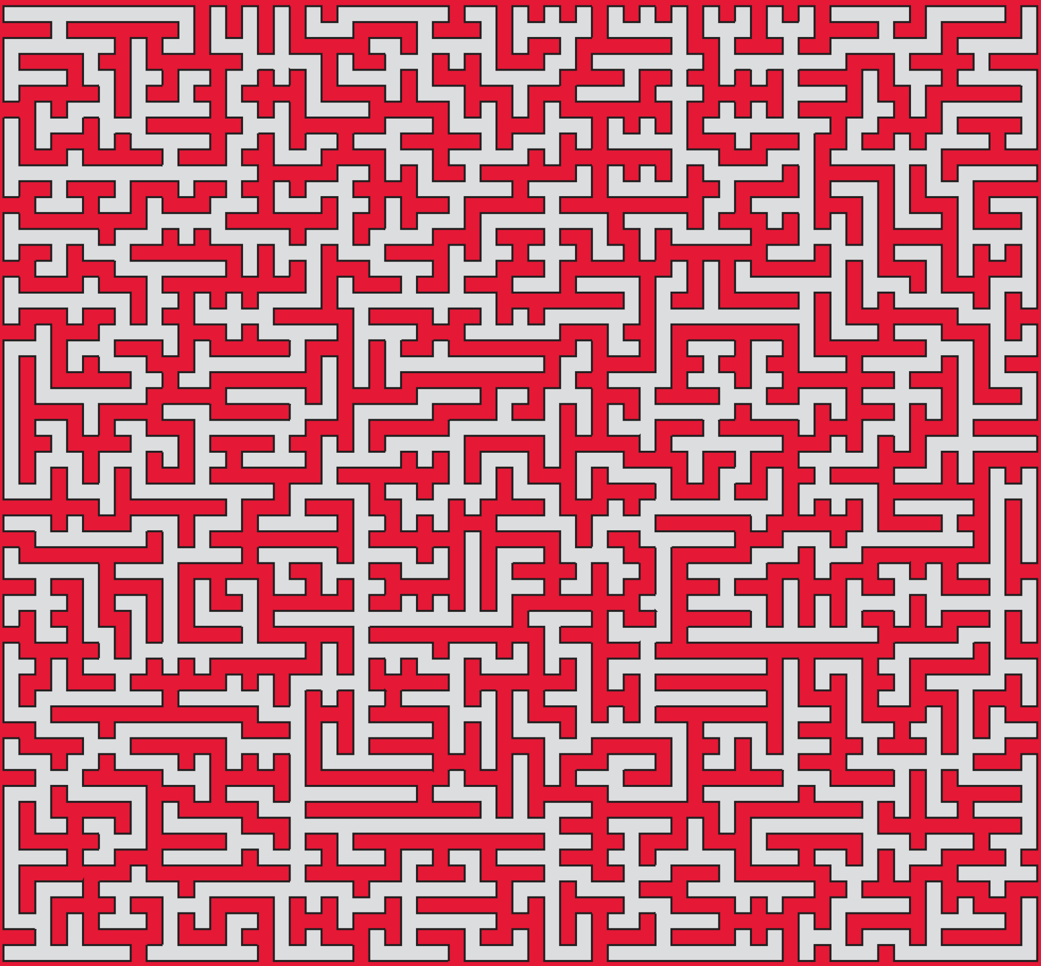
Vol. 85, No. 3, June 2012

MATHEMATICS MAGAZINE



This 60 × 65 rectangle has been divided into $60 \cdot 65 = 3900$ unit squares with their corners at the integer lattice points. Some of the edges of the squares have been removed, leaving a simple closed curve that passes through all of the lattice points. What is the area of the curve's interior?

A Timely Problem (see page 220)

- Why Ellipses Are Not Elliptic Curves
- Double Fun with Double Factorials
- Identities, Triangles, Subgroups, and Integrals

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

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 83, at pages 73–74, and is available at the *Magazine*'s website www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Please submit new manuscripts by email directly to the editor at mathmag@maa.org. A brief message containing contact information and with an attached PDF file is preferred. Word-processor and DVI files can also be considered. Alternatively, manuscripts may be mailed to Mathematics Magazine, 132 Bodine Rd., Berwyn, PA 19312-1027. If possible, please include an email address for further correspondence.

Cover image by Hugh Montgomery.

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Lancaster, PA, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

Subscription correspondence and notice of change of address should be sent to the Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to

MAA Advertising
1529 Eighteenth St. NW
Washington DC 20036

Phone: (877) 622-2373
E-mail: tmarmor@maa.org

Further advertising information can be found online at www.maa.org

Change of address, missing issue inquiries, and other subscription correspondence:

MAA Service Center, *maahq@maa.org*

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Periodicals postage paid at Washington, D.C. and additional mailing offices.

Postmaster: Send address changes to Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

Printed in the United States of America

Vol. 85, No. 3, June 2012

MATHEMATICS MAGAZINE

LETTER FROM THE EDITOR

In this issue we address some recurring themes.

Elliptic curves are everywhere these days. In our first article, Adrian Rice and Bud Brown trace the continuous half of their history, which is also the history of elliptic functions and elliptic integrals. They take us back to Euclid and Apollonius, who knew a lot about ellipses but not so much about their arc lengths, and who would have benefited from reading Eisenstein and Weierstrass. They conclude with a topological test, which draws a stark distinction between elliptic curves and the ellipses that inspired their study.

Combinatorial identities are a recurring theme. In this issue Henry Gould and Jocelyn Quaintance provide a festival of identities involving double factorials, some of them driven by analogies to ordinary factorials. In the Notes, Team Frumosu gives us an identity involving partitions (really, compositions) and reciprocals that magically sum to zero. And to everything we know about middle binomial coefficients, Christian Aebi and Grant Cairns add a remarkable congruence. Here is a fact: the numbers $4^6 + \binom{6}{3}$, $4^{10} + \binom{10}{5}$, and $4^{18} + \binom{18}{9}$ are all divisible by cubes. Why would that be?

Raymond Boute takes us back to the Brachistochrone problem (our theme from February, 2010) and treats it as geometry, echoing techniques from optics and acoustics. Jean Nganou tells us about subgroups of index 2, which are normally interesting, and finds many groups that have none of them.

If you want to use integration by parts—to say that $\int Fg = FG - \int fG$, for example—you need for F and G to be differentiable, right? Maybe not! According to Vicente Munoz, sometimes if f and g make a good-faith effort to be the derivatives of F and G, you can give them full credit.

Triangles are a recurring theme. Do you know the Steiner-Lehmus Theorem? It draws a conclusion from the lengths of two angle bisectors in a triangle. The theorem doesn't work for just any pair of cevians (cevian = segment connecting a vertex to a point of the opposite side). What makes the angle bisectors special? Not bisecting angles, says Victor Oxman; it is all about where the cevians intersect. By recognizing this, Oxman is able to generalize the Steiner-Lehmus Theorem to certain other pairs of intersecting cevians.

Clark Kimberling and Peter Moses started with triangles, but they branched out, and discovered a class of inequalities that transcend their original context.

Finally, in the Letters Section you'll see evidence that some themes have been recurring for longer than we might have guessed. It's another reason why we love our subject.

Walter Stromquist, Editor

Why Ellipses Are Not Elliptic Curves

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

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

After circles, ellipses are probably the most familiar curves in all of mathematics. Like circles, they are a special subclass of the so-called conic sections, or curves obtained by slicing a cone with a plane, and their applications are many and varied. For example, thanks to Johannes Kepler and his laws of planetary motion, astronomers know ellipses as the orbits of planets and many comets about the sun. In acoustics, architects have used the reflection property of ellipses—namely, that a light ray originating at one focus is reflected off the ellipse to the other focus—to construct whispering galleries in such places as St. Paul's Cathedral in London and Statuary Hall in the U. S. Capitol. Ellipses even find their way into modern medicine, where the reflection property is the basis for lithotripsy, a medical procedure for treating kidney stones and gall stones without invasive surgery. They also crop up from time to time in bad jokes mathematicians often like to tell: "What shape is a kiss?" "A lip tickle!"

In analytic geometry we learn that an ellipse is the set of all points in the plane the sum of whose distances from two fixed points is a given positive constant. Using this definition along with the distance formula, we may derive equations for ellipses which, in general, are of the form $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$, where $B^2 - 4AC < 0$. However, we may translate and rotate the axes as necessary to obtain the familiar equation of an ellipse centered at the origin with semimajor axis a and semiminor axis b, namely,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

which looks like FIGURE 1.

This beautiful object is certainly a curve, and its shape is evidently elliptical, so you would think that mathematicians would call it an "elliptic curve." But they do not. The name "elliptic curve" is reserved for a very different class of curves and, as with ellipses, we can define these in more than one way.

We can start with curves in the real plane $\mathbb{R}^2$, but much of the theory of elliptic curves depends on seeing them in $\mathbb{C}^2$. All in good time: for now, we give a simplified definition, namely that an elliptic curve is the set of solutions to an equation of the form $E(x, y) = 0$, where $E(x, y)$ is a cubic polynomial in x and y. We require further that

Math. Mag. **85** (2012) 163–176. doi:10.4169/math.mag.85.3.163. © Mathematical Association of America

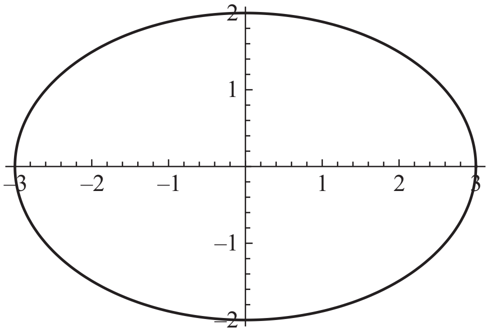


Figure 1 An ellipse with semimajor axis $a = 3$ and semiminor axis $b = 2$

$E(x, y)$ is *nonsingular*, which means that at no point do the function E and its partial derivatives E_x and E_y all vanish simultaneously. We may apply transformations, such as translations and other more complicated ones, to show that for our purposes, elliptic curves all have the form

$$y^2 = p(x),$$

where $p(x)$ is a cubic polynomial with no repeated roots. FIGURE 2 has two examples.

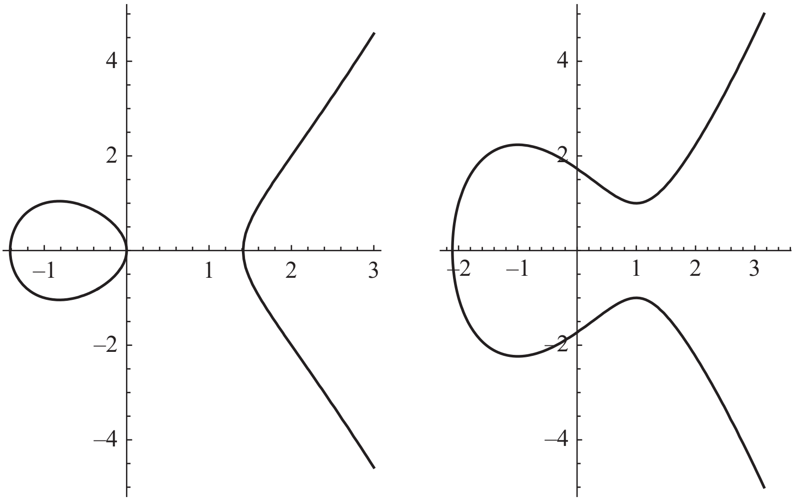


Figure 2 Two elliptic curves: $y^2 = x^3 - 4x$ (left) and $y^2 = x^3 - 3x + 3$ (right)

Elliptic curves come from algebraic geometry, and their applications show up in various systems of public key cryptography, in the factorization of large integers, in primality testing, and most famously in the proof of Fermat's Last Theorem. There are two principal facets to the study of elliptic curves, namely the discrete (which arises mainly from problems in number theory and abstract algebra) and the continuous (coming principally from the realms of calculus and complex analysis). The paper "Three Fermat trails to elliptic curves" [**5**] is a look at the history of the subject's discrete side via the congruent numbers problem, Fermat's Last Theorem, and the search for nontrivial integer solutions of $x^4 + ax^2y^2 + y^4 = z^2$, showing how each of these problems found solutions by recasting them as questions involving elliptic curves.

This paper has a very different focus and motivation. It deals with the history of the continuous side of the subject, from attempts to rectify the ellipse all the way up to the Weierstrass $\wp$-function. It thus serves as a companion piece to [**5**]. But why is such an article necessary? Well, when comparing FIGURES 1 and 2, many might be tempted to ask why it is that ellipses and elliptic curves look nothing like each other, yet have names that sound so similar. And they are quite right to wonder, because elliptic curves have almost nothing to do with ellipses at all. Why then are they called elliptic curves?

The answer lies in the word *almost*. There *is* a connection between ellipses and elliptic curves, but it's not at all obvious and is the result of a connected but distinctly nonlinear sequence of mathematical events. The simplest mathematical reason why ellipses are not elliptic curves is that their algebraic forms are fundamentally different: as we have seen, ellipses are quadratic, elliptic curves are cubic.

But this is not a particularly interesting answer. Nor does it explain how such different geometrical objects ended up with such similar-sounding names. To *really* answer the question properly, we need to look back at the history and development of these concepts. We will therefore take a stroll through the history of mathematics, encountering first the ellipse, moving on to elliptic integrals, then to elliptic functions, jumping back to elliptic curves, and eventually making the connection between elliptic functions and elliptic curves. We will then finally be in a position to find out why no elliptically-shaped planar curves may ever be called elliptic curves.

From ellipses to elliptic integrals

It all started, as many mathematical stories do, in ancient Greece and with one of the three classical construction problems, known as the Duplication Problem: given a cube with a certain volume V, construct a cube of volume $2V$ using only a compass and a straightedge. In modern notation, if a is the edge of the original cube, the goal is to construct a line segment of length $a\sqrt[3]{2}$. One early geometrical solution, ascribed to Hippocrates of Chios (ca. 460–380 BCE), involved determining two lengths x and y that satisfy the proportions

$$a : x = x : y = y : 2a.$$

Considering the three proportions separately and treating x and y as variables, the 4th century BCE mathematician Menaechmus showed that these proportions yield the curves $x^2 = ay$, $y^2 = 2ax$, and $xy = 2a^2$, which we recognize as equations of two parabolas and a hyperbola. For, if we multiply the first equation by x and the third equation by a, we are led to the equations $x^3 = axy = 2a^3$; hence, $x = a\sqrt[3]{2}$.

Menaechmus showed that any two of these equations imply the third, and that the lengths x and y are indeed the lengths required to produce the length $a\sqrt[3]{2}$. (We note that the Greek geometers developed several other constructions for finding $a\sqrt[3]{2}$, including Archytas of Tarentum's ingenious method involving the intersection of a cylinder, a torus with zero interior diameter, and a right circular cone. Pierre Wantzel (1814–1848) finally proved in 1837 that constructing $a\sqrt[3]{2}$ using only compass and straightedge is impossible—but that's another story.) Menaechmus went on (some say) to describe these as conic sections, discovering the ellipse in the process. Around 300 BCE Euclid wrote *Conics*, a major work, now known only through later commentaries, and containing a number of theorems on various properties of ellipses.

But it was Apollonius of Perga (ca. 262–190 BCE) who, in his eight-volume treatise *On Conics* [**10**], provided the most exhaustive study to date of the subject, as well as giving them the name ellipse, from the Greek *elleipsis*, meaning "falling short." So comprehensive was Apollonius's work that for nearly two millennia, it contained the majority of what was known on the subject. But there were gaps, and by the 17th century, mathematicians finally began to develop techniques that could fill them.

One question that Apollonius could not answer precisely was how to find the arc length of an ellipse. Geometric techniques were insufficient, as ellipses are curved shapes, but the invention of the integral calculus in the 1660s and 1670s provided a marvelous new tool for answering this question. With the introduction of this new technique, the question of finding the precise lengths of various curves, including the ellipse, became a major open problem for mathematicians. The arc length formula is one of the standard topics in courses on integral calculus: if $y = f(x)$ is continuous and has a continuous derivative on the interval $[a, b]$, then the length L_a^b of the curve is given by

$$L_a^b = \int_a^b \sqrt{1 + (f'(x))^2}\, dx.$$

But the first attempts from that era to find the arc length of an ellipse involved series, not integrals. For example, in 1669, Isaac Newton (1642–1727) expressed the arc length of an ellipse as an infinite series; other series-based expressions followed from the great Swiss genius Leonhard Euler (1707–1783) in 1733 and the Scottish mathematician Colin Maclaurin (1698–1746) in 1742. Why did they avoid integration?

To understand why, let's try using integration to find the arc length of an ellipse between, say, x_0 and x_1 and see what happens. Let a and b be positive numbers with $a > b$, and consider the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

We solve this equation for y and take the positive square root, which yields the function $y = f(x) = b\sqrt{a^2 - x^2}/a$. We calculate $\sqrt{1 + (f'(x))^2}$ and simplify the resulting messy expression by setting $k = \sqrt{a^2 - b^2}/a$; this transforms the resulting arc length integrand into $\sqrt{(a^2 - k^2x^2)/(a^2 - x^2)}$, and the arc length formula becomes

$$L_{x_0}^{x_1} = \int_{x_0}^{x_1} \sqrt{\frac{a^2 - k^2x^2}{a^2 - x^2}}\, dx.$$

Thus the total arc length, L, of the ellipse is given by

$$L = 4\int_0^a \frac{\sqrt{a^2 - k^2x^2}}{\sqrt{a^2 - x^2}}\, dx.$$

Unfortunately, this integral cannot be evaluated directly. The same is true if we use trigonometric functions to parameterize the curve as $x = a\sin t$, $y = b\cos t$, for $0 \leq t \leq 2\pi$, when the integral becomes

$$L = 4a\int_0^{\pi/2} \sqrt{1 - k^2\sin^2 t}\, dt.$$

This integral, in either the algebraic or the trigonometric form, is commonly known as an *elliptic integral.* To be more precise, this particular integral is called an *elliptic integral of the second kind.* The first part of this name arises because we are trying to determine the arc length of an *ellipse* via *integration.* But why "of the second kind"? And how many kinds are there?

The terminology and classification were introduced by the French mathematician Adrien-Marie Legendre (1752–1833). Legendre was fascinated by various interesting types of integrals which could not be computed by regular means—today we would call these "non-elementary integrals." These are integrals of functions $f(x)$ for which f does not have an antiderivative expressible in terms of elementary functions—polynomial, rational, algebraic, trigonometric, logarithmic or exponential. Starting in the 1750s, Euler had derived a great many results about these kinds of integrals, but it was Legendre who turned the subject into a systematic theory.

For 40 years from 1786, Legendre worked with many kinds of nonelementary integrals. He finally realized that the integrals arising from the above arc-length calculations could be expressed as one of three fundamental types, which we now define as *elliptic integrals of the first, second, and third kind,* respectively:

$$\begin{aligned} F(\phi) &= \int_0^{\phi} \frac{dt}{\sqrt{1-k^2\sin^2 t}}, \\ E(\phi) &= \int_0^{\phi} \sqrt{1-k^2\sin^2 t}\,dt, \quad \text{and} \\ \Pi(\phi) &= \int_0^{\phi} \frac{dt}{\left(1+n\sin^2 t\right)\sqrt{1-k^2\sin^2 t}}. \end{aligned}$$

Here, k, or the *modulus*, is a value in $[0, 1]$. Strictly speaking, the modulus k is a real constant such that k^2 is not equal to 0 or 1, but most texts on elliptic integrals restrict k such that $0 < k < 1$. The upper limit of the elliptic integrals, the amplitude ϕ, can be any real number, although it makes sense to focus on values in $[0, \pi/2]$. As for the word *amplitude*, this expression arose from Legendre's usage in referring to the physical applications which drove much of his work on elliptic integrals. Finally, in the elliptic integral of the third kind, n is taken to be a real constant, usually assumed to be nonzero because the case $n = 0$ reduces to the elliptic integral of the first kind.

Between 1825 and 1828, Legendre published a three-volume treatise [**16**] on these elliptic integrals (which, confusingly for us, he called elliptic *functions*), containing much of his four decades of work on the subject. How ironic, then, that just as Legendre was finishing his life's work, two young mathematicians were just beginning theirs with ideas that would render many of Legendre's techniques obsolete. Those two mathematicians were Niels Henrik Abel (1802–1829) and Carl Gustav Jacobi (1804–1851).

From elliptic integrals to elliptic functions

Both Abel and Jacobi wrote Legendre's elliptic integrals using the substitution $x = \sin t$, to give

$$\begin{aligned} F(u) &= \int_0^{u} \frac{dx}{\sqrt{\left(1-x^2\right)\left(1-k^2x^2\right)}}, \\ E(u) &= \int_0^{u} \frac{1-k^2x^2}{\sqrt{(1-x^2)(1-k^2x^2)}}\,dx, \quad \text{and} \\ \Pi(u) &= \int_0^{u} \frac{dx}{\left(1+nx^2\right)\sqrt{\left(1-x^2\right)\left(1-k^2x^2\right)}} \quad (\text{where } |u| \le 1), \end{aligned}$$

as the elliptic integrals of the first, second, and third kind, respectively. Indeed, to this day, elliptic integrals are still defined as those whose integrands are rational functions involving square roots of cubic or quartic polynomials. But it was Abel who realized that these integrals, although interesting and important, were not the most significant thing to be studying. (Gauss had realized this 30 years before, but did not publish his findings.) Consider the well-known integral

$$u = f(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}},$$

which, as we learn in calculus, is the inverse sine function. Abel argued that the function $f(x)$ defined by this integral was not as convenient to use as its inverse, $x = f^{-1}(u) = \sin u$. Likewise, he said, we should turn our attention from elliptic integrals to their inverses, which we now call *elliptic functions*.

Jacobi took this idea and ran with it [**12**]. He noticed that if $k = 0$ in the first kind of elliptic integral, we would simply get the inverse sine function. So for nonzero k, he defined the inverse of the first elliptic integral to be what he called the "sine amplitude" or sn u. Now, just as in regular trigonometry, where everything else can be built on the sine function, Jacobi went on to define further elliptic functions, such as the "cosine amplitude" function $\operatorname{cn} u = \sqrt{1 - \operatorname{sn}^2 u}$, and the "delta amplitude" function $\operatorname{dn} u = \sqrt{1 - k^2 \operatorname{sn}^2 u}$. He soon found that his new elliptic functions had many similar properties to the familiar trigonometric functions. For example, the regular sine function is periodic with period 2π, so that for any integer n, $\sin(x + 2\pi n) = \sin x$. But Jacobi's sine amplitude function was *doubly periodic;* in other words, there were two distinct numbers α and β (both complex, with $\alpha/\beta \notin \mathbb{R}$) such that

$$\operatorname{sn}(u + m\alpha) = \operatorname{sn}(u + n\beta) = \operatorname{sn}\, u.$$

In 1835, Jacobi proved that no single-valued function that is either analytic or *meromorphic* (that is, analytic except possibly at locations called *poles*, where a denominator vanishes to finite order) could ever have more than two independent periods. In fact, the only such functions to have two such periods were the elliptic functions. By 1847, a young German prodigy by the name of Ferdinand Gotthold Eisenstein (1823–1852) had taken the innovative step of starting with the periods to define the elliptic functions via infinite series (see [**7**], [**21**]). From there, he proved a startling result that made a connection between elliptic functions and a particular kind of cubic curve, to whose history we now turn.

The pre-history of elliptic curves

Having traced the study of the ellipse—particularly its arc length—to what we now call elliptic functions, let's back up and trace another story which originated with the Greeks and helps us understand elliptic curves.

In our introduction, you learned that an elliptic curve is a curve of the form $y^2 = p(x)$, where $p(x)$ is a cubic polynomial with no repeated roots. Although such cubic curves were not studied in detail until the late 1600s, two different problems from the apparently unrelated area of number theory, both going back many centuries, mark the origin of questions involving these curves. (In what follows, we will call these cubic curves *elliptic curves*, although they did not receive this name until the early twentieth century.)

The first problem comes from Diophantus of Alexandria's *Arithmetica* [**11**], written some time during the third or fourth century CE. Problem 24 of Book IV reads as follows: “To divide a given number into two numbers such that their product is a cube minus its side.” If we call Diophantus’ given number a, the task is to find X and Y such that

$$Y(a - Y) = X^3 - X.$$

Diophantus solved the problem for $a = 6$ by substituting $X = kY - 1$ and choosing the value $k = 3$; this causes the resulting polynomial in Y to have only a cubic and quadratic term. Ignoring the double root $Y = 0$, he obtained $Y = 26/27$ and $X = 17/9$. Therefore, the two numbers called for in the problem are $Y = 26/27$ and $a - Y = 136/27$ (since $26/27 + 136/27 = 6$), and the product of those two numbers is $(17/9)^3 - (17/9)$. (For additional details, see [**4**, pp. 34–35].)

We note that Diophantus’ curve $Y(a - Y) = X^3 - X$ is actually an elliptic curve in disguise, for the linear substitution $y = Y - a/2, x = -X$ leads to $y^2 = x^3 - x + (a/2)^2$. Now, it is important to stress that Diophantus had no concept of analytic geometry or modern algebraic notation, and certainly no idea about elliptic curves. Nevertheless, his work marked the beginning of a chain of inquiry that was to have wide-reaching and deep consequences many centuries later.

The second problem related to elliptic curves dates from certain Arabic manuscripts of roughly the eighth century, and Leonardo of Pisa, better known as Fibonacci (ca. 1175–1250), made it famous in Europe. He encountered the problem in question at the court of the Holy Roman Emperor Frederick II—namely, to find a rational number r such that both $r^2 - 5$ and $r^2 + 5$ are rational squares. Fibonacci found such a number, namely $r = 41/6$: sure enough, $r^2 - 5 = (31/6)^2$, $r^2 = (41/6)^2$ and $r^2 + 5 = (49/6)^2$ are indeed all squares. In his 1225 book *Liber quadratorum* (The Book of Squares) [**19**], Fibonacci called the positive integer n a *congruent number* if $u^2 - n$, u^2 and $u^2 + n$ are all nonzero squares for some rational number u.

The connection with elliptic curves lies in the fact that if n is a congruent number, then the *product* of the three nonzero rational squares $u^2 - n$, u^2 and $u^2 + n$ is also a nonzero rational square, say, v^2. In modern terminology, this implies that (u^2, v) is a point on the curve $E_n : y^2 = x^3 - n^2x = x(x - n)(x + n)$ with rational coordinates that are not both zero—a so-called nonzero *rational point*. Now for every positive integer n, $x(x - n)(x + n)$ is a cubic polynomial with distinct roots, which implies that E_n is an elliptic curve. Thus, if n is a congruent number, then the elliptic curve E_n contains a nonzero rational point. (For more information about congruent numbers, see the companion paper [**5**]; in addition, Koblitz uses congruent numbers as a unifying theme throughout his excellent book [**15**] on elliptic curves.)

Both of these ancient problems resurfaced in the early seventeenth century, when the French mathematician Claude-Gaspar Bachet de Meziriac (1581–1638) made a Latin translation of Diophantus’s *Arithmetica* and published it in 1621 [**3**]. This translation contained an appendix, which included Fibonacci’s congruent numbers problem, as well as some original results about Diophantine equations. One of the latter was the following theorem, which we give in modern notation. Fix an integer c and consider the equation $y^2 = x^3 + c$. If (x, y) is a solution to this equation with x and y both rational numbers, i.e., a *rational solution*, then

$$\left(\frac{x^4 - 8cx}{4y^2}, \frac{-x^6 - 20cx^3 + 8c^2}{8y^3}\right)$$

is also a rational solution.

The impact that Bachet’s translation of the *Arithmetica* had on the history of mathematics is a direct consequence of the fact that Pierre de Fermat (1601–1665) acquired a copy of it around 1630. Fermat’s chief mathematical contributions lie in his work in number theory: his introduction of the ideas of divisibility and primality gave the subject its definitive flavor, its most tantalizing question, and the direction it has taken for the better part of the last four centuries. Fermat's copy of Bachet's translation, which was reproduced and published by his son Samuel in 1670 [**9**], contained copious notes, annotations, and conjectures, including the famous Fermat Conjecture that if the integer n is greater than 2, then the equation $x^n + y^n = z^n$ has no integer solutions with $xyz \neq 0$—ultimately proved by Andrew Wiles and Richard Taylor in 1994.

Among Fermat's collected works we find several references to problems involving what we would now call elliptic curves, in particular, his conjecture that the only integers satisfying the equation $y^2 = x^3 - 2$ are $(x, y) = (3, \pm 5)$, and that the only integers satisfying $y^2 = x^3 - 4$ are $(x, y) = (2, \pm 2)$ and $(x, y) = (5, \pm 11)$.

Fermat's remarkable work might have gone unnoticed, except that during the 1730s, Leonhard Euler obtained a copy of Fermat's collected works. He was so struck by this body of mathematics that he proceeded to verify nearly all of Fermat's conjectures, including Fermat's statement about integer points on the curves $y^2 = x^3 - 2$ and $y^2 = x^3 - 4$. Euler expanded the scope of number theory far beyond Fermat's work, and his influence gave number theory its status as a legitimate field of mathematical inquiry [**8**]. Euler also did quite a bit of work on the congruent numbers problem and, as noted earlier, derived many results about elliptic integrals. The latter included formulas for adding these integrals, which provided a starting point for Legendre's work on the same subject in the 1780s.

In the meantime, during the 1670s, Newton used the recently developed tools of analytic geometry to try to classify cubic curves, in particular those of the form $y^2 = ax^3 + bx^2 + cx + d$ [**17**]. In doing so, he explained the mysteries behind both Diophantus' *Arithmetica* problem and Bachet's theorem about rational solutions to $y^2 = x^3 + c$. He pointed out that both Diophantus and Bachet were essentially intersecting a line with a cubic curve, and that in general, such an intersection consists of three points. However, if the line is tangent to the curve, then two of those three points are the same. FIGURE 3 tells the story.

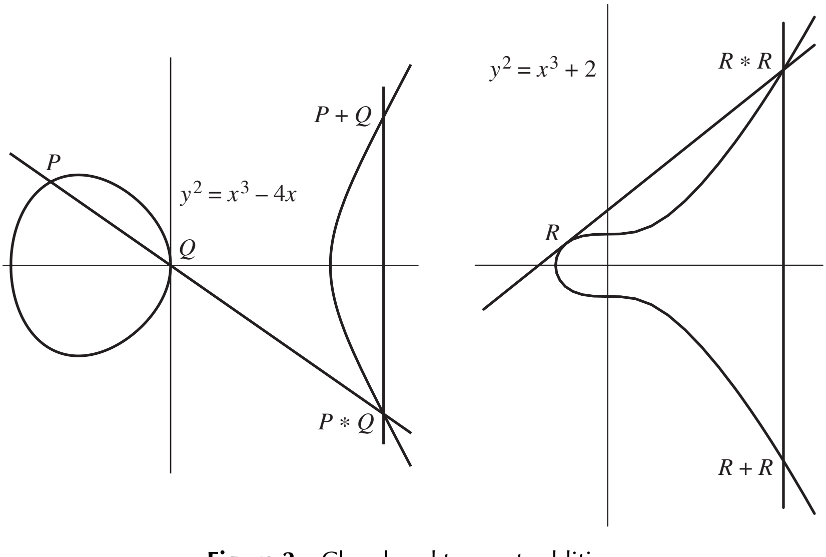


Figure 3 Chord and tangent addition

For the curve on the left, the line through P and Q intersects the curve in the third point $P * Q$; for the curve on the right, the tangent to the curve at R intersects the curve in the "third point" $R * R$. Using this star $*$ operation we can define an addition of the points on our curve. Namely, we define the sum $P + Q$ and $R + R$ to be the reflections of $P * Q$ and $R * R$, respectively, in the x-axis. This so-called chord and tangent addition gives the elliptic curve a group structure, which future researchers would find extremely useful—but that's another story, for which see [**5**].

Newton's insight ultimately led to general formulas for the addition of points on curves of the form $y^2 = ax^3 + bx^2 + cx + d$ (see [**14**, p. 10], for details), but it would first require the discovery of an amazing connection between such curves and elliptic functions. And it is precisely that connection which brings us back to the groundbreaking work of Gotthold Eisenstein in 1847.

From elliptic functions to elliptic curves

In order to appreciate Eisenstein's work, let's begin with an infinite series. Now it is probably not obvious, but it is true that

$$\sum_{m=-\infty}^{\infty} (z + m\pi)^{-2} = (\sin z)^{-2}.$$

(To begin to informally convince yourself of this, note that replacing z by $z + 2\pi$ on both sides of the equation leaves the relationship unchanged. For a formal proof, see [**1**, p. 11].) And since all trigonometry is ultimately based on the sine function, the whole subject could in theory be founded just as well on the above infinite series. Using this as his inspiration, Eisenstein constructed a new function out of a *doubly* infinite series, namely

$$\sum_{m,n=-\infty}^{\infty} (z + m\omega_1 + n\omega_2)^{-2}$$

where $\omega_1, \omega_2 \in \mathbb{C}$, and $\omega_1/\omega_2 \notin \mathbb{R}$. A bit of algebra reveals that this convergent series has two distinct periods, ω_1 and ω_2.

As previously noted, Jacobi had proved that the only single-valued meromorphic (i.e., analytic everywhere except for poles) functions with two linearly independent periods are the elliptic functions. Indeed, the modern definition of an elliptic function is a single-valued, meromorphic function f, defined on $\mathbb{C}$, for which there exist two distinct complex numbers ω_1 and ω_2 such that ω_1/ω_2 is not a real number and $f(z + \omega_1) = f(z + \omega_2) = f(z)$. And since Eisenstein's series-based function is defined over $\mathbb{C}$, is single-valued, meromorphic and doubly-periodic, it therefore has to be an elliptic function.

It was then that Eisenstein came up with a massively important result (see [**21**, pp. 22–24]). He proved that all elliptic functions of the form

$$y(z) = \sum_{m,n=-\infty}^{\infty} (z + m\omega_1 + n\omega_2)^{-2} - \sum_{\substack{m,n=-\infty \\ (m,n)\neq(0,0)}}^{\infty} (m\omega_1 + n\omega_2)^{-2}$$

must satisfy differential equations of the form

$$\left[y'(z)\right]^2 = p(y(z)),$$

where p is a cubic polynomial (depending on ω_1 and ω_2) with no repeated roots.

Does the phrase "a cubic polynomial with no repeated roots" ring a bell? If it does not, go back and re-read our introductory section: we'll wait for you.

$*\quad *\quad *$

That's right—we defined an elliptic curve to be a curve of the form $y^2 = p(x)$, where $p(x)$ is a cubic polynomial with no repeated roots. We thus see that Eisenstein's work connects elliptic functions with elliptic curves.

A decade and a half later, in 1863, the famous analyst Karl Weierstrass (1815–1897) used this to define perhaps the most famous elliptic function of all, the Weierstrass $\wp$-function [**20**]:

$$\wp(z) = \wp(z;\omega_1,\omega_2) = z^{-2} + \sum_{\substack{m,n=-\infty \\ (m,n)\neq(0,0)}}^{\infty} \left[(z - m\omega_1 - n\omega_2)^{-2} - (m\omega_1 + n\omega_2)^{-2}\right].$$

As one would expect of an elliptic function, the $\wp$-function is doubly periodic with periods ω_1 and ω_2. But so are its derivative $\wp'(z)$, and its second derivative $\wp''(z)$, and so on. In fact all of its derivatives are elliptic functions with periods ω_1 and ω_2. Furthermore—and this is the amazing bit—*every* single elliptic function with periods ω_1 and ω_2 can be written as a rational function of $\wp(z)$ and $\wp'(z)$. (For a proof of this standard result, see [**2**, p. 189].) In other words, just as the sine function is the basis for all other trigonometric functions, so is the Weierstrass $\wp$-function the basis of all other elliptic functions.

By means of a clever argument using series, Weierstrass was able to show that the differential equation that his function satisfied was indeed a cubic, just as Eisenstein had proved, namely

$$\left[\wp'(z)\right]^2 = 4\wp^3(z) - g_2\wp(z) - g_3,$$

where g_2 and g_3 are special constants depending only on ω_1 and ω_2. It is therefore not hard to see that the point $(\wp(z), \wp'(z))$ lies on the cubic curve

$$y^2 = 4x^3 - g_2x - g_3.$$

Now in calculus, you learn about parametric equations and how they can describe a curve. By a *parameterization* of a curve C, we mean a continuous bijection from a set of numbers to the set of all points on C. For example, letting $x = a\sin t$ and $y = b\cos t$ with $t \in [0, 2\pi)$ gives a familiar parameterization of the standard ellipse $x^2/a^2 + y^2/b^2 = 1$. In the same way, we see that setting $x = \wp(z)$ and $y = \wp'(z)$ gives a parameterization of the cubic curve $y^2 = 4x^3 - g_2x - g_3$.

(You may wonder about the domain of the set of complex numbers z needed for the parameterization of the cubic. We'll get to that later. We also note that in order to rigorously prove parameterization, one must show the existence of a continuous, bijective map. Those interested in the details should consult [**15**, pp. 22–26].)

Now, cubic curves of the form $y^2 = ax^3 + bx^2 + cx + d$ had been well known for years. Indeed, as we mentioned earlier, Isaac Newton carried out a major study of them in the 1670s. But it wasn't until 1834 that Jacobi pointed out a possible connection between cubic curves and elliptic functions [**13**], followed by Eisenstein's proof of such a relationship in 1847. Then in 1864, a German mathematician by the name of Alfred Clebsch (1833–1872) introduced the idea above of using elliptic functions to parameterize cubic curves [**6**], and Weierstrass linked a clever addition formula for elliptic functions to the addition of points on these cubic curves. Finally, in a landmark paper of 1901 [**18**], Henri Poincaré (1854–1912) tied all these ideas together, effectively marking the birth of a new area of study. And because they require elliptic functions for their parameterization, these curves became known as elliptic curves.

Why ellipses are not elliptic curves

We have thus seen the historical path that led from the ellipse, first of all to elliptic integrals (one of which expresses the arc length of an ellipse), then to elliptic functions (obtained by inverting an elliptic integral), and finally to elliptic curves (which require elliptic functions for their parameterization). All of this leads to the question we posed at the beginning: why are ellipses not elliptic curves? The answer lies firstly in extending the domain of both curves from the reals to the complex numbers, and secondly in the matter of their respective parameterizations. We have already mentioned that ellipses may be parameterized by trigonometric functions, and this holds as much for ellipses in $\mathbb{C}^2$ as it does for those in $\mathbb{R}^2$. But in $\mathbb{C}^2$ such curves are best described, not as curves at all, but as *surfaces*. In particular, curves parameterizable by singly-periodic complex-valued functions are topologically equivalent to spheres. Here is one way to see this.

As functions in $\mathbb{R}^2$, the single periodicity of the sine and its derivative, cosine, means that their domain is $\mathbb{R}$ mod $2\pi\mathbb{Z}$. Now on the real line, $2\pi\mathbb{Z}$ is a one-dimensional lattice, so geometrically, $\mathbb{R}$ mod $2\pi\mathbb{Z}$ is a circle. If we change the domain to $\mathbb{C}$ and map to $\mathbb{C}^* \times \mathbb{C}^*$, where $\mathbb{C}^* = \mathbb{C} \cup \{\infty\}$, here's what happens. Since the ellipse is parameterized by $\sin z$ and $\cos z$, its (complex) domain is divided into infinitely long vertical strips with real width 2π, as in the following figure, where on the left we see the plane with the vertical lines $\text{Re}(z) = \pi$ and $\text{Re}(z) = -\pi$ drawn. The periodicity of sine and cosine means that every distinct point in the complex plane corresponds to a distinct point in this strip; thus the ellipse, since it is parameterized by these two functions, is completely described by how they map the points in this strip. Since we are mapping to a compact set, we can identify all points in the strip such that $\text{Im}(z) = \pm\infty$ with a single point, ∞. If we further identify the two lines $\text{Re}(z) = \pi$ and $\text{Re}(z) = -\pi$, this transforms the strip into a surface with ∞ represented by a point at the top, the origin by a point at the bottom, and a meridian line, corresponding to the identified vertical lines, joining the two points along the surface. As we see on the right, this resulting geometric figure is topologically equivalent to a sphere.

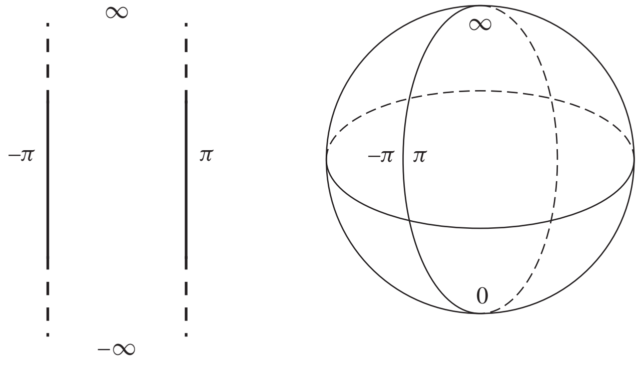


Figure 4 The complex co-domain of an ellipse is a sphere

However, just as the elliptic integral representing the arc length of an ellipse cannot be evaluated using regular calculus techniques, elliptic curves cannot be parameterized by elementary functions. The simplest functions that will successfully parameterize elliptic curves are elliptic functions, and it is this parameterization that is the key to understanding why ellipses are not elliptic curves.

Suppose we have an elliptic curve

$$y^2 = ax^3 + bx^2 + cx + d.$$

Then the ordered pairs (x, y) that work in this equation can be written as $(f(z), f'(z))$, where $f(z)$ is an elliptic function with periods ω_1 and ω_2. Since $f(z)$ is elliptic, it is a rational function of $\wp(z)$ and $\wp'(z)$, as is its derivative $f'(z)$, which is also an elliptic function with the same periods. Not only that, but it can be shown that the correspondence $(x, y) \leftrightarrow (f(z), f'(z))$ between all points (x, y) on the elliptic curve and complex numbers $z \in \mathbb{C}/\Lambda$, where

$$\Lambda = \{m\omega_1 + n\omega_2 : m, n \in \mathbb{Z}\},$$

is one-to-one and onto; see, for example, [**15**, pp. 22–26]. In other words, the functions $x = f(z)$ and $y = f'(z)$ parameterize the elliptic curve, and the fact that f and f' have periods ω_1 and ω_2 means that there is a one-to-one correspondence between all points on the curve and the equivalence classes

$$z + \Lambda = \{z + m\omega_1 + n\omega_2 : m, n \in \mathbb{Z}\}.$$

The double-periodicity of f and f' thus partitions the complex plane into a lattice that looks something like FIGURE 5.

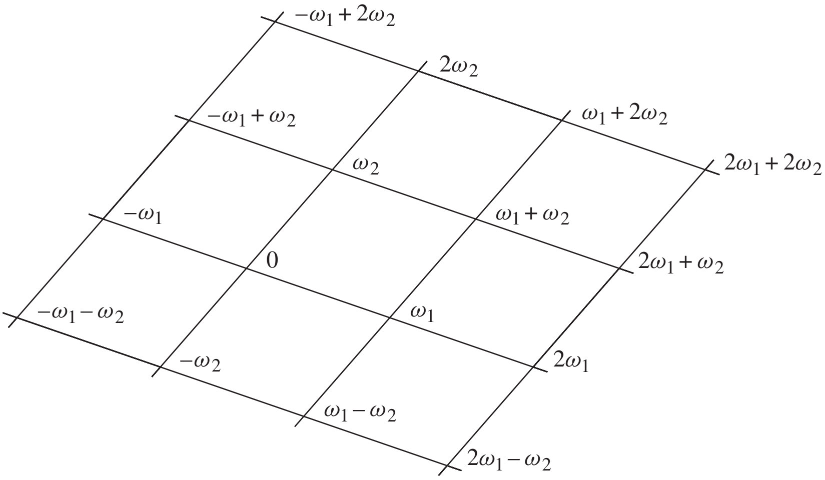


Figure 5 The lattice Λ generated by ω_1 and ω_2

Elliptic curves, which require elliptic functions for their parameterization, are thus isomorphic to the space $\mathbb{C}/\Lambda$ of these equivalence classes. On the other hand, as we have seen, ellipses, since they can be parameterized by elementary trigonometric functions, are isomorphic to the sphere. But $\mathbb{C}/\Lambda$ is not topologically equivalent to a sphere at all. Instead, it is a torus! To see this, look at FIGURE 6. The top left shows one of the infinitely many parallelograms which make up the lattice in FIGURE 5. (Since the parallelograms are all congruent to each other, it doesn't really matter which one we choose.) The top right shows this parallelogram rolled up into a cylinder, so that edge C meets edge A. Imagine that this resulting cylinder is made of material flexible enough

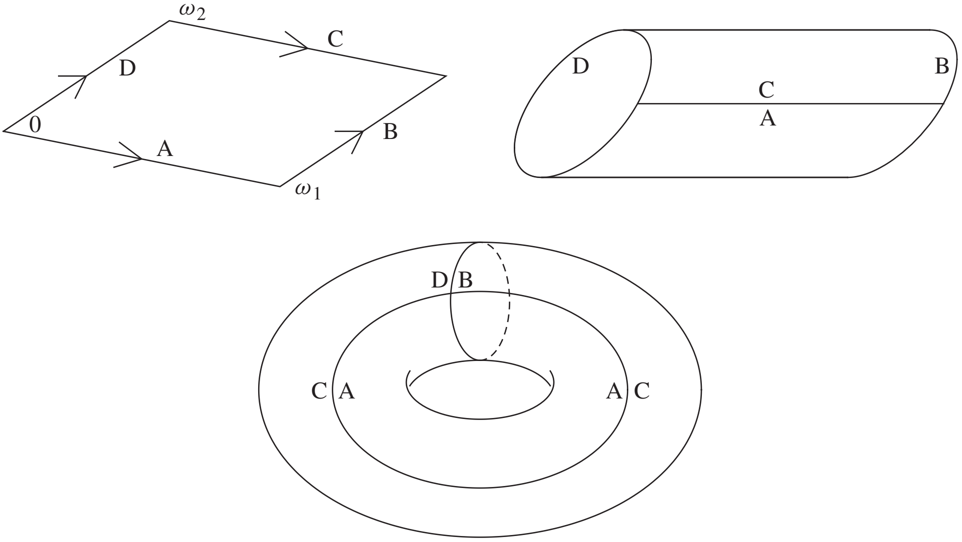


Figure 6 $\mathbb{C}/\Lambda$ is a torus

to be bent round so that edge B and edge D can be joined together, as in the bottom figure. It is clear that the resulting surface formed by this rolled-up parallelogram is a torus. Thus, since every point on an elliptic curve can be mapped to one of these "periodic parallelograms," which in turn can be transformed into tori, every elliptic curve is topologically equivalent to a torus.

Now, a torus is a surface that is completely incapable of being (legitimately) transformed into a sphere, meaning that no curve isomorphic to a sphere could possibly belong to a set of objects isomorphic to $\mathbb{C}/\Lambda$. This gives a visually very obvious—and mathematically, very profound—reason why elliptic curves are not parameterizable by any elementary functions.

It also tells us why ellipses are not (and never could be) elliptic curves!

REFERENCES

1. G. E. Andrews, R. Askey, and R. Roy, *Special Functions*, Cambridge University Press, 2000.
2. J. V. Armitage and W. F. Eberlein, *Elliptic Functions*, Cambridge University Press, 2006.
3. C.-G. Bachet, *Diophanti Alexandrini Arithmeticorum*, Sebastiani Cramoisy, Paris, 1621.
4. I. G. Bashmakova, *Diophantus and Diophantine Equations*, Mathematical Association of America, 1997.
5. E. Brown, Three Fermat trails to elliptic curves, *College Math. J.* **31** (May, 2000) 162–172. http://dx.doi.org/10.2307/2687483
6. A. Clebsch, Über einen Satz von Steiner und einige Punkte der Theorie der Curven dritter Ordnung, *J. für die reine und angewandte Mathematik* **63** (1864) 94–121. http://dx.doi.org/10.1515/crll.1864.63.94
7. F. G. Eisenstein, Beiträge zur Theorie der elliptischen Funktionen, *J. für die reine und angewandte Mathematik* **35** (1847) 137–274. http://dx.doi.org/10.1515/crll.1847.35.137
8. L. Euler, *Leonhardi Euleri Opera Omnia*, Ser. 1, vol. 2–5, B. G. Teubner, Leipzig and Berlin, 1911–13.
9. P. Fermat, *Oeuvres*, vol. 1. Gauthier-Villars, Paris, 1891–1896.
10. M. N. Fried and S. Unguru, *Apollonius of Perga's Conica: Text, Context, Subtext*, Brill, Leiden, 2001.
11. T. L. Heath, *Diophantus of Alexandria*, Cambridge University Press, 1910.
12. C. G. J. Jacobi, *Fundamenta nova functionarum ellipticarum*, Borntraeger, 1829.
13. ———, De usu theoriae integralium ellipticorum et integralium abelianorum in analysi diophantea, *Werke* **2** (1834) 53–55.
14. A. W. Knapp, *Elliptic Curves*, Princeton University Press, 1992.
15. N. Koblitz, *Introduction to Elliptic Curves and Modular Forms,* Springer-Verlag, 1984.
16. A.-M. Legendre, *Traité des fonctions elliptiques*, 3 vols., Huzard-Courcier, Paris, 1825–28.

17. I. Newton, De resolutione quaestionum circa numeros, *Mathematical Papers* **4** (1670s) 110–115.
18. H. Poincaré, Sur les propriétés arithmétiques des courbes algébriques, *Journal de mathématiques pures et appliquées* (series 5) **7** (1901) 161–233.
19. L. E. Sigler (transl.), *Fibonacci's Liber Quadratorum*, Academic Press, 1987.
20. K. Weierstrass, Vorlesungen über die Theorie der elliptischen Funktionen, *Mathematische Werke*, vol. 5, 1863.
21. A. Weil, *Elliptic Functions According to Eisenstein and Kronecker*, Springer-Verlag, 1976.

Summary Elliptic curves are a fascinating area of algebraic geometry with important connections to number theory, topology, and complex analysis. As their current ubiquity in mathematics suggests, elliptic curves have a long and fascinating history stretching back many centuries. This paper presents a survey of key points in their development, via elliptic integrals and functions, closing with an explanation of why no elliptically-shaped planar curved line may ever be called an elliptic curve.

ADRIAN RICE is Professor of Mathematics at Randolph-Macon College in Ashland, Virginia. His research specialty is the history of mathematics, focusing on nineteenth- and early twentieth-century mathematics in particular. His most recent book, *Mathematics in Victorian Britain*, co-edited with Raymond Flood and Robin Wilson, was published by Oxford University Press in 2011. In his spare time, he enjoys music, travel, and spending time with his wife and three-year-old son.

EZRA (BUD) BROWN grew up in New Orleans and has degrees from Rice and LSU. Since 1969, he has been at Virginia Tech in Blacksburg, Virginia, where he is currently Alumni Distinguished Professor of Mathematics. He does research in number theory and combinatorics, and his book, *Biscuits of Number Theory*, co-edited with Art Benjamin, was published by the MAA in 2009. He plays piano jazz, has been in six operas, goes kayaking with his wife, and occasionally bakes biscuits for his students.

Double Fun with Double Factorials

HENRY GOULD
JOCELYN QUAINTANCE
West Virginia University
Morgantown, WV 26506-6310
gould@math.wvu.edu
quaintan@math.rutgers.edu

All mathematicians know of the factorial, defined by $n! = n(n-1)(n-2)\cdots(2)(1)$ for integers $n \geq 1$ and $0! = 1$. In this paper we consider the double factorial: If n is a positive integer we define

$$n!! = (n)(n-2)(n-4)\cdots(4)(2)$$

if n is even, and

$$n!! = (n)(n-2)(n-4)\cdots(3)(1)$$

if n is odd. We define $0!! = 1$; also, it will be convenient to define $(-1)!! = 1$. Some examples:

n	0	1	2	3	4	5	6	7	8
$n!!$	1	1	2	3	8	15	48	105	384

Note that $n!!$ is not the same as the iterated factorial $(n!)!$, which grows much faster.

We do not know precisely when, where, or by whom, the double factorial notation was devised. It was used by Meserve [7] in 1948, but it is not mentioned by Cajori in his very detailed 1928–1929 history of mathematical notations [2]. Thus, we surmise that the notation was introduced at some time between 1928 and 1948.

Double factorials can be written in terms of ordinary factorials. If $n = 2m$ is even, then

$$\begin{aligned}(2m)!! &= (2m)(2m-2)(2m-4)\cdots(4)(2)\\ &= 2(m)2(m-1)\cdots 2(2)2(1) = 2^m m!,\end{aligned} \tag{1}$$

and if $n = 2m-1$ is odd, then

$$\begin{aligned}(2m-1)!! &= (2m-1)(2m-3)(2m-5)\cdots(3)(1)\left[\frac{(2m)(2m-2)\cdots(4)(2)}{(2m)(2m-2)\cdots(4)(2)}\right]\\ &= \frac{(2m)!}{2^m m!}.\end{aligned}$$

Double factorials can also be defined recursively. Just as we can define the ordinary factorial by $n! = n\cdot(n-1)!$ for $n \geq 1$ with $0! = 1$, we can define the double factorial by

$$n!! = n\cdot(n-2)!!$$

Math. Mag. **85** (2012) 177–192. doi:10.4169/math.mag.85.3.177. © Mathematical Association of America

for $n \geq 2$ with initial values $0!! = 1!! = 1$. With our convention that $(-1)!! = 1$, the recursion is valid for all positive integers n.

Are double factorials useful? One way in which these notations arose historically was as a simple way to present the integral formulas

$$\int_0^{\pi/2} \sin^{2n} x\,dx = \frac{(2n-1)!!}{(2n)!!}\frac{\pi}{2}$$

and

$$\int_0^{\pi/2} \sin^{2n-1} x\,dx = \frac{(2n-2)!!}{(2n-1)!!}.$$

Some handbooks of integrals, such as the CRC Handbook [**6**] and others by Burington, Dwight, etc., write these formulas as

$$\int_0^{\pi/2} \sin^n x\,dx = \int_0^{\pi/2} \cos^n x\,dx = \begin{cases} \frac{(1)(3)(5)\cdots(n-1)}{(2)(4)(6)\cdots(n)}\frac{\pi}{2} = \frac{(n-1)!!}{(n)!!}\frac{\pi}{2}, & n \text{ even}, n \geq 2 \\ \frac{(2)(4)(6)\cdots(n-1)}{(1)(3)(5)\cdots(n)} = \frac{(n-1)!!}{(n)!!}, & n \text{ odd}, n \geq 3 \end{cases}$$

In either case

$$\int_0^{\pi/2} \sin^n x\,dx = \int_0^{\pi/2} \cos^n x\,dx = \frac{1}{2}\sqrt{\pi}\,\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2}+1)}, \qquad n \geq -1,$$

where $\Gamma(z)$ is defined as in [**11**, p. 15].

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt, \qquad \mathrm{Re}(z) > 0.$$

These integral formulas are due to John Wallis (1616–1703) who in 1655 [**13**] discovered the remarkable infinite product formula

$$\frac{\pi}{2} = \frac{2\cdot 2\cdot 4\cdot 4\cdot 6\cdot 6\cdots}{1\cdot 3\cdot 3\cdot 5\cdot 5\cdot 7\cdots} = \lim_{n\to\infty} \frac{((2n)!!)^2}{(2n-1)!!\,(2n+1)!!}.$$

Double factorials have arisen also in enumerative combinatorics. A *two-colored permutation* of the set $[n] = \{1, 2, \dots, n\}$ is an ordinary permutation of $[n]$ with each entry colored, say, red or blue. If $n = 2$, for example, there are eight such permutations, namely (1_r2_r), (1_r2_b), (1_b2_r), (1_b2_b), (2_r1_r), (2_r1_b), (2_b1_r), (2_b1_b). In general the number of these permutations is $n!$ (for the choice of permutation) times 2^n (for the color choices). From equation (1) we recognize this number to be $(2n)!!$.

A *matching* in the set $[2n] = \{1, 2, \dots, 2n\}$ is a partition of $[2n]$ into n unordered pairs. For example, if $n = 4$ there are three matchings; and it turns out that in general that the number of matchings in $[2n]$ is $(2n-1)!!$.

More enumerative applications can be found at the OEIS website [**9**]. For example, in a note at that site, David Callan reports that $(2n)!!$ counts the permutations of the multiset $\{1, 1, 2, 2, \dots, n, n, n+1, n+1\}$ with the property that, for each $i = 1, \dots, n$, between the two occurrences of i there is exactly one entry larger than i.

Higher-order factorials. Factorials can be further generalized to multi-factorials. If m is a positive integer, we define the multi-factorial $(n)!_m$, or m-factorial of n, as

$$(n)!_m = (n)(n-m)(n-2m)\cdots(*)$$

where the last term $(*)$ can have any of the values $1, \dots, m$. We can get the same result recursively by defining

$$(n)!_m = n \cdot (n-m)!_m$$

and taking as initial values

$$(n)!_m = 1 \quad \text{when } n = 0, -1, -2, \dots, -(m-1).$$

It is common to write $(n)!_3$ as $n!!!$ and call it the triple factorial of n, and similarly $(n)!_4$ (quadruple factorial) may be written as $n!!!!$, etc. TABLE 1 gives the values of $(n)!_m$ for $m = 1, 2, 3, 4, 5$ and $n = 1$ to 10.

TABLE 1: The Sequences $n!$, $(n)!!$, $(n)!!!$, $(n)!!!!$, and $(n)!!!!!$ for $n = 1, 2, 3, \dots, 10$

n	$n!$	$(n)!!$	$(n)!!!$	$(n)!!!!$	$(n)!!!!!$
1	1	1	1	1	1
2	2	2	2	2	2
3	6	3	3	3	3
4	24	8	4	4	4
5	120	15	10	5	5
6	720	48	18	12	6
7	5040	105	28	21	14
8	40320	384	80	32	24
9	362880	945	162	45	36
10	3628800	3840	280	120	50

In this paper we focus on the role of double factorials in combinatorial identities. Our purpose is to illustrate how double factorials add new insight to this field. We will also derive some amusing identities involving double factorials and what we call double factorial binomial coefficients (DFBC's), defined by

$$\left(\!\!\binom{n}{k}\!\!\right) = \frac{(n)!!}{(k)!!\,(n-k)!!}.$$

Two basic identities involving double factorials

We start with two well known identities involving factorials, and in this section we generalize them using double factorials and multi-factorials.

The first identity is

$$\sum_{k=0}^{n} (k)\,k! = (n+1)! - 1. \tag{2}$$

This identity has a "one-line proof," although we need to typeset it in two lines:

$$\begin{aligned}\sum_{k=0}^{n}(k)\,k! &= \sum_{k=0}^{n}(k+1)\,k! - \sum_{k=0}^{n}k! = \sum_{k=0}^{n}(k+1)! - \sum_{k=0}^{n}k!\\ &= \sum_{k=1}^{n+1}k! - \sum_{k=0}^{n}k! = (n+1)! - 0!.\end{aligned}$$

Essentially it is a telescoping argument. The proof adapts easily to the case of double factorials:

$$\begin{aligned}\sum_{k=0}^{n}(k+1)\,k!! &= \sum_{k=0}^{n}(k+2)\,k!! - \sum_{k=0}^{n}k!! = \sum_{k=0}^{n}(k+2)!! - \sum_{k=0}^{n}k!!\\ &= \sum_{k=2}^{n+2}k!! - \sum_{k=0}^{n}k!! = (n+1)!! + (n+2)!! - 0!! - 1!!,\end{aligned}$$

and so the generalization of equation (2) to double factorials takes the form

$$\sum_{k=0}^{n}(k+1)\,k!! = (n+1)!! + (n+2)!! - 2. \tag{3}$$

There is something we can do with the double factorials in (3) that we could not do with the ordinary factorials in (2). We can treat the even and odd terms separately. The argument given for (3) works with no change if we apply it to just the terms with k even, or to just the terms with k odd. When n itself is even, the consequences are these:

$$\sum_{\substack{k=0\\ k\text{ even}}}^{n}(k+1)\,k!! = (n+2)!! - 1 \qquad (n\text{ even}) \tag{4}$$

$$\sum_{\substack{k=0\\ k\text{ odd}}}^{n}(k+1)\,k!! = (n+1)!! - 1 \qquad (n\text{ even}) \tag{5}$$

There is an intimate connection between the parts and the whole: Adding equations (4) and (5) gives equation (3).

If n is odd, the separation still works, but the results are subtly different:

$$\sum_{\substack{k=0\\ k\text{ even}}}^{n}(k+1)\,k!! = (n+1)!! - 1 \qquad (n\text{ odd})$$

$$\sum_{\substack{k=0\\ k\text{ odd}}}^{n}(k+1)\,k!! = (n+2)!! - 1 \qquad (n\text{ odd})$$

The double factorials on the right have traded places! One way to reconcile the last four equations is to note that in every case, the double factorial on the right is the one that would have appeared next in the sum on the left. Another way is to note that the argument of the double factorial on the right, $n+1$ or $n+2$, always has the same parity as the indices k on the left.

These results generalize easily to multi-factorials. For any positive integer m, the full identity becomes

$$\begin{aligned}\sum_{k=0}^{n}(k+m-1)\,(k)!_m &= (n+1)!_m + (n+2)!_m + \cdots + (n+m)!_m \\ &\qquad - (0)!_m - (1)!_m - (2)!_m - \cdots - (m-1)!_m \\ &= (n+1)!_m + (n+2)!_m + \cdots + (n+m)!_m - 1 - \binom{m}{2} \qquad (6)\end{aligned}$$

where, in the last line, we are using the facts that $(n)!_m = n$ for $n = 1, \ldots, m-1$ and $1 + 2 + \cdots + (m-1) = \binom{m}{2}$. The proof follows the same pattern as for the earlier results.

Just as in the case of double factorials, this result can be separated into parts. But now there is one part for each residue class mod m. When n is a multiple of m, the part corresponding to indices k that are multiples of m is

$$\sum_{\substack{k=0 \\ k\equiv 0 \bmod m}}^{n} (k+m-1)(k)!_m = (n+m)!_m - 1 \qquad (n \text{ a multiple of } m)$$

and the part corresponding to any other residue $r = 1, \ldots, m-1$ is

$$\sum_{\substack{k=0 \\ k\equiv r \bmod m}}^{n} (k+m-1)\,(k)!_m = (n+r)!_m - r \qquad (n \text{ a multiple of } m).$$

The partition still works for any nonnegative integer n, but some care must be taken. One way to express the result is as follows. For any $n \geq 0$, any $m \geq 1$, and any $r = 0, 1, \ldots, m-1$, we have

$$\sum_{\substack{k=0 \\ k\equiv r \bmod m}}^{n} (k+m-1)\,(k)!_m = (**)!_m - (r)!_m \qquad (7)$$

where $(**)$ is whichever of the numbers $n+1, n+2, \ldots, n+m$ is congruent to r mod m. That means that the entry $(**)$ is always in the same congruence class as the indices k in the sum, and, in fact, $(**)!_m$ is always the multi-factorial that would have appeared next in the sequence of multi-factorials in the sum. If we add the versions of equation (7) for all residue classes r, we get the full sum, equation (6).

It is a common custom to write sums like these in a way that avoids the extra restriction on the index k. We summarize our results so far in a theorem which honors this custom.

THEOREM 1. *Let $n \geq 0$, $m \geq 1$, and let r satisfy $0 \leq r \leq m-1$. Then*

$$\sum_{k=0}^{n}((k+1)m + r - 1)\,(km+r)!_m = ((n+1)m+r)!_m - (r)!_m.$$

It might take a long time to track all of the indices, but the sums described in the theorem turn out to be exactly the same sums as are described by equation (7).

Before proceeding to our second well known identity, we should note an alternating-sum variation of the full double-factorial identity, equation (3). It is obtained by subtracting equation (5) from equation (4):

$$\sum_{k=0}^{n}(-1)^k(k+1)\,k!! = (-1)^n\left((n+2)!! - (n+1)!!\right).$$

It holds whether n is odd or even.

Our next well known factorial identity is a reciprocal form of equation (2), namely,

$$\sum_{k=0}^{n}\frac{k}{(k+1)!} = 1 - \frac{1}{(n+1)!}. \tag{8}$$

This identity, too, has a "one-line" proof:

$$\begin{aligned}\sum_{k=0}^{n}\frac{k}{(k+1)!} &= \sum_{k=0}^{n}\frac{k+1}{(k+1)!} - \sum_{k=0}^{n}\frac{1}{(k+1)!}\\ &= \sum_{k=0}^{n}\frac{1}{k!} - \sum_{k=1}^{n+1}\frac{1}{k!} = \frac{1}{0!} - \frac{1}{(n+1)!}.\end{aligned}$$

As was the case with the proof of equation (2), the proof can be generalized to the context of multi-factorials. We will show, in detail, the double-factorial generalization of equation (8) and then provide a sketch of the multi-factorial generalization. To derive the double-factorial generalization, note that the recurrence relation $(k+2)!! = (k+2)k!!$ implies

$$\begin{aligned}\sum_{k=0}^{n}\frac{k+1}{(k+2)!!} &= \sum_{k=0}^{n}\frac{k+2}{(k+2)!!} - \sum_{k=0}^{n}\frac{1}{(k+2)!!}\\ &= \sum_{k=0}^{n}\frac{1}{k!!} - \sum_{k=2}^{n+2}\frac{1}{k!!} = \frac{1}{0!!} + \frac{1}{1!!} - \frac{1}{(n+1)!!} - \frac{1}{(n+2)!!}.\end{aligned}$$

and so the generalization of equation (8) to double factorials takes the form

$$\sum_{k=0}^{n}\frac{k+1}{(k+2)!!} = 2 - \frac{1}{(n+1)!!} - \frac{1}{(n+2)!!}. \tag{9}$$

Just as for the earlier identity, we can partition the sums into even and odd terms. When n is even, the results are

$$\sum_{\substack{k=0\\ k\text{ even}}}^{n}\frac{k+1}{(k+2)!!} = 1 - \frac{1}{(n+2)!!}, \qquad (n\text{ even}) \tag{10}$$

$$\sum_{\substack{k=0\\ k\text{ odd}}}^{n}\frac{k+1}{(k+2)!!} = 1 - \frac{1}{(n+1)!!} \qquad (n\text{ even}). \tag{11}$$

These formulas can be made to work with odd n as well, as long as we ensure that the double factorial on the right is always the same as the last one to appear on the left.

Equation (9) is, of course, the sum of equations (10) and (11). Furthermore, if we subtract equation (11) from equation (10), we obtain

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

These results can be generalized to multi-factorials. Let $m \geq 1$; then

$$\sum_{k=0}^{n}\frac{k+m-1}{(k+m)!_m}=1+\sum_{k=1}^{m-1}\frac{1}{k}-\sum_{k=1}^{m}\frac{1}{(n+k)!_m}. \qquad (12)$$

For example, if $m = 3$ this identity becomes

$$\sum_{k=0}^{n}\frac{k+2}{(k+3)!_3}=\frac{5}{2}-\frac{1}{(n+1)!_3}-\frac{1}{(n+2)!_3}-\frac{1}{(n+3)!_3}.$$

If $m = 1$, this identity becomes equation (8), while if $m = 2$, this identity becomes equation (9).

If, in what has now become the standard proof, we use our agreed convention that $(n)!_m = 1$ when $-m-1 \leq n \leq 0$, then we can obtain this variant of equation (12):

$$\sum_{k=0}^{n}\frac{k}{(k+1)!_m}=m-\sum_{k=0}^{m-1}\frac{1}{(n-k+1)!_m}.$$

If $m = 2$ this variant becomes

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

which is a variant of equation (9).

Using residue classes mod m, we can split equation (12) into m separate equations. When n is a multiple of m they take the form

$$\sum_{\substack{k=0\\ k\equiv r \bmod m}}^{n}\frac{k+m-1}{(k+m)!_m}=\frac{1}{(r)!_m}-\frac{1}{(n+r)!_m}. \qquad (n \text{ a multiple of } m)$$

As before, we would like to follow the custom of avoiding extra restrictions on the index. We therefore restate the most recent formulas in the form of a theorem, using different indices.

THEOREM 2. *Let $n \geq 0$, $m \geq 1$, and let r satisfy $0 \leq r \leq m-1$. Then*

$$\sum_{k=0}^{n}\frac{(k+1)m-r-1}{((k+1)m-r)!_m}=1-\frac{1}{((n+1)m-r)!_m}.$$

If one takes the trouble to track all of the index changes, one can verify that Theorem 2 encapsulates all of our results concerning the second identity.

Generating functions involving double factorials

In this section we turn to generating functions as another tool to obtain double-factorial identities. Given any sequence $\{a_n\}_{n=0}^{\infty}$ of real or complex numbers, define its generating function to be the power series

$$\sum_{k=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots . \tag{13}$$

One may choose to take (13) as a formal power series, which means that the x^n term is simply a placeholder for the nth term of the original series. Or, if the series converges to a function in some interval—which means that (13) is the MacLaurin series for the function—then we may take advantage of that fact and simplify the representation.

For a first example, let $\{a_n\}_{n=0}^{\infty}$ be the sequence of all ones. Then its generating function is

$$\sum_{k=0}^{\infty} a_n x^n = \sum_{k=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}, \tag{14}$$

a familiar geometric series, which converges when $|x| < 1$. As another example, take the sequence $\left\{\frac{1}{n!}\right\}_{n=0}^{\infty}$. Its generating function is

$$\sum_{k=0}^{\infty} a_n x^n = \sum_{k=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = e^x,$$

the well known exponential series, which converges for all x.

Generating functions play an important role in enumerative combinatorics since they provide an elegant way of obtain functional expressions for many combinatorial sequences. An excellent resource for generating functions is the book *generatingfunctionology* by Herb Wilf [**15**].

We take a moment to recall the definition of $\binom{x}{k}$ when k is a nonnegative integer and x is any complex number. First define $\binom{x}{0} = 1$. When $k \geq 1$, define

$$\binom{x}{k} = \frac{\prod_{i=0}^{k-1}(x-i)}{k!} = \frac{x(x-1)\cdots(x-k+1)}{k!}.$$

If x is an integer $m \geq n$, then this definition coincides with the usual definition of $\binom{m}{n} = \frac{m!}{n!\,(n-m)!}$. But the new definition works when x is not an integer. For example, we can compute

$$\binom{-1/2}{3} = \frac{(-1/2)(-3/2)(-5/2)}{3!} = -\frac{5}{16}.$$

With some effort, we can generalize this example to obtain a formula we will need later:

$$\binom{-1/2}{n} = \frac{(-1)^n}{2^{2n}}\binom{2n}{n}. \tag{15}$$

Newton showed, using Taylor series, that

$$(1+y)^x = \sum_{k=0}^{\infty} \binom{x}{k} y^k, \tag{16}$$

when x is any complex number and y is a complex number with $|y| < 1$. We will also need this formula, which (when x is not required to be a positive integer) is called Newton's Binomial Theorem.

We will now derive two generating functions, each associated with a sequence of double factorials, and use these two generating functions to answer Problem 11406 of *The American Mathematical Monthly* [**3**]. The problem asked (in effect) for a proof of the fact that we record as Theorem 3. (Concise solutions by other methods were given by Andersen and Abel at [**1**].)

THEOREM 3. *When n is a nonnegative integer,*

$$\sum_{k=0}^{n}\binom{n}{k}(2k-1)!!\,(2n-2k-1)!! = (2n)!!.$$

Proof. First, we construct the generating function for $\left\{\frac{(2n)!!}{n!}\right\}_{n=0}^{\infty}$. Recalling that $(2n)!! = 2^n n!$, we find that

$$\sum_{n=0}^{\infty}\frac{x^n}{n!}(2n)!! = \sum_{n=0}^{\infty}(2x)^n = \frac{1}{1-2x}, \tag{17}$$

which converges when $|x| < \frac{1}{2}$.

Next, we construct the generating function for $\left\{\frac{(2n-1)!!}{n!}\right\}_{n=0}^{\infty}$. Now we recall that $(2n-1)!! = \frac{(2n)!}{2^n n!}$. Using that and equation (15), along with the fact that $(-1)!! = 1$, we have

$$\sum_{n=0}^{\infty}\frac{x^n}{n!}(2n-1)!! = \sum_{n=0}^{\infty}\left(\frac{x}{2}\right)^n\binom{2n}{n} = \sum_{n=0}^{\infty}(-2x)^n\binom{-\frac{1}{2}}{n} = \frac{1}{\sqrt{1-2x}}. \tag{18}$$

The last equality in equation (18) is an application of equation (16) with $x = -\frac{1}{2}$ and $y = -2x$. Squaring equation (18) gives

$$\left[\sum_{n=0}^{\infty}\frac{x^n}{n!}(2n-1)!!\right]^2 = \frac{1}{1-2x}. \tag{19}$$

Note that the right side of equation (19) matches the right side of equation (17). This means

$$\sum_{n=0}^{\infty}\frac{x^n}{n!}(2n)!! = \left[\sum_{n=0}^{\infty}\frac{x^n}{n!}(2n-1)!!\right]^2. \tag{20}$$

If we expand the square, we have two equivalent formal power series, or if we prefer, two series that converge on the inverval $|x| < \frac{1}{2}$. With either interpretation, this means that for any n, the coefficient of x^n on left side must equal the coefficient of x^n on the right side. The question remains how to find the coefficient of x^n on the right side.

The answer is provided by the Cauchy convolution, which describes how to multiply formal power series. Given any two generating functions $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=0}^{\infty} b_n x^n$, the Cauchy convolution [**15**] states that

$$\begin{aligned}\left[\sum_{n=0}^{\infty}a_n x^n\right]\left[\sum_{n=0}^{\infty}b_n x^n\right] &= (a_0 + a_1 x + a_2 x^2 + \cdots)(b_0 + b_1 x + b_2 x^2 + \cdots)\\ &= \sum_{n=0}^{\infty} c_n x^n,\end{aligned}$$

where

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}. \tag{21}$$

Now the square in equation (20) can be simplified. Using the Cauchy convolution with $a_n = \left\{\frac{(2n-1)!!}{n!}\right\}_{n=0}^{\infty} = b_n$, we find that

$$\begin{aligned} c_n = \sum_{k=0}^{n} a_k b_{n-k} &= \sum_{k=0}^{n} \frac{(2k-1)!!}{k!}\frac{(2n-2k-1)!!}{(n-k)!} \\ &= \frac{1}{n!}\sum_{k=0}^{n} \frac{n!\,(2k-1)!!\,(2n-2k-1)!!}{(n-k)!\,k!} \\ &= \frac{1}{n!}\sum_{k=0}^{n} \binom{n}{k}(2k-1)!!\,(2n-2k-1)!!. \end{aligned}$$

Thus, we have

$$\begin{aligned} \left[\sum_{n=0}^{\infty} \frac{x^n}{n!}(2n-1)!!\right]^2 &= \sum_{n=0}^{\infty} c^n x^n \\ &= \sum_{n=0}^{\infty} \frac{x^n}{n!}\sum_{k=0}^{n} \binom{n}{k}(2k-1)!!\,(2n-2k-1)!!. \end{aligned} \tag{22}$$

Comparing the coefficients of $\frac{x^n}{n!}$ in equations (20) and (22) proves Theorem 3. ■

A variation of the technique we used to prove Theorem 3 provides insight into Formula (3.90) of H. W. Gould's *Combinatorial Identities* [**4**]. First note

$$\sum_{n=0}^{\infty} \frac{(2n-1)!!}{(2n)!!}x^n = \sum_{n=0}^{\infty} \frac{(2n)!\,x^n}{2^n n!\,2^n n!} = \sum_{n=0}^{\infty} \left(\frac{x}{4}\right)^n \binom{2n}{n}$$

Using equation (15) we are able to simplify the rightmost sum and obtain

$$\sum_{n=0}^{\infty} \frac{(2n-1)!!}{(2n)!!}x^n = \sum_{n=0}^{\infty} \binom{-\frac{1}{2}}{n}(-x)^n = \frac{1}{\sqrt{1-x}}. \tag{23}$$

By squaring equation (23), simplifying the square using the Cauchy convolution, and comparing the coefficients of x^n to those in equation (14), we find that

$$\sum_{j=0}^{n} \frac{(2j-1)!!}{(2j)!!}\frac{(2n-2j-1)!!}{(2n-2j)!!} = 1.$$

Since $(2n)!! = 2^n n!$ and $(2n-1)!! = \frac{(2n)!}{2^n n!}$, we can rewrite the left sum and obtain

$$\frac{1}{2^{2n}}\sum_{j=0}^{n} \binom{2j}{j}\binom{2n-2j}{n-j} = 1,$$

an equivalent formulation of Formula (3.90).

Another example of generating function involving double factorials involves the sequence $\left\{\frac{1}{(2n)!!}\right\}_{n=0}^{\infty} = \left\{\frac{1}{(2^n n!)}\right\}_{n=0}^{\infty}$. For this particular sequence we find that

$$\sum_{n=0}^{\infty} \frac{x^n}{(2n)!!} = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{x}{2}\right)^n = e^{x/2} \tag{24}$$

By squaring equation (24), applying the Cauchy convolution, and finally comparing the coefficients of x^n we obtain

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

Our next goal is to generalize equation (24). If we consider reciprocal coefficients involving the convolution of even double factorials, equation (24) becomes

$$\begin{aligned}\sum_{k=0}^{n} \frac{x^k}{(2k)!!\,(2n-2k)!!} &= \sum_{k=0}^{n} \frac{x^k}{2^k k!\,2^{n-k}(n-k)!}\\ &= \frac{1}{2^n n!}\sum_{k=0}^{n} \frac{n!\,x^k}{k!\,(n-k)!} = \frac{(1+x)^n}{n!\,2^n}.\end{aligned}$$

We should remark that Weisstein, in MathWord [**14**], writes our relation (24) in the form

$$\sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!!} = e^{x^2/2}$$

and notes that

$$\sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!!} = \sqrt{\frac{\pi}{2}}\,\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) e^{x^2/2},$$

where

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt.$$

He is then able to combine these into a single relation from which he notes the approximation

$$\sum_{n=0}^{\infty} \frac{1}{(n)!!} = 3.0594074053425761445\ldots.$$

Finally, Weisstein notes the remarkable formula of Ramanujan

$$\sum_{n=0}^{\infty}(-1)^n \left[\frac{(2n-1)!!}{(2n)!!}\right]^3 = \left[\frac{\Gamma(9/8)}{\Gamma(5/4)\Gamma(7/8)}\right]^2.$$

This is cited by Hardy [**5**, p. 106] among many results due to Ramanujan. Another specially pleasing Ramanujan formula cited by Hardy [**5**, p. 7] is this really amazing gem:

$$\sum_{n=0}^{\infty}(-1)^n(4n+1)\left[\frac{(2n-1)!!}{(2n)!!}\right]^3=\frac{2}{\pi}.$$

Proofs of these last four results are beyond the scope of our work here. Ramanujan's formula is certainly remarkable, and he found other similar relations. We mention them to whet the reader's appetite for truly astonishing formulas. Hardy and Ramanujan did not use the $(n)!!$ notation.

Double factorial binomial coefficients

Since factorials play such an important role in the algebraic definition of $\binom{n}{k}$, it seems only natural to define an equivalent quantity for the double factorial. The *double factorial binomial coefficient* (DFBC) is defined as

$$\left(\!\!\binom{n}{k}\!\!\right)=\frac{(n)!!}{(k)!!\,(n-k)!!}.$$

Unfortunately, this does not have a nice combinatorial meaning, since for the most part, the DFBC is not an integer. However, one special exception to this observation is when upper and lower indices are both even, since

$$\left(\!\!\binom{2n}{2k}\!\!\right)=\frac{(2n)!!}{(2k)!!\,(2n-2k)!!}=\frac{2^n n!}{2^k k!\,2^{n-k}(n-k)!}=\binom{n}{k}. \tag{25}$$

Fortunately, certain other DFBC's can be written in terms of binomial coefficients, with the aid of Table Z of *Combinatorial Identities* [**4**]. In particular, we can use Formulas (Z.45) and (Z.46) of [**4**] for this purpose. For ease of exposition, we reproduce these formulas here.

$$\binom{n-\frac{1}{2}}{k}=\binom{2n}{n}\binom{n}{k}\frac{1}{2^{2k}\binom{2n-2k}{n-k}} \tag{Z.45}$$

$$\binom{n+\frac{1}{2}}{n-k}=\frac{2n+1}{2k+1}\binom{2n}{n}\binom{n}{k}\frac{2^{2k-2n}}{\binom{2k}{k}} \tag{Z.46}$$

To use equation (Z.46), we write the right side of equation (Z.46) in terms of factorials and simplify. In particular, we have

$$\begin{aligned}\binom{n+\frac{1}{2}}{n-k}&=\frac{2n+1}{2k+1}\frac{(2n)!}{n!\,n!}\frac{n!}{k!\,(n-k)!}\frac{k!\,k!\,2^{2k-2n}}{(2k)!}\\&=\frac{(2n+1)!}{n!\,2^n}\frac{2^k k!}{(2k+1)!}\frac{1}{2^{n-k}(n-k)!}\\&=\frac{(2n+1)!!}{(2k+1)!!\,(2n-2k)!!}=\left(\!\!\binom{2n+1}{2k+1}\!\!\right)\end{aligned} \tag{26}$$

If we let $k=0$ in equation (Z.46), equation (26) becomes

$$\binom{n+\frac{1}{2}}{n}=\frac{(2n+1)!!}{(2n)!!}=\frac{(2n+1)(2n-1)!!}{(2n)!!},$$

which is $(2n+1)$ times the coefficient of x^n in the leftmost sum of equation (23).

Equations (25) and (26) show that if n and k have the same parity (i.e., are both even or both odd), then the DFBC is an ordinary binomial coefficient. These observations allow us to examine the associated generating functions. For example,

$$\sum_{k=0}^{n} \left(\!\!\binom{2n}{2k}\!\!\right) x^k = \sum_{k=0}^{n} \binom{n}{k} x^k = (1+x)^n$$

and

$$\sum_{k=0}^{\infty} \left(\!\!\binom{2n+1}{2k+1}\!\!\right) x^k = \sum_{k=0}^{n} \left(\!\!\binom{2n+1}{2k+1}\!\!\right) x^k = \sum_{k=0}^{n} \binom{n+\frac{1}{2}}{n-k} x^k = \sum_{k=0}^{n} \binom{n+\frac{1}{2}}{k} x^{n-k}.$$

However, the sums in this equation cannot be put into closed form.

It remains to analyze the case of DFBC's having different parities. Equation (Z.45) will help us with this situation. By writing the right side of equation (Z.45) in terms of factorials, we find that

$$\begin{aligned}\binom{n-\frac{1}{2}}{k} &= \frac{(2n)!}{n!\,n!}\,\frac{n!}{k!\,(n-k)!}\,\frac{(n-k)!\,(n-k)!}{(2n-2k)!\,2^{2k}}\\ &= \frac{(2n)!}{2^n n!}\,\frac{(n-k)!\,2^{n-k}}{(2n-2k)!}\,\frac{1}{k!\,2^k}\\ &= \frac{(2n-1)!!}{(2n-2k-1)!!\,(2k)!!} = \left(\!\!\binom{2n-1}{2k}\!\!\right).\end{aligned}\tag{27}$$

Using equation (Z.45) we can determine the generating function as follows.

$$\sum_{k=0}^{\infty} \left(\!\!\binom{2n-1}{2k}\!\!\right) x^k = \sum_{k=0}^{\infty} \binom{n-\frac{1}{2}}{k} x^k = (1+x)^{n-1/2}$$

Unfortunately, the remaining case of DFBC's where n is even and k is odd is not easily found in Table Z of [**4**]. However, using factorials, we can write the following

$$\begin{aligned}\left(\!\!\binom{2n}{2k+1}\!\!\right) &= \frac{(2n)!!}{(2k+1)!!\,(2n-2k-1)!!} = \frac{2^n n!\,2^k k!\,2^{n-k}(n-k)!}{(2k+1)!\,(2n-2k)!}\\ &= \frac{2^{2n}}{2k+1}\,\frac{\binom{n}{k}}{\binom{2k}{k}\binom{2n-2k}{n-k}}\end{aligned}\tag{28}$$

A useful variation of equation (Z.45) is

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

Restating gives us

$$\binom{2k}{k}\binom{2n-2k}{n-k} = (-1)^{n-k}2^{2n}\binom{n}{k}\binom{k-\frac{1}{2}}{n}.$$

Substituting this into equation (28) gives us

$$\left(\!\!\binom{2n}{2k+1}\!\!\right) = \frac{(-1)^{n-k}}{2k+1}\,\frac{1}{\binom{k-\frac{1}{2}}{n}}.\tag{29}$$

A useful restatement of equation (29) is

$$(-1)^{n-k}\binom{k-\frac{1}{2}}{n}(2k+1)=\left(\!\!\binom{2n}{2k+1}\!\!\right)^{-1}. \tag{30}$$

Equation (30) can be used to evaluate the *reciprocal* generating function, namely

$$\begin{aligned}\sum_{n=0}^{\infty}x^{n}\left(\!\!\binom{2n}{2k+1}\!\!\right)^{-1}&=(-1)^{k}(2k+1)\sum_{n=0}^{\infty}\binom{k-\frac{1}{2}}{n}(-x)^{n}\\&=(-1)^{k}(2k+1)(1-x)^{k-\frac{1}{2}}.\end{aligned}$$

We mention that Formula (2.26) of [**4**], combined with equations (29) and (27), provides further identities involving double factorial binomial coefficients. In particular, we have

$$\sum_{k=0}^{p}\frac{1}{\binom{k-\frac{1}{2}}{n}}=\frac{n}{n-1}\left(\frac{1}{\binom{-\frac{1}{2}-1}{n-1}}-\frac{1}{\binom{p-\frac{1}{2}}{n-1}}\right)=\sum_{k=0}^{p}(-1)^{n-k}(2k+1)\left(\!\!\binom{2n}{2k+1}\!\!\right),$$

and

$$\sum_{k=0}^{p}\frac{1}{\binom{n-\frac{1}{2}}{k}}=\frac{k}{k-1}\left(\frac{1}{\binom{-\frac{1}{2}-1}{k-1}}-\frac{1}{\binom{p-\frac{1}{2}}{k-1}}\right)=\sum_{n=0}^{p}\left(\!\!\binom{2n-1}{2k}\!\!\right)^{-1}.$$

Double factorials and Stirling numbers

If x is any real or complex number, we define the nth *falling power* of x to be

$$(x)_{n}=x(x-1)(x-2)\cdots(x-n+1)\qquad (n\text{ factors}).$$

This can be seen as a generalization of $n!$, since $n!=(n)_{n}$. This definition makes $(x)_{n}$ into an nth-degree polynomial in x. What are its coefficients?

The answer is given by the (signed) *Stirling numbers of the first kind*, called $s(n,k)$ in the popular notation of Riordan [**12**]. These numbers may be defined by the identity

$$\sum_{k=0}^{n}s(n,k)x^{k}=(x)_{n}=n!\binom{x}{n}. \tag{31}$$

In the definition of $(x)_{n}$, we required the factors to descend by one each time. This one-step difference in the factors is related to the definition of $n!$ and the recurrence relation $n!=n(n-1)!$. If want to adapt the concept of $(x)_{n}$ to the context of double factorials, we should let the factors decrease in steps of two, since $(n)!!=n(n-2)!$. Thus, we define $((x))_{n}$ to be the n-factor product $((x))_{n}\equiv(x)(x-2)\cdots(x-2n+2)$. With this definition, we see that

$$((x))_{n}=\prod_{k=0}^{n-1}(x-2k)=2^{n}\prod_{k=0}^{n-1}\left(\frac{x}{2}-k\right)=2^{n}\binom{\frac{x}{2}}{n}n!=2^{n}\left(\frac{x}{2}\right)_{n}. \tag{32}$$

Using equation (31), we expand the binomial coefficient in equation (32) to obtain

$$((x))_{n}=2^{n}\sum_{k=0}^{n}s(n,k)\frac{x^{k}}{2^{k}}=\sum_{k=0}^{n}s(n,k)2^{n-k}x^{k}. \tag{33}$$

This equation implies that the coefficient of x^k in the expansion of $((x))_n$ is $s(n, k)2^{n-k}$. We can invert equation (33) by using Stirling numbers of the second kind. Recall from [**12**] that

$$x^n = \sum_{k=0}^{n} S(n, k)(x)_k \tag{34}$$

with the additional condition of $S(n, 0) = 0$ for $n > 0$.

Equation (32), along with equation (34), implies

$$\frac{x^n}{2^n} = \sum_{k=0}^{n} S(n, k)\left(\frac{x}{2}\right)_k = \sum_{k=0}^{n} S(n, k)2^{-k}((x))_k. \tag{35}$$

Equation (35) provides our desired inversion of equation (33) since we can multiply both sides by 2^{-n} to obtain

$$x^n = \sum_{k=0}^{n} 2^{n-k} S(n, k)((x))_k.$$

Conclusion

We hope this paper has given you an interest into the fascinating realm of double factorials. If you want to continue to explore ways double factorials appear in combinatorial identities, we suggest perusing the tables of [**4**] and seeing how many of these formulas can be transformed into equivalent expressions involving DFBC. For example, any expression involving $\binom{x}{n-k}$ can be rewritten using equation (26) if you let $x = n + \frac{1}{2}$. Try doing this with Identities (3.4), (3.14), and (3.15) of [**4**]. Also, you can use equation (29) to transform any expression involving $\binom{x}{k}$ whenever $x = n - \frac{1}{2}$. Apply this transformation to Identities (1.9), (1.10) and (1.11) of [**4**] and see what you get. Of course, these suggestions for further exploration are only the tip of the iceberg. Play around and see what you discover. Happy Explorations!

REFERENCES

1. Kenneth F. Andersen and Ulrich Abel, Solutions to problem 11406, *Amer. Math. Monthly* **117** (2010) 935.
2. Florian Cajori, *History of Mathematical Notations*, Two Volumes, 1928–1929. Reprinted 1993, Courier Dover Publications.
3. A. A. Dzhumadil'daeva, Problem 11406, *Amer. Math. Monthly* **116** (2009) 82.
4. H. W. Gould, *Combinatorial Identities, A Standardized Set of Tables Listing 500 Binomial Coefficient Summations*, 2nd Ed., viii + 106 pp. Published by the author, Morgantown, WV, 1972.
5. G. H. Hardy, *Ramanujan, Twelve Lectures*, AMS Chelsea Publishing, 1999.
6. Charles D. Hodgman, Ed., *CRC Standard Mathematical Tables*, 12th ed., Chemical Rubber Publishing Co., Cleveland, 1954.
7. Bruce E. Meserve, Double factorials, *Amer. Math. Monthly* **55** (1948) 425–426. http://dx.doi.org/10.2307/2306136.
8. R. Ondrejka, Tables of double factorials, *Math. Comp.* **24** (1960).
9. The On-Line Encyclopedia of Integer Sequences, http://oeis.org; see A006882, A001147, A000165.
10. Jocelyn Quaintance and Harris Kwong, Permutations and combinations of colored multisets, *Journal of Integer Sequences* **3** (2010) Article 10.2.6.
11. Earl D. Rainville, *Special Functions*, Macmillan, New York, 1960.
12. John Riordan, *Introduction to Combinatorial Analysis*, Wiley, 1958.
13. John Wallis, *Arithmetica Infinitorum*, London, 1655.
14. Eric W. Weisstein, "Double Factorial," from MathWorld, a Wolfram web resource, http://mathworld.wolfram.com/DoubleFactorial.html.
15. H. S. Wilf, *generatingfunctionology*, Academic Press, 1994.

Summary The *double factorial* of n may be defined inductively by $(n+2)!! = (n+2)(n)!!$ with $(0)!! = (1)!! = 1$. Alternatively we may define this notion by the two relations $(2n)!! = 2 \cdot 4 \cdot 6 \cdot 8 \cdots (2n) = 2^n n!$ and $(2n-1)!! = 1 \cdot 3 \cdot 5 \cdot 7 \cdots (2n-1) = (2n)!/2^n n!$. Our object is to exhibit some properties and identities for the double factorials. Furthermore, we extend the notion of double factorial to the binomial coefficients by introducing double factorial binomial coefficients. The double factorial binomial coefficient is defined as

$$\left(\!\!\binom{n}{k}\!\!\right) = \frac{(n)!!}{(k)!!\,(n-k)!!}.$$

We derive identities and generating functions involving these double factorial binomial coefficients.

HENRY W. GOULD received his B.A. (1954) and M.A. (1956) at the University of Virginia. He taught full time at West Virginia University for 49 years, where since 2007 he is Professor Emeritus. He has published over 200 papers in 18 countries. He is a Fellow of the American Association for the Advancement of Science and the Institute of Combinatorics and Its Applications. His specialties are combinatorics, number theory, and history of mathematics. He has been an associate editor of the Fibonacci Quarterly since its founding in 1962, and is an associate editor of journals in China, Serbia, and Canada. His many hobbies include amateur radio, genealogy, hiking, and philosophy.

JOCELYN QUAINTANCE received her Ph.D. in mathematics from the University of Pittsburgh in 2002. Her main research interests involve combinatorial identities, enumerative combinatorics, and experimental mathematics. She is working on transcribing Professor Gould's Theory of Series and is a visiting scholar at Rutgers University.

NOTES

The Brachistochrone Problem Solved Geometrically: A Very Elementary Approach

RAYMOND T. BOUTE
INTEC, Ghent University
Ghent, Belgium
Raymond.Boute@pandora.be

One of the beauties of mathematics is that old problems never lose their freshness. They also pose the continuous challenge of finding ever simpler solutions.

A noted example is the *brachistochrone* ("shortest time") problem, found in printed sources [**1, 4, 5, 8, 9, 11**] and on the web [**12**]. Quoting Kline [**9**, Ch. 24], *"The problem is to determine the path down which a particle will slide from one given point to another not directly below in the shortest time"* (FIGURE 1). Johann Bernoulli posed this problem as a public contest. Solutions were given by Leibniz, de l'Hospital, Newton, Johann Bernoulli and his brother Jakob Bernoulli [**4**].

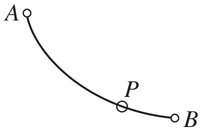


Figure 1 Illustrating the problem statement

Clearly the problem is challenging. Still, Newton is reputed to have solved it in one day.

All solutions found in the literature use analysis [**1, 4, 5, 8, 9, 11, 12**] and many comment that this is crucial. For instance, each of the Bernoulli brothers, by a different method, sets up a differential equation and recognizes it as that of the cycloid [**1**]. Modern texts [**5**] present the problem as an exercise in optimization, but using mathematical machinery (calculus of variations) not yet available at the end of the 17th century, except in embryonic form.

Instead, we use geometry only, which is arguably simpler than analysis. The derivation is direct, constructive and self-contained. Our attitude is problem solving rather than theorem proving. Indeed, exercises on other topics [**2, 3**] show that geometry often points the way by exposing the essence of a problem, and even helps us in discovering solutions rather than just proving them *a posteriori*. Here the cycloid is not taken as a proof goal, but is the solution that emerges.

The derivation is accessible to high school students—at least those not victim to curriculum reforms eliminating geometry! Notions of triangle similarity and Thales's theorem suffice.

Math. Mag. **85** (2012) 193–199. `doi:10.4169/math.mag.85.3.193`. © Mathematical Association of America

Deriving the solution

Why is the fastest path not simply a straight line? If velocity were constant, the least-time path would indeed be straight. However, gravity accelerates the particle, and a steeper initial downward slope might allow gaining velocity faster. To quantify this effect, we briefly recall some basic physics.

Physics of the problem. Assume a particle of mass m moves without friction along a path in constant gravity g. At speed v, the kinetic energy is $mv^2/2$. As the particle moves down over a level difference h, regardless of the shape of the path, gravity contributes an amount of energy mgh. Since there is no friction, this is fully imparted as an increase in kinetic energy. If the initial velocity is zero, then $mv^2/2 = mgh$. As a result, the velocity v at "depth" or *level* h below the starting point (not necessarily vertically) satisfies $v^2/2 = gh$.

Solution strategy. The shape of the fastest path is derived geometrically in three steps.

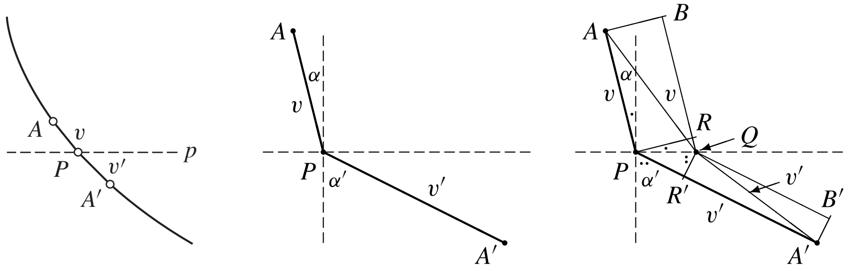
(a) Modeling local behavior, i.e., the motion along a short segment of the complete path.

(b) Deriving an invariant of the motion. It turns out that this can be linked to a circle with constant radius.

(c) Decomposing the velocity in a translational and a rotational component for that circle. Magnitudes turn out to be equal, making the trajectory a cycloid by definition.

Step (a): Local least time model. Replacing any segment of a path with a faster segment shortens the total time. Such a *local* change assumes the velocities at the endpoints of the segment unchanged, which is indeed so since velocity depends on the level only. Hence any segment of the fastest path is itself a fastest path. To model changes in speed and in the angle with the vertical, consider a short arc AA' and approximate it by two chords AP and PA', as in FIGURE 2(a). As discussed later, the usual presentation involving layers of constant velocity (changing at the boundaries) is methodologically unsound, which is often overlooked because it yields the same end result. Instead, we use a simple fact, probably known to Galilei [**1**]:

THEOREM. *The average velocity of a particle moving along a segment under (any) constant acceleration is the arithmetic mean of the velocities at the endpoints.*



(a) Short segment AA' (b) Conventions for v, α (c) Minimizing time

Figure 2 Deriving a local least-time property

Proof. Let S and F be the start and finish points, a the acceleration, t the travel time. Then $SF = V_S t + at^2/2 = (V_S + V_F)t/2$, hence $SF/t = (V_S + V_F)/2$. ■

Since V_P depends on level only, the averages defined by $v := (V_A + V_P)/2$ and $v' := (V_P + V_{A'})/2$ are unchanged if P is displaced on the horizontal p. The crux is placing P to minimize time.

A similar fastest path problem models light refraction at the boundary between regions with different (constant) light speeds. This problem was first solved by Fermat [**1, 11**] using a form of derivatives, but we allow only geometry. A geometric proof due to Huygens is given in [**11**]. For the sake of completeness, we present a more structured version.

For a polygonal path such as XY or XYZ, we write $T(XY)$ or $T(XYZ)$ for travel time, so that $T(AP) = AP/v$ etc.

Let Q be a point different from P on the same horizontal, defining the path AQA', as in FIGURE 2(c). To compare $T(APA')$ and $T(AQA')$, we decompose paths into segments. Let B and B' be the projections of A and A' on the lines through Q parallel to PA and PA' respectively.

For Q as shown w.r.t. P, let R and R' be such that $PABR$ and $QB'A'R'$ are rectangles. We calculate $T(AQA') - T(APA')$ by chaining equalities and inequalities [**7**] in steps small enough to be self-explanatory with reference to FIGURE 2(c). The third step uses $AQ > BQ$ and $QA' > QB'$.

$$\begin{aligned}
T(AQA') - T(APA') &= T(AQ) + T(QA') - T(AP) - T(PA') \\
&= AQ/v + QA'/v' - AP/v - PA'/v' \\
&> BQ/v + QB'/v' - AP/v - PA'/v \\
&= (BR + RQ)/v + QB'/v' - AP/v - (PR' + R'A')/v' \\
&= (AP + RQ)/v + R'A'/v' - AP/v - (PR' + R'A')/v' \\
&= RQ/v - PR'/v' \\
&= PQ(\sin\alpha/v - \sin\alpha'/v')\ ,
\end{aligned}$$

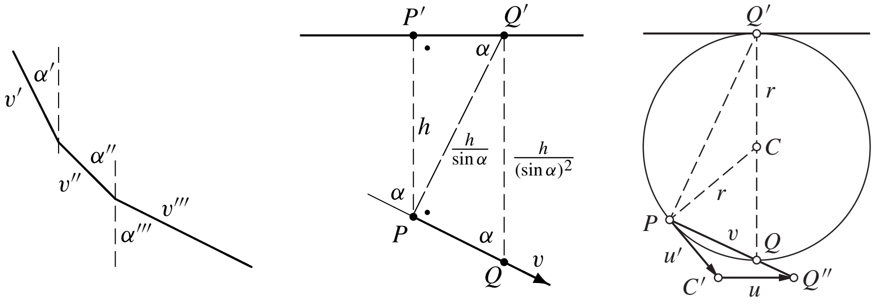
With Q on the other side, $T(AQA') - T(APA') > PQ(\sin\alpha'/v' - \sin\alpha/v)$. In either case, if P satisfies $\sin\alpha/v = \sin\alpha'/v'$ then $T(AQA') - T(APA') > 0$, so AQA' takes more time (this covers any point with $\sin\alpha/v \neq \sin\alpha'/v'$, granting the converse as well). Hence the position of P making APA' the fastest path is given by

$$\frac{v}{\sin\alpha} = \frac{v'}{\sin\alpha'} \tag{1}$$

Huygens's argument to model refraction of light assumes *local* speed that is (i) constant in each region and (ii) may jump at the boundary. Both are essential in canceling terms to obtain $RQ/v - PR'/v'$. The usual presentation of the brachistochrone borrows this argument by considering thin adjacent layers of constant local speed to make (i) and (ii) valid. However, along the brachistochrone local speed varies smoothly! For *average* velocities as shown, Huygens's argument obtains a free generalization, with formally the same calculations.

Step (b): Deriving invariants. In a dynamical process (our sliding particle), an *invariant* is a quantity that stays constant and represents a nontrivial characteristic of the motion, thus supporting further derivation or construction. In the geometry of motion, an invariant typically is a constructible line segment whose length remains constant during the motion.

Applying (1) to adjacent segments as in FIGURE 3(a), by transitivity of equality: $v'''/\sin\alpha''' = v''/\sin\alpha'' = v'/\sin\alpha'$ irrespective of chord lengths and velocities (indeterminacies discussed later). Hence, for the limit, the path, we get directly $v/\sin\alpha = w$ where w is a constant.



(a) Minimal time path (b) Geometry at a point (P) (c) Geometry in motion

Figure 3 Geometric derivation of the brachistochrone

Eliminating v from $v/\sin\alpha = w$ and $v^2/2 = gh$ yields the invariant $h/(\sin\alpha)^2$ since

$$\frac{h}{(\sin\alpha)^2} = \frac{w^2}{2g} \tag{2}$$

Let P be an arbitrary position along the path and h be the depth (below the starting point). A segment of length $h/(\sin\alpha)^2$ is easily constructed via two successive right triangles, as in FIGURE 3(b). Given the evident nature of the construction, we forgo the custom of describing it in words. Since $\triangle Q'PQ$ is a right triangle, P is on the circle with constant diameter $Q'Q$, by Thales's theorem. Let C be the center, as in FIGURE 3(c).

Step (c): Generating the trajectory. Since P is always on a circle with constant diameter $w^2/2g$ and tangent to the horizontal at $h = 0$, the trajectory can be traced by considering P to be attached to the circle, allowing C to move horizontally and the circle to rotate around C. This can be done for any motion confined to the strip between $h = 0$ and $h = w^2/2g$, and determines the instantaneous translational velocity u (of C) and rotational velocity u' at the rim (relative motion around C) uniquely, except when the directions are parallel, i.e., if P is at Q or Q').

Consider now the geometry of these velocities for the motion of interest, as in FIGURE 3(c). To avoid clutter, we draw the velocity vectors just as arrows marked with magnitudes. $\triangle PC'Q''$ is similar to $\triangle PCQ'$ by perpendicularity of corresponding sides. Hence $CP = CQ'$ implies $u' = u$.

By this equality, at Q' the rotational and translational velocities cancel each other (same magnitude, opposite direction). Hence the point of the moving circle that touches the zero-height horizontal is not moving, so the circle rolls without slipping. Therefore, by definition, the movement of P is cycloidal.

FIGURE 4 intuitively depicts a small displacement PP', decomposable into PT due to rotation and TP' due to translation. The length of segment(TP') equals the length of arc(PT) which, for small displacements, is approximated by the length of chord(PT).

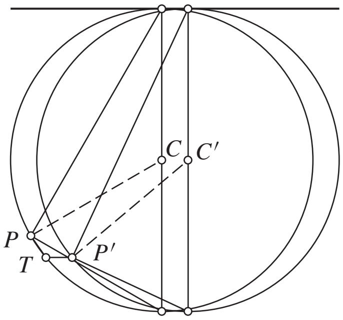


Figure 4 Illustration: small displacement by rolling without slipping

A bonus: the velocity of the particle. The translational velocity u of C is obtained as a bonus by the aforementioned similarity of $\triangle PC'Q''$ and $\triangle PCQ'$ in FIGURE 3(c):

$$\frac{u}{r} = \frac{C'Q''}{CQ'} = \frac{PQ''}{PQ'} = \frac{v}{2r\sin\alpha} = \frac{w}{2r}.$$

Hence the motion of the particle is the same as that of a point on (the circumference of) a circle rolling at constant velocity (without sliding): $u = w/2$.

The geometry of FIGURE 3(c) turns out to directly entail Roberval's construction of the tangent of the cycloid which, as noted in [**1**], is very accessible to students.

Scaling the cycloid. The boundary conditions for the trajectory are its starting point and its ending point. Let the circle start rolling with P at the starting point. The constant velocity w is a scale factor with which one can adjust the size of the circle ($D = w^2/2g$) and hence of the cycloid in order to pass through the given ending point.

Resolving indeterminacies. Indeterminacies arise in $v/\sin\alpha$ for $v = 0$ and $\alpha = 0$ (starting point) and in the decomposition of the velocity if the horizontal and tangential directions coincide ($v = u + u'$ at Q and $v = u - u'$ at Q'). These cases are similar to $\frac{\sin x}{x}$ and are resolved in the usual way: by filling the one-point gap with a value that makes the function continuous. Constructions in classical geometry usually resolve such points in their stride, as done here.

Geometry of the tautochrone

Christiaan Huygens observed in the 17th century that, if the bob of a pendulum clock can be made to follow a cycloidal path, the duration of a full swing, starting from rest in any initial position, is independent of that initial position [**6**]. This is called the *tautochrone* ("same time") property. For an ordinary pendulum hanging from a fixed point and moving in a circular arc, the tautochrone property holds only approximately and provided the swing is small.

In a comment about an earlier version of this paper, Gary Lawlor encouraged the author to verify geometrically that the cycloid is a tautochrone path.

FIGURE 4 is redrawn as FIGURE 5(a), in the direction a swinging pendulum is usually shown and less cramped, but remembering that PT is small. Let β be the angle between the normal to the path and the vertical or, equivalently, between the tangent and the horizontal.

Let us first find the arc length starting from the lowest point of the cycloid ($\beta = 0$).

For small PT, approximately $TP' = PT$ and hence $\angle P'PT = (\beta + \beta')/2$ (hint: calculate angles w.r.t. the horizontal). Let PV and CC' be the projections of PT and QQ' (respectively) on PP', so $CC' = PV = VP'$. Therefore

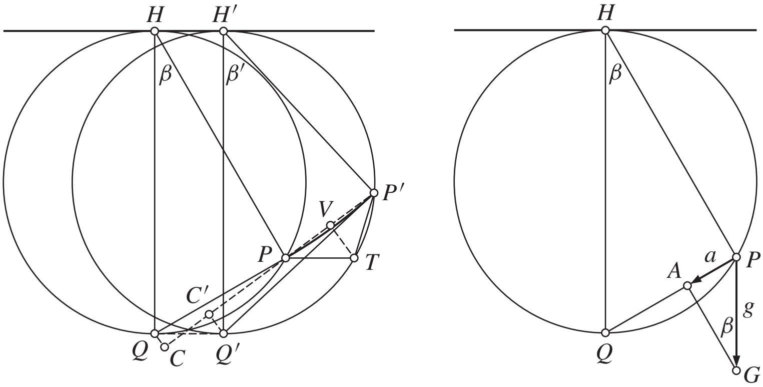
$$PP' = 2 \cdot VP' = 2(C'P' - C'V) = 2(C'P' - CP).$$

Since $\angle QPC = \angle Q'P'C' = (\beta' - \beta)/2$ and $\beta' - \beta$ is small, $C'P' = Q'P'$ and $CP = QP$ (approximately), so

$$PP' = 2(Q'P' - QP) = 4r(\sin\beta' - \sin\beta).$$

Hence for the arc length $\Delta s = PP' = 2 \cdot \Delta\, QP = 4r \cdot \Delta(\sin\beta)$ and finally, starting from the lowest point, $s = 2 \cdot QP = 4r \sin\beta$, geometrically and trigonometrically (in that order). The total arc length (full swing) of the cycloid is $8r$.

Note in passing that, beside simplicity, other advantages over arguments based on $ds = \sqrt{(dx)^2 + (dy)^2}$ are the absence of square roots and the preservation of the sense of direction.



(a) Deriving the arc length (b) Deriving the acceleration

Figure 5 Tautochrone property of the cycloid

The acceleration a of the bob in the direction of the path (i.e., tangentially, FIGURE 5b) is given by $a = -g\sin\beta$ (vectorially: $\overrightarrow{PA} = -g \cdot \overrightarrow{QP}/HQ$), where g is the acceleration of gravity. This is exactly proportional to the distance s along the path from the lowest point, more specifically $a = -gs/(4r)$. Consider now a swing starting from rest at a given initial position. For a swing starting from rest at any other position, arc lengths, accelerations and velocities are scaled linearly by the same factor. Hence the swing durations are the same.

For readers who wonder how motion might be coerced along a cycloidal path, this is shown FIGURE 6, borrowed with permission from Lawlor's paper [**10**]. The lower cycloid describes the path of the bob K, provided the string on which the bob is suspended is coerced by the upper cycloid. That the string in any position is normal to the lower cycloid and tangent to the upper cycloid follows by simple inspection from FIGURE 3(c), matching KJ to PQ' and LJ to PQ.

More details and a different approach to the brachistochrone problem are given in [**10**].

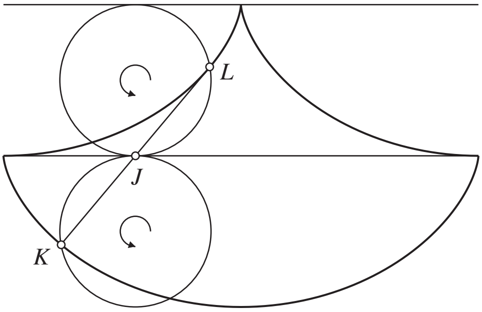


Figure 6 Cycloidal pendulum motion

Conclusion

A simple geometric solution was given to a once very challenging problem. As far as could be verified, it is perhaps the first one without calculus, and yields the cycloid by direct construction. The geometric solution is very accessible to students without calculus background, and provides additional insight to those who have had calculus, on the basis of the principle "if you have not solved the problem in more than one way, you have not solved the problem".

REFERENCES

1. Jeff Babb and James Currie, The Brachistochrone Problem: Mathematics for a broad audience via a large context problem, *The Montana Mathematics Enthusiast* **5** (no. 2&3) (2008) 169–183. http://www.math.umt.edu/TMME/vol5no2and3/TMME_vol5nos2and3_a1_pp.169_184.pdf
2. Raymond T. Boute, Simple geometric solutions to De l'Hospital's pulley problem, *College Math. J.* **30** (1999) 311–314. http://dx.doi.org/10.2307/2687672.
3. Raymond T. Boute, Moving a rectangle around a corner—geometrically, *Amer. Math. Monthly* **111** (2004) 435–437. http://dx.doi.org/10.2307/4145272.
4. Carl B. Boyer and Uta C. Merzbach, *A History of Mathematics*, 2nd ed., Wiley, 1991.
5. Arthur E. Bryson and Yu-Chi Ho, *Applied Optimal Control*, Hemisphere, 1975.
6. Hansjörg Geiges, Christiaan Huygens and Contact Geometry, *Nieuw Archief voor Wiskunde, ser. 5*, **6** (2005) 117–123.
7. David Gries and Fred B. Schneider, Teaching math more effectively, through calculational proofs, *Amer. Math. Monthly* **102** (1995) 691–697. http://dx.doi.org/10.2307/2974638.
8. Ernst Hairer and Gerhard Wanner, *Analysis by Its History*, Springer, 2000.
9. Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, 1972.
10. Gary Lawlor, A New Minimization Proof for the Brachistochrone, *Amer. Math. Monthly* **103** (1996) 242–249. http://dx.doi.org/10.2307/2975375.
11. Vladimir M. Tikhomirov, *Stories about Maxima and Minima* (Mathematical World, Vol. 1), AMS–MAA, 1990.
12. Eric S. Weisstein, "Brachistochrone Problem", in *MathWorld,* a Wolfram Web Resource, http://mathworld.wolfram.com/BrachistochroneProblem.html

Summary The brachistochrone is the path of swiftest descent for a particle under gravity between points not on the same vertical. The problem of finding it was posed in the 17th century, and only analytical solutions appear to be known. Here a geometrical solution is given requiring only basic properties of triangles, and the result is the cycloid. The cycloid is also shown by geometry to be Huygens's tautochrone. The presentation style is tutorial, and the geometric arguments are accessible to high school students.

My Favorite Rings

JOHN KILTINEN
Northern Michigan University
Marquette, MI

KAREN AUCOIN
McNeese State University
Lake Charles, LA

(Tune: “My favorite things” from “The Sound of Music”)

1
Integers, all types, plus, zero, and minus,
Matrices square, equal width to their highness,
Integers modulo all finite things,
These are a few of my favorite rings.

Add to a base ring a number of unknowns,
X one, two, three, of each other they are clones,
Add them and multiply, it makes me sing,
Lengthens the list of my favorite rings.

Refrain
When I’m asked to
Prove a theorem,
And I’m feeling blue,
I simply remember my favorite rings,
And then I know what to do!

2
Take any ring, on it take all the functions,
Add them and multiply, with no compunctions,
Do the ops pointwise, it is just the thing,
Now I have built a new favorite ring.

Rings do not have to be like this, quixotic,
Just take a number set far less exotic
Rationals, reals, or more complex things,
Put on my list of my favorite rings,

Refrain
On a down day,
I just sort them,
These commute, these don’t,
I sort and I pile up my favorite rings,
And then staying blue, I won’t!

Math. Mag. **85** (2012) 200. doi:10.4169/math.mag.85.3.200. © Mathematical Association of America

A Remarkable Combinatorial Identity

MIHAIL FRUMOSU
mfrumosu@gmail.com

ALEXANDER TEODORESCU-FRUMOSU
Clark University
Worcester, MA
afrumosu@clarku.edu

The following identity holds for all integers $m \geq 2$ and all positive integers p:

$$\sum_{p=1}^{m} \left(\frac{(-1)^p}{p!} \sum_{k_1+\cdots+k_p=m} \frac{1}{k_1 \cdot \cdots \cdot k_p} \right) = 0. \tag{1}$$

The inner sum is over all p-term *ordered partitions* of m; that is, over all sequences $k_1, k_2, \ldots, k_p$ of positive integers such that $k_1 + \cdots + k_p = m$.

Example. We illustrate the identity in the case $m = 4$.

$$\begin{aligned}
&\sum_{p=1}^{4} \left(\frac{(-1)^p}{p!} \sum_{k_1+\cdots+k_p=4} \frac{1}{k_1 \cdot \cdots \cdot k_p} \right) \\
&\quad = -\frac{1}{1!}\left(\frac{1}{4}\right) + \frac{1}{2!}\left(\frac{1}{1\cdot 3} + \frac{1}{3\cdot 1} + \frac{1}{2\cdot 2}\right) \\
&\qquad - \frac{1}{3!}\left(\frac{1}{1\cdot 1\cdot 2} + \frac{1}{1\cdot 2\cdot 1} + \frac{1}{2\cdot 1\cdot 1}\right) + \frac{1}{4!}\left(\frac{1}{1\cdot 1\cdot 1\cdot 1}\right) \\
&\quad = -\frac{1}{4} + \frac{1}{2!}\left(\frac{1}{1\cdot 3}\cdot\frac{2!}{1!\,1!} + \frac{1}{2\cdot 2}\cdot 1\right) - \frac{1}{3!}\left(\frac{1}{1\cdot 1\cdot 2}\cdot\frac{3!}{2!\,1!}\right) + \frac{1}{4!}\left(\frac{1}{1\cdot 1\cdot 1\cdot 1}\cdot 1\right) \\
&\quad = -\frac{1}{4} + \frac{11}{24} - \frac{1}{4} + \frac{1}{24} = 0
\end{aligned}$$

In the second (split) line of the example, every ordered partition of 4 appears in one denominator. In the third line, partitions that differ only by order are grouped together. For example, when $p = 3$, each of the triplets $(1, 1, 2)$, $(1, 2, 1)$, $(2, 1, 1)$ gives the same product, so they are grouped together.

We briefly recall some facts about ordered partitions.

Ordered partitions. Let S be a set of n elements, and suppose that $n = n_1 + n_2 + \cdots + n_k$. An ordered partition is said to be of type $(n_1, n_2, \ldots, n_k)$ if it contains n_1 ones, n_2 twos, and in general n_k entries equal to k for each k. The number of ordered partitions of S of type $(n_1, n_2, \ldots, n_k)$ is

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_k!}.$$

By this definition, the tuples $(1, 3)$ and $(3, 1)$ in the $p = 2$ case of the example are ordered partitions of type $(1, 0, 1)$; their number is given by $\frac{2!}{1!\,1!} = 2$. On the other hand,

Math. Mag. **85** (2012) 201–205. doi:10.4169/math.mag.85.3.201. © Mathematical Association of America

in the $p = 3$ case, the triples $(1, 1, 2)$, $(1, 2, 1)$ and $(2, 1, 1)$ are ordered partitions of type $(2, 1)$; their number is $\frac{3!}{2!\,1!} = 3$.

We are now ready to prove the identity.

Proof of formula (1). By calculus—either by a Taylor expansion about $x = 0$ or a term-by-term integration of a geometric series—we can establish the following power series expansion for $(-\ln(1 - x))$:

$$-\ln(1-x) = \int \frac{1}{1-x}\,dx = \int \left(\sum_{n=0}^{\infty} x^n\right) dx = \sum_{n=0}^{\infty} \frac{x^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{x^n}{n}.$$

The series converges absolutely for $|x| < 1$.

We next look at powers of this infinite series, as these will be needed below. Thus, for $p \geq 2$,

$$\begin{aligned}(-\ln(1-x))^p &= \left(\sum_{n=1}^{\infty} \frac{x^n}{n}\right)^p \\ &= \left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right)\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) \cdot \cdots \cdot \left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) \\ &= \sum_{m=2}^{\infty} \left(\sum_{k_1+\cdots+k_p=m} \frac{1}{k_1 \cdot \cdots \cdot k_p}\right) x^m. \qquad (2)\end{aligned}$$

The last line follows because, for any fixed m, there is a contribution to the coefficient of x^m from each selection of terms $1/k_1, \ldots, 1/k_p$ from the various factors for which $k_1 + \cdots + k_p = m$.

Using the power series expansion for e^x—known to converge everywhere—we can write (for all $|x| < 1$):

$$\begin{aligned}1 - x = e^{\ln(1-x)} &= \sum_{p=0}^{\infty} \frac{(\ln(1-x))^p}{p!} \qquad (3) \\ &= 1 + \sum_{p=1}^{\infty} \frac{(-1)^p}{p!} \left(\sum_{n=1}^{\infty} \frac{x^n}{n}\right)^p \\ &= 1 - \frac{1}{1!}\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) \\ &\quad + \frac{1}{2!}\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right)\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) - \cdots \\ &= 1 - x - \left(\frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) + \frac{1}{2!}\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right)^2 - \cdots \\ &= 1 - x + \sum_{m=2}^{\infty} \left[\sum_{p=1}^{m} \left(\frac{(-1)^p}{p!} \sum_{k_1+\cdots+k_p=m} \frac{1}{k_1 \cdot \cdots \cdot k_p}\right)\right] x^m \qquad (4)\end{aligned}$$

as noted in equation (2).

Equation (4) implies now that

$$\sum_{m=2}^{\infty}\left[\sum_{p=1}^{m}\left(\frac{(-1)^p}{p!}\sum_{k_1+\cdots+k_p=m}\frac{1}{k_1\cdot\cdots\cdot k_p}\right)\right]x^m=0$$

for all x inside the interval of convergence. It follows therefore that

$$\sum_{p=1}^{m}\left(\frac{(-1)^p}{p!}\sum_{k_1+\cdots+k_p=m}\frac{1}{k_1\cdot\cdots\cdot k_p}\right)=0,\qquad\text{for all } m\geq 2. \qquad ■$$

Déjà-vu. We would like to highlight some historical notes that seem connected to this combinatorial identity. We begin by recalling the number of partitions of a set $\{1,2,\ldots,m\}$ into p blocks, also known as Stirling number of the second kind [**1**].

$$S(m,p)=\frac{1}{p!}\sum_{k_1+\cdots+k_p=m}\binom{m}{k_1,\ldots,k_p} \tag{5}$$

These numbers satisfy the identity (see [**2**])

$$\sum_{p=1}^{m}(-1)^p(p-1)!\,S(m,p)=0. \tag{6}$$

Combining (5) and (6) we arrive at

$$\sum_{p=1}^{m}\frac{(-1)^p}{p}\sum_{k_1+\cdots+k_p=m}\binom{m}{k_1,\ldots,k_p}=0 \tag{7}$$

so that upon factoring $m!$ all the way out and dividing (7) by it we reach

$$\sum_{p=1}^{m}\frac{(-1)^p}{p}\sum_{k_1+\cdots+k_p=m}\frac{1}{k_1!\cdots k_p!}=0. \tag{8}$$

When comparing (8) to the identity (1), we find the resemblance between these two identities quite striking. In fact, we can now write:

$$\sum_{p=1}^{m}\frac{(-1)^p}{p!}\sum_{k_1+\cdots+k_p=m}\frac{1}{k_1\cdot\cdots\cdot k_p}=\sum_{p=1}^{m}\frac{(-1)^p}{p}\sum_{k_1+\cdots+k_p=m}\frac{1}{k_1!\cdots k_p!}=0$$

allowing this equality to be paraphrased as a “factorial commute” between the inner k_i’s and the outer p. We need to stress that while these two identities are closely related visually, we are unaware of any deeper connection between them.

Power series of a composition. Suppose that f has a power series expansion

$$f(x)=\sum_{n=1}^{\infty}a_nx^n,\qquad\text{for } |x|<R_1$$

and that g has its own power series

$$g(x) = \sum_{k=1}^{\infty} b_k x^k, \qquad \text{for } |x| < R_2$$

If $R_2 < R_1$, we can write the composition $f \circ g$ as follows:

$$\begin{aligned}(f \circ g)(x) &= f(g(x)) \\ &= \sum_{n=1}^{\infty} a_n \left(\sum_{k=1}^{\infty} b_k x^k \right)^n \\ &= \sum_{m=1}^{\infty} \left[\sum_{p=1}^{m} \left(a_p \sum_{k_1+\cdots+k_p=m} b_{k_1} \cdot \cdots \cdot b_{k_p} \right) \right] x^m \qquad (9)\end{aligned}$$

which establishes an identity similar to that in (2), but in more generality.

We note that the inner series for $g(x)$ is a unit series in this setup, i.e., a series with no constant term. This is helpful, for the infinite series induced by the free term of g could stand in the way of convergence of the composite $f \circ g$. The following example illustrates this last remark.

Let

$$f(x) = \sum_{n=1}^{\infty} nx^n, \qquad \text{for } |x| < 1,$$

and

$$g(x) = 1 + f(x), \qquad \text{for } |x| < 1.$$

A straightforward calculation can, in fact, show that $f(x) = \frac{x}{(1-x)^2}$ so that

$$(f \circ f)(x) = \frac{x(1-x)^2}{(x^2 - 3x + 1)^2}$$

with the associated power series

$$(f \circ f)(x) = \left(x + 2x^2 + 3x^3 + \cdots\right) + 2\left(x + 2x^2 + 3x^3 + \cdots\right)^2 + \cdots .$$

This series remains convergent in any interval centered at zero, as long as the root $\frac{3-\sqrt{5}}{2}$ is avoided. On the other hand,

$$(f \circ g)(x) = \frac{(x^2 - x + 1)(1-x)^2}{x^2}$$

with the power series

$$\begin{aligned}(f \circ g)(x) &= (1 + x + 2x^2 + \cdots) + 2(1 + x + 2x^2 + \cdots)^2 \\ &\quad + 3(1 + x + 2x^2 + \cdots)^3 + \cdots \\ &= (1 + 2 + 3 + \cdots) + (x\text{-terms}).\end{aligned}$$

The rearrangement of this series is certainly permitted for nonnegative x, as nonnegative series are commutative—even if divergent. In addition, the series fails to converge if $x = 0$ and therefore in any interval centered at *zero*.

Bell polynomials. The multinomial coefficients that emerge from counting the number of $(k_1, \dots, k_p)$-orderings in the Formula (1), as described in the beginning of the article, are also showing in the so-called Bell polynomials—named after Eric Temple Bell in combinatorial mathematics:

$$B_{n,k}(x_1, \dots, x_{n-k+1}) = \sum \frac{n!}{j_1! \cdots j_{n-k+1}!} \left(\frac{x_1}{1!}\right)^{j_1} \cdot \cdots \cdot \left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}}$$

where the sum extends over all sequences of non-negative integers $j_1, \dots, j_{n-k+1}$ with the property that $j_1 + j_2 + \cdots = k$ and $j_1 + 2j_2 + 3j_3 + \cdots = n$.

Food for thought. Although our proof of the Proposition was strictly calculus-based, it would be desirable, in the least, to seek and obtain a more discrete, combinatorially based argument for this identity.

REFERENCES

1. Warren P. Johnson, The curious history of Faà di Bruno's formula, *Amer. Math. Monthly* **109** (March, 2002) 217–234. http://dx.doi.org/10.2307/2695352.
2. John Riordan, *Combinatorial Identities*, Wiley, 1979.

Summary We state and prove a combinatorial identity. Given an integer m, one forms all ordered partitions (or compositions) of m, then forms the product of the entries in each ordered partition. The identity combines the inverses of these products. We prove the identity by analytic methods and relate it to other identities with similar constructions.

Morley's Other Miracle

CHRISTIAN AEBI
Collège Calvin
Geneva, Switzerland 1211
christian.aebi@edu.ge.ch

GRANT CAIRNS
La Trobe University
Melbourne, Australia 3086
G.Cairns@latrobe.edu.au

The title refers to a remarkable contribution by Frank Morley to number theory:

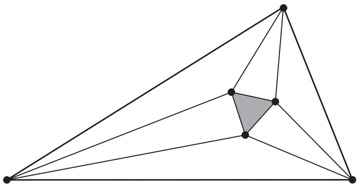
$$4^{p-1} \equiv \pm\binom{p-1}{\frac{p-1}{2}} \pmod{p^3}.$$

if p is a prime and $p \geq 5$.

In geometry, *Morley's miracle* says that in every planar triangle the adjacent angle trisectors meet at the vertices of an equilateral triangle. Morley obtained this wonderful result in 1899, and to this day it continues to attract interest. There are now many known proofs; see the cut-the-knot web site [**1**]. Perhaps the most celebrated ones are those due to Alain Connes [**2**, pp. 43–46] and John Conway (unpublished, yet

Math. Mag. **85** (2012) 205–211. doi:10.4169/math.mag.85.3.205. © Mathematical Association of America

accessible at [**1**]). A proof in the same spirit as Connes' was published earlier by Liang-shin Hahn [**6**]; see also [**4**]. Conway's proof is perhaps the simplest and nicest one; a somewhat longer proof having the same general approach was given by Coxeter [**3**] and attributed to Raoul Bricard; see also [**10, 12**].



Morley's geometric miracle was by no means his sole surprising discovery. In number theory, he published the following result in the Annals of Mathematics 1894–1895 [**9**].

MORLEY'S CONGRUENCE. *If p is prime and $p \geq 5$, then*

$$(-1)^{(p-1)/2} \cdot \binom{p-1}{\frac{p-1}{2}} \equiv 2^{2p-2} \pmod{p^3}.$$

To appreciate the "miraculous" nature of this congruence, one first needs to compare it with other congruences known at the time. Some famous ones for primes p include:

- Fermat's little theorem: $2^{p-1} \equiv 1 \pmod{p}$.
- Wilson's theorem: $(p-1)! \equiv -1 \pmod{p}$.
- Lucas' theorem: If $0 \leq n, j < p$, then $\binom{pm+n}{pi+j} \equiv \binom{m}{i}\binom{n}{j} \pmod{p}$.

The above three congruences are modulo p, while Morley's congruence is modulo p^3. The difference between mod p^3 and mod p is analogous to having a result to three significant figures, rather than just one significant figure.

The other striking aspect of Morley's congruence was the nature of his original proof, which made an ingenious use of integration of trigonometric sums. First he used the Fourier series:

$$\begin{aligned} 2^{2n}\cos^{2n+1}x &= \cos(2n+1)x + (2n+1)\cos(2n-1)x \\ &\quad + \frac{(2n+1)2n}{1.2}\cos(2n-3)x + \cdots + \frac{(2n+1)2n\ldots(n+2)}{n!}\cos x. \end{aligned}$$

He integrated this term by term and compared it with the following formula, which can be obtained by induction using integration by parts:

$$\int_0^{\frac{1}{2}\pi} \cos^{2n+1}x\,dx = \frac{2n(2n-2)\ldots 2}{(2n+1)(2n-1)\ldots 3}. \tag{1}$$

This established his result modulo p^2, where $p = 2n+1$. To obtain the result modulo p^3, Morley then used (1) again to integrate the following power series in $\cos x$, known from "treatises on trigonometry":

$$\begin{aligned} (-1)^{\frac{p-1}{2}}\cos px &= p\cos x - \frac{p(p^2-1^2)}{3!}\cos^3 x \\ &\quad + \frac{p(p^2-1^2)(p^2-3^2)}{5!}\cos^5 x - \cdots + (-1)^{\frac{p-1}{2}}2^{p-1}\cos^p x. \end{aligned}$$

Subsequently, two alternate proofs were given that used the properties of Bernoulli numbers: the 1913 Royal Danish Academy of Sciences paper by Niels Nielsen [**11**, p. 353] and the 1938 Annals of Mathematics paper by Emma Lehmer [**8**, p. 360].

The main aim of this note is to establish Morley's congruence by entirely elementary number theory arguments. The key to this approach is the following basic congruence modulo p that curiously, we have not seen in the literature. A proof appears below.

LEMMA 1. *If p is prime and $p \geq 5$, then*

$$\sum_{\substack{0<j<i<p \\ i \text{ even},\ j \text{ odd}}} \frac{1}{ij} \equiv 0 \pmod{p}.$$

Here, $\frac{1}{ij}$ denotes the multiplicative inverse of ij modulo p. Throughout this note, p is a prime greater than 3 and $\frac{1}{i}$ denotes either the fraction $1/i$ or the multiplicative inverse of i modulo p or modulo p^2, according to the context. Thus, modulo p, the symbol $\frac{1}{i}$ denotes the unique integer x satisfying $1 \leq x \leq p-1$ and $ix \equiv 1 \pmod{p}$. Similarly, modulo p^2, the symbol $\frac{1}{i}$ denotes the unique integer x satisfying $1 \leq x \leq p^2 - 1$ and $ix \equiv 1 \pmod{p^2}$.

After we have established Morley's congruence, we will show in the final section that it can also be deduced from Granville's elegant proof of Skula's conjecture [**5**].

Reduction of the problem

We will use the following well known facts [**7**, Theorem 117], that we prove for completeness.

LEMMA 2. *If p is prime and $p \geq 5$, then*

$$\sum_{i=1}^{(p-1)/2} \frac{1}{i^2} \equiv 0 \pmod{p}.$$

Proof. As $\frac{1}{i^2} \equiv \frac{1}{(p-i)^2} \pmod{p}$, one has

$$2\sum_{i=1}^{(p-1)/2} \frac{1}{i^2} \equiv \sum_{i=1}^{p-1} \frac{1}{i^2} = \sum_{j=1}^{p-1} j^2 = \frac{(p-1)p(2p-1)}{6} \equiv 0 \pmod{p}.$$

The second step uses the substitution $j = \frac{1}{i}$; note that j takes on the same set of values as i. ■

LEMMA 3. *If p is prime and $p \geq 5$, then*

$$\sum_{i=1}^{p-1} \frac{(-1)^i}{i} \equiv \sum_{i=1}^{(p-1)/2} \frac{1}{i} \pmod{p^2}.$$

Proof. When $0 < i \leq \frac{p-1}{2}$, one has $i(p-i) + i^2 \equiv -p(p-i) \pmod{p^2}$, and dividing by $i^2(p-i)$ gives $\frac{1}{i} + \frac{1}{p-i} \equiv -\frac{p}{i^2} \pmod{p^2}$. Summing and using Lemma 2 gives $\sum_{i=1}^{p-1} \frac{1}{i} \equiv 0 \pmod{p^2}$, which is known as Wolstenholme's theorem. Thus (using $i = 2j$)

$$\sum_{i=1}^{p-1} \frac{(-1)^i}{i} \equiv 2 \sum_{\substack{i=2 \\ i \text{ even}}}^{p-1} \frac{1}{i} \equiv \sum_{j=1}^{(p-1)/2} \frac{1}{j} \pmod{p^2}. \qquad \blacksquare$$

Turning to the terms in Morley's congruence, first note that

$$\begin{aligned} \binom{p}{i} &= \frac{p \cdot (p-1) \cdot (p-2) \cdots (p-(i-1))}{i \cdot 1 \cdot 2 \cdots (i-1)} \\ &= (-1)^{i-1} \cdot \frac{p}{i} \cdot \left(1 - \frac{p}{1}\right) \cdot \left(1 - \frac{p}{2}\right) \cdots \left(1 - \frac{p}{i-1}\right). \end{aligned} \tag{2}$$

Thus $\binom{p}{i} \equiv (-1)^i \cdot \left(-\frac{p}{i} + p^2 \cdot \sum_{j=1}^{i-1} \frac{1}{ij}\right) \pmod{p^3}$ and so $2^p = 2 + \sum_{i=1}^{p-1} \binom{p}{i}$ gives

$$2^{p-1} \equiv 1 - \frac{p}{2} \cdot \sum_{i=1}^{p-1} \frac{(-1)^i}{i} + \frac{p^2}{2} \cdot \sum_{0<j<i<p} \frac{(-1)^i}{ij} \pmod{p^3}.$$

Squaring, and using Lemma 3, we have

$$2^{2p-2} \equiv 1 - p \cdot \sum_{i=1}^{(p-1)/2} \frac{1}{i} + p^2 \cdot \left(\frac{1}{4} \left(\sum_{i=1}^{(p-1)/2} \frac{1}{i} \right)^2 + \sum_{0<j<i<p} \frac{(-1)^i}{ij} \right) \pmod{p^3}. \tag{3}$$

From (2) we also have

$$(-1)^{i-1} \binom{p-1}{i-1} = (-1)^{i-1} \frac{i}{p} \binom{p}{i} = \left(1 - \frac{p}{1}\right) \cdot \left(1 - \frac{p}{2}\right) \cdots \left(1 - \frac{p}{i-1}\right).$$

Taking $i = \frac{p+1}{2}$ gives

$$(-1)^{\frac{p-1}{2}} \cdot \binom{p-1}{(p-1)/2} \equiv 1 - p \cdot \sum_{i=1}^{(p-1)/2} \frac{1}{i} + p^2 \cdot \sum_{0<j<i\le\frac{p-1}{2}} \frac{1}{ij} \pmod{p^3},$$

or equivalently, using Lemma 2,

$$(-1)^{\frac{p-1}{2}} \cdot \binom{p-1}{(p-1)/2} \equiv 1 - p \cdot \sum_{i=1}^{(p-1)/2} \frac{1}{i} + \frac{p^2}{2} \cdot \left(\sum_{i=1}^{(p-1)/2} \frac{1}{i} \right)^2 \pmod{p^3}. \tag{4}$$

Comparing (3) and (4), we observe that Morley's congruence is therefore valid mod p^2. In order to obtain it mod p^3, it suffices to prove that

$$\frac{1}{4} \left(\sum_{i=1}^{(p-1)/2} \frac{1}{i} \right)^2 \equiv \sum_{0<j<i<p} \frac{(-1)^i}{ij} \pmod{p},$$

or equivalently,

$$\left(\sum_{\substack{0<i<p \\ i \text{ even}}} \frac{1}{i} \right)^2 \equiv \sum_{0<j<i<p} \frac{(-1)^i}{ij} \pmod{p}. \tag{5}$$

The considerations so far have reduced Morley's congruence modulo p^3 to a congruence modulo p.

Completion of the proof

In the remainder of the proof, all congruences are taken modulo p. First notice that as

$$\sum_{\substack{0<i<p\\ i\text{ even}}} \frac{1}{i} = \sum_{\substack{0<j<p\\ j\text{ odd}}} \frac{1}{p-j} \equiv -\sum_{\substack{0<j<p\\ j\text{ odd}}} \frac{1}{j},$$

the left hand side of (5) is

$$\Bigg(\sum_{\substack{0<i<p\\ i\text{ even}}} \frac{1}{i}\Bigg)^2 \equiv -\Bigg(\sum_{\substack{0<i<p\\ i\text{ even}}} \frac{1}{i}\Bigg)\Bigg(\sum_{\substack{0<j<p\\ j\text{ odd}}} \frac{1}{j}\Bigg) \equiv -\sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij} - \sum_{\substack{0<j<i<p\\ i\text{ odd, }j\text{ even}}} \frac{1}{ij}.$$

On the other hand,

$$\sum_{\substack{0<j<i<p\\ i,j\text{ odd}}} \frac{1}{ij} = \sum_{\substack{0<i<j<p\\ i,j\text{ even}}} \frac{1}{(p-i)(p-j)} \equiv \sum_{\substack{0<i<j<p\\ i,j\text{ even}}} \frac{1}{ij}$$

and so the right hand side of (5) is

$$\begin{aligned}
\sum_{0<j<i<p} \frac{(-1)^i}{ij} &= \Bigg(\sum_{\substack{0<j<i<p\\ i,j\text{ even}}} \frac{1}{ij} - \sum_{\substack{0<j<i<p\\ i,j\text{ odd}}} \frac{1}{ij}\Bigg) - \sum_{\substack{0<j<i<p\\ i\text{ odd, }j\text{ even}}} \frac{1}{ij} + \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij}\\
&\equiv (0) - \sum_{\substack{0<j<i<p\\ i\text{ odd, }j\text{ even}}} \frac{1}{ij} - \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij} + 2\sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij}\\
&\equiv \Bigg(\sum_{\substack{0<i<p\\ i\text{ even}}} \frac{1}{i}\Bigg)^2 + 2\sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij}.
\end{aligned}$$

Hence (5) follows from Lemma 1, and so the proof of Lemma 1 is our final task.

Proof of Lemma 1. In this calculation we make the substitutions $j = i - k$ and $i = p + j - \ell$ respectively, and then return to the original index letters and add.

$$\begin{aligned}
3\sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij} &= \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij} + \sum_{\substack{0<k<i<p\\ i\text{ even, }k\text{ odd}}} \frac{1}{i(i-k)} + \sum_{\substack{0<j<\ell<p\\ \ell\text{ even,}j\text{ odd}}} \frac{1}{(p+j-\ell)j}\\
&= \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{ij} + \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{i(i-j)} + \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{1}{(p+j-i)j}\\
&= \sum_{\substack{0<j<i<p\\ i\text{ even, }j\text{ odd}}} \frac{p}{j(i-j)(p+j-i)} \equiv 0 \pmod{p}.
\end{aligned}$$

This gives the desired result, as $p > 3$. ■

The connection with Skula's conjecture

Consider the Fermat quotient $F = \frac{2^{p-1}-1}{p}$, and note that

$$2^{2p-2} = 1 + 2Fp + F^2p^2. \tag{6}$$

Adopting the notation of [**5**], set

$$q(x) = \frac{x^p - (x-1)^p - 1}{p}, \qquad g(x) = \sum_{i=1}^{p-1} \frac{x^i}{i}, \qquad G(x) = \sum_{i=1}^{p-1} \frac{x^i}{i^2}.$$

Note that $F = q(2)/2$. The following remarkable identity was established in [**5**]:

$$-G(x) \equiv \frac{1}{p}(q(x) + g(1-x)) \pmod{p}, \tag{7}$$

from which Granville deduced Skula's conjecture: $F^2 \equiv -G(2) \pmod{p}$. From (7),

$$2F \equiv -g(-1) - G(2)p \equiv -g(-1) + F^2p \pmod{p^2}.$$

Hence, substituting in (6), we obtain

$$2^{2p-2} = 1 + 2Fp + F^2p^2 \equiv 1 - g(-1)p + \frac{1}{2}g(-1)^2p^2 \pmod{p^3}. \tag{8}$$

From Lemma 3, $g(-1) \equiv \sum_{i=1}^{\frac{p-1}{2}} \frac{1}{i} \pmod{p^2}$, and so from (8)

$$2^{2p-2} \equiv 1 - \left(\sum_{i=1}^{\frac{p-1}{2}} \frac{1}{i}\right) p + \frac{1}{2}\left(\sum_{i=1}^{\frac{p-1}{2}} \frac{1}{i}\right)^2 p^2 \pmod{p^3}.$$

Together with (4), this gives Morley's congruence once again.

REFERENCES

1. Alexander Bogomolny, *Cut the knot*, http://www.cut-the-knot.org/triangle/Morley/.
2. Alain Connes, A new proof of Morley's theorem, *Les relations entre les mathématiques et la physique théorique*, Inst. Hautes Études Sci., Bures, 1998.
3. H. S. M. Coxeter, *Introduction to Geometry*, Wiley Classics Library, John Wiley, New York, 1989, Reprint of the 1969 edition.
4. Hansjörg Geiges, Beweis des Satzes von Morley nach A. Connes, *Elem. Math.* **56**(4) (2001) 137–142. http://dx.doi.org/10.1007/PL00000548.
5. Andrew Granville, The square of the Fermat quotient, *Integers* **4** (2004) A22, 3 pp. (electronic), http://www.emis.de/journals/INTEGERS/papers/e22/e22.pdf.
6. Liang-shin Hahn, *Complex Numbers and Geometry*, MAA Spectrum, Mathematical Association of America, Washington, DC, 1994.
7. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, fifth ed., The Clarendon Press Oxford University Press, New York, 1979.
8. Emma Lehmer, On congruences involving Bernoulli numbers and the quotients of Fermat and Wilson, *Ann. Math.* **39**(2) (1938) 350–360. http://dx.doi.org/10.2307/1968791.
9. F. Morley, Note on the congruence $2^{4n} \equiv (-)^n(2n)!/(n!)^2$, where $2n+1$ is a prime, *Ann. Math.* **9**(1–6) (1894/95) 168–170. http://dx.doi.org/10.2307/1967516.
10. D. J. Newman, The Morley miracle, *Math. Intelligencer* **18**(1) (1996) 31–34. http://dx.doi.org/10.1007/BF03024813.
11. Niels Nielsen, Recherches sur les nombres de Bernoulli, *Danske Vidensk. Selsk. Skr.* (7) **10** (1913) 285–366.
12. Gerhard Wanner, Elementare Beweise des Satzes von Morley, *Elemente der Mathematik* **59**(4) (2004) 144–150. http://dx.doi.org/10.1007/s00017-003-0194-z.

Summary Frank Morley is famous for his theorem concerning the angle trisectors of a triangle. This note gives an elementary proof of another result of Morley's, which relates the middle binomial coefficient to a certain power of two. The striking thing about Morley's congruence is that it is valid modulo the third power of the prime being considered.

Integration by Parts without Differentiation

VICENTE MUÑOZ
Universidad Complutense de Madrid
Madrid, Spain
vicente.munoz@mat.ucm.es

The usual rule of integration by parts taught in calculus states that

$$\int_a^b FG'\,dt = FG\Big|_a^b - \int_a^b F'G\,dt, \tag{1}$$

for differentiable functions $F, G : [a, b] \to \mathbb{R}$. This is a straightforward consequence of the product rule for derivatives: $(FG)' = F'G + FG'$, and a fundamental theorem of calculus:

$$FG\Big|_a^b = \int_a^b (FG)'\,dt = \int_a^b (F'G + FG')\,dt. \tag{2}$$

If

$$F(x) = c_1 + \int_a^x f(t)\,dt \qquad \text{and} \qquad G(x) = c_2 + \int_a^x g(t)\,dt \tag{3}$$

for continuous functions f, g on $[a, b]$ and constants c_1, c_2, then F, G have derivatives $F' = f$, and $G' = g$ (by the other fundamental theorem of calculus). Thus equation (2) can be written as

$$FG\Big|_a^b = \int_a^b (fG + Fg)\,dt. \tag{4}$$

However, if f, g are not continuous and F, G are not differentiable everywhere, then the product rule for derivatives cannot be used directly to prove integration by parts. It is interesting to note that the more general version in equation (4) can be proved without using differentiability. All that is needed is that f, g are "integrable" functions on $[a, b]$. Moreover, the proof is simple and accessible to undergraduates.

For showing (4) for integrable functions $f, g : [a, b] \to \mathbb{R}$, where F, G are defined in (3), we take two steps. First we assume that $c_1 = c_2 = 0$. Then we perform the following integration:

$$\begin{aligned}\int_a^b (Fg + Gf)\,dt &= \int_a^b g(t)\left(\int_a^t f(s)\,ds\right)dt + \int_a^b f(t)\left(\int_a^t g(s)\,ds\right)dt \\ &= \int_a^b \int_a^y f(x)g(y)\,dx\,dy + \int_a^b \int_a^x f(x)g(y)\,dy\,dx\end{aligned}$$

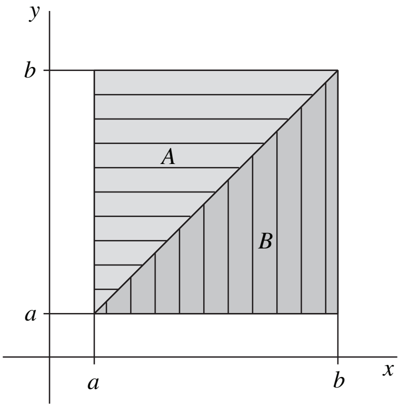
Math. Mag. **85** (2012) 211–213. doi:10.4169/math.mag.85.3.211. © Mathematical Association of America

$$= \int_a^b \int_a^b f(x)g(y)\,dx\,dy$$

$$= \left(\int_a^b f(x)\,dx\right)\left(\int_a^b g(y)\,dy\right)$$

$$= F(b)\,G(b) = FG\Big|_a^b$$

In the above, the first equality is by definition of F, G. In the second equality we have set $x = s$, $y = t$ in the first summand, and $x = t$, $y = s$ in the second summand. In the third equality we have added up the integration regions as in the following figure. The other equalities are clear.



For the second step, let c_1, c_2 be any real numbers. Then let $F = \tilde{F} + c_1$, $G = \tilde{G} + c_2$, with $\tilde{F}(x) = \int_a^x f(t)\,dt$, $\tilde{G}(x) = \int_a^x g(t)\,dt$, which are already known to satisfy the formula $\tilde{F}\,\tilde{G}\big|_a^b = \int_a^b (\tilde{F}g + \tilde{G}f)\,dt$. Then

$$\begin{aligned}\int_a^b (Fg + Gf)\,dt &= \int_a^b ((\tilde{F} + c_1)\,g + (\tilde{G} + c_2)\,f)\,dt \\ &= \int_a^b (\tilde{F}\,g + \tilde{G}\,f)\,dt + c_1 \int_a^b g + c_2 \int_a^b f \\ &= \tilde{F}\,\tilde{G}\Big|_a^b + c_1(G(b) - G(a)) + c_2(F(b) - F(a)) \\ &= (F(b) - F(a))\,(G(b) - G(a)) + F(a)\,(G(b) - G(a)) \\ &\quad + G(a)\,(F(b) - F(a)) \\ &= F(b)G(b) - F(a)G(a) = FG\Big|_a^b\end{aligned}$$

The fourth equality uses $\tilde{F}(b) = F(b) - c_1 = F(b) - F(a)$, $\tilde{G}(b) = G(b) - c_2 = G(b) - G(a)$, and $c_1 = F(a)$, $c_2 = G(a)$.

Some comments. If f, g are Riemann integrable, then they are bounded and they are continuous almost everywhere. More generally, the result holds if f, g are Lebesgue integrable. In this case, the functions F, G are absolutely continuous and therefore differentiable almost everywhere. In fact, a function F on $[a, b]$ is absolutely continuous if and only if it has the form $F(x) = c + \int_a^x f(t)\,dt$ for some Lebesgue integrable function f and constant c. (See Theorem 3.36 in [**2**], Theorem 6.85 in [**5**] or Theorem 18.17 in [**4**].)

Equation (4) for Lebesgue integrable functions f, g is Theorem 18.19 in [**4**]. They show that FG is also absolutely continuous, and use the product rule for the derivative of FG, which works almost everywhere. Their Corollary 18.20 gives the integration by parts formula (1) for absolutely continuous functions F, G. This is also Theorem 6.90 in [**5**], Exercise 7.9.2 in [**1**], and Exercise 35, §3.5 in [**2**].

One can also look at the derivatives of functions F, G as *distributions* (see [**3**, chapter 6] for foundational material on the theory of distributions). The derivative of F is a distribution $D(F)$ satisfying

$$\int_a^b D(F)\,h = Fh|_a^b - \int_a^b F(t)\,h'(t)\,dt \tag{5}$$

for any C^∞ function $h : [a, b] \to \mathbb{R}$. If $F(x) = c + \int_a^x f(t)\,dt$, then $D(F) = f$ almost everywhere. This follows from the integration by parts rule that we have proved, applied to f, h', and the defining equation (5). By what we said above, for absolutely continuous functions F, G, there is a Leibniz rule for derivatives $D(FG) = D(F)\,G + F\,D(G)$. This gives the result, using the rule $\int_a^b D(F) = F|_a^b$ (which is just equation (5) with $h = 1$).

REFERENCES

1. A. M. Bruckner, J. B. Bruckner, and B. S. Thomson, *Real Analysis*, Prentice-Hall, 1996.
2. G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., Wiley-Interscience, 1999.
3. W. Rudin, *Functional Analysis*, 2nd ed., McGraw-Hill, 1991.
4. E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, 1965.
5. K. R. Stromberg, *An Introduction to Classical Real Analysis*, Wadsworth, 1981.

Summary We prove the classical integration-by-parts formula without using differentiation. This provides an elementary illustration that integration by parts is valid for many non-differentiable functions. The proof is simple and accessible to undergraduates.

Two Cevians Intersecting on an Angle Bisector

VICTOR OXMAN
The Western Galilee College
Acre, Israel
victor.oxman@gmail.com

Let ABC be any triangle with cevians AA_1, BB_1 intersecting at a point O on the angle bisector CC_1 (FIGURE 1). If O happens to be the incenter of the triangle, then AA_1 and BB_1 are also angle bisectors and the classical Steiner-Lehmus theorem [**1**] tells us that if they are equal, then the triangle is isosceles.

But what if O is some other point of the angle bisector CC_1? Then does $AA_1 = BB_1$ still imply that $AC = BC$? The following theorem gives a positive answer to this

Math. Mag. **85** (2012) 213–215. doi:10.4169/math.mag.85.3.213. © Mathematical Association of America

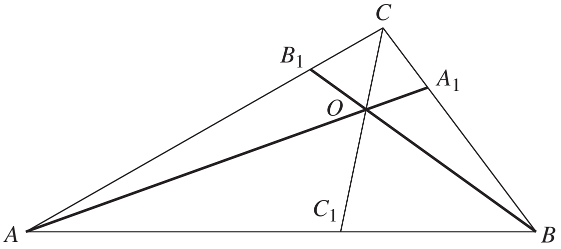


Figure 1 CC_1 bisects angle C

question. In this way we generalize the Steiner-Lehmus theorem, and we give it a new, purely geometric proof as well.

THEOREM. *Let A_1, B_1, C_1 be points on the sides of triangle ABC, and let CC_1 bisect angle C. If the cevians AA_1, BB_1, and CC_1 are concurrent and $AA_1 = BB_1$, then $AC = BC$ and the triangle is isosceles.*

Proof. The proof depends on a lemma, which is a congruence result in the spirit of SAS, SSS, etc.

LEMMA. *If a side, its opposite angle and the bisector of that angle in triangle are respectively equal to a side, its opposite angle and the bisector of that angle in another triangle, then the triangles are congruent.*

Proof. For definiteness we can assume that the triangles have the common side AB, the opposite vertices C and C' reside on the same halfplane in relation to line AB, angles $\angle ACB$ and $\angle AC'B$ are equal, and the bisector CD of $\angle ACB$ is equal to the bisector $C'D'$ of $\angle AC'B$. It is well known that the locus of all the points on a halfplane, from which a given segment AB is seen at a given angle, is an arc of the circle passing through points A and B (FIGURE 2).

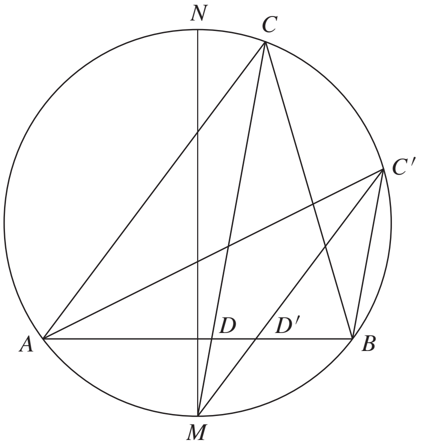


Figure 2 Proof of the Lemma

Let MN be the circle diameter perpendicular to AB. Let us suppose that C and C' are not the same point. Without loss of generality we can assume that C and C' reside on arc NB in order N, C, C', B. Then the lines CD and $C'D'$ intersect the circle at point M, and $D'C' < DC$ because

(1) The arc MC' is smaller than the arc MC, forcing $MC' < MC$; and

(2) The angle $\angle MDD'$ of $\triangle MDD'$ is obtuse, so $MD < MD'$.

Thus we get a contradiction with $CD = C'D'$. So C and C' are the same point and triangles ACB and $AC'B$ are congruent. ■

Now returning to the proof of the theorem we can conclude that for $AA_1 = BB_1$ we have $\triangle BCB_1 \cong \triangle ACA_1$ and so $BC = AC$. ■

With some extra care one can prove that if the two cevians AA_1 and BB_1 intersect on an angle bisector CC_1 and $AA_1 > BB_1$, then $AC > BC$.

REFERENCES

1. H. S. M. Coxeter, *Introduction to Geometry*, John Wiley, New York, 1969.

Summary We prove the next generalization of the Steiner-Lehmus theorem: if two equal cevians intersect with each other on the angle bisector of the third triangle vertex, then the triangle is isosceles.

How Rare Are Subgroups of Index 2?

JEAN B. NGANOU
University of Oregon
Eugene, OR 97403
nganou@uoregon.edu

The first and most famous counterexample to the converse of Lagrange's Theorem is the alternating group A_4, which consists of the even permutations on 4 letters. The group has order 12, but it has no subgroup of order six. Equivalently, it has no subgroup of index 2. Brennan and MacHale gave a proof of this fact in this MAGAZINE [**1**], and then ten more proofs for good measure. That article motivated us to learn more about the existence of subgroups of index 2 in other groups. More precisely, we considered the questions of the essential ingredients that drive the arguments in [**1**] and whether they can be generalized to other groups.

In this paper we consider especially the role played by the subgroup of squares in determining whether a group has a subgroup of index 2. Using this approach we identify a larger class of groups that do not have subgroups of index 2; that is, a larger class of counterexamples to Lagrange's Theorem.

Subgroups of index 2 are of special interest because they are always normal. In fact, if H has index 2 in G then H has only one coset other than itself. That is enough to force $Ha = aH$ for every $a \in G$, which makes H normal in G.

We hope that this paper is accessible to readers with very little background. We do not use any results or concepts beyond the first part of Gallian's textbook [**3**].

All groups in this article are finite.

Subgroups to watch for

We start by defining the subgroup of squares and verifying its main properties. Recall that if X is any subset of a group G, the subgroup $\langle X\rangle$ of G generated by X is the smallest subgroup of G containing X. It easy to verify that $\langle X\rangle$ consists of the identity

Math. Mag. **85** (2012) 215–220. doi:10.4169/math.mag.85.3.215. © Mathematical Association of America

element, together with the set of elements that are formed from "finite words" in the elements of X:

$$\langle X\rangle = \{e\} \cup \{x_1 \cdots x_n \mid x_i \in X; n \geq 1\}$$

In fact, it is easily seen that the set on the right is non-empty and closed under multiplication. The closure under inverses follows from $(x_1 \cdots x_n)^{-1} = x_n^{-1} \cdots x_1^{-1}$, and the fact that since G is finite each x_i^{-1} is equal to $x_i^{k_i}$ for some $k_i \geq 1$. Thus, by the two-step subgroup test [**3**, Theorem 3.2] the set is a subgroup of G.

Suppose in addition that X is closed under conjugation; that is for every $x \in X$ and $a \in G$, $axa^{-1} \in X$. Then for every $a \in G$, and $x_1, x_2, \ldots, x_n \in X$,

$$a(x_1x_2 \cdots x_n)a^{-1} = (ax_1a^{-1})(ax_2a^{-1}) \cdots (ax_na^{-1}) \in \langle X\rangle.$$

Hence $\langle X\rangle$ is normal in G. Therefore, if X is closed under conjugation, then $\langle X\rangle$ is a normal subgroup of G.

A close look at some of the proofs in [**1**] motivates the consideration of the subgroup generated by the squares and subgroup generated by elements of odd orders.

For any group G, we denote by G^2 the subgroup of G generated by squares of elements in G, that is $G^2 = \langle\{x^2 : x \in G\}\rangle$. We say that G is generated by squares if $G = G^2$.

Note that G^2 is normal in G. In fact, from the observation above, it is enough to justify that the set of squares in G is closed under conjugation. But this is clear because, for every $x, a \in G$, $ax^2a^{-1} = (axa^{-1})^2$.

For any group G, we denote by G^{odd} the subgroup of G generated by elements of odd order in G. Since every element a of odd order satisfies the equation $a = a^{2k}$ for some integer k, then G^{odd} is a subgroup of G^2. Note that G^{odd} can be a proper subgroup of G^2, for instance the subgroup of $\mathbb{Z}_4$ generated by elements of odd order is trivial while its subgroup of squares is $2\mathbb{Z}_4$.

Just like G^2, the subgroup G^{odd} is normal in G. This follows from the observation above and the fact that orders of elements are preserved under conjugation.

How many subgroups of index 2?

By Lagrange's Theorem, groups of odd order do not have subgroups of index 2, therefore only groups of even orders are relevant here. To answer our question on the existence of subgroups index 2, we consider a more general question: How many subgroups of index 2 are there in the group? The following key result shifts the question to one about finite vector spaces over $\mathbb{Z}_2$, where we have better tools to answer the question.

THEOREM 1. *The groups G and G/G^2 have the same number of subgroups of index* 2.

To see this, we start by the following observation. If H is a subgroup of index 2 in G, then H is normal in G and the factor group G/H has order 2. Therefore, by a consequence of the Lagrange Theorem [**2**, Corollary 7.4], $(xH)^2 = H$ for all $x \in G$. Hence, $x^2 \in H$ for all $x \in G$, and it follows that $G^2 \subseteq H$. Since G^2 is normal in G, then it is normal in H and we can consider the factor group $\mathcal{H} := H/G^2$. This is a subgroup of G/G^2 and a simple calculation shows that $[G/G^2 : \mathcal{H}] = [G : H] = 2$. We define a map φ from the set of subgroups of G of index 2 to the set of subgroups of G/G^2 of index 2 by $\varphi(H) = \mathcal{H}$. This is well defined and we leave it as an exercise to verify that it is a bijection by showing that its inverse is defined as follows. Given a subgroup $\mathcal{H}$ of G/G^2, then $\varphi^{-1}(\mathcal{H}) := \{x \in G : xH \in \mathcal{H}\}$. Since there is a bijection between the set of subgroups of index 2 in G and the set of subgroups of index 2 in G/G^2, then these sets have the same cardinality as claimed. ■

We now examine the structure of the quotient group G/G^2.

Every group G satisfying $x^2 = e$ for all $x \in G$ is Abelian. In fact, if $x^2 = e$ for all $x \in G$, then $x^{-1} = x$ for all $x \in G$. Therefore, for every $a, b \in G$, $ab = (ab)^{-1} = b^{-1}a^{-1} = ba$.

It follows from this observation that G/G^2 is Abelian since $(xG^2)^2 = G^2$ for all $x \in G$. Therefore, G/G^2 is a finite Abelian group in which every element other than the identity has order 2. Thus by the Fundamental Theorem of Finite Abelian groups, G/G^2 is isomorphic to a direct product of cyclic groups. But, since every non-identity in G/G^2 has order 2, then the only cyclic group that can show up as a factor in its direct product decomposition is $\mathbb{Z}_2$. That means there exists an integer $n \geq 0$ such that

$$G/G^2 \cong \bigoplus_{i=1}^{n} \mathbb{Z}_2 \tag{1}$$

The right-hand of (1) is to be interpreted as the trivial group when $n = 0$ and this situation arises exactly when $G = G^2$.

In addition to being a group, $\bigoplus_{i=1}^{n} \mathbb{Z}_2$ has a natural structure of a vector space over $\mathbb{Z}_2$ with vector addition being the usual group addition and scalar multiplication defined in the natural way across components.

For readers that are not familiar with vector spaces over fields other than $\mathbb{R}$ or $\mathbb{C}$, it should be noted that the axioms remain the same. For instance, if we require the same axioms of vector spaces over $\mathbb{R}$, but use $\mathbb{Z}_2$ as the set of scalars, we obtain the notion of vector spaces over $\mathbb{Z}_2$. The definitions of concepts such as basis, dimension and subspace are the same. Moreover, every n-dimensional vector space V over $\mathbb{Z}_2$ is isomorphic to $\bigoplus_{i=1}^{n} \mathbb{Z}_2$ where the isomorphism is obtained by taking coordinates with respect to some fixed basis; from which it follows that V is finite as a set and $|V| = 2^n$.

Moreover, the subspaces and the subgroups of $\bigoplus_{i=1}^{n} \mathbb{Z}_2$ coincide. In fact, it is clear that a subspace is a subgroup and that every subgroup is closed under addition. On the other hand, every subgroup is closed under scalar multiplication because there are only two scalars, 0, 1. A multiplication by 0 yields the zero vector, while a multiplication by 1 has no effect. Finally, since a k-dimensional subspace has order 2^k, we see that subgroups of index 2 (and so order 2^{n-1}) must correspond to subspaces of dimension $n - 1$. Recall that an $(n - 1)$-dimensional subspace of an n-dimensional vector space V is called a hyperplane of V. We now count the hyperplanes of finite dimensional vector spaces over $\mathbb{Z}_2$.

THEOREM 2. *Every n-dimensional vector space over $\mathbb{Z}_2$ has exactly $2^n - 1$ hyperplanes.*

Without loss of generality, we may assume that $V = \bigoplus_{i=1}^{n} \mathbb{Z}_2 = \mathbb{Z}_2^n$. Using simple combinatorics and elementary linear algebra, we show that there are $2^n - 1$ hyperplane(s) in V.

First we count the ordered linearly independent sets of $n - 1$ vectors in $\mathbb{Z}_2^n$. To construct such a set, we need to carefully choose $n - 1$ vectors $v_1, \ldots, v_{n-1}$.

For the choice of v_1, the only vector to avoid is the zero vector of $\mathbb{Z}_2^n$, leaving us with $2^n - 1$ ways of choosing v_1. Once v_1 is chosen, we must choose $v_2 \notin \text{Span}\{v_1\} = \{0, v_1\}$, and there are $2^n - 2$ such choices. Suppose we have chosen v_1 through v_k, then we need to choose $v_{k+1} \notin \text{Span}\{v_1, \ldots, v_k\}$. An ordered linearly independent set of k vectors spans a subspace containing 2^k vectors and these must all be avoided if one wants to extend the ordered set and preserve linear independence. It follows that there are $2^n - 2^k$ choices for the $(k + 1)$st vector, thus there are $2^n - 2^k$ possibilities. In total, we have $(2^n - 1)(2^n - 2)\cdots(2^n - 2^{n-2})$ linearly independent subsets of $n - 1$ vectors. Similarly, since each hyperplane of $\mathbb{Z}_2^n$ contains 2^{n-1} vectors, the same argument above shows that each hyperplane of $\mathbb{Z}_2^n$ has $(2^{n-1} - 1)(2^{n-1} - 2)\cdots(2^{n-1} - 2^{n-2})$ linearly independent subsets with $n - 1$ vectors. Each of these sets is the basis of a hyperplane, but different bases may give rise to the same hyperplane. In fact, the argument above shows that each hyperplane in $\mathbb{Z}_2^n$ has $(2^{n-1} - 1)(2^{n-1} - 2)\cdots(2^{n-1} - 2^{n-2})$ distinct bases. Therefore, the number of hyperplanes in $\mathbb{Z}_2^n$ is:

$$\frac{(2^n - 1)(2^n - 2)\cdots(2^n - 2^{n-2})}{(2^{n-1} - 1)(2^{n-1} - 2)\cdots(2^{n-1} - 2^{n-2})}$$

Which simplifies to $2^n - 1$ as we needed. ■

Combining Theorem 1 and Theorem 2 we now obtain a formula for the number of subgroups of index 2 in a group G.

COROLLARY 1. *There are exactly* $2^n - 1$ *subgroups of index* 2 *in G where n is the integer in the isomorphism (1).*

By the isomorphism (1), subgroups of G/G^2 are of the same number as subgroups of $\mathbb{Z}_2^n$. From the discussion preceding Theorem 2, subgroups of G/G^2 correspond to subspaces of $\mathbb{Z}_2^n$ and subgroups of index 2 in G/G^2 correspond to subspaces of $\mathbb{Z}_2^n$ of dimension $n - 1$ (the hyperplanes). By Theorem 2, $\mathbb{Z}_2^n$ has $2^n - 1$ hyperplanes. Therefore, there are $2^n - 1$ subgroup(s) of index 2 in G/G^2, and by Theorem 1, we conclude that there are $2^n - 1$ subgroups of index 2 in G. ■

It is worth pointing out that an alternate proof of the preceding result that uses more advanced tools can be found in [**2**]. In addition, readers interested in more details on counting subspaces of a fixed dimension can consult [**6**], [**7**]. The following results are easy consequences of Corollary 1.

COROLLARY 2. *A group has no subgroup of index* 2 *if and only if it is generated by squares.*

COROLLARY 3. *A group G has a unique subgroup of index* 2 *if and only if* G^2 *has index* 2 *in G.*

Since $G^{\text{odd}} \subseteq G^2$, it follows from the above Corollaries and Lagrange's Theorem that if more than half of the elements in G have odd order, then G has no subgroups of index 2. (Brennan and MacHale present this result as a theorem [**1**].) It is worth pointing out that this condition on elements of odd order is not a necessary condition for a group to have no subgroups of index 2. For instance, the linear group $SL(2, 3)$ of 2×2 matrices in $\mathbb{Z}_3$ with determinant 1 is a group of order 24 that has no subgroups of index 2, but has only nine elements of odd order [**5**].

Applications

We present a few applications of the results from the previous section.

EXAMPLE. For every $n \geq 2$, the alternating group A_n on n symbols has no subgroups of index 2.

This fact follows (when $n \geq 5$) from the well known fact that A_n is a simple group for $n \geq 5$, but we prefer a direct proof.

By Corollary 2, it is enough to show that A_n is generated by squares. For this, recall that every even permutation is a product of 3-cycles [**2**, Ex. 5.47]. Therefore, it is enough to show that every 3-cycle is in A_n^2, but this is clear because if α is a 3-cycle, then $\alpha = \alpha^{-2} = (\alpha^{-1})^2$ and $\alpha^{-1} \in A_n$.

As another application of the Corollaries, we prove that the alternating group A_n is the only subgroup of index 2 in the symmetry groups S_n.

EXAMPLE. S_n has a unique subgroup of index 2, which is A_n.

By Corollary 3 it is enough to show that $A_n = S_n^2$. As seen in the previous example, $A_n = A_n^2$ and since $A_n^2 \subseteq S_n^2$, then $A_n \subseteq S_n^2$. Conversely, since the square of every permutation is an even permutation, then $S_n^2 \subseteq A_n$. Therefore, $A_n = S_n^2$ as required.

We want to apply the results of the previous section to the dihedral groups D_n ($n > 2$). We recall the definition of the dihedral groups. Let $\mathrm{Isom}(\mathbb{R}^2)$ be the set of isometries of $\mathbb{R}^2$ (also thought of as the xy-plane), then $\mathrm{Isom}(\mathbb{R}^2)$ is a group under the composition of maps. Let R be the rotation counter-clockwise in $\mathbb{R}^2$ of $360°/n$ about the origin and S be the reflection about the x-axis. Then D_n is the subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ generated by $\{R, S\}$ in the sense discussed in the introductory paragraphs. Since $R^n = id$ and $RS = SR^{n-1}$, it is easy to see that $D_n = \{S^i R^j : i = 0, 1;\ j = 0, 1, \dots, n-1\}$, in particular D_n has order $2n$. We know that the subgroup $\langle R\rangle$ of rotations has index 2 in D_n. Therefore, we are interested in knowing if there are other subgroups of index 2 or if this is the only one. We have the following result.

THEOREM 3. *D_n has a unique subgroup of index* 2 *if n is odd and three subgroups of index* 2 *when n is even.*

With the description of D_n preceding the theorem, one sees that $D_n^2 = \langle R^2\rangle$. In fact the elements outside of $\langle R\rangle$ have the form SR^i and are reflections of order 2, which means that when they are squared they do not contribute any non-trivial elements to the subgroup D_n^2.

On the other hand, recall that in every group, if a is an element of order $o(a) = m$, then for every integer $k > 0$, $o(a^k) = m/\gcd(m, k)$ [**3**, Thm. 4.2]. Since $o(R) = n$, we have

$$o(R^2) = \begin{cases} n/2 & \text{if } n \text{ is even} \\ n & \text{if } n \text{ is odd} \end{cases}$$

Thus,

$$|D_n/D_n^2| = |D_n/\langle R^2\rangle| = \frac{2n}{o(R^2)} = \begin{cases} 4 & \text{if } n \text{ is even} \\ 2 & \text{if } n \text{ is odd} \end{cases}$$

And the isomorphism (1) for the group D_n corresponds to:

$$D_n/D_n^2 \cong \begin{cases} \mathbb{Z}_2 \bigoplus \mathbb{Z}_2 & \text{if } n \text{ is even} \\ \mathbb{Z}_2 & \text{if } n \text{ is odd} \end{cases}$$

The result is then clear from either Theorem 1 or Theorem 2.

The following result (whose easy proof we omit) leads to a large class of counterexamples to the converse of the Lagrange's Theorem.

THEOREM 4. *For every group G_1 and G_2,*

$$(G_1 \bigoplus G_2)^2 = G_1^2 \bigoplus G_2^2$$

Now it follows from Corollary 2 and Theorem 3 that if neither G_1 nor G_2 has a subgroup of index 2, then $G_1 \bigoplus G_2$ does not have a subgroup of index 2. In particular, for every $n \geq 4$ and every odd natural number m, the group $A_n \bigoplus \mathbb{Z}_m$ does not have a subgroup of index 2. This provides a counterexample of order $\frac{n!m}{2}$ (when $n \geq 4$ and $m \geq 1$ odd) to the converse of Lagrange's Theorem. Other counterexamples are $SL(2, 3) \bigoplus A_n$ with $n \geq 4$ and $SL(2, 3) \bigoplus \mathbb{Z}_n$ with n odd.

Note that using the fact that more than half the elements of A_4 have odd order, the authors of [**1**] were able to conclude that there are counterexamples of order $12m$ with m odd to Lagrange's Theorem. Unfortunately, it is not true in general that more than half the elements of A_n have odd order. For instance, A_{10} has only 893025 elements of odd orders while A_{10} has 1814400 elements. Therefore, our approach produces a class of counterexamples beyond those in [**1**].

REFERENCES

1. M. Brennan, D. MacHale, Variations on a theme: A_4 definitely has no subgroup of order six!, *Math. Mag.* **73** (2000), 36–40; http://dx.doi.org/10.2307/2691487.
2. R. R. Crawford, K. D. Wallace, On the number of subgroups of index two—An application of Goursat's Theorem for groups, *Math. Mag.* **48** (1975), 172–174; http://dx.doi.org/10.2307/2689703.
3. J. Gallian, *Contemporary Abstract Algebra,* 7th ed., Brooks/Cole, 2010.
4. D. MacHale, Minimum counterexamples in group theory, *Math. Mag.* **54** (1981), 23–28; http://dx.doi.org/10.2307/2689377.
5. G. Mackiw, The linear group $SL(2, 3)$ as a source of examples, *Math. Gazette* (March 1997), 64–67; http://dx.doi.org/10.2307/3618770.
6. A. Prasad, Counting Subspaces of a Finite Vector Space—1, *Resonance* **15** (2010), No. 11, 977–987; http://www.ias.ac.in/resonance/November2010/p977-987.pdf, http://dx.doi.org/10.1007/s12045-010-0114-5
7. ———, Counting Subspaces of a Finite Vector Space—2, *Resonance* **15** (2010), No. 12, 1074–1083; http://www.ias.ac.in/resonance/December2010/p1074-1083.pdf, http://dx.doi.org/10.1007/s12045-010-0120-7

Summary Using elementary combinatorics and linear algebra, we compute the number of subgroups of index 2 in any finite group. This leads to necessary and sufficient conditions for groups to have no subgroups of index 2, or to have a unique subgroup of index 2. Illustrative examples are provided, along with a class of counterexamples to the converse of Lagrange's Theorem.

A Timely Problem (from the cover)

First Solution. Pick's formula applies to any polygon bounded by a simple closed curve with vertices at integer lattice points (even if the edges are not all horizontal and vertical). The formula gives the area as $A = I + B/2 - 1$ where I is the number of interior lattice points and B is the number of lattice points on the boundary. In this case $I = 0$ and $B = 61 \cdot 66 = 4026$, so $A = 2012$.

Second Solution. To find the area without Pick's formula, walk around the curve in a counterclockwise direction, keeping the interior on the left. Let L denote the number of lattice points at which the curve turns left, S the number of lattice points at which it continues straight, and R the number of lattice points at which it turns to the right. Completing one circuit means turning 360° to the left, so that $L - R = 4$, and the length of the curve means that $L + S + R = 61 \cdot 66 = 4026$.

At each lattice point count the number of interior unit squares with a corner at that lattice point. If the curve turns left at a vertex, then there is 1 such square; if the curve is straight then there are 2; and if the curve turns right there are 3. Each interior unit square is counted 4 times, so that $4A = L + 2S + 3R = 2(L + S + R) - (L - R) = 2 \cdot 4026 - 4$ and $A = 2012$.

—Problem and solutions by Hugh Montgomery

Inequalities Involving Six Numbers, with Applications to Triangle Geometry

CLARK KIMBERLING
University of Evansville
Evansville, IN 47722
ck6@evansville.edu

PETER MOSES
Moparmatic Co., 1154 Evesham Road, Astwood Bank
Redditch, Worcestershire B96 6DT, England
mows@mopar.freeserve.co.uk

In a search for geometric points that minimize certain functions, we encountered a surprise. We were looking for inequalities involving “triangle numbers”—that is, numbers a, b, c that are sidelengths of a triangle, which is to say that

$$b + c > a, \quad c + a > b, \quad a + b > c. \tag{1}$$

The surprise is that for some of the inequalities that we found, the minimizations hold not only for triangles, but for many other values of a, b, c. A second surprise is that proofs of the theorems are little else than nice applications of the two-variable second-derivative test.

We begin with a lemma involving four real numbers:

LEMMA 1. *If* $-1 \le t \le 2$, *then*

$$x^2 + y^2 + z^2 + t(yz + zx + xy) \ge 0 \tag{2}$$

for all real numbers x, y, z.

Proof. For all real x, y, z, t,

$$x^2 + y^2 + z^2 + t(yz + zx + xy) = S_1 + S_2,$$

where

$$S_1 = \frac{2-t}{6}[(y-z)^2 + (z-x)^2 + (x-y)^2],$$
$$S_2 = \frac{1+t}{3}(x+y+z)^2,$$

and clearly $S_1 \ge 0$ and $S_2 \ge 0$ if $-1 \le t \le 2$, so that (2) holds. ■

In order to apply Lemma 1, some facts about trilinear coordinates will be helpful. Suppose that ABC is a triangle with sidelengths $a = |BC|$, $b = |CA|$, $c = |AB|$. For any point X in the plane of ABC, let α be the directed distance from X to the line BC; that is, $\alpha > 0$ if X lies on the same side of BC as A, but $\alpha = 0$ if $X \in BC$, and $\alpha < 0$ otherwise. Let β and γ be the directed distances from X to lines CA and AB, respectively. The position of X is then given by (α, β, γ), these being the *actual trilinear distances* of X; any numbers x, y, z respectively proportional to α, β, γ are called *homogeneous trilinear coordinates,* or simply *trilinears,* and we write $X = x : y : z$. (For more on trilinear coordinates, see [**3, 4**].)

Math. Mag. **85** (2012) 221–228. doi:10.4169/math.mag.85.3.221. © Mathematical Association of America

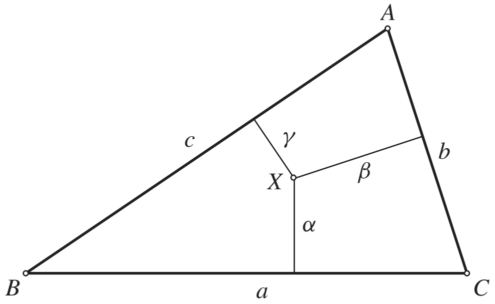


Figure 1 Trilinear distances α, β, γ

Given a point $X = x : y : z$, we can determine (α, β, γ) by noting that the area of ABC is given by

$$\sigma = (a\alpha + b\beta + c\gamma)/2, \tag{3}$$

since σ is the sum of (directed, if X lies outside ABC) areas $a\alpha/2$, $b\beta/2$, $c\gamma/2$ of the triangles BCX, CAX, ABX. Now $\alpha = hx$, $\beta = hy$, $\gamma = hz$ for some h, and substituting these expressions into (3) and solving for h gives

$$h = 2\sigma/(ax + by + cz).$$

So we have

$$\alpha = \frac{2\sigma x}{ax + by + cz}, \quad \beta = \frac{2\sigma y}{ax + by + cz}, \quad \gamma = \frac{2\sigma z}{ax + by + cz}. \tag{4}$$

(If $ax + by + cz = 0$, then $x : y : z$ lies on the line at infinity, a condition that is excluded in the sequel.)

The area of ABC is also given by Heron's formula,

$$16\sigma^2 = (a + b + c)(-a + b + c)(a - b + c)(a + b - c). \tag{5}$$

A sufficient condition for the right side in (5) to be positive is that the inequalities (1) hold.

Some special points of the triangle ABC can be defined by their trilinear coordinates. Among them are the incenter, $I = 1 : 1 : 1$, and the symmedian point, $K = a : b : c$. These and other triangle centers are indexed in a *Mathematics Magazine* article [**4**] and its extension, the Encyclopedia of Triangle Centers [**5**]. For example, I is indexed as X_1 and K as X_6.

Geometrically, I is the point of intersection of the internal angle bisectors at the vertex angles A, B, C. A construction of K is conveniently an extension of the construction of I. Simply reflect each median line in the matching angle bisector; the reflected lines concur in K. The line KI, given by the equation $(b - c)x + (c - a)y + (a - b)z = 0$, is shown in Figure 4.

A family of inequalities

We shall soon prove the fact, mentioned just above, that the symmedian point of an arbitrary triangle ABC is the point $X = \alpha : \beta : \gamma$ that minimizes the function $\alpha^2 + \beta^2 + \gamma^2$, where α, β, γ are the actual trilinear distances for X with respect to ABC. The same proof will also cover the fact that the incenter of ABC is the point

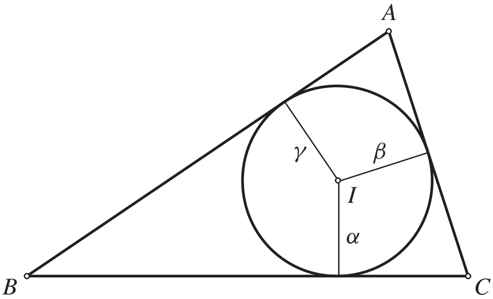


Figure 2 The incenter: $\alpha = \beta = \gamma$, so that $I = 1 : 1 : 1$.

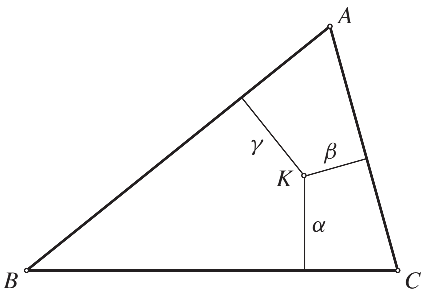


Figure 3 The symmedian point: $\alpha = a/(2\sigma)$, $\beta = b/(2\sigma)$, $\gamma = c/(2\sigma)$, so that $K = a : b : c$.

X that minimizes the function $\alpha^2 + \beta^2 + \gamma^2 - (\beta\gamma + \gamma\alpha + \alpha\beta)$. (Recently, Robert Smither [7] demonstrated a clever application of the symmedian point K that uses its minimization of the sum $\alpha^2 + \beta^2 + \gamma^2$).

These two special cases lead us to examine the more general function f given by

$$f(\alpha, \beta, \gamma) = \alpha^2 + \beta^2 + \gamma^2 + t(\beta\gamma + \gamma\alpha + \alpha\beta), \tag{6}$$

anticipating that the minimizing points X will, for several values of t, be interesting triangle centers. Solving this minimization problem means finding a point P such that $f(\alpha, \beta, \gamma) \geq f(P)$ whenever (α, β, γ) are the actual trilinear distances of a point. The problem is easily recast using trilinears $X = x : y : z$ and $P = p : q : r$ and (4): we seek P such that

$$\begin{aligned}&\frac{4\sigma^2[x^2 + y^2 + z^2 + t(yz + zx + xy)]}{(ax + by + cz)^2}\\ &\qquad \geq \frac{4\sigma^2[p^2 + q^2 + r^2 + t(qr + rp + pq)]}{(ap + bq + cr)^2}\end{aligned} \tag{7}$$

for all points X.

A solution, we shall show, is given parametrically by

$$P = 2a + t(a - b - c) : 2b + t(b - c - a) : 2c + t(c - a - b). \tag{8}$$

so that the inequality (7) holds for these values of p, q, r. The right side of (7) becomes

$$\frac{(2 - t)(t + 1)}{(t + 2)(a^2 + b^2 + c^2) - 2t(bc + ca + ab)}.$$

We have reached this point with the assumption that a, b, c are triangle numbers; that is, that they satisfy (1). But nothing about the inequality (7) forces us to make this assumption, and that opens the intriguing possibility that the inequality might hold for other values of a, b, c. This possibility leads straight to Theorem 1.

THEOREM 1. *Suppose that $-1 \leq t \leq 2$. Then the inequality*

$$\begin{aligned}&\frac{x^2 + y^2 + z^2 + t(yz + zx + xy)}{(ax + by + cz)^2}\\ &\qquad \geq \frac{(2 - t)(t + 1)}{(t + 2)(a^2 + b^2 + c^2) - 2t(bc + ca + ab)}\end{aligned} \tag{9}$$

holds for all real numbers x, y, z, a, b, c for which the denominators are not zero.

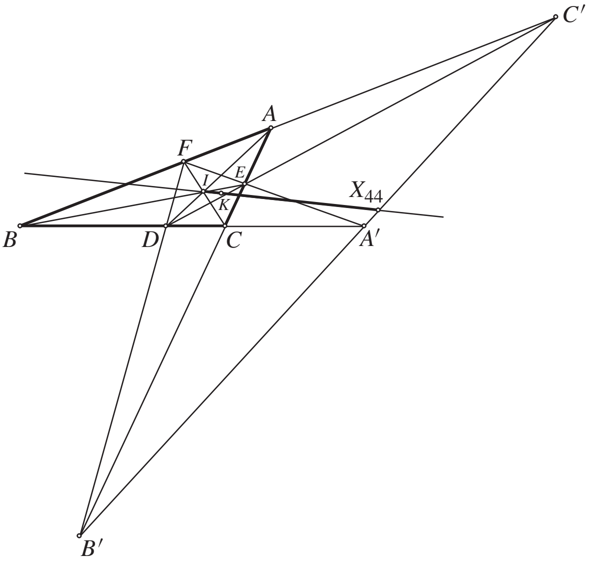


Figure 4 The lines KI and $x + y + z = 0$

Proof. It suffices to put $z = 1$ and to minimize the function

$$D = \frac{x^2 + y^2 + 1 + t(y + x + xy)}{(ax + by + c)^2}$$

obtained from the left side of (9). Setting the partial derivatives D_x and D_y equal to 0 leads directly to $x : y : z = P$ as in (8). Although the derivatives D_{xx}, D_{yy}, and D_{xy}, all evaluated at $(p/r, q/r, 1)$, are lengthy, the test-number

$$T = D_{xx}D_{yy} - D_{xy}^2,$$

is compactly expressible as

$$T = \frac{(2 - t)\,(-2c + at + bt - ct)^6}{h^5} \tag{10}$$

where $h = (t + 2)(a^2 + b^2 + c^2) - 2t(bc + ca + ab) \geq 0$ by Lemma 1; so that $h > 0$ since the denominators in (9) are not zero.

Case 1: $t \neq 2$. Here, clearly $T > 0$. Turning to $D_{xx} = D_{xx}(p/r, q/r, 1)$, we find that $D_{xx} = 2F_1^2F_2h^{-3}$, where

$$F_1 = -2c + at + bt - ct,$$
$$F_2 = a^2t^2 + b^2t + c^2t + 2b^2 + 2c^2 - 2t(bc + ca + ab).$$

Writing F_2 as $h + a^2(t - 2)(t + 1)$ shows that $F_2 > 0$. Consequently, $D_{xx} > 0$, as required.

Case 2: $t = 2$. In this case, $T = 0$, so that the second derivative test does not apply. However, the right side of (9) is 0, and the left side is

$$\frac{(x + y + z)^2}{(ax + by + cz)^2},$$

so that (9) holds. ■

More minimizers

Rewriting P in (8) as

$$\begin{aligned} P(t) = \quad & 2(t+1)a - t(a+b+c) \\ : \quad & 2(t+1)b - t(a+b+c) \\ : \quad & 2(t+1)c - t(a+b+c) \end{aligned}$$

shows that for $-\infty < t < \infty$, the locus of P is the line of the points $I = P(-1)$ and $K = P(0)$. Thus, the minimizers in Theorem 1 form a segment within the line IK, specifically, the segment from I to $X_{44} = P(2)$, corresponding to the interval $-1 \leq t \leq 2$. Next, we shall show that if a, b, c satisfy (1), then many more points on the line IK are minimizers. We begin with a lemma much like Lemma 1 in statement, if not in proof.

LEMMA 2. *If a, b, c are sidelengths of a triangle, then*

$$a^2 + b^2 + c^2 - 2(bc + ca + ab) < 0.$$

Proof.

$$\begin{aligned} & a^2 + b^2 + c^2 - 2(bc + ca + ab) \\ & \quad = -[a(b+c-a) + b(c+a-b) + c(a+b-c)]. \qquad \blacksquare \end{aligned}$$

THEOREM 2. *Suppose that $t \leq 2$, that x, y, z are real numbers, and that a, b, c are sidelengths of a triangle. Then the inequality (9) holds.*

Proof. By Theorem 1, the stated conclusion is already proved unless $t < -1$, so assume that $t < -1$. By Lemma 2,

$$(t+2)(a^2 + b^2 + c^2) - 2t(bc + ca + ab) > 0.$$

Consequently, by (10), we have $T > 0$. Also $D_{xx} > 0$ because $F_2 > 0$ and $F_3 > 0$, by Lemma 2. By the second-derivative test, (9) holds. ■

Examples of minimizers—points P in (8)—are tabulated here:

Minimizers for various values of *t*

t	2	1	0	$-1/2$	-1	$-2/3$	$-3/2$	-2	-4
P	X_{44}	X_{1743}	X_6	X_{1449}	X_1	X_{1100}	X_{3247}	X_{37}	X_{45}

Each of these nine triangle centers has properties given in the encyclopedia [**5**]. Note that the minimal value (the right side of (9)) is 0 for $t = -1$ or $t = 2$.

When $t = 2$, there are other solutions to (7) than (8). It is easy to check that they are the points on the line $x + y + z = 0$. This line, shown in Figure 4, is constructed as follows: let

$$\begin{aligned} D &= AI \cap BC, & E &= BI \cap CA, & F &= CI \cap AB, \\ A' &= EF \cap BC, & B' &= FD \cap CA, & C' &= DE \cap AB; \end{aligned}$$

then A', B', C' are collinear on the line $x + y + z = 0$, known as the antiorthic axis of ABC.

Still more minimizers

The second-derivative test also applies to other minimization proofs. One example, for which the proof is so similar to that of Theorem 2 that we state it without proof, is that X_9 minimizes $yz + zx + xy$. That is, if a, b, c are sidelengths of a triangle, then

$$yz + zx + xy \geq \frac{(ax + by + cz)^2}{2(ab + ac + bc) - a^2 - b^2 - c^2}$$

for all real x, y, z.

Yet another minimization, for which there may exist a better proof than the second-derivative test, is that the triangle center given by

$$3a + \sqrt{3(a^2 + b^2 + c^2)} : 3b + \sqrt{3(a^2 + b^2 + c^2)} : 3b + \sqrt{3(a^2 + b^2 + c^2)}$$

minimizes the function

$$\frac{x^2 + y^2 + z^2}{(ax + by + cz)(x + y + z)}.$$

A symmetric version of Theorem 1

The inequality (9) can be written in symmetric form as follows. If s and t both lie in the interval $[-1, 2]$ and $(s + 2)(t + 2) = 4$, then for any six real numbers a, b, c, x, y, z,

$$\begin{aligned}&[a^2 + b^2 + c^2 + s(bc + ca + ab)][x^2 + y^2 + z^2 + t(yz + zx + xy)]\\ &\quad \geq (ax + by + cz)^2(1 + s)(1 + t). \qquad (11)\end{aligned}$$

When $s = t = 0$, this is the classical Cauchy-Schwarz inequality, but neither (11) nor (9) is found in the large collection [**6**]. There is, however, an intriguing similarity-in-appearance of (11) to Wagner's inequality [**8, 1, 2**]. For $n = 3$, the latter can be written as follows: if $0 \leq u \leq 1$, then

$$\begin{aligned}&[a^2 + b^2 + c^2 + 2u(bc + ca + ab)][x^2 + y^2 + z^2 + 2u(yz + zx + xy]\\ &\quad \geq [ax + by + cz + u[a(y + z) + b(z + x) + c(x + y)]]^2 \qquad (12)\end{aligned}$$

holds for all values of a, b, c, x, y, z.

For $n > 3$, the inequality (12) generalizes to Wagner's inequality, whereas (11) appears to generalize in a different way, as stated here:

CONJECTURE (MOSES). Suppose that $n \geq 3$ and that the following conditions hold:

(1) s and t both lie in the interval $\left[\dfrac{-2}{n-1}, 2\right]$;

(2) $\left(s + \dfrac{n+1}{n-1}\right)\left(t + \dfrac{n+1}{n-1}\right) = \dfrac{3n-1}{n-1}$;

(3) $a_1, a_2, \ldots, a_n, x_1, x_2, \ldots, x_n$ are real numbers.

Let $B = \dfrac{3n-1}{4(n-1)^3}[2 + s(n-1)][2 + t(n-1)]$. Then

$$\left(\sum_{1 \leq j \leq n} a_j^2 + s \sum_{1 \leq j < k \leq n} a_j a_k\right)\left(\sum_{1 \leq j \leq n} x_j^2 + t \sum_{1 \leq j < k \leq n} x_j x_k\right) \geq B \left(\sum_{1 \leq j \leq n} a_j x_j\right)^2.$$

REFERENCES

1. P. S. Bullen, *A Dictionary of Inequalities*, Addison Wesley Longman, Harlow, England, 1998.
2. Pietro Cerone and Sever S. Dragomir, *Mathematical Inequalities*, CRC Press, New York, 2011.
3. Harold Scott MacDonald Coxeter, Some applications of trilinear coordinates, *Linear Algebra* **226**(8) (1995) 375–388. http://dx.doi.org/10.1016/0024-3795(95)00169-R.
4. Clark Kimberling, Central points and central lines in the plane of a triangle, *Math. Mag.* **67** (1994) 163–187. http://dx.doi.org/10.2307/2690608.
5. Clark Kimberling, *Encyclopedia of Triangle Centers—ETC*, http://faculty.evansville.edu/ck6/encyclopedia/ETC.html
6. D. S. Mitrinović, J. E. Pečarić, and V. Volenec, *Recent Advances in Geometric Inequalities*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.
7. Robert K. Smither, The symmedian point: constructed and applied, *College Math. J.* **42** (2011) 115–117. http://dx.doi.org/10.4169/college.math.j.42.2.115.
8. S. S. Wagner, A generalization of Cauchy's inequality, *Notices Amer. Math. Soc.* **12** (1965) 220.

Summary The Cauchy-Schwarz inequality is generalized in a new way. The method is motivated by extremal properties of the incenter and symmedian point in the plane of a triangle with sidelengths a, b, c. Remarkably, in some of the inequalities involving six real numbers a, b, c, x, y, z, the numbers a, b, c need not be sidelengths of a triangle.

PROBLEMS

BERNARDO M. ÁBREGO, *Editor*
California State University, Northridge

Assistant Editors: SILVIA FERNÁNDEZ-MERCHANT, California State University, Northridge; JOSÉ A. GÓMEZ, Facultad de Ciencias, UNAM, México; EUGEN J. IONASCU, Columbus State University; ROGELIO VALDEZ, Facultad de Ciencias, UAEM, México; WILLIAM WATKINS, California State University, Northridge

PROPOSALS

To be considered for publication, solutions should be received by November 1, 2012.

1896. *Proposed by Timothy Hall, PQI Consulting, Cambridge, MA.*

Find with proof the value of

$$\int_0^\infty \frac{\cos(\sqrt{x})}{\sqrt{x}} \cos x \, dx.$$

1897. *Proposed by H. A. ShahAli, Tehran, Iran.*

Let m and n be positive integers such that $m < n$. Determine necessary and sufficient conditions for a sequence $\{x_j\}_{j=1}^n$ of real numbers to satisfy that

$$\left|\sum_{j\in S} x_j\right| = \left|\sum_{\substack{1\le j\le n\\ j\notin S}} x_j\right|,$$

for every m-element subset S of $\{1, 2, \ldots, n\}$.

1898. *Proposed by Mark Bowron, Laughlin, NV.*

A subset E of a topological space X is called a *Kuratowski* 14-*set* if 14 distinct sets can be obtained by repeatedly applying closure and complement to E in some order. It is known that Kuratowski 14-sets E with $|E| = 3$ exist. Do any exist with $|E| < 3$?

Math. Mag. **85** (2012) 229–236. doi:10.4169/math.mag.85.3.229. © Mathematical Association of America

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any 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.

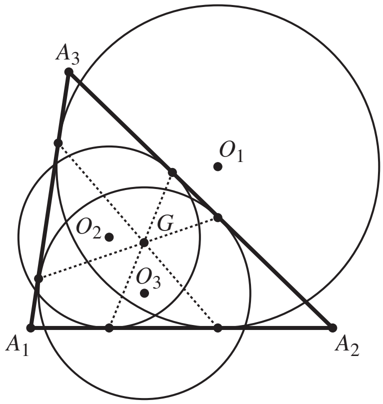
Solutions should be written in a style appropriate for this MAGAZINE.

Solutions and new proposals should be mailed to Bernardo M. Ábrego, Problems Editor, Department of Mathematics, California State University, Northridge, 18111 Nordhoff St, Northridge, CA 91330-8313, or mailed electronically (ideally as a LaTeX or pdf file) to mathmagproblems@csun.edu. All communications, written or electronic, should include **on each page** the reader's name, full address, and an e-mail address and/or FAX number.

1899. *Proposed by Michel Bataille, Rouen, France.*

Let $A_1A_2A_3$ be a triangle with centroid G. For $i \in \{1, 2, 3\}$, the circle $\mathcal{C}_i$ with center O_i and radius r_i is tangent to the two lines through A_i spanned by the sides of the triangle; moreover, the points of tangency and G are collinear. Prove that

$$\frac{r_1}{r_1 + GO_1} + \frac{r_2}{r_2 + GO_2} + \frac{r_3}{r_3 + GO_3} = 2.$$



1900. *Proposed by Greg Oman, University of Colorado at Colorado Springs, Colorado Springs, CO.*

Let X be a set, and let S_X denote the set of all functions $f : X \to \mathbb{Z}$. The set S_X becomes a ring via the operations $(f + g)(x) := f(x) + g(x)$ and $(f \cdot g)(x) := f(x)g(x)$. Let B_X be the subring of S_X consisting of the functions f whose images in $\mathbb{Z}$ are finite. Does there exist an infinite set X such that the rings B_X and S_X are isomorphic?

Quickies

Answers to the Quickies are on page 236.

Q1021. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let $\{x_n\}$ be a sequence of real numbers with at least one cluster point and let

$$K = \{x : x \text{ is a cluster point of } \{x_n\}\}.$$

(Recall that a cluster point of $\{x_n\}$ is the limit of a subsequence of $\{x_n\}$.) Suppose that $(x_{n+1} - x_n) \to 0$. Prove or disprove that K is a connected set of real numbers.

Q1022. *Proposed by Andrew Simoson, King College, Bristol, TN.*

Show that for any $a \geq 1$, the function

$$f(t) = \frac{(1-t)(\sqrt{a^2-a}+a)-1}{(1+t)(\sqrt{a^2-a}-a)+1}$$

is constant for all t except when the denominator vanishes.

Solutions

A stationary point in disguise **June 2011**

1871. *Proposed by Cosmin Pohoata, Princeton University, Princeton, NJ.*

Let f, g be two differentiable real functions such that $g(x) \neq 0$ for all real numbers x. Suppose that c is a real number such that

$$f(c)\int_a^b g(x)\,dx \neq g(c)\int_a^b f(x)\,dx,$$

for all pairwise distinct real numbers a and b. Prove that $(f/g)'(c) = 0$.

Solution by Robert Calcaterra, University of Wisconsin-Platteville, Platteville, WI. We argue by way of contradiction. The key observation we need is the following simple statement (whose proof is included for completeness at the end of our solution):

> *For every continuous function h defined on an open interval I, differentiable at $c \in I$, and such that $h(c) = 0$ and $h'(c) \neq 0$, there exist distinct a and b in I such that*
>
> $$\int_a^b h(x)\,dx = 0.$$

Define $h(x) = f(x)g(c) - g(x)f(c)$ for $x \in \mathbb{R}$. Then $h(c) = 0$ and $h'(x) = f'(x)g(c) - g'(x)f(c)$ for $x \in \mathbb{R}$. Since $(f/g)'(c) = h'(c)/(g(c))^2 \neq 0$ implies $h'(c) \neq 0$, the observation above gives the existence of some distinct a and b satisfying

$$\int_a^b h(x)\,dx = 0 \quad \Leftrightarrow \quad g(c)\int_a^b f(x)\,dx = f(c)\int_a^b g(x)\,dx.$$

This contradicts the hypothesis of the problem and so we conclude that $(f/g)'(c) = 0$.

To prove the key observation, assume without loss of generality that $h'(c) = \ell > 0$. From the definition of the derivative, there exists $\delta > 0$ so that

$$\left|\frac{h(x)}{x-c} - \ell\right| < \frac{\ell}{2} \text{ for all } x \in [c-\delta, c+\delta] \subseteq I,\ x \neq c.$$

Thus $h(x) > 0$ if $x \in (c, c+\delta]$ and $h(x) < 0$ if $x \in [c-\delta, c)$. If we choose $t \in (0, \alpha]$ with $\alpha = \min\{\int_c^{c+\delta} h(x)\,dx, -\int_{c-\delta}^{c} h(x)\,dx\} > 0$, then by the Intermediate Value Theorem there are $b \in (c, c+\delta]$ and $a \in [c-\delta, c)$ such that $\int_c^b h(x)\,dx = t$ and $\int_a^c h(x)\,dx = -t$. Therefore $\int_a^b f(x)\,dx = 0$ and $a \neq b$.

Editor's Note. A similar solution was given by Haryono Tandra. Most of our solvers used a direct argument showing that c is actually an extremum of the function (f/g). One of the ideas along this line is to observe that the continuous function of two variables

$$(a, b) \to \frac{\int_a^b f(x)\,dx}{\int_a^b g(x)\,dx} - \frac{f(c)}{g(c)},\ a \neq b,$$

keeps a constant sign on the connected set $\{(a, b) \in \mathbb{R}^2 : b > a\}$. Then letting a go to b, implies that c is an extremum of (f/g). Another method was to observe that the function $\psi(x) = f(c)\int_c^x g(t)\,dt - g(c)\int_c^x f(t)\,dt$ is one-to-one in $\mathbb{R}$, and then conclude that ψ must be strictly monotone.

Also solved by Michel Battaile (France), Michael W. Botsko, Bill Cowieson, William R. Green, Eugene A. Herman, John C. Kieffer, Raymond Mortini (France), Paolo Perfetti (Italy), Mher Safaryan (Armenia), Joel Schlosberg, Haryono Tandra, Marian Tetiva (Romania), and the proposer. There were two incomplete and three incorrect submissions.

A sum decomposition modulo *m* June 2011

1872. *Proposed by Gregor Olšavský, Penn State University/the Behrend College, Erie, PA.*

Let m be an integer greater than 1. Show that every integer n can be written as $n \equiv a + b \pmod{m}$ where a is an integer that is relatively prime to m, and b is an integer such that $b^2 \equiv b \pmod{m}$.

Solution by Brian D. Beasley, Department of Mathematics, Presbyterian College, Clinton, SC.
Given an integer $m > 1$, we write m in its canonical form $m = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$ for primes $p_1 < p_2 < \dots < p_r$ and positive integers k_i. Let n be an arbitrary integer. Using the Chinese Remainder Theorem, we may choose an integer b such that for each i in $\{1, 2, \dots, r\}$,

$$\begin{cases} b \equiv 0 \pmod{p_i^{k_i}} & \text{if } p_i \text{ does not divide } n, \\ b \equiv 1 \pmod{p_i^{k_i}} & \text{if } p_i \text{ divides } n. \end{cases}$$

Then $b^2 \equiv b \pmod{m}$, since $b^2 \equiv b \pmod{p_i^{k_i}}$ for each i in $\{1, 2, \dots, r\}$. Next, we note that $\gcd(n - b, m) = 1$. Suppose some p_i divides $n - b$. Then either p_i divides n, in which case p_i must divide b, a contradiction to the fact that $b \equiv 1 \pmod{p_i^{k_i}}$; or p_i does not divide n, in which case p_i divides b by our choice of b, yet p_i also divides $n - b$, yielding a contradiction. Hence the result follows by taking $a = n - b$.

Also solved by Vahagn Aslanyan (Armenia), Michel Bataille (France), Elton Bojaxhiu (Germany) and Enkel Hysnelaj (Australia), Charles Burnette, Robert Calcaterra, John Christopher, Con Amore Problem Group (Denmark), Bill Cowieson, Tamás Dékány (Hungary), Dmitry Fleischman, Jaime Gutiérrez (Republic of Panama), Dorn Hetzel, Omran Kouba (Syria), Kathleen E. Lewis (Republic of the Gambia), László Lipták, Peter McPolin (Northern Ireland), Missouri State University Problem Solving Group, ONU-SOLVE problem group, J. M. Pacheco and Ángel Plaza (Spain), Prapanpong Pongsriiam and Tammatada Pongsriiam (Thailand), John P. Robertson, Mher Safaryan (Armenia), Joel Schlosberg, Nicholas C. Singer, John H. Smith, Marian Tetiva (Romania), and the proposer.

A trigonometric inequality for triangles June 2011

1873. *Proposed by Elias Lampakis, Kiparissia, Greece.*

Let ABC be a triangle with $a = BC$, $b = AC$, and $c = AB$. Prove that

$$2a\cos(\angle A) + 2b\cos(\angle B) + 2c\cos(\angle C) \le 3\sqrt[3]{abc}.$$

Solution by Enkel Hysnelaj, University of Technology, Sydney, Australia, and Elton Bojaxhiu, Kriftel, Germany.
Denote by Δ and s the area and the semiperimeter of $\triangle ABC$, respectively. By the Law of Cosines, the left hand side of the inequality in the problem is equal to

$$\begin{aligned}
&2a\cdot\frac{b^2+c^2-a^2}{2bc}+2b\cdot\frac{c^2+a^2-b^2}{2ca}+2c\cdot\frac{a^2+b^2-c^2}{2ab}\\
&=\frac{a(b^2+c^2-a^2)}{bc}+\frac{b(c^2+a^2-b^2)}{ca}+\frac{c(a^2+b^2-c^2)}{ab}\\
&=\frac{(a+b+c)(a+b-c)(b+c-a)(c+a-b)}{abc}\\
&=\frac{16s(s-a)(s-b)(s-c)}{abc}=\frac{16\Delta^2}{abc},
\end{aligned}$$

where the last equality is given by Heron's formula. Thus after taking square roots, the requested inequality is equivalent to $4\Delta\le\sqrt{3}\sqrt[3]{(abc)^2}$.

Let R be the circumradius of $\triangle ABC$. It is well-known that $\Delta=abc/4R$ and $a+b+c\le 3\sqrt{3}R$. Using these facts together with the Arithmetic Mean–Geometric Mean Inequality gives

$$4\Delta=\frac{abc}{R}\le\frac{3\sqrt{3}abc}{a+b+c}\le\frac{3\sqrt{3}abc}{3\sqrt[3]{abc}}=\sqrt{3}\sqrt[3]{(abc)^2},$$

which completes the proof.

Editor's Note. Using the Law of Sines, the inequality $a+b+c\le 3\sqrt{3}R$ is equivalent to $\sin(\angle A)+\sin(\angle B)+\sin(\angle C)\le\frac{3}{2}\sqrt{3}$, which follows by the concavity of the $\sin x$ function on $[0,\pi]$. Several submissions point out that the equality is reached for the equilateral triangles. Con Amore problem group, John Heuver, and Math 490 @ LSU Shreveport prove a slightly stronger result. They prove that the left hand side of the inequality given is less than or equal to $9abc/(ab+bc+ca)$.

Also solved by George Apostolopoulos (Greece); Dionne Bailey, Elsie Campbell, and Charles Diminnie; Herb Bailey; Michel Bataille (France); Rich Bauer; Erhard Braune (Austria); Scott H. Braun; Robert Calcaterra; Minh Can; ConAmore problem group (Denmark); Chip Curtis; Dan Daly and Haohao Wang; Prithwijit De (India); Marian Dincă (Romania); Dorottya Fekete (Hungary); Dmitry Fleischman; Allyson Fredrickson; Michael Goldenberg and Mark Kaplan; Eugene A. Herman; John G. Heuver (Canada); Joe Howard; Hidefumi Katsuura and Samih Obaid; L. R. King; Omran Kouba (Syria); Kee-Wai Lau (China); Ji Hu Lee (Korea); Leighann M. Leyerle; Quanquan Li; Weiping Li; Math 490 @ LSU Shreveport; Mathramz problem solving group; Donald J. Moore; Marcus Näslund (Sweden); ONU-SOLVE problem group; Pedro Perez and Alin Stancu; C. R. Selvaraj and Suguna Selvaraj; Earl A. Smith; Neculai Stanciu (Romania); Marian Tetiva (Romania); Hawpe Gamage Tharaka; Michael Vowe (Switzerland); Albert R. Whitcomb; John Zacharias; and the proposer. There was one incorrect submission.

The minimum polynomial of cos(3π/17) cos(5π/17) June 2011

1874. *Proposed by Michel Bataille, Rouen, France.*

Show that $\alpha=\cos(3\pi/17)\cos(5\pi/17)$ is algebraic over $\mathbb{Q}$. In addition, find the minimum polynomial and the algebraic conjugates of α.

Solution by Con Amore Problem Group, formerly all employed at the Department of Mathematics, Pedagogical University of Denmark, Copenhagen, Denmark.
Consider the cyclotomic polynomial

$$\Phi_{17}(x)=x^{16}+x^{15}+\ldots+x+1=\prod_{k=1}^{16}\left(x-\left(e^{iu}\right)^k\right)=\prod_{k=1}^{16}\left(x-\xi^k\right),$$

where $u = 2\pi/17$. Using the trigonometric product formulas, we find that

$$4\alpha = 2\cos(2\pi/17) + 2\cos(8\pi/17) = \xi + \xi^{16} + \xi^4 + \xi^{13} = \beta_1.$$

The splitting field $\mathbb{Q}(\xi)$ of $\Phi_{17}(x)$ over $\mathbb{Q}$ has exactly 16 automorphisms determined by

$$\xi \mapsto \xi^k, k = 1, 2, 3, \ldots, 16.$$

Using these on β_1, we find that β_1 has four conjugates over $\mathbb{Q}$ (each appearing four times):

$$\begin{aligned}
&\beta_1 = 4\alpha = 4\cos(3\pi/17)\cos(5\pi/17),\\
&\beta_2 = \xi^3 + \xi^5 + \xi^{14} + \xi^{12} = 2(\cos 3u + \cos 5u) = 4\cos(2\pi/17)\cos(8\pi/17),\\
&\beta_3 = \xi^9 + \xi^{15} + \xi^8 + \xi^2 = 2(\cos 2u + \cos 8u) = 4\cos(6\pi/17)\cos(10\pi/17), \text{ and}\\
&\beta_4 = \xi^{10} + \xi^{11} + \xi^7 + \xi^6 = 2(\cos 6u + \cos 7u) = 4\cos(\pi/17)\cos(13\pi/17).
\end{aligned}$$

Putting $\gamma_1 = \beta_1 + \beta_3$ and $\gamma_2 = \beta_2 + \beta_4$, and using the root expressions for β_1, β_2, β_3, β_4, γ_1, and γ_2, and the fact that $\gamma_1 + \gamma_2 = -1$ (since the coefficient of x^{15} in $\Phi_{17}(x)$ is 1), we find by easy (but tedious) calculations that the minimum polynomial for β_1 over $\mathbb{Q}$ is (since $\mathbb{Q}$ is a perfect field)

$$\begin{aligned}
(x-\beta_1)(x-\beta_3)(x-\beta_2)(x-\beta_4) &= \left(x^2 - \gamma_1 x - 1\right)\left(x^2 - \gamma_2 x - 1\right)\\
&= x^4 + x^3 + (\gamma_1\gamma_2 - 2)x^2 - x + 1\\
&= x^4 + x^3 - 6x^2 - x + 1. \qquad (1)
\end{aligned}$$

Because the conjugates of α over $\mathbb{Q}$ are

$$\begin{aligned}
\alpha = \beta_1/4 &= \cos(3\pi/17)\cos(5\pi/17),\\
\beta_2/4 &= \cos(2\pi/17)\cos(8\pi/17),\\
\beta_3/4 &= \cos(6\pi/17)\cos(10\pi/17), \text{ and}\\
\beta_4/4 &= \cos(\pi/17)\cos(13\pi/17),
\end{aligned}$$

it follows that α is algebraic over $\mathbb{Q}$, and from Equation (1) that the minimum polynomial of α over $\mathbb{Q}$ is

$$x^4 + \frac{1}{4}x^3 - \frac{3}{8}x^2 - \frac{1}{64}x + \frac{1}{256}.$$

Also solved by Rich Bauer, Paul Budney, Robert Calcaterra, Dmitry Fleischman, Michael Goldenberg and Mark Kaplan, Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Germany), Jaime Gutiérrez (Republic of Panama), Eugene A. Herman, Joel Iiams, Omran Kouba (Syria), Elias Lampakis (Greece), Peter McPolin (Northern Ireland), Howard Cary Morris, Northwestern University Math Problem Solving Group, Marian Tetiva (Romania), and the proposer. There were 3 incomplete or incorrect submissions.

The generating function of up-sawtooth permutations June 2011

1875. *Proposed by Éric Pité, Paris, France.*

Let $a_0 = a_1 = 1$ and for $n \geq 2$ define a_n as the number of permutations σ of $\{1, 2, \ldots, n\}$ such that

$$\begin{aligned}
&\sigma(1) < \sigma(2), \sigma(3) < \sigma(4), \ldots, \sigma(2j-1) < \sigma(2j) \text{ with } j = \lfloor n/2 \rfloor, \text{ and}\\
&\sigma(2) > \sigma(3), \sigma(4) > \sigma(5), \ldots, \sigma(2k) > \sigma(2k+1) \text{ with } k = \lfloor (n-1)/2 \rfloor.
\end{aligned}$$

Prove that for every $z \in \mathbb{C}$ such that $|z| < 1$,

$$\sum_{n=0}^{\infty} \frac{a_n}{n!} z^n = \frac{\sin z + 1}{\cos z}.$$

Solution by Jonathan Kariv and David Lonoff (Graduate Students), Department of Mathematics, University of Pennsylvania, Philadelphia, PA.

We call a permutation of the type described in the problem an *up-sawtooth* permutation. Similarly, a permutation which begins with a fall and then alternates rise, fall, etc. will be called *down-sawtooth*. We show that for $n \geq 2$,

$$a_n = \sum_{\substack{k=2 \\ k \text{ even}}}^{n} \binom{n-1}{k-1} a_{k-1} a_{n-k} = \sum_{\substack{k=1 \\ k \text{ odd}}}^{n} \binom{n-1}{k-1} a_{k-1} a_{n-k}.$$

To establish the first equation, let $k = \sigma^{-1}(n)$ (i.e., k is the position of n when σ is written in one-line notation). Note that k must be even, otherwise $n = \sigma(k) < \sigma(k-1)$ if $k > 1$, or $n = \sigma(k) < \sigma(k+1)$ if $k = 1$. Then the $k-1$ elements coming before n can be chosen in $\binom{n-1}{k-1}$ ways and arranged to form an up-sawtooth in a_{k-1} ways. The remaining $n-k$ elements that come after n can be arranged into an up sawtooth in a_{n-k} ways. Thus there are $\binom{n-1}{k-1} a_{k-1} a_{n-k}$ up-sawtooths with $k = \sigma^{-1}(n)$. Adding over all even k gives the first equation. The second part follows similarly by letting $k = \sigma^{-1}(1)$. The only difference is that the elements coming after 1 must form a down-sawtooth. Note that by replacing $\sigma(j)$ with $n + 1 - \sigma(j)$, we have that a_n also counts the number of down-sawtooths of length n, and so the above argument carries through to establish the second equation. Adding the two equations and dividing by 2 gives

$$a_n = \frac{1}{2} \sum_{k=1}^{n} \binom{n-1}{k-1} a_{k-1} a_{n-k}. \tag{1}$$

Set $S(z) = \sum_{n=0}^{\infty} \frac{1}{n!} a_n z^n$. Convergence for $|z| < 1$ is guaranteed by the fact that $a_n < n!$. Multiplying Equation (1) by $\frac{1}{n!} z^n$ and adding over all $n \geq 2$ gives

$$S(z) - z - 1 = \sum_{n=2}^{\infty} \frac{a_n}{n!} z^n = \frac{1}{2} \sum_{n=2}^{\infty} \sum_{k=1}^{n} \binom{n-1}{k-1} \frac{a_{k-1} a_{n-k}}{n!} z^n.$$

Differentiating both sides, and then re-indexing slightly, we find that

$$\begin{aligned} S'(z) - 1 &= \frac{1}{2} \sum_{n=2}^{\infty} \sum_{k=1}^{n} \binom{n-1}{k-1} \frac{a_{k-1} a_{n-k}}{(n-1)!} z^{n-1} \\ &= \frac{1}{2} \sum_{n=1}^{\infty} \sum_{k=0}^{n} \binom{n}{k} \frac{a_k a_{n-k}}{n!} z^n = \frac{1}{2} \left(S(z)^2 - 1 \right), \end{aligned}$$

which implies that $S'(z)/(S(z)^2 + 1) = 1/2$. Now integrating both sides, we find that $\arctan(S(z)) = z/2 + C$, and $S(0) = 1$ implies that $C = \arctan(1) = \pi/4$. Finally, the identity $\tan(\alpha + \beta) = (\sin 2\alpha + \sin 2\beta)/(\cos 2\alpha + \cos 2\beta)$ shows that

$$S(z) = \tan\left(\frac{z}{2} + \frac{\pi}{4}\right) = \frac{\sin z + 1}{\cos z}.$$

Editor's Note. Many solvers indicated that this problem is well-known. The permutations described in the problem are called alternating permutations and the number of such permutations a_n are called Euler zigzag numbers or Up/down numbers. The determination of the numbers a_n is called Andre's problem. As noted by Cecil Rousseau, the solution to this problem was published in 1879. [André, D. Développements de sec x et tan x. *Comptes Rendus Acad. Sci.*, Paris 88, 965–967, 1879.] Other references include H. Dorie, *100 Great Problems of Elementary Mathematics*, pp. 64–69; Graham, Knuth, and Patashnik, *Concrete Mathematics*, ex. 7.1, p. 546; and I. Tomescu, *Problems in Combinatorics and Graph Theory*. ex. 12.19, p. 61.

Also solved by Michel Bataille (France), Robert Calcaterra, Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Germany), Moubinool Omarjee (France), Robert W. Pratt, Cecil Rousseau, Joel Schlosberg, Marian Tetiva (Romania), and the proposer.

Answers

Solutions to the Quickies from page 230.

A1021. K is a connected set. Let a and b belong to K with $a < b$. Let $c \in (a, b)$. Suppose that c is not a cluster point of $\{x_n\}$. Let $r = \min\{c - a, b - c\}$. Since c is not a cluster point of $\{x_n\}$, there exist a positive number $\varepsilon < r$ and a positive integer n_0 such that $|x_n - c| \geq \varepsilon$ for all $n \geq n_0$. Since $(x_{n+1} - x_n) \to 0$, there exists $N > n_0$ such that $|x_{n+1} - x_n| < \varepsilon$ for $n > N$. Since a is a cluster point, there exists $m > N$ such that $x_m \leq c - \varepsilon$. If $x_{m+1} \geq c + \varepsilon$, then $x_{m+1} - x_m \geq 2\varepsilon$ which is a contradiction. Given that $|x_{m+1} - c| \geq \varepsilon$, it follows that $x_{m+1} \leq c - \varepsilon$. By induction this argument shows that $x_k < c - \varepsilon$ for all $k \geq m$, which contradicts the fact that b is a cluster point. Thus c is a cluster point of $\{x_n\}$ and therefore K is connected.

A1022. Both numerator and denominator are linear polynomials in t. Because both of them have the same root $t_0 = \sqrt{1 - 1/a}$, it follows that $f(t)$ is constant for all $t \neq t_0$. Thus $f(t) = f(-1) = 2a + 2\sqrt{a^2 + a} - 1$ for all $t \neq 0$.

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.

Perkins, David, *Calculus and Its Origins*, MAA, 2012; xiv + 165 pp, $60.95 (MAA member: $49.95), $28 ebook. ISBN 978–0–88385–575–1 (print edition), 978–0–61444–508–1 (electronic edition).

Should students have "the chance to learn calculus as a pursuit rather than as a toolkit"? This book tells a tale of the slow intellectual pursuit of calculus, culminating in the definitions of fundamental concepts (limit, derivative, integral, convergence, continuity) and in important results (fundamental theorem, series expansions). Rather than the book presenting a single thread of historical narrative, each chapter follows a theme (e.g., curves, indivisibles, quadrature). Each chapter ends with exercises that are meant to be read even if not worked. The book eschews "routine practice problems" and biographical information ("widely available"). As could be expected, the reader needs tolerance for geometric arguments (e.g., similar triangles), as well as facility in algebra and trigonometry.

Chang, Kenneth, Pasta graduates from alphabet soup to advanced geometry, *New York Times* (10 January 2012). `http://www.nytimes.com/2012/01/10/science/pasta-inspires-scientists-to-use-their-noodle.html`.

Legendre, George L., *Pasta by Design*, Thames and Hudson, 2011; 208 pp, $29.95. ISBN 978–0–500–51580–8.

Tiee, Chris, Math 20E LectureB, `http://www.math.ucsd.edu/~ctiee/math20e-w06/`.

Now you can have your math and eat it too. Sander Huisman (University of Twente, Netherlands) has recreated pasta shapes in Mathematica. Meanwhile, architect Legendre "experienced a similar epiphany" and classified 92 types of pasta in a kind of family tree, giving for each "a mathematical equation, a mouthwatering picture, and a paragraph of suggestions, like sauces to eat it with." Chris Tiee (UC San Diego) offers students in vector calculus a chance to match images of pasta varieties to parametrizations of them.

Lagarias, Jeffrey C. (ed.), *The Ultimate Challenge: The* $3x + 1$ *Problem*, American Mathematical Society, 2010; xiv + 341 pp, $59 (AMS member price: $47.20). ISBN 978–0–8218–4940–8.

Erdős: "Hopeless. Absolutely hopeless." Well, despite that discouraging remark to editor Lagarias, we have here two overviews by Lagarias (one reprinted from 1985 and a new one); survey papers on connections to dynamical systems, to ergodic theory, and to the theory of computation; two papers on stochastic models; half a dozen reprinted early papers; and a 75-page extensively annotated bibliography 1963–1999 (updated through 2009 at an indicated Website). Lagarias observes that the $3x + 1$ problem "comes close to being a 'perfect' problem in the Hilbert sense": clear and simply stated, and difficult—but perhaps not accessible. "It mocks our efforts!" It may even be undecidable.

Math. Mag. **85** (2012) 237–238. doi:10.4169/math.mag.85.3.237. © Mathematical Association of America

Cook, William J., *In Pursuit of the Traveling Salesman: Mathematics at the Limits of Computation*, Princeton University Press, 2012; ix + 228 pp, $27.95. ISBN 978–0–691–15270–7.

Schuessler, Jennifer, Arts Beat: Willy Loman, where are you?, *New York Times* (15 March 2012) C8, http://www.nytimes.com/interactive/2012/03/01/theater/201203-death-of-a-salesman-interactive.html#Math.

This is a luscious book, popularized mathematics at its best. The traveling salesman problem (TSP) is simple to state, has myriads of everyday applications, has stimulated considerable mathematics, and pushes computers to their limits. The book treats its history, describes applications, demonstrates solution methods, explains computational complexity, explores how humans solve small instances, and reproduces art works based on TSP. Only the chapter on linear programming contains equations. The book is beautifully laid out, with color illustrations and photographs on almost every page. "The ultimate goal is to encourage readers to take up their own pursuit of the salesman, with the hope that a knockout blow will come from an as yet unknown corner." The article by Schuessler notes a connection to the currently-revived *Death of a Salesman* by Arthur Miller: Its protagonist is a desperate unsuccessful traveling salesman; perhaps "a firmer grip on the [TSP] might have spared him at least some exhaustion."

Goirely, Alain and Derek E. Moulton, The mathematics behind *Sherlock Holmes: A Game of Shadows*, *SIAM News* 45 (3) (April 2012) 1, 8; http://www.siam.org/news/news.php?issue=0045.03 .

The TV show *Numb3rs* (2005–2010, now in syndication) used real-life mathematician consultants to get contemporary mathematics right, in solving cases and also in the equations on the blackboards and glassboards shown. The recent film *Sherlock Holmes: A Game of Shadows* did the same but with a twist: The mathematics for mathematician Prof. Moriarty's blackboard had to be relevant for the late 1890s and also play into the plot. The board contains the key to an enciphered code based on the Pascal triangle (author Arthur Conan Doyle wrote that Moriarty had composed "A treatise on the binomial theorem") and equations for the n-body problem (Moriarty also wrote *The Dynamics of an Asteroid*), plus partial differential equations toward developing a laser. The *SIAM News* article displays the blackboard with pointers to some of the mathematics.

Murray, Michael, Alan Carey, and Peter Bouwknegt, The Yang-Mills existence and mass gap problem. http://theconversation.edu.au/millennium-prize-the-yang-mills-existence-and-mass-gap-problem-3848.

This is the last in a series of succinct and non-technical descriptions by Australian mathematicians of the Clay Mathematics Institute's Millennium Prize Problems. Links are provided to the descriptions of the other six problems; each runs two to four pages, and just one has equations.

McGuire, Gary, Bastian Tugemann, and Gilles Civario, There is no 16-clue sudoku: Solving the sudoku minimum number of clues problem, http://arxiv.org/abs/1201.0749.

The authors resolve the question of the smallest number of clues that a sudoku problem can have and still have a unique solution. They proceeded by exhaustive search for a 16-clue puzzle among the non-equivalent 5,472,730,538 completed sudoku grids, using a new "hitting set" algorithm. (A *hitting set* for a collection of subsets is a set that intersects every one of the subsets; finding a minimal such set is an NP-complete problem.) The paper is eminently readable by students in mathematics or computer science.

Johnson, Carolyn Y., A food pyramid made of cookies, *Boston Globe* (11 December 2011), http://www.bostonglobe.com/ideas/2011/12/11/food-pyramid-made-cookies/Ev66x0eHjUIcBSEAl1EIYI/story.html.

Michael Brenner (Harvard University), an applied mathematician, with help from a graduate student and an undergraduate, has plotted thousands of recipes for baked goods onto the surface of a tetrahedron whose axes are sugar, flour, eggs, and liquid. The points form clusters of cookies, pancakes, brownies, and other families of delights (all colored differently)—but with lots of white space where as-yet-unexploited recipes may lie. The article provides a colorful tetrahedral net to cut out and assemble.

NEWS AND LETTERS

Laguerre, Pólya, Szegő, and the Lost Cousin

In a recent note [**5**] in this MAGAZINE, Tossavainen proves what he calls the "lost cousin of the fundamental theorem of algebra." This cousin bounds the number of zeros of certain linear combinations of exponential functions; the proof uses Rolle's theorem. The statement and proof of the cousin (for exponential functions) are similar in spirit to Descartes' rule of signs (for polynomial functions). The final paragraphs of [**5**] observe that Descartes' rule of signs does not generalize in a similar fashion to logarithm functions. In another recent note [**4**] in this MAGAZINE, Sznajder derives a variant of the lost cousin which he denotes the "generalized polynomial theorem." This variant is also proved using Rolle's theorem.

The well-known problem collection [**3**] of George Pólya and Gabor Szegő was published in German back in 1924; its English edition appeared in 1972. Chapter 1 of part V of [**3**] is called "Rolle's theorem and Descartes' rule of signs"; section 6 of this chapter deals with "Laguerre's proof of Descartes' rule of signs," and section 7 is called "What is the basis of Descartes' rule of signs?" Exercise 77 in the same chapter (page 46 of [**3**]) reads as follows:

> ***77.*** *Let $a_1, a_2, \ldots, a_n, \lambda_1, \lambda_2, \ldots, \lambda_n$ be real constants, $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. Denote by Z the number of real zeros of the entire function*
>
> $$F(x) = a_1 e^{\lambda_1 x} + a_2 e^{\lambda_2 x} + \cdots + a_n e^{\lambda_n x}$$
>
> *and by C the number of changes of sign in the sequence of numbers $a_1, a_2, \ldots, a_n$. Then $C - Z$ is a non-negative even integer.*

The solution of Exercise 77 (stated on page 47 of [**3**]) is based on Rolle's theorem. Section 7 on pages 50–52 of [**3**] discusses further generalizations of Descartes' rule of signs to other function sequences (beyond polynomial and exponential functions). Exercise 87 (on page 50 of [**3**]; solution on pages 226 and 227) provides a concise characterization of such function sequences in terms of certain Wronskian determinants. Pólya and Szegő attribute all these results (and many others) to the French mathematician Edmond Laguerre (1834–1896). The collected works of Laguerre were edited by Charles Hermite, Henri Poincaré, and Eugène Rouché, and appeared around the end of the nineteenth century; see [**2**].

The relationship between [**2, 3**] on the one side and [**5, 4**] on the other side is as follows. Exercise 77 of [**3**] summarizes and strengthens all results derived in the recent notes [**5, 4**]. Exercise 87 of [**3**] supersedes the observation of [**5**] on logarithm functions, as logarithms do not satisfy the condition on the Wronskians. The message of all this should be clear: Read the old masters! Certainly Pólya and Szegő, and also the very old master Laguerre!

GERHARD J. WOEGINGER
University of Eindhoven
5600 MB Eindhoven, Netherlands
gwoegi@win.tue.nl

Looking Back on the Lost Cousin

When I considered submitting the manuscript of "Lost Cousin" to this MAGAZINE, I was a little worried about that my efforts might be in vain. The result I had discovered seemingly belonged to the folklore of mathematics; it was similar to those numerous exercises that we see

Math. Mag. **85** (2012) 239–240. doi:10.4169/math.mag.85.3.239. © Mathematical Association of America

in textbooks but we do not know who have originally defined them.The theorem might be only a consequence of a more general result that all matured mathematicians knew but I did not.

So, I realized the potential problem but I did not know how to solve it. Searching MathSciNet or other databases—and the bookshelves of our library, too!—did not help me because "exponential function," "fundamental theorem of algebra" and other similar search terms guided me to something else than what I was looking for. For instance, "Descartes' rule of signs" does not link to [**3**] in MathSciNet.

Then I discovered a possible way to get around the problem. As said on the first page of [**5**], the article is first and foremost a story about a personal learning process, not only about the product of such a process. Whether my result was new or not, it would be anyway meaningful to discuss it as a concrete example of what learning and doing mathematics can be.

Actually, there is a short conclusion to the story of "Lost Cousin." Quite soon after [**5**] came out, I noticed that I had forgotten to discuss the power functions. This is a little embarrassing to say but I left that simply because I focused only on examining whether my proof applies to all differentiable and monotone functions or not. Luckily, Sznajder noticed my mistake and corrected it in [**4**].

A year later a friend of mine contacted me telling that he had read "Lost Cousin" and noticed the linkage to Descartes' rule of signs. This led us to examine the connection more thoroughly and it resulted in a short note [**1**]. But, as Woeginger demonstrates, it is possible that we were again reinventing the wheel. He is right: we should read the old masters, indeed!

Timo Tossavainen
School of Applied Educational Science and Teacher Education
University of Eastern Finland
Savonlinna, Finland
timo.tossavainen@uef.fi

From the Editor

It is often hard to find all the precedents for a result, and we certainly wouldn't want to be denied a nice bit of mathematics because of an author's doubts on that score. But when the connections are finally revealed, it's a great opportunity. I challenge readers to look for this subject in Laguerre's papers. You may learn something you couldn't have learned in any other way!

A related matter—Professor Sznajder reports an error in his statement of the Generalized Polynomial Theorem [**4**]: It should conclude, "has no more than n positive roots." The word "positive" was omitted, although the context makes it clear that only positive roots are being considered. We are glad to put this correction on record.

Walter Stromquist

REFERENCES

1. P. Haukkanen and T. Tossavainen, A generalization of Descartes' rule of signs and the fundamental theorem of algebra, *Applied Mathematics and Computation* **218** (2011) 1203–1207; http://dx.doi.org/10.1016/j.amc.2011.05.107.
2. E. Laguerre, *Oeuvres,* Paris, Gauthier-Villars (1898–1905).
3. → G. Pólya and G. Szesö, *Problems and Theorems in Analysis II*, Springer, 1976.
4. → R. Sznajder, More on the lost cousin of the fundamental theorem of algebra. *Math. Mag.* **82** (2009) 218–219; http://dx.doi.org/10.4169/193009809X468850.
5. T. Tossavainen, The lost cousin of the fundamental theorem of algebra. *Math. Mag.* **80** (2007) 290–294.

MAA
MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth St., NW • Washington, DC 20036

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS