to $x_1 + x_2 + x_3 \equiv 1 \pmod 2$ are $(x_1, x_2, x_3) = (0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$, and $(1, 1, 1)$. In each case, the number of solutions is $2^{3-1} = 4$. It seems to be cumbersome to write down the solutions for general p, k, and n.

Equation (1) is the same as

$$x_1 + x_2 + \cdots + x_n = k + jp \quad (0 \le j < n), \tag{2}$$

which can be regarded as an extension of the classic combinatorics problem of showing that the number of nonnegative integer solutions to

$$x_1 + x_2 + \cdots + x_n = k \tag{3}$$

is $\binom{n+k-1}{k}$, as in [2]. For example, the solution $x_1 = 2$, $x_2 = 1$, $x_3 = 3$ to $x_1 + x_2 + x_3 = 6$ can be represented as $11 * 1 * 111$, where the six 1s are allocated among the $8\ (= 6 + 3 - 1)$ places. This is equivalent to the allocation of $2\ (= 3 - 1)$ "stars" (given by the *s) into the eight places. (This method is referred to as the "stars and bars" method in [1].) The number of ways to allocate the k 1s among the $n + k - 1$ places is $\binom{n+k-1}{k}$. However, the restriction of the x_i's to $[p-1]$ seems to preclude this approach to determine the number of solutions to equation (1).

A more general problem, sometimes called the "donut shop problem," could be useful to find the number of nonnegative integer solutions to equation (3) under inequality constraints, where $0 \le a_i \le x_i \le b_i$ for $i = 1, \ldots, n$, as considered in [1, 8]. This is equivalent to the problem of finding the number of nonnegative integer solutions to $y_1 + y_2 + \cdots + y_n = k - s$, where $0 \le y_i \le b_i - a_i$ for each i and $s = \sum_{i=1}^{n} a_i$. To obtain the solution, we can use the inclusion–exclusion principle, as in [3], to get the following formula

$$|A_1 \cup \cdots \cup A_n| = \sum_{\emptyset \ne I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |A_I|, \tag{4}$$

where $|\cdot|$ denotes the number of elements in a set and $A_I = \cap_{i \in I} A_i$.

Math. Mag. **92** (2019) 123–125. doi:10.1080/0025570X.2019.1549890
MSC: Primary 60C05, Secondary 05A19

For example, when $p = 6$, $x_1 + x_2 \equiv 1 \pmod 6$, and $x_i \in [5]$ for each i, there are two cases: $x_1 + x_2 = 1$ and $x_1 + x_2 = 7$. In the former case, the number of cases is $|\mathcal{U}| = \binom{2+1-1}{1} = 2$, where $|\mathcal{U}|$ represents the number of solutions without any restrictions; here $\mathcal{U}$ can be thought of as the universal set. In the latter case, $|\mathcal{U}| = \binom{2+7-1}{7} = 8$, $|A_1| = \binom{2+(7-6)-1}{7-6} = 2 = |A_2|$, and $|A_1 \cap A_2| = \emptyset$, thus the number of solutions is $8 - 2 - 2 = 4$. By combining these two cases, the total number of solutions is $2 + 4 = 6$.

Similarly, when $x_1 + x_2 \equiv 2 \pmod 6$, we deal with two cases: $x_1 + x_2 = 2$ and $x_1 + x_2 = 8$. In the former case, the number of solutions is $|\mathcal{U}| = \binom{2+2-1}{2} = 3$. In the latter case, $|\mathcal{U}| = \binom{2+8-1}{8} = 9$, $|A_1| = \binom{2+(8-6)-1}{8-6} = 3 = |A_2|$, and $|A_1 \cap A_2| = \emptyset$, thus the number of solutions is $9 - 3 - 3 = 3$. By combining these two cases, there are $3 + 3 = 6$ solutions.

This method seems to be cumbersome, too. Determining the number of solutions to equation (1) seems to be tough to solve using a combinatorial approach.

Consider the following probabilistic interpretation of $x_1 + x_2 \equiv k \pmod 6$. We define a rule to move counterclockwise around the unit circle starting at $(1, 0)$. Roll a six-sided die and when it shows j dots, then move $(j - 1)(2\pi/6)$ $(j = 1, 2, \ldots, 6)$ counterclockwise along the circumference from your current position. What is the probability that you will be at position $k(2\pi/6)$ $(k = 0, 1, \ldots, 5)$ along the unit circle after throwing the die two times?

In the next section, I show how counting the number of solutions to equation (1) is easier in the probabilistic context.

Probability problem

Suppose that the probability question from the previous section is generalized to a p-sided die, where rolling a $j \in [p - 1]$ results in moving counterclockwise $(j - 1)(2\pi/p)$ along the circumference. Each x_i from equation (1) corresponds to a roll of the p-sided die. Take a second to think about how determining the probability of landing on $k(2\pi/p)$ after n rolls is equivalent to the number of solutions to equation (1).

Theorem 1 (Probability)**.** *Let X_i be the random variable for which $P(X_i = j) = \frac{1}{p}$ for $j = 0, 1, \ldots, p - 1$. Define $S_n = \sum_{i=1}^{n} X_i$. Then, $P(S_n \equiv k \pmod p) = \frac{1}{p}$ for $k = 0, 1, \ldots, p - 1$.*

Proof. When you throw a p-sided die one time, the probability that you are at the position $k(2\pi/p)$ is $\frac{1}{p}$ for all k $(k = 0, 1, \ldots, p - 1)$. This means that a second throw provides equal probability to each position you proceed regardless of the last position; this is the Markov property. Therefore, throwing a die n times also provides equal probability of $\frac{1}{p}$ to each position of $k(2\pi/p)$ for all k $(k = 0, 1, \ldots, p - 1)$. ■

Since each X_i in Theorem 1 has p possibilities, the total number of cases of $X_1, X_2, \ldots, X_n$ is p^n. Because the probability $P(S_n \equiv k \pmod p) = \frac{1}{p}$, then the total number of cases in which $S_n \equiv k \pmod p$ is p^{n-1} for all $k \in [p - 1]$.

Combinatorial problem

To interpret the probability problem to the combinatorial problem, we only regard $x_i \in \{0, 1, \ldots, p - 1\}$, $(i = 1, \ldots, n)$ in equation (1) as the samples from a discrete uniform distribution. This delivers the answer to the combinatorial problem.

Theorem 2 (Combinatorics)**.** *For $k \in [p-1]$ and $x_i \in [p-1]$ for each i, there are p^{n-1} solutions $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to*

$$x_1 + x_2 + \cdots + x_n \equiv k \pmod{p}.$$

Conclusion

In this note, we provided a relatively easy solution to a problem in combinatorics via a probabilistic approach. There are other instances in which probabilistic proofs of nonprobabilistic results are easier, too. See, for example, Rosalsky [7], Peterson [6], and Kataria [4, 5].

Acknowledgment The author would like to thank the Editor and referees for valuable comments.

REFERENCES

[1] Brualdi, R. A. (2017). *Introductory Combinatorics*, 5th ed. Englewood Cliffs, NJ: Pearson.
[2] Fomin, D., Genkin, S., Itenberg, I. V. (1996). *Mathematical Circles: Russian Experience*. Mathematical World, Vol. 7. Providence, RI: American Mathematical Society.
[3] Jukna, S. (2011). *Extremal Combinatorics*, 2nd ed. Heidelberg: Springer.
[4] Kataria, K. K. (2016). A probabilistic proof of the multinomial theorem. *Amer. Math. Monthly*. 123: 94–96.
[5] Kataria, K.K. (2017). Some probabilistic interpretations of the multinomial theorem. *Math. Mag.* 90: 221–224.
[6] Peterson, J. (2013). A probabilistic proof of a binomial identity. *Amer. Math. Monthly.* 120: 558–562.
[7] Rosalsky, A. (2007). A simple and probabilistic proof of the binomial theorem. *Amer. Statist.* 61: 161–162.
[8] University of North Dakota Mathematics Department (2010). Math408: Combinatorics, p. 9. http://arts-sciences.und.edu/math/_files/docs/courses/supp/math408notesold.pdf

Summary. A problem that seems to be tough in a field sometimes becomes easy to solve by looking at it from a different field. In this note, a problem in combinatorics is framed as a problem in probability, where it becomes easier to solve.

HIDEO HIROSE (MR Author ID: 620160) received his Ph.D. in Engineering from Nagoya University. He held visiting positions at Stanford University before joining the faculty at Kyushu Institute of Technology, where he was eventually appointed vice dean of the university. Then, he moved to Hiroshima Institute of Technology, and was appointed director of Data Science Research Center.

Proof Without Words: Independent Sets in Grid graphs and Tilings of Aztec Diamonds

STEVE BUTLER
Iowa State University
Ames, IA 50011
butler@iastate.edu

The *Aztec diamond* of order n is the set of lattice squares $[a, a+1] \times [b, b+1]$ $(a, b \in \mathbb{Z})$ that lie completely inside the tilted square $|x| + |y| \leq n + 1$. An independent set in a graph is a collection of vertices so that no pair is adjacent.

Theorem 1. *The number of ways to tile the Aztec diamond of order n with rectangles one of whose dimensions is one equals the number of independent sets in the graph* $P_{2n} \square P_{2n}$ *(the Cartesian product of the path graph on* $2n$ *vertices with itself).*

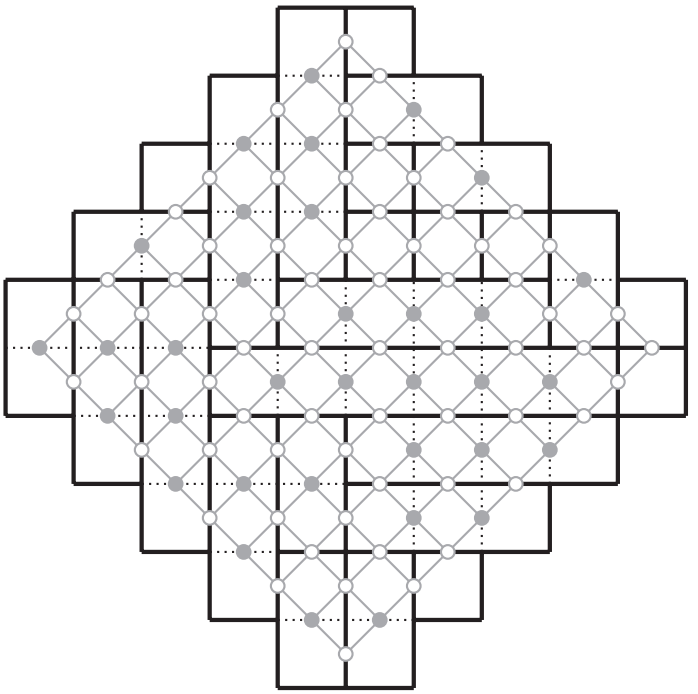


Figure 1 The tilted graph $P_{10} \square P_{10}$ is laid on top of the Aztec diamond of order 5; the gray dots in the graph form an independent set.

More information about these numbers is in sequence A006506 in the OEIS [3]. A classic result for tiling of Aztec diamonds and dominoes is given by Elkies et al. [1,2].

Math. Mag. **92** (2019) 126–127. doi:10.1080/0025570X.2018.1541688
MSC: Primary 05A15, Secondary 05C30; 05C69

REFERENCES

[1] Elkies, N., Kuperberg, G., Larsen, M., Propp, J. (1992). Alternating sign matrices and domino tilings I. *J. Algebraic Combin.* 1: 111–132.
[2] Elkies, N., Kuperberg, G., Larsen, M., Propp, J. (1992). Alternating sign matrices and domino tilings II. *J. Algebraic Combin.* 2: 219–234.
[3] OEIS Foundation, Inc. (2019). *The On-Line Encyclopedia of Integer Sequences*. oeis.org.

Summary. A visual proof that the number of independent sets in the grid graph $P_{2n} \square P_{2n}$ equals the number of tilings of the Aztec diamond A_n with rectangles where at least one of the dimensions is one.

STEVE BUTLER (MR Author ID: 801542) is the Barbara J. Janson Professor of Mathematics at Iowa State University. His interests in mathematics includes mathematics of juggling and card shuffling (and more generally fun mathematical toys).

Math Bite: A Simple Proof of the RMS–AM Inequality

The following quadratic equation has no real roots:

$$(x+a)^2 + (x+b)^2 = 0, a, b \in \mathbb{R}, a \neq b. \tag{1}$$

If we expand (1) and use the fact that it's discriminant is negative, we get $2ab < a^2 + b^2$, which is a geometric mean–root-mean square (RMS) inequality for two numbers. We modify this simple observation to prove the RMS–arithmetic mean (AM) inequality of the following theorem.

Theorem. *For real numbers* $a_1, \ldots, a_n$, $\sqrt{\dfrac{a_1^2 + \cdots + a_n^2}{n}} \geq \dfrac{a_1 + \cdots + a_n}{n}$.

Proof. The quadratic equation

$$(x+a_1)^2 + \cdots + (x+a_n)^2 = nx^2 + 2(a_1 + \cdots + a_n)x + a_1^2 + \cdots + a_n^2 = 0$$

has a real solution if and only if $x = -a_1 = \cdots = -a_n$. Hence, the quadratic equation has at most one solution and therefore it has a discriminant $D \leq 0$. Because $D \leq 0$, we get $4(a_1 + \cdots + a_n)^2 - 4n(a_1^2 + \cdots + a_n^2) \leq 0$ which is equivalent to $(a_1 + \cdots + a_n)^2 \leq n(a_1^2 + \cdots + a_n^2)$, proving the result. ■

—Contributed by Konstantinos Gaitanas (MR Author ID: 1103989)
National Technical University of Athens, 15780 Athens, Greece
raffako@hotmail.com

Divisibility Tests Unified: Stacking the Trimmings for Sums

EDWIN O'SHEA
James Madison University
Harrisonburg, VA 22807
osheaem@jmu.edu

The most well-known divisibility tests are the last digits tests for 2 and 5, the sum of digits test for 9, and the alternating digit sum for 11, but the oldest divisibility test is one for deciding divisibility by 7. That test is at least fifteen hundred years old and is prescribed in the Talmud [**8**, Abodah Zarah 9b] as follows: "If one does not know what the year is in the Sabbatical cycle of seven years, let him... put aside the hundreds... and convert the remainder into Sabbatical Cycles [of seven years each] after adding thereto two years for every complete century; what is left over will give him the number of the given year in the current Sabbatical Cycle." In algebraic notation, the remainder when 7 is divided into a given integer, written as $x + 100 \cdot y$, equals that when 7 is divided into $x + 2 \cdot y$. For example, to remainder when 7 divides $32184 = 84 + 100 \cdot 321$ equals that when 7 divides $84 + 2 \cdot 321 = 726$.

The Talmud's test is the first of seventy or so listed in Dickson's encyclopedic *History of the Theory of Numbers* [**3**, Chapter XII] and includes tests by luminaries such as Fibonacci, Lagrange, Pascal, and Sylvester. Tests that reinterpret those recorded by Dickson can be found in a number of relatively recent papers [**2**], [**4**], [**6**], [**9**] and the sources referenced therein. Among the tests is one for 7 by Zbikowski [**10**] asserting that an integer a, written in the form $a = 10\bar{a} + a_0$, is divisible by 7 if and only if 7 divides the integer $\bar{a} - 2a_0$. For example, the test reduces 32184 to $3218 - 2 \cdot 4 = 3210$, which can be applied again, reducing 3210 to 321, and again, reducing 321 to 30; since 7 does not divide 30 it does not divide 32184.

This *trimming* procedure, the given integer a being "trimmed" to another with one digit less, is universally presented as being cut from a different cloth from the sum of digits tests for 9 and 11. We claim that this is not so by deriving a family of summing tests, due to Khare [**5**], from Zbikowski's family of trimming tests. We can also show that the best known summing tests, the binomial tests, can also be derived from an adapted form of Zbikowski's tests. To the best of our knowledge this marriage of trimming and summing tests is new. *

In homage to the school venue where many of us were first exposed to divisibility tests, we will only require basic properties of the integers with a dash of the induction axiom; we will use neither the binomial theorem nor modular arithmetic. Our central tool is *stacking*, a decimal representation that is flexible enough to respect a six-year old's choosing of 10 pennies over one dime. The well-known sum and alternating sum of digits tests for 9 and 11 follow as corollaries. We close with a brief comparative

Math. Mag. **92** (2019) 128–135. doi:10.1080/0025570X.2019.1562293
MSC: Primary: 11A07, Secondary: 11A51

*This paper was first submitted in June 2014 with the first referee report returned in June 2018. In the intervening period, Eric L. McDowell independently arrived at many of the same results in "Divisibility Tests: A History and User's Guide" which appeared in May 2018 in *MAA Convergence*. What we call "stacking" is what that paper refers to as "flowing." McDowell's presentation deserves a wide readership and contains many references not addressed in this present paper. Editor's note: This manuscript was lost in the Editorial Manager system during the first year of its use. As a consequence, it was given a 2016 manuscript number. The Editor takes responsibility and apologizes to the author and to readers of THIS MAGAZINE for the delay.

analysis of Khare's tests, the binomial summing tests and Zbikwoski's trimming tests, and how these tests in base 10 generalize to any base.

Defining Divisibility Tests

Rather than operate under a Justice Potter-like assumption [7], that we all know a divisibility test when we see it, let us propose a decent definition. In most basic terms, a *divisibility test for an integer* q should be a function $f_q : \mathbb{Z} \to \mathbb{Z}$ such that q divides a if and only if q divides $f_q(a)$ for every integer a. The identity function $f_q(a) = a$ is easy to compute but q dividing $f_q(a)$ is no easier to decide than if q divides a. The computation of the remainder in the classical division theorem, $f_q(a) = r$, might have the property that q divides r is easier to decide than q divides a but the computation of $f_q(a)$ is likely to be mentally difficult. We'd like to propose that a divisibility test $f_q(a)$ should be *easy* to compute and it ought to be *easier* to decide if q divides $f_q(a)$ than if q divides a. The terms "easy" and "easier" are ambiguous but one criterion for "easy" is that $f_q(a)$ is computable with relative ease. "Easier" could also mean a number of things but a desirable property might be that the number of digits in $f_q(a)$ is less than that in a. Note that any test f_q is iterative, with $f_q^2(a) = f_q(f_q(a))$ being a test too for q dividing a, and $f_q^3(a) = f_q(f_q(f_q(a)))$ too, etc.

We promised to only use basic divisibility properties to derive our tests; no modular arithmetic or binomial theorem. To that effect, the following claims appear in number theory texts like Andrews [1].

(1) If two integers a and s are both divisible by q then their sum and difference, $a \pm s$ are also divisible by q.

(2) If q is relatively prime to 10 and q divides $10 \cdot m$ then q divides m.

We can write an integer a as $a = a_n a_{n-1} \ldots a_2 a_1 a_0 = \sum_{k=0}^{n} 10^k \cdot a_k$, where each $0 \le a_k \le 9$. For shorthand, we denote the number of digits of a, $n+1$, as length(a). Letting $a_{[k,l]} := a_k a_{k-1} \ldots a_{l+1} a_l$, we can always write $a = 10^k \cdot a_{[n,k]} + a_{[k-1,0]}$. As a special case, let $\bar{a}$ denote $a_{[n,1]}$ and write

$$a = 10 \cdot \bar{a} + a_0 \quad \text{and similarly,} \quad q = 10 \cdot \bar{q} + q_0.$$

For example, if $a = 32184$ then $a_4 = 3$, $a_3 = 2$, $a_2 = 1$, $a_1 = 8$ and $a_0 = 4$. The length of a is 5. We can write 32184 in a variety of ways including $10^2 \cdot a_{[4,2]} + a_{[1,0]} = 10^2 \cdot 321 + 84$ and $10 \cdot \bar{a} + a_0 = 10 \cdot 3218 + 4$.

It is left as an exercise to apply claim **(1)** to derive the last digit tests $f_2(a) = f_5(a) = a_0$. More generally, $f_{2^k}(a) = f_{5^k}(a) = a_{[k-1,0]}$ are divisibility tests for $q = 2^k$ and $q = 5^k$. For example, 8 divides $a = 32184$ because 8 divides a's last three digits, $f_{2^3}(32184) = 184$. With the above notation, the Talmud test is $\text{Tal}_7(a) = 2 \cdot a_{[n,2]} + a_1 a_0$. It too can be proved using claim **(1)**: letting $s = 98 \cdot a_{[n,2]}$, 7 divides $a = 100 \cdot a_{[n,2]} + a_1 a_0$ if and only if it divides $a - s = 2 \cdot a_{[n,2]} + a_1 a_0$.

Zbikowski's Trimming Tests as One Test

Zbikowski's test for 7 is $T_7(a) = \bar{a} - 2 \cdot a_0$. On an example like $T_7(32184) = 3210$ we see that T_7 takes a given a and "trims" it to another integer of length one less than the original a. This motivates the following definition:

(Trimming) A divisibility test f_q is called a *trimming* test if the length of $f_q(a)$ is one less than the length of a, for almost every a.

We say "almost" because if a is already a single digit there is nothing to be done and there are instances, like $T_7(49) = -14$, where the test maps a two-digit number to

another two-digit number. We leave it as an exercise to show that if length$(a) \geq 3$ then $T_7(a)$ has shorter length than a.

Here's why T_7 works on our running example of $a = 32184$. By claim **(1)**, we can subtract any multiple of 7 from 2184 and the result will be divisible by 7 if 32184 itself is divisible by 7, so choose a multiple of 7 that when subtracted from 32184 leaves a zero in the last digit. Clearly, 21 times the last digit of 32184, namely $21 \cdot 4$ will serve this role. The difference is $32184 - 21 \cdot 4 = (10 \cdot 3218 + 4) - ((20 + 1) \cdot 4)$. The 4's cancel leaving a multiple of 10. By claim **(2)**, we can trim that right-most zero from $32184 - 21 \cdot 4 = 32100$ to get 3210 and our decision of whether 7 divides 32184 becomes equivalent to deciding if 7 divides 3210.

Zbikowski [10] extended this argument for every a and for any q with last digit equal to 1, 3, 7, or 9. These tests have received considerable attention in recent papers by Zazkis [9], Cherniavsky and Mouftakhov [2], and Ganzell [4] and the reader can see a derivation of these tests there. We will not derive these tests here but wish to recast these tests as one test. First, we consider Zbikowski's tests as four different cases, followed by examples.

Theorem 1. (Zbikowski [10]) *For every q with last digit equal to either* 1, 3, 7, *or* 9, *there is a trimming test $T_q(a)$ given in the following table.*

q_0	1	3	7	9
c_q	1	-3	3	-1
$T_q(a) = t_q(a, c_q)$	$\bar{a} - \bar{q}a_0$	$\bar{a} + (3\bar{q} + 1)a_0$	$\bar{a} - (3\bar{q} + 2)a_0$	$\bar{a} + (\bar{q} + 1)a_0$

The following examples demonstrate Zbikowski's tests.

- If $q = 21$ then $\bar{q} = 2$ and $T_{21}(a) = \bar{a} - 2 \cdot a_0 = \bar{a} - 2a_0$.
  For $a = 32184$, $T_{21}(32184) = 3218 - 2 \cdot 4 = 3210$ and $T_{21}(3210) = 321$.
- If $q = 13$ then $\bar{q} = 1$ and $T_{13}(a) = \bar{a} + (3 \cdot 1 + 1)a_0 = \bar{a} + 4a_0$.
  For $a = 32184$, $T_{13}(32184) = 3218 + 4 \cdot 4 = 3234$ and $T_{13}(3234) = 339$.
- If $q = 17$ then $\bar{q} = 1$ and $T_{17}(a) = \bar{a} - (3 \cdot 1 + 2)a_0 = \bar{a} - 5a_0$.
  For $a = 32184$, $T_{17}(2184) = 3218 - 5 \cdot 4 = 3198$ and $T_{17}(3198) = 279$.
- If $q = 39$ then $\bar{q} = 3$ and $T_{39}(a) = \bar{a} + (3 + 1)a_0 = \bar{a} + 4a_0$.
  For $a = 32184$, $T_{39}(32184) = 3218 + 4 \cdot 4 = 3234$ and $T_{39}(3234) = 339$.

As expected the above examples trim the length by one per iteration. The examples are for q's with two digits but q can be of any length, like $T_{181} = \bar{a} - 18 \cdot a_0$.

Absent from previous expositions on Zbikowski's tests is that the four tests reduce to one. First, it appears that $T_{13} = T_{39}$ and $T_7 = T_{21}$. Using the table above, one can show the following.

(3) If $q_0 = 3$ or 7 then $T_q(a) = T_{3q}(a)$.

This reduces our four tests to only two, those T_q's for which $q_0 = 1$ or $q_0 = 9$. With $[x]$ denoting the nearest integer to x we leave it to the reader, using the table above, to confirm the following.

$$\textbf{(4)} \quad T_q(a) = \bar{a} + \omega_q \cdot a_0 \quad \text{where} \quad \omega_q = \left\{ \begin{array}{rll} -[q/10] & \text{if} & q_0 = 1 \\ [q/10] & \text{if} & q_0 = 9 \end{array} \right\}.$$

In summary, Zbikowksi's test reads easily as one test: *If an odd divisor q ends in* 1 *or* 9 *then divide q by* 10 *and round the result to the nearest integer; attach a sign of minus or plus to the result depending on whether you have rounded down or up*

for the signed weight ω_q. If q ends in 3 or 7 then triple q and do as before; that is, $\omega_q = \omega_{3q}$. Zbikowski's test for q dividing a is then everything but the last digit of a plus the signed weight ω_q times the last digit of a.

For example, to write a divisibility test for $q = 17$ we triple 17 to get 51. For the signed weight ω_{17}, divide 51 by 10 and round to the nearest integer to produce 5; since we rounded down the signed weight must be negative and so -5 is the weight for the test for $q = 17$. That is, $T_{17}(a) = \bar{a} - 5a_0$. Likewise, $T_{79} = \bar{a} + 8a_0$ since 79/10 rounds to 8 and the weight is positive since we rounded up (not down) to 8.

Using Zbikowski's trimming test $T_q(a) = \bar{a} + \omega_q \cdot a_0$ we shall derive Khare's general weighted sum of digits tests [5]. Khare's summing tests S_q match the usual tests for 9 and 11 but differ from the better known binomial tests for all other q. Nonetheless, we can also derive the usual binomial tests by adapting Zbikowski's tests to trim from the left rather than the right. This is all achieved by a form of child's play we call stacking.

Stacking: Preferring Pennies to Dimes

The trimming tests $T_9(a) = \bar{a} + a_0$ and $T_{11}(a) = \bar{a} - a_0$ are not the same yet look similar to the sum and alternating sum of digits tests, respectively. These sum of digits tests are usually verified by modular arithmetic—geometric series suffice too—but the trimming tests have only used the basic divisibility properties **(1)** and **(2)**. From the trimming tests T_q we will derive the usual tests for 9 and 11 and Khare's *summing* tests for every q. We should first define what we mean by a summing test.

(Summing) A divisibility test $f_q(a) = \sum_{j=0}^{n} \gamma_j a_{n-j}$ is called a *summing* test for q if each $\gamma_j \in \mathbb{Z}$.

Let's investigate the trimming test $T_9(a) = \bar{a} + a_0$ with our running example $a = 32184$ and see if we can get some ideas on how to derive the sum of digits test $3 + 2 + 1 + 8 + 4$. The trimming test applied iteratively is

$$32184 \xrightarrow{T_9} 3218 + 4 = 3222 \xrightarrow{T_9} 322 + 2 = 324 \xrightarrow{T_9} 32 + 4 = 36 \xrightarrow{T_9} 3 + 6.$$

The summing and recursive trimming tests yield a different final output. We claim that they are equal provided that a "stacking" procedure intervenes. To explain the main idea, let's start with a non-trivial theorem, that of *every positive integer has a unique base* 10 *representation*. This is mathematically respected but colloquially malleable. When writing checks we are allowed to express 1562 in unambiguous but different ways, both as "one thousand, five hundred and sixty-two" and as "fifteen hundred and sixty-two." The former is in keeping with strict mathematical practice yet the latter is customary even though 15, the coefficient (allowing ourselves to call it that) of one hundred in the latter is not between 0 and 9.

In the same vein, when adults add two integers, like $3218 + 4 = 3222$ that result from $T_9(32184)$, we simplify in concordance with unique representability. Computing the sum $3218 + 4$ is equivalent to giving an adult 3218 cents as 321 dimes and 8 pennies and giving them a further 4 pennies, with which the adult opts to exchange $8 + 4 = 12$ pennies for 1 dime and 2 pennies for a total of 322 dimes and 2 pennies. We are raised to value efficiency: the fewer coins, the better. However, given the same choice, a six-year old may opt to keep the 12 pennies. She knows that 10 pennies and 1 dime both equal 10 cents but 10 pennies are far more fun to play with and easier to share than a dime and so she chooses to stack the pennies together. In other words, she might opt for 321 dimes and $8 + 4 = 12$ pennies, that is, $3218 + 4 = 10 \cdot 321 + 8 + 4 = 10 \cdot 321 + (8 + 4)$. Depending on her mathematical formalism, she may define stacking the pennies as follows.

(Stacking) Given an integer $r = 10\bar{r} + r_0$ and a (possibly empty) sum of single-digit integers s write the *stacking* of their sum $r + s \overset{\text{Stack}}{=\!=} 10\bar{r} + (r_0 + s)$.

For short, we write the stacking of r and s as $\text{Stack}(r + s)$. For example, stacking 3218 and 4 together equals the representation $\text{Stack}(3218 + 4) = 10 \cdot 321 + (8 + 4)$. Since stacking is nothing more than an alternative representation of $r + s$, it follows that q divides $r + s$ if and only if q divides $\text{Stack}(r + s)$.

Stacking Zbikowski Trimmings for Khare's Summing Tests

With stacking in mind, let's iteratively trim as before with T_9 but now follow each trimming with a stacking.

$$\begin{aligned}
32184 &\xrightarrow{T_9} 3218 + 4 \overset{\text{Stack}}{=\!=} 10 \cdot 321 + (8 + 4) \\
&\xrightarrow{T_9} 321 + (8 + 4) \overset{\text{Stack}}{=\!=} 10 \cdot 32 + (1 + 8 + 4) \\
&\xrightarrow{T_9} 32 + (1 + 8 + 4) \overset{\text{Stack}}{=\!=} 10 \cdot 3 + (2 + 1 + 8 + 4) \\
&\xrightarrow{T_9} 3 + (2 + 1 + 8 + 4) \overset{\text{Stack}}{=\!=} (3 + 2 + 1 + 8 + 4).
\end{aligned}$$

The above says that

$$(\text{Stack} \circ T_9)^4(32184) = (3 + 2 + 1 + 8 + 4) = 18 =: S_9(32184),$$

where the latter denotes the usual sum of the digits test for 9. Let us see if iteratively trimming and stacking with $T_7(a) = \bar{a} + (-2)a_0$ can provide a sum-like test for $q = 7$ using our running example $a = 32184$.

$$\begin{aligned}
32184 &\xrightarrow{T_7} 3218 + (-2) \cdot 4 \\
&\overset{\text{Stack}}{=\!=} 10 \cdot 321 + (8 + (-2) \cdot 4) \\
&\xrightarrow{T_7} 321 + (-2) \cdot (8 + (-2) \cdot 4) \\
&\overset{\text{Stack}}{=\!=} 10 \cdot 32 + (1 + (-2) \cdot (8 + (-2) \cdot 4)) \\
&\xrightarrow{T_7} 32 + (-2) \cdot (1 + (-2) \cdot (8 + (-2) \cdot 4)) \\
&\overset{\text{Stack}}{=\!=} 10 \cdot 3 + 2 + (-2) \cdot (1 + (-2) \cdot (8 + (-2) \cdot 4)) \\
&\xrightarrow{T_7} 3 + (-2)(2 + (-2) \cdot (1 + (-2) \cdot (8 + (-2) \cdot 4))) \\
&\overset{\text{Stack}}{=\!=} 3 + (-2)(2 + (-2) \cdot (1 + (-2) \cdot (8 + (-2) \cdot 4))).
\end{aligned}$$

In other words, $(\text{Stack} \circ T_7)^4(32184) = 3 + (-2)^1 2 + (-2)^2 \cdot 1 + (-2)^3 \cdot 8 + (-2)^4 \cdot 4$. The above examples for $q = 7$ and $q = 9$ with $a = 32184$ suggest summing tests with $\gamma_j = (-2)^j = \omega_7^j$ and $\gamma_j = 1 = \omega_9^j$ for 7 and 9, respectively. We claim this holds in general.

Theorem 2. *If $T_q = \bar{a} + \omega_q a_0$ is a trimming test for q then $S_q(a) := \sum_{j=0}^{n} \omega_q^j a_{n-j}$ is a summing test for q.*

The tests S_q were presented in 1997 by Khare [5] but their modular arithmetic proof does not involve trimming tests. Briefly, Khare's construction begins by choosing γ_q as the minimum residue representative of the inverse of 10 modulo q. That is, $\gamma_q \equiv 10^{-1} \mod q$ of smallest size. Khare then proposes $S_q = \sum_{j=0}^{n} \gamma_q^j a_{n-j}$ is a test by virtue of

$$\gamma_q^n a = \sum_{j=0}^{n} \gamma_q^n 10^j a_j \equiv S_q(a) \mod q.$$

It is straightforward to check that Khare's γ_q equals Zbikowski's ω_q. Our derivation of Khare's tests from Zbikowski's tests uses neither modular arithmetic nor the binomial theorem and it unifies the trimming and summing families. Before proving the result, let's appreciate Khare's tests for some examples on $a = 32184$:

- $S_7(32184) = 3 + (-2) \cdot 2 + (-2)^2 \cdot 1 + (-2)^3 \cdot 8 + (-2)^4 \cdot 4 = 3$
- $S_9(32184) = 3 + 1 \cdot 2 + 1^2 \cdot 1 + 1^3 \cdot 8 + 1^4 \cdot 4 = 18.$
- $S_{11}(32184) = 3 + (-1) \cdot 2 + (-1)^2 \cdot 1 + (-1)^3 \cdot 8 + (-1)^4 \cdot 4 = -2.$
- $S_{17}(32184) = 3 + (-5) \cdot 2 + (-5)^2 \cdot 1 + (-5)^3 \cdot 8 + (-5)^4 \cdot 4 = 1518.$
- $S_{39}(32184) = 3 + 4 \cdot 2 + 4^2 \cdot 1 + 4^3 \cdot 8 + 4^4 \cdot 4 = 1563.$

Proof of Theorem 2 by Trimming and Stacking. We will show, by induction on the length of a, that $S_q(a) = (\text{Stack} \circ T_q)^n(a)$ whenever a has length $n + 1$.

If $n = 1$ then $a = a_1a_0$ has length two and $\text{Stack}(T_q(a_1a_0)) = \text{Stack}(\bar{a} + \omega_q a_0) = a_1 + \omega_q a_0$ as claimed. Assume that $S_q(a') = (\text{Stack} \circ T_q)^{n-1}(a')$ for every a' with length n and consider any integer $a = a_n a_{n-1} \ldots a_2 a_1 a_0$ with length $n + 1$. Applying $\text{Stack} \circ T_q$ to this a results in $\text{Stack}(T_q(a)) = \text{Stack}(\bar{a} + \omega_q a_0) = 10a_{[n,2]} + (a_1 + \omega_q a_0)$, an integer with n digits with last digit equal to $(a_1 + \omega_q a_0)$ to which the induction hypothesis applies; hence,

$$\begin{aligned}
(\text{Stack} \circ T_q)^n(a) &= (\text{Stack} \circ T_q)^{n-1}(\text{Stack} \circ T_q(a)) \\
&= (\text{Stack} \circ T_q)^{n-1}(10a_{[n,2]} + (a_1 + \omega_q a_0)) \\
&= S_q(10a_{[n,2]} + (a_1 + \omega_q a_0)) \\
&= \sum_{j=0}^{n-2} \omega_q^j a_{n-j} + \omega_q^{n-1}(a_1 + \omega_q a_0) \\
&= \sum_{j=0}^{n-2} \omega_q^j a_{n-j} + \omega_q^{n-1} a_1 + \omega_q^n a_0 = \sum_{j=0}^{n} \omega_q^j a_{n-j} = S_q(a). \qquad \square
\end{aligned}$$

(Left) Stacking the (Left) Trimmings for Binomial Summing Tests

It would be remiss not to mention the most well-known summing tests, those that follow from the binomial identity. We wish to derive the binomial tests from an adapted form of Zbikowski's tests that trim from the left instead of the right, further solidifying the unification of trimming and summing tests.

The binomial tests are developed by applying the binomial theorem to the standard expression for a modulo q,

$$a = \sum_{j=0}^{n} (q + (10 - q))^j a_j \equiv \sum_{j=0}^{n} (10 - q)^j a_j =: B_q(a) \mod q.$$

The well-known tests for 9 and 11 are B_9 and B_{11} and are usually motivated in this fashion. The binomial test for 7 is $B_7(a) = \sum_{j=0}^{n} 3^j a_j$ and for, say, 39 it is $B_{39}(a) = \sum_{j=0}^{n} (-29)^j a_j$. We claim that these tests can be developed via a recursive trimming and stacking procedure akin to the derivation of Khare's tests from Zbikowski's.

On our main example, testing if 7 divides 31284, notice that we can rewrite 32184 as $10^3((7 + 3) \cdot 3 + 2) + 184$. The term in brackets is regarded as a non-traditional coefficient of 10^3 just as we did in stacking (on the right) earlier. For testing divisibility by

7 we can cast off the 7 in the bracketed term before distributing, so 7 divides 32184 if and only if 7 divides $10^3((3)\cdot 3+2)+184=(11)184$. As before, this last number might be how we would write a check, writing the integer longhand as "eleven thousand, one hundred, and eighty-four."

Repeating again, $(11)184=10^2(11(7+3)+1)+84$ reduces to $10^2(11(3)+1)+84=(34)84$, or "thirty four hundred and eighty-four". Repeating once more, $(34)84$ reduces to $(3\cdot 34+8)4=(110)4$ which, repeating again, reduces to (334). In other words, 7 divides 32184 if it divides 334. We can repeat this process again on 334 itself, should we wish, and it would equal $3^2\cdot 3+3^1\cdot 3+4=40$. We can conclude that 7 does not divide 32184.

The example motivates an adapted version of Zbikowski's tests T_q and the Stack function, which we will call *left trim*, LT_q and *left stack*, LStack. It is immediate that

$$LT_q(a):=10^{n-1}(10-q)a_n+a_{[n-1,0]}$$

is a test for q and that

$$\text{LStack}(10^{n-1}(10-q)a_n+10^{n-1}a_{n-1}+a_{[n-2,0]})=((10-q)a_n+a_{n-1})a_{[n-2,0]}$$

provides the same flexibility that the original Stack function provided. Here is a more careful presentation of our main example with this notation.

$$\begin{aligned}
32184 &\xrightarrow{LT_7} 10^3\cdot(3\cdot 3)+2184\\
&\xlongequal{\text{LStack}} (3\cdot 3+2)184\\
&\xrightarrow{LT_7} 10^2\cdot 3\cdot(3\cdot 3+2)+184\\
&\xlongequal{\text{LStack}} (3^2\cdot 3+3\cdot 2+1)84\\
&\xrightarrow{LT_7} 10^1\cdot 3\cdot(3^2\cdot 3+3\cdot 2+1)+84\\
&\xlongequal{\text{LStack}} (3^3\cdot 3+3^2\cdot 2+3\cdot 1+8)+4\\
&\xrightarrow{LT_7} 10^0\cdot 3\cdot(3^3\cdot 3+3^2\cdot 2+3\cdot 1+8)+4\\
&\xlongequal{\text{LStack}} (3^4\cdot 3+3^3\cdot 2+3^2\cdot 1+3\cdot 8+4)=334.
\end{aligned}$$

Theorem 3. *The binomial test* $B_q(a)$ *equals* $(\text{LStack}\circ LT_q)^n(a)$.

The proof is very similar to that of Theorem 2, inducting on the length of a and trimming and stacking on the left as we did previously on the right. We leave the details as an exercise.

Closing Remarks

Starting with the test for 7 and using only elementary tools, we reduced Zbikowski's tests to a single trimming test for all integers. From Zbikowski's tests we derived Khare's summing tests as well as the binomial tests, adding only a dash of the induction axiom to our basic divisibility criteria. The two families of divisibility tests, trimming and summing, are much closer than initially meets the eye.

Khare's tests are vastly preferable to the binomial tests and, in practice, the trimming tests are superior to both summing tests. The weights in Khare's tests *scale down* the original divisor q by a factor of 10 or 10/3 whereas the binomial tests have weights that are the *difference* of q with 10. For example, Khare's $S_{39}(a)=\sum_{j=0}^{n}4^j a_j$ is preferable to the binomial $B_{39}(a)=\sum_{j=0}^{n}(-29)^j a_j$. The practice of Zbikowski's trimming is better than both as it avoids the mental computation of high powers of ω_q, relying only on multiplying the last digit of an integer a by ω_q followed by a straightforward subtraction and then recursively repeating this procedure.

For Zbikowski's tests $a \equiv 10^{-1}T_q(a) \mod q$ and since stacking changes the representation of the number a but not a itself, then $a \equiv \omega_q T_q(a)$ and $a \equiv \omega_q^n S_q(a) \mod q$ whenever a has length $n+1$. In contrast, part of the appeal of the binomial tests B_q is its preservation of remainders.

Khare also generalized the base $b = 10$ to tests S_q for q in any base b. If q and b are co-prime then the ω_q term is precisely the least residue of $\omega_q \equiv b^{-1} \mod q$ and there are last-digits tests for all factors of b. Indeed, this article could be written for a general base b and the results would hold as one would expect.

Finally, while most tests are of the trimming and summing variety, there are tests that are not equivalent to those outlined here, like the Talmud test Tal_7. Dickson [**3**, Chapter XII] has many gems not discussed here and independently deriving each of them and understanding the some of the original sources would make for an excellent senior project.

Acknowledgments. I am grateful for conversations with Elizabeth Brown, Ezra "Bud" Brown, Brant Jones, Rachel Quinlan, and Jason Rosenhouse throughout my thinking and writing about divisibility tests. My student, Cameron Stopak suggested trimming from the left as an adaption to Zbikowski's trimming from the right. I was stuck for some time on how to jump from trimming to summing by elementary means and it was from playing shop with my children that I realized that "stacking pennies" was exactly what was needed so thanks to e- and f-too.

REFERENCES

[1] Andrews, G. E. (1994). *Number Theory*. Mineola, NY: Dover.
[2] Cherniavsky, Y., Mouftakhov, A. (2014). Zbikowski's divisibility criterion. *College Math. J.* 45(1): 17–21. DOI: 10.4169/college.math.j.45.1.017
[3] Dickson, L. E. (2005). *History of the Theory of Numbers, Volume I: Divisibility and Primality*. Mineola, NY: Dover Publications.
[4] Ganzell, S. (2017). Divisibility tests, old and new. *College Math. J.* 48(1): 36–40. DOI: 10.4169/college.math.j.48.1.36
[5] Khare, A. (1997). Divisibility tests. *Furman Univ. Electron. J. Undergraduate Math*. 3(1):1–5.
[6] Renault, M. (2006). Stupid divisibility tricks. *Math. Horizons*. 14(2): 19–21, 42.
[7] Jacobellis v. Ohio. 378 US 184 – Supreme Court, 1964.
[8] *The Babylonian Talmud* (1935–1948).Translated into English with notes, glossary, and indices under the editorship of I. Epstein. London: Soncino Press.
[9] Zazkis, R. (1999). Divisibility: A problem solving approach through generalizing and specializing. *Humanistic Math. Netw. J*. 21(1): 34–38.
[10] Zbikowski, A. (1861). Note sur la divisibilité des nombres. *Bull. Acad. Sci. St. Pétersbourg* 3: 151–153.

Summary. Divisibility tests are algorithms that can quickly decide if one integer is divisible by another. There are many tests but most are either of the *trimming* or *summing* variety. Our goals are to present Zbikowski's family of trimming tests as one test and to unify the trimming and summing tests. We do the latter by showing, first, that the most effective summing tests, due to Khare, can be derived directly from the Zbikowski's test and, second, that the best known summing tests—the binomial tests—can be derived from an adapted form of Zbikowski's tests. We introduce the notion of *stacking*, the claim that a six-year old would always choose 10 pennies over a dime, and use only basic divisibility properties to achieve our goals.

EDWIN O'SHEA (MR Author ID: 749078) received his PhD in mathematics from the University of Washington. Originally from the northside of Cork City in Ireland, he is currently an associate professor at James Madison University in Virginia. His training is in computational algebra and combinatorics and he has additional scholarly interests in geometry and history. Outside of mathematics he enjoys baking, billiards, chatting with other people, and walking with his dog at night.

The Instructor's Guide to Real Induction

PETE L. CLARK
University of Georgia
Athens, GA 30605
plclark@gmail.com

In this paper, we pursue inductive principles of ordered sets. To get a sense of what this means, consider the principle of mathematical induction. When applied, one thinks in terms of families of statements $P(n)$ indexed by the natural numbers $\mathbb{N}_0 = \mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$, but the cleanest enunciation is in terms of subsets. We call a subset $S \subseteq \mathbb{N}_0$ *inductive* if it satisfies both of the following properties:

(MI1) We have $0 \in S$.
(MI2) For all $n \in \mathbb{N}_0$, if $n \in S$, then also $n + 1 \in S$. The principle of mathematical induction is that $\mathbb{N}_0$ has no proper inductive subset. (In this form, induction appears as the last and most important of the Peano Axioms.) To prove that $P(n)$ holds for all $n \in \mathbb{N}_0$ "by induction" one shows that the set $S = \{n \in \mathbb{N}_0 \mid P(n) \text{ holds}\}$ is inductive, and thus $S = \mathbb{N}_0$.

In the next section, we work in a closed, bounded interval $[a, b]$ on the real line. We define an inductive subset $S \subset [a, b]$ and state and prove the *principle of real induction* (Theorem 1): there are no proper inductive subsets of $[a, b]$. Just as mathematical induction is a powerful technique for proving families of statements indexed by the natural numbers, real induction can be used to prove families of statements indexed by intervals on the real line. This has applications in elementary analysis, especially to the basic interval theorems concerning a continuous function $f : [a, b] \to \mathbb{R}$. Students in a first real analysis course often find the standard proofs of these results hard to absorb, understand and remember. Proofs by mathematical induction have a common scaffolding that gives students a place to start, and so too do proofs by real induction: if one can "find the induction hypothesis," then the proof dissects into more manageable goals.

Comparing the real induction proofs of these results to the more standard proofs, one gets the sense that real induction functions as a sort of alternative to Dedekind's completeness axiom: every subset of $\mathbb{R}$ that is nonempty and bounded above has a supremum.

Recent years have seen the rise of a program that Propp has called real analysis in reverse [**24**] (see also [**7**, **30**] and the references cited therein)—given a result of real analysis that can be enunciated in any ordered field, one asks whether the truth of that theorem in an ordered field implies Dedekind completeness—in other words, forces that ordered field to be isomorphic to $\mathbb{R}$. (More ambitiously, one could seek to characterize the class of ordered fields in which that theorem holds.) Our definition of an inductive subset of $[a, b]$ makes sense for any elements $a < b$ in an ordered field, and it turns out that the absence of proper inductive subsets of $[a, b]$ is equivalent to Dedekind completeness. In other words, real induction characterizes $\mathbb{R}$ among ordered fields.

After the section on real induction, we go further, giving a definition of an inductive subset of any ordered set and showing the *principle of ordered induction*: the nonexistence of proper inductive subsets is equivalent to Dedekind completeness. Since well-

Math. Mag. **92** (2019) 136–150. doi:10.1080/0025570X.2019.1549902
MSC: Primary 06A05, Secondary 26A03

ordered sets are Dedekind complete, real induction holds in any well-ordered set, and this recovers the *principle of transfinite induction*. Especially, $\mathbb{N}_0$ is well-ordered, and this extra special case recovers the principle of mathematical induction. An ordered space can be endowed with a canonical order topology, and ordered induction is a natural tool for exploring the interplay between certain topological properties of order topologies and completeness properties of the order. This material could be used in a general topology course.

And it has: in 2015 I taught such a course at the advanced undergraduate level in which I began with the topology of $\mathbb{R}$, including real induction, and spent some time on both order topologies and metric spaces—each of which generalizes and abstracts the real numbers, but in different ways—before moving on to topological spaces in general. This material also has a certain dialectic appeal, as it effects a synthesis of discrete and continuous induction, and in this regard could be of broad interest. The material on ordered induction is more abstract than that on real induction—necessarily so, since the unification afforded by abstraction is the major payoff. We have strived to make it accessible to the broadest possible audience. All that is assumed is the notion of a topological space; order theory and the connections between order and topology are developed from scratch.

Although it is natural to speak of these various forms of induction as axioms, our interest in them is not meta-mathematical. Rather, we seek to expose new proof techniques. The potential applicability of a good proof technique should be open-ended, not circumscribed in advance. In this regard, we have left some applications of real induction as challenges to the reader and also stated some problems for which I do not know definitive solutions. I hope thereby to entice the reader into the pleasure of independent or novel discovery, a pleasure this topic has afforded me several times over the years.

Real induction

Let $a < b$ be real numbers. We define a subset $S \subseteq [a, b]$ to be *inductive* if:
(RI1) We have $a \in S$.
(RI2) If $a \leq x < b$, then $x \in S \Rightarrow [x, y] \subseteq S$ for some $y > x$.
(RI3) If $a < x \leq b$ and $[a, x) \subseteq S$, then $x \in S$.

Theorem 1 (Principle of real induction)**.** *For a subset $S \subseteq [a, b]$, the following are equivalent:*
(i) S is inductive.
(ii) $S = [a, b]$.

Proof. (i) $\Rightarrow$ (ii). Let $S \subseteq [a, b]$ be inductive. Seeking a contradiction, suppose $S' = [a, b] \setminus S$ is nonempty, so $\inf S'$ exists and lies in $[a, b]$.

Case 1. Suppose $\inf S' = a$. By (RI1) we have $a \in S$, so by (RI2), there exists $y > a$ such that $[a, y] \subseteq S$. Thus y is a greater lower bound for S' than $a = \inf S'$, a contradiction.

Case 2. Suppose $a < \inf S' \in S$. If $\inf S' = b$, then $S = [a, b]$. Otherwise, by (RI2) there exists $y > \inf S'$ such that $[\inf S', y] \subseteq S$. Also, because $[a, \inf S') \subset S$, then $[a, y] \subset S$ and thus again y is a greater lower bound for S' than $\inf S'$, a contradiction.

Case 3. Suppose $a < \inf S' \in S'$. Then $[a, \inf S') \subseteq S$, so (RI3) gives $\inf S' \in S$, a contradiction.

The opposite direction, (ii) $\Rightarrow$ (i), is immediate. ∎

A little history Theorem 1 was first published by Hathaway [**12**]. I came up with it independently in 2010 [**6**] as a variation of Kalantari's induction over the continuum

[13, Section 3]. I was scheduled to give a seminar in my department that afternoon on a different topic, but instead I spoke on real induction. The audience shared my enthusiasm, which encouraged me to further develop and disseminate the material.

The enunciation of an inductive principle for subintervals of $\mathbb{R}$ is far from new. The earliest instance I know of is a 1923 work of Khinchin [15]. There is an earlier borderline case: a 1919 work of Chao [4] gives an inductive criterion for a subset of $[a, \infty)$ to be all of $[a, \infty)$. However, Chao's criterion involves a "discrete increment" $\Delta > 0$ and in fact can be proved by conventional mathematical induction.

In addition to the works [4, 12, 13, 15], each of the following papers introduces some form of continuous induction, in many cases without reference to past precedent: [2, 8–10, 16, 18, 19, 23, 25, 26].

Often for a mathematical principle, implicit use predates explicit formulation. The first explicit use of mathematical induction was in Pascal's 1665 *Traité du triangle arithmétique*, but most agree that Euclid's celebrated Proposition IX.20—There are infinitely many primes—of circa 300 BCE contains the crucial implicit use of an inductive principle. (Strictly speaking, Euclid assumes there are three primes and produces a fourth. Evidently some more general principle is intended.) Later we will encounter an important implicit use of real induction that predates the work of Khinchin and even of Chao.

Applications in analysis Let us see real induction in action.

Theorem 2 (Intermediate value theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function, and let $L \in \mathbb{R}$ be in between $f(a)$ and $f(b)$. Then there is $c \in [a, b]$ such that $f(c) = L$.*

Proof. Replacing f by $\pm(f - c)$, we reduce to the following special case: if $f : [a, b] \to \mathbb{R} \setminus \{0\}$ is continuous and $f(a) > 0$, then $f(b) > 0$. Let $S = \{x \in [a, b] \mid f(x) > 0\}$, so $f(b) > 0$ if and only if $b \in S$. We will use real induction to show that $S = [a, b]$. Thus $f(b) > 0$, completing the proof.

(RI1) Since $f(a) > 0$, we have $a \in S$.

(RI2) Let $x \in S$, $x < b$, so $f(x) > 0$. Since f is continuous at x, there exists $\delta > 0$ such that f is positive on $[x, x + \delta]$, and thus $[x, x + \delta] \subseteq S$.

(RI3) Let $x \in (a, b]$ be such that $[a, x) \subseteq S$, i.e., f is positive on $[a, x)$. We claim that $f(x) > 0$. Indeed, since $f(x) \neq 0$, the only other possibility is $f(x) < 0$, but if so, then by continuity there would exist $\delta > 0$ such that f is negative on $[x - \delta, x]$, i.e., f is both positive and negative at each point of $[x - \delta, x]$, a contradiction. ■

Theorem 3. *A continuous function $f : [a, b] \to \mathbb{R}$ is bounded.*

Proof. Let $S = \{x \in [a, b] \mid f : [a, x] \to \mathbb{R}$ is bounded$\}$. We will use real induction to show that $S = [a, b]$.

(RI1): Evidently $a \in S$.

(RI2): Suppose $x \in S$, so that f is bounded on $[a, x]$. But then f is continuous at x, so is bounded near x: for instance, there exists $\delta > 0$ such that for all $y \in [x - \delta, x + \delta]$, $|f(y)| \leq |f(x)| + 1$. So f is bounded on $[a, x]$ and also on $[x, x + \delta]$ and thus on $[a, x + \delta]$.

(RI3): Suppose $x \in (a, b]$ and $[a, x) \subseteq S$. Since f is continuous at x, there exists $0 < \delta < x - a$ such that f is bounded on $[x - \delta, x]$. Since $a < x - \delta < x$, f is bounded on $[a, x - \delta]$, so f is bounded on $[a, x]$. ■

When using real induction, one must beware the following pitfall. Often we have a family of statements $P(I)$ indexed by subintervals I of $[a, b]$. In the proof of Theorem 3, $P(I)$ is: f is bounded on I. In the proof of Theorem 2, $P(I)$ is: f is positive

at all points of I. It can be tempting to construe (RI3) as: for all $a < x \leq b$, assume $P([a, x))$ holds and prove $P([a, x])$. But this is not correct: we must assume $P([a, y])$ holds for all $a \leq y < x$ and prove $P([a, x])$. Sometimes the distinction is immaterial: a function positive on $[a, y]$ for all $a \leq y < x$ is positive on $[a, x)$. But sometimes it matters: a function bounded on $[a, y]$ for all $a \leq y < x$ need not be bounded on $[a, x)$.

Here is an instance in which finding the right inductive hypothesis requires insight.

Theorem 4 (Integrability theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is Darboux integrable: for all $\epsilon > 0$, there is a partition $\mathcal{P} = \{a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b\}$ of $[a, b]$ such that the difference between the associated upper sum*

$$U(f, \mathcal{P}) = \sum_{i=0}^{n-1} \sup(f, [x_i, x_{i+1}])(x_{i+1} - x_i)$$

and the associated lower sum

$$L(f, \mathcal{P}) = \sum_{i=0}^{n-1} \inf(f, [x_i, x_{i+1}])(x_{i+1} - x_i)$$

is less than ϵ.

Proof. For $\epsilon > 0$, let $S(\epsilon) = \{x \in [a, b] \mid$ there exists a partition $\mathcal{P}_x$ of $[a, x]$ where $U(f, \mathcal{P}_x) - L(f, \mathcal{P}_x) < (x - a)\epsilon\}$. We will use real induction to show that for all $\epsilon > 0$, we have $S(\epsilon) = [a, b]$. Then $b \in S(\frac{\epsilon}{b-a})$, completing the proof.

(RI1) As usual, this is clear.

(RI2) Suppose that for $x \in [a, b)$ we have $[a, x] \subseteq S(\epsilon)$, so that there is a partition $\mathcal{P}_x$ of $[a, x]$ such that $U(f, \mathcal{P}_x) - L(f, \mathcal{P}_x) < (x - a)\epsilon$. Since f is continuous at x, there is $\delta > 0$ such that $\sup(f, [x, x + \delta]) - \inf(f, [x, x + \delta]) < \epsilon$. Now let $y \in [x, x + \delta]$ and take the partition $\mathcal{P}_y = \mathcal{P}_x \cup \{y\}$ of $[a, y]$. Then

$$\begin{aligned} &U(f, \mathcal{P}_y) - L(f, \mathcal{P}_y) \\ &\quad = (U(f, \mathcal{P}_x) + (y - x)\sup(f, [a, y])) - (L(f, \mathcal{P}_x) + (y - x)\inf(f, [a, y])) \\ &\quad < (x - a)(\epsilon) + (y - x)(\epsilon) = (y - a)(\epsilon). \end{aligned}$$

(RI3) Suppose that for $x \in (a, b]$ we have $[a, x) \subseteq S(\epsilon)$. Since f is continuous at x, there is $\delta > 0$ such that $\sup(f, [x - \delta, x]) - \inf(f, [x - \delta, x]) < \epsilon$. Since $x - \delta < x$, $x - \delta \in S(\epsilon)$, there is a partition $\mathcal{P}_{x-\delta}$ of $[a, x - \delta]$ such that $U(f, \mathcal{P}_{x-\delta}) = L(f, \mathcal{P}_{x-\delta}) = (x - \delta - a)\epsilon$. Let $\mathcal{P}_x = \mathcal{P}_{x-\delta} \cup \{x\}$. Then as above we get

$$U(f, \mathcal{P}_x) - L(f, \mathcal{P}_x) < (x - \delta - a)\epsilon + \delta\epsilon = (x - a)\epsilon. \qquad \blacksquare$$

Applications in topology

Theorem 5 (Bolzano–Weierstrass)**.** *Each infinite subset $\mathcal{A}$ of $[a, b]$ has a limit point: there is $L \in [a, b]$ such that for all $\delta > 0$, the set $(L - \delta, L + \delta) \cap \mathcal{A}$ is infinite.*

Proof. Let S be the set of x in $[a, b]$ such that *if* $\mathcal{A} \cap [a, x]$ is infinite, it has a limit point. It suffices to show $S = [a, b]$, which we will do by real induction.

(RI1) is clear.

(RI2) Suppose $x \in [a, b) \cap S$. If $\mathcal{A} \cap [a, x]$ is infinite, then it has a limit point and hence so does $\mathcal{A} \cap [a, b]$: thus $S = [a, b]$. If for some $\delta > 0$, $\mathcal{A} \cap [a, x + \delta]$ is finite, then $[x, x + \delta] \subseteq S$. Otherwise $\mathcal{A} \cap [a, x]$ is finite but $\mathcal{A} \cap [a, x + \delta]$ is infinite for all $\delta > 0$, and then x is a limit point for $\mathcal{A}$ and $S = [a, b]$ as above.

(RI3) If $[a, x) \subseteq S$, then either $\mathcal{A} \cap [a, y]$ is infinite for some $y < x$, so $x \in S$; or $\mathcal{A} \cap [a, x]$ is finite, so $x \in S$; or $\mathcal{A} \cap [a, y]$ is finite for all $y < x$ and $\mathcal{A} \cap [a, x]$ is infinite, so x is a limit point of $\mathcal{A} \cap [a, x]$ and $x \in S$. ■

Recall that a subset U of $[a, b]$ is *open* if for all $x \in U$, there is $\delta > 0$ such that

$$\begin{cases} (x - \delta, x + \delta) \subset U & x \notin \{a, b\} \\ [a, x + \delta) \subset U & x = a \\ (x - \delta, b] \subset U & x = b \end{cases}.$$

A subset is closed if its complement is open.

Theorem 6. *The interval $[a, b]$ is connected: if U and V are disjoint open subsets of $[a, b]$ such that $U \cup V = [a, b]$, then $U = [a, b]$ or $V = [a, b]$.*

Proof. Suppose $[a, b] = U \cup V$, with U and V open and $U \cap V = \emptyset$. We assume $a \in U$ and prove by real induction that $U = [a, b]$: (RI1) is immediate, (RI2) holds because U is open, and (RI3) holds because U is closed. We're done! ■

Theorem 7 (Heine–Borel)**.** *The interval $[a, b]$ is compact: if $\{U_i\}_{i \in I}$ is a family of open subsets of $[a, b]$ such that $\bigcup_{i \in I} U_i = [a, b]$, then there is a finite subset $J \subset I$ such that $\bigcup_{i \in J} U_i = [a, b]$.*

Proof. For an open covering $\mathcal{U} = \{U_i\}_{i \in I}$ of $[a, b]$, let

$$S = \{x \in [a, b] \mid \mathcal{U} \cap [a, x] \text{ has a finite subcovering}\}.$$

We prove $S = [a, b]$ by real induction. (RI1) is clear. (RI2): If $U_1, \ldots, U_n$ covers $[a, x]$, then some U_i contains $[x, x + \delta]$ for some $\delta > 0$. (RI3): If $[a, x) \subseteq S$, then $x \in U_i$ for some $i \in I$; let $y < x$ be such that $(y, x] \in U_i$. There is a finite $J \subseteq I$ with $\bigcup_{i \in J} U_i \supset [a, y]$, so $\{U_i\}_{i \in J} \cup U_i$ covers $[a, x]$. We're done! ■

Some real induction proofs for the reader Here are more results amenable to real induction. The proofs are left to you.

Theorem 8 (Mean value inequality)**.** *Let $f : [a, b] \to \mathbb{R}$ be differentiable. Suppose that there exists $M \in \mathbb{R}$ such that for all $x \in [a, b]$ we have $f'(x) \geq M$. Then for all $x < y \in \mathbb{R}$, we have $f(y) - f(x) \geq M(y - x)$.*

Theorem 9 (Uniform continuity theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is uniformly continuous on $[a, b]$.*

Theorem 10 (Cantor intersection theorem)**.** *Let $\{F_n\}_{n=1}^{\infty}$ be a decreasing sequence of closed subsets of $[a, b]$. Put $F = \bigcap_n F_n$. Then* either *$F \neq \varnothing$ or there exists $n \in \mathbb{Z}^+$ such that $F_n = \varnothing$.*

Theorem 11 (Lebesgue number lemma)**.** *If $\{U_i\}_{i \in I}$ is an open covering of $[a, b]$, then there is $\delta > 0$ such that if $A \subseteq [a, b]$ has diameter at most δ, then $A \subseteq U_i$ for some $i \in I$.*

Theorem 12 (Dini's lemma)**.** *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of continuous real-valued functions on the interval $[a, b]$ that is pointwise decreasing: for all $x \in [a, b]$ and all $n \in \mathbb{Z}^+$, $f_{n+1}(x) \leq f_n(x)$. If $f : [a, b] \to \mathbb{R}$ is continuous and $f_n \to f$ pointwise, then $f_n \to f$ uniformly.*

Theorem 13 (Arzelà–Ascoli)**.** *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of continuous functions on $[a, b]$ such that:*
(i) There is $M \in \mathbb{R}$ such that for all $n \in \mathbb{Z}^+$ and all $x \in [a, b]$, $|f_n(x)| \leq M$, and
(ii) For all $x \in [a, b]$ and all $\epsilon > 0$, there exists $\delta > 0$ such that if $|x - y| < \delta$, then for all $n \in \mathbb{Z}^+$, $|f_n(x) - f_n(y)| < \epsilon$.
Then there is a subsequence $\{f_{n_k}\}$ that is uniformly convergent on $[a, b]$.

Theorem 8 is a consequence of the mean value theorem. It is one of several results that have been advocated (by some; we will not weigh in on this issue) as being pedagogically preferable to the mean value theorem. The other variants can also be proved by real induction. But what about the mean value theorem itself?

Problem 1. The standard proof of the mean value theorem is a deduction from the extreme value theorem. Either prove the mean value theorem directly by real induction or explain why it is not possible to do so.

Problem 2. Find other theorems that can be proved via real induction.

Comments and complements Our proof of Theorem 2 is not so different from the usual proof using suprema. That proof is probably even cleaner: it suffices to assume that $f(a) > 0$ and $f(b) < 0$ and show that there is $c \in (a, b)$ with $f(c) = 0$. For this, let $c = \sup \{x \in [a, b] \mid f(x) \leq 0\}$. Then—as follows from the definition of continuity—we must have $f(c) = 0$.

But this proof has within it a germ of the idea for real induction. In fact, one can motivate real induction in a classroom setting by asking for a proof of Theorem 3 along the lines of the above proof of Theorem 2: i.e., we start by defining $c = \sup \{x \in [a, b] \mid f \text{ is bounded on } [a, x]\}$. Then it emerges naturally that we want to show that $c = b$ and that we can establish this by showing (RI2) and (RI3). (Here, as in every application I know of, (RI1) is obvious.) It is then an interesting exercise to see how to modify the standard proof of Theorem 2 to get the proof by real induction.

We have used real induction to prove the interval theorems of elementary real analysis (cf. [**28**, Chapter 7]) with one exception: we are missing the extreme value theorem, which asserts that every continuous $f : [a, b] \to \mathbb{R}$ assumes its maximum and minimum values. This result may be deduced from Theorem 3 by an easy argument using suprema: by Theorem 3, $M = \sup(f, [a, b])$ is finite. If M were not attained on the interval $[a, b]$, then the function $g : x \mapsto \frac{1}{M - f(x)}$ would be continuous and unbounded on $[a, b]$, contradicting Theorem 3. To reiterate: we do not advocate using real induction *in place of* Dedekind's completeness axiom but rather—when helpful!—as a proof technique.

Theorem 4 is usually proved using the uniform continuity of continuous functions $f : [a, b] \to \mathbb{R}$. In [**28**, pp. 292–293], Spivak gives a different proof, establishing equality of the upper and lower integrals by differentiation. This method goes back at least to M.J. Norris [**22**]. Our proof seems different from both of these.

Standard proofs of Theorem 5 use monotone subsequences, dissection/nested intervals or the compactness of $[a, b]$. Our proof appears to be new.

Perhaps the best argument for real induction in the classroom is the proofs it affords for Theorems 6 and 7: not only are they short and simple, but initiates in real induction will find them easily.

On the one hand this suggests that the concepts of connectedness and compactness may be inherently inductive in some sense. There seems to be something to this: see, e.g., induction on connectedness and induction on compactness in [**31**]. We will give a different kind of generalization in the next section when we explore connectedness and compactness in order topologies.

On the other hand, it raises the question of why this proof technique—which, recall, has appeared in many variations in more than a dozen prior works—is not more popular. The situation becomes even more curious once one learns that our proof of Theorem 7 is essentially the same as one given in 1904 by Henri Lebesgue. In [**17**], Lebesgue proves the result as follows: he says that $x \in [a, b]$ is "reached" if there is a finite subcovering of the interval $[a, x]$, and proceeds by considering the supremum of the set of all points x that are reached. This is the last of the early proofs of Theorem 7 surveyed in [**1**]; they also discuss proofs by Borel, Cousin, Schoenflies, and Young. The authors are quite enthusiastic about Lebesgue's proof, writing "This is the one! The proof is thoroughly modern and simple to follow. In comparison, all previous arguments are cumbersome and overly complicated."

Of course Theorem 2 and the extreme value theorem are quick consequences of Theorems 6 and 7, via the following basic result.

Proposition 1. *Let $f : X \to Y$ be a continuous surjection of topological spaces.*
*(a) [**20**, Theorem 23.5] If X is connected, then so is Y.*
*(b) [**20**, Theorem 26.5] If X is compact, then so is Y.*

Ordered induction

In this section we pursue induction in ordered sets, obtaining a common generalization of mathematical induction and real induction.

Ordered sets An *ordered set* (sometimes called a linearly ordered or totally ordered set) is a set X endowed with a binary relation $\leq$ that satisfies:

- reflexivity: for all $x \in X$, $x \leq x$;
- anti-symmetry: for all $x, y \in X$, if $x \leq y$ and $y \leq x$, then $x = y$;
- transitivity: for all $x, y, z \in X$, if $x \leq y$ and $y \leq z$, then $x \leq z$; and
- totality: for all $x, y \in X$, at least one of $x \leq y$ and $y \leq x$ holds.

Our distinguished example is the interval $[a, b] \subseteq \mathbb{R}$.

A *top element* (resp. a *bottom element*) of an ordered set $(X, \leq)$ is an element $\top$ (resp. $\bot$) such that $x \leq \top$ (resp. $\bot \leq x$) for all $x \in X$. Clearly X can have at most one top (resp. bottom) element. If X lacks a top element, then we can simply adjoin such an element, denoted $\top$—that is, $\top$ is not an element of X and decreed to satisfy $x < \top$ for all $x \in X$. Similarly, if X lacks a bottom element we can adjoin one, denoted $\bot$. We denote by $\tilde{X}$ the set X extended by a top element if it lacks one and extended by a bottom element if it lacks one. Applying this construction to the real numbers, we get the extended real numbers, in which *every* subset has a supremum and an infimum.

If X and Y are ordered sets, a map $f : X \to Y$ is *isotone* (also called order-preserving), increasing or monotone, though the latter is used in analysis also for antitone (order-reversing) maps if for all $x_1 \leq x_2$ in X, we have $f(x_1) \leq f(x_2)$ in Y. A map $f : X \to Y$ is an *order-isomorphism* if it is isotone and admits an isotone inverse $g : Y \to X$. (An isotone bijection is an order-isomorphism, but defining an isomorphism as a bijective morphism is certainly wrong in other contexts, e.g., for topological spaces or partially ordered sets.)

Next we define intervals in an ordered set. The empty set is decreed to be an open interval in X. A closed, bounded interval in X is either the empty set or a subset of the form $[a, b] = \{x \in X \mid a \leq x \leq b\}$ for elements $a \leq b$ in X. A nonempty subset $I \subseteq X$ is an interval if $\inf I$ and $\sup I$ both exist in $\tilde{X}$ and $I \cup \{\inf I, \sup I\}$ is a closed, bounded interval in $\tilde{X}$. A nonempty interval I is open if the following hold:

(i) if $\inf I \in I$ then $\inf I$ is the bottom element of X and (ii) if $\sup I \in I$ then $\sup I$ is the top element of X. For $x \in X$, we explicitly define the following intervals:

$$^{<}x = \{y \in X \mid y < x\},\ ^{>}x = \{y \in X \mid y > x\},$$
$$^{\leq}x = \{y \in X \mid y \leq x\},\ ^{\geq}x = \{y \in X \mid y \geq x\}.$$

The open intervals form a base for a topology on X, called the *order topology*. The bounded open intervals—those of the form (a, b) for elements $a < b$ of X, $[\bot, b)$ for $b \in X$ if X has a bottom element, $(a, \top]$ for $a \in X$ if X has a top element and $X = [\bot, \top]$ if X has both top and bottom elements—are a base for the same topology. If $X = \mathbb{R}$ this is the usual Euclidean topology.

In some ways, order topologies are closer relatives to $\mathbb{R}$ than an arbitrary metric space, while in other ways they are more exotic. For example, they need not be metrizable or even first countable. Order topologies are always Hausdorff, so a compact subset must be closed. Moreover a compact subset C must be bounded—that is, contained in a closed, bounded interval: for each $x \in C$, let I_x be a bounded open interval containing x. Then there is a finite subset $Y \subseteq X$ such that $C \subseteq \bigcup_{x \in Y} I_x$ and $C \subseteq [\min_{x \in Y} \inf I_x, \max_{x \in Y} \sup I_x]$.

An ordered set is *Dedekind complete* if every nonempty subset that is bounded above has a supremum. This holds if and only if every nonempty subset that is bounded below has an infimum. An ordered set is complete if *every* subset has a supremum (if and only if every subset has an infimum).

Proposition 2. *Let X be an ordered set. Then,*
(a) X is Dedekind complete if and only if $\tilde{X}$ is complete.
(b) X is complete if and only if it is Dedekind complete and has top and bottom elements.

Proof. We observe that $\inf \varnothing$ exists if and only if X has a top element—in which case $\inf \varnothing = \top$—and $\sup \varnothing$ exists if and only if X has a bottom element—in which case $\inf \varnothing = \bot$. The rest is straightforward and left to the reader. ■

Ordered induction A subset S of an ordered set $(X, \leq)$ is *inductive* if it satisfies all of the following:

(IS1) There is $a \in X$ such that $^{\leq}a \subseteq S$.
(IS2) For all $x \in S$, either $x = \top$ or there is $y > x$ such that $[x, y] \subseteq S$.
(IS3) For all $x \in X$, if $^{<}x \subseteq S$, then $x \in S$.

Theorem 14 (Principle of ordered induction)**.** *For a nonempty ordered set X, the following are equivalent:*
(i) X is Dedekind complete.
(ii) The only inductive subset of X is X itself.

Proof. (i) $\Rightarrow$ (ii). Let $S \subseteq X$ be inductive. Seeking a contradiction, we suppose $S' = X \setminus S$ is nonempty. Fix $a \in X$ satisfying (IS1). Then a is a lower bound for S', so by hypothesis S' has an infimum, say y. Any element less than y is strictly less than every element of S', so $^{<}y \subset S$. By (IS3), $y \in S$. If $y = \top$, then $S' = \{\top\}$ or $S' = \varnothing$: both are contradictions. So $y < \top$, and then by (IS2) there exists $z > y$ such that $[y, z] \subseteq S$ and thus $^{\leq}z \subseteq S$. Thus z is a lower bound for S' that is strictly larger than y, a contradiction.

(ii) $\Rightarrow$ (i). Let $T \subseteq X$ be nonempty and bounded below by a. Let S be the set of lower bounds for T. Then $^{\leq}a \subseteq S$, so S satisfies (IS1).

Case 1. Suppose S does not satisfy (IS2): there is $x \in S$ with no $y \in X$ such that $[x, y] \subseteq S$. Since S is downward closed, x is the top element of S and $x = \inf T$.

Case 2. Suppose S does not satisfy (IS3): there is $x \in X$ such that ${}^{\prec}x \in S$ but $x \notin S$, i.e., there exists $t \in T$ such that $t < x$. Then also $t \in S$, so t is the least element of T: in particular $t = \inf T$.

Case 3. If S satisfies (IS2) and (IS3), then $S = X$, $T = \{\top\}$ and $\inf T = \top$. ■

Transfinite induction An ordered set X is well-ordered if every nonempty subset has a bottom element. If X is well-ordered and $\varnothing \subsetneq Y \subset X$ is bounded above, then (as usual) if Y has a top element $\top_Y$ then $\sup Y = \top_Y$; otherwise there is an element $x \in X$ such that $x > y$ for all $y \in Y$ and thus, by well ordering, a least such element, which is $\sup Y$. That is, well-ordered subsets are Dedekind complete, and thus, in view of Theorem 14, the only inductive subset of a well-ordered set X is X itself.

Let $x < y$ be elements of an ordered set X. If $[x, y] = \{x, y\}$ then we say that y is the successor of x and that x is the predecessor of y. If X is a nonempty well-ordered set, then every $y \neq \top$ has a successor. Clearly $\bot$ has no predecessor. The natural numbers form an infinite well-ordered set in which every $x \neq \bot$ has a predecessor, and this characterizes $\mathbb{N}_0$ up to order-isomorphism.

In a well-ordered set, (IS2) is equivalent to

(IS2′) For all every $x \in S$, either $x = \top$ or the successor of x also lies in S.

Thus we recover the following important result.

Theorem 15 (Principle of transfinite induction)**.** *Let X be a nonempty well-ordered set. Let S be a subset of X such that:*
(T1) We have $\bot \in S$.
(T2) If $x \in X$, either $x = \top$ or the successor of x also lies in S.
(T3) For all $y \in X$, if ${}^{\prec}y \subseteq S$, then $y \in S$.
Then $S = X$.

This statement is in fact rather redundant: applying (T3) with $y = \bot$ we get (T1); applying (T3) with y the successor of a non-top element x, we get (T2). The redundancy could be eliminated by requiring (T3) only for non-bottom elements that have no predecessors. But it is harmless, and moreover it is often natural to treat the three cases separately. (For instance, the three cases correspond to the three types of ordinal numbers.) As mentioned above, in $\mathbb{N}_0$ there is no non-bottom element without a predecessor, and thus applied therein Theorem 15 becomes the principle of mathematical induction.

Transfinite induction does not seem to have the ubiquitous presence in mathematics students' toolkits that it once did. If true, that is both unfortunate and beyond the scope of this article to remedy. However, we can recommend [**14**] which gives this topic the elegant presentation, context and range of applications that it deserves.

Completeness of subsets Let X be a Dedekind complete ordered set, and let $\varnothing \neq Y \subseteq X$. Then Y is an ordered set in its own right—when is it Dedekind complete? The analogy with metric spaces (and Cauchy completeness: that is, in which every Cauchy sequence converges) suggests that this holds if and only if Y is closed, but a little thought shows that this cannot be quite right. For example, the arctangent function gives an order isomorphism from $\mathbb{R}$ to $(\frac{-\pi}{2}, \frac{\pi}{2})$, so $(\frac{-\pi}{2}, \frac{\pi}{2})$ is Dedekind complete but not closed in $\mathbb{R}$. The precise answer is as follows: as usual, let $\tilde{X}$ be X augmented with a bottom element and/or a top element if and only if X lacks them. Then $\tilde{X}$ is complete by Proposition 2, so $\inf Y$ and $\sup Y$ exist in $\tilde{X}$. This allows us to view $\tilde{Y} = Y \cup \{\inf Y, \sup Y\}$ as a subset of the complete ordered set $\tilde{X}$.

Proposition 3. *Let Y be a subset of a Dedekind complete ordered set $(X, \leq)$ Then Y is Dedekind complete if and only if $\tilde{Y}$ is closed in $\tilde{X}$. It follows that every interval in a Dedekind complete ordered set is Dedekind complete.*

Proof. Suppose that Y is a Dedekind complete subset of the ordered set X, and let $x \in \tilde{X}$ be a point such that every neighborhood U of x in $\tilde{X}$ meets Y. Seeking a contradiction, we suppose that $x \notin \tilde{Y}$. Then at least one of the following holds: (i) for all elements $x' < x$ of X, we have $(x', x) \cap Y \neq \varnothing$, in which case $x = \sup\{y \in Y \mid y \leq x\} \in \tilde{Y}$, or (ii) for all elements $x'' > x$ of X we have $(x, x'') \cap Y \neq \varnothing$, in which case $x = \inf\{y \in Y \mid x \leq y\} \in \tilde{Y}$. This contradiction shows that $\tilde{Y}$ is closed in $\tilde{X}$.

Now suppose that X is Dedekind complete and that $Y \subset X$ is such that $\tilde{Y}$ is closed in $\tilde{X}$. Let A be a nonempty subset of Y that is bounded above by $M \in Y$, and let $a \in A$. Since X is Dedekind complete, $\sup A$ exists in X, and clearly $\sup A \in [a, M]$. Since $A \subset Y$, every open interval in X centered at $\sup A$ contains points of Y, so $\sup A \in \tilde{Y} \cap [a, M] \subset Y$, and Y is Dedekind complete.

The empty interval is Dedekind complete, and for every nonempty interval $I \subseteq X$, the set $\tilde{I}$ is a closed interval in $\tilde{X}$, so I is Dedekind complete. ■

Ordered fields An *ordered field* F is a field that is endowed with an ordering $\leq$ that satisfies the following compatibilities:

$$\text{for all } x, y, z \in F,\ x \leq y \text{ implies } x + z \leq y + z \text{ and}$$

$$\text{for all } x, y \in F,\ x \geq 0,\ y \geq 0 \text{ implies } xy \geq 0.$$

The real numbers $\mathbb{R}$ form an ordered field. Every subfield of an ordered field is again an ordered field, so one gets many examples by taking subfields of $\mathbb{R}$. The standard orderings on $\mathbb{R}$ and $\mathbb{Q}$ are in fact the unique orderings on these fields [**5**, Section 15.2]. Every ordered field has characteristic 0, so admits $\mathbb{Q}$ as an ordered subfield [**5**, *loc. cit.*].

Here is another example. Let $\mathbb{R}(t)$ be the field of rational functions $\frac{p(t)}{q(t)}$, that is, $p(t)$ and $q(t)$ are real polynomial functions and $q(t)$ is not identically zero. If $r, s \in \mathbb{R}(t)$ are distinct rational functions, then either $r(x) > s(x)$ for all sufficiently large x, which we denote $r > s$, or $r(x) < s(x)$ for all sufficiently large x, which we denote $r < s$. This makes $\mathbb{R}(t)$ into an ordered field. An element x of an ordered field is called infinitely large if $x > n$ for all $n \in \mathbb{Z}^+$. In the field $\mathbb{R}(t)$, the element t is infinitely large.

An ordered field that admits an infinitely large element is called non-Archimedean. The subfields of $\mathbb{R}$ are Archimedean, and conversely every Archimedean ordered field admits a unique isotone field embedding into $\mathbb{R}$ and thus may be identified with a subfield of $\mathbb{R}$ [**5**, Corollary 15.48]. If F is a proper subfield of $\mathbb{R}$, then F contains $\mathbb{Q}$ so is dense in $\mathbb{R}$. It follows that $\tilde{F}$ is not closed in the extended real numbers $\tilde{\mathbb{R}}$, so F is not Dedekind complete by Proposition 3. On the other hand, if F is non-Archimedean, then the upper bounds for $\mathbb{N}_0$ are precisely the infinitely large elements, which exist by definition, but if x is infinitely large then so is $x - 1$, so $\mathbb{N}_0$ has no supremum in F. We conclude that any Dedekind complete ordered field is isomorphic to $\mathbb{R}$ [**5**, Theorem 15.56].

Ordered induction is also a characteristic property of $\mathbb{R}$ among ordered fields, as the following corollary indicates.

Corollary 1. *In an ordered field $(F, +, \cdot, \leq)$, the following are equivalent:*
(i) F is Dedekind complete (and thus isomorphic to $\mathbb{R}$).

(ii) For all $a < b$ in F, the interval $[a, b]$ is complete.
(ii′) For all $a < b$ in F, the only inductive subset of $[a, b]$ is $[a, b]$.
(iii) There are $a < b$ in F such that the interval $[a, b]$ is complete.
(iii′) There are $a < b$ in F such that the only inductive subset of $[a, b]$ is $[a, b]$.

Proof. (i) $\Rightarrow$ (ii) by Proposition 3.
(ii) $\Leftrightarrow$ (ii′) and (iii) $\Leftrightarrow$ (iii′) by Theorem 14.
(ii) $\Rightarrow$ (iii) is clear.
(iii) $\Rightarrow$ (i). Suppose $[a, b]$ is complete. Let $S \subset F$ be any subset that is nonempty and bounded above, say by B. Let $A \in S$. Then S has a supremum in F if and only if $T = \{x \in S \mid A \leq x\} \subseteq [A, B]$ does. The map

$$\ell : [a, b] \to [A, B], \; x \mapsto \frac{B - A}{b - a}(x - a) + A$$

is an order-isomorphism, so $[A, B]$ is complete and T has a supremum in F. ∎

Problem 3. Characterize the inductive subsets of $[a, b] \cap \mathbb{Q}$.

Completeness and connectedness A subset Y of an ordered set $(X, \leq)$ is convex if for all $x, z, y \in X$ with $x < z < y$, if $x, y \in Y$ then also $z \in Y$. In any ordered set $(X, \leq)$, both intervals and connected sets are convex. The former is clear; as for the latter, if $Y \subseteq X$ is not convex, there are $x < z < y \in X$ with $x, y \in Y$ and $z \notin Y$, and then $Y_1 = {}^{<}z \cap Y$, $Y_2 = {}^{>}z \cap Y$ is a separation of Y. The converse implications depend on completeness, as indicated in the following proposition.

Proposition 4. *In an ordered set $(X, \leq)$, the following are equivalent:*
(i) X is Dedekind complete.
(ii) Every convex subset $Y \subseteq X$ is an interval.

Proof. (i) $\Rightarrow$ (ii). We may assume that Y is nonempty. Consider $\tilde{Y} \subseteq \tilde{X}$. We have $\tilde{Y} \subseteq [\inf Y, \sup Y]$. Conversely, if $\inf Y < z < \sup Y$ then there are $x, y \in Y$ with $x < z < y$, so $z \in Y$. Thus $\tilde{Y} = [\inf Y, \sup Y]$, so Y is an interval.
(ii) $\Rightarrow$ (i). We proceed by contraposition. Suppose X is not Dedekind complete, and let $Y \subseteq X$ be nonempty, bounded above and without a supremum in X. Let

$$D(Y) = \{x \in X \mid x \leq y \text{ for some } y \in Y\}.$$

Then $D(Y)$ is convex, bounded above and has no supremum, so not an interval. ∎

The next question is when intervals are connected. For this, even completeness is not sufficient. For example, a finite ordered set with more than one element is complete but not connected: the order topology is discrete. The extra condition we need is as follows: an ordered set $(X, \leq)$ is densely ordered if for all $x < y$ in X there is $z \in (x, y)$. A convex subset of a densely ordered set is again densely ordered.

Theorem 16. *For an ordered set X, the following are equivalent:*
(i) X is densely ordered and Dedekind complete.
(ii) X is connected in the order topology.

Proof. (i) $\Rightarrow$ (ii). Step 1. We suppose $\perp \in X$. Since X is densely ordered, a subset $S \subseteq X$ which contains $\perp$ and is both open and closed in the order topology is inductive. Since X is Dedekind complete, by Theorem 14, $S = X$. This shows X is connected.

Step 2. We may assume $X \neq \varnothing$ and choose $a \in X$. By Proposition 3, Step 1 applies to show $^{\geq}a$ is connected. A similar downward induction argument shows $^{\leq}a$ is connected. Since $X = {}^{\leq}a \cup {}^{\geq}a$ and ${}^{\leq}a \cap {}^{\geq}a \neq \varnothing$, X is connected.

(ii) $\Rightarrow$ (i). We proceed by contraposition. First, if X is not densely ordered, there are $a < b$ in X with $[a, b] = \{a, b\}$, so $A = {}^{\leq}a$, $B = {}^{\geq}b$ is a separation of X.

Next, suppose that X is densely ordered but there is a subset $S \subseteq X$ that is nonempty, bounded below by a and with no infimum. Let L be the set of lower bounds for X. Since S has no infimum, for all $\ell \in L$ there is $\ell' \in L$ with $\ell' > \ell$, and thus

$$\ell \in {}^{<}\ell' \subset L.$$

This shows that L is open. Now let $x \in X \setminus L$. Then x is not a lower bound for S, so there is $s \in S$ with $s < x$. Since X is densely ordered, there is $y \in X$ with $s < y < x$, and then

$$x \in {}^{>}y \subset X \setminus L.$$

This shows that L is closed. Since $a \in L$ and X is connected, we must have $L = X$. Thus $L \cap S = S$ is nonempty, but any element of $L \cap S$ is an infimum for S, a contradiction. ■

Corollary 2. *Let $(X, \leq)$ be densely ordered and Dedekind complete. For a subset $Y \subseteq X$, the following are equivalent:*
(i) Y is connected in the order topology.
(ii) Y is convex.
(iii) Y is an interval.

Proof. (i) $\Rightarrow$ (ii) was shown above for any order topology.

(ii) $\Rightarrow$ (iii) by Proposition 4.

(iii) $\Rightarrow$ (i): Being an interval, Y is a convex subset of a densely ordered set, so Y is densely ordered. By Proposition 3, Y is Dedekind complete, so by Theorem 16, Y is connected in the order topology. ■

Above we saw that an ordered field F is Dedekind complete if and only if there are $a < b$ in F such that the interval $[a, b]$ is complete. This has the following consequence.

Corollary 3. *Let $(F, +, \cdot, <)$ be an ordered field. The following are equivalent:*
(i) F is Dedekind complete (and thus isomorphic to $\mathbb{R}$).
(ii) Every closed interval $[a, b]$ of F is connected in the order topology.
(iiii) For some $a < b$ in F, the interval $[a, b]$ is connected in the order topology.

It follows that if $I \subseteq \mathbb{R}$ is an interval and $f : I \to \mathbb{R}$ is continuous, then $f(I)$ is again an interval. If I is closed and bounded, then it is compact, so $f(I)$ is again closed and bounded. Conversely, it is a nice exercise to show that if $I, J \subseteq \mathbb{R}$ are intervals, each consisting of more than one point, and J is closed and bounded if I is, then there is a continuous function $f : I \to \mathbb{R}$ with $f(I) = J$.

Completeness and compactness

Theorem 17. *For an ordered set X, the following are equivalent:*
(i) X is complete.
(ii) X is compact in the order topology.

Proof. (i) $\Rightarrow$ (ii). Let $\mathcal{U} = \{U_i\}_{i\in I}$ be an open covering of X. Let S be the set of $x \in X$ such that the covering $\mathcal{U} \cap [\bot, x]$ of $[\bot, x]$ admits a finite subcovering. We have $\bot \in S$, so S satisfies (IS1). Suppose $U_1 \cap [\bot, x], \ldots, U_n \cap [\bot, x]$ covers $[\bot, x]$. If there exists $y \in X$ such that $[x, y] = \{x, y\}$, then adding to the covering any element U_y containing y gives a finite covering of $[\bot, y]$. Otherwise some U_i contains x and hence also $[x, y]$ for some $y > x$. So S satisfies (IS2). Now suppose that $x \neq \bot$ and $[\bot, x) \subseteq S$. Let $i_x \in I$ be such that $x \in U_{i_x}$, and let $y < x$ be such that $(y, x] \subseteq U_{i_x}$. Since $y \in S$, there is a finite $J \subseteq I$ with $\bigcup_{i\in J} U_i \supset [a, y]$, so $\bigcup_{i\in J} U_i \cup U_{i_x} \supset [a, x]$. Thus $x \in S$ and S satisfies (IS3). Thus S is inductive; since X is Dedekind complete, we have $S = X$. In particular $\top \in S$, hence the covering has a finite subcovering.

(ii) $\Rightarrow$ (i). For each $x \in X$ there is a bounded open interval I_x containing x. If X is compact, $\{I_x\}_{x\in X}$ has a finite subcovering, so X is bounded, i.e., has $\bot$ and $\top$. Let $S \subseteq X$. Since $\inf \varnothing = \top$, we may assume $S \neq \varnothing$. Let L be the set of lower bounds for S. For each $(b, s) \in L \times S$, consider the closed interval $C_{b,s} := [b, s]$. For any finite subset $\{(b_1, s_1), \ldots, (b_n, s_n)\}$ of $L \times S$, we have

$$\bigcap_{i=1}^{N} [b_i, s_i] \supset [\max b_i, \min s_i] \neq \varnothing.$$

Since X is compact, there is $y \in \bigcap_{L\times S}[b, s]$ and then $y = \inf S$. ■

Corollary 4 (Generalized Heine–Borel)**.** *(a) For an ordered set X, the following are equivalent:*
(i) X is Dedekind complete.
(ii) A subset S of X is compact in the order topology if and only if it is closed and bounded.
(b) For an ordered field F, the following are equivalent:
(i) F is Dedekind complete (and thus isomorphic to $\mathbb{R}$).
(ii) Every closed bounded interval $[a, b] \subseteq F$ is compact.
(iii) For some $a < b$ in F, the interval $[a, b]$ is compact.

Proof. (a) (i) $\Rightarrow$ (ii). A compact subset of any ordered space is closed and bounded. Conversely, if X is Dedekind complete and $S \subseteq X$ is closed and bounded, then by Proposition 3, S is complete and then by Theorem 17, S is compact.

(ii) $\Rightarrow$ (i). If $S \subseteq X$ is nonempty and bounded above, let $a \in S$. Then $S' = S \cap {}^{\geq}a$ is bounded, so $\overline{S'}$ is compact and thus $\overline{S'}$ is complete by Theorem 17. The least upper bound of $\overline{S'}$ is also the least upper bound of S.

(b) This follows immediately from part (a) and Corollary 1. ■

Comments and complements The notion of Dedekind completeness goes back to Dedekind's construction of $\mathbb{R}$ using Dedekind cuts. In an ordered set X, a *Dedekind cut* is a pair (L, R) of subsets of X such that R is the set of upper bounds for L and L is the set of lower bounds for R. Then $L \cup R = X$. Indeed, let $x \in X$; if $x \notin R$ then there is $\ell \in L$ with $x < \ell$, and if $x \notin L$ then there is $r \in R$ with $r < x$, but then $r < x < \ell$, so $r \in R$ is not an upper bound for L. Similarly, $L \cap R$ is either empty or consists of a single point $x = \top_L = \bot_R$. In the latter case we say the cut is principal.

Example 1. If the ordered set X has a bottom element $\bot$, then $(\{\bot\}, X)$ is a principal cut in X; otherwise $(\varnothing, X)$ is a nonprincipal cut in X. Similarly, if X has a top element $\top$, then $(X, \{\top\})$ is a principal cut in X; otherwise $(X, \varnothing)$ is a nonprincipal cut in X.

Example 2. In $\mathbb{Q}$,

$$L = \{x \in \mathbb{Q} \mid x < \sqrt{2}\}, \ R = \{x \in \mathbb{Q} \mid \sqrt{2} < x\}$$

is a nonprincipal cut. To define a similar cut in $\mathbb{R}$ we must place $\sqrt{2}$ in both L and R, giving a principal cut. However $(\varnothing, \mathbb{R})$ and $(\mathbb{R}, \varnothing)$ are nonprincipal cuts in $\mathbb{R}$.

An ordered set X is complete (resp. Dedekind complete) if and only if every cut (resp. every cut different from $(\varnothing, X)$ and $(X, \varnothing)$) is principal. (We leave this for the interested reader to prove—by the principle of ordered induction or otherwise!) Let $\mathcal{C}(X)$ be the set of cuts in X. For (L_1, R_1), (L_2, R_2) in X, we put $(L_1, R_1) \leq (L_2, R_2)$ if and only if $L_1 \subseteq L_2$. This makes $\mathcal{C}(X)$ into a complete ordered set. For $x \in X$, we define

$$\iota(x) = (\{y \in X \mid y \leq x\}, \{z \in X \mid x \leq z\}).$$

Then $\iota : X \hookrightarrow \mathcal{C}(X)$ is an isotone injection. Thus every ordered set can be canonically embedded in a complete ordered set, which may expand the range of applicability of the Principle of Ordered Induction.

Problem 4. Use the embedding $\iota : X \hookrightarrow \mathcal{C}(X)$ to give applications of the principle of ordered induction to ordered sets X that are *not* Dedekind complete.

The principality of Dedekind cuts may not be the most initially appealing completeness axiom, but it can be an elegant proof technique. Propp puts it to good use in [**24**].

Most of the above results can be found piecemeal in various places. The implication (i) $\Rightarrow$ (ii) in Theorem 17 is due to Frink [**11**]. This is probably the more interesting direction. A different proof of (ii) $\Rightarrow$ (i) goes by contraposition: if X is not complete, then there is a nonprincipal Dedekind cut (L, R), which one can use to construct an open cover without a finite subcover: cf. [**29**, p. 67]. The implication (i) $\Rightarrow$ (ii) of Theorem 16 is treated by Munkres [**20**, Theorem 24.1]. Similarly, [**20**, Theorem 27.1] gives a portion of Corollary 4.

A subtlety arises when considering the topology on a subset Y of an ordered set $(X, \leq)$. On the one hand, restricting $\leq$ to Y makes Y an ordered set in its own right, and thus it gets an order topology. On the other hand, we can endow Y with the topology it inherits as a subspace of the order topology on X. In general the order topology is coarser than the subspace topology, and they need not coincide. For example, consider $Y = \{0\} \cup (1/2, 1] \subseteq \mathbb{R}$. Then $\{0\}$ is open in the subspace topology on Y but not in the order topology—with the order topology, Y is homeomorphic to $[\frac{1}{2}, 1]$. Moreover there is no ordering on Y that induces the subspace topology. Thus in the above results we were careful to specify "in the order topology."

For a convex subset $Y \subseteq X$ the two topologies coincide. We give an example in which the distinction matters. An ordered field F that is *not* Dedekind complete is totally disconnected in the order topology. That is, if $Y \subseteq F$ consists of more than a single point, then Y is not connected *in the subspace topology*. However, if $F \supset \mathbb{R}$—e.g., $F = \mathbb{R}(t)$—then the subset $\mathbb{R}$ is connected in the order topology.

Acknowledgments The author thanks François Dorais, William G. Dubuque, Harold Erazo, Bryce Glover, Joel D. Hamkins, Niles Johnson, Iraj Kalantari, Paul Pollack, James Propp, and Alex Rice for helpful conversations, for pointers to the literature, and for identifying typos.

REFERENCES

[1] Andre, N. R., Engdahl, S. M., Parker, A. E. (2013). An analysis of the first proofs of the Heine-Borel theorem. *Convergence*. doi.org/10.4169/loci003890

[2] Bereková, H. (1982). The principle of induction in continuum and related methods. *Acta Math. Univ. Comenian.* 40(41): 97–100.

[3] Bosák, J. (1958). Generalization of the method of complete induction. *Acta Fac. Nat. Univ. Comenian. Math.* 2: 255–256.

[4] Chao, Y. R. (1919). A note on "Continuous mathematical induction." *Bull. Amer. Math. Soc.* 26: 17–18.

[5] Clark, P. L. Field theory. http://math.uga.edu/~pete/FieldTheory.pdf

[6] Clark, P. L. https://math.stackexchange.com/q/4204

[7] Deveau, M., Teismann, H. (2013/4). 72+42: characterizations of the completeness and Archimedean properties of ordered fields. *Real Anal. Exchange*. 39: 261–303.

[8] Dowek, G. (2003). *Preliminary Investigations on Induction over Real Numbers*, preprint.

[9] Duren, W. L., Jr. (1957). Mathematical induction in sets. *Amer. Math. Monthly*. 64: 19–22.

[10] Ford, L. R. (1957). Interval-additive propositions. *Amer. Math. Monthly*. 64: 106–108.

[11] Frink, O., Jr. (1942) Topology in lattices. *Trans. Amer. Math. Soc.* 51: 569–582.

[12] Hathaway, D. (2011). Using continuity induction. *College Math. J.* 42: 229–231.

[13] Kalantari, I. (2007). Induction over the continuum. In *Induction, Algorithmic Learning Theory, and Philosophy*. Logic, Epistemology, and the Unity of Science, Vol. 9, Dordrecht, Germany: Springer, pp. 145–154.

[14] Kaplansky, I. (1977). *Set Theory and Metric Spaces.* 2nd ed. New York: Chelsea Publishing Co.

[15] Khinchin, A. (1923). Das Stetigkeitsaxiom des Linearkontinuums als Induktionsprinzip betrachtet. *Fund. Math.* 4: 164–166.

[16] Khinchin, A. (1949). The simplest linear continuum. *Uspehi Matem. Nauk (N.S.)* 4(30): 180–197.

[17] Lebesgue, H. *Leçons sur l'intégration et la recherche des fonctions primitives.* Paris.

[18] Leinfelder, H. (1982/83). A unifying principle in real analysis. *Real Anal. Exchange*. 8(2): 511–518.

[19] Moss, R. M. F., Roberts, G. T. (1968). A creeping lemma. *Amer. Math. Monthly.* 75: 649–652.

[20] Munkres, J. R. (2000). *Topology*. 2nd ed. Upper Saddle River, NJ: Prentice Hall, Inc.

[21] Neubrunnová, A. (1986/87). On a unifying principle in real analysis. *Acta Math. Univ. Comenian.* 48/49: 123–126.

[22] Norris, M. J. (1952). Integrability of continuous functions. *Amer. Math. Monthly.* 59: 244–245.

[23] Perron, O. (1926). Was sind und sollen die irrationalen Zahlen? *Jber. Deutsch. Math. Verein.* 35: 194–203.

[24] Propp, J. (2013). Real analysis in reverse. *Amer. Math. Monthly.* 120: 392–408.

[25] Šalát, T. (1984/85). Remarks on unifying principles in real analysis. *Real Anal. Exchange*. 10(2): 343–348.

[26] Shanahan, P. (1972). A unified proof of several basic theorems of real analysis. *Amer. Math. Monthly*. 79: 895–898.

[27] Shanahan, P. (1974). Addendum to: A unified proof of several basic theorems of real analysis. *Amer. Math. Monthly*. 81: 890–891.

[28] Spivak, M. (2009). *Calculus*. 4th ed. Houston, TX: Publish or Perish Press.

[29] Steen, L. A., Seebach, J. A., Jr. (1978). *Counterexamples in Topology.* 2nd ed. New York/Heidelberg: Springer-Verlag.

[30] Teissmann, H. (2013). Toward a more complete list of completeness axioms. *Amer. Math. Monthly*. 120: 99–114.

[31] Viro, O. Ya., Ivanov, O. A., Netsvetaev, O. A. Yu., Kharlamov, V. M. (2008). *Elementary Topology. Problem Textbook.* Providence, RI: American Mathematical Society.

[32] Willard, S. (1970). *General Topology*. Reading, MA: Addison-Wesley Publishing Co.

Summary. We introduce real induction, a proof technique analogous to Mathematical Induction but applicable to statements indexed by an interval on the real line. We apply these principles to give streamlined, conceptual proofs of basic results in elementary real analysis and topology. Then we pursue inductive principles in arbitrary ordered sets. Applications are given, e.g., to "real analysis in reverse."

PETE L. CLARK (MR Author ID: 767639) is a professor of mathematics at the University of Georgia. His primary research interest is number theory.

PROBLEMS

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

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

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

Proposals

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

2066. *Proposed by George Stoica, New Brunswick, Canada.*

Is there a function $f : \mathbb{R} \to \mathbb{R}$ satisfying $|f(x+y)| \geq |f(x)+f(y)|$ for all $x, y \in \mathbb{R}$, with strict inequality for at least some x, y? How about a function satisfying the reverse inequality $|f(x+y)| \leq |f(x)+f(y)|$ everywhere, strictly somewhere?

2067. *Proposed by Elton Bojaxhiu, Eppstein am Taunus, Germany and Enkel Hysnelaj, Sydney, Australia.*

Chord $\overline{XY}$ of a circle $\mathcal{C}$ is not a diameter. Let P, Q be two different points strictly inside $\overline{XY}$ such that Q lies between P and X. Chord $\overline{MN}$ is perpendicular to the diameter of $\mathcal{C}$ through Q, where $MP < NP$. Prove that $(MQ - PQ) \cdot XY \geq 2 \cdot QX \cdot PY$, and characterize those cases in which equality holds.

2068. *Proposed by Ovidiu Furdui and Alina Sîntămariăn, Technical University of Cluj-Napoca, Cluj-Napoca, Romania.*

Prove that the series

$$\sum_{n=1}^{\infty} \frac{3 \cdot 6 \cdots (3n)}{7 \cdot 10 \cdots (3n+4)} \cdot \frac{1}{3n+7}$$

converges, and find its sum.

2069. *Proposed by Eugene Delacroix, Lycee Therese dAvila, France and Su Pernu Mero, Valenciana GTO, Mexico.*

Math. Mag. **91** (2) 151–158. doi:10.1080/0025570X.2019.1569894.

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.

Three points are chosen uniformly and independently at random in the unit interval $[0, 1]$. These points divide the interval into four segments of lengths a, b, c, and d. Find the expected value and standard deviation of the random variable $X = \max(a, b, c, d)$.

2070. *Proposed by Enrique Treviño, Lake Forest College, Lake Forest, IL.*

Fix a prime p. For any integer $n \geq p$, let S_n be the number of ways of coloring n points using p distinct colors, each at least once. Characterize those n such that S_n is not a multiple of p^2.

Quickies

1089. *Proposed by Richard Stephens, Columbus State University, Columbus, GA.*

Is there a function $f : \mathbb{R} \to \mathbb{R}$ such that (i) f is discontinuous everywhere, and (ii) $f(f(x)) = -x$ for all $x \in \mathbb{R}$?

1090. *Proposed by Konstantinos Gaitanas, Volos, Greece.*

Find all primes $p > 2$ such that the range of the sequence $\{a_n\}$ in $\mathbb{Z}_p$ (the ring of integers modulo p) defined recursively by

- $a_0 = 1$, and
- $a_{n+1} = 2^{a_n} \pmod{p}$ for $n \geq 0$,

is equal to $\mathbb{Z}_p - \{0\}$.

Solutions

2041. *Proposed by Vadim Mitrofanov, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine.*

Let $ABCD$ be a quadrilateral that circumscribes a circle of radius r and is also inscribed in a circle of radius R. Let s be the semiperimeter of $ABCD$. Prove the inequality $s^2 \leq 6R^2 + 4r^2$.

Solution by Elton Bojaxhiu, Eppstein am Taunus, Germany and Enkel Hysnelaj, Sydney, Australia.
Let $a = AB$, $b = BC$, $c = CD$ and $d = DA$ be the side lengths of $ABCD$ and $s = (a + b + c + d)/2$ its semiperimeter. Since $ABCD$ is a circumscribed quadrilateral, its area is $K = rs$, and we further have $s = a + c = b + d$. Since $ABCD$ is cyclic, the relations

$$K = \sqrt{(s-a)(s-b)(s-c)(s-d)} \qquad \text{(Brahmagupta's formula), and}$$

$$4KR = \sqrt{(ab+cd)(ac+bd)(ad+bc)} \qquad \text{(Parameshvara's formula)}$$

also hold, hence

$$R^2 = \frac{(ab+cd)(ac+bd)(ad+bc)}{16(s-a)(s-b)(s-c)(s-d)} = \frac{(ab+cd)(ac+bd)(ad+bc)}{16abcd}, \quad (1)$$

$$r^2 = \frac{K^2}{s^2} = \frac{(s-a)(s-b)(s-c)(s-d)}{s^2} = \frac{abcd}{s^2}. \quad (2)$$

By homogeneity, we may henceforth assume $s = 1$ without loss of generality. Let $x = ac$ and $y = bd$. Since $a + c = b + d = s = 1$, we have $x, y \le 1/4$. Now let $u = ac + bd = x + y \le 1/2$ and $v = \sqrt{abcd} = \sqrt{xy}$. By the AM–GM inequality, we have $v \le u/2 \le 1/4$, hence $3 - 8v \ge 1$, so $(3-8v)^2 \ge 1$, i.e., $0 \le 8v^2 - 6v + 1 = 3(1-v)^2 - (2-5v^2)$. Again since $u/(2v) \ge 1$, it follows that

$$2 - 5v^2 \le 3(1-v)^2 \le 3\left(\frac{u}{2v} - v\right)^2 \quad \Rightarrow \quad 1 \le \frac{3u(u-4v^2)}{8v^2} + 4v^2. \quad (3)$$

From $a + c = 1 = b + d$, we get

$$u - 4v^2 = (ac \cdot 1 + bd \cdot 1) - 4abcd = ac(b+d)^2 + bd(a+c)^2 - 4abcd$$
$$= ab^2c + acd^2 + a^2bd + bc^2c = (ab+cd)(ad+bc).$$

Since $s = 1$, inequality (3) and equations (1), (2) give

$$s^2 = 1 \le 6 \cdot \frac{(ab+cd)(ac+bd)(ad+bc)}{16abcd} + 4abcd = 6R^2 + 4r^2.$$

Also solved by Michel Bataille, Robin Chapman (UK), Kyle Gatesman, Subhankar Gayen (India), Omran Kouba (Syria), Elias Lampakis (Greece), Joel Schlosberg, Achilleas Sinefakopoulos (Greece), Michael Vowe, and the proposer. There was 1 incomplete or incorrect solution.

2042. *Proposed by Rick Mabry and Debbie Shepherd, Louisiana State University Shreveport, Shreveport, LA.*

Recursively define random variables $X_0, X_1, \ldots, X_n, \ldots$ and $Y_0, Y_1, \ldots, Y_n, \ldots$ taking values in $[0, 1]$ as follows:

- $X_0 = 0$ and $Y_0 = 1$ are constants;
- for $n = 0, 1, 2, \ldots$, X_{n+1} and Y_{n+1} are chosen uniformly and independently in the closed interval with endpoints X_n, Y_n.

Prove that, with probability 1, the limits $\widetilde{X} = \lim_{n\to\infty} X_n$ and $\widetilde{Y} = \lim_{n\to\infty} Y_n$ both exist and are equal, and find their common distribution.

Solution by Northwestern University Math Problem Solving Group, Evanston, IL.
We will prove:

1. The limits $\widetilde{X}$ and $\widetilde{Y}$ exist and are equal with probability 1.
2. The common distribution of $\widetilde{X}$ and $\widetilde{Y}$ has probability density $g(t) = 6t(1-t)$ on $[0, 1]$.

1. Denote by $[\![X_n, Y_n]\!]$ the closed interval with endpoints X_n, Y_n, i.e., $[\![X_n, Y_n]\!] = [X_n, Y_n]$ if $X_n \le Y_n$, and $[\![X_n, Y_n]\!] = [Y_n, X_n]$ if $X_n > Y_n$. Since X_{n+1}, Y_{n+1} are chosen in $[\![X_n, Y_n]\!]$, we have $[\![X_{n+1}, Y_{n+1}]\!] \subseteq [\![X_n, Y_n]\!]$ for every $n \ge 0$, so $\{[\![X_n, Y_n]\!]\}_{n\ge 0}$ is a sequence of nested intervals. By the nested interval theorem their intersection is nonempty; it consists of a unique point (equal to a common limit of X_n and Y_n) precisely if the interval lengths tend to zero. We prove next that this is the case with probability 1.

Let $L_n = |Y_n - X_n|$ be the length of $[\![X_n, Y_n]\!]$, and let $\delta \in (0, 1)$. Since X_{n+1}, Y_{n+1} are independently and uniformly chosen in $[\![X_n, Y_n]\!]$, it is easy to verify that

$$\mathbf{P}[L_{n+1} \ge \delta \mid L_n] = \begin{cases} \left(1 - \dfrac{\delta}{L_n}\right)^2, & \delta \le L_n; \\ 0, & \delta > L_n. \end{cases}$$

Since $L_n \le 1$, we have $(1 - \delta/L_n)^2 \le (1 - \delta)^2$ whenever $L_n \ge \delta$. It follows by routine induction that $\mathbf{P}[L_n \ge \delta] \le (1 - \delta)^{2n}$ for every $n \ge 0$. Indeed, this assertion holds for $n = 0$ since $L_0 = |Y_0 - X_0| = |1 - 0| = 1 \ge \delta$ occurs with probability $1 \le (1 - \delta)^0$. Assuming next that $\mathbf{P}[L_n \ge \delta] \le (1 - \delta)^{2n}$ for some $n \ge 0$, we have

$$\begin{aligned} \mathbf{P}[L_{n+1} \ge \delta] &= \mathbf{P}[L_n \ge \delta] \cdot \mathbf{P}[L_{n+1} \ge \delta \mid L_n \ge \delta] \le (1 - \delta)^{2n}(1 - \delta/L_n)^2 \\ &\le (1 - \delta)^{2n}(1 - \delta)^2 = (1 - \delta)^{2(n+1)}, \end{aligned}$$

completing the inductive proof. Since $0 < \delta < 1$, the upper bound $(1 - \delta)^{2n}$ approaches 0 as n tends to infinity, so we see that the event $\lim_{n\to\infty} L_n = 0$ has probability 1. This shows that (with probability 1) the sequence of nested intervals $\{[\![X_n, Y_n]\!]\}_{n\ge 0}$ converges to a single point, and the intervals' endpoints X_n, Y_n converge to that same limit $\widetilde{X} = \widetilde{Y}$.

2. We show that the common distribution of $\widetilde{X}$ and $\widetilde{Y}$ has probability density $g(t) = 6t(1 - t)$ on $[0, 1]$. For each integer $n \ge 0$ define a new random variable Z_n in $[\![X_n, Y_n]\!]$ whose conditional probability density given the event $\{X_n = x, Y_n = y\}$ is

$$f_{Z_n|\{X_n=x,Y_n=y\}} = g_{x,y}(t) = \frac{1}{|y - x|} \cdot g\left(\frac{t - x}{y - x}\right) = \frac{6(t - x)(y - t)}{|y - x|^3} \qquad \text{on } [\![x, y]\!].$$

(We may take $Z_n = X_n$ when $X_n = Y_n$, but this event has zero probability and may as well be ignored.) Since $Z_n \in [\![X_n, Y_n]\!]$, the limit $\widetilde{Z} = \lim_{n\to\infty} Z_n$ satisfies $\widetilde{Z} = \widetilde{X} = \widetilde{Y}$ with probability 1, and its distribution coincides with that of both $\widetilde{X}$ and $\widetilde{Y}$.

In what follows, we write $f^{(n)}$ for f_{Z_n}, and $f^{(n)}_{\mathcal{E}}$ for the conditional density $f_{Z_n|\mathcal{E}}$ of Z_n given an event $\mathcal{E}$. Next, we shall show by induction that for every $n \ge 0$ the (unconditional) probability density of Z_n on $[0, 1]$ is $f^{(n)} = g$. For $n = 0$ we have $X_0 = 0$, $Y_0 = 1$ are constant, so $f^{(0)} = f^{(0)}_{\{X_0=0,Y_0=1\}} = g_{0,1} = g$. Next, fix $n \ge 0$ and assume that the probability density of Z_n on $[0, 1]$ is $f^{(n)} = g$. It is clear from the recursive nature of the problem that the conditional distribution of Z_{n+1} given an arbitrary outcome $\{X_1 = x, Y_1 = y\}$ of X_1, Y_1 (for $x, y \in [0, 1]$) is exactly the distribution as that of Z_n given $\{X_0 = x, Y_0 = y\}$, i.e., the distribution that Z_n would have provided that the entire process were started setting the deterministic variables X_0, Y_0 to the initial values $X_0 = x, Y_0 = y$ instead of $X_0 = 0, Y_0 = 1$. (This is true whether $x \le y$ or $y > x$.) The inductive hypothesis is thus subsumed in the assertion that $f^{(n+1)}_{\{X_1=x,Y_1=y\}} = g_{x,y}$ for $x, y \in [0, 1]$. For fixed $t \in [0, 1]$, the probability density function $f^{(n+1)}$ of Z_{n+1} is the

expected value (average) of the probability density functions $f^{(n)}_{\{X_1=x,Y_1=y\}}$ of Z_{n+1}, so we have

$$f^{(n+1)}(t) = \int_0^1 \int_0^1 f^{(n+1)}_{\{X_1=x,Y_1=y\}}(t)\,dy\,dx = \int_0^1 \int_0^1 g_{x,y}(t)\,dy\,dx$$
$$= 2\int_0^t \int_t^1 g_{x,y}(t)\,dy\,dx,$$

since the fact that $g_{x,y} = g_{y,x}$ implies that the double integral over $[0,1]^2$ is twice the integral over the triangle $\{0 \le x \le y \le 1\}$, and moreover in this region the integrand is actually supported on the rectangle $\{0 \le x \le t\} \times \{t \le y \le 1\}$ because $g_{x,y}$ is supported on $[\![x, y]\!]$. Routine computation now gives

$$f^{(n+1)}(t) = 2\int_0^t \int_t^1 \frac{6(t-x)(y-t)}{(y-x)^3}\,dy\,dx = 12\int_0^t (t-x)\cdot \int_{t-x}^{1-x} \frac{z-(t-x)}{z^3}\,dz\,dx$$
$$= 12\int_0^t (t-x)\cdot\left[\frac{(t-x)}{2z^2} - \frac{1}{z}\right]_{z=t-x}^{z=1-x} dx = 6\int_0^t \left(\frac{1-t}{1-x}\right)^2 dx$$
$$= 6\left[\frac{(1-t)^2}{1-x}\right]_{x=0}^{x=t} = 6t(1-t) = g(t).$$

This completes the inductive step of the proof that $f_{Z_n} = f^{(n)} = g$ on $[0,1]$ for every integer $n \ge 0$. Thus, the limit $\widetilde{Z} = \lim_{n\to\infty} Z_n$ and hence also $\widetilde{X}, \widetilde{Y}$ must have the same distribution, with probability density $g(t) = 6t(1-t)$ on $[0,1]$.

Editor's Note. Stephen Herschkorn remarked that the problem may be regarded as one starting from uniformly distributed (i.e., Beta(1, 1)) variables X_n, Y_n to produce Beta(2, 2)-distributed variables $\widetilde{X} = \widetilde{Y}$. He also verified that for $n = 1, 2, 3, 4, 5$, if X_n, Y_n are Beta(n, n)-distributed in the analogous sequence of nested intervals, then the corresponding limit $\lim_{n\to\infty} X_n = \lim_{n\to\infty} Y_n$ has distribution Beta$(2n, 2n)$. This observation suggests a more general underlying question.

Also solved by Elton Bojaxhiu (Germany) & Enkel Hysnelaj (Australia), Robert Calcaterra, Kyle Gatesman, Stephen J. Herschkorn, Omran Kouba (Syria), Yang Liu, Albert Natian, Kenneth Schilling, Nicholas C. Singer, Mark Wildon (UK), and the proposer.

2043. *Proposed by Greg Oman, University of Colorado, Colorado Springs and Adam Salminen, University of Evansville, Evansville, IN.*

Find all commutative rings R with unity such that:

(i) R contains some element x that is neither nilpotent nor a unit (i.e., $x^n \ne 0$ for all $n \ge 1$ and $xy \ne 1$ for all $y \in R$), and

(ii) every proper nonzero ideal of R is maximal.

Solution by Tom Jager, Calvin College, Grand Rapids, MI.

We prove that such rings R are precisely direct sums $F_1 \oplus F_2$ of two fields F_1, F_2.

First assume R is a direct sum $F_1 \oplus F_2$ of fields. Units of R are elements (a, b) with nonzero $a \in F_1$, $b \in F_2$. Clearly, the element $x = (1, 0)$ of R is neither nilpotent nor a unit. Let I be a nonzero proper ideal of R. If (a, b) is a nonzero element of I, then

(a, b) is necessarily a non-unit since I is proper, so exactly one of a, b is nonzero. If $a \neq 0$ and $b = 0$ then $I \supseteq R(a, 0) = F_1 \oplus \langle 0 \rangle$. However, the ideal $F_1 \oplus \langle 0 \rangle$ is maximal in R since $R/(F_1 \oplus \langle 0 \rangle) \cong F_2$ is a field, so necessarily $I = F_1 \oplus \langle 0 \rangle$ in this case. If $a = 0$ and $b \neq 0$ we conclude $I = \langle 0 \rangle \oplus F_2$ similarly.

Conversely, assume that R satisfies the conditions of the problem. Let x be a non-unit, non-nilpotent element of R. Clearly, so is x^2. Thus, the ideals $\langle x \rangle$ and $\langle x^2 \rangle$ are both proper and nonzero, hence maximal by hypothesis; however, $\langle x \rangle$ includes $\langle x^2 \rangle$, so we must have $\langle x \rangle = \langle x^2 \rangle$ by maximality. In particular, $x \in \langle x^2 \rangle = Rx^2$ implies that $x = ax^2$ for some $a \in R$. It follows that $x(1 - ax) = 0$. Letting $y = 1 - ax$ we have $xy = 0$; in addition, y is not a unit since $x \neq 0$. Furthermore, $ax \in \langle x \rangle$ but $1 \notin \langle x \rangle$, so $y \notin \langle x \rangle$. Thus, $\langle y \rangle$ is nonzero proper ideal of R, hence maximal by hypothesis. Elements of the intersection $\langle x \rangle \cap \langle y \rangle$ are of the form $z = ux = vy$ for some $u, v \in R$; thus, $z = ux = u(ax^2) = ax(ux) = ax(vy) = av(xy) = 0$. Hence, $\langle x \rangle \cap \langle y \rangle = \langle 0 \rangle$, and we further have $1 = ax + y \in \langle x \rangle + \langle y \rangle$, so $\langle x \rangle + \langle y \rangle = R$. The elements $e_1 = ax$ and $e_2 = y$ satisfy $e_1 e_2 = 0$, $e_1 + e_2 = 1$, and also $e_1 x = axx = ax^2 = x$ and $e_2 y = (1 - ax)y = y - axy = y - ax(1 - ax) = y - ax + a^2x^2 = y - ax + ax = y = e_2$. Thus, $F_1 = \langle e_1 \rangle$ and $F_2 = \langle e_2 \rangle$ are rings, with unities e_1 and e_2, respectively, such that $R = F_1 + F_2$ and $F_1 \cap F_2 = \langle 0 \rangle$. If I is any ideal of F_1 and $r \in R$ is arbitrary, we have $r = ce_1 + de_2$ for some $c, d \in R$ (since $F_1 + F_2 = R$), hence for all $z \in I$ we have $rz = ce_1 z + de_2 z = cz \in I$, since $e_1 z = z$ (as $z \in F_1$) and $e_2 z = 0$ (as $e_2 z \in F_2 \cap F_1 = \langle 0 \rangle$). It follows that a nonzero ideal I of F_1 is actually a (necessarily proper) nonzero ideal of R, thus maximal in R by hypothesis, and therefore $I = F_1$ since I and F_1 are both maximal in R. This shows that F_1 is a field, and analogously so is F_2. We have already shown that R is their direct sum, concluding the proof.

Also solved by Paul Budney, Robert Calcaterra, Souvik Dey, Joseph DiMuro, Abhay Goel, Missouri State University Problem Solving Group, Greg Oman & Adam Salminen, Francisco Perdomo and Ángel Plaza (Spain), Michael Reid, John H. Smith, Mark Wildon (UK), and the proposer.

2044. *Proposed by George Stoica, Saint John, New Brunswick, Canada.*

Find all continuous functions $f : [0, 1] \to [0, \infty)$ such that

$$\lim_{x \to 0^+} e^{1/x} f(x) = 0 \qquad \text{and} \qquad f(x) \leq \int_0^x \frac{f(t)}{t^2} dt \quad \text{for all } x \in [0, 1].$$

Solution by Eugene A. Herman, Grinnell College, Grinnell, IA.
We prove that the only function satisfying the given conditions is the zero function on $[0, 1]$. For all $x \geq 1$, let

$$g(x) = f\left(\frac{1}{x}\right), \qquad \text{and} \qquad G(x) = \int_x^\infty g(t)\, dt.$$

Then the inequality

$$-G'(x) = g(x) = f\left(\frac{1}{x}\right) \leq \int_0^{1/x} \frac{f(t)}{t^2}\, dt = \int_x^\infty f\left(\frac{1}{u}\right) du = G(x)$$

also holds for all $x \geq 1$, and so

$$\frac{d}{dx}(e^x G(x)) = e^x (G(x) + G'(x)) \geq 0.$$

Thus, $e^x G(x)$ is nonnegative and nondecreasing on $[1, \infty)$. By hypothesis, we have $\lim_{x\to\infty} e^x g(x) = \lim_{x\to 0^+} e^{1/x} f(x) = 0$. Thus, given $\epsilon > 0$ arbitrary, there exists $M \geq 1$ such that $e^x g(x) \leq \epsilon$ for all $x \geq M$; therefore,

$$e^x G(x) \leq e^x \int_x^\infty \epsilon e^{-t}\, dt = \epsilon$$

also holds for all $x \geq M$. Thus, the nonnegative nondecreasing function $e^x G(x)$ satisfies $\lim_{x\to\infty} e^x G(x) = 0$, so G is identically zero. Hence, g is also identically zero on $[1, \infty)$, and so is f.

Also solved by Michel Bataille, Elton Bojaxhiu & Enkel Hysnelaj, Robert Calcaterra, Omran Kouba (Syria), Kee-Wai Lau (Hong Kong, China), Northwestern University Math Problem Solving Group, and the proposer. There was one incomplete or incorrect solution.

2045. *Proposed by Kenneth Levasseur and Nicholas Raymond (student), University of Massachusetts Lowell, Lowell, MA.*

Let n be a positive integer. For any base b (a positive integer greater than 1) consider the set $R_{n,b}$ consisting of all 2^n nonnegative integers r whose base-b expansion has (at most) n digits, each either 0 or 1. Given $n > 1$, for what bases b is $R_{n,b}$ a complete system of residues to the modulus 2^n?

Solution by Michael Reid, University of Central Florida, Orlando, FL.
We prove that $R_{n,b}$ is a complete set of residues modulo 2^n if and only if $b \equiv 2 \pmod 4$, i.e., if b is divisible by 2 but not by 4.

Assume first that $R_{n,b}$ is a complete set of residues (CSR) modulo 2^n. Consider the primitive 2^n-th root of unity $\zeta = \exp(2\pi i/2^n)$. Since $R_{n,b}$ is a CSR, we have

$$\begin{aligned}(1+\zeta)\left(1+\zeta^b\right)\left(1+\zeta^{b^2}\right)\cdots\left(1+\zeta^{b^{n-1}}\right)&\\ = \sum_{\varepsilon_0\in\{0,1\}}\sum_{\varepsilon_1\in\{0,1\}}\cdots\sum_{\varepsilon_{n-1}\in\{0,1\}} \zeta^{\varepsilon_0+\varepsilon_1 b+\varepsilon_2 b^2+\cdots+\varepsilon_{n-1}b^{n-1}} &= \sum_{r\in R_{n,b}} \zeta^r\\ = \sum_{t=0}^{2^n-1}\zeta^t = \frac{\zeta^{2^n}-1}{\zeta-1} &= 0.\end{aligned}$$

Thus, one of the factors in the product above must vanish, say $1+\zeta^{b^k}$ with $0 \leq k \leq n-1$; thus, $\exp(2\pi i b^k/2^n) = \zeta^{b^k} = -1 = \exp(\pi i)$, so $b^k \equiv 2^{n-1} \pmod{2^n}$. Since $n \geq 2$ by hypothesis, we must have $b^k \equiv 2^{n-1} \equiv 0 \pmod 2$; therefore, $k \geq 1$, b^k is even, and hence so is b. Since $R_{n,b}$ is a CSR, $r = 0$ is its only element in the class 0 (mod 2^n); thus, the element b^{n-1} of $R_{n,b}$ satisfies $b^{n-1} \not\equiv 0 \pmod{2^n}$, hence b is not divisible by 4 (for otherwise b^{n-1} would be divisible by $4^{n-1} = 2^{2n-2}$, thus *a fortiori* by 2^n since $n \geq 2$). We conclude that $b \equiv 2 \pmod 4$.

Conversely, suppose $b \equiv 2 \pmod 4$. Write $b = 2c$ (with c an odd integer). By uniqueness of base-b representations, $R_{n,b}$ has cardinality exactly 2^n. In order to show that $R_{n,b}$ is a CSR, it suffices to prove that $R_{n,b}$ does not represent any residue class modulo 2^n more than once, which we presently prove by induction for all $n \geq 1$. Clearly, $R_{1,b} = \{0, 1\}$ is a CSR modulo 2^1, so the assertion holds for $n = 1$ (this is true whether or not $b \equiv 2 \pmod 4$). Assume next that the statement holds for some $n \geq 1$, and let $r, s \in R_{n+1,b}$ be congruent modulo 2^{n+1}. By definition of $R_{n+1,b}$, the numbers r, s have expressions $r = \sum_{i=0}^n \delta_i b^i$, $s = \sum_{i=0}^n \varepsilon_i b^i$, with $\delta_i, \varepsilon_i \in \{0, 1\}$ for

$i = 0, 1, \ldots, n$. *A fortiori,* $r \equiv s \pmod{2^n}$; since $b^n = 2^n c^n \equiv 0 \pmod{2^n}$, we have $\sum_{i=0}^{n-1} \delta_i b^i \equiv r \equiv s \equiv \sum_{i=0}^{n-1} \varepsilon_i b^i \pmod{2^n}$. By the inductive hypothesis, the congruence above implies $\delta_i = \varepsilon_i$ for $0 \le i \le n-1$. The congruence $r \equiv s \pmod{2^{n+1}}$ now gives $2^n c^n \delta_n = b^n \delta_n \equiv b^n \varepsilon_n = 2^n c^n \varepsilon_n \pmod{2^{n+1}}$, so $c^n \delta_n \equiv c^n \varepsilon_n \pmod 2$. Since c is odd, cancellation gives $\delta_n \equiv \varepsilon_n \pmod 2$, and since $\delta_n, \varepsilon_n \in \{0, 1\}$, we obtain $\delta_n = \varepsilon_n$. This completes the inductive proof that $R_{n,b}$ represents no class modulo 2^n more than once for $n \ge 1$. Hence, $R_{n,b}$ is a CSR modulo 2^n for all $n > 1$ and $b \equiv 2 \pmod 4$.

Also solved by Robert Calcaterra, Joseph DiMuro, Dmitry Fleischman, Kyle Gatesman, Eugene A. Herman, Laura Queipo (student) & José H. Nieto (Venezuela), Nicholas C. Singer, and the proposer.

Answers

Solutions to the Quickies from page 152.

A1089. We construct such a function f. Let $f(0) = 0$.

- For x rational, $0 < |x| < \sqrt{2}$, let $f(x) = 2/x$.
- For x rational, $|x| > \sqrt{2}$, let $f(x) = -2/x$.
- For x irrational, $|x| < 1$, let $f(x) = 1/x$.
- For x irrational, $|x| > 1$, let $f(x) = -1/x$.

Note that f maps rational numbers to rational numbers, and irrational to irrational. We have $f(f(0)) = f(0) = 0 = -0$. For rational $x \neq 0$ we have $f(f(x)) = -2/(2/x) = -x$, while for x irrational, $f(f(x)) = -1/(1/x) = -x$. Clearly, $|f(x)| \to \infty$ as $x \to 0$. For $a \neq 0$, $|f(x)|$ has two different sequential limits as $x \to a$, namely $1/|a|$ and $2/|a|$. Thus, f is discontinuous everywhere.

A1090. Assume that $\{a_n\} = \mathbb{Z}_p - \{0\}$ for some $p \ge 3$. *A fortiori,* by the recursive definition of $\{a_n\}$, it is clear that 2 is a primitive element modulo p. In particular, $2^{p-1} \equiv 1$ and $2^{(p-1)/2} \equiv -1 \pmod p$. If $a_n = (p+1)/2$ holds for some n, then $a_{n+1} \equiv 2^{a_n} \equiv 2^{(p+1)/2} = 2 \cdot 2^{(p-1)/2} \equiv -2 \pmod p$, and hence $a_{n+1} = p - 2$. Conversely, if $a_{n+1} = p - 2$, since 2 is primitive modulo p, we must have $a_n \equiv (p+1)/2 \pmod{p-1}$, and so $a_n = (p+1)/2$. Similarly, one sees that $a_{m+1} = (p+1)/2$ if and only if $a_m = p - 2$. Since the sequence $\{a_n\}$ includes the terms $(p+1)/2$ and $p-2$ by hypothesis, it must be purely periodic with range $\{(p+1)/2, p-2\} = \{a_n\} = \mathbb{Z}_p - \{0\}$. Thus, this can only hold when $p = 3$. Reciprocally, it is immediate to verify that $p = 3$ does indeed satisfy the required conditions.

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.

Brams, Steven J., and Mehmet S. Ismail, Making the rules of sports fairer, *SIAM Review* 60 (1) (2018) 181–202.

_____, D. Marc Kilgour, and Walter Stromquist, Catch-Up: A rule that makes service sports more competitive, *American Mathematical Monthly* 125 (9) (November 2018) 771–796. http://as.nyu.edu/content/dam/nyu-as/faculty/documents/Competition%20in%20Service%20Sports.pdf.

Isaksen, Aaron, Mehmet Ismail, Steven J. Brams, and Andy Nealen, Catch-Up: A game in which the lead alternates, *Game and Puzzle Design Journal* 1 (2) (2015) 38–49. http://game.engineering.nyu.edu/wp-content/uploads/2015/10/catch-up-a-game-in-which-the-lead-alternates-2015.pdf. Online playable version at http://game.engineering.nyu.edu/projects/catch-up/.

Many sports have rules that are not "fair," in the sense that "they do not ensure that equally skilled competitors have the same probability of winning." In particular, rules for the order of play in tie-breaking in soccer, American professional football, most racquet sports (except tennis), and volleyball are unfair, in that they systematically advantage one of the two sides (e.g., in soccer, the side that kicks first in a shootout). Brams et al. urge adoption of a Catch-Up Rule: A contestant who is behind after a round is "advantaged" for the next round. They compare this rule to the Win-by-Two Rule and other rules. They show that the Catch-Up Rule is "fairer" (except in tennis), generally strategy-proof, and does not change the probability of winning in racquet sports or volleyball (but tends to increase the length of the game). Isaksen et al. offer a simple but interesting game played with integers, called Catch-Up, and analyze it in part.

Suzuki, Jeff, *Patently Mathematical: Picking Partners, Passwords, and Careers by the Numbers*, Johns Hopkins University Press, 2019; 283 pp, $34.95. ISBN 978-1-4214-2705-8.

A publicity editor must have picked the subtitle of this book, since it scarcely agrees with the author's announced theme: "Under what conditions should a device based on a mathematical algorithm be patentable?" All of the following have been the subject of patents based mainly on mathematics: indexing documents, search engine algorithms, distinguishing between images, ranking attractiveness of photographs of people, rating potential compatibility of couples, opening a browser window, evaluating security of a password, retaining customers, influencing people, fitting medical devices to patients, compressing data, determining insurance premiums, and encrypting data. A surprising amount of mathematics appears in the book: vectors, matrices, logs, linear combinations, graphs with edge weighting, entropy, likelihood ratios, simulated annealing, error-correcting codes, fractals, cellular automata, public-key encryption, modulo arithmetic, zero-knowledge proofs, elliptic curves, and projective coordinates. Author Suzuki slickly exposits the mathematics involved, expresses mixed emotions about the patents, and offers suggestions for patent policy. (The index does not do justice to the topics in the book; and the book should have included a list of, and page numbers for, the patents discussed.)

Math. Mag. **92** (2019) 159–160. doi:10.1080/0025570X.2019.1580087

Lin, Thomas (ed.), *The Prime Number Conspiracy: The Biggest Ideas in Math from* Quanta, MIT Press, 2018; xx+309 pp, $19.95(P). ISBN 978-0-262-53635-6.

Honner, Patrick, Unscrambling the hidden secrets of superpermutations, https://www.quantamagazine.org/unscrambling-the-hidden-secrets-of-superpermutations-20190116/.

Quanta Magazine is an online science magazine that vows not to "report on anything you might actually find useful," such as medical or technological breakthroughs. That leaves lots of room, of course, for mathematics. The magazine, available at www.quantamagazine.org, is sponsored by the Simons Foundation. This volume collects 37 articles from its first 5 years, out of the 324 so far on mathematical topics. The title refers to evidence that "primes seem to avoid being followed [immediately] by another prime with the same final digit," a feature reflected in the prime assembly-line mechanism depicted on the book's cover. But the volume is by no means limited to articles about primes. Instead, it features stories also about other contemporary mathematics, including discoveries about proofs, the nature of mathematical thinking, connections with computing, new findings about infinity, and the lives of mathematicians. Fewer than half a dozen equations appear. This is an exciting book, by first-rate expositors, of up-to-the-minute developments in mathematics. The more-recent article by Honner is a perfect example of what to expect. (*Quanta* has also published a companion volume, *Alice and Bob Meet the Wall of Fire: The Biggest Ideas in Science*.)

Cubitt, Toby S., David Pérez-García, and Michael Wolf, The un(solv)able problem, *Scientific American* 319 (4) (October 2018) 29–37.

The spectral gap problem in physics asks whether there are discrete "gaps" between energy levels in a material, and the answer has consequences for quantum phase transitions of the material. The Yang-Mills Millennial Prize Problem in mathematics deals with a similar question about a "mass gap." The authors were able to show that the spectral gap problem is undecidable in general, by encoding many copies of the same Turing machine into the quantum ground state of a material; the ground state energy rises if the Turing machines halt. So the proof rests on the undecidability of the halting problem. Curiously, the proof also uses aperiodic tiles. Reinterpreting back into the physical world: "[E]ven a perfect, complete description of the microscopic interactions between a material's particles is not always enough to deduce its macroscopic properties."

Savage, Neil, Always out of balance, *Communications of the Association for Computing Machinery* 61 (4) (April 2018) 12–14.

More negative results! John Nash showed that every finite non-cooperative game has a Nash equilibrium, a state in which no player can do better by changing strategy. Unfortunately, there is no easy (polynomial-time) way to find Nash equilibria in general, or even approximate them.

Pukelsheim, Friedrich, *Proportional Representation: Apportionment Methods and Their Applications*, 2nd ed., Springer, 2017; xxvii+342 pp, $79.99(P). ISBN 978-3-319-64706-7.

In 2018, Democrats received 54% of the votes for legislators in Wisconsin but won only 36 of 99 seats in its lower house. Part of the mismatch was due to severe gerrymandering and part to geographical concentration of voters of like mind (leading to "wasted votes"). Author Pukelsheim starts from rules for rounding numbers, describes various divisor and quota methods, investigates biases introduced, explores apportionment criteria, discusses practical implementations, and details the method of double proportionality (apportioning seats to districts proportionally to population figures and to parties proportionally to vote counts). Use of that method, successfully promoted by Pukelsheim for use in several parts of the world, would have ameliorated the unfairness in Wisconsin. New to the second edition of the book is a 20-page "Biographical Digest" about individuals who contributed to apportionment methods. Though the book illustrates the methods with concrete real-life examples, it is not for non-mathematicians; a shorter companion textbook (but without exercises), *Sitzzuteilungsmethoden: Ein Kompaktkurs...*, is less mathematically demanding. (Disclosure: Author Pukelsheim is a personal friend who sponsored me on several occasions as a guest professor at the University of Augsburg, and I offered him comments for the first edition of the book.)