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 issue of the *Magazine*, 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 Susan Stromquist

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Hanover, 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: (866) 821-1221
Fax: (202) 387-1208
E-mail: advertising@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. 83, No. 1, February 2010

MATHEMATICS MAGAZINE

LETTER FROM THE EDITOR

Last year, as Editor-Elect of this MAGAZINE, I had the pleasure of working with the Editor, Frank Farris. Frank has now completed his second term as Editor, but his influence continues to be felt. He remains active in the MAA's publications arena. Many of the articles and notes we publish in 2010 will reflect his selections and editing. He has been very helpful to me in the transition. So has former Editor Allen Schwenk, and I am grateful to them both.

With this issue we welcome our new Problems Editor, Bernardo Ábrego. He succeeds Elgin Johnston, and I want to add my voice to Frank's in thanking Elgin for his nine years of service.

I hope you enjoy the articles in this issue. Ken Ross's article on repeating decimals applies rigorous mathematical techniques to a very elementary topic. If you are teaching a prodigy who is ready to study detailed proofs but not yet steeped in subject matter, this article may be your text.

Shirley Yap shines a unifying light on differential equation techniques. Can you find a connection between her paper and Gary Brookstone's note on the brachistochrone problem? Martin Griffiths gives us a case study in functional equations, and Olympia Nicodemi tells us the history of uniformly accelerated motion. Galileo connected it to gravity, but would his exposition of it have met the MAGAZINE's standards? (Of Galileo and Oresme, who is mentioned in two of our articles?)

Ethan Bolker's note is a definitive treatment of a famous card trick. Some of the shorter notes expand on topics raised in earlier issues of the MAGAZINE. Four of our notes authors are students. We know that our readers will enjoy the Reviews column, the Problems section, and the Putnam feature in this issue.

MATHEMATICS MAGAZINE is always on the lookout for interesting articles. I want especially to encourage authors who can describe the mathematics of practical applications or connections between mathematics and other disciplines. Our Guidelines for Authors are reprinted periodically, and the latest version appears in this issue. We prefer email now—otherwise, there is little change from 2001.

The strengths of the MAGAZINE arise from the efforts of its authors and of its many referees. Lists of our referees have been appearing in our December issues. If you like what you read here, perhaps you can find some of your colleagues on these lists and thank them personally. If you are a referee yourself, know that your work is appreciated.

Walter Stromquist, Editor

ARTICLES

Differential Equations—Not Just a Bag of Tricks!

SHIRLEY LLAMADO YAP
California State University, East Bay
Hayward, CA 94544
shirley.yap@csueastbay.edu

Typical first courses in differential equations comprise a variety of techniques to solve specific equations—homogeneous, exact, separable, and so on. After taking such a course, a student might justifiably conclude that the subject is just a bag of tricks. However, there is a beautiful and deep theory that unifies and extends most of these seemingly unrelated methods. The theory, introduced by the Norwegian mathematician Sophus Lie in the mid-19th century, exploits the symmetries of differential equations. The method finds a coordinate system in which the differential equation is easier to solve.

Sophus Lie started his mathematical career studying geometry. Differential equations entered his studies in 1869, when he observed that a geometric condition related to the symmetries of tetrahedral complexes translated into a first-order partial differential equation. Upon learning of this connection between Lie's geometric work and differential equations, his colleague Felix Klein communicated to Lie that his method of integrating differential equations using transformations was analogous to the way Abel and Galois used symmetries to solve polynomial equations [**9**, p. 22]. Lie then attempted to develop for differential equations what Galois had done for algebraic equations—classify and solve them using group theory.

The general mathematical community did not fully appreciate Lie's work during his lifetime—a fact that he once lamented in a letter to his colleague Adolf Mayer: "If I only knew how I could get mathematicians interested in transformation groups and the treatment of differential equations that arises from them. I am certain, absolutely certain, that, at some point in the future, these theories will be recognized as fundamental" [**7**, p. 7]. At the time of Lie's death, the theory of continuous transformations veered towards the global, abstract tendencies of modern differential geometry and away from Lie's original applications in differential equations [**14**, p. xvi].

In the past few decades, scientists have resurrected Lie's program and rejuvenated research in the field. In 1950, G. Birkhoff applied Lie's methods to engineering in "Hydrodynamics: A Study in Logic, Fact and Similitude" [**1**]. In the 1980s, L. V. Ovsiannikov and others successfully applied symmetry methods to solve problems in fluid mechanics, gas dynamics, classification of second-order linear equations, conservation law theories, and other physical problems. For most of the 20th century, engineers viewed Lie's ideas as little more than a theoretical curiosity because of the intractable computations involved in the process. However, the dramatic improvement of computer algebra software in the past thirty years has made previously impossible symmetry computations easy. Current research in the field is burgeoning with applica-

Math. Mag. **83** (2010) 3–14. doi:10.4169/002557010X479947.

tions in gas dynamics [**17**], evolutionary biology [**8**], quantum chemistry [**2**], hydrodynamics [**4**], signal processing [**5**], and many other areas of science and engineering.

Knowledge of elementary differential equations and multivariable calculus provides enough background to understand the salient features of the symmetry method. The main idea is that complex motions can be reduced to simple translations, if we look carefully enough.

Symmetries

Symmetries of algebraic equations In calculus, students learn that the graph of $f(x) = x^2$ is symmetric with respect to reflection across the y-axis, the graph of $f(x) = x^3$ is symmetric about the origin, and the graph of $\sin(x)$ is symmetric with respect to horizontal translation by 2π. These transformations are symmetries of f because they map the graph of f to itself. In general, for a function $f : \mathbb{R} \to \mathbb{R}$, a symmetry of f is a continuous map from $\mathbb{R}^2$ to $\mathbb{R}^2$ that maps the graph of f to itself and has a continuous inverse.

For example, for any nonzero $t \in \mathbb{R}$, $\phi_t : (x, y) \mapsto (tx, t^2y)$ is a symmetry of $y = x^2$ because if (a, b) is a point on the graph of $y = x^2$, $t^2b = (ta)^2$, which means that $\phi_t(a, b) = (ta, t^2b)$ is also on the graph of $y = x^2$. FIGURE 1 shows this symmetry for selected values of t. If we imagine t to represent time, then the letters and curves in the figure allow us to follow the motion of space under these symmetries at various points in time. A similar calculation shows that $\phi_t : (x, y) \mapsto (x + t, e^t y)$ is a symmetry of $y = e^x$, as shown in FIGURE 2.

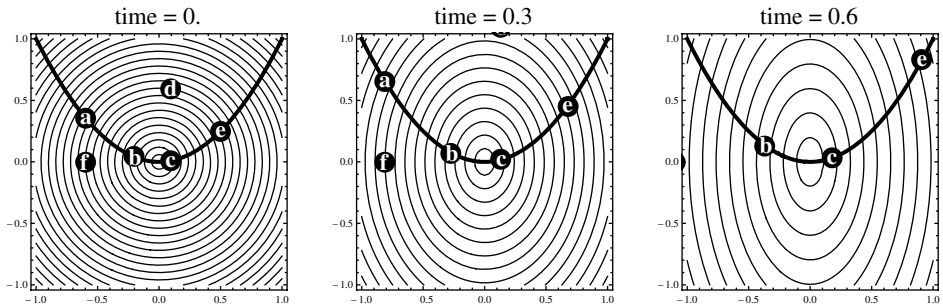


Figure 1 These pictures show how $\phi_t : (x, y) \mapsto (tx, t^2y)$ maps the graph of $y = x^2$ to itself. For reference, we show how ϕ_t transforms concentric circles and various points on and off the graph.

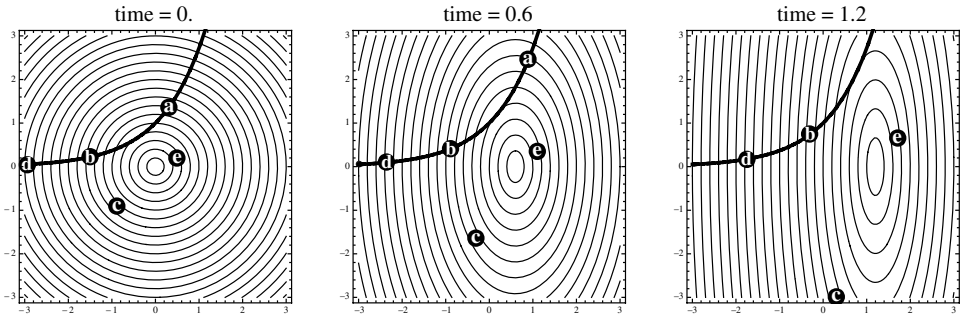


Figure 2 $\phi_t : (x, y) \mapsto (x + t, e^t y)$ is a symmetry of $f(x) = e^x$

If f is a function from $\mathbb{R}^2$ to $\mathbb{R}$, a symmetry of f is a transformation of $\mathbb{R}^3$ that sends any point that satisfies $z = f(x, y)$ to another point that satisfies the same equation. For example, the solution set of $z = x^2 + y^2$ is a paraboloid, which can be thought of as a union of circles parallel to the xy-plane. The family of maps $\phi_t : (x, y, z) \mapsto (tx, ty, t^2 z)$ moves circles up the paraboloid as t increases. FIGURE 3 shows a few snapshots.

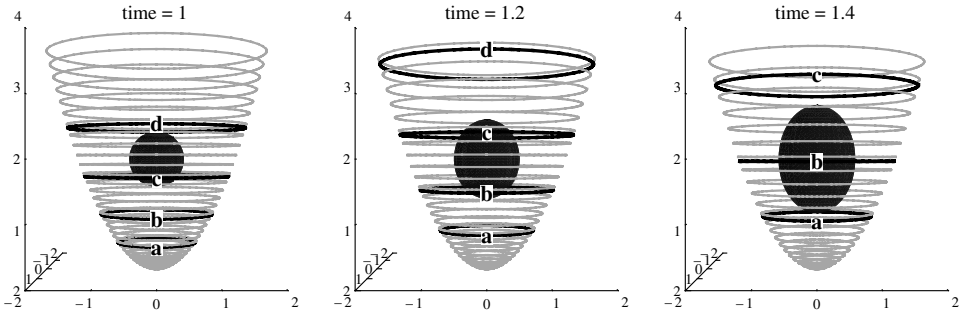


Figure 3 Solutions of $z = x^2 + y^2$ move up the parabola under the transformation $\phi_t : (x, y, z) \mapsto (tx, ty, t^2 z)$. For reference, we show how a sphere is transformed under ϕ_t.

Symmetries of ordinary differential equations Symmetries of ordinary differential equations (ODEs) also permute the solutions of the equations, but in the case of ODEs, a solution is an entire curve, not just a point.

Given the first order ODE

$$\frac{dy}{dx} = \omega(x, y),$$

with solutions defined in a domain D of $\mathbb{R}^2$, we look for transformations ϕ from D to D (or some subset $S \subset D$ to itself) that map solutions to other solutions. For reasons to be described, we also need ϕ to be differentiable and have a differentiable inverse. These conditions are equivalent to a transformation $\phi(x, y) = (u(x, y), v(x, y))$ having a nonzero Jacobian:

$$u_x\, v_y - v_x\, u_y \neq 0. \tag{1}$$

EXAMPLE 1. The graphs of solutions of the simplest ODE

$$\frac{dy}{dx} = 0 \tag{2}$$

are horizontal lines in the plane. For any real number t, $\phi_t : (x, y) \mapsto (e^t x, e^t y)$ is a symmetry of (2) because it maps horizontal lines to other horizontal lines. Any of these transformations with $t \neq 0$ will stretch or shrink the lines, but horizontal lines are preserved as sets.

EXAMPLE 2. Each solution of

$$\frac{dy}{dx} = \frac{2y}{x} \tag{3}$$

is a parabola passing through $(0, 0)$. For any real number t, $\phi_t : (x, y) \mapsto (u(t), v(t)) = (x, e^t y)$ is a symmetry of (3) because ϕ_t maps the curve $y = cx^2$ to $v = (ce^t)u^2$, which is another parabola passing through the origin.

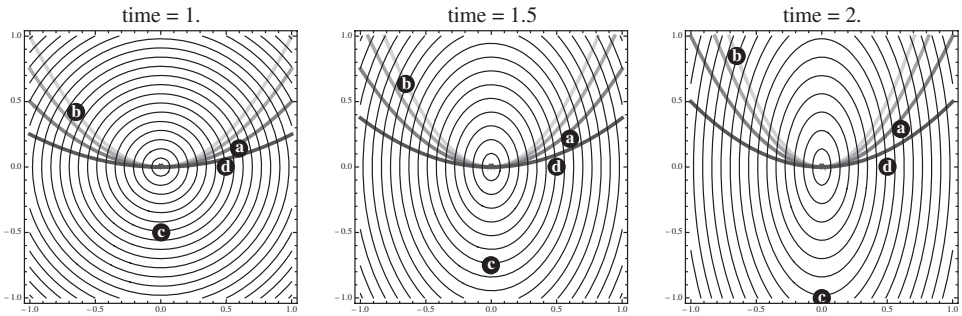


Figure 4 $\phi_t : (x, y) \mapsto (x, ty)$ is a symmetry of (3)

EXAMPLE 3. The nonconstant solutions to

$$\frac{dy}{dx} = \begin{cases} \frac{1-y^2}{x} & x \neq 0 \\ 0 & x = 0 \end{cases} \tag{4}$$

are the curves $y_c(x) = \frac{cx^2-1}{cx^2+1}$, where c is a positive constant. For any t, $\phi_t : (x, y) \mapsto (u, v) := (e^t x, y)$ is a symmetry of (4) because the image $v(u)$ of a solution $y(x)$ is another solution:

$$\left(x, \frac{cx^2-1}{cx^2+1}\right) \to \left(e^t x, \frac{cx^2-1}{cx^2+1}\right) = (u, v), \quad \text{so}$$

$$v(u) = \frac{c(u/e^t)^2 - 1}{c(u/e^t)^2 + 1} = \frac{(c/e^{2t})x^2 - 1}{(c/e^{2t})x^2 + 1} = y_{ce^{-2t}}(u).$$

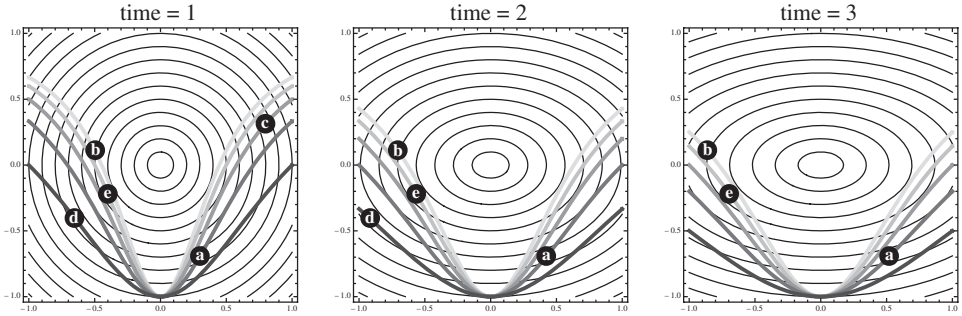


Figure 5 $\phi_t : (x, y) \mapsto (u, v) := (e^t x, y)$ is a symmetry of (4)

Each of these symmetry examples involves a parameter t and so represents a whole *family* of symmetries parametrized by t. Although there are other types of symmetries, we are mainly interested in these so-called *one-parameter Lie groups*. Olver [**14**, p. 34] gives a formal definition of this phrase.

Since the solutions to equations (2), (3), and (4) were already known, it was easy to check if a given transformation was a symmetry of those differential equations. However, the greatest utility of symmetries is to help *find* solutions. To that end, we describe how to compute symmetries of an ODE without knowing the solutions.

Computing symmetries We develop a formula that allows us to check whether a transformation (or really a family of transformations) is a symmetry of an ODE, but our real goal is to use that formula to *find* symmetries. If $y(x)$ is a solution to the first order ordinary differential equation

$$\frac{dy}{dx} = \omega(x, y), \tag{5}$$

a symmetry of (5) maps a solution $y(x)$ to another solution $v(u)$. In other words, $v(u)$ also solves (5): $\frac{dv}{du} = \omega(u, v)$. Expanding the differentials dv and du in terms of dx and dy results in a partial differential equation (PDE) for u and v:

$$\omega(u, v) = \frac{dv}{du} = \frac{v_x\,dx + v_y\,dy}{u_x\,dx + u_y\,dy} = \frac{v_x\,dx + v_y\,y'(x)\,dx}{u_x\,dx + u_y\,y'(x)\,dx} = \frac{v_x + v_y\,\omega(x, y)}{u_x + u_y\,\omega(x, y)}.$$

Thus, any two functions $u(x, y)$ and $v(x, y)$ that solve the PDE

$$\omega(u, v) = \frac{v_x + v_y\,\omega(x, y)}{u_x + u_y\,\omega(x, y)} \tag{6}$$

and the change of variable condition (1) fit together as the components of a symmetry of (5). In general, (6) may be a complicated PDE for u and v. However, we can impose additional conditions on u and v that simplify (6).

For example, the symmetry condition for the differential equation

$$\frac{dy}{dx} = y$$

is

$$v = \frac{v_x + v_y\,y}{u_x + u_y\,y}. \tag{7}$$

We do not seek *all* solutions of (7), merely some; one way is to set $v := y$, which reduces it to

$$u_x + yu_y = 1. \tag{8}$$

Setting $u_x = 0$ or $u_y = 0$ could further simplify (8). Since setting $u_x = 0$ would produce a degenerate transformation, violating condition (1), we set $u_y = 0$, which reduces (8) to $u_x = 1$. This equation is satisfied by $u = x + t$, where t is a constant of integration. Thus, the family of translations $\phi_t(x, y) = (x + t, y)$ all solve (8).

In fact, a similar analysis shows that if $\omega(x, y)$ is independent of x, then (5) is symmetric under the translations $(x, y) \mapsto (x + t, y)$. More generally, higher order ODEs that are missing the independent variable, such as $y_{xx}y_x = y^2$, are also symmetric under translations in the x direction. Similarly, ODEs that are missing the dependent variable are symmetric under the translations $(x, y) \mapsto (x, y + t)$. This is just the familiar "$+C$" from calculus.

Transformations that scale the independent or dependent variable are also common symmetries of differential equations. For example, according to (6), the symmetries of (4) satisfy

$$\frac{v_x + v_y(\frac{1-y^2}{x})}{u_x + u_y(\frac{1-y^2}{x})} = \frac{1 - v^2}{u}. \tag{9}$$

We again impose the conditions $v_x = u_y = 0$ and $v = y$ to reduce (9) to $x\,u_x = u$, with solution $u = e^t\,x$, where t is a constant of integration. Thus, $(x, y) \mapsto (e^t\,x, y)$ is a symmetry of (4).

An example of a PDE with a more complex symmetry arises from the ODE

$$\frac{dy}{dx} = \frac{y+1}{x} + \frac{y^2}{x^3}, \tag{10}$$

which is symmetric under the family of maps $\phi_t : (x, y) \mapsto (u, v) = (\frac{x}{1-tx}, \frac{y}{1-tx})$. Although the computations that derive these symmetries are too lengthy to include in this paper, we check the symmetry condition $v(u) = \omega(u, v)$:

$$\begin{aligned}
\frac{dv}{du} &= \frac{\frac{ty}{(1-tx)^2}\,dx + \frac{1}{1-tx}\,dy}{\frac{1}{(1-tx)^2}\,dx} = ty + (1-tx)\frac{dy}{dx} \\
&= ty + (1-tx)\left(\frac{y+1}{x} + \frac{y^2}{x^3}\right) = \frac{y+(1-tx)}{x} + (1-tx)\left(\frac{y^2}{x^3}\right) \\
&= \frac{\frac{y}{1-tx} + 1}{\frac{x}{1-tx}} + \frac{\left(\frac{y}{1-tx}\right)^2}{\left(\frac{x}{1-tx}\right)^3} = \frac{v+1}{u} - \frac{v^2}{u^3}.
\end{aligned}$$

In general, (6) is difficult to solve because it is nonlinear. However, the first-order term of the Taylor series expansion of (6) is a linear PDE and can be integrated to construct the symmetry.

Using symmetries to solve differential equations

The way symmetries are used to solve differential equations highlights a major theme in mathematics: Complicated problems can become simple when viewed in the right coordinate system. We first show how to solve ODEs with translational symmetries and then show how a general symmetry can be turned into a translational symmetry.

Suppose that $y'(x) = \omega(x, y)$ is symmetric with respect to translations in the y-direction: $\phi_t(x, y) = (u(t), v(t)) := (x, y + t)$. Then

$$\omega(x, y+t) = \omega(u(t), v(t)) = \frac{dv}{du} = \frac{d(y+t)}{dx} = \frac{dy}{dx} = \omega(x, y),$$

which shows that ω is independent of y. Therefore, $y'(x) = \omega(x)$, which is readily solved by integration: $y(x) = \int \omega(x)\,dx$. These computations show that any differential equation with a translational symmetry in the y direction is separable.

Converting a general symmetry into a translational symmetry Now suppose that $y'(x) = \omega(x, y)$ is invariant under *some* one-parameter Lie group that is not necessarily a translational symmetry. Then, by what is sometimes called a "straightening out theorem," there is a differentiable change of variables $(x, y) \mapsto (r, s)$, defined in some domain of $\mathbb{R}^2$, for which $y'(x) = \omega(x, y)$ is invariant under translations in the s direction [**14**, p. 30]. In the next few paragraphs, we show how to compute the *canonical coordinates* r and s and then express $y'(x) = \omega(x, y)$ in the new coordinate system.

Given a point (x, y) in the plane, the set of points $\{\phi_t(x, y) | a < t < b\}$ traces a curve in the plane called an *orbit* of ϕ_t. We call ϕ_t a *flow* because each point on the plane can be thought of as a molecule of fluid traveling along the trajectory defined by ϕ_t. For example, the orbits of $(x, y) \mapsto (e^t x, y)$ are horizontal lines. Differentiating

$(u(t), v(t))$ and evaluating at $t = 0$ results in the direction field $(\xi(x, y), \eta(x, y)) := (u'(0), v'(0))$ of the flow. If $\xi(x, y) \equiv 0$, then the orbits are vertical lines. Otherwise, the orbits are are exactly the integral curves of the vector field of the flow and are the solutions to

$$\frac{dy}{dx} = \frac{\eta(x, y)}{\xi(x, y)}. \tag{11}$$

For example, the vector field corresponding to the flow $\phi_t(x, y) = (u(t), v(t)) := \left(\frac{x}{1-tx}, \frac{y}{1-tx}\right)$ is $(\xi(x, y), \eta(x, y)) = (x^2, yx)$. The integral curves $y(x)$ of (x^2, yx) are the solutions to

$$\frac{dy}{dx} = \frac{yx}{x^2} = \frac{y}{x},$$

which are the rays $y = cx$ $(x \neq 0)$ emanating from the origin. Thus, for each constant c, the set of points (x, y) such that $yx^{-1} = c$ constitutes one orbit. In general, the constant of integration from (11) parametrizes the orbits of ϕ_t.

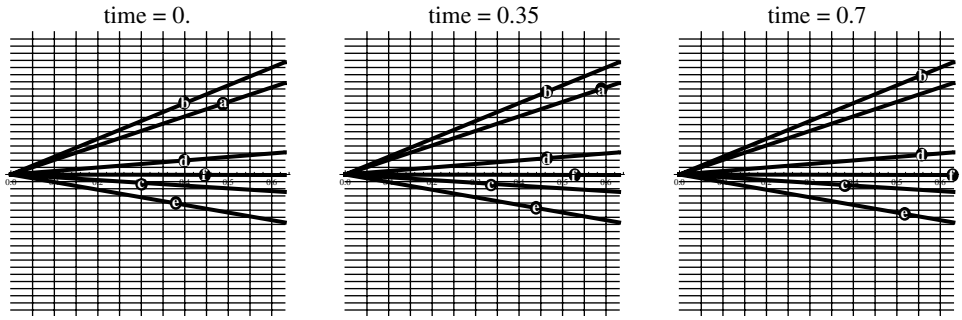


Figure 6 The orbits of various points under $\phi_t : (x, y) \mapsto (\frac{x}{1-tx}, \frac{y}{1-tx})$

Geometrically, converting an arbitrary symmetry into a translational symmetry means transforming the orbits of the symmetry into the orbits of the translational symmetry $(x, y) \mapsto (x, y + t)$, which are just vertical trajectories of unit speed. Algebraically, straightening the curves means that r is constant on the orbits and s travels at unit speed speed along the orbits: $s(u(t), v(t)) = s(\phi_t(x, y)) = s(x, y) + t$. The function $r(x, y)$ is easily computed by solving for the constant of integration in (11). For example, since the orbits of the symmetry $(x, y) \mapsto (\frac{x}{1-tx}, \frac{y}{1-tx})$ are given by $y = cx$, $r(x, y) := yx^{-1}$ is constant on the orbits.

Differentiating the normalizing condition $s(x, y) + t = s(u(t), v(t))$ with respect to t and evaluating at $t = 0$ yields the following PDE for s:

$$1 = s_x u'(0) + s_y v'(0) = s_x \xi(x, y) + s_y \eta(x, y). \tag{12}$$

In the (r, s) coordinate system, $y'(x) = \omega(x, y)$ becomes

$$\frac{ds}{dr} = \frac{s_x\, dx + s_y\, dy}{r_x\, dx + r_y\, dy} = \frac{s_x + s_y\, y'}{r_x + r_y\, y'}, \tag{13}$$

where $\frac{s_x + s_y\, y'}{r_x + r_y\, y'}$ is independent of s because its symmetries in the (r, s) coordinate system are translations in the s direction. After integrating (13), we can express the

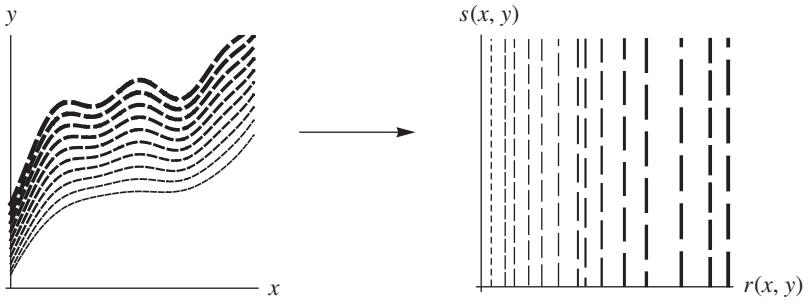


Figure 7 A change of variables turns the orbits of ϕ_t into vertical lines

solution $s(r)$ in the original xy coordinate system if the transformation $(x, y) \to (r(x, y), s(x, y))$ is invertible. Therefore, the Jacobian determinant of the transformation must be nonzero:

$$r_x s_y - r_y s_x \neq 0. \tag{14}$$

Thus, equations (11), (12), and inequality (14) determine the change of variables $(x, y) \mapsto (r, s)$ that convert a symmetry to a translational symmetry.

EXAMPLE 1. The orbits of the symmetry $(x, y) \mapsto (e^t x, y)$ of (4) are horizontal lines in the plane. Thus, $r := y$ is constant on the orbits of ϕ_t and (12) is equal to $s_x x = 1$ with solution $s = \ln|x|$.

In the r-s coordinate system, (4) becomes

$$\frac{ds}{dr} = \frac{\frac{1}{x}dx}{dy} = \frac{1}{x}\frac{dx}{dy} = \frac{1}{x}\frac{x}{1-y^2} = \frac{1}{1-y^2} = \frac{1}{1-r^2}, \tag{15}$$

with solution $s = \ln\left(\sqrt{\frac{1+r}{1-r}}\right) + c$, where c is a constant of integration. Substituting $r = y$ and $s = \ln(x)$ yields the family of solutions $y = \frac{Cx^2-1}{Cx^2+1}$, where C is an arbitrary constant.

EXAMPLE 2. We have already shown that (10) is symmetric under the maps $\phi_t(x, y) = (\frac{x}{1-tx}, \frac{y}{1-tx})$ and that $r(x, y) := yx^{-1}$ is constant on the orbits of ϕ_t. To solve for s, we set $s_y = 0$ so that (12) becomes $s_x = \frac{1}{x^2}$, whose solution is $s(x) = -\frac{1}{x}$.

In the (r, s) coordinate system, (10) becomes

$$\frac{ds}{dr} = -\frac{1}{x^2}\left[-\frac{y}{x^2} + \frac{1}{x}\left(\frac{y+1}{x} + \frac{y^2}{x^3}\right)\right]^{-1} = \frac{1}{1+\frac{y^2}{x^2}} = \frac{1}{1+r^2}.$$

This is easily solved to yield $s(r) = \tan^{-1}(r) + c$, which, after substituting $r = yx^{-1}$ and $s = -x^{-1}$, becomes $y = x\tan\left(-x^{-1} + c\right)$. Several solutions are shown in FIGURE 8.

Standard integration techniques seen in the light of symmetry

We now show how various techniques in differential equations are just specific instances of finding canonical coordinates for the equation.

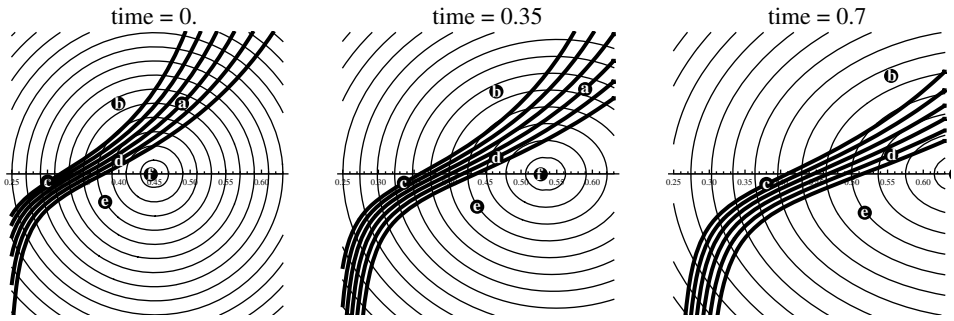


Figure 8 The behavior of several solutions of (10) under $\phi_t : (x, y) \mapsto (\frac{x}{1-tx}, \frac{y}{1-tx})$ at various points in time. For reference, we also show how ϕ_t transforms circles in the plane.

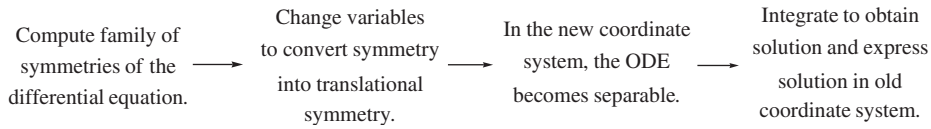


Figure 9 Summary of the symmetry method

Homogeneous coordinates A homogeneous equation has the form

$$\frac{dy}{dx} = G\left(\frac{y}{x}\right), \tag{16}$$

where G only depends on the ratio of y to x.

The maps $\phi_t : (x, y) \to (e^t x, e^t y)$ are symmetries of (16) whose orbits are rays emanating from the origin as well as the orbit consisting of the origin. The function $r := yx^{-1}$ is constant on those orbits and $s := \ln|x|$ satisfies (12), and (14). Under this change of coordinates $(x, y) \mapsto (r, s)$, (16) becomes the separable equation

$$\frac{ds}{dr} = \frac{s_x dx + s_y dy}{r_x dx + r_y dy} = \frac{\frac{1}{x}}{-\frac{y}{x^2} + G(\frac{y}{x})\frac{1}{x}} = \frac{1}{G(r) - r}.$$

Integrating factor Recall that the solution to the inhomogeneous first-order linear equation

$$y' + F(x)y = G(x) \tag{17}$$

is

$$y = -e^{\int_0^x F\,d\tau} \int e^{\int_0^x F\,d\tau} G(x)\,dx,$$

which is computed by multiplying (17) by $e^{\int_0^x F\,d\tau}$ and integrating both sides of the equality.

The symmetry method explains the presence of $e^{\int_0^x F\,d\tau}$ in the solution of (17): Since $y_h := e^{\int_0^x F\,d\tau}$ is a solution to the homogeneous equation

$$y' + F(x)y = 0,$$

(17) is symmetric under the family of transformations $\phi_t : (x, y) \mapsto (x, y + ty_h(x))$. In other words, if $y(x)$ is a solution to (17), then so is its image under ϕ_t, $y(x) + ty_h(x)$, because

$$\begin{aligned}\frac{d}{dx}\big[y(x) + ty_h(x)\big] + F(x)\big[y(x) + ty_h(x)\big] \\ &= y(x) + F(x)y(x) + \big[ty_h(x)' + tF(x)y_h(x)\big] \\ &= y(x) + F(x)y(x) = G(x).\end{aligned}$$

Since the orbits of ϕ_t are already vertical lines, we set $r := x$. Since $\xi = 0$ and $v'(0) = y_h(t)$, (12) becomes $s_y = y_h(x)^{-1}$ with solution $s(x, y) = yy_h(x)^{-1}$. In the (r, s) coordinate system, (17) becomes

$$\begin{aligned}\frac{ds}{dr} &= \frac{s_x dx + s_y dy}{r_x dx + r_y dy} = \frac{-\frac{yy_h'}{y_h^2}dx + \frac{1}{y_h}dy}{dx} \\ &= -\frac{y\,(-F(x)y_h)}{y_h^2} + \frac{1}{y_h}\left(\frac{dy}{dx}\right) \\ &= -\frac{-F(x)y}{y_h} + \frac{1}{y_h}\big[G(x) - F(x)y\big] = \frac{G(x)}{y_h(x)} = \frac{G(r)}{y_h(r)},\end{aligned}$$

with solution $s(r) = \int \frac{G(r)}{y_h(r)}\,dr = \int \frac{G(r)}{e^{\int_0^x F\,d\tau}}\,dr$.

Reduction of order In this example, symmetry is used to lower the order of an ODE from two to one.

The homogeneous second order linear equation

$$y_{xx} + p(x)y_x + q(x)y = 0 \tag{18}$$

is invariant under the family of transformations $\phi_t : (x, y) \mapsto (x, e^t y)$ because the scalar factors out of each term:

$$(e^t y)_{xx} + p(x)(e^t y)_x + q(x)(e^t y) = e^t\big(y_{xx} + p(x)y_x + q(x)y\big) = 0.$$

The orbits of ϕ_t are already vertical so that $r := x$ is constant on the orbits. Since $\xi = 0$ and $\eta = y$, (12) has solution $s := \ln|y|$. In these coordinates, (18) becomes

$$s_{xx} + s_x^2 + p(x)s_x + q(x) = 0. \tag{19}$$

Since (19) is independent of s, we can let $z := s_x$ and reduce the order of (19) to the first order ODE

$$z_x + z^2 + p(x)z + q(x) = 0.$$

Converting a partial differential equation into an ordinary differential equation For ODEs, a family of symmetries reduces the degree of the ODE by one. Analogously, symmetries can reduce a PDE of n variables to a PDE of $n - 1$ variables.

For example, the so-called *transport equation*

$$z_x + z_y = k, \quad k \in \mathbb{R} \tag{20}$$

is symmetric under $\phi_t : (x, y, z) \mapsto (u, v, z) = (x + t, y + t, z)$ because

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x} = \frac{\partial z}{\partial u} \quad \text{and} \quad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial y} = \frac{\partial z}{\partial v}$$

so that $z_u + z_v = k$ whenever $z_x + z_y = k$. Geometrically, this means that ϕ_t maps the solution *surface* given by $z(x, y)$ to another solution surface. The functions $r := y - x$ and $s := \frac{1}{2}y + \frac{1}{2}x$ and $q(x, y, z)$ are canonical coordinates for $\phi_t : (x, y, z) \mapsto (x + t, y + t, z)$.

In the (r, s, q) coordinate system,

$$u_x = \frac{\partial u}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial u}{\partial s}\frac{\partial s}{\partial x} = -u_r + \frac{1}{2}u_s \quad \text{and}$$
$$u_y = \frac{\partial u}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial u}{\partial s}\frac{\partial s}{\partial y} = u_r + \frac{1}{2}u_s.$$

Therefore, $k = u_x + u_y = u_s$ and in the (r, s, q) coordinate system, (20) is equivalent to the ODE $k = u_s$.

Resources for further reading

We have only touched upon some of the fundamental ideas of the symmetry method: computing symmetries, transforming arbitrary symmetries to translational symmetries, and reducing first-order ODEs to integration. There are many facets of the method that have not been discussed here, such as the linearization of the symmetry condition (6) and the applications to higher order ODEs and PDEs. Fortunately, there are many resources available for the interested reader, such as Olver's classic text *Applications of Lie Groups to Differential Equations* [**14**], which provided material for the sections on homogeneous coordinates and reduction of order.

For the reader who has had vector calculus and elementary differential equations, we recommend Hydon's *Symmetry Methods for Differential Equations* [**10**], from which equations (4), (7), and (10) appear as examples or exercises. Readers who have had some exposure to differential geometry may enjoy Bluman and Kumei's *Symmetries of Differential Equations* [**3**], from which the explanation on integrating factors is taken. Experienced researchers will find the *CRC Handbook of Lie Group Analysis of Differential Equations* [**11**], edited by Ibragimov, a useful resource that covers Lie-Bäcklund, conditional and non-classical symmetries, approximate symmetry groups for equations with a small parameter, group analysis of differential equations with distributions, integro-differential equations, recursions, and symbolic software packages.

Miller's book *Symmetry and Separation of Variables* discusses symmetry applied to the Helmholtz, heat, wave, Laplace, and Schrodinger's equation and other equations of mathematical physics. Hawkins' *Emergence of the Theory of Lie Groups*, from which much of the introduction to this article is based, provides a thorough history of the symmetry method and the theory of transformation groups. Connections between Lie and Klein can be found in *Felix Klein and Sophus Lie* by I. M. Yaglom.

REFERENCES

1. G. Birkhoff, *Hydrodynamics: A Study in Logic, Fact, and Similitude*, Greenwood Press, 1978.
2. Rafael D. Benguria and Cecilia Yarur, Symmetry properties of the solutions to Thomas-Fermi-Dirac-Von Weizsäcker type equations, *Trans. Amer. Math. Soc.*, **320**:2 (1990) 665–675. doi:10.2307/2001695
3. G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Springer-Verlag, New York, 1989.
4. Y. Boutros, M. Abd-El-Malek, N. Badran, and H. Hassan, Lie-group method for unsteady flows in a semi-infinite expanding or contracting pipe with injection or suction through a porous wall, *J. Comput. Appl. Math.* **197**(2) (2006) 465–494. doi:10.1016/j.cam.2005.11.031
5. Harry H. Denman and William J. Larkin III, Invariance conditions on ordinary differential equations defining smoothing functions, *SIAM J. Appl. Math.* **17**(6) (1969) 1246–1257. doi:10.1137/0117116

6. P. P. Boyle, W. Tian, and F. Guan, The Riccati equation in mathematical finance, *J. Symbolic Comput.* **33**(3) (2002) 343–355. doi:10.1006/jsco.2001.0508
7. B. Fritzsche, Sophus Lie: A sketch of his life and work, *J. Lie Theory* **9** (1999) 1–38.
8. W. Getz, Invariant structures in a system of competitive logistic ordinary differential equations, *SIAM J. Appl. Math.* **36**(2) (1979) 321–333. doi:10.1137/0136026
9. T. Hawkins, *Emergence of the Theory of Lie Groups: An Essay in the History of Mathematics, 1869–1926*, Springer, New York, 2000.
10. P. Hydon, *Symmetry Methods for Differential Equations: A Beginner's Guide*, Cambridge Texts in Applied Mathematics, Cambridge Univ. Press, New York, 2000.
11. N. H. Ibragimov, ed., *CRC Handbook of Lie Group Analysis of Differential Equations*, CRC Press, Boca Raton, FL, 1994.
12. L. P. Lystad, P. Nyman, and R. Høibakk, The Riccati equation—an economic fundamental equation which describes marginal movement in time, *Model. Identif. Control* **27**(1) (2006) 3–21. doi:10.4173/mic.2006.1.1
13. Willard Miller, Jr., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.
14. P. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1986.
15. F. Schwarz, Symmetries of differential equations: from Sophus Lie to computer algebra, *SIAM Rev.* **30**(3) (1988) 450–481. doi:10.1137/1030094
16. D. A. Sánchez, *Ordinary Differential Equations: A Brief Eclectic Tour*, MAA, Washington, DC, 2002.
17. C. Wang and Z. Rusak, Similarity solutions of $\phi_x{}^3\phi_{xx} = \phi_{\tilde{y}\tilde{y}}$ with applications to transonic aerodynamics of dense gasses, *SIAM J. Appl. Math.* **59**(2) (1998) 514–528. doi:10.1137/S0036139997314824
18. I. M. Yaglom, *Felix Klein and Sophus Lie: Evolution of the Idea of Symmetry in the Nineteenth Century*, Birkhauser, Boston, 1988.
19. Z. Yan, The Riccati equation with variable coefficients expansion algorithm to find more exact solutions of nonlinear differential equations, *Comput. Phys. Comm.* **152**(1) (2003) 1–8. doi:10.1016/S0010-4655(02)00756-7
20. M. Zelikin, *Control Theory and Optimization I: Homogeneous Spaces and the Riccati Equation in the Calculus of Variations*, Springer, New York, 2000.

Summary Differential equations often seems like a hodgepodge of techniques designed to solve very specific equations. There is, however, a unifying theme for many of those techniques, which is to find the right coordinate system with which to express the equation. In the mid-nineteenth century, Sophus Lie discovered how to unify and extend many of the techniques that existed at his time by using the symmetries of differential equations to find good coordinate systems. This paper explains the symmetry method at a level suitable for undergraduates who have taken vector calculus and differential equations.

SHIRLEY LLAMADO YAP received her Ph.D. from the University of Pennsylvania in differential geometry and is now an assistant professor at the California State University, East Bay. Her research interests include differential geometry, differential equations on manifolds, mathematical biology and mathematical finance. Her nonmathematical interests include modern ballet, jazz, lindy-hop, ultimate frisbee, Aikido and working with the homeless. She was born in the Philippines and considers herself very lucky to have become, against many odds, a mathematician.

A Permutation of Features

Sharp-eyed readers will notice that we have moved the author information from the inside front cover to the ends of the Articles, and that we are now including summaries (abstracts) at the ends of Articles and Notes. We hope you don't mind, but these changes are mainly for the convenience of electronic storage systems. We want those systems to make the summaries widely searchable, and to keep the author information with the rest of the content. We're including a "citation line" on the first page of each feature for similar reasons. —*Ed.*

$f(f(x)) = -x$, Windmills, and Beyond

MARTIN GRIFFITHS
University of Manchester
Manchester, M13 9PL, United Kingdom
martin.griffiths@manchester.ac.uk

Before reading any further, I invite you to find a function $f : \mathbb{R} \to \mathbb{R}$ satisfying the equation $f(f(x)) = -x$ for all $x \in \mathbb{R}$.

After a few moments thought you might come to the same conclusion as I did, namely that this is not an entirely trivial problem. It is an example of a *functional equation*, for which the solutions are functions of x rather than values of x. Such an equation generally expresses a relationship between the value of a function at some point and its values at other points. We tend only to use the term 'functional equation' for equations that are not, in a simple sense at least, reducible to algebraic equations.

Of course, if we allowed f to be complex-valued then the solution $f(z) = iz$ is immediate. This highlights a common trait amongst functional equations; namely that the ease of solution often depends on the domain of the function. For example, Cauchy's functional equation, $f(x + y) = f(x) + f(y)$, is easy to solve if we restrict x and y to the rationals (see [**4**, p. 1]), but its solution over the real numbers turns out to be a rather more difficult problem. Recurrence relations, such as the Fibonacci recurrence $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$ with $F_1 = F_2 = 1$, may be thought of as functional equations for which the domain of the function is $\mathbb{N}$. When such restrictions are imposed on a function they may, in addition to affecting the ease of solution, determine the number of solutions possessed by a particular functional equation. If it has any solutions at all, there may be a finite number of them, or infinitely many.

In this article we explore the equation $f(f(x)) = -x$ and some generalizations. This functional equation is appealing because it is non-trivial yet accessible, and has some noteworthy results associated with it. Of particular interest to us is the way that the enforced graphical symmetry of any solution to the equation implies that it must possess a certain non-obvious analytic property. Indeed, a central theme here is this interplay of the visual with the analytic aspects of the problem. We consider the general properties shared by functions satisfying this equation, as well as specific solutions.

Some history

Functional equations may, in spirit if not in name, be traced back to antiquity. Indeed, Archimedes used recurrence relations to obtain ever better approximations to π. In the 14th century the French philosopher and mathematician Nicole Oresme used what was essentially a functional equation to give an indirect definition of linear functions. Possibly the most well-known functional equation is

$$g(1 + z) = zg(z), \tag{1}$$

which is satisfied by Euler's gamma function $\Gamma(z)$ for all $z \in \mathbb{C}$ except the non-positive integers. It is worth noting however, that the gamma function is not the only solution to this equation. Indeed, the function $g(z) = 0$ for all $z \in \mathbb{C}$ also fits the bill, among

Math. Mag. **83** (2010) 15–23. doi:10.4169/002557010X479956.

others. Nonetheless, it is true that $\Gamma(z)$ is the unique solution to the system of three functional equations consisting of (1) and the two further equations given by:

$$\frac{2^{2z-1}}{\sqrt{\pi}} g(z) g\left(z+\frac{1}{2}\right) = g(2z) \quad \text{and} \quad g(z)g(1-z) = \frac{\pi}{\sin(\pi z)}.$$

Another famous historical example,

$$h(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) h(1-s),$$

is satisfied by the Riemann zeta function $\zeta(z)$. A form of this equation can be seen in Riemann's seminal 1859 paper on the distribution of the primes [**7**]. Papers on functional equations appeared sporadically throughout the first half of the 20th century, but it was not until 1968 that the first book explicitly devoted to the subject was published [**3**]. A short biography of the author of this book, the Polish mathematician Marek Kuzcma, can be found on the website [**5**]. Nowadays of course, this area is very much more in the mainstream of mathematics; indeed, even bright teenagers are exposed to functional equations via mathematical olympiads [**2, 6**]. These equations have applications in fields as diverse as geometry, engineering, economics, probability and statistics. See [**1**] for further details concerning the history of functional equations and their widespread applications.

A few simple properties of f

The following function is tantalizingly close to satisfying our requirements:

$$f(x) = \begin{cases} -\frac{1}{x} & \text{if } x \geq 1 \\ \frac{1}{x} & \text{if } 0 < x < 1 \\ 0 & \text{if } x = 0 \\ \frac{1}{x} & \text{if } -1 \leq x < 0 \\ -\frac{1}{x} & \text{if } x < -1 \end{cases}.$$

It gives $f(f(x)) = -x$ for all real numbers except for $x = -1$, and maps onto all real numbers except for 1. The graph of $y = f(x)$ is shown in FIGURE 1. Is it possible to tweak f slightly in order to obtain a solution? This is something we consider in due course. Let us first ascertain some simple properties necessarily possessed by any function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(f(x)) = -x$ for all $x \in \mathbb{R}$, assuming such a function exists!

Possibly the most obvious property of f is that it is onto. It must also be true that

$$f(0) = 0.$$

In order to show this, assume to the contrary that $f(0) = t$ for some $t \neq 0$. Then $f(t) = f(f(0)) = 0$ and thus $f(f(t)) = f(0) = t$, contradicting the definition of f.

A clue to what is essentially the defining property of the graph of f may be gleaned from FIGURE 1. Indeed:

$$\text{The graph } y = f(x) \text{ has rotational symmetry of order 4,} \tag{2}$$

where the center of rotation is at the origin. This statement is taken to mean that in rotating the graph of $y = f(x)$ about the origin (either clockwise or anticlockwise),

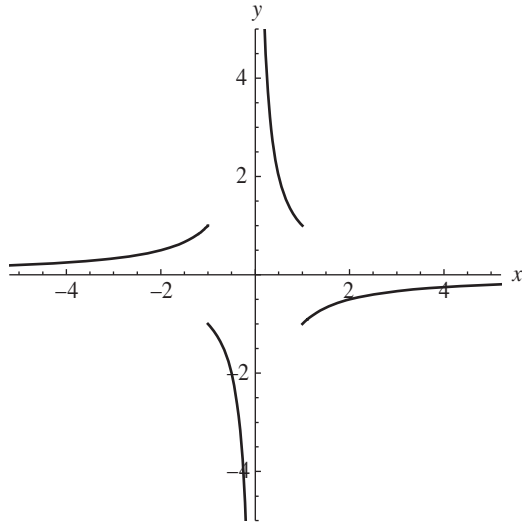


Figure 1 A function that almost works

we obtain an exact copy of $y = f(x)$ precisely when the angle of rotation is either $\pi/2$, π, $3\pi/2$ or 2π. To see that (2) is in fact true, choose any $a \neq 0$. Then $f(a) = b$ for some $b \in \mathbb{R}$. From the definition of f, $f(b) = f(f(a)) = -a$, $f(-a) = f(f(b)) = -b$ and $f(-b) = f(f(-a)) = a$. Thus the graph contains the points (a, b), $(b, -a)$, $(-a, -b)$ and $(-b, a)$. These points form the vertices of a square whose center is the origin, as required.

From (2) and the fact that $f(0) = 0$, it follows immediately that the only point of intersection of $y = f(x)$ with the x-axis occurs at the origin. This in turn implies that

$$f(x) \text{ is discontinuous for at least one value of } x,$$

as we now explain. Suppose, to the contrary, that $f(x)$ is continuous for all $x > 0$. Then, since the curve $y = f(x)$ cannot cross the x-axis when $x > 0$, it must, for $x > 0$, lie either entirely in the first quadrant or entirely in the fourth quadrant. This, however, contradicts (2), as required.

It must also be the case that

$$f(x) \text{ is a bijective function.}$$

Since $f : \mathbb{R} \to \mathbb{R}$ is onto, all we need do is show that f is necessarily one-to-one. If $f(a) = f(b)$ then $-a = f(f(a)) = f(f(b)) = -b$ so that $a = b$, showing that f is indeed one-to-one.

Let us now return to our initial attempt at finding a suitable function f. We can eliminate the problem at $x = -1$ by redefining f on the non-zero integers as follows. With $n \in \mathbb{N}$ then

$$f(k) = \begin{cases} -2n & \text{if } k = 2n - 1 \\ -2n + 1 & \text{if } k = -2n \\ 2n & \text{if } k = -2n + 1 \\ 2n - 1 & \text{if } k = 2n \end{cases}.$$

Of course, this in turn causes a problem with numbers of the form $\frac{1}{n}$ where $n \in \mathbb{Z}$ such that $n \geq 2$ or $n \leq -2$. However, this is easily dealt with. With $n \in \mathbb{N}$ we redefine f by

$$f(k) = \begin{cases} -\frac{1}{2n+1} & \text{if } k = \frac{1}{2n} \\ -\frac{1}{2n} & \text{if } k = -\frac{1}{2n+1} \\ \frac{1}{2n+1} & \text{if } k = -\frac{1}{2n} \\ \frac{1}{2n} & \text{if } k = \frac{1}{2n+1} \end{cases},$$

giving a function that does indeed satisfy $f(f(x)) = -x$.

Another way of defining f is to set $f(0) = 0$ and then cycle the other numbers amongst themselves within certain intervals. For example, the numbers in $[-2, 0) \cup (0, 2]$ can be mapped to each other as follows. If $x \in (0, 1]$ then $f(x) = x + 1$, $f(x + 1) = -x$, $f(-x) = -x - 1$ and $f(-x - 1) = x$. In general, with m a positive integer and $x \in (2m - 2, 2m - 1]$, the given mapping can be used for the numbers in $[-2m, -2m + 2) \cup (2m - 2, 2m]$. The graph of this function looks a bit like the sails of an infinite windmill (with gaps), part of which is shown in FIGURE 2.

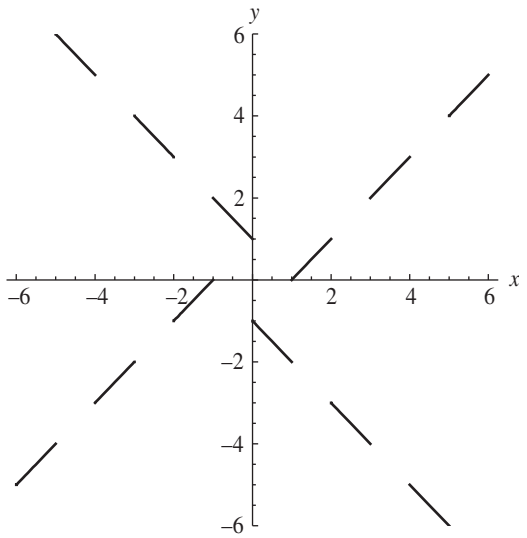


Figure 2 A 'windmill' function

Delving a little deeper

Both of the solutions found in the previous section possessed infinitely many discontinuities. We may next ask ourselves whether it is possible to find a solution that has only finitely many. The answer turns out to be an emphatic "No." In fact,

> Every function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(f(x)) = -x$ for all $x \in \mathbb{R}$ has infinitely many discontinuities.

To explain why this is so requires rather more intricate reasoning than that used thus far, and is indeed the primary aim of this section.

In order both to set the scene and to provide further motivation for the constructions appearing later in this article, we start by outlining the main ideas behind the proof. Suppose, for example, that f is continuous on some open interval (a, b). Then, under a rotation of $\pi/2$, the induced image of (a, b) is some disjoint interval (c, d). Successive rotations then give induced images $(-b, -a)$, $(-d, -c)$ and (a, b). The simplest scenario might be something like $(0, 1)$ to $(1, \infty)$ to $(-1, 0)$ to $(-\infty, -1)$. Since $f(0) = 0$ we just need to deal with the points $x = 1$ and $x = -1$ in this case. However, the addition of one more point will result in the addition of another three, so this will not quite work. The detailed proof that follows merely expands on this idea of counting points and intervals under the assumption that there are only finitely many discontinuities. All potential arrangements of intervals are accounted for, and it is shown that such a mismatch always occurs under this assumption, thereby providing a contradiction.

We already know that $f(x)$ is discontinuous for at least one positive value of x. Let us now suppose that f has only a finite number of discontinuities. Then there exists some positive $r \in \mathbb{R}$ such that f is discontinuous at $x = r$ but continuous for all $x > r$. Since the graph of f does not intersect the positive x-axis we may assume, without loss of generality, that $f(x) > 0$ for all $x > r$, for if this were not the case then we would simply replace f with $-f$. The portion of the graph $y = f(x)$ lying in the first quadrant, $\{(x, y) : x > 0, y > 0\}$, consists of:

(i) A finite number, m say, of isolated points. If $m \geq 1$ we denote these points by $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$.

(ii) The infinite continuous curve C_∞, which is of the form $\{(x, f(x)) : x \geq r\}$ or $\{(x, f(x)) : x > r\}$. We define I_{C_∞} either as $[r, \infty)$ or (r, ∞) accordingly.

(iii) A finite number, k say, of continuous curve segments, each of the form $\{(x, f(x)) : a \leq x \leq b\}$, $\{(x, f(x)) : a < x \leq b\}$, $\{(x, f(x)) : a \leq x < b\}$ or $\{(x, f(x)) : a < x < b\}$ for some $a, b \in \mathbb{R}$ with $0 \leq a < b \leq r$. Let us denote these curves and their associated intervals by C_j and I_j respectively, $j = 1, 2, \ldots, k$. The interval I_j is either (a, b), $(a, b]$, $[a, b)$ or $[a, b]$ for some $a, b \in \mathbb{R}$ with $a < b$, depending on the form of C_j. We say, for example, that the interval $(a, b]$ has an *open lower endpoint* and a *closed upper endpoint*.

Note that the existence of C_∞ in conjunction with (2) precludes each C_j from being infinite, and also that (2) implies that any given C_j will be strictly monotone over I_j.

From (2) we know that on rotating the portion of the graph of $y = f(x)$ in the first quadrant through $\pi/2$ clockwise about the origin we obtain the portion of the graph of $y = f(x)$ in the fourth quadrant. Under this rotation isolated points get mapped to isolated points, finite continuous curve segments defined on closed-closed intervals get mapped to finite continuous curve segments defined on closed-closed intervals, and so on (there is of course the possibility that curves defined on open-closed intervals get mapped to curves defined on closed-open intervals, and vice versa). We shall use $\overline{C}_j$ and $\overline{I}_j$ to denote the image of C_j and the induced image of I_j respectively, and $\overline{(x_j, y_j)}$ to denote the image of (x_j, y_j).

We next need to consider the image of C_∞ in the fourth quadrant, $\overline{C}_\infty$ say, under the aforementioned rotation. Since f is continuous and one-to-one on (r, ∞), it must be strictly monotone for $x > r$. If f is decreasing then it is bounded below by 0. If, on the other hand, f is increasing then, from (2), it certainly cannot exceed r. Either way, we must have that

$$\lim_{x \to \infty} f(x) = a$$

for some $a \in \mathbb{R}$ with $a \geq 0$. Note that the fact that f is strictly monotone on (r, ∞) means that $f(x)$ can never actually equal a. Thus the interval, $I_{\overline{C}_\infty}$ say, over which the infinite continuous curve $\overline{C}_\infty$ is defined, is either a finite closed-open, open-closed or open-open interval (the latter if, and only if, C_∞ is of the form $\{(x, f(x)) : x > r\}$).

For the remainder of the argument, it is important to realize that the intervals defined in (ii) and (iii) each have two associated endpoints. Even if two intervals are contiguous, such as $(a, b]$ and (b, c) or (a, b) and (b, c) for example, they would still have total of four endpoints between them (three open and and one closed for the first pair, and four open for the second pair).

Let S denote the set of all $2k + 2$ intervals defining segments of continuous curves in the graph $y = f(x)$ for $x > 0$. The intervals in S may be ordered according to where they lie on the positive x-axis. To this end, let J_1 be the unique element of S with g.l.b. equal to zero, J_2 be the unique element of S whose g.l.b. is equal to the l.u.b. of J_1, and so on. Suppose that for some $i \in \{1, 2, \ldots, 2k + 1\}$ the l.u.b. of J_i is a. Then there is an isolated point with x-coordinate a if, and only if, the upper endpoint of J_i and lower endpoint of J_{i+1} are both open. If, on the other hand, there is no such isolated point then either the upper endpoint of J_i is open and the lower endpoint of J_{i+1} is closed or vice versa. Here is an example of a possible ordering of the first few intervals/isolated points, where $0 < x_1 < x_2 < \cdots$:

$$(0, x_1]\,(x_1, x_2)\,x_2\,(x_2, x_3)\,[x_3, x_4)\,[x_4, x_5]\ldots$$

We are now in a position to obtain a contradiction and hence complete the proof of this result. Here is a summary of the key points that arise, under the assumption that f has only a finite number of discontinuities (remembering that infinity does count as an endpoint for an interval):

(a) The total number of endpoints of all the intervals contained in S is $2(2k + 2) = 4k + 4$, which is a multiple of 4.

(b) The number of closed endpoints amongst all the intervals in S is even, say $2n$ for some non-negative integer n (since, by (2), there will be equal contributions from curve segments in the first and fourth quadrants). These $2n$ closed endpoints correspond to $2n$ open endpoints from elements in S.

(c) The $2m$ isolated points in the graph of $y = f(x)$ for $x > 0$ correspond to $4m$ open endpoints from elements in S.

(d) Both the lower endpoint of J_1 and the upper endpoint of J_{2k+2} are open.

From (b), (c) and (d) there must be $2n + 4m + 2$ open endpoints amongst all the intervals in S. The total number of endpoints in S is therefore $2n + (2n + 4m + 2) = 4(n + m) + 2$, which is not a multiple of 4. This, however, contradicts (a), so our supposition that f has only a finite number of discontinuities must be false.

We now go on to show, constructively, that for each interval containing zero there are in fact uncountably many solutions to $f(f(x)) = -x$ possessing only countably many discontinuities such that all of these discontinuities lie within this interval. Let $c \in \mathbb{R}$ with $c > 0$. For any $0 < \epsilon < 1/\sqrt{c}$ we may construct a strictly increasing sequence $\{a_n\}$ such that

$$\frac{1}{\sqrt{c}} - \epsilon < a_1 < \frac{1}{\sqrt{c}} < \frac{1}{a_1 c} < \frac{1}{\sqrt{c}} + \epsilon \quad \text{and} \quad \lim_{n\to\infty} a_n = \frac{1}{\sqrt{c}}.$$

First we set $f\left(\frac{1}{\sqrt{c}}\right) = a_1$, $f(a_1) = -\frac{1}{\sqrt{c}}$, $f\left(-\frac{1}{\sqrt{c}}\right) = -a_1$, $f(-a_1) = \frac{1}{\sqrt{c}}$ and $f(0) = 0$. Next, for $k \in \mathbb{N}$ we define

$$f(x) = \begin{cases} a_{2k+1} & \text{if } x = a_{2k} \\ -a_{2k} & \text{if } x = a_{2k+1} \\ -a_{2k+1} & \text{if } x = -a_{2k} \\ a_{2k} & \text{if } x = -a_{2k+1} \end{cases}$$

and, with $b_i = 1/(a_i c)$ for $i = 1, 2, \ldots$,

$$f(x) = \begin{cases} b_{2k} & \text{if } x = b_{2k-1} \\ -b_{2k-1} & \text{if } x = b_{2k} \\ -b_{2k} & \text{if } x = -b_{2k-1} \\ b_{2k-1} & \text{if } x = -b_{2k} \end{cases}$$

Finally, for any values of x not included above, we set

$$f(x) = \begin{cases} -\frac{1}{cx} & \text{if } x > \frac{1}{\sqrt{c}} \\ \frac{1}{cx} & \text{if } 0 < x < \frac{1}{\sqrt{c}} \\ \frac{1}{cx} & \text{if } -\frac{1}{\sqrt{c}} < x < 0 \\ -\frac{1}{cx} & \text{if } x < -\frac{1}{\sqrt{c}} \end{cases}.$$

This defines a function f that maps the real numbers onto the real numbers and satisfies the equation $f(f(x)) = -x$. Given any open interval containing zero we may, by choosing sufficiently large c and sufficiently small ϵ, restrict the discontinuities of f to lie within this interval. Furthermore, there are uncountably many values of c for which this could be done.

Generalizing somewhat

It is also true that any function $f : \mathbb{R} \to \mathbb{R}$ satisfying the equation $f(f(x)) = a - x$ necessarily has infinitely many discontinuities, and that for each interval containing $a/2$ there are uncountably many such functions possessing only countably many discontinuities such that all of these discontinuities lie within this interval. To see this, suppose that $f(b) = c$. Then $f(c) = f(f(b)) = a - b$, $f(a - b) = f(f(c)) = a - c$ and $f(a - c) = f(f(a - b)) = b$. Thus the graph contains the points (b, c), $(c, a - b)$, $(a - b, a - c)$ and $(a - c, b)$. These points form the vertices of a square whose center is at $(a/2, a/2)$. The claim now follows from the results obtained in the previous two sections, on noting that the graph of $y = f(x)$ has rotational symmetry of order 4 about the point $(a/2, a/2)$.

We next give two examples of functions $f : \mathbb{R} \to \mathbb{R}$ that satisfy the equation $f(f(x)) = -cx$ for $c > 0$. Let $f(a) = b$ where a and b are positive real numbers such that $a < b$. From the property $f(f(x)) = -cx$, the points

$$\left(a(-c)^n, b(-c)^n\right) \quad \text{and} \quad \left(b(-c)^n, a(-c)^{n+1}\right), \quad n \in \mathbb{Z},$$

all lie on the graph of $y = f(x)$. Suppose, for the time being at least, that $c > 1$, and define the points A_{2m}, B_{2m}, A_{2m+1} and B_{2m+1} to have coordinates

$$\left(a(-c)^m, b(-c)^m\right), \quad \left(b(-c)^m, a(-c)^{m+2}\right), \quad \left(b(-c)^m, a(-c)^{m+1}\right),$$
$$\text{and} \quad \left(a(-c)^{m+2}, b(-c)^{m+1}\right)$$

respectively, where $m \in \mathbb{Z}$. Now let $\overline{A_{2m}B_{2m}}$ represent the line segment connecting A_{2m} and B_{2m}, and containing A_{2m} but not B_{2m}. The line segment $\overline{A_{2m+1}B_{2m+1}}$ is defined

similarly. It is easily checked that f induces a mapping from $\overline{A}_{2m}B_{2m}$ to $\overline{A}_{2m+1}B_{2m+1}$, and from $\overline{A}_{2m+1}B_{2m+1}$ to $\overline{A}_{2(m+1)}B_{2(m+1)}$. Then, along with the point $f(0) = 0$, the lines

$$\overline{A}_{2m}B_{2m} \quad \text{and} \quad \overline{A}_{2m+1}B_{2m+1}, \quad m \in \mathbb{Z},$$

give the graph of a function f satisfying all the requirements. An example of such a graph is given in FIGURE 3. In this case $a = 2$, $b = 3$ and $c = 2$. There are several things worth noting here:

1. All the line segments forming the graph of $y = f(x)$ in quadrants 1 and 3 have the same gradient, and similarly for the line segments in quadrants 2 and 4.
2. For any positive number t, there exist infinitely many disjoint intervals of the form $[s, s + t]$, where $s \in \mathbb{R}$, such that f is continuous over each of these intervals.
3. For any $\epsilon > 0$, f has infinitely many discontinuities on the interval $(-\epsilon, \epsilon)$.

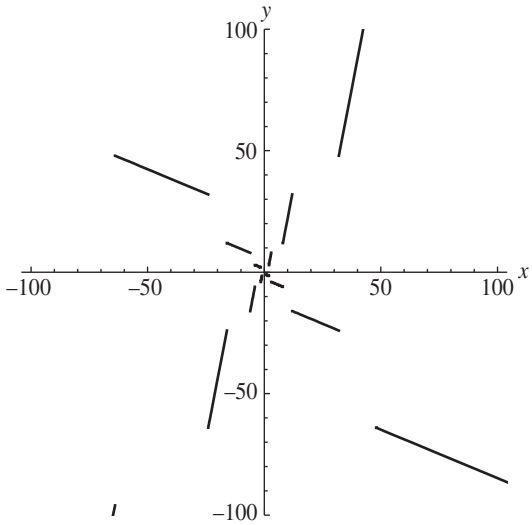


Figure 3 A typical 'distorted windmill' function

If $a > b$ then we define the graph of an alternative function $g(x)$, as follows. The points D_{2m}, E_{2m}, D_{2m+1} and E_{2m+1} are defined to have coordinates

$$\big(b(-c)^m, a(-c)^{m-2}\big), \quad \big(a(-c)^m, b(-c)^m\big), \quad \big(a(-c)^{m-2}, b(-c)^{m+1}\big),$$
$$\text{and} \quad \big(b(-c)^m, a(-c)^{m+1}\big)$$

respectively. Now let $D_{2m}\overline{E}_{2m}$ represent the line segment connecting D_{2m} and E_{2m}, and containing E_{2m} but not D_{2m}. The line segment $D_{2m+1}\overline{E}_{2m+1}$ is defined similarly. Then, along with the point $g(0) = 0$, the lines

$$D_{2m}\overline{E}_{2m} \quad \text{and} \quad D_{2m+1}\overline{E}_{2m+1}, \quad m \in \mathbb{Z},$$

give the graph of a function g satisfying all the requirements. With $a = 3$, $b = 2$ and $c = 2$ we obtain the graph shown in FIGURE 4. It is possible to obtain similar graphs when $0 < c < 1$.

We leave the reader to explore the functional equation $f(f(x)) = d - cx$ for the possible existence of even stranger windmills and beyond.

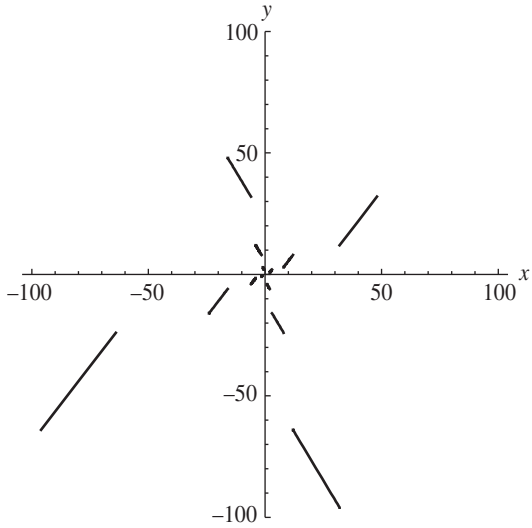


Figure 4 Another 'distorted windmill' function

REFERENCES

1. J. Aczel, *Functional Equations: History, Applications and Theory*, D. Reidel, Dordrecht, 1984.
2. A. Gardiner, *The Mathematical Olympiad Handbook*, Oxford University Press, Oxford, 1997.
3. M. Kuczma, *Functional Equations in a Single Variable*, PWN-Polish Scientific, Warsaw, 1968.
4. K. Y. Li, Functional equations, *Mathematical Excalibur* **8**(1) (February 2003); available at `http://www.math.ust.hk/excalibur/v8_n1.pdf`.
5. J. J. O'Connor and E. F. Robertson, Biography of Marek Kuczma, `http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Kuczma.html`.
6. M. Radovanović, Functional equations, The IMO Compendium Group (2007); available at `http://www.imomath.com/tekstkut/funeqn_mr.pdf`.
7. B. Riemann, Über die Anzahl der Primzahlen unter einer gegebenen Grösse, *Monatsberichte der Berliner Akademie* (November 1859) 671–680.

Summary In this article we consider the solutions of a particular functional equation and some generalizations. This equation possesses non-trivial yet accessible solutions, and is rather appealing because of the way that the enforced graphical symmetry of any solution implies that it must possess a certain non-obvious analytic property. Indeed, a central theme here is this interplay of the visual with the analytic aspects of the problem. After giving a brief historical overview of functional equations, we go on to study the general properties shared by functions satisfying our particular equation, as well as specific solutions. Finally, solutions to a generalization of the equation are obtained.

MARTIN GRIFFITHS is a Lecturer in Mathematics Education at the University of Manchester, having previously been simultaneously both a part-time Lecturer in Mathematics at the University of Essex and Head of Mathematics at a high school in Colchester. Prior to all of this, he had a career in the British Army.

His diverse interests range from mathematical epidemiology to aspects of combinatorics associated with partitions, and he has had almost sixty articles and papers published or accepted for publication. From the point of view of mathematics education, he is, amongst many other things, interested in how we may enthuse and challenge our most able students both at school and at university.

Galileo and Oresme: Who Is Modern? Who Is Medieval?

OLYMPIA NICODEMI
SUNY Geneseo
Geneseo, NY 14454
nicodemi@geneseo.edu

Some celebrate 1638 as the date that marks the marriage of mathematics and physics. That was the year Galileo published his *Discorsi e Dimostrazione Matematiche intorno a due Nuove Scienze*, or, as we shall call it, the *Two New Sciences*. In that work Galileo proves the law of free fall, familiar to us in the form $x(t) = gt^2/2$. As a mathematical relation, it describes how far an object would travel if it were uniformly accelerated from rest: The distance traveled is proportional to the square of the time elapsed; the proportionality constant g is not necessarily due to gravity and the direction is not necessarily down. As physics, this law makes a bolder and more tangible statement: This is how far an object released from your hand falls in time t (at least if it's near the earth and experiences no air resistance).

The mathematical aspect of this law has the longer history, dating back to the middle ages. In this article, we compare Galileo's mathematical treatment of uniform acceleration to that of Nicole Oresme, a 14th century scholastic philosopher and mathematician. Whose mathematics we grasp more easily may come as a surprise.

Galileo's law of free fall

The *Two New Sciences* is structured as a conversation held among educated friends over a period of four days. The friends, named Simplicio, Salviati, and Sagredo, discuss the writings of a "wise author" (Galileo himself). Days 3 and 4 are dedicated to the discussion of moving bodies, which they call *moveables*. On Day 3, the friends discuss the wise author's Theorem 1 and its corollary, Theorem 2, which together establish what we call the law of free fall [**3**, p. 165 ff.].

THEOREM 1. *The time in which a certain space is traversed by a moveable in uniformly accelerated movement from rest is equal to the time in which the same space would be traversed by the same moveable carried in uniform motion whose degree of speed is one-half the maximum and final degree of speed of the previous, uniformly accelerated, motion.*

THEOREM 2. *If a moveable descends from rest in uniformly accelerated motion, the spaces run through in any times whatever are to each other as the duplicated ratio of their times; that is, are as the squares of those times.*

Galileo was an accomplished draftsman who filled his works with many finely drawn and instructive diagrams. FIGURE 1 is similar to that which accompanies his Theorem 1. As we often do in the MAGAZINE, let's think of this figure as a "proof without words" and fill in the details. First, focus on the rectangle *GABF* and ignore the segment *CD* for now. When an object travels at a constant velocity v over an interval of time of length t, the distance traveled is $v \cdot t$. If we let v be the length of the base *FB* of the rectangle *GABF* and let t be its height (as Galileo did), then the distance traveled

Math. Mag. **83** (2010) 24–32. doi:10.4169/002557010X479965.

is its area. Now imagine an object that starts from rest but accelerates at a constant rate to a final velocity of v_f at time t. We can model this scenario with the right triangle *ABE* in FIGURE 1. Let v_f be its base. If we again reason that the distance traveled is the area of the figure, then that distance is $v_f \cdot t/2$. But now notice that if $v = v_f/2$, then $v_f \cdot t/2$ is exactly how far an object traveling at *constant* velocity $v_f/2$ would travel in time t. In Galileo's words, a uniformly accelerated object travels as far as it would have, had it been "carried in uniform motion whose degree of speed is one-half the maximum and final degree of speed." (Modern readers can, of course, prove the same thing with calculus.)

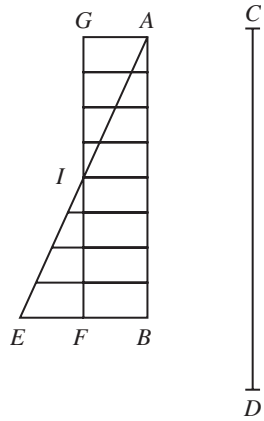


Figure 1 Galileo's diagram, where vertical distances represent time and horizontal ones represent velocity

To establish Theorem 2, we simply need to extend FIGURE 1 and count rectangles at each tick of a clock. Take a look at FIGURE 2. At the first tick, there is one rectangle; at the second tick, there are 3; at the third tick, there are 5, and so on. At the end of n ticks, there are $1 + 3 + \cdots + (2n - 1) = n^2$ rectangles. So at the end of t units of time, the distance traveled is t^2 times the amount traveled in the first unit of time.

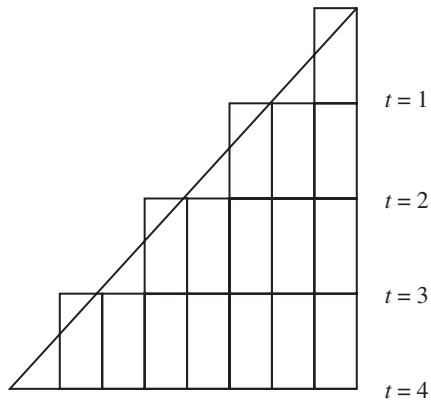


Figure 2 A diagram for Theorem 2, Galileo's law of free fall

It certainly is an intellectual treat to realize that such a basic physical law emerges simply and beautifully from elementary geometry. Our inner Plato smiles. But why

is something so simple a pivotal moment? At the time of Galileo, free fall was not assumed to be an example of uniformly accelerated motion. In fact, uniform acceleration itself had been correctly analyzed by medieval scholastics, as we will explain. It was Galileo that first showed that free fall—the motion of a body that is dropped from your hand—is indeed an example of uniformly accelerated motion. He did this with his famous inclined plane experiments, which are discussed later in the *Two New Sciences*. It is this link of mathematical theory to experimental verification that overturned the then prevalent Aristotelian tenet that mathematics was too exacting to model the complexity of nature. Within one hundred years, Newton would use the experimental observations of Kepler (that planets have elliptical orbits) and the mathematical strides of Descartes (the merging of algebra and geometry) to deduce the Universal Law of Gravitation.

A closer look Galileo postulated the free fall law when he was a young professor at the University of Padua around 1607, but he revisited the problem frequently over his long career. It seems that he felt he needed a deeper mathematical proof. We find that proof in the *Two New Sciences*, the culminating work of his career. So the very simplicity of the arguments presented above prompts us to question whether our facile deduction of Galileo's law reflects his own thinking. Surely the geometry we invoked was no hurdle for Galileo. So where is the nexus of proof for him? To gain some insight, we turn to Galileo's own words. They will sound strange to modern ears. Please experience the strangeness and please refer to FIGURE 1 again.

Galileo's Proof of Theorem 1. Let line *AB* represent the time in which the space *CD* is traversed by a moveable in uniformly accelerated movement from rest at *C*. Let *EB*, drawn in any way upon *AB*, represent the maximum and final degree of speed increased in the instants of time *AB*. All the lines reaching *AE* from single points of the line *AB* and drawn parallel to *BE* will represent the increasing speed after the instant *A*. Next, I bisect *BE* at *F*, and I draw *FG* and *AG* parallel to *BA* and *BF*; the parallelogram *AGFB* will [thus] be constructed, equal to the triangle.

Now if the parallels in triangle *AEB* are extended as far as *IG*, we shall have the aggregate of all parallels contained in the quadrilateral equal to the aggregate of those included in the triangle *AEB*, for those in triangle *IEF* are matched by those contained in triangle *GIA*, while those in the trapezium *AIFB* are common. Since each and all instants of time *AB* correspond to each point and all points of the line *AB*, from which points the parallels drawn and included within the triangle *AEB* represent increasing degrees of the increased speed, while the parallels contained within the parallelogram represent in the same way just as many degrees of speed not increased but equable, it appears that there are just as many momenta of speed consumed in the accelerated motion according to the increasing parallels in triangle *AEB*, as in the equable motion according to the parallels of the parallelogram *GB*. For the deficit of momenta in the first half of the accelerated motion (the momenta represented by the parallels in the triangle *AGI* falling short) is made up by the momenta represented by the parallels of the triangle *IEF*.

It is therefore evident that equal spaces will be run through in the same time by two moveables, of which one is moved with a motion uniformly accelerated from rest, and the other with equable motion having a momentum one-half the momentum of the maximum speed of the accelerated motion; which was [the proposition] intended. ■

The notion of area is not central to Galileo's proof. For him, velocity was not a quantity computed as quotient of distance and time. It was a basic primitive quality of a moving thing, a quality that had a continuous spectrum of possible intensities. If two objects in motion had velocities with unchanging intensities, distances traversed

in equal times would be in proportion to their intensities, but that distance would not be measured, computed, or quantified as *velocity* × *time* and so would not be quantified as the area of a rectangle. The issue of changing velocity was more complicated. For Galileo and his contemporaries, the distance traveled by an object with changing velocity is proportional to the *total* velocity accumulated, or as he says, the momenta accumulated and consumed as it traveled. We do not have such a concept in modern physics.

The key to Galileo's argument comes in the second paragraph of the proof. It begins with the assertion that the "aggregate of all parallels contained in the quadrilateral [is] equal to the aggregate of those included in the triangle *AEB*." To show that these aggregates (infinite sets really) are equal, he establishes what we would call a one-to-one correspondence between them: At each instant of the interval of time (represented by the segment *AB* in FIGURE 1), we can match two intensities, the degree (or intensity) of the speed of uniform motion and the degree of the speed of uniformly accelerated motion. These intensities are represented by the horizontal lines drawn in FIGURE 1. The aggregates are in one-to-one correspondence via the continuum of time "[s]ince each and all instants of time *AB* correspond to each point and all points of the line *AB*." By comparing the horizontal extents of the intensities in each case, Galileo sees that the excess of the intensities that lie within the top half of the rectangle but outside the triangle are matched by the excess of the intensities that lie within the bottom half of the triangle but not the rectangle. In conclusion, there is just as much "speed consumed" in the motion represented by the quadrilateral as that represented by the triangle. Thus, for Galileo, the distances traveled are the same for both the motions.

Galileo also refers to "uniformly accelerated motion" as "naturally accelerated motion," by which he means free fall. He finds it quite reasonable that nature should employ uniform acceleration for free fall because "she habitually employs the first, simplest, and easiest means" [**3**, p. 153]. But his contemporaries thought otherwise. So Galileo brackets Theorems 1 and 2. He precedes them with a mathematical refutation of the then current model of free fall and follows them with experimental evidence for his model.

First, the refutation. In their on-going reading of the wise author's words, the friends learn of his (Galileo's) assertion that the speed of an object dropped from rest increases in proportion to the time elapsed. The speaker Salviati, who represents Galileo's point of view among the discoursing friends, must refute the view prevalent in Galileo's time that in free fall, velocity is proportional to the distance fallen rather than to the time elapsed. Today, we would model the older view by the differential equation $dx/dt = cx$ where distance x is measured from the point of release and where c is a constant. The solution is $x(t) = Ae^{ct}$ where A depends on the initial condition, $x(0) = 0$. Thus $A = 0$ and the solution is $x(t) = 0$. No motion! The object stays put at the point of release. (This is rather like Road-Runner running off a cliff before he notices it is a cliff.) To dismiss this idea, Salviati first notes that in the case of uniform or constant velocity, if velocity is proportional to the distance to be traversed, the time it takes to travel "two braccia" or two arm-lengths would be the same as the time it takes to travel "four braccia". (If you travel 30 miles at 15 mph, it takes two hours. Keeping the same 2 to 1 proportion, but going 60 miles at 30 mph, it still takes two hours.) He extrapolates that argument to the case when velocity varies in proportion to distance traveled. Here is Salviati's (really Galileo's) argument [**3**, p. 160]:

> When speeds have the same ratio as the spaces passed or to be passed, those spaces come to be passed in equal times; if the speeds with which the falling body passed a space of four braccia were the doubles of the speeds with which it passes the first two braccia as one space is double the other space, then the times

> of those passages are equal; but for the same moveable to pass four braccia and two in the same time cannot take place.

Galileo makes a one-to-one correspondence between the intensities of the speeds that occur during a free fall that drops a distance of 2 braccia with the intensities during a fall that drops a distance 4 braccia. In the second case, each intensity is twice the intensity of the first. So the *total velocity* is twice the first case, and hence the amount of time it takes to go the double distance is the *same* amount of time as it takes to go the smaller distance. Clearly impossible! (And Galileo anticipates the use of the one-to-one correspondence used in the proof of Theorem 1.)

After he refutes the prevailing view of free fall, and after the presentation of the wise author's mathematical proof that, in naturally accelerated motion, distance is proportional to the square of the time elapsed, Salviati brings in the experimental evidence, Galileo's famous inclined plane experiment [**3**, p. 169].

> In a wooden beam or rafter about twelve braccia long, half a braccia wide, and three inches thick, a channel was rabbeted in along the narrowest dimension, a little over an inch wide and made very straight; so that this would be clean and smooth, there was glued within it a piece of vellum, as much smoothed and cleaned as possible. In this there was made to descend a very hard bronze ball, well rounded and polished.... [We] noted the time that it consumed in running all the way, repeating the process many times, in order to be quite sure as to the amount of time, in which we never found a difference of even the tenth part of a pulse-beat.... [We] made the same small ball descend only one-quarter of this channel, and the time of its descent being measured, this was found to be precisely one-half the other...

Galileo continues to elaborate, but our short quote is enough to verify the t^2 law for free fall. If $x = t^2$ (with units chosen so that $g = 2$), then $t = \sqrt{x}$. So the ball should go a quarter of the way in half the time, which he verified. Galileo's Proof of Theorem 1 and Salviati's arguments give us an interesting historical perspective. With them, Galileo laid the ground work for modern mathematical physics and its methodology: He developed a mathematical theory for a physical phenomenon and confirmed it with repeatable experiments. Yet he did so with physical concepts that are strange to us and mathematical techniques that are strangely employed.

Nicole Oresme

Now we look into the thinking of one of Galileo's medieval predecessors, Nicole Oresme (pronounced Orezzmay), an extraordinarily forward-looking French scholastic philosopher, who also addressed the issue of uniform acceleration. Oresme lived and worked in the 14th century, the same century that brought us the poetry of Dante and Chaucer, the art of Duccio and Giotto, the music of Guillaume de Machaut and Francesco Landini, and the Black Death. It was a period of intense scholarship, creativity, and tragedy.

Oresme's work, *De configurationibus qualitatum et motuum* or *The Geometry of Qualities and Motions*, is a treatise on the use of geometry both to imagine (model) and to compare qualities (aspects of an object) that can change in intensity. Velocity is such a quality. When a quality (like velocity) has constant intensity over an extension (in space or over an interval of time), Oresme called it *uniform*. If a quality changes in intensity, it said to be *difform*. There are two ways to be difform. If change is uniform

(as when velocity changes and acceleration is constant) then a quality is said to be *uniformly difform*. Otherwise, it is said to be *difformly difform*. What wonderful terminology! Oresme studies all three cases; we are interested in the quality of velocity as it changes intensity uniformly over time—uniformly difform velocity.

Oresme models his concepts of change much as we would. The span of time over which change occurs is represented by a line segment. The intensity of a quality at a certain moment in time is represented by a line segment perpendicular to the time interval. Different intensities at different times are modeled by line segments that have heights in the same ratio as the intensities. Taking all the intensities at all moments of time in a given interval, he observes that that their topmost points trace out a curve. The quality as a whole is thus modeled by a two dimensional figure. A quality with uniform or constant intensity is modeled by a rectangle and a uniformly difform motion is modeled by a right triangle or trapezoid as in FIGURE 3a; difformly difform motions are represented by the various other shapes as in FIGURE 3b.

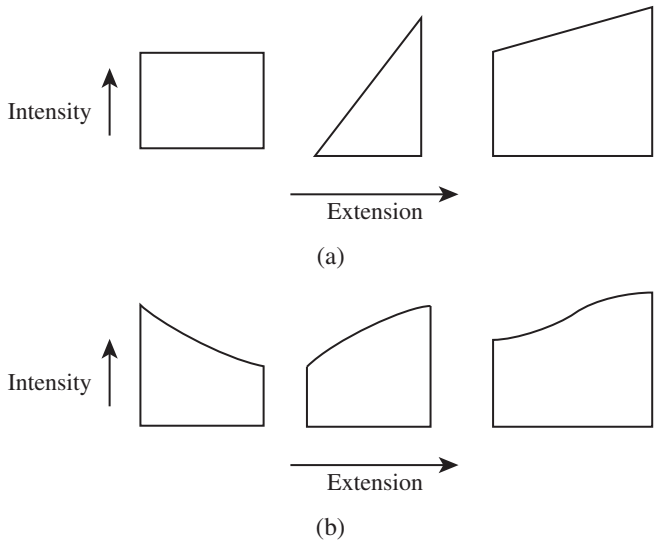


Figure 3 Uniform and difform motions, where we interpret extension as time and intensity as velocity

Here's what Oresme said about his modeling process [**4**, p. 175]:

> [S]omething is more quickly and perfectly understood when it is explained by a visible example. Thus it seems quite difficult for certain people to understand the nature of a quality that is uniformly difform. But what is easier to understand than that the altitude of a right triangle is uniformly difform?.... [T]hen one recognizes with ease in such a quality its difformity, disposition, figuration, and *measure*.

Throughout his work, Oresme is concerned about how the mind conceptualizes. But we are interested in the last word of the quote, *measure* (my italics). What Oresme is interested in measuring (or at least comparing) is the "quantity of the quality." How much of a quality do you have? When applied to the quality of velocity, the quantity of the quality is what Galileo would have referred to as the total velocity. Unlike Galileo, Oresme measures by comparing areas [**4**, p. 405].

> The universal rule is this, that the measure or ratio of any two linear or surface qualities or velocities is as that of the figures by which they are comparatively and mutually imagined.... Therefore, in order to have measures and ratios of qualities and velocities one must have recourse to geometry.

Here's how Oresme stated and proved the Mean Velocity Law [**4**, p. 409], which is equivalent to Galileo's Theorem 1. (Oresme's diagram is reproduced in FIGURE 4. Note that the time interval is the bottom horizontal line.)

> [Proposition] Every quality, if it is uniformly difform, is of the same quantity as would be a quality of the same or equal subject that is uniform according to the degree of the middle point of the same subject.

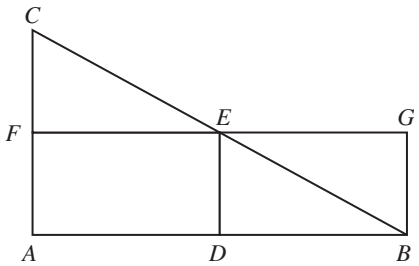


Figure 4 Oresme's simple diagram for the mean velocity law

> [Proof] [L]et there be a quality imaginable by *ABC*, the quality being uniformly difform and terminated at no degree in point *B*. And let *D* be the middle point of the subject line. The degree of this point, or its intensity, is imagined by the line *DE*. Therefore, the quality which would be uniform throughout the whole subject at degree *DE* is imaginable by the rectangle *AFGB*.... Therefore, it is evident by the 26th [proposition] of [Book] I [of the Elements] of Euclid that the two small triangles *EFC* and *EGB* are equal. Therefore the larger *BAC*, which designates the uniformly difform quality, and the rectangle *AFGB*, which designates the quality uniform of degree in the middle point, are equal. And this is what has been proposed.

Oresme goes on to emphasize that his assertion applies to the quality of velocity and, like Galileo, he assumes that distance traveled is directly proportional to total velocity. So Oresme's 14th-century mathematical approach is very familiar: When we find the total distance traveled by integrating velocity over time, we find area.

A digression, just for fun Turning a few pages further in Oresme's work [**4**, p. 413], we find another intriguing and forward looking discussion of a diagram that serves as a proof for the sum of an infinite series. (Have we found a medieval Euler?)

We would write the series that Oresme investigates as $\sum_{i=1}^{\infty} i/2^i$. He proves that its sum is 2 via the diagram reproduced in FIGURE 5: Two unit squares in FIGURES 5a and 5b are divided on the base at 1/2, 1/4, 1/8, etc. The second box is dismantled and piled on the first as in FIGURE 5c, which has a total area of 2, the sum of the areas of the two reassembled boxes. Now look at FIGURE 5d in which we have extended the vertical lines from the base at positions 1/4, 1/8, etc. Notice how the boxes stack up. Starting from the left, there is one rectangle of width 1/2. Then there are two

rectangles, one on top of the other, of width $1/4$; then three of width $1/8$; then four of width $1/16$, and so forth.

The total area of the two unit boxes (namely, 2) is thus represented as the sum $1/2 + 2/2^2 + 3/2^3 + 4/2^4 + \cdots$. As Oresme puts it, the total area must be four times the area of the first part, that is, four times the area of the section labeled E in FIGURE 5b. (Modern readers can compute the sum via geometric series: $\sum_{i=1}^{\infty} i/2^i$ as $\sum_{j=1}^{\infty} \sum_{i=j}^{\infty} 1/2^i = \sum_{j=1}^{\infty} 1/2^{j-1} = 2$.)

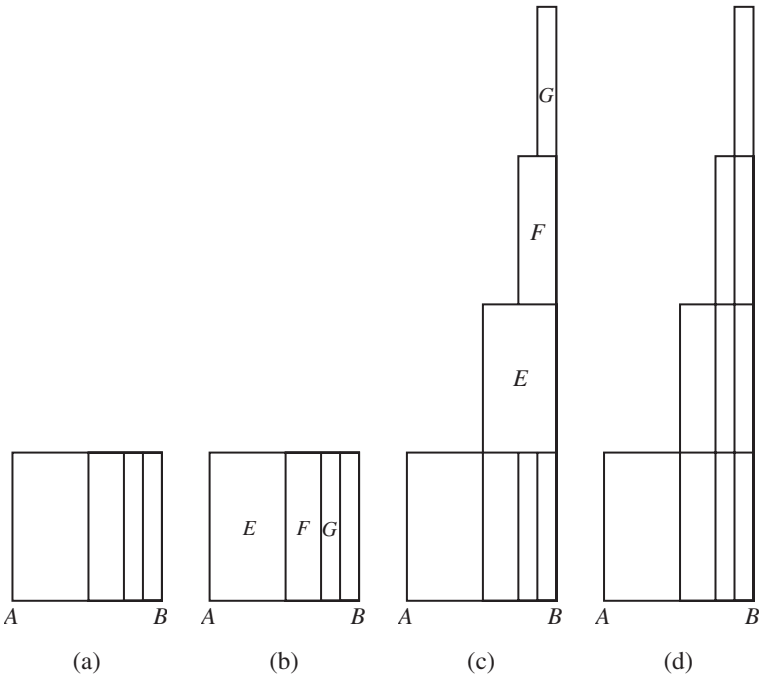


Figure 5 Oresme's proof without words

Oresme says, "A finite [in area] surface can be made as long as we wish, or as high, by varying the extension without increasing the size." We are acquainted with this notion through convergent improper integrals such as $\int_0^1 \frac{1}{\sqrt{1-x}}\,dx$. But Oresme is really interested in how his observation about areas informs his theory of velocity. He goes on to say [**4**, p. 415]:

> In the same way, if some mobile were moved with a certain velocity in the first proportional part of some period of time, divided in such a way, and in the second part were moved twice as rapidly, and in the third three times as fast, and in the fourth four times, and increasing successively to infinity, the total velocity would be precisely four times the [total] velocity of the first part, so that the mobile in the whole hour would traverse precisely four times what it traversed in the first half of the hour. . . and yet it would be moved infinitely fast.

The notion of (instantaneous) velocity that is infinite is difficult to grasp even today. Consider the motion defined by $x(t) = 1 - \sqrt{1-t}$. The total distance traversed in one unit of time starting at $t = 0$ is $x(1) = 1$. The average velocity is therefore 1 but instantaneous velocity is given by $v(t) = 1/2\sqrt{1-t}$ and $\lim_{t\to 1} 1/2\sqrt{1-t} = +\infty$. (When I asked a physicist about this, the response was, "We ignore the endpoints.")

Conclusion

When we look at how Galileo uses mathematics, he seems more remote than Oresme. However, what separates Galileo and Oresme is deeper and dispels any idea that Oresme is the more modern scientist. In Oresme's work, motion and velocity are treated as an application of a more general theory that offers a context for modeling anything that could be considered to have intensity. To him, the model works equally well for qualities such as whiteness, sweetness, and pain, nouns for which he had no measure and therefore no recourse to experimentation [**4**, p. 211].

> [O]ne quality of the body—say, its hotness—can be figured in one way, and perhaps another quality of the same body, such as its whiteness, can be figured in another way, and perhaps another of its qualities—possibly its sweetness—can be figured in a still different way, and similarly for the other [qualities].

Galileo was a quantifier who adamantly insisted that theory be born out by experiment. He did not deal with "qualities" and theories that could not be verified physically. On the other hand, he set the stage for the physics of "ideal" situations. For instance, through his inclined plane experiment, he slowed the effects of gravity by having his object travel down a ramp rather than fall straight down. On a ramp at an angle θ degrees from the horizontal, $x(t) = g\sin(\theta)t^2/2$. So on the vertically tilted ramp ($\theta = 90°$), $x(t) = gt^2/2$, the law of free fall. Further, he determined that, in so far as his experimental results differed from his theoretical results, they differed because of air resistance and friction. He extrapolated to the ideal situation free of these restrictions. Galileo took the leap of faith: natural motion could be modeled by mathematics. Metaphysics (the theory of why things move) became physics (science of how things move), an experimental science with mathematics both at its service and at its helm.

REFERENCES

1. Marshall Clagett, *Science of Mechanics in the Middle Ages*, University of Wisconsin, Madison, 1959.
2. Stillman Drake, *Galileo at Work: His Scientific Biography*, University of Chicago Press, Chicago and London, 1978.
3. Galileo Galilei, *Two New Sciences*, trans. Stillman Drake, University of Wisconsin Press, Madison, 1974.
4. Nicole Oresme, *The Geometry of Qualities and Motions*, translation and commentary by Marshall Clagett, University of Wisconsin, Madison, Milwaukee, and London, 1968.

Summary If a body is constantly accelerated from rest to a final velocity v, then the distance it goes in time t is the same as if it had travelled at constant velocity $v/2$. This fact, the Mean Velocity Law, was proved by Galileo on his way to proving the Law of Free Fall, a milestone in the evolution of modern physics. But the Mean Velocity Law itself was well known as far back as the Middle Ages. In this article, we compare Galileo's mathematical approach to that of the medieval philosopher and mathematician, Nicole Orseme, and we also look at the scientific context—just what information the law gave—in each era, with some surprising results.

OLYMPIA NICODEMI, after completing a Ph.D. at the University of Rochester, went on to teach the wonderful students just down the road at SUNY Geneseo. ("Geneseo" is a Seneca word for pleasant valley.) Her mathematical interests range from old (what was Oresme thinking?) to new (wavelets). Otherwise, she enjoys family, music, gardening and stalking the perfect gelato in its native habitat.

Repeating Decimals: A Period Piece

KENNETH A. ROSS
University of Oregon
Eugene, OR 97403
rossmath@pacinfo.com

Since retiring I have been working with younger people, mostly ages 9–12. They tend to find long division boring, so I show them that long division is interesting by studying repeating decimals. Some good questions from the kids led me to look into the results in this paper. Most of the results are known—as noted in the history section at the end of this article—but they have not been readily accessible in a general form.

The first interesting repeating decimal is the decimal expansion for $\frac{1}{7} = 0.\overline{142857}$. I have known forever that the repeating portions of $\frac{2}{7}$, $\frac{3}{7}$, $\frac{4}{7}$, $\frac{5}{7}$, and $\frac{6}{7}$ are all cyclic permutations of 142857, but I only recently stumbled upon another property of 142857: If you break its set of digits into 2 strings of equal length, then the numbers add to 999: $142 + 857 = 999$. I will call this the 2-block property. The number 142857 also has the 3-block property: $14 + 28 + 57 = 99$. If we look at $\frac{3}{7} = 0.428571$, then again $428 + 571 = 999$ but in this case $42 + 85 + 71 = 198$. That is not quite as nice as 99, but 198 is twice 99 and we regard this as close enough to say that $3/7$ has the 3-block property. With this understanding, $1/7$ and $3/7$ also satisfy the 6-block property, because the sums of their digits are multiples of 9.

These nice block properties are surprisingly common. In this paper, we give two simple theorems that explain nearly all appearances of this phenomenon. First, we illustrate with special cases, stating the theorems and a corollary along the way. Proofs follow the examples, and help to clarify what is going on.

For any fraction, the length of the smallest repeating portion of its decimal is called the *period,* for which we always write ℓ. Thus the periods of $\frac{t}{7}$ for $t = 1, 2, 3, 4, 5, 6$ are all 6. We say that the fraction satisfies the m-block property if m divides ℓ and, when we break the repeating portion into m blocks of equal length $k = \ell/m$, the sum of the m blocks is a string of k nines or an integer multiple of a string of k nines. TABLE 1 contains some examples where the denominators are primes.

These illustrate the following corollary, which is a corollary to both Theorems 1 and 2 below. As we will see in the history section, the prime 487 is especially interesting in our story.

COROLLARY. *Consider a prime $p \geq 7$, and let ℓ be the period of t/p where $1 \leq t < p$. (Then $\ell \leq p - 1$.) If m divides ℓ where $m > 1$, then the m-block property holds for t/p.*

For $m = 2$, the sum of the two blocks, A and B, is exactly a string of nines. Just add A and B in the standard way, working from right to left, and you will see that each column adds to 9—no carrying. This property is sometimes referred to as the "nines property." In this case, the sum is exactly a string of nines. This is also true for $1/p$ for $m = 3$; the numerator being 1 is crucial here as a glance at $3/7$ shows. We explain this after proving Theorem 2.

The m-block property holds for $1/n$ in many cases when n is not prime. See TABLE 2 below. You're invited to momentarily ignore the third column and verify some of the claims in the table about m-block properties. The challenge is to see why some m-block properties hold, while others don't.

Math. Mag. **83** (2010) 33–45. doi:10.4169/002557010X479974.

TABLE 1

denominator	ℓ	fraction	m-block property holds for
7	6	$1/7 = 0.\overline{142857}$	$m = 2, 3, 6$
7	6	$3/7 = 0.\overline{428571}$	$m = 2, 3, 6$
13	6	$1/13 = 0.\overline{076923}$	$m = 2, 3, 6$
13	6	$11/13 = 0.\overline{846153}$	$m = 2, 3, 6$
17	16	$1/17 = 0.\overline{0588235294117647}$	$m = 2, 4, 8, 16$
19	18	$1/19 = 0.\overline{052631578947368421}$	$m = 2, 3, 6, 9, 18$
31	15	$1/31 = 0.\overline{032258064516129}$	$m = 3, 5, 15$
31	15	$11/31 = 0.\overline{354838709677419}$	$m = 3, 5, 15$
73	8	$17/73 = 0.\overline{23287671}$	$m = 2, 4, 8$
487	486	$1/487 = 0.\overline{0020533\cdots}$	$m = 2, 3, 6, 9, 18, 27, 54, 81, 162, 243, 486$

Do you see a pattern for when the m-block property holds? The key is to look at $k = \ell/m$ and see whether it is a multiple of any of the periods of the reciprocals of the prime factors in the denominator. Those periods are listed in the third column of TABLE 2. For example, for $77 = 7 \cdot 11$, the periods of $1/7$ and $1/11$ are 6 and 2, respectively. For $m = 2$, 3 and 6, the corresponding values of k are 3, 2 and 1. Of these k's, only 2 is a multiple of 6 or 2, and the m-block property only fails for $m = 3$ where $k = 2$. Here is the general theorem, which is Theorem 3 in Harold Martin's 2007 paper [**24**].

THEOREM 1. *Let $n = p_1^{a_1} \cdots p_r^{a_r}$ where the primes $p_j \geq 7$ are distinct. Let ℓ be the period of t/n, where $0 < t < n$ and t is relatively prime to n, so that the fraction t/n is reduced. For each prime p_j, we write $\ell(p_j)$ for the period of $1/p_j$. If $\ell = mk$, where $m > 1$ and k is an integer, and if none of the periods $\ell(p_j)$ divides k, then t/n has the m-block property.*

TABLE 2

				m-block property	
denominator	ℓ	$\ell(p_j)'s$	fraction	holds for	fails for
$77 = 7 \cdot 11$	6	6, 2	$1/77 = 0.\overline{012987}$	$m = 2, 6$	$m = 3$
$91 = 7 \cdot 13$	6	6, 6	$1/91 = 0.\overline{010989}$	$m = 2, 3, 6$	no m
$143 = 11 \cdot 13$	6	2, 6	$19/143 = 0.\overline{132867}$	$m = 2, 6$	$m = 3$
$259 = 7 \cdot 37$	6	6, 3	$19/259 = 0.\overline{073359}$	$m = 3, 6$	$m = 2$
$407 = 11 \cdot 37$	6	2, 3	$19/407 = 0.\overline{046683}$	$m = 6$	$m = 2, 3$
$1001 = 7 \cdot 11 \cdot 13$	6	6, 2, 6	$151/1001 = 0.\overline{150849}$	$m = 2, 6$	$m = 3$
$803 = 11 \cdot 73$	8	2, 8	$1/803 = 0.\overline{00124533}$	$m = 8$	$m = 2, 4$
$451 = 11 \cdot 41$	10	2, 5	$1/451 = 0.\overline{0022172949}$	$m = 10$	$m = 2, 5$
$1147 = 31 \cdot 37$	15	15, 3	$1/1147 = 0.\overline{000871839581517}$	$m = 3, 15$	$m = 5$
$1241 = 17 \cdot 73$	16	16, 8	$1/1241 = 0.\overline{0008058017727639}$	$m = 4, 8, 16$	$m = 2$

All of the examples in TABLE 2 are easy to verify directly. All of the m-block properties that hold, hold by Theorem 1. Each failure in TABLE 2 occurs when at least one of the $\ell(p_j)$'s divides k. Nevertheless, we are morally obligated to check the failures, because the converse of Theorem 1 is not true: The m-block property can hold, even if some period $\ell(p_j)$ divides k. Here is the simplest example: For $n = 253 = 11 \cdot 23$ and $\ell = 22 = mk$ where $m = 11$ and $k = 2$, we have

$$\frac{1}{253} = 0.\overline{00395256916996047430 83},$$

the 11-block sum is $594 = 6 \cdot 99$, and yet $k = 2$ is a multiple of $\ell(11) = 2$. A class of examples follows the proof of Theorem 2.

Why do we restrict primes to be bigger than 5? First, avoiding factors of 2 and 5 in the denominators is a major simplification, because they are also factors of 10. Second, we can still get to these cases indirectly. If n is divisible by 2 or 5, then the repeating portion of t/n also occurs as the repeating portion of another fraction t^*/n^* where 2 and 5 are not factors of n^*. Consider, for example, $17/280$. Since $280 = 2^3 \cdot 5 \cdot 7$, to eliminate the 2's and 5 in the denominator, we multiply the fraction by 10^3 and obtain

$$10^3 \cdot \frac{17}{280}, \quad \text{which reduces to} \quad \frac{425}{7} = 60 + \frac{5}{7} = 60.\overline{714285}.$$

Therefore $\frac{17}{280} = 0.060\overline{714285}$, and the repeating portion of the decimal expansion is the same as for $5/7$. **For these reasons, we assume that *n* is relatively prime to 10**.

We also avoid allowing 3 as a factor of the denominator, because the block property involving nines rarely holds in this case, essentially because 3 divides 9. Moreover, if 3 is a factor of n in Theorem 1, the hypotheses never hold: k is always a multiple of $\ell(3) = 1$. There are some patterns, though, if the denominator has at most two factors of 3; consider

$$\frac{23}{117} = \frac{23}{3^2 \cdot 13} = 0.\overline{196581} \quad \text{and} \quad \frac{65}{219} = \frac{65}{3 \cdot 73} = 0.\overline{29680365}.$$

But this is another story. **We do not allow multiples of 3 in the denominator**, since the extra complications tend to obscure the main ideas of this article.

We are now ready to address denominators that are powers of a prime. We avoided these fractions in TABLE 2, in part because they have long periods. For prime powers, the following theorem gives easily-checked conditions that are *equivalent* to the m-block property.

THEOREM 2. *Consider a prime power p^a where $p \geq 7$, and consider an integer t where $0 < t < p^a$ and t is relatively prime to p. Let ℓ be the period of t/p^a and suppose $\ell = mk$ where $m > 1$ and k is an integer. The following are equivalent:*

(a) *The m-block property holds for t/p^a.*

(b) *m is not a power of p.*

(c) *The period $\ell(p)$ of $1/p$ does not divide k.*

TABLE 3 gives examples where the denominators are powers of primes and the block property fails. Note that, in each case, neither (b) nor (c) in Theorem 2 holds, since m is a power of p and $\ell(p)$ divides k.

TABLE 3

denominator	ℓ	fraction	m	k	$\ell(p)$	m-block sum
$49 = 7^2$	42	$1/49=0.\overline{020408163265307\cdots}$	7	6	6	3,142,854
$121 = 11^2$	22	$1/121 = 0.\overline{00826446280\cdots}$	11	2	2	504
$169 = 13^2$	78	$1/169 = 0.\overline{005917\cdots}$	13	6	6	6,076,917
$343 = 7^3$	294	$1/343 = 0.\overline{002915\cdots}$	7	42	6	$\cdots$ 77,548
$343 = 7^3$	294	$1/343 = 0.\overline{002915\cdots}$	49	6	6	24,442,833

The last column in TABLE 3 shows how uninteresting the m-block sums can be. TABLE 4 lists some examples where the m-block property holds for $1/p^a$.

TABLE 4

denominator	ℓ	fraction	m-block property holds for
$49 = 7^2$	42	$1/49 = 0.\overline{020408163265307\cdots}$	$m = 2, 3, 6, 14, 21, 42$
$121 = 11^2$	22	$1/121 = 0.\overline{00826446280\cdots}$	$m = 2, 22$
$169 = 13^2$	78	$1/169 = 0.\overline{0059171597\cdots}$	$m = 2, 3, 6, 26, 39, 78$
$343 = 7^3$	294	$1/343 = 0.\overline{002915\cdots}$	$m = 2, 3, 6, 14, 21, 42, 98, 147, 294$
$237,169 = 487^2$	486	$1/237169 = 0.\overline{000004216\cdots}$	$m = 2, 3, 6, 9, 18, 27, 54, 81, 162, 243, 486$

Midy's Theorem

The 2-block property for fractions with prime denominators is the earliest published result, and it is called Midy's Theorem [**25**, 1836]. With prime denominator, all that is required is that ℓ be even. Theorem 2 shows that Midy's 2-block property also holds for powers of primes $p \geq 7$, since 2 is not a power of p. Theorem 1 asserts that Midy's 2-block property holds for $n = p_1^{a_1} \cdots p_r^{a_r}$ provided $\ell/2$ is not a multiple of any of the $\ell(p_j)$'s. TABLE 2 gives four simple examples where Midy's 2-block property holds and five where it fails. Similar remarks apply to the 3-block property, which has received some attention.

Jones and Pearce [**18**] take a different and interesting approach using fractal-like "graphical analysis graphs" of fractions, which are based on decimal expansions to various bases. Under our general hypotheses (in Theorem 1), t/n satisfies Midy's 2-block property if and only if the graphical analysis graph is rotationally symmetric in base 10. This follows from Theorem 3 in [**18**] and our Lemma 3 below.

EXERCISE 1. Prove Theorem 1 for the case $m = \ell$ and $k = 1$.

EXERCISE 2. With the notation as in Theorem 1, show that if $\ell = 2k$ and k is an integer, then t/n has Midy's 2-block property if and only if none of the periods $\ell(p_j)$ divides k. See Theorem 8 in [**24**].

Now we prepare for the proofs of Theorems 1 and 2.

Modular arithmetic

The basics of modular arithmetic in $\mathbb{Z}(n) = \{0, 1, 2, \ldots, n-1\}$ will be our main tool. What we need can be found in most number theory books. See, for example, Chapter 2 in [**27**] on congruences in $\mathbb{Z}(n)$, or Chapters 3 and 9 in the new book, [**9**]. Much less sophisticated treatments, such as found in §14.3 of [**16**], Chapter 4 in [**4**] or sections 4 and 9 in [**7**], are sufficient for this article. The group $\mathbb{U}(n)$ of units, described below, is defined on page 747 of [**16**], right after Example 16.4 about $\mathbb{U}(9)$.

Here's a quick overview. For integers a and b, we will write $a \equiv b \pmod{n}$ if their difference is divisible by n, i.e., if they have the same remainders when they are divided by n. The sum or product in $\mathbb{Z}(n)$ is the remainder of the ordinary sum or product when it's divided by n. For example, $4 + 5 = 2$ and $4 \cdot 5 = 6$ in $\mathbb{Z}(7)$, since $4 + 5 \equiv 2 \pmod{7}$ and $4 \cdot 5 \equiv 6 \pmod{7}$. Similarly $8 + 11 = 4$ and $8 \cdot 11 = 13$ in $\mathbb{Z}(15)$, since $8 + 11 \equiv 4 \pmod{15}$ and $8 \cdot 11 \equiv 13 \pmod{15}$.

In many ways, modular arithmetic is very similar to the arithmetic of integers, but there are at least two important differences. In $\mathbb{Z}(n)$, $ab = 0$ need not imply that either $a = 0$ or $b = 0$, unless n is prime. For example, $9 \cdot 5 \equiv 0 \pmod{15}$. So, we have to be careful not to slip into using this property, unless n is prime of course.

On the positive side, in many $\mathbb{Z}(n)$, the numbers 1 and its negative $n - 1$ are not the only numbers with multiplicative inverses. Numbers with inverses are also called units, so we write $\mathbb{U}(n)$ for the set of numbers in $\mathbb{Z}(n)$ that have inverses. $\mathbb{U}(n)$ consists exactly of the nonzero numbers in $\mathbb{Z}(n)$ that are relatively prime to n. It is a group and is called the group of units; see Theorem 2.47 (page 121) in [**27**]. This implies that $\mathbb{U}(n)$ is closed under multiplication (modulo n). As an example, $\mathbb{U}(15) = \{1, 2, 4, 7, 8, 11, 13, 14\}$. The inverses of elements in $\mathbb{U}(15)$ can be seen from the extended identity: $2 \cdot 8 \equiv 4^2 \equiv 7 \cdot 13 \equiv 11^2 \equiv 14^2 \equiv 1 \pmod{15}$. For a prime p, we have $\mathbb{U}(p) = \mathbb{Z}(p) \setminus \{0\}$; for example, $\mathbb{U}(7) = \{1, 2, 3, 4, 5, 6\}$ and the inverses in $\mathbb{Z}(7)$ can be read from $2 \cdot 4 \equiv 3 \cdot 5 \equiv 6^2 \equiv 1 \pmod{7}$.

Long division

Let's now discuss how we calculate t/n using the division algorithm. As usual, we assume that $0 < t < n$ and that the fraction t/n is reduced, so that t and n are relatively prime and t is in $\mathbb{U}(n)$.

We write r_0 for t, so r_0 is an honorary remainder at the beginning of the process. The standard division algorithm provides digits $d_1, d_2, d_3, \ldots$ (so that $t/n = 0.d_1d_2d_3\cdots$) and remainders $r_1, r_2, r_3, \ldots$ satisfying

$$d_j \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \quad \text{and} \quad r_j \in \mathbb{Z}(n)$$

for all $j = 1, 2, 3, \ldots$. Both d_j and r_j are determined by r_{j-1} and n. First, d_j is the integer part of $10 \cdot r_{j-1}/n$, and then r_j is given by $r_j = 10 \cdot r_{j-1} - d_j n$. Observe that $r_j = 10 \cdot r_{j-1} \pmod{n}$, from which it follows, by induction, that

$$r_j \equiv t \cdot 10^j \pmod{n} \quad \text{for all} \quad j.$$

Periods of t/n

Here are some well-known facts about the periods of decimal fractions.

LEMMA 1. *Suppose that 2 and 5 are not divisors of n, let ℓ be the period of $1/n$, and consider t in $\mathbb{U}(n)$.*

(a) *The period ℓ of $1/n$ is also the period of t/n for all $t \in \mathbb{U}(n)$.*

(b) *The repeating portions of the decimal expansions for t/n are purely periodic; that is, they start at d_1.*

(c) *The period ℓ for t/n is the smallest positive integer ℓ satisfying $10^\ell \equiv 1 \pmod{n}$. Thus $10^k \equiv 1 \pmod{n}$ if and only if k is a multiple of ℓ.*

Proof. (a) The period ℓ of $1/n$ is the smallest number ℓ such that $r_{j+\ell} = r_j$ for some j. We'll write i for the smallest value of j such that $r_{j+\ell} = r_j$. Therefore all $r_0, r_1, \ldots, r_{i+\ell-1}$ are distinct and $r_{i+\ell} = r_i$. Let s_j be the corresponding remainders for t/n. Then

$$s_j \equiv t \cdot 10^j \pmod{n} \equiv t \cdot r_j \pmod{n}$$

for each j. Since $\mathbb{U}(n)$ is closed under multiplication (modulo n), and since t and the prime factors of 10^j are in $\mathbb{U}(n)$, we see that the remainders r_j and s_j are also in $\mathbb{U}(n)$. On the group $\mathbb{U}(n)$, the map $r \to t \cdot r \pmod{n}$ is one-to-one, so i is the smallest integer so that $s_{i+\ell} = s_i$ and $s_0, s_1, \ldots, s_{i+\ell-1}$ are distinct.

(b) We want to show that i in the proof of part (a) equals 0. It suffices to show this for $1/n$. We have $10^i \equiv 10^{i+\ell} \pmod{n}$, and since 10^i has an inverse (modulo n), this implies that $10^0 \equiv 10^{0+\ell} \pmod{n}$. Since i was minimal, i must be 0.

(c) follows from the proof of part (b). ∎

Part (a) of Lemma 1 is implicit in Leavitt's papers [**21**, 1967] and [**22**, 1984], and part (c) is Theorem 1 in Leavitt's paper [**22**, 1984]. All of Lemma 1 follows immediately from Theorem 135 in Hardy & Wright's Number Theory book [**17**].

The following fact, an easy consequence of Lemma 1(c), is helpful for determining the period of t/n. If $n = p_1^{a_1} \cdots p_r^{a_r}$, where the primes p_j are distinct and do not include 2 or 5, then the period of $1/n$ is the least common multiple of the periods of $1/p_j^{a_j}$, taken over $j = 1, \ldots, r$. (The periods for $1/p^a$, where p is prime, are described in Lemma 4.)

Two lemmas

As always $t \in \mathbb{U}(n)$, 2 and 5 are not factors of n, and $\ell = mk$ where $m > 1$. The repeating part of the decimal expansion for t/n, namely $d_1 d_2 \cdots d_\ell$, breaks into m consecutive blocks of integers, each of length k: $A_1, A_2, \ldots, A_m$. Thus $A_1 = d_1 d_2 \cdots d_k$, $A_2 = d_{k+1} d_{k+2} \cdots d_{2k}$, etc., and $A_m = d_{(m-1)k+1} \cdots d_{mk}$. (Here, juxtaposition denotes a string of digits, not multiplication. Also, we refer to the strings as integers, which we can sum, even if they begin with some zeros.)

LEMMA 2. *The following statements are equivalent and, when they hold, they all hold for the same integer K, which satisfies $K \leq m - 1$.*

(a) $A_1 + A_2 + \cdots + A_m = K(10^k - 1)$. (*Note that each integer being added has k digits and $10^k - 1$ is a string of k nines, so this is the m-block property for t/n.*)

(b) $A_1 A_2 \cdots A_m + A_2 A_3 \cdots A_m A_1 + \cdots + A_m A_1 \cdots A_{m-1} = K(10^\ell - 1)$. (*Note that each integer being summed here has ℓ digits and that $10^\ell - 1$ is a string of ℓ nines.*)

(c) $0.\overline{A_1 A_2 \cdots A_m} + 0.\overline{A_2 A_3 \cdots A_m A_1} + \cdots + 0.\overline{A_m A_1 \cdots A_{m-1}} = K$. (*These are repeating decimals; for example, $0.\overline{A_1 A_2 \cdots A_m}$ is exactly t/n.*)

Proof. Let A, B, and C be the sums in parts (a), (b), and (c), respectively.
(a) $\Longleftrightarrow$ (b). Since

$$A_1A_2\cdots A_m = A_m + A_{m-1}\cdot 10^k + A_{m-2}\cdot 10^{2k} + \cdots + A_1\cdot 10^{(m-1)k},$$

with similar sums for $A_2A_3\cdots A_mA_1$, etc., we see that

$$\begin{aligned} B &= A + A\cdot 10^k + A\cdot 10^{2k} + \cdots + A\cdot 10^{(m-1)k} \\ &= A(1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k}) = A\cdot\frac{10^{km}-1}{10^k-1} = A\cdot\frac{10^\ell-1}{10^k-1}. \end{aligned}$$

It follows that $B = K(10^\ell - 1)$ if and only if $A = K(10^k - 1)$.
(b) $\Longleftrightarrow$ (c). We have

$$0.\overline{A_1A_2\cdots A_m} = \frac{A_1A_2\cdots A_m}{10^\ell - 1},$$

with similar numerators for the other decimals. Summing over all the decimals yields

$$C = \frac{A_1A_2\cdots A_m + A_2A_3\cdots A_mA_1 + \cdots + A_mA_1\cdots A_{m-1}}{10^\ell - 1} = \frac{B}{10^\ell - 1},$$

so that $B = K(10^\ell - 1)$ if and only if $C = K$. ∎

LEMMA 3. *The m-block property holds for each t/n if and only if*

$$1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k} \equiv 0 \,(mod\, n).$$

Proof. Recall that (a) of Lemma 2 is the m-block property. We first assume that $1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k} \equiv 0 \,(\mathrm{mod}\, n)$ and prove (c) of Lemma 2. Note that

$$\frac{t}{n} = \frac{r_0}{n} = 0.\overline{d_1d_2\cdots d_\ell} = 0.\overline{A_1A_2\cdots A_m}.$$

From the long division algorithm it's evident that

$$\frac{r_1}{n} = 0.\overline{d_2d_3\cdots d_\ell d_1},$$
$$\frac{r_2}{n} = 0.\overline{d_3\cdots d_\ell d_1d_2},$$

etc., and that

$$\frac{r_k}{n} = 0.\overline{d_{k+1}d_{k+2}\cdots d_k} = 0.\overline{A_2\cdots A_mA_1}.$$

In general, for $0 \le j \le m-1$, we have

$$\frac{r_{jk}}{n} = 0.\overline{d_{jk+1}d_{jk+2}\cdots d_{jk}} = 0.\overline{A_{j+1}\cdots A_{j-1}A_j}.$$

To show that the sum of the repeating decimals is an integer, it suffices to show that

$$\frac{t + r_k + r_{2k} + \cdots + r_{(m-1)k}}{n} \quad \text{is an integer.}$$

Since $r_{jk} \equiv t\cdot 10^{jk} \,(\mathrm{mod}\, n)$, it suffices to show

$$t\cdot(1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k}) \equiv 0 \,(\mathrm{mod}\, n),$$

and this follows from

$$1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k} \equiv 0 \,(\mathrm{mod}\, n),$$

so that (c) of Lemma 2 holds.

The steps in the proof are reversible: if t/n satisfies the m-block property for some t in $\mathbb{U}(n)$, then

$$t \cdot (1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k}) \equiv 0 \,(\mathrm{mod}\, n),$$

and since t and n are relatively prime, this implies

$$1 + 10^k + 10^{2k} + \cdots + 10^{(m-1)k} \equiv 0 \,(\mathrm{mod}\, n). \qquad \blacksquare$$

It follows from the proof of Lemma 3 that if some t/n has the m-block property, then all t/n do, for t in $\mathbb{U}(n)$. (Lemma 3 is given on page 92 in Shrader-Frechette [**32**, 1978].)

Proof of Theorem 1

By Lemma 3, it suffices to show

$$1 + x + x^2 + \cdots + x^{m-1} \equiv 0 \,(\mathrm{mod}\, n),$$

where $x = 10^k$. Note that $x^m \equiv 10^{km} \equiv 10^{\ell} \equiv 1 \,(\mathrm{mod}\, n)$, so that $x^m - 1 \equiv 0 \,(\mathrm{mod}\, n)$. Fix p_j. Since n divides $x^m - 1$, we have that

$$p_j^{a_j} \quad \text{divides} \quad x^m - 1 = (x-1)(1 + x + x^2 + \cdots + x^{m-1}).$$

By hypothesis, k is not a multiple of $\ell(p_j)$, so $x = 10^k \not\equiv 1 \,(\mathrm{mod}\, p_j)$ by Lemma 1(c). In other words, p_j does not divide $x - 1$. It follows that

$$p_j^{a_j} \quad \text{divides} \quad 1 + x + x^2 + \cdots + x^{m-1}.$$

This is true for all j, so n divides $1 + x + x^2 + \cdots + x^{m-1}$; hence

$$1 + x + x^2 + \cdots + x^{m-1} \equiv 0 \,(\mathrm{mod}\, n). \qquad \blacksquare$$

Note that a special case of Theorem 1 is the key implication (c) $\Longrightarrow$ (a) of Theorem 2. To complete the proof of Theorem 2, we need to know more about periods of prime powers.

Periods of prime powers

Dickson [**6**, page 164] credits Lemma 4 below to Thibault [**34**, 1843]. It was proved by Prouhet [**28**, 1846] and again by Muir [**26**, 1875]. Since the proof does not seem readily accessible or all that simple, I provide a proof here, starting with a lemma from Muir [**26**].

MUIR'S LEMMA. *If $p \geq 3$ is prime and M, n and a are positive integers, then*

$$M \equiv 1 \,(mod\ p^a) \quad \textit{if and only if} \quad M^{p^n} \equiv 1 \,(mod\ p^{a+n}).$$

Proof. The first implication is easy. If $M \equiv 1 \pmod{p^a}$, then $M - 1 = rp^a$ for some integer r, and

$$M^p - 1 = (rp^a + 1)^p - 1 = p(rp^a) + \sum_{k=2}^{p} \binom{p}{k} (rp^a)^k.$$

Now p^{a+1} divides each term on the right, so $M^p \equiv 1 \pmod{p^{a+1}}$. Now apply induction on n. Note that p doesn't need to be prime for this implication.

For the harder implication, it suffices to prove that $M^p \equiv 1 \pmod{p^{a+1}}$ implies $M \equiv 1 \pmod{p^a}$, for then

$$\begin{aligned} M^{p^n} \equiv 1 \pmod{p^{a+n}} &\Rightarrow M^{p^{n-1}} \equiv 1 \pmod{p^{a+n-1}} \\ &\Rightarrow \cdots \Rightarrow M^p \equiv 1 \pmod{p^{a+1}} \Rightarrow M \equiv 1 \pmod{p^a}. \end{aligned}$$

So, suppose $M^p \equiv 1 \pmod{p^{a+1}}$, and let $y = M - 1$. It suffices to prove that $y \equiv 0 \pmod{p^a}$. We have

$$M^p - 1 = (y + 1)^p - 1 = \sum_{k=1}^{p-1} \binom{p}{k} y^k + y^p.$$

Since p is prime, p divides each $\binom{p}{k}$ for $1 \le k \le p - 1$, and so p divides every term on the right-side of the equality except possibly y^p. But since p divides $M^p - 1$, we see that p divides y^p too. Hence p divides y itself. Now let $b \ge 1$ be the biggest exponent so that p^b divides y. If $b \ge a$, then $y \equiv 0 \pmod{p^a}$, and we're done.

So assume $1 \le b \le a - 1$. Then $y = p^b r$ where r is relatively prime to p, and we can write

$$\begin{aligned} M^p - 1 = (y + 1)^p - 1 &= (p^b r + 1)^p - 1 \\ &= p(p^b r) + \binom{p}{2} (p^b r)^2 + \sum_{k=3}^{p} \binom{p}{k} (p^b r)^k. \end{aligned}$$

Since p divides $\binom{p}{2}$ and $2b + 1 \ge b + 2$, p^{b+2} divides $\binom{p}{2}(p^b r)^2$. Since $b + 2 \le a + 1$ and $M^p \equiv 1 \pmod{p^{a+1}}$, p^{b+2} divides $M^p - 1$. Thus p^{b+2} divides every term in the last displayed formula except for $p(p^b r)$, and this is a contradiction. ∎

Note The last paragraph of the last proof doesn't work if $p = 2$. In fact, $N^2 \equiv 1 \pmod{2^{a+1}}$ does not imply $N \equiv 1 \pmod{2^a}$. Consider $N = 7$ and $a = 3$, or $N = 31$ and $a = 5$.

In Lemma 4, we'll write $\ell(p^a)$ for the period of $1/p^a$.

LEMMA 4. *For a prime p not equal to 2 or 5, and for $a \ge 1$, we have $\ell(p^a) = p^s \ell(p)$ for some $s \le a - 1$. In fact, if w is the largest power of p such that $\ell(p^w) = \ell(p)$, then $\ell(p^a) = \ell(p) p^{a-w}$ for $a > w$.*

Proof. Since $10^{\ell(p)} \equiv 1 \pmod{p}$, the easy implication in Muir's Lemma (with $M = 10^{\ell(p)}$) gives $10^{\ell(p)p^{a-1}} \equiv 1 \pmod{p^a}$. Thus $\ell(p^a)$ divides $\ell(p)p^{a-1}$, by Lemma 1(c). Now, $10^{\ell(p^a)} \equiv 1 \pmod{p^a}$ forces $10^{\ell(p^a)} \equiv 1 \pmod{p}$, so $\ell(p^a)$ is a multiple of $\ell(p)$, again using Lemma 1(c). Since $\ell(p)$ is relatively prime to p, being less than p, we see that $\ell(p^a) = \ell(p)p^s$ for some $s \le a - 1$.

Now suppose $\ell(p^w) = \ell(p)$ and $\ell(p^{w+1}) > \ell(p)$. By the preceding paragraph, $\ell(p^a) = \ell(p)p^s$ for some $s \le a - 1$, and we want to show that $s = a - w$ when $a > w$. Since $10^{\ell(p)} \equiv 1 \pmod{p^w}$, the easy Muir implication (with $a = w$) gives

$10^{\ell(p)p^n} \equiv 1 \pmod{p^{n+w}}$ and so (with $n = a - w$), we see that $10^{\ell(p)p^{a-w}} \equiv 1 \pmod{p^a}$. Thus $\ell(p)p^{a-w}$ must be a multiple of $\ell(p)p^s$, and so $s \le a - w$. Now assume that $s < a - w$ for some $a > w$. Then $10^{\ell(p)p^s} \equiv 1 \pmod{p^a}$ and the harder Muir implication gives $10^{\ell(p)} \equiv 1 \pmod{p^{a-s}}$. Since $a - s \ge w + 1$, we conclude $10^{\ell(p)} \equiv 1 \pmod{p^{w+1}}$, so that $\ell(p^{w+1}) \le \ell(p)$, contrary to our supposition about w. ∎

As noted after Theorem 9 in [**22**, 1984], w is almost always 1, though $p = 3$, $p = 487$ and $p = 56{,}598{,}313$ are exceptions where $w = 2$. In 1984, that was all that was known, though w was known for all primes less than 300 million.

Proof of Theorem 2

First, we observe that the implications (b) $\Longleftrightarrow$ (c) follow immediately from Lemma 4: Since $\ell = p^s\ell(p)$ for some $s \ge 0$, $mk = p^s\ell(p)$, so m is a power of p if and only if k is a multiple of $\ell(p)$.

Next, the implication (c) $\Longrightarrow$ (a) is a special case of Theorem 1.

Finally, suppose that (a) holds, and assume (b) and (c) fail. By Lemma 3, we have

$$1 + x + \cdots + x^{m-1} \equiv 0 \pmod{p^a},$$

where $x = 10^k$. Since (c) fails, k is a multiple of $\ell(p)$. Thus $10^k \equiv 1 \pmod{p}$ by Lemma 1(c), so p divides $x - 1$. Since $x^m - 1 = (x - 1)(1 + x + \cdots + x^{m-1})$ and p^a divides the sum, p^{a+1} divides $x^m - 1$, so that $x^m \equiv 1 \pmod{p^{a+1}}$. Since (b) fails, $m = p^u$ for some $u \ge 1$. Therefore,

$$x^{p^u} \equiv 1 \pmod{p^{a+1}}.$$

By Muir's Lemma, we obtain

$$x^{p^{u-1}} \equiv 1 \pmod{p^a} \quad \text{or} \quad 10^{kp^{u-1}} \equiv 1 \pmod{p^a}.$$

Since $kp^{u-1} < kp^u = km = \ell$, this contradicts the fact that ℓ is the minimal power of 10 equivalent to 1 mod (p^a). We conclude that (a) implies (b) and (c). ∎

Theorem 2 assures us that the 2-block and 3-block properties always hold for $1/p^a$, $p \ge 7$. As promised after our statement of the Corollary, we now explain why the 3-block sum is *exactly* a string of nines for $1/p^a$. This is the case in Lemma 3's proof where $t = 1 = r_0$. Then $1 + r_k + r_{2k} \le 1 + (p^a - 1) + (p^a - 1) = 2p^a - 1$, so the sum must be p^a. This implies that, in this case, the constant K in Lemma 2 is equal to 1.

An extension of Theorem 1

EXERCISE 3. With the notation of Theorem 1, show that the m-block property holds provided that, for each prime p_j, either (i) $\ell(p_j)$ does not divide k, or else (ii) $\ell(p_j)$ divides k, the exponent $a_j = 1$, and p_j divides m.

This extension is an endless source of examples of the m-block property not covered by Theorem 1. As noted after first stating Theorem 1, the simplest one involves $n = 11 \cdot 23 = 253$. The most complicated one that I've also verified directly is for $n = 11 \cdot 89^2 = 87{,}131$ where $\ell = 3916$, $m = 979$, and $k = 4$. The 979-block sum is $488 \cdot 9999$.

Some history

Periods of decimal expansions of fractions were studied extensively in the latter half of the 1700s; see [**3**]. One project was the creation of tables and another involved theoretical questions such as the calculation of the periods. The Corollary for $m = 2$ and $1/p$, known as Midy's theorem, implies that the digits of the second half of the decimal expansion for $1/p$, p a prime, can be calculated instantly from the first half, provided the period is even. C. F. Hindenburg used this fact, at least when the period is $p - 1$, to simplify calculations of such decimal expansions. He communicated this idea to J. H. Lambert in December 1776 after he learned from Lambert about the connection between periodic decimal expansions and Fermat's Little Theorem.

Working in isolation, Henry Goodwyn [**14**, 1802] was another prodigious calculator who also knew about this application of Midy's theorem. Related problems were addressed by Lambert, Jean Bernoulli, John Robertson, and others. Then in 1793, 16-year-old C. F. Gauss [**11**] learned about the problems and, by 1797–1801, he had solved most of them based on his new work on the foundations of number theory.

Even though the Corollary for $m = 2$ and $1/p$ was known at least as far back as 1776 (Hindenburg), it is known as Midy's theorem because of the pamphlet [**25**, 1836], and this label appears in the titles of the articles [**12**], [**23**], and [**24**]. See Dickson [**6**, page 163] for a very brief summary of what Midy did. Perhaps Midy didn't give a convincing proof, since Dickson states that Lafitte [**20**, 1846] provided a proof.

Dickson [**6**, pp. 161–173] gives a complete history of the study of periodic decimals from 1770 to 1891. A short useful history is given in Shrader-Frechette [**32**], whose modern references begin with R. E. Green [**15**, 1963] who "considered only reciprocals of primes, but seems to be the first person since Midy to examine the notion of breaking the period into several blocks." Here we add to these histories, with a focus on the results in our paper.

After the turn of the 19th century, Midy's theorem seems to have been forgotten until the 1960s, though in the problems section of the 1912 American Mathematical Monthly E. B. Escott [**10**] states the theorem, asks for a proof and asks for what other fractions the result holds. Neither the problem posed, nor any of the three solutions published, mention Midy.

Theorem 2 of Leavitt's paper [**21**, 1967] gives a proof of Midy's theorem close in spirit to our proofs. Both of his papers, [**21**] and [**22**, 1984], have lots of results related to Theorem 2. See, especially, his Theorems 2, 10, and 11 in [**22**]. Maurice Shrader-Frechette's paper [**32**, 1978] is full of interesting facts, including a version of Theorem 1. Lemma 3 and related ideas are also (somewhat hidden) in the paper. This paper has been under-appreciated, because of its unique terminology and notation (especially for the purposes of our presentation) and because its essay format makes it difficult to infer precise theorems and proofs. Another related article is Ecker [**8**, 1983].

Our Theorem 1 is explicitly stated (and proved) in Theorem 3 of Harold Martin's article [**24**, 2007]. Moreover, his Theorem 8 characterizes n for which $1/n$ satisfies Midy's 2-block property.

Several recent papers contain special cases of the results in our paper. Theorem 1 in Dan Kalman's very nice paper [**19**, 1996] is a special case of our Theorem 2. Brian Ginsberg [**12**, 2004] provides a proof of the Corollary for $m = 2$, which is the same as that in [**33**, 2003]. He goes on to prove the result for $m = 3$ and $1/p$. Theorem 3 in Joseph Lewittes' paper [**23**, 2006] is a special case of our Theorem 2, and his Theorem 4 is a characterization related to our Theorem 1. The Corollary is stated for $1/p$ in Jane Arledge and Sarah Tekanski's recent article [**1**, 2008]. They also obtain related results for $1/p$ and $1/p^2$.

Finally, we give a little more history regarding the periods of p^a. On pages 294–295 of the book [**5**, 1852], Desmarest states that if $p < 1000$ is a prime, and if $p \neq 3$ and $\neq 487$, then the period for $1/p^2$ is $p\ell(p)$ where $\ell(p)$ is the period of $1/p$. According to Dickson, Shanks [**29**, 1874] stated that the period of $1/p^a$ is $p^{a-1}\ell(p)$ for $p > 5$, without mentioning 487. Perhaps as penance, in [**30**, 1877], for $p = 487$ he verified that $1/p^2$ has the same period as $1/p$, namely 486. He gave two arguments and avoided giving the full decimal expansions. Glaisher [**13**, 1878] gave the full decimal expansions for $1/p$ and $1/p^2$ (where $p = 487$ and $\ell(p) = 486$), thus verifying Desmarest's suggestion that 487 is an exceptional prime. He lamented that Desmarest didn't show his work: "These words and others distinctly imply that 3 and 487 are the only exceptions to the general rule up to 1000. In order to establish this and to find that 487 was an exception, Desmarest must have performed the divisions (or employed some equivalent process); but it seems strange that if he had actually performed this heavy work he should not have stated the fact explicitly. On the other hand, I have been able to find no allusion to the property of the number 487 prior to the date of Desmarest's work; and it is scarcely to be supposed that Desmarest would adopt so important a statement as that quoted above without giving his authority."

Acknowledgment Many thanks to Rod Nillsen, Bill Kantor, and Dick Koch for help and encouragement.

REFERENCES

1. Jane Arledge and Sarah Tekansik, A new property of repeating decimals, *College Math. J.* **39** (2008) 107–111.
2. Lawrence Brenton, Remainder wheels and group theory, *College Math. J.* **39** (2008) 129–135.
3. Maarten Bullynck, Decimal periods and their tables: A German research topic (1765–1801), *Historia Math.* **36** (2009) 137–160. doi:10.1016/j.hm.2008.09.004
4. David M. Burton, *Elementary Number Theory*, 6th ed., McGraw-Hill, New York, 2005.
5. E. Desmarest, *Théorie des nombres*, L. Hachette, Paris, 1852.
6. L. E. Dickson, *History of the Theory of Numbers*, vol. 1, 1919.
7. Underwood Dudley, *Elementary Number Theory*, 2nd ed., 1978, Dover edition 2008.
8. Michael W. Ecker, The alluring lore of cyclic numbers, *Two-Year College Math. J.* **14** (1983) 105–109. doi:10.2307/3026586
9. Harold M. Edwards, *Higher Arithmetic: an Algorithmic Introduction to Number Theory*, American Mathematical Society, Providence, RI, 2008.
10. E. B. Escott, F. H. Safford, and C. A. Laisant, Solution of problem 370, *American Mathematical Monthly* **19** (1912) 130–132.
11. Carl Fredrick Gauss, *Disquisitiones Arithmeticae* (trans. Arthur A. Clarke), Yale University Press, New Haven, CT, 1965.
12. Brian D. Ginsberg, Midy's (nearly) secret theorem—an extension after 165 years, *College Math. J.* **35** (2004) 26–30. doi:10.2307/4146879
13. J. W. L. Glaisher, On circulating decimals, *Proc. Cambridge Phil. Soc.* **3** (1878) 185–206.
14. Henry Goodwyn, Curious properties of prime numbers, taken as the divisors of unity, *J. of Natural Philosophy, Chemistry, and the Arts* **1** (1802) 314–316.
15. R. E. Green, Primes and recurring decimals, *Math. Gaz.* **47** (1963) 25–33. doi:10.2307/3612039
16. Ralph P. Grimaldi, *Discrete and Combinatorial Mathematics*, 5th ed., Addison-Wesley, Boston, 2003.
17. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 4th ed., Oxford University Press, London, 1960.
18. Rafe Jones and Jan Pearce, A postmodern view of fractions and the reciprocals of Fermat primes, *Math. Mag.* **73** (2000) 83–97. Reprinted in the anthology *Biscuits of Number Theory*, ed. by Arthur T. Benjamin and Ezra Brown, MAA, Washington, DC, 2008.
19. Dan Kalman, Fractions with cycling digit patterns, *College Math. J.* **27** (1996) 109–115. doi:10.2307/2687398
20. P. Lafitte, Théoreme sur les fractions périodiques, *Nouvelles Ann. Math.* **5** (1846) 397–399.
21. W. G. Leavitt, A theorem on repeating decimals, *Amer. Math. Monthly* **74** (1967) 669–673. doi:10.2307/2314251
22. W. G. Leavitt, Repeating decimals. *College Math. J.* **15** (1984) 299–308. doi:10.2307/2686394

23. Joseph Lewittes, Midy's theorem for periodic decimals, *Integers: Electronic J. Combinatorial Number Theory* **7**(1) (2007) #A02 (11 pages).
24. Harold W. Martin, Generalizations of Midy's theorem on repeating decimals, *Integers: Electronic J. Combinatorial Number Theory* **7** (2007) #A03 (7 pages).
25. E. Midy, *De quelques proprietés des nombres et des fractions décimales périodiques*, Nantes, 1836, 21 pages.
26. Thomas Muir, Theorems on congruences bearing on the question of the number of figures in the periods of the reciprocals of the integers, *Messenger Math.* **4** (1875) 1–5.
27. Ivan Niven, Herbert S. Zuckerman, and Hugh L. Montgomery, *An Introduction to the Theory of Numbers*, 5th ed., John Wiley, New York, 1991.
28. E. Prouhet, *Nouv. Ann. Math.* **5** (1846) 661.
29. William Shanks, *Messenger Math.* **3** (1874) 52–55.
30. William Shanks, Remarks chiefly on $487^2 \equiv 486$, *Proc. Roy. Soc. London* **25** (1877) 551–553.
31. James K. Schiller, A theorem in the decimal representation of rationals, *American Mathematical Monthly* **66** (1959) 797–798. doi:10.2307/2310471
32. M. Shrader-Frechette, Complementary rational numbers, *Math. Mag.* **51** (1978) 90–98.
33. V. Subramanyam, Cyclic decimal expansions, *Resonance* **8** (2003) 75–80. doi:10.1007/BF02837924
34. Thibault, *Nouv. Ann. Math.* **2** (1843) 80–89.

Summary Consider the repeating decimal of a reduced fraction t/n between 0 and 1, where none of the primes 2, 3 or 5 is a factor of n. The fraction satisfies the m-block property if m divides the period ℓ and, when the repeating portion is broken into m blocks of equal length, the sum of the m blocks is a string of nines or an integer multiple of a string of nines. An easily-verified sufficient condition is given that implies t/n has the m-block property. The condition is not necessary. For n equal to a prime power p^a ($p \geq 7$), the condition is shown to be sufficient as well. Moreover, t/p^a has the m-block property if and only if m is not a power of p. (These general results for prime powers seem to be new.) The 2-block property for primes is known as Midy's theorem (1836). It holds for all prime powers, but not in general. Many examples and a brief history are included.

KENNETH A. ROSS holds a B.S., University of Utah (1956), and also a Ph.D., University of Washington (1960). He taught at the University of Rochester, 1961–1964, and the University of Oregon, 1965–2000. In the 1980s and 1990s, he served as MAA Secretary, Associate Secretary, and President. His mathematical interests include abstract commutative harmonic analysis, elementary probability, and expository writing. His books include *Elementary Analysis: The Theory of Calculus*, originally published in 1980 by Springer. His most recent publication was written jointly with James D. Harper: Stopping Strategies and Gambler's Ruin, *Mathematics Magazine* **78** (2005) 255–268. Ken has been a member of both the AMS and MAA for over fifty years. His sedentary hobbies in retirement include baseball, mathematics, and mentoring young kids in mathematics.

NOTES

Gergonne's Card Trick, Positional Notation, and Radix Sort

ETHAN D. BOLKER
Departments of Mathematics and Computer Science
University of Massachusetts Boston
Boston, MA 02125-3393
eb@cs.umb.edu

The three pile trick Hold a deck of cards face down and deal 27 cards face up in rows of three, creating three piles each nine high. Overlap the cards in each pile so that your audience can see the values of the cards and so that you can easily pick them up while preserving their order.

Ask a spectator to think of one of the cards, remember it, and tell you which pile it's in. Announce that you will magically move her card to the middle of the deck.

Pick up the three piles, turning them over so that they are face down, quietly making sure that the pile containing the chosen card is *in the middle*. Accompany this action by any patter you choose.

Do this twice more, each time putting the chosen pile in the middle. Then count out the deck to the middle card and turn it over, to your audience's surprise and applause.

I found this classic trick in Rouse Ball [**11**, p.138] while searching for self-working magic that depends on mathematics rather than dexterity. I have taught it to a fourth grade mathematics club and to precocious first graders. Writing this paper led me to lots of other references, starting with Gardner [**4**]. You can find several discussions on the internet [**2**],[**10**]; Bogomolny [**1**] provides a Java applet. The trick is named for Joseph Diaz Gergonne (1771–1859), who first published an analysis in *Annales de Mathématiques*, the journal he founded [**5**, iv, 1813–1814, pp. 276–284]. Mathematicians regularly return to the problems it raises, sometimes rediscovering or reproving theorems known to previous authors—for example, Dickson in 1895 in the first volume of the *Bulletin of the American Mathematical Society* [**3**] and Harrison, Brennan and Gapinski much more recently in *Discrete Applied Mathematics* [**6**]. The treatment here explains the trick as a special case of the radix sorting algorithm from computer science.

Base three arithmetic Most discussions of the trick go on to describe a generalization that clearly depends on base three arithmetic.

Ask the spectator where she wants you to make her chosen card appear in the deck. Tell her that for your magic to succeed she must start counting at 0, not at 1—so the first card is the 0th and the last is the 26th.

Expand the chosen position as a three digit number in base 3. Read the digits from right to left as you pick up the piles and turn them from face up to face down, using the appropriate digit to determine the position of the chosen pile. For example, if the

Math. Mag. **83** (2010) 46–49. doi:10.4169/002557010X479983.

target position is $15 = 120_3$, the first time the chosen pile goes on top (none above it), the second time on the bottom (two above it), the third time in the middle (one above it, one below).

Now count out the face down deck, starting from 0, and turn over the spectator's card at number 15.*

Because the count starts at 0 the middle card in the deck is number $13 = 111_3$. Those three 1's tell you why the chosen pile always goes in the middle in the original version of the trick. In Gergonne's analysis, repeated by Rouse Ball and others, counting starts at 1. Then the middle card is at number 14, and the discussion of the generalization is cluttered with mysterious 1's to be added and subtracted.† Counting from 0 makes the connection with base three arithmetic much clearer, and makes a nice piece of patter for the budding magician.

Radix sort The previous sections described *how* to do the trick. The question *Why does it work?* has several answers. What's new about the one that follows is the connection to *radix sort*, a well known algorithm for putting things in order, for example, in a computer [**7**, pp. 170–173].

Suppose you shuffle a deck of 27 cards numbered 0, 1, . . . , 26. (Counting from 0 is standard practice in computer science.) To restore them to numerical order:

- Express each of the card values in base three.
- Deal the cards into three piles labelled 0, 1, and 2, putting each card in the pile that matches its rightmost digit. Pick up the cards with pile 0 on top, then pile 1, then pile 2 on the bottom, preserving the order of the cards in each pile.
- Repeat, this time using the middle digit to place the cards in piles.
- Repeat, using the leftmost digit.

The cards are in order. To see why, note that after the first pass the cards with numbers that end in 0 are above those that end in 1, which are in turn above those that end in 2. As you deal the next pass, they retain that partial order in each pile of nine, so, for example, the cards in pile 0 are the ones with numbers ending in "00", "01" and "02" in that order (whatever their leftmost digits). The final pass sorts by the leftmost digit.

Think back to Gergonne's trick. Only the card the spectator chose has a prescribed position in the "sorted" deck. So when you deal out the cards you need not assign them to labelled piles as in radix sort, you just play them as they come. After the deal you label the single pile the spectator identifies with the appropriate digit 0, 1 or 2. Put that one in its proper place as you pick up the piles.

When I teach kids this trick they find the base three expansion of the desired final position of the card by taking out as many nines as they can (0, 1 or 2), then as many threes as possible from the remainder. What's left is the units digit. The standard computer algorithm cleverly finds the digits in the opposite order, from right to left, but this way is easier for kids to grasp.

Generalizing the trick and the sort Gergonne knew he could do his trick with a deck of n^n cards but only $n = 3$ is practical: 2^2 is a trivial 4, and $4^4 = 256$ is too many cards to handle. Martin Gardner discusses—in principle only—magician Mel Stover's gargantuan 10^{10} = ten billion card trick, dealing ten times into ten piles of a billion cards each [**4**, ch. 3], [**8**, p. 21].

*Before you try teaching this to children, work it for yourself several times. Just reading the description doesn't educate your hands. In fact I find I that mine forget the manipulations after a while.

†Some authors finesse the problem by counting from the bottom of the deck.

But radix sort and hence Gergonne's trick will work with an n^k card deck for any k. Since the numbers from 0 to $n^k - 1$ have k digit base n expansions you will need k passes to move the selected card to the selected place. In particular, decks of 8, 16 and 32 cards work well to teach binary notation. Expand the target position as a string of three or four or five zeroes and ones. At each pass deal the cards into just two piles and use the digits from right to left to determine which pile to pick up first.

In fact, radix sort works even when the size of the deck isn't a power of the base. k passes will do the job for m cards when $n^{k-1} < m \leq n^k$. For Gergonne's trick the piles must be the same height, so m must be a multiple of n. A great-nephew of mine showed me a version with $m = 21$ and $n = k = 3$: the piles are seven cards high. Since there are 27 three digit numbers in base three that you can use to specify the position of the designated pile in each pass and only 21 positions, it takes some work to see precisely how the final position of the selected card depends on the order in which you pick up the piles. For some of the three digit strings that position depends on where the card started. For example, you can show that if the selected card is in position 0, 1 or 2 then "001" moves it to to position 0 (the top of the deck) but any other selected card ends up in position 1. Of course "000" moves any selected card to the top of the deck. But there is no single three digit string that can move any selected card to position 1.

The following table shows the good target positions, to which you can move every card, and the base three string (to be read from right to left) that tells you how to pick up the piles.

position	0	2	3	6	7	10	12	13	17	18	20
code	000	010	011	022	100	111	122	200	211	212	222

Note that "111" is still the code for the middle (10th) position, even though it represents 13 in base three.

When $m = 12$ every target position is good, and often you have a choice of how to get there, which may make the trick a little less transparent. But there's no middle to the deck: "111" moves some selected cards to position 5 and others to position 6.

position	0	1	2	3	4	5	6	7	8	9	10	11
code	000 001	010	012	021 022	100 101	110	112	121 122	200 201	210	222	221 222

I leave to the reader the verification of these tables and the formulation and proof of any theorems they suggest. Finding them yourself will be more fun than searching for them in the literature.

There's a generalization of radix sort that points to more Gergonne tricks. Suppose you wish to sort a shuffled full deck of 52 cards so that they end up in the order

$$\spadesuit A, \spadesuit 2, \ldots, \spadesuit K, \heartsuit A, \heartsuit 2, \ldots, \heartsuit K, \diamondsuit A, \diamondsuit 2, \ldots, \diamondsuit K, \clubsuit A, \clubsuit 2, \ldots, \clubsuit K$$

Simply deal first into 13 piles of 4 cards, one for each of the values A, 2, ... , K. Pick up the piles in order. Then deal into four piles of 13, one for each suit. Finally, assemble the piles in the suit order you wish. This amounts to describing the cards using a *mixed radix* in which the "units" digit is one of the thirteen possible card values and the "tens" digit is the suit.

Using a mixed radix you can do Gergonne's trick with a deck of, say, 15 cards, in two passes rather than three by labelling the positions p between 0 and 14 with digit strings "xy" where $0 \leq x \leq 4$ and $0 \leq y \leq 2$ so that $p = 3x + y$. First deal the cards into three piles of five, then into five piles of three. Use the digits y and then

x to determine how many piles to put above the selected pile in each pass. Then you can do the trick the other way, with piles of three followed by piles of five—write $p = 5x + y$ with $0 \le x \le 2$ and $0 \le y \le 4$. Dickson [**3**] and Onnen[**9**] wrote about this generalization.

When you've mastered it, move on to a mixed radix with three-digit numbers. But if you are serious about doing mixed radix Gergonne tricks, it pays to learn the right to left digit algorithm for expressing numbers in the strange bases you invent.

REFERENCES

1. A. Bogomolny, The computer as a magician, http://www.maa.org/editorial/knot/ComputerAsMagician.html.
2. David Britland, Cardopolis, http://cardopolis.blogspot.com/2008_01_20_archive.html.
3. L. E. Dickson, Gergonne's pile problem, *Bull. Amer. Math. Soc.* **1** (1895) 184–186, http://projecteuclid.org/euclid.bams/1183414376.
4. M. Gardner, *Mathematics, Magic and Mystery*, Dover, 1956.
5. J. D. Gergonne, Récréations mathématiques. Recherches sur un tour de cartes, *Annales de Mathématiques* **iv** (1813–1814) 276–283, http://archive.numdam.org/item?id=AMPA_1813-1814__4__276_1.
6. J. Harrison, T. Brennan, and S. Gapinski, The Gergonne p-pile problem and the dynamics of the function $x \to \lfloor (x + r)/p \rfloor$, *Discrete Applied Mathematics* **82** (1998) 103–113. doi:10.1016/S0166-218X(97)00132-7
7. D. Knuth, *The Art of Computer Programming, Volume 3: Sorting and Searching*, 3rd ed., Addison-Wesley, 1997.
8. Max Maven, in D. Wolfe and T. Rodgers, eds., *Puzzlers' Tribute: A Feast for the Mind*, AK Peters, 2001.
9. H. Onnen, Gergonne's pile problem, *Bull. Amer. Math. Soc.* **16** (1909) 121–130. http://projecteuclid.org/euclid.bams/1183420544. doi:10.1090/S0002-9904-1909-01874-9
10. Eric Shrader and Keith Riding, Cut the knot, http://www.cut-the-knot.org/arithmetic/rapid/CardTrick.shtml.
11. W. W. Rouse Ball and H. S. M. Coxeter, *Mathematical Recreations and Essays*, 13th ed., Dover, 1987.

Summary Gergonne's three pile card trick has been a favorite of mathematicians for nearly two centuries. This new exposition uses the radix sorting algorithm well known to computer scientists to explain why the trick works, and to explore generalizations. The presentation suggests strategies for introducing the trick and base three arithmetic to elementary school students.

A GM-AM Ratio

CONWAY XU
Montgomery Blair High School
Silver Spring, MD 20901
conway.xu@gmail.com

Let $\mathrm{GM}(a_1, \ldots, a_n)$ and $\mathrm{AM}(a_1, \ldots, a_n)$ denote the geometric and arithmetic mean of positive real numbers $a_1, \ldots, a_n$, respectively. Kubelka [**1**] proves that for any $s > 0$,

$$\lim_{n\to\infty} \frac{\mathrm{GM}(1^s, \ldots, n^s)}{\mathrm{AM}(1^s, \ldots, n^s)} = \frac{s+1}{e^s}. \tag{1}$$

He uses the squeeze theorem and a Riemann sum argument. In pursuing a simpler method to show (1), I realized that $\ln[\mathrm{GM}(1^s, \ldots, n^s)/n^s]$ is a Riemann sum

Math. Mag. **83** (2010) 49–50. doi:10.4169/002557010X479992.

for the area between the x-axis and $y = \ln(x^s)$ from $x = 0$ to $x = 1$ and that $\mathrm{AM}(1^s, \ldots, n^s)/n^s$ is also a Riemann sum for the area between the x-axis and $y = x^s$ from $x = 0$ to $x = 1$. These observations yield the following two stronger results: For any real number $s > -1$,

$$\lim_{n\to\infty} \frac{\mathrm{GM}(1^s, \ldots, n^s)}{n^s} = e^{-s}, \tag{2}$$

$$\lim_{n\to\infty} \frac{\mathrm{AM}(1^s, \ldots, n^s)}{n^s} = \frac{1}{s+1}. \tag{3}$$

To show (2), it is equivalent to show that

$$\lim_{n\to\infty} \Big[\ln[\mathrm{GM}(1^s, \ldots, n^s)] - \ln(n^s) \Big] = -s.$$

In fact, using the Riemann sum yields that

$$\lim_{n\to\infty} \Big[\ln[\mathrm{GM}(1^s, \ldots, n^s)] - \ln(n^s) \Big] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{i}{n}\right)^s = \int_0^1 \ln(x^s)dx = -s.$$

Similarly, (3) follows from

$$\lim_{n\to\infty} \frac{\mathrm{AM}(1^s, \ldots, n^s)}{n^s} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{i}{n}\right)^s = \int_0^1 x^s dx = \frac{1}{s+1}.$$

Now (1) follows from (2) and (3) by division. This method works for $s > -1$, while Kubelka [**1**] assumes $s > 0$.

Some other limits that involve $\mathrm{GM}(1^s, \ldots, n^s)$ and $\mathrm{AM}(1^s, \ldots, n^s)$, and that do not follow directly from (1), can be evaluated by using (2) and (3). For example, using (2) and (3) we can show that

$$\lim_{n\to\infty} \frac{[\mathrm{GM}(1^s, \ldots, n^s)]^2}{[\mathrm{AM}(1^s, \ldots, n^s)]^2 + n^s\ \mathrm{GM}(1^s, \ldots, n^s)} = \frac{e^{-s}(s+1)^2}{e^s + (s+1)^2}.$$

REFERENCE

1. R. P. Kubelka, Means to an end, *Mathematics Magazine* **74** (2001) 141–142.

Summary If $s > -1$, then the limit of the ratio of the geometric mean of $1^s, \ldots, 2^s$ to their arithmetic mean, as n increases to infinity, is $(s+1)/e^s$. This is proved using Riemann sums. Similar limits are also established for the arithmetic and geometric means separately.

$\lim_{m\to\infty} \sum_{k=0}^{m} (k/m)^m = e/(e-1)$

FINBARR HOLLAND
School of Mathematical Sciences
University College Cork
Cork, Ireland
f.holland@ucc.ie

In his interesting article [**4**], Michael Spivey illustrates the utility and importance of the Euler-Maclaurin formula by examining the asymptotic behavior of the power sums

$$1^m + 2^m + 3^m + \cdots + m^m$$

as $m \to \infty$. In particular, he uses this formula to evaluate the limit mentioned in the title of this Note. Here we describe two alternative ways to determine this limit. The first of these is elementary, but *ad-hoc*, while the second demonstrates the power and elegance of the Lebesgue integral, and is, perhaps, more appealing.

To set the scene, note that, if $m \geq 1$, then by reversing the sum we see that

$$\begin{aligned}\sum_{k=0}^{m} \left(\frac{k}{m}\right)^m &= \sum_{k=0}^{m} \left(\frac{m-k}{m}\right)^m \\ &= \sum_{k=0}^{m} \left(1 - \frac{k}{m}\right)^m \\ &= \sum_{k=0}^{\infty} u_m(k),\end{aligned}$$

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

$$u_m(k) = \begin{cases} \left(1 - \frac{k}{m}\right)^m, & \text{if } 0 \leq k \leq m, \\ 0, & \text{if } m \leq k. \end{cases}$$

Since the geometric-mean of a finite set of positive numbers does not exceed the arithmetic-mean of the same set of numbers, it follows that, if $1 \leq k \leq m$,

$$\sqrt[m+1]{\left(1 - \frac{k}{m}\right)^m \cdot 1} \leq \frac{m(1 - \frac{k}{m}) + 1}{m+1} = 1 - \frac{k}{m+1},$$

whence, for all $k \geq 0$,

$$0 \leq u_m(k) \leq u_{m+1}(k), \quad m = 1, 2, \ldots .$$

Also, it's familiar that

$$\lim_{m\to\infty} u_m(k) = e^{-k}, \quad k = 0, 1, 2 \ldots .$$

Thus, for each integer $k \geq 0$, the sequence $m \to u_m(k)$ increases to e^{-k}.

Math. Mag. **83** (2010) 51–54. doi:10.4169/002557010X485111.

Armed with these facts, let us now return to the task of evaluating the limit

$$\lim_{m\to\infty}\sum_{k=0}^{\infty}u_m(k).$$

If it were legitimate to interchange the limit operation and the summation displayed here, without qualification, we could conclude without further ado that

$$\lim_{m\to\infty}\sum_{k=0}^{\infty}u_m(k)=\sum_{k=0}^{\infty}\lim_{m\to\infty}u_m(k)=\sum_{k=0}^{\infty}e^{-k}=\frac{e}{e-1}. \tag{1}$$

However, if performed blindly, this maneuver may lead to an absurdity. As a cautionary example, if for $m = 1, 2, \ldots$,

$$a_m(k)=\begin{cases}\frac{1}{m+1}, & \text{if } 0\le k\le m,\\ 0, & \text{if } k\ge m,\end{cases}$$

then

$$1=\lim_{m\to\infty}\sum_{k=0}^{\infty}a_m(k)\neq\sum_{k=0}^{\infty}\lim_{m\to\infty}a_m(k)=0.$$

Therefore, the crux of the matter is the justification of the interchange employed in (1). We do this in two ways.

First we present an elementary approach. Since $0 \le u_m(k) \le e^{-k}$, we see that if $1 \le n \le m$, then

$$\sum_{k=0}^{n}u_m(k)\le\sum_{k=0}^{\infty}u_m(k)\le\sum_{k=0}^{\infty}e^{-k}=\frac{e}{e-1}, \tag{2}$$

But because $u_m(k) \le u_{m+1}(k)$ for all k, m, the sequence

$$m\to\sum_{k=0}^{\infty}u_m(k)$$

is monotonic increasing. It is also bounded above by $e/(e-1)$ as (2) shows. Hence, it is convergent, and its limit is finite. Also, $\lim_{m\to\infty} u_m(k) = e^{-k}$, and so, keeping n fixed, and making m tend to infinity in (2), we deduce that

$$\sum_{k=0}^{n}e^{-k}\le\lim_{m\to\infty}\sum_{k=0}^{\infty}u_m(k)\le\frac{e}{e-1}. \tag{3}$$

Finally, letting n tend to infinity in (3), we see that (1) holds. In other words, the limit

$$\lim_{m\to\infty}\sum_{k=0}^{m}\left(\frac{k}{m}\right)^m$$

exists, and is equal to $e/(e-1)$.

(As the reader may care to confirm, essentially the same argument establishes the following general result: Suppose $(a_m(n))$ is a double sequence of nonnegative real numbers such that, for all natural numbers m, n, $a_m(n) \le a_{m+1}(n)$. Then

$$\lim_{m\to\infty}\sum_{n=1}^{\infty}a_m(n)=\sum_{n=1}^{\infty}\lim_{m\to\infty}a_m(n),$$

even if not all of the limits are finite. Incidentally, the same conclusion holds under the weaker hypothesis that, for all $n \geq 1$, $\lim_{m\to\infty} a_m(n)$ exists and exceeds $a_m(n)$ for all $m \geq 1$.)

Another way to justify (1) is to set the problem in the context of the Lebesgue integral [**1**], and to use one of its crowning glories, namely, Lebesgue's Monotone Convergence Theorem. This theorem states that if (X, μ) is a measure space, and (f_n) is a sequence of measurable functions such that, for all $x \in X$,

$$0 \leq f_n(x) \leq f_{n+1}(x), \quad n = 1, 2, \ldots, \quad \text{and} \quad f(x) = \lim_{n\to\infty} f_n(x),$$

then f is measurable, and

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f_n \, d\mu.$$

This can be brought into play by noting that

$$\sum_{k=0}^{\infty} u_m(k) = \int_{\mathbb{N}_0} u_m \, d\nu,$$

where ν stands for the counting measure on the set of nonnegative integers $\mathbb{N}_0$, so that, if E is any subset of $\mathbb{N}_0$, whether finite or infinite, then $\nu(E)$ is the cardinal number of E. Hence, making a direct appeal to the Monotone Convergence Theorem, we deduce that

$$\begin{aligned}
\lim_{m\to\infty} \sum_{k=0}^{m} \left(\frac{k}{m}\right)^m &= \lim_{m\to\infty} \int_{\mathbb{N}_0} u_m \, d\nu \\
&= \int_{\mathbb{N}_0} e^{-k} \, d\nu(k) \\
&= \sum_{k=0}^{\infty} e^{-k} \\
&= \frac{e}{e-1}.
\end{aligned}$$

Similarly, it can be proved by either of the above methods that, if s is any positive real number, then

$$\lim_{m\to\infty} \sum_{k=0}^{m} \left(\frac{k}{m}\right)^{sm} = \frac{e^s}{e^s - 1}. \tag{4}$$

This raises the question: what's the story if s is complex? Answer: (4) continues to hold provided the real part of s is positive. The quickest way to see this is to use Lebesgue's Dominated Convergence Theorem, which states that if (f_n) is a sequence of measurable functions on a measure space (X, μ) such that, for all $x \in X$, $|f_n(x)| \leq F(x)$, where $\int_X F \, d\mu < \infty$, and $f(x) = \lim_{n\to\infty} f_n(x)$, then

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f_n \, d\mu.$$

We leave it as an exercise for the reader to fill in the details. Alternatively, the same result can be achieved by invoking Tannery's Theorem [**2**], [**3**], which arguably is a forerunner of Lebesgue's convergence theorems.

REFERENCES

1. Robert G. Bartle, *The Elements of Integration and Lebesgue Measure*, John Wiley, New York, 1995.
2. E. T. Copson, *An Introduction to the Theory of Functions of a Complex Variable*, Oxford University Press, New York, 1935.
3. T. J. I'A. Bromwich, *An Introduction to the Theory of Infinite Series*, 2nd ed., Macmillan, London, 1926.
4. Michael Z. Spivey, The Euler-Maclaurin formula and sums of powers, *Mathematics Magazine* **79** (2006) 61–65.

Summary Two proofs are given of the limit relation $\lim_{m\to\infty} \frac{1}{m^m} \sum_{k=0}^{m} k^m = \frac{e}{e-1}$, a result due to Michael Spivey. One is elementary, and suitable for discussion in an introductory course on Analysis; the other is more sophisticated, and uses the machinery of the Lebesgue integral. Generalizations of the result are left as exercises for the reader.

Letter to the Editor

MICHAEL Z. SPIVEY
University of Puget Sound
Tacoma, WA 98416
mspivey@ups.edu

In my paper [**1**] I prove

$$\lim_{m\to\infty} \left[\left(\frac{1}{m}\right)^m + \left(\frac{2}{m}\right)^m + \cdots + \left(\frac{m-1}{m}\right)^m \right] = \frac{1}{e-1}. \tag{1}$$

This result can be generalized. For any fixed integer k.

$$\lim_{m\to\infty} \left[\left(\frac{1}{m}\right)^m + \left(\frac{2}{m}\right)^m + \cdots + \left(\frac{m+k}{m}\right)^m \right] = \frac{e^{k+1}}{e-1}.$$

A simple way to prove this is to observe that, for $k \geq 0$,

$$\lim_{m\to\infty} \left[\left(\frac{m}{m}\right)^m + \left(\frac{m+1}{m}\right)^m + \cdots + \left(\frac{m+k}{m}\right)^m \right] = 1 + e + \ldots + e^k = \frac{e^{k+1}-1}{e-1}, \tag{2}$$

and then sum Equations (1) and (2). A similar argument holds for $k < 0$.

There is also a mistake in my paper, first communicated to the editors by Vito Lampret. Near the end of the paper I wish to show that

$$\lim_{m\to\infty} \sum_{k=1}^{m} \left[\frac{B_k}{k!} m^{1-k} (m(m-1)\cdots(m-k+2)) \right] = \sum_{k=1}^{\infty} \frac{B_k}{k!}. \tag{3}$$

First, I express the left-hand side of Equation (3) as

$$\lim_{m\to\infty} \sum_{k=1}^{m} \frac{B_k}{k!} \left[1 + O\left(\frac{1}{m}\right) \right].$$

This is correct. However, the big-O notation disguises the fact that the implicit constant in $O(1/m)$ is dependent on k. Since k ranges from 1 to m over the sum, the maximum constant for the $O(1/m)$ expressions might depend on m, with the result

that the maximum of the $O(1/m)$ expressions might not actually be $O(1/m)$. Thus, a few lines later, it is invalid to make the claim

$$\lim_{m\to\infty} \sum_{k=1}^{m} \frac{B_k}{k!} \left[O\left(\frac{1}{m}\right)\right] = \lim_{m\to\infty} O\left(\frac{1}{m}\right) \sum_{k=1}^{m} \frac{B_k}{k!} = 0,$$

even though $\sum_{k=1}^{\infty} B_k/k!$ converges.

However, this can be corrected fairly easily. Pick $\epsilon > 0$ and find p such that $\sum_{k>p} \frac{|B_k|}{k!} < \epsilon$. Then write

$$\begin{aligned}
&\lim_{m\to\infty} \sum_{k=1}^{m} \left[\frac{B_k}{k!} m^{1-k} (m(m-1)\cdots(m-k+2)) \right] \\
&\quad = \lim_{m\to\infty} \sum_{k=1}^{p} \left[\frac{B_k}{k!} m^{1-k} (m(m-1)\cdots(m-k+2)) \right] \\
&\qquad + \lim_{m\to\infty} \sum_{k>p} \left[\frac{B_k}{k!} m^{1-k} (m(m-1)\cdots(m-k+2)) \right].
\end{aligned}$$

The sum in the first term to the right of the equals sign has only p terms in it, and so the method in [**1**] is valid for this term. Thus the first term is within ϵ of $\sum_{k=1}^{\infty} B_k/k!$. Also, it is easy to see that the second term is within ϵ of 0. Equation (3) follows.

REFERENCE

1. Michael Z. Spivey, The Euler-Maclaurin formula and sums of powers, *Mathematics Magazine* **79** (2006) 61–65.

Integer-Coefficient Polynomials Have Prime-Rich Images

BRYAN BISCHOF
Kansas State University
Manhattan KS 66506-2602
Schof@math.ksu.edu

JAVIER GOMEZ-CALDERON
The Pennsylvania State University–New Kensington
New Kensington, PA 15068-1765
jxg11@psu.edu

ANDREW PERRIELLO
University of Pittsburgh
Pittsburgh, PA 15260
acp27@Pitt.edu

Since the time of Euler, mathematicians have been investigating polynomials with integer coefficients and the values they take on at integer points. It is well known, for example, that a nonconstant polynomial $f(x)$ with integer coefficients produces at least one composite image [**1**, p. 46].

Math. Mag. **83** (2010) 55–57. doi:10.4169/002557010X480008.

In this note, we use Taylor expansions to improve this elementary result, showing that $f(x)$ takes an infinite number of composite values. Given a positive integer n, we show that $f(x)$ takes an infinite number of values that are divisible by at least n distinct primes, and an infinite number of values that are divisible by p^n for some prime p.

Notation For a nonzero integer c, let $\omega(c)$ denote the number of distinct prime numbers that divide c. For example, $\omega(700) = \omega(2^2 \times 5^2 \times 7) = 3$. Similarly, for an integer c and a prime p, let $\psi_p(c)$ denote the highest power of p that divides c, that is, $\psi_p(c) = e$ if and only if p^e divides c but p^{e+1} does not divide c. For example, $\psi_7(3773) = \psi_7(7^3 \times 11) = 3$.

Now, and for the rest of this paper, let $f(x)$ denote a polynomial with integer coefficients and degree $d > 1$. In this notation, we show that given a positive integer n, there are infinitely many integers b such that $\omega(f(b)) > n$ and, for some prime p, infinitely many integers b such that $\psi_p(f(b)) > n$.

Abundant prime factors Most classic number theory textbooks use the Taylor polynomial to show that $f(x)$ produces at least one composite image; that is, $\omega(f(b)) > 1$ for at least one composite image; that is, either $\omega(f(b)) > 1$ for some integer b or $\phi_p(f(b)) > 1$ for some integer b and some prime p. Following this lead we apply the Taylor polynomial to show that $\omega(f(b)) > n$ for an infinite number of integers b. On route to a contradiction, let us assume that there is a constant N such that $\omega(f(a)) < N$ for all integer a such that $f(a) \neq 0$. Thus, the set of positive integers $\Omega = \{\omega(f(a)) : f(a) \neq 0, a \in \mathbb{Z}\}$ has a largest element, call it $n = \omega(f(b)) > 1$. Suppose that the prime factorization of $f(b) = c$ is given by

$$f(b) = c = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}.$$

Then the dth Taylor polynomial for f based at b is

$$\begin{aligned} f(b + tc^2) &= c + \frac{f^{(1)}(b)}{1!}tc^2 + \frac{f^{(2)}(b)}{2!}t^2c^4 + \cdots + \frac{f^{(d)}(b)}{d!}t^d c^{2d} \\ &= c + c^2\left(\frac{f^{(1)}(b)}{1!}t + \frac{f^{(2)}(b)}{2!}c^2t^2 + \cdots + \frac{f^{(d)}(b)}{d!}c^{2d-2}t^d\right). \end{aligned}$$

Inspection shows that $f(b + tc^2)$ is divisible by c for every integer t. One also sees, since $f(x)$ is not constant, that the expression in parenthesis is not zero.

From the first of these observations, since we assume that $\omega(f(b + tc^2)) \leq n$, the primes that divide c are exactly those that divide $f(b + tc^2)$, provided $f(b + tc^2) \neq 0$, and the powers of those primes in the factorization of $f(b + tc^2)$ are at least as high as the powers in c. This allows us to write $f(b + tc^2) = \pm p_1^{e_1} p_2^{e_2} \cdots p_n^{e_n}$, where $\alpha_i \leq e_i$ for all $1 \leq i \leq n$, as long as $f(b + tc^2) \neq 0$.

Furthermore, if $\alpha_i < e_i$, then $p_i^{\alpha_i+1}$ divides c^2 and $f(b + tc^2)$ and consequently also divides $c = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}$, which contradicts the Fundamental Theorem of Arithmetic. Hence, $f(b + tc^2) = \pm c$, if $f(b + tc^2) \neq 0$, which contradicts the Fundamental Theorem of Algebra. Therefore, the set $\Omega = \{\omega(f(a)) : a \in \mathbb{Z}, f(a) \neq 0\}$ is not bounded above and we have proved the following result.

$$\textit{Given an integer } n > 0, \omega\big(f(b)\big) > n \textit{ for an infinite number of integers } b. \qquad (1)$$

High powers of some primes Now assume that $f(x)$ is irreducible. Then $f(x)$ and its derivative $f^{(1)}(x)$ are relatively prime and, therefore,

$$h(x)f(x) + g(x)f^{(1)}(x) = r$$

for some polynomials $h(x)$ and $g(x)$ with integer coefficients, and a nonzero integer r. Then, applying (1), we choose an integer b and a prime p so that p divides $f(b) \neq 0$ but does not divide r. Hence, $h(b)f(b) + g(b)f^{(1)}(b) = r$ and so $f^{(1)}(b)$ is not divisible by p.

Therefore, p and $f^{(1)}(b)$ are relatively prime and so

$$px + f^{(1)}(b)y = 1$$

for some integers x and y. We now write $f(b) = mp^e$ where $\psi_p(f(b)) = e \geq 1$ and work as we did before to obtain

$$\begin{aligned}
&f(b - myp^e) \\
&\quad = mp^e + \frac{f^{(1)}(b)(-myp^e)}{1!} + \frac{f^{(2)}(b)(-myp^e)^2}{2!} + \cdots + \frac{f^d(b)(-myp^e)^d}{d!} \\
&\quad = p^e \left(m\left(1 - f^{(1)}(b)y\right) + \frac{f^{(2)}(b)p^e(-my)^2}{2!} + \cdots + \frac{f^{(d)}(b)p^{e(d-1)}(-my)^d}{d!} \right) \\
&\quad = p^e \left(m(px) + \frac{f^{(2)}(b)p^e(-my)^2}{2!} + \cdots + \frac{f^{(d)}(b)p^{e(d-1)}(-my)^d}{d!} \right) \\
&\quad = p^{e+1} \left(mx + \frac{f^{(2)}(b)p^{e-1}(-my)^2}{2!} + \cdots + \frac{f^{(d)}(b)p^{e(d-1)-1}(-my)^d}{d!} \right).
\end{aligned}$$

Hence,

$$1 \leq \psi_p\big(f(b)\big) = e < e + 1 \leq \psi_p\big(f(b - mp^e)\big).$$

Further, since $f^{(1)}(b - myp^e) = f^{(1)}(b) \neq 0 \pmod{p}$, the process can be repeated as many times as we please. Therefore, we have proved that the set

$$\big\{\psi_p\big(f(a)\big) : a \in \mathbb{Z},\ p \text{ prime}\big\}$$

is not bounded above. Thus, we have proved the following result.

Given $n > 0$, $\psi_p\big(f(b)\big) > n$ for a prime p and an infinite number of integers b. (2)

We complete our paper with the following immediate consequences of (1) and (2).

For an infinite number of integers b, $f(b)$ is not a prime power. (3)

For an infinite number of integers b, $f(b)$ is not square-free. (4)

There are infinitely many prime numbers. (5)

REFERENCE

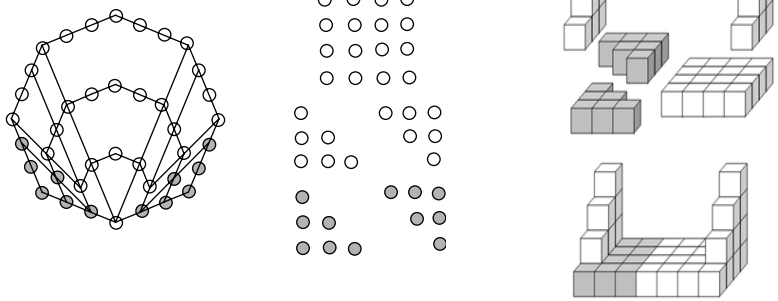
1. Gareth A. Jones and J. Mary Jones, *Elementary Number Theory*, Springer, London, 1998.

Summary It is well known that a nonconstant polynomial $f(x)$ with integer coefficients produces, for integer values of x, at least one composite image. In this note, we use Taylor expansions to improve this elementary result, showing that $f(x)$ takes an infinite number of composite values. Given a positive integer n, we show that $f(x)$ takes an infinite number of values that are divisible by at least n distinct primes, and an infinite number of values that are divisible by p^n for some prime p.

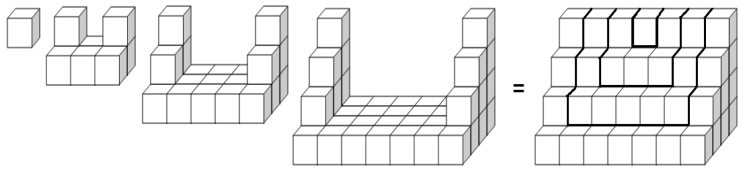
Proof Without Words: Sums of Octagonal Numbers

$$Q_k = 1 + 7 + 13 + \cdots + (6k-5) = k(3k-2) \Rightarrow \sum_{k=1}^{n} Q_k = \frac{n(n+1)(2n-1)}{2}$$

For $k = 4$:



For $n = 4$:



REFERENCE

1. J. O. Chilaka, Sums of octagonal numbers, *Mathematics in Computer Education* **33**(1) (Winter 1999) 62. Also appears in *Proofs Without Words II*, R. Nelsen, ed., MAA, 2000, p.104.

—Hasan Unal
Yildiz Technical University
34210, Istanbul, Turkey

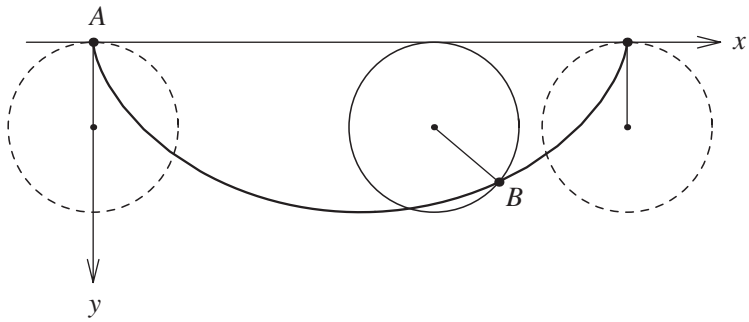
Yet Another Elementary Solution of the Brachistochrone Problem

GARY BROOKFIELD
California State University, Los Angeles
Los Angeles, CA 90032-8204
gbrookf@calstatela.edu

In 1696 Johann Bernoulli issued a famous challenge to his fellow mathematicians:

> Given two points A and B in a vertical plane, find the curve connecting the two points such that an object, starting with zero velocity at A, slides without friction along the curve to B in the least possible time.

Such a curve is called a *brachistochrone*. Newton, Leibniz, l'Hôpital, Jakob Bernoulli (Johann's brother), and the challenger were able to show that a brachistochrone is a segment of a cycloid arc. By a cycloid arc we mean the curve traced out by a point on the rim of a disk as it rolls once along a line. The graph shows the cycloid arc formed by a disk rolling underneath a horizontal line, which is the orientation appropriate for our problem.



Since the object starts with zero speed at A, this point lies at one end of the cycloid arc. With the coordinate system shown, the cycloid arc is given by the equations

$$x = R(\theta - \sin\theta) \quad \text{and} \quad y = R(1 - \cos\theta), \tag{1}$$

with R being the radius of the disk and θ the angle that the disk has rotated from its starting position at A.

The brachistochrone problem is considered to be the beginning of the calculus of variations [**3, 4**], and a modern solution [**8**] would make use of general methods from that branch of mathematics: the Euler, Lagrange, and Jacobi tests, the Weierstrass excess function and more. Even so, many solutions that avoid the calculus of variations have been published [**1, 2, 6**]. The solution we present here amounts to little more than a change of coordinate systems, and is general enough that we prove that the cycloid arc yields the minimum travel time, not just among curves that are smooth, but also among curves that have loops and corners.

To begin, we set up a Cartesian coordinate system for the vertical plane containing A and B as above, with the x-axis horizontal and the positive y-axis pointing down. The coordinates of the sliding object (x, y) are functions of time t on an interval $[0, T]$ such

Math. Mag. **83** (2010) 59–63. doi:10.4169/002557010X480017.

that $A = (0, 0) = (x(0), y(0))$ and $B = (x(T), y(T))$. The number T is, of course, the travel time of the object and the quantity we want to minimize.

Since we assume that there is no friction, the sum of the kinetic energy and the gravitational potential energy, $E = \frac{1}{2}mv^2 - mgy$, is a constant of the motion. Here v is the speed of the object, m is the mass of the object, and $g = 9.8\ \text{m/s}^2$ is the acceleration due to gravity at the Earth's surface. By construction, we have $y = 0$ and $v = 0$ at A, so $E = 0$, and

$$2gy = v^2 \tag{2}$$

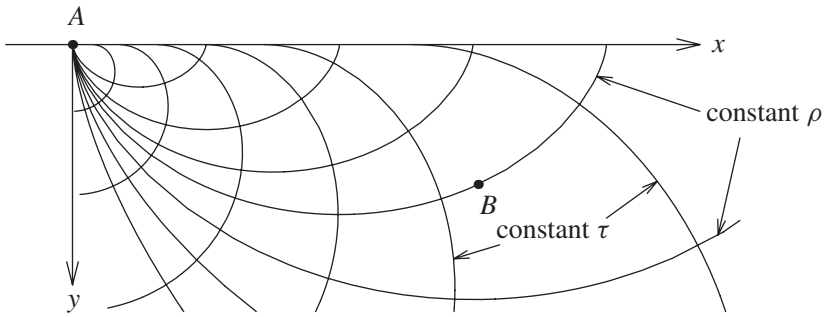
throughout the object's motion.

In what follows, we require that $y > 0$ and $x > 0$ except at A and (possibly) B. It is plausible that any trajectory which minimizes travel time will satisfy these conditions. Indeed, $y \geq 0$ follows from (2).

We now change to a coordinate system in which the expected cycloid solutions to the brachistochrone problem are built in. Specifically, we introduce new coordinates ρ and τ, which are related to x and y by

$$x = \rho\tau - \rho^2 \sin\frac{\tau}{\rho} \quad \text{and} \quad y = \rho^2\left(1 - \cos\frac{\tau}{\rho}\right) \tag{3}$$

where $0 < \rho$ and $0 \leq \tau \leq 2\pi\rho$. These equations are just (1) with $R = \rho^2$ and $\theta = \tau/\rho$. In particular, for a fixed $\rho > 0$, the curve parametrized by τ is a cycloid arc made by rolling a disk of radius $R = \rho^2$ along the x-axis. The graph shows several of these cycloid arcs, as well as some constant τ curves, and makes plausible the fact, which we prove later, that (3) represents a change of coordinate systems.



We now suppose that all relevant trajectories of the object are given by functions ρ and τ of time on the interval $[0, T]$ that determine the Cartesian coordinates (x, y) of the object by (3). Notice from (3) that, since $\rho > 0$, the point A has zero τ-coordinate, and so $\tau(0) = 0$. We write $\dot{x}$, $\dot{y}$, $\dot{\tau}$, and $\dot{\rho}$ for the derivatives of these functions with respect to time. Using the chain rule we can express $\dot{x}$ and $\dot{y}$ in terms of $\dot{\tau}$ and $\dot{\rho}$:

$$\dot{x} = \frac{\partial x}{\partial \tau}\dot{\tau} + \frac{\partial x}{\partial \rho}\dot{\rho} = \left(\rho - \rho\cos\frac{\tau}{\rho}\right)\dot{\tau} + \left(\tau + \tau\cos\frac{\tau}{\rho} - 2\rho\sin\frac{\tau}{\rho}\right)\dot{\rho},$$

$$\dot{y} = \frac{\partial y}{\partial \tau}\dot{\tau} + \frac{\partial y}{\partial \rho}\dot{\rho} = \left(\rho\sin\frac{\tau}{\rho}\right)\dot{\tau} + \left(2\rho - 2\rho\cos\frac{\tau}{\rho} - \tau\sin\frac{\tau}{\rho}\right)\dot{\rho}.$$

With a bit of calculation, which the reader may enjoy carrying out, (2) can be also be written in terms of $\dot{\tau}$ and $\dot{\rho}$.

$$\begin{aligned} 2gy = v^2 &= \dot{x}^2 + \dot{y}^2 \\ &= 2\rho^2 \left(1 - \cos\frac{\tau}{\rho}\right)\dot{\tau}^2 \\ &\quad + 2\left(4\,\rho^2\left(1 - \cos\frac{\tau}{\rho}\right) - 4\,\rho\,\tau\,\sin\frac{\tau}{\rho} + \tau^2\left(1 + \cos\frac{\tau}{\rho}\right)\right)\dot{\rho}^2 \\ &= 2y\,\dot{\tau}^2 + 4\left(2\rho\,\sin\frac{\tau}{2\rho} - \tau\,\cos\frac{\tau}{2\rho}\right)^2\dot{\rho}^2 \end{aligned} \tag{4}$$

Using this equation it is now easy to solve the brachistochrone problem. The term in $\dot{\rho}^2$ is nonnegative, so $2y\dot{\tau}^2 \le 2gy$ and, since $y > 0$ except at A and (possibly) B, we have $\dot{\tau} \le \sqrt{g}$ except perhaps at $t = 0$ and $t = T$. Integrating this inequality on the interval $[0, T]$ gives

$$\tau(T) = \int_0^T \dot{\tau}\,dt \le \int_0^T \sqrt{g}\,dt = \sqrt{g}\,T, \tag{5}$$

or $\tau(T) \le \sqrt{g}\,T$. Thus the time taken for the object to travel from A to B is bounded below (except for the factor $\sqrt{g}$) by the τ-coordinate of B.

One obvious way to obtain this minimum travel time is to set $\dot{\tau} = \sqrt{g}$ and $\dot{\rho} = 0$, since then (4) holds and we get equality in (5). This, of course, means that ρ is a constant and the path of the object is a cycloid arc.

When ρ is a constant, we have $\dot{\theta} = \dot{\tau}/\rho = \sqrt{g}/\rho = \sqrt{g/R}$, and so the object's motion is the same as a point on the rim of a disk of radius R rolling along the x-axis with constant angular velocity $\omega = \dot{\theta} = \sqrt{g/R}$.

For example, suppose that the points A and B are L units apart at the same elevation. Then A and B are the end points of a cycloid arc made by one complete rotation of a disk of radius $R = L/2\pi$. Thus $\omega = \dot{\theta} = \sqrt{2\pi g/L}$ and the time taken to travel from A to B is $T = 2\pi/\omega = \sqrt{2\pi L/g}$. If L is 100 meters and $g = 9.8\ \text{m/s}^2$, then $T \approx 8$ seconds—faster than the world record time for sprinters over the same distance.

There are some important questions left unanswered by our discussion so far. Is there a unique cycloid arc joining the given points A and B? (As noted by Zeng [**9**], this question is often overlooked in the literature.) Which points (x, y) in the plane can be expressed in the form (3) for some uniquely determined τ and ρ? Is the cycloid arc the only way to get equality in (5) and hence minimum travel time?

To answer these questions, we first state some simple trigonometric inequalities, which the reader can easily prove using calculus and double angle identities.

LEMMA 1. *The following inequalities hold for* $0 < \theta < 2\pi$*:*

(1) $0 < \theta - \sin\theta$.

(2) $0 < \sin\dfrac{\theta}{2} - \dfrac{\theta}{2}\cos\dfrac{\theta}{2}$.

(3) $0 < 2(1 - \cos\theta) - \theta\sin\theta$.

Inequalities (1) *and* (2) *hold also when* $\theta = 2\pi$.

We now use Lemma 1 to show that each relevant point in the plane has unique ρ-τ coordinates that satisfy (3), and so there is a unique cycloid arc joining A and B.

LEMMA 2. *If $x > 0$ and $y \geq 0$, then there are unique $R > 0$ and $0 \leq \theta \leq 2\pi$ satisfying* (1)*, as well as unique $\rho > 0$ and $0 \leq \tau \leq 2\pi\rho$ satisfying* (3)*.*

Proof. The function

$$h(\theta) = \frac{1 - \cos\theta}{\theta - \sin\theta}$$

is defined on $(0, 2\pi]$, by the first inequality in Lemma 1, and has the derivative

$$h'(\theta) = \frac{2(\cos\theta - 1) + \theta\sin\theta}{(\theta - \sin\theta)^2}.$$

By the third part of Lemma 1, $h'(\theta) < 0$ on $(0, 2\pi)$ and so h is strictly decreasing on $(0, 2\pi]$.

We also have $h(2\pi) = 0$, and, by l'Hôpital's rule, $\lim_{\theta\to 0^+} h(\theta) = \infty$. Since $y/x \geq 0$ and h is continuous on $(0, 2\pi]$, the Intermediate Value Theorem guarantees the existence of some θ in $(0, 2\pi]$ such that $h(\theta) = y/x$. Since h is strictly decreasing, θ is unique. Now it is easy to check that θ and $R = x/(\theta - \sin\theta)$ is the unique solution of (1), and that $\rho = \sqrt{R}$ and $\tau = \theta\rho$ is the unique solution of (3). ∎

Since each point (x, y) with $x > 0$ and $y \geq 0$ corresponds to uniquely determined ρ and τ, the equations in (3) represent a change of coordinate systems for the first quadrant of the xy-plane. So long as B is in this quadrant, its ρ coordinate determines a unique cycloid arc of the form (1) that passes through it. Moreover, if the object remains in this quadrant, its motion can be described by functions ρ and τ of time.

We claimed at the beginning of this note that our proof shows that the cycloid arc yields the minimum travel time among curves that may have loops or corners. Since we are using two functions of time, x and y, to describe possible paths of the object, loops are not a problem. At a corner in the path, however, the derivatives of x, y, ρ, and τ may not exist. Since we have so far implicitly assumed that these functions are differentiable, we now need to see whether our discussion can be generalized to functions which are not differentiable everywhere. For our arguments to work, we need that the integral in (5) exists and that τ is the indefinite integral of $\dot{\tau}$. This happens if and only if τ is *absolutely continuous*. Royden [**7**, Chapter 5] gives the definition and properties of absolutely continuous functions—including the fact that such functions are differentiable almost everywhere and are the indefinite integrals of their derivatives.

It is then natural to suppose that τ and ρ are both absolutely continuous, and that (4) holds almost everywhere. With these assumptions, we can prove that the cycloid arc is the only brachistochrone. If the concept of absolute continuity is unfamiliar, the reader can show that the argument below works with the stronger assumption that τ and ρ have continuous, or piecewise continuous, derivatives.

Suppose that, for some absolutely continuous functions τ and ρ, equality is attained in (5):

$$\int_0^T \dot{\tau}\, dt = \int_0^T \sqrt{g}\, dt.$$

Since (4) holds almost everywhere, we also have $\dot{\tau} \leq \sqrt{g}$ almost everywhere. These conditions imply that $\dot{\tau} = \sqrt{g}$ almost everywhere. Plugging this result into (4), and using the second fact from Lemma 1 to show that the coefficient of $\dot{\rho}^2$ is nonzero, we get that $\dot{\rho} = 0$ almost everywhere. This implies that ρ is a constant function, and hence, that the minimum travel time is attained only by the cycloid arc.

REFERENCES

1. D. C. Benson, An elementary solution of the brachistochrone problem, *Amer. Math. Monthly* **76** (1969) 890–894. doi:10.2307/2317941
2. G. A. Bliss, *Calculus of Variations*, MAA, 1925.
3. H. Erlichson, Johann Bernoulli's brachistochrone solution using Fermat's principle of least time, *Eur. J. Phys.* **20** (1999) 299–304. doi:10.1088/0143-0807/20/5/301
4. H. H. Goldstine, *A History of the Calculus of Variations from the 17th through the 19th Century,* Springer Verlag, New York, 1980.
5. Nils Johnson, The brachistochrone problem, *College Math J.* **35** (2004) 192–197. doi:10.2307/4146894
6. W. Hrusa and J. L. Troutman, Elementary characterization of classical minima, *Amer. Math. Monthly* **88** (1981) 321–327. doi:10.2307/2320106
7. H. Royden, *Real Analysis*, 3rd ed., Prentice Hall, 1988.
8. F. Y. M. Wan, *An Introduction to the Calculus of Variations and Its Applications*, 2nd ed., Chapman & Hall, New York, 1995.
9. Jim Zeng, A note on the brachistochrone problem, *College Math J.* **27** (1996) 206–208. doi:10.2307/2687169

Summary This note provides an elementary solution of the brachistochrone problem. This problem is to find the curve connecting two given points so that an object slides without friction along the curve from one point to the other point in the least possible time. The key is to introduce a coordinate system where the expected cycloid solutions are built in.

A Property Characterizing the Catenary

EDWARD PARKER
Brown University
Providence, RI 02912
Edward_Parker@brown.edu

Let $y(x)$ be any strictly positive C^1 function, and consider the curve which is the graph of $y(x)$ over an interval $[a, b]$ in the function's domain. This curve has a well-defined arc length and there is a well-defined area under it. Are there any functions which have the property that the ratio of the area under the curve to the curve's arc length is independent of the interval over which they are measured?

In order for this property to hold, we must have [**1**, page 279]

$$\int_a^b y(x)\,dx = k\int_a^b \sqrt{1+y'(x)^2}\,dx,$$

where k is a positive constant independent of a and b. In order for this to be true for all intervals $[a, b]$ in the function's domain, the integrands must be identically equal. Bringing k inside the right-hand integral, setting the integrands equal, and solving for $y'(x)$ yields

$$y'(x) = \pm\frac{\sqrt{y(x)^2-k^2}}{k}. \tag{1}$$

Clearly $y(x) = k$ is a solution. When $y(x) \neq k$, we can separate variables and find a more surprising result:

$$y(x) = k\cosh\left(\frac{x-c}{k}\right),$$

Math. Mag. **83** (2010) 63–64. doi:10.4169/002557010X485120.

which is the well-known catenary curve. Therefore catenaries and constant functions are the only curves that are twice-differentiable everywhere and have the property that they bound an area proportional to their arc length over any horizontal interval. (If we relax our smoothness requirements, we could also allow curves that are defined piecewise to be either constant or catenary curves over different intervals).

We have obtained the geometric result that at every point on a catenary $y\,dx = k\,ds$, where ds is the arc length differential. This property leads directly to the interesting result that for every interval $[a, b]$, the geometric centroid of the area under a catenary curve defined on this interval is the midpoint of the perpendicular segment connecting the centroid of the curve itself and the x-axis. Note that the centroid of the curve lies above the curve itself.

The result that the content of a region bounded by a catenary is proportional to the content of the boundary itself over any interval extends directly to the three-dimensional case. If a surface of revolution has the property that the ratio of the volume it encloses to its surface area is independent of the interval on which it is defined, then it must obey the equation [**1**, pp. 326 and 466]

$$\int_a^b \pi y(x)^2\,dx = k\int_a^b 2\pi y(x)\sqrt{1+y'(x)^2}\,dx.$$

After setting the integrands equal as before, a factor of $\pi y(x)$ cancels from both sides and we can rearrange to obtain equation (1) again, but with each k replaced by the term $2k$. Therefore the surface of revolution generated by

$$y(x) = 2k\cosh\left(\frac{x-c}{2k}\right),$$

which is the famous catenoid surface, is the only twice-differentiable surface of revolution other than the cylinder of radius $2k$ which encloses a volume that is k times its surface area over every horizontal interval.

REFERENCE

1. Jerrold E. Marsden and Anthony J. Tromba, *Vector Calculus*, 5th ed., W. H. Freeman, New York, 2003.

Summary We show that the area under a catenary curve is proportional to its length in the following sense: given a catenary curve, we can take any horizontal interval and examine the ratio of the area under the curve to the length of the curve on that interval, and we find that the resulting ratio is independent of the chosen interval. This property extends to the three-dimensional case as well: the volume contained by a horizontal interval of a catenoid surface is proportional to its surface area in the same sense. We also show from this property that the centroid of the area under an interval of a catenary is the midpoint of the segment connecting the centroid of the catenary and the x-axis.

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; ROGELIO VALDEZ, Facultad de Ciencias, UAEM, México; WILLIAM WATKINS, California State University, Northridge

PROPOSALS

To be considered for publication, solutions should be received by July 1, 2010.

1836. *Proposed by Michael Wolterman, Washington and Jefferson College, Washington, PA.*

Let $n \geq 3$ be a natural number. Find how many pairwise non-congruent triangles are there among the $\binom{n}{3}$ triangles formed by selecting three vertices of a regular n-gon.

1837. *Proposed by Duong Viet Thong, Nam Dinh University of Technology Education, Nam Dinh City, Vietnam.*

Let $f : [1, 2] \to \mathbb{R}$ be a continuous function such that $\int_1^2 f(x)\,dx = 0$. Prove that there exists a real number c in the open interval $(1, 2)$, such that $cf(c) = \int_c^2 f(x)\,dx$.

1838. *Proposed by Costas Efthimiou, University of Central Florida, Orlando, FL.*

Compute the sum $\displaystyle\sum_{n=0}^{\infty}\sum_{m=1}^{\infty}(-1)^{n+m}\frac{\ln(m+n)}{m+n}$.

1839. *Proposed by Robert A. Russell, New York, NY.*

Consider a sphere of radius 1 and three points A, B, and C on its surface, such that the area of the convex spherical triangle ABC is π. Let L, M, and N be the midpoints of the shortest arcs AB, BC, and CA. Give a characterization of the spherical triangle LMN.

1840. *Proposed by Tuan Le, 12th grade, Fairmont High School, Anaheim, CA.*

Let a, b, and c be nonnegative real numbers such that no two of them are equal to zero. Prove that

$$\frac{a}{b+c}+\frac{b}{c+a}+\frac{c}{a+b}+\frac{3\sqrt[3]{abc}}{2(a+b+c)} \geq 2.$$

Math. Mag. **83** (2010) 65–70. doi:10.4169/002557010X480026.

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.

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.

1834 (corrected). *Proposed by Cosmin Pohoata, student, National College "Tudor Vianu," Bucharest, Romania.*

Let $ABCD$ be a cyclic quadrilateral that also has an inscribed circle with center I, and let ℓ be a line tangent to the incircle. Let A', B', C', and D', respectively, be the projections of A, B, C, and D onto ℓ. Prove that

$$\frac{AA' \cdot CC'}{AI \cdot CI} = \frac{BB' \cdot DD'}{BI \cdot DI}.$$

(This problem was originally published without an essential hypothesis. We apologize to the proposer and to solvers. Solutions will be accepted through July 1, 2010.)

Quickies

Answers to the Quickies are on page 70.

Q997. *Proposed by Éric Pité, Paris, France*

Let p be a prime and n a positive integer. Show that $n!$ divides $\prod_{k=0}^{n-1}(p^n - p^k)$.

Q998. *Proposed by Ovidiu Furdui, Campia Turzii, Cluj, Romania.*

Let $f, g : [0, 1] \to \mathbb{R}$ be two continuous functions. Prove that

$$\lim_{n\to\infty} \int_0^1 f(x^n)g(x)\,dx = f(0)\int_0^1 g(x)\,dx.$$

Solutions

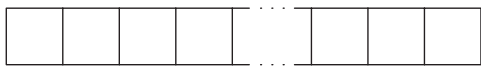
Calculating the Wiener index **February 2009**

1811. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.*

Given a connected graph G with vertices $v_1, v_2, \ldots, v_n$, let $d_{i,j}$ denote the distance from v_i to v_j. (That is, $d_{i,j}$ is the minimal number of edges that must be traversed in traveling from v_i to v_j.) The Wiener index $W(G)$ of G is defined by

$$W(G) = \sum_{1\le i<j\le n} d_{i,j}.$$

a. Find the Wiener index for the grid-like graph



on $2n$ vertices.

b. Find the Wiener index for the comb-like graph

on $2n$ vertices.

Solution by G.R.A.20 Problems Group, Rome, Italy.

Let

$$f(n) = \sum_{i=1}^{n}\sum_{j=i+1}^{n} |i - j| = \binom{n+1}{3}.$$

For $1 \le i \le n$, let $(1,\ i)$ denote the ith vertex from the right in the bottom row of either graph and let $(2,\ i)$ be the ith vertex from the right in the top row of either graph.

a. Let G_1 be the grid-like graph. In this graph, the distance between $(1,\ i)$ and $(2,\ j)$ is $|i - j| + 1$, and the distance between $(1,\ i)$ and $(1,\ j)$ or $(2,\ i)$ and $(2,\ j)$ is $|i - j|$. Hence

$$W(G_1) = \sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|+1) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}|i-j|$$
$$= 4f(n) + n^2 = \frac{n(n+2)(2n-1)}{3}.$$

b. Let G_2 be the comb-like graph. The the distance between $(1,\ i)$ and $(2,\ j)$ is $|i - j| + 1$, distance between $(1,\ i)$ and $(1,\ j)$ is $|i - j|$, and for $i \ne j$, the distance between $(2,\ i)$ and $(2,\ j)$ is $|i - j| + 2$. Hence

$$W(G_2) = \sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|+1) + \sum_{i=1}^{n}\sum_{j=i+1}^{n}|i-j| + \sum_{i=1}^{n}\sum_{j=i+1}^{n}(|i-j|+2)$$
$$= 4f(n) + n^2 + 2\binom{n}{2} = \frac{n(2n^2+6n-5)}{3}.$$

Also solved by Steve Abbott, Michael Andrioli, Michel Bataille (France), J. C. Binz (Switzerland), Robert Calcaterra, Mark Crawford, Chip Curtis, A. K. Desai, Robert L. Doucette, Joeseph Fredette, Fejéntaláltuka Szeged Problem Solving Group (Hungary), Dmitry Fleischman, David Getling (New Zealand), Sharan Gopal, (India), Russell Jay Hendel, Santhosh Karnik, Andrew Krull, Harris Kwong, David P. Lang, Jeremy Lee, Thomas C. Lominac, Reiner Martin, David Nacin, José H. Nieto (Venezuela), José M. Pacheco and Ángel Plaza (Spain), Robert Patenaude, Vadim Ponomarenko, Robert Pratt, Joel Schlosberg, Harry Sedinger, Nicholas C. Singer, Skidmore College Problem Group, John H. Smith, Albert Stadler (Switzerland), Philip D. Straffin, John Sumner and Aida Kadic-Galeb, Texas State University Problem Solvers Group, Bob Tomper, Alexey Vorobyov, Todd G. Will, Michael Woltermann, and the proposer.

A flooral arrangement **February 2009**

1812. *Proposed by Bob Tomper, University of North Dakota, Grand Forks, ND.*

Let m and n be relatively prime positive integers. Prove that

$$\sum_{k=1}^{n} k^2 \left\lfloor \frac{km}{n} \right\rfloor = n\sum_{k=1}^{n} k \left\lfloor \frac{km}{n} \right\rfloor - \frac{n(n^2-1)(m-1)}{12}.$$

Solution by Albert Stadler, Herrliberg, Switzerland.

Because m and n are relatively prime, km/n is never an integer for $1 \le k \le n-1$. Hence

$$n\sum_{k=1}^{n} k \left\lfloor \frac{km}{n} \right\rfloor - \sum_{k=1}^{n} k^2 \left\lfloor \frac{km}{n} \right\rfloor = \sum_{k=1}^{n-1} k(n-k) \left\lfloor \frac{km}{n} \right\rfloor$$
$$= \frac{1}{2}\sum_{k=1}^{n-1} k(n-k)\left(\left\lfloor \frac{km}{n} \right\rfloor + \left\lfloor \frac{(n-k)m}{n} \right\rfloor\right)$$
$$= \frac{1}{2}\sum_{k=1}^{n-1} k(n-k)(m-1) = \frac{n(n^2-1)(m-1)}{12}.$$

This completes the proof.

Also solved by Michel Bataille (France), Robert Calcaterra, Chip Curtis, David Doster, Robert L. Doucette, Eric Egge, Dmitry Fleischman, E. S. Friedkin, G.R.A.20 Problem Solving Group (Italy), Omran Kouba (Syria), Peter W. Lindstrom, Kim McInturff, Rituraj Nandan, José H. Nieto (Venezuela), Éric Pité (France), Nicholas C. Singer, John Sumner and Aida Kadic-Galeb, Marian Tetiva (Romania), Giang Tran, Serge Varjabedian (France), Francisco Vial (Chile), Alexey Vorobyov, John B. Zacharias, and the proposer.

An inequality. **February 2009**

1813. *Proposed by Elton Bojaxhiu, Albania, and Enkel Hysnelaj, Australia.*

Let a, b, and c be positive real numbers. Prove that

$$\frac{1}{a(1+b)}+\frac{1}{b(1+c)}+\frac{1}{c(1+a)}\geq\frac{3}{\sqrt[3]{abc}(1+\sqrt[3]{abc})}.$$

Solution by Young Ho Kim, Yonsei University, Seoul, Korea.

The required inequality is equivalent to

$$\begin{aligned}L&=\frac{1}{3}\left(\frac{1+abc}{a(1+b)}+\frac{1+abc}{b(1+c)}+\frac{1+abc}{c(1+a)}\right)\\&\geq\frac{1+abc}{\sqrt[3]{abc}(1+\sqrt[3]{abc})}=\frac{1}{\sqrt[3]{abc}}-1+\sqrt[3]{abc}.\end{aligned}$$

Because

$$1+\frac{1+abc}{a(1+b)}=\frac{1+a}{a(1+b)}+\frac{b(1+c)}{1+b},$$

it follows that

$$L+1=\frac{1}{3}\left(\frac{1+a}{a(1+b)}+\frac{1+b}{b(1+c)}+\frac{1+c}{c(1+a)}+\frac{b(1+c)}{1+b}+\frac{c(1+a)}{1+c}+\frac{a(1+b)}{1+a}\right).$$

Finally, the Arithmetic Mean–Geometric Mean Inequality implies that

$$\begin{aligned}L+1&\geq\sqrt[3]{\frac{1+a}{a(1+b)}\cdot\frac{1+b}{b(1+c)}\cdot\frac{1+c}{c(1+a)}}+\sqrt[3]{\frac{b(1+c)}{1+b}\cdot\frac{c(1+a)}{1+c}\cdot\frac{a(1+b)}{1+a}}\\&=\frac{1}{\sqrt[3]{abc}}+\sqrt[3]{abc},\end{aligned}$$

which completes the proof.

Note. Some readers pointed out that the problem has appeared before. The references given include T. Andreescu, V. Cirtoaje, G. Dospinescu, M. Lascu, *Old and New Inequalities*, GIL Publishing House, 2004; and Problem 2977, proposed by V. Cirtoaje, in *Crux Mathematicorum with Mathematical Mayhem*.

Also solved by Geroge Apostolopoulos (Greece), Michel Bataille (France), Robert Calcaterra, Minh Can, Charles R. Diminnie, Fisher Problem Solving Group, Tom Leong, Omran Kouba (Syria), Valmir Krasniqi (Republic of Kosova), Paolo Perfetti (Italy), Henry Ricardo, C. R. Selvaraj and Suguna Selvaraj, Nicholas C. Singer, Tony Tam, Stan Wagon, and the proposers.

A characterization of the Euler line **February 2009**

1814. *Proposed by Michael Goldenberg and Mark Kaplan, The Ingenuity Project, Baltimore Polytechnic Institute, Baltimore, MD.*

Let $A_1A_2A_3$ be a triangle with circumcenter O, and let B_1 be the midpoint of A_2A_3, B_2 be the midpoint of A_3A_1, and B_3 be the midpoint of A_1A_2. For $-\infty<t\leq\infty$ and

$k = 1, 2, 3$, let $B_{k,t}$ be the point defined by $\overrightarrow{OB_{k,t}} = t\overrightarrow{OB_k}$, (where by $B_{k,\infty}$ we mean the point at infinity in the direction of $\overrightarrow{OB_k}$.) Prove that for any $t \in (-\infty, \infty]$, the lines $A_kB_{k,t}$, $k = 1, 2, 3$, are concurrent, and that the locus of all such points of concurrent is the Euler line of triangle $A_1A_2A_3$.

Solution by Herb Bailey, Rose Hulman Institute of Technology, Terre Haute, IN.

We assume that the triangle is not equilateral, since otherwise the Euler line does not exist. We prove a more general result:

> Let P be a point on the perpendicular bisector of side A_1A_2, and let $B_{k,t}$ be the point defined by $\overrightarrow{PB_{k,t}} = t\overrightarrow{PB_k}$. Then lines $A_kB_{k,t}$, $1 \le k \le 3$, are concurrent and the locus of all such points of concurrence is a line. This line is the Euler line if and only if P is chosen to be the circumcenter O.

Define an x, y-coordinate system with origin at P, x-axis parallel to side A_1A_2, and vertices $A_1(-a, d)$, $A_2(a, d)$, and $A_3(b, c)$, where $a > 0$ and $c \neq d$. Hence the coordinates of midpoint B_1 are $(x_{B_1}, y_{B_1}) = ((a + b)/2, (d + c)/2)$, with similar expressions for midpoints B_2 and B_3. The coordinates of $B_{k,t}$ are then (tx_{B_k}, ty_{B_k}), $k = 1, 2, 3$. Let ℓ_k denote the line $A_kB_{k,t}$, $1 \le k \le 3$ and let Q_{ij}, $1 \le i < j \le 3$ be the intersection of ℓ_i and ℓ_j. Calculations show that

$$Q_{12} = Q_{13} = Q_{23} = \left(\frac{bt}{t+2}, \frac{(c+2d)t}{t+2}\right).$$

Thus ℓ_1, ℓ_2, and ℓ_3 are concurrent. As t varies, these points generate a line provided $c + 2d$ and b are not both 0, that is, provided P is not the centroid of the triangle. If $c + 2d$ and b are not both 0, then the locus of concurrency is the line with equation

$$y = \frac{c+2d}{b}x,$$

where the line is vertical if $b = 0$. The coordinates of the centroid of triangle $A_1A_2A_3$ are $(b/3, (c + 2d)/3)$. If this triangle is not equilateral, then the circumcenter $O = (0, e)$ is distinct from the centroid, and the Euler line has equation

$$y = e + \frac{c+2d-3e}{b}x.$$

This line coincides with the locus of concurrency if and only if $e = 0$, that is, if and only if $P = O$.

Also solved by Michel Bataille (France), Robert Calcaterra, Robert L. Doucette, Dmitry Fleischman, L. R. King, José H. Nieto (Venezuela), Joel Schlosberg, Raul A. Simon (Chile), Albert Stadler (Switzerland), John Sumner and Aida Kadic-Galeb, Texas State University Problem Solvers Group, Alexey Vorobyov, and the proposers.

Subrings of $\mathbb{Q}$. **February 2009**

1815. *Proposed by Stephen J. Herschkorn, Rutgers University, New Brunswick, NJ.*

It is well known that if R is a subring of the ring $\mathbb{Z}$ of integers, then there is a unique positive integer m such that $R = m\mathbb{Z}$. Determine a similar unique characterization for any subring of the ring $\mathbb{Q}$ of rational numbers. What is the cardinality of the class of all subrings of $\mathbb{Q}$? (We do not assume that a ring has a multiplicative identity.)

Solution by Robert Calcaterra, University of Wisconsin–Platteville, Platteville, WI.

For any set P of prime numbers and any positive integer k relatively prime to every element of P, let

$$S_{P,k} = \left\{ \frac{km}{n}, m, n \in \mathbb{Z},\ n > 0, \text{and such that every prime factor of } n \text{ is in } P \right\}.$$

It is easy to check that $S_{P,k}$ is a subring of $\mathbb{Q}$. We will show that every nontrivial subring of $\mathbb{Q}$ is of this form.

Let $R \neq \{0\}$ be a subring of $\mathbb{Q}$, let D be the set of positive denominators of elements in R when they are expressed in lowest terms, and let P be the set of all primes that divide at least one element of D. Note that if $\frac{m}{n} \in R$ with $n \in \mathbb{Z}^+$, then $m = n \cdot \frac{m}{n}$ is also in R. Thus, if $R \neq \{0\}$, then $R \cap \mathbb{Z}^+$ is nonempty. Let k be the minimum element of $R \cap \mathbb{Z}^+$.

Now let $r \in R$. Then there exist an $x \in D$ and an $a \in \mathbb{Z}$ such that x and a are relatively prime and $r = \frac{a}{x}$. Thus $a \in R$, and the remainder when a is divided by k is also in R. By the minimality of k, it follows that this remainder is 0, so a is a multiple of k. Therefore, $r \in S_{P,k}$, so $R \subseteq S_{P,k}$.

Next assume that n is a product of primes in P and let p be a prime divisor of n. Then p is a divisor of y for some $y \in D$. If $\frac{y}{p}$ copies of a fraction with denominator y are added, the result is a fraction with denominator p. Hence $p \in D$. Moreover, since R is closed under multiplication, n must also belong to D. Hence there is a $b \in \mathbb{Z}$ such that b and n are relatively prime and $\frac{b}{n} \in R$. The argument in the previous paragraph implies that $b = kc$ for some integer c, and the Euclidean algorithm implies that $sc + tn = 1$ for some integers s and t. Therefore

$$\frac{km}{n} = sm \cdot \frac{b}{n} + tm \cdot k \in R,$$

for every integer m. This completes the proof that $R = S_{P,k}$, and completes the characterization of the nontrivial subrings of $\mathbb{Q}$.

Because the cardinality of the set of prime numbers is $\aleph_0$, the countably infinite cardinal, the number of subrings of $\mathbb{Q}$ is at least $2^{\aleph_0}$. On the other hand, the number of subsets of $\mathbb{Q}$ is $2^{\aleph_0}$. Hence the cardinality of the set of all subrings of $\mathbb{Q}$ is $2^{\aleph_0}$.

Also solved by Michel Bataille (France), Paul Budney, Fisher Problem Solving Group, Dmitry Fleischman, David P. Lang, Tom Moore, Northwestern University Math Problem Solving Group, Phill Schultz, Nicholas C. Singer, John Sumner and Aida Kadic-Galeb, Tony Tam, Marian Tetiva (Romania), Texas State Problem Solvers Group, Alexey Vorobyov, and the proposer.

Answers

Solutions to the Quickies from page 66.

A997. The group $\mathrm{GL}(n, \mathbb{F}_p)$, consisting of all the invertible $n \times n$ matrices with entries in $\mathbb{F}_p$, is of order $\prod_{k=0}^{n-1}(p^n - p^k)$. The group of $n \times n$ permutation matrices is of order $n!$ and it is a subgroup of $\mathrm{GL}(n, \mathbb{F}_p)$. Hence, by Lagrange Theorem, $n!$ divides $\prod_{k=0}^{n-1}(p^n - p^k)$.

A998. Let $h_n : [0, 1] \to \mathbb{R}$ be given by $h_n(x) = f(x^n)g(x)$ and let $h(x) = \lim_{n\to\infty} h_n(x)$ be the point-wise limit function of $h_n(x)$. Because f and g are continuous, it follows that $h(x) = f(0)g(x)$ for $x \in [0, 1)$, $h(1) = f(1)g(1)$, and the functions $\{h_n(x)\}$ are uniformly bounded, i.e., $|h_n(x)| \leq M$ for all n and all $x \in [0, 1]$. Thus, by the Bounded Convergence Theorem,

$$\lim_{n\to\infty} \int_0^1 h_n(x)\,dx = \int_0^1 \lim_{n\to\infty} h_n(x)\,dx = \int_0^1 h(x)\,dx = f(0)\int_0^1 g(x)\,dx.$$

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles and books 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.

Suzuki, Jeff, *Mathematics in Historical Context*, MAA, 2009; x + 409 pp, $58.50 (member price: $46.95). ISBN 978-0-88385-570-6.

I have complained elsewhere about how calculus is usually taught without any connections pointed out to the history of science or to intellectual history in general (or to much of anything outside of mathematics). Some reasons why are that students perceive calculus pragmatically as just a tool (they are not seeking connections), and that calculus instructors don't know connections (in going to a new edition, one well-known calculus book stripped out most applications because teaching assistants didn't understand them). This welcome book considers parts of the mathematical whole up to the middle of the twentieth century and weaves those parts largely into *political* history. How accurate the political history is, how adequate its interpretation, and how close or tenuous the mathematical tie-ins, I must leave to more learned reviewers to evaluate. Wherever I dipped into the book, though, I was captivated and learned a connection or insight that I had not been aware of. I was reminded of James Burke's 1978 documentary series *Connections*, about technological innovations, and his follow-on books, video series, and columns for *Scientific American*. (The book's index is serviceable for major proper names but is completely inadequate for terms: cossist? Enigma? cube and cosa? group theory? *calculus*?)

Inselberg, Alfred, *Parallel Coordinates*, Springer, 2009; xxvi + 554 pp plus Windows CD-ROM, $79.95. ISBN 978-0-387-21507-5.

The basic idea of parallel coordinates is to represent a space of any number of dimensions by lining up their coordinate axes in the plane of the page. A point in the space is represented by a polygonal line crossing through its coordinates on the axes. The research question is how to discern, detect, and recognize relationships from such a display of a data set: "Parallel coordinates . . . transforms the search for multivariate relations in a data set into a pattern-recognition problem." The analysis, however, begins with inverse questions: How, for example, is a line through the points manifested in the display? Software packages for exploratory data analysis and data mining include parallel coordinates displays, with some of the classification automated. The book includes case studies of data visualization through application of parallel coordinates to collision avoidance (in air traffic control), improved production of computer chips, computer vision, analyzing networks, recognizing a truck from its noise signature (a technique employed in a recent episode of the TV series "Numb3rs"), and—germane even to pure mathematicians—visualizing functions over the complex plane (by Yoav Yaari). The bulk of the book is devoted to the underlying theory, beginning with transformations in projective geometry, how transformations are manifested in the plane with parallel coordinates, and continuing through analysis of lines, hyperplanes, hypersurfaces, and proximity relations. The main prerequisite is linear algebra. (The Preface begins with an alarmingly false version of the tale of John Snow's identification of the source of cholera in London in 1854: Snow did not replace the pump handle but at his instigation the local council removed it; and cholera was not transmitted from touching the handle but from drinking the water. Mathematicians: If you write outside your specialty, to avoid losing credibility get your facts straight! Would-be book editors: Take lots of science in college, so you will be able to recognize such blunders.)

Math. Mag. **83** (2010) 71–72. doi:10.4169/002557010X485157.

Between the Folds. Video, 54 min., plus 30 min. of outtakes. Region 1 DVD, NTSC or PAL format. Directed by Vanessa Gould, edited by Kristi Barlow; executive producer Sally Rosenthal. Available from Green Fuse Films; order form at `http://www.greenfusefilms.com/store.html`. $55 for classroom and non-profit use; $25 for home/private use, also available subtitled in German or in Italian. Discussion guide and downloadable folds at `http://www.pbs.org/independentlens/between-the-folds/getinvolved.html`.

Origami metamorphosed in the middle of the last century from a simple craft to a sophisticated form of sculpture, thanks to Akira Yoshizawa's "breathing life into the paper." Some current origami pieces involve hundreds of diagram steps, thousands of folds, and dozens of hours to create. In the current century, origami has drawn the attention of engineers (airbag design), biologists (protein folding), computer scientists (algorithms for flattening), and mathematicians (what can be done?). This film has utterly entrancing sequences showing artists folding astonishing pieces (and even making the paper); you can see how the "life in the paper," and the emotion in the creations, flows from the vitality in their eyes and speech. Featured toward the end is Erik Demaine (MIT), who demonstrates a "post-modernist" variation: make folds, then make a single cut through. He then mentions that such a procedure can be proved capable of making any (straight-edged) shape. Another speaker emphasizes that origami math is not just compartmentalized into geometry but involves "all of math," mentioning abstract algebra, linear algebra, matrices, and more. Why do origami? "Because," says Demaine, "it's fun." (The discussion guide misspells Yoshizawa's name.)

Hearn, Robert A., and Erik D. Demaine, *Games, Puzzles, and Computation*, A K Peters, 2009; ix + 237 pp, $45. ISBN 978-1-56881-322-6.

Puzzles and games are fun because they are challenging. How challenging? Only a computer scientist, analyzing them in terms of their complexity, can say for sure. And that is what Hearn and Demaine do for a variety of games, using what they call *constraint logic*. They model a game as an oriented graph with edge weights. The constraints are minimum flows at each vertex, a legal move is the reversal of an edge's orientation, and the goal of the game is to reverse a particular edge. The many games considered include the commercial puzzles TipOver (NP-complete), Mastermind (also NP-complete), and Rush Hour (PSPACE-complete), and the two-player games Hex, Amazons, Othello, Gomoku, and Konane (all PSPACE-complete); there is no mention of mancala games. The authors include open questions about many games. (Note to the book's editors: "course grain" on p. 2 should be "coarse grain.")

Tan, Ming To., Guo-Liang Tan, and Man-Lai Tang, Sample surveys with sensitive questions: A nonrandomized response approach, *American Statistician* 63 (1) (February 2009) 9–16.

Randomized response techniques try to assure anonymity in surveys with sensitive questions (e.g, use of drugs). The respondent uses private randomization (e.g., a coin toss) to determine which of two questions (one "sensitive," one "harmless") to answer. Using known probability for the randomization, a researcher can estimate proportions in the population, and their variances. Tan et al. dispense with the coin, using an event with known probability in the population (e.g., being born in the summer). They would have the subject respond "true" or "false" to "I don't use drugs and I was born in the summer." This method is more efficient than traditional randomized response, though nuances of how to use it in practice still need to be worked out.

Sudan, Madhu, Probabilistically checkable proofs, *Communications of the Association for Computing Machinery* 52 (3) (March 2009) 76–84.

"Can a proof be checked without reading it?" Author Sudan discusses two approaches to construct probabilistically checkable proofs (PCPs). The goal is a format in which a researcher can write a proof, such that if the proof is correct, the reviewer will be convinced, and if it is incorrect, the reviewer will reject it "with overwhelming probability." Moreover, the reviewer's verification algorithm would spare the reviewer from reading the entire proof, because the algorithm aspires to be only probabilistically correct. Sudan summarizes research on interactive proofs and holographic proofs. Verifiers for PCPs can be constructed, but an obstacle is the size of PCPs: For an n-bit classical proof, a PCP would have size about $\mathcal{O}\big(n(\log n)\big)^{\mathcal{O}(1)}$.

NEWS AND LETTERS

Guidelines for Authors

What do *you* like to read? What kind of writing can grab the interest of an undergraduate mathematics major? How can MATHEMATICS MAGAZINE serve to remind us all why we chose to study mathematics in the first place? If you keep these questions firmly in mind, you will be well on the way to meeting our editorial guidelines.

General information MATHEMATICS MAGAZINE is an expository journal of undergraduate mathematics. In this section, we amplify our meaning of these words.

Articles submitted to the MAGAZINE should be written in a clear and lively *expository* style. The MAGAZINE is not a research journal; papers in a terse "theorem-proof" style are unsuitable for publication. The best contributions provide a context for the mathematics they deliver, with examples, applications, illustrations, and historical background. We especially welcome papers with historical content, and ones that draw connections among various branches of the mathematical sciences, or connect mathematics to other disciplines.

Every article should contain interesting *mathematics.* Thus, for instance, articles on mathematical pedagogy alone, or articles that consist mainly of computer programs, would be unsuitable.

The MAGAZINE is an *undergraduate* journal in the broad sense that its intended audience is teachers of collegiate mathematics and their students. One goal of the MAGAZINE is to provide stimulating supplements for undergraduate mathematics courses, especially at the upper undergraduate level. Another goal is to inform and refresh the teachers of these courses by revealing new connections or giving a new perspective on history. We also encourage articles that arise from undergraduate research or pose questions to inspire it. In writing for the MAGAZINE, make your work attractive and accessible to non-specialists, including well-prepared undergraduates.

Writing and revising MATHEMATICS MAGAZINE is responsible first to its readers an then to its authors. A manuscript's publishability therefore depends as much on the quality of exposition as the mathematical significance. Our general advice is simple. Say something new in an appealing way, or say something old in a refreshing, new way. But say it clearly and directly, assuming a minimum of background. Our searchable database of past pieces from the MAGAZINE and the *College Mathematics Journal* is reachable from the MAGAZINE'S website and can help you check the novelty of your idea.

Make your writing vigorous, expressive, and informal, using the active voice. Give plenty of examples and minimize computations. Help the reader understand your motivation and share your insights. Illustrate your ideas with visually appealing graphics, including figures, tables, drawings, and photographs.

First impressions are vital. Choose a short, descriptive, and attractive title; feel free to make it funny, if that would draw the reader in. Be sure that the opening sentences provide a welcoming introduction to the entire paper. Readers should know why they ought to invest time reading your work.

Our referees are asked to give detailed suggestions on style, as well as check for mathematical accuracy. In practice, almost every paper requires a careful revision by the author,

Math. Mag. **83** (2010) 73–74. doi:10.4169/002557010X485869.

followed by further editing in our office. To shorten this process, be sure to read your own work carefully, possibly after putting it away for a cooling-off period.

Provide a generous list of references to invite readers—including students—to pursue ideas further. Bibliographies may contain suggested reading along with sources actually referenced. In all cases, cite sources that are currently and readily available.

Since 1976, the Carl B. Allendoerfer Prize has been awarded annually to recognize expository excellence in the MAGAZINE. In addition to these models of style, many useful references are available. Some are listed at the end of these guidelines.

Style and format We assume that our authors are at least sometime-readers of the MAGAZINE, with some knowledge of its traditions. If so, they know that most papers are published either as Articles or as Notes. Articles have a broader scope than Notes and usually run longer than 2000 words. Notes are typically shorter and more narrowly focused. Articles should be divided into a few sections, each with a carefully chosen title. Notes, being shorter, usually need less formal sectioning. Footnotes and subsectioning are almost never used in the MAGAZINE .

In addition to expository pieces, we accept a limited number of Math Bites, poems, cartoons, Proofs Without Words, and other miscellanea.

List references either alphabetically or in the order cited in the text, adhering closely to the MAGAZINE'S style for capitalization, use of italics, etc.

We recommend using simple, unadorned LaTeX in the preparation of your manuscript. Whatever technology you use, try to follow MAGAZINE style in small matters, but space your manuscript generously and leave large margins for the benefit of reviewers. Include the title and all authors' names, addresses, and email addresses at the top of the first page. Templates with further stylistic details are posted at our website in a variety of formats. Number the pages. Whether LaTeX is used or not, we hope to receive some electronic version of every article accepted.

Authors who prefer "blind refereeing" may omit author information from their manuscripts. (Make sure you include it in a covering message!) In this case we will avoid revealing the authors' identity to reviewers. We cannot promise absolute security in this regard.

For initial submission, graphical material may be interspersed with text. Each figure should be numbered, and referenced by number in the text. Authors themselves are responsible for providing images of suitable quality. If a piece is to appear in the MAGAZINE, separate copies of all illustrations may be needed, both with and without added lettering. We hope authors will be able to provide electronic versions of all figures. EPS is the ideal format for line art. EPS, TIFF, or high-resolution JPEG are acceptable for photos and screen captures.

Submitting manuscripts We encourage electronic submissions. Please send new manuscripts my email directly to the editor at mathmag@maa.org. The same address is appropriate for inquiries and general correspondence. A brief message containing contact information with an attached PDF file is best. Word-processor or DVI files can also be considered. Alternatively, manuscripts may be mailed to Mathematics Magazine, 132 Bodine Road, Berwyn, PA 19312-1027. If possible, please include an email address for further correspondence.

Suggested Reading

1. R. P. Boas, Can we make mathematics intelligible? *Amer. Math. Monthly* **88** (1981) 727–731.
2. Paul Halmos, How to write mathematics, *Enseign. Math.* **16** (1970) 123–152. Reprinted in Paul Halmos, *Selecta, Expository Writings*, Vol. 2, Springer, New York, 1983, 157–186.
3. Andrew Hwang, Writing in the age of LaTeX, *AMS Notices* **42** (1995) 878–882.
4. D.E. Knuth, T. Larrabee, and P. M. Roberts, *Mathematical Writing*, MAA Notes #14, 1989.
5. Steven G. Krantz, *A Primer of Mathematical Writing*, American Mathematical Society, Providence, RI, 1997.
6. N. David Mermin, *Boojums All the Way Through*, Cambridge Univ. Press, Cambridge, UK, 1990.

70th Annual William Lowell Putnam Mathematical Competition

Editor's Note: Additional solutions will be printed in the *Monthly* later in the year.

PROBLEMS

A1. Let f be a real-valued function on the plane such that for every square $ABCD$ in the plane, $f(A) + f(B) + f(C) + f(D) = 0$. Does it follow that $f(P) = 0$ for all points P in the plane?

A2. Functions f, g, h are differentiable on some open interval around 0 and satisfy the equations and initial conditions

$$f' = 2f^2gh + \frac{1}{gh}, \qquad f(0) = 1,$$

$$g' = fg^2h + \frac{4}{fh}, \qquad g(0) = 1,$$

$$h' = 3fgh^2 + \frac{1}{fg}, \qquad h(0) = 1.$$

Find an explicit formula for $f(x)$, valid in some open interval around 0.

A3. Let d_n be the determinant of the $n \times n$ matrix whose entries, from left to right and then from top to bottom, are $\cos 1, \cos 2, \ldots, \cos n^2$. (For example, $d_3 = \begin{vmatrix} \cos 1 & \cos 2 & \cos 3 \\ \cos 4 & \cos 5 & \cos 6 \\ \cos 7 & \cos 8 & \cos 9 \end{vmatrix}$. The argument of cos is always in radians, not degrees.) Evaluate $\lim_{n\to\infty} d_n$.

A4. Let S be a set of rational numbers such that

(a) $0 \in S$;

(b) If $x \in S$ then $x + 1 \in S$ and $x - 1 \in S$; and

(c) If $x \in S$ and $x \notin \{0, 1\}$, then $\frac{1}{x(x-1)} \in S$.

Must S contain all rational numbers?

A5. Is there a finite abelian group G such that the product of the orders of all its elements is 2^{2009}?

A6. Let $f\colon [0, 1]^2 \to \mathbb{R}$ be a continuous function on the closed unit square such that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist and are continuous on the interior $(0, 1)^2$. Let $a = \int_0^1 f(0, y)\,dy$, $b = \int_0^1 f(1, y)\,dy$, $c = \int_0^1 f(x, 0)\,dx$, and $d = \int_0^1 f(x, 1)\,dx$. Prove or disprove: There must be a point (x_0, y_0) in $(0, 1)^2$ such that

$$\frac{\partial f}{\partial x}(x_0, y_0) = b - a \qquad \text{and} \qquad \frac{\partial f}{\partial y}(x_0, y_0) = d - c.$$

B1. Show that every positive rational number can be written as a quotient of products of factorials of (not necessarily distinct) primes. For example, $\frac{10}{9} = \frac{2! \cdot 5!}{3! \cdot 3! \cdot 3!}$.

Math. Mag. **83** (2010) 75–80. doi:10.4169/002557010X485878.

B2. A game involves jumping to the right on the real number line. If a and b are real numbers and $b > a$, the cost of jumping from a to b is $b^3 - ab^2$. For what real numbers c can one travel from 0 to 1 in a finite number of jumps with total cost exactly c?

B3. Call a subset S of $\{1, 2, \ldots, n\}$ *mediocre* if it has the following property: Whenever a and b are elements of S whose average is an integer, that average is also an element of S. Let $A(n)$ be the number of mediocre subsets of $\{1, 2, \ldots, n\}$. [For instance, every subset of $\{1, 2, 3\}$ except $\{1, 3\}$ is mediocre, so $A(3) = 7$.] Find all positive integers n such that $A(n + 2) - 2A(n + 1) + A(n) = 1$.

B4. Say that a polynomial with real coefficients in two variables x, y is *balanced* if the average value of the polynomial on each circle centered at the origin is 0. The balanced polynomials of degree at most 2009 form a vector space V over $\mathbb{R}$. Find the dimension of V.

B5. Let $f : (1, \infty) \to \mathbb{R}$ be a differentiable function such that

$$f'(x) = \frac{x^2 - f(x)^2}{x^2(f(x)^2 + 1)} \qquad \text{for all } x > 1.$$

Prove that $\lim_{x\to\infty} f(x) = \infty$.

B6. Prove that for every positive integer n, there is a sequence of integers $a_0, a_1, \ldots, a_{2009}$ with $a_0 = 0$ and $a_{2009} = n$ such that each term after a_0 is either an earlier term plus 2^k for some nonnegative integer k, or of the form $b \bmod c$ for some earlier positive terms b and c. [Here $b \bmod c$ denotes the remainder when b is divided by c, so $0 \le (b \bmod c) < c$.]

SOLUTIONS

Solution to A1. Yes. Given a point P, choose a square $ABCD$ with center P. The midpoints of the sides form another square $A'B'C'D'$. Segments $A'C'$ and $B'D'$ divide $ABCD$ into four smaller squares. Summing the relation for these smaller squares, and subtracting the relation for $ABCD$ and twice the relation for $A'B'C'D'$, yields $4f(P) = 0$, so $f(P) = 0$.

Solution to A2. First note that

$$\begin{aligned}(fgh)' &= f'gh + fg'h + fgh' \\ &= \left(2f^2gh + \frac{1}{gh}\right)gh + \left(fg^2h + \frac{4}{fh}\right)fh + \left(3fgh^2 + \frac{1}{fg}\right)fg \\ &= 6f^2g^2h^2 + 6.\end{aligned}$$

Therefore, $F(x) = f(x)g(x)h(x)$ satisfies the separable equation $F' = 6(F^2 + 1)$. Solving this and using the initial condition $F(0) = 1$, we get $F(x) = \tan(6x + \pi/4)$.

Meanwhile, the first of the given differential equations can be written as

$$f' = 2fF + \frac{f}{F}$$

which is also separable, and we get

$$\frac{f'(x)}{f(x)} = 2F(x) + \frac{1}{F(x)} = 2\,\frac{\sin(6x + \pi/4)}{\cos(6x + \pi/4)} + \frac{\cos(6x + \pi/4)}{\sin(6x + \pi/4)},$$

which integrates to

$$\log f(x) = -\frac{1}{3}\log\cos(6x+\pi/4) + \frac{1}{6}\log\sin(6x+\pi/4) + C$$

on an interval around $x = 0$ where all the quantities whose logarithms are involved are positive. From $f(0) = 1$ we get $C = \frac{1}{6}\log\frac{\sqrt{2}}{2} = \log\bigl(1/\sqrt[12]{2}\bigr)$. Therefore, the answer is

$$f(x) = \frac{1}{\sqrt[12]{2}}\frac{\sqrt[6]{\sin(6x+\pi/4)}}{\sqrt[3]{\cos(6x+\pi/4)}}.$$

Solution to A3. Observe that

$$\frac{1}{2}\bigl(\cos k + \cos(2n+k)\bigr) = \cos n \cdot \cos(n+k),$$

which shows that for $n \geq 3$, the average of the first and third rows of the matrix is a nonzero multiple of the second row. Thus, $d_n = 0$ for $n \geq 3$, so $\lim_{n\to\infty} d_n = 0$.

Solution to A4. No. Experimentation shows that $\{m \pm 1/n : m \in \mathbb{Z}, n \in \mathbb{Z}_{>0}\} \subseteq S$, but suggests that perhaps the simplest numbers not of this form, such as $2/5$, need not be in S. So we try letting S be the set of rational numbers that are not of the form $m + 2/5$ with $m \in \mathbb{Z}$. Then S trivially satisfies the first two conditions. Now suppose that $x = a/b \in \mathbb{Q}$ where $a, b \in \mathbb{Z}$ satisfy $\gcd(a, b) = 1$ and $b > 0$, and suppose that $\frac{1}{x(x-1)} = m + 2/5$ for some $m \in \mathbb{Z}$. Then $\frac{b^2}{a(a-b)} = m + \frac{2}{5}$, but $a > a - b$, so $(a, a-b)$ must be one of $(5, 1), (5, -1), (1, -5), (-1, -5)$. So $x \in \{5/4,\ 5/6,\ 1/6,\ -1/4\}$ and $\frac{1}{x(x-1)}$ is not of the form $m + 2/5$. Thus S is a counterexample.

Solution to A5. No. If $|G|$ is divisible by an odd prime p, then $|G|$ has an element of order p, but this contradicts the product of the orders being 2^{2009}, so $G \cong \mathbb{Z}/2^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/2^{e_k}\mathbb{Z}$ for some integers $e_i > 0$. Let m_r be the number of cyclic order-r subgroups of G. Grouping elements by the subgroup they generate shows that the number of elements of order r in G is $\phi(r)m_r$. Thus

$$\begin{aligned} 2^{2009} &= 2^{m_2}4^{2m_4}8^{4m_8}16^{8m_{16}}\cdots \\ 2009 &= m_2 + 2\cdot 2m_4 + 3\cdot 4m_8 + 4\cdot 8m_{16} + \cdots \end{aligned} \tag{1}$$

In particular, $m_2 \equiv 1 \pmod 4$. But $m_2 = 2^k - 1$, so $k = 1$. Thus G is cyclic! So the sequence $m_2, m_4, m_8, \ldots$ has the form $1, 1, \ldots, 1, 0, 0, 0, \ldots$. To reach a sum of 2009 in (1), we must certainly have $m_2 = m_4 = m_8 = 1$. But then (1) yields the contradiction

$$2009 = 1 + 2\cdot 2 + 3\cdot 4 \pmod{16}.$$

Solution to A6. The statement is false. Let $g(x)$ be a C^1 function on $[0, 1]$ such that $g(x) \equiv 0$ on $[0, 1/4]$ and $[3/4, 1]$, $g'(x) > 0$ on $(1/4, 1/2)$ and $g'(x) < 0$ on $(1/2, 3/4)$. Let $h(x) = \frac{1}{2}g\bigl(\frac{x}{2} + \frac{1}{4}\bigr)$. Then $\int_0^1 g(x)\,dx = \int_0^1 h(x)\,dx$, $h'(x) > 0$ on $(0, 1/2)$ and $h'(x) < 0$ on $(1/2, 1)$, and $h(0) = h(1) = 0$. Let

$$f(x, y) = (1-y)g(x) + yh(x).$$

Then $a = b = 0$ and $c = d = \int_0^1 g(x)\,dx$. If the claim were true, there would be a point $(x, y) \in (0, 1)^2$ where $\partial_x f = \partial_y f = 0$. But $\partial_y f(x, y) = -g(x) + h(x)$, so we would have $g(x) = h(x)$. This can only occur on one of the intervals $(1/4, 1/2)$ and $(1/2, 3/4)$. But if $x \in (1/4, 1/2)$, then $\partial_x f(x, y) = (1-y)g'(x) + yh'(x) > 0$. Similarly, if $x \in (1/2, 3/4)$, then $\partial_x f < 0$.

Solution to B1. It's enough to show that any prime can be written in this form, and this can be done by strong induction. First, $1 = 2!/2!$, $2 = 2!$, and if all primes smaller than a given prime p have expressions of the desired form, we can write $p = \frac{p!}{(p-1)!}$, factor the denominator into primes (which are less than p), and use the induction hypothesis.

Solution to B2. Answer: Total cost c is possible if and only if $1/3 < c \leq 1$.

Since $b^3 - ab^2 = b^2(b-a)$, the total cost is an upper Riemann sum for the strictly increasing function x^2; that is, the area of a union of rectangles contained in $[0, 1]^2$ and containing the area under the graph of $y = x^2$ for $x \in [0, 1]$. Therefore $\int_0^1 x^2 dx < c \leq 1$. Thus $1/3 < c \leq 1$ is necessary.

Suppose that c satisfies $1/3 < c \leq 1$. We can find a subdivision $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$ whose upper Riemann sum r is less than c. For $0 \leq t \leq 1$, let $f(t)$ be the upper Riemann sum associated with the subdivision $0 \leq tx_1 \leq tx_2 \leq \cdots \leq tx_n \leq 1$. Then f is continuous on $[0, 1]$ with $f(0) = 1$ and $f(1) = r$, so by the Intermediate Value Theorem, there exists $t \in [0, 1]$ with $f(t) = c$. Jumping from 0 to tx_1 to $\ldots$ to tx_n to 1 (and ignoring jumps of length 0) yields a path with total cost c.

Solution to B3. Answer: $\{1, 3, 7, 15, \ldots\}$, the set of powers of 2 minus 1.

Fix n. Let $B(n+2)$ be the number of mediocre subsets of $\{1, 2, 3, \ldots, n+2\}$ that include both 1 and $n+2$. Let's count all the mediocre subsets of $\{1, 2, 3, \ldots, n+2\}$: There are $B(n+2)$, plus the number of mediocre subsets of $\{1, 2, 3, \ldots, n+1\}$ (there are $A(n+1)$ of these), plus the number of mediocre subsets of $\{2, \ldots, n+2\}$ (these are the same sets shifted by 1, so there are $A(n+1)$ of these, too), minus the number of mediocre subsets of $\{2, 3, \ldots, n+1\}$ (there are $A(n)$ of these, and we must subtract them because we have added them twice). That is,

$$A(n+2) = B(n+2) + 2A(n+1) - A(n),$$

or equivalently,

$$A(n+2) - 2A(n+1) + A(n) = B(n+2).$$

The problem is asking when $B(n+2) = 1$.

If $n+1$ has an odd factor $f > 3$ then $\{1, 2, 3, \ldots, n+2\}$ and $\{1, 1+f, 1+2f, \ldots, n+2\}$ are two mediocre subsets that include 1 and $n+2$, so $B(n+2) \geq 2$. On the other hand, if $n+1$ is a power of 2 is is easy to see that $\{1, 2, 3, \ldots, n+2\}$ is the only such subset, so $B(n+2) = 1$ exactly when $n+1$ is a power of 2.

Solution to B4. For $0 \leq i \leq 2009$, let H_i be the space of homogeneous polynomials of degree i. It is spanned by $x^i, x^{i-1}y, \ldots, y^i$, so $\dim H_i = i + 1$. A polynomial $P = \sum_{i=0}^{2009} P_i$ with $P_i \in H_i$ is balanced if and only if for each $r > 0$, any of the following equivalent conditions holds.

$$\int_0^{2\pi} P(r\cos\theta, r\sin\theta)d\theta = 0,$$

$$\int_0^{2\pi} \sum_{i=0}^{2009} P_i(r\cos\theta, r\sin\theta)d\theta = 0,$$

$$\sum_{i=0}^{2009} \left(\int_0^{2\pi} P_i(\cos\theta, \sin\theta)d\theta \right) r^i = 0.$$

Therefore, P is balanced if and only if

$$\int_0^{2\pi} P_i(\cos\theta, \sin\theta)d\theta = 0 \qquad \text{for all } i \text{ in the range } 0 \le i \le 2009.$$

Let V_i be the kernel of the linear transformation $\phi_i : H_i \to \mathbb{R}$ defined by

$$\phi_i(f) = \int_0^{2\pi} f(\cos\theta, \sin\theta)d\theta$$

for $f \in H_i$. If i is even, then $f(x, y) = (x^2 + y^2)^{i/2}$ has nonzero image, so $\dim V_i = \dim H_i - 1 = i$. If i is odd, then the average value of $\phi(x^j y^{i-j})$ on the unit circle is 0 by the symmetry $(x, y) \mapsto (-x, y)$ if j is odd, and $(x, y) \mapsto (x, -y)$ if j is even; thus ϕ is identically zero, so $\dim V_i = \dim H_i = i + 1$. Finally, $V = \bigoplus_{i=0}^{2009} V_i$, so

$$\dim V = \sum_{i=0}^{2009} \dim V_i = \sum_{i=0}^{2009}(i+1) - \sum_{\substack{0\le i\le 2009\\ i \text{ even}}} 1 = \frac{2010 \cdot 2011}{2} - \frac{2010}{2} = 2020050.$$

Solution to B5. We first prove that $f(x)$ has a limit (finite or infinite) at infinity. We have

$$f'(x) = \frac{1}{f(x)^2+1} - \frac{f(x)^2}{x^2(f(x)^2+1)} \ge \frac{1}{f(x)^2+1} - \frac{1}{x^2}$$

so that

$$0 \le f'(x) + \frac{1}{x^2} = \frac{d}{dx}\left(f(x) - \frac{1}{x}\right).$$

It follows that $f(x) - 1/x$ is increasing, therefore has a limit L (which may be infinity). But $1/x \to 0$ as $x \to \infty$, hence we also have $f(x) \to L$.

Suppose that L is finite. Then for some constant M we have $f(x) \le M$ for all $x \ge 2$, and

$$\begin{aligned} f(x) &= f(2) + \int_2^x \frac{t^2 - f(t)^2}{t^2(f(t)^2+1)}\,dt \\ &= f(2) + \int_2^x \frac{t^2}{t^2(f(t)^2+1)}\,dt - \int_2^x \frac{f(t)^2}{t^2(f(t)^2+1)}\,dt \\ &\ge f(2) + \int_2^x \frac{1}{M^2+1}\,dt - \int_2^\infty \frac{1}{t^2}\,dt \\ &\ge f(2) + \frac{x-2}{M^2+1} - \frac{1}{2}, \end{aligned}$$

which contradicts the assumption that f is bounded. Thus, $\lim_{x\to\infty} f(x) = \infty$.

Solution to B6. (based on a student paper) We begin with two lemmas.

LEMMA 1. *If we've obtained b and c, $b > c$, then we can get $b - c$ in three more steps.*

Choose an integer m such that $2^m > b$. Then extend the sequence with (i) $b + 2^m$, (ii) $c + 2^m$, (iii) $b - c = b + 2^m \pmod{2^m + c}$.

LEMMA 2. *If we've obtained $b > 0$, then for any positive integer k, we can obtain b^k is six more steps.*

Choose an integer m such that $2^m > 2b^k$. Then (i) $2^m = a_0 + 2^m$, (ii) $2^{km} = a_0 + 2^{km}$, (iii, iv, v) $2^m - b$ from Lemma 1, (vi) $b^k = 2^{km} \pmod{2^m - b}$.

So now we can use the following sequence. First choose m so that $2^m > 2n$. Then

$$\begin{aligned}
a_0 &= 0 \\
a_1 &= a_0 + 2^{2m} = 2^{2m} \\
a_2 &= a_1 + 2^0 = 2^{2m} + 1 \\
a_8 &= (2^{2m} + 1)^{n-1} \quad \text{(Lemma 2)} \\
a_9 &= a_0 + 2^{4m} = 2^{4m} \\
a_{10} &= (n-1)2^{2m} + 1 = (2^{2m} + 1)^{n-1} \pmod{2^{4m}} \\
a_{11} &= a_0 + 2^0 = 1 \\
a_{14} &= 2^{2m} - 1 \quad \text{(Lemma 1)} \\
a_{15} &= n = (n-1)2^{2m} + 1 \pmod{2^{2m} - 1} \\
a_i &= a_0 + a_{15} = n \quad \text{for } i \le 16 \le 2009.
\end{aligned}$$