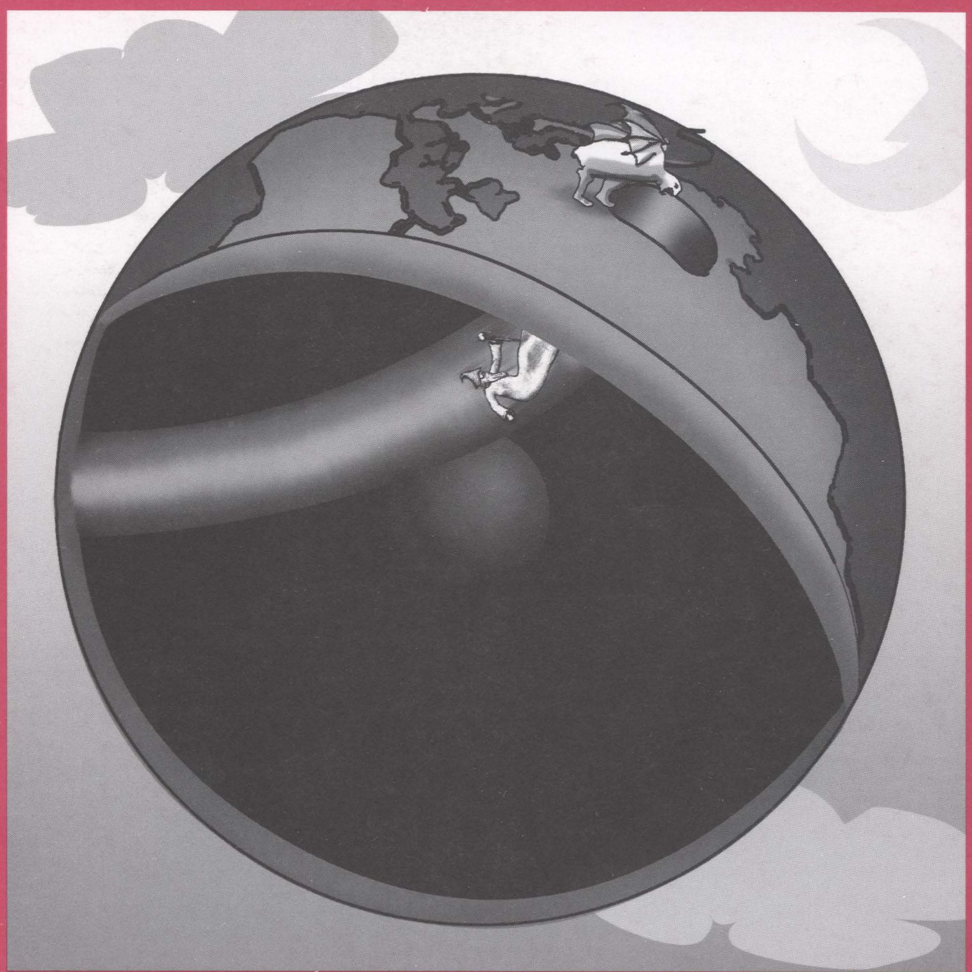
Vol. 77, No. 3, June 2004

MATHEMATICS MAGAZINE



Gandalf Falling through the Earth

- Falling down a Hole through the Earth
- Upper Bounds on the Sum of Principal Divisors of an Integer

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

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

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at 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.

Submit new manuscripts to Frank A. Farris, Editor, Mathematics Magazine, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0373. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

Cover image, *Gandalf Falling through the Earth*, by Lauren Gregory and Jason Challas. If Gandalf kept falling down the hole in the Mines of Moria without banging into the sides, his path would resemble the one illustrated on the cover. Gandalf would fall along a curved trajectory that misses the center of the Earth by a wide margin, and ultimately resurface somewhere on the other side. Such a fall gives a whole new meaning to the term *Middle Earth*.

Lauren Gregory is a Senior Marketing major at Santa Clara University, where Jason Challas teaches computer art with gravity.

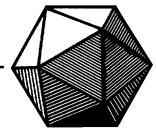
AUTHORS

Andrew J. Simoson is professor of mathematics at King College. When first seeing that his equations of motion outlandishly implied that a pebble dropped in northern climes falls further south than east away from the vertical, he subsequently laid the work aside, vowing to find the error some other day. Months later, the confusion cleared when he realized that for falling objects, the physicists' term *straight down* means the direction as given by a plumb bob, which at 45° latitude generates a tenth of a degree disparity with *truly down*. No wonder the world's affairs seem out of kilter at times.

Roger Eggleton and **William Galvin** are Australian mathematicians whose recent research collaboration produced several papers on inequalities, especially polynomial inequalities. Their paper in this MAGAZINE develops some number theoretic inequalities, proved somewhat surprisingly by methods that scarcely appeal to number theory. Eggleton's university teaching and research career, spanning four decades, has included more than a year in each of Australia, Brunei, Canada, Israel, and the U.S., with shorter professional visits to twice as many more countries. He has published some 60 research papers, mainly in graph theory, combinatorics and number theory. His Erdős number is 1 (several times).

Galvin's four-decade career began teaching high school mathematics, followed by many years in mathematics teacher-training in Australian tertiary-level institutions. Following early retirement, he culminated a career of distinguished mathematical service with a three-year stint as co-editor of the Australian Mathematical Society's *Gazette*. After a long battle with cancer, he passed away on December 12, 2003, barely two months after completing his editorial work. An obituary appears in *Austral. Math. Soc. Gaz.* **31** (2004), pp. 4–5.

Vol. 77, No. 3, June 2004



MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, 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 Frank Peterson (*FPetersonj@aol.com*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 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

Falling down a Hole through the Earth

ANDREW J. SIMOSON
King College
Bristol, Tennessee 37620
ajsimoso@king.edu

Drop a pebble into a hole in the Earth. Better yet, allow the pebble to drill its own hole as it falls. Along what path will it fall? How close to the center of the Earth will it approach? How fast will it do so? Keep in mind that the Earth rotates. Ignore all issues involving resistance.

In answering these questions, we first give some color and history behind this classic problem. Then we derive the requisite differential equation governing the pebble's motion when dropped from the surface of the rotating Earth. Assuming a linear gravitational field within the Earth, we generate analytic solutions, so demonstrating that the pebble follows an ellipse whose center is the center of the Earth. When the pebble is dropped at the Equator it misses the center by over 300 km. In doing so, the pebble moves in accordance with a familiar parametrization of the ellipse. Since the Earth is rotating while the pebble falls, the pebble's actual route through the Earth is another curve, whose shape we determine. We then solve the same problem using other gravitational fields for the Earth, and close with a few questions for further inquiry.

Some colorful history

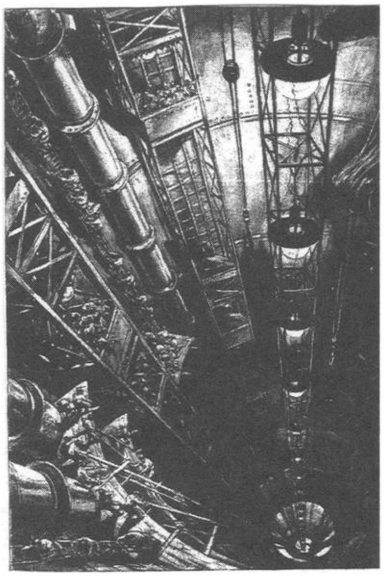
Deep holes elicit mystery. For who among us when encountering a well with shadowed bottom is not tempted to drop a pebble and wait, listening for a splash? Witness the romance of the deep hole in popular literature, such as Alice falling down a rabbit hole for ever so long that she grows sleepy, and thinks she must surely have long since fallen beyond the Earth's center [**3**, pp. 12–14], or the more recent Gandalf the Grey of *The Lord of the Rings* falling down an abyss that "none has measured," falling down unto "the uttermost foundations of stone" [**18**, p. 490].

In 1632, Galileo in his monumental defense of a rotating Earth model argued that a cannon ball dropped down a perforation in the Earth will follow a path of simple harmonic motion, oscillating forever between the drop site and its antipode [**8**, p. 227]. FIGURE 1a is the frontispiece from Galileo's book, showing his three debaters trying to convince each other about how objects move in space.

Replying to a letter from Robert Hooke informing him of a current theory suggesting that celestial storms in outer space kept the planets orbiting the Sun, Isaac Newton, in a letter of November 28, 1679, described an experiment for Hooke to present to the Royal Society using a falling ball to prove that the Earth rotates. Newton went on to speculate that should the ball fall down a hole at the Equator it would spiral around to the Earth's center. Hooke countered in a letter of January 6, 1680, that, in the absence of resistance, the trajectory should be a closed elliptical-like loop with the Earth's center at the center of the loop. Arnol'd [**1**, pp. 15–26] gives a more detailed account of this correspondence between Newton and Hooke.

Many popular mathematics and physics essays delight in discussing the problem of falling through the Earth. For example, in *The Strand Magazine* of 1909, a French astronomer advocates construction of a deep hole in England (not France!), using convicts and peace-time armies for labor; FIGURE 1b [**5**, p. 351], is a fanciful view of this

a. Discussing the hole



b. Working in the hole

Figure 1 Thinking and doing

hole as work is progressing. Other essays [**2**] lightheartedly suggest constructing transportation tunnels through the Earth with one proposed route connecting the antipodal points of Honolulu and the Kalahari Desert of Botswana. Furthermore, almost every differential equations text, such as Simmons [**14**, p. 24], or general physics text, such as Shortley [**13**, p. 258], contains at least one exercise involving a body falling from the surface to the center of the Earth, albeit these usually and explicitly describe the hole as running from pole to pole. Now let's see what occurs when this constraint is lifted.

Deriving a differential equation

We wish to drop a pebble P at the Equator of a rotating Earth whose density is spherically symmetric. The differential equations that govern its motion are derived by a slight modification of what can be found in many introductory calculus texts [**4**, pp. 827–830]. For ease of reference, we summarize the approach.

Introduce polar coordinates (r, θ) in a plane through the Equator with Earth's center at the origin so that the North Pole is above the plane. Two natural orthogonal unit vectors are the radial and angular vectors $\mathbf{u}_r = \cos(\theta)\,\mathbf{i} + \sin(\theta)\,\mathbf{j}$ and $\mathbf{u}_\theta = -\sin(\theta)\,\mathbf{i} + \cos(\theta)\,\mathbf{j}$.

Note that $d\mathbf{u}_r/d\theta = \mathbf{u}_\theta$ and $d\mathbf{u}_\theta/d\theta = -\mathbf{u}_r$. Since P's position can be written as $r\mathbf{u}_r$, then P's velocity $\mathbf{v}$ is

$$\mathbf{v} = \frac{d\,r\mathbf{u}_r}{dt} = r\mathbf{u}_\theta \frac{d\theta}{dt} + \mathbf{u}_r \frac{dr}{dt}.$$

By similar reasoning, P's acceleration $\mathbf{a}$ is given by

$$\mathbf{a} = \left(r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt}\right)\mathbf{u}_\theta + \left(\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2\right)\mathbf{u}_r. \tag{1}$$

Assume that pebble P's only acceleration is that of the gravitational field of the Earth, which is directed entirely in the $\mathbf{u}_r$ direction. Let $f(r)$ be the magnitude of this gravitational acceleration r units from the center. From (1), this assumption leads us to

$$0 = r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt} = \frac{1}{r}\frac{d(r^2\frac{d\theta}{dt})}{dt} \quad \text{and} \quad f(r) = \frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2. \tag{2}$$

The first of the equations in (2) is the *law of the conservation of angular momentum*; that is, since $1/r$ is never 0,

$$r^2\frac{d\theta}{dt} = h, \tag{3}$$

where h is a constant. The second equation in (2) is written in terms of h as

$$\boxed{f(r) = \frac{d^2r}{dt^2} - \frac{h^2}{r^3}.} \tag{4}$$

In the absence of friction and any other forces, the falling pebble's path is found by solving (4) along with the initial conditions,

$$\boxed{r = R \text{ when } t = 0 \text{ and } \theta = 0 \quad \text{and} \quad \frac{dr}{dt}(0) = 0 \text{ and } \frac{d\theta}{dt}(0) = \frac{2\pi}{Q},} \tag{5}$$

where R is the radius of the Earth and Q is the period of one revolution of the earth about its axis. Note that because the pebble is dropped at the Equator, the initial rotation rate is that of the Earth, $d\theta/dt(0) = 2\pi/Q$. Therefore the constant angular momentum, h, in (3) is given by

$$\boxed{h = \frac{2\pi R^2}{Q}.} \tag{6}$$

A useful trick to solve (4) for r in terms of θ is to let $z = 1/r$. Then

$$\frac{dr}{dt} = \frac{d\left(\frac{1}{z}\right)}{d\theta}\frac{d\theta}{dt} = -\frac{1}{z^2}\frac{dz}{d\theta}\frac{d\theta}{dt} = -r^2\frac{d\theta}{dt}\frac{dz}{d\theta} = -h\frac{dz}{d\theta}.$$

By similar reasoning,

$$\frac{d^2r}{dt^2} = -h^2z^2\frac{d^2z}{d\theta^2}.$$

Thus, an alternate version of (4) is

$$\boxed{\frac{d^2z}{d\theta^2} + z = -\frac{1}{h^2z^2}f\left(\frac{1}{z}\right).} \tag{7}$$

At this point, the reader may recall that when the force function f is the familiar inverse square law, namely, $f(r) = -k/r^2$, where k is some positive constant, then an orbiting pebble's path is described by Kepler's three laws, which we list for the sake of later contrast.

Kepler's three laws where $f(r) = -\dfrac{k}{r^2}$

i. Pebble P's path is an ellipse E, with one focus at the origin, the center of the Earth.
ii. P's position vector (from the origin) sweeps out area at a constant rate.
iii. The square of the period is proportional to the cube of E's semi-major axial length. (8)

The linear model

We now return to our main problem of interest and imagine a homogeneously dense Earth, a simple, natural model. From this assumption, Newton derived the corresponding gravitational force function culminating in *Corollary III* of *Proposition 91* of *Book I* of the *Principia*. We describe this briefly.

Think of the Earth as composed of concentric spheres. Newton showed that the net gravitational attraction of any sphere on any body located anywhere inside the sphere is zero and that the net gravitational attraction of any sphere on any body located anywhere outside the sphere is exactly the same as that of a point of identical mass located at the center of the sphere. So as a body passes through the Earth, the only part of the Earth that attracts the body consists of those spheres whose radii are less than or equal to the distance of the object. Thus the mass acting on the body is proportional to the cube of its distance from the center of the Earth. The force is proportional to the mass and inversely proportional to the square of the distance, and so the force $f(r)$ is directly proportional to the distance, giving $f(r) = -kr$, where k is some positive constant. (See [**6**, Chapter 6] or [**15**, pp. 336–339, 341], for more formal, self-contained arguments.)

With $f(r) = -kr$, (4) and (7) become

$$\boxed{\frac{d^2r}{dt^2} + kr = \frac{h^2}{r^3} \quad \text{and} \quad \frac{d^2z}{d\theta^2} + z = \frac{k}{h^2z^3}.} \tag{9}$$

Both of these nonlinear differential equations have the form

$$\frac{d^2w}{du^2} + \alpha^2 w = \frac{\beta^2}{w^3}, \tag{10}$$

where α and β are constants. Since we wish to solve (9) for the initial conditions, $r(0) = R = 1/z(0)$, $dr/dt(0) = 0$, and $dz/d\theta(0) = 0$, the initial conditions for (10) are $w(0) = \delta$ for some $\delta > 0$ and $dw/du(0) = 0$. To solve this equation, use the method of reduction of order, letting $p = dw/du$, which means that $d^2w/du^2 = dp/du = dp/dw \cdot dw/du = p\,dp/dw$. Hence (10) becomes

$$p\,dp = \left(\frac{\beta^2}{w^3} - \alpha^2 w\right) dw, \tag{11}$$

where $p = 0$ when $w = \delta$. Integrating (11) and using the boundary conditions and solving for p gives

$$\frac{dw}{du} = p = \pm\sqrt{\beta^2\left(\frac{1}{\delta^2} - \frac{1}{w^2}\right) - \alpha^2(w^2 - \delta^2)}.$$

Via the substitution $q = w^2 - \delta^2$, this becomes

$$\frac{\delta}{2\sqrt{\beta^2 q - \alpha^2\delta^2 q(q+\delta^2)}}\,dq = \pm du,$$

where $q = 0$ when $u = 0$. Integrate and use the initial condition, then rewrite in terms of w and u to find

$$\sin^{-1}\left(\frac{2\alpha^2\delta^2}{\beta^2 - \alpha^2\delta^4}\left(w^2 - \frac{\beta^2 + a^2\delta^4}{2\alpha^2\delta^2}\right)\right) = \pm 2\alpha u - \frac{\pi}{2},$$

which in turn simplifies to

$$w^2 = \frac{\beta^2 + \alpha^2\delta^4}{2\alpha^2\delta^2} - \frac{\beta^2 - \alpha^2\delta^4}{2\alpha^2\delta^2}\cos(2\alpha u).$$

Via the double angle formula for cosine and liberal use of the fundamental trigonometric identity and because w is positive near $u = 0$, this expression simplifies to

$$w = \sqrt{\delta^2\cos^2(\alpha u) + \left(\frac{\beta}{\alpha\delta}\right)^2 \sin^2(\alpha u)}. \tag{12}$$

This means that the solutions to (9) are, respectively,

$$\boxed{r(t) = \sqrt{R^2\cos^2(\sqrt{k}\,t) + \frac{h^2}{kR^2}\sin^2(\sqrt{k}\,t)},} \tag{13}$$

with $\delta = R$, $\alpha = \sqrt{k}$, and $\beta = h$, and

$$\boxed{r(\theta) = \frac{1}{\sqrt{\frac{1}{R^2}\cos^2\theta + \frac{kR^2}{h^2}\sin^2(\theta)}},} \tag{14}$$

with $\delta = 1/R$, $\alpha = 1$, $\beta = \sqrt{k}/h$, and $r(\theta) = 1/z(\theta)$.

The polar plot of (14) can be seen to be an ellipse of semi-axial lengths R and $h/(R\sqrt{k})$. To do so, let

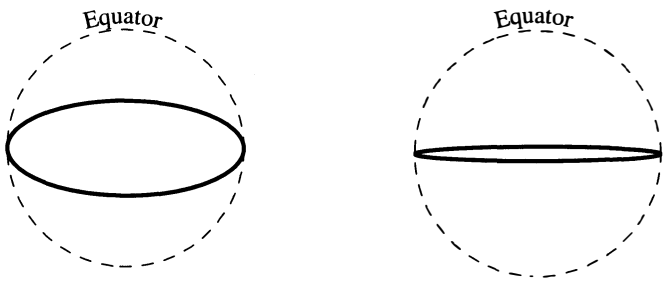
$$\big(X(\theta),\ Y(\theta)\big) = r(\theta)(\cos\theta,\ \sin\theta), \tag{15}$$

and observe that $X^2/R^2 + Y^2/(h/R\sqrt{k})^2 = 1$. For example, FIGURE 2a is the ellipse given by (15) when $h = 0.4R^2\sqrt{k}$.

If we assume that the tangential velocity of pebble P at its moment of being dropped is such that r begins to decrease, then it is clear that the nearest P gets to Earth's center is

$$\boxed{\frac{h}{R\sqrt{k}}.} \tag{16}$$

The ellipse of FIGURE 2b shows the pebble's path for our Earth, with $g = 9.8$ m/s^2, $R = 6400$ km, $k = g/R$, and $Q = 86400$ seconds. Thus the nearest that P approaches Earth's center as it falls is a bit more than 376 km.



a. A pebble's path b. The path with respect to our Earth

Figure 2 Elliptical solutions in the Equatorial plane of the Earth

Kepler analogs Let $A(t)$ be the area swept out by P's position vector from time 0 to time t. Via (3) the change in area ΔA per change in time Δt is given by

$$\Delta A = \int_t^{t+\Delta t} \frac{r^2}{2}\frac{d\theta}{d\tau}\,d\tau = \int_t^{t+\Delta t} \frac{h}{2}\,d\tau = \frac{h}{2}\Delta t, \tag{17}$$

where τ is a dummy variable. That is, P's position vector sweeps out area at a constant rate with respect to the ellipse's center rather than with respect to its focus. This argument is independent of the gravitational field $f(r)$, so Kepler's second law applies to any gravity field radiating from the origin.

A useful form of our solution (13) is the parametrization

$$\boxed{\big(x(t),\ y(t)\big) = \left(R\cos(\sqrt{k}\,t),\ \frac{h}{R\sqrt{k}}\sin(\sqrt{k}\,t)\right).} \tag{18}$$

To see that this parametrization is valid, first observe that both (15) and (18) parametrize the same ellipse. To demonstrate that they both trace out this ellipse in time in the same way, define the angle $\upsilon(t)$ and the length $s(t)$ so that $\tan(\upsilon) = y/x$ and $s(t) = \sqrt{x^2 + y^2}$. Therefore

$$\frac{s^2}{x^2}\frac{d\upsilon}{dt} = \sec^2\upsilon\frac{d\upsilon}{dt} = \frac{xy' - yx'}{x^2},$$

where $x' = dx/dt$ and $y' = dy/dt$. So

$$\begin{aligned} s^2\,d\upsilon &= (xy' - yx')\,dt \\ &= \left(R\cos(\sqrt{k}\,t)\frac{h}{R\sqrt{k}}\sqrt{k}\cos(\sqrt{k}\,t) - \frac{h}{R\sqrt{k}}\sin(\sqrt{k}\,t)(-R\sqrt{k})\sin(\sqrt{k}\,t)\right)dt \\ &= h\,dt, \end{aligned}$$

which means that the area swept out by x-y's position vector from t to $t + \Delta t$ is

$$\Delta\mathcal{A} = \frac{1}{2}\int_t^{t+\Delta t} s^2\frac{d\upsilon}{d\tau}\,d\tau = \frac{1}{2}\int_t^{t+\Delta t} h\,d\tau = \frac{h}{2}\Delta t, \tag{19}$$

where τ is a dummy variable for time. By (17) and (19), from time 0 to t, the parametrizations (15) and (18) both sweep out area $ht/2$; and since the parametrization (18) starts at the initial condition $(R, 0)$ of (5) and proceeds in the counterclockwise direction, then the arc lengths generated by these two parametrizations over this

time period are the same, which means that (18) faithfully describes the motion of the falling pebble.

Is there an analogous third Keplerian law governing the period? From (14), we see that the angle lapse between perigee and apogee is $\pi/2$, which by (13) means that $(\pi/2)/\sqrt{k}$ is one-fourth of the period; that is, the period T of the ellipse is

$$\boxed{T = \frac{2\pi}{\sqrt{k}}.} \tag{20}$$

For Earth, this period is about 84.6 minutes, which is the period of a pebble dropped at the North Pole through a homogeneously dense Earth. Table 11.1 in an essay, *How to get anywhere in about 42 minutes* in [**2**, p. 110], contains such period data for all the planets in our solar system. Observe that if a pebble is dropped initially at distance ρ from the origin in this gravitational field with initial rotation rate about the origin as $2\pi/Q$, then the pebble follows the path of an ellipse whose semi-major axis is ρ and whose semi-minor axis is

$$\frac{2\pi\rho^2}{\rho Q\sqrt{k}} = \frac{2\pi\rho}{Q\sqrt{k}}$$

by (6) and (16). Since the ratios of these semi-axial lengths is a constant, then this entire family of ellipses shares the same eccentricity. Note also that by (20), the period of orbits are independent of the initial rotation rate of the pebble about the origin.

To recap, the path of a pebble falling through the Earth has the following three properties, which the reader may contrast with Kepler's three planetary laws of motion in (8).

Kepler's three laws where $f(r) = -kr$ (21)

i. Pebble P's path is an ellipse whose center is the center of the Earth. This is *Corollary I* of *Proposition 10* of *Book I* of the *Principia*.

ii. P's position vector (from the origin) sweeps out area at a constant rate. This is *Proposition I* of the *Principia*, valid for any radial gravity field.

iii. The family of orbits pertaining to pebbles dropped in this field shares the same period. This is *Corollary II* of *Proposition 10* of the *Principia*.

If the pebble is dropped from the Earth's surface somewhere other than at the Equator the solution is generated as before, except for the constant of angular momentum. To indicate this new value, let $\hat{h}$ be the angular momentum of the pebble about the Earth's center when it is dropped at latitude ψ, where $-\pi/2 \le \psi \le \pi/2$. In FIGURE 3, the letters A, B, C, D, O, and P represent the points $(R, 0, 0)$, $(0, R, 0)$, the North Pole, the pebble's drop point $(R\cos\psi, 0, R\sin\psi)$, the center of the Earth, and the pebble's position at time t. The curve through A and B is the Equator; the curve through B and D is a great circle. Observe that the tangential speed v of a point at latitude ψ on the surface of the Earth about the Earth's axis is $2\pi R\cos\psi/Q$, where R is the radius of the Earth and Q is the Earth's period about its axis; this tangential speed v of the pebble is also $v = R\,d\theta/dt(0)$ where $\theta(t)$ is the angle between D and P. Hence $d\theta/dt(0) = 2\pi\cos\psi/Q$, which means that $\hat{h} = 2\pi R^2\cos\psi/Q = h\cos\psi$. The pebble's path thus is parametrized by $(R\cos(\sqrt{k}\,t), (h/R\sqrt{k})\cos\psi\sin(\sqrt{k}\,t))$, and lies in the plane through B, D, and O. In terms of the standard x-y-z coordinate system, the pebble's path is obtained

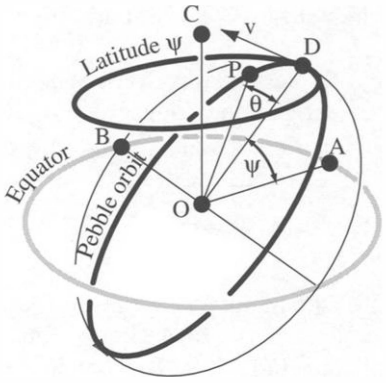


Figure 3 Pebble dropped at latitude ψ

by rotating the vectors $(R\cos(\sqrt{k}\,t),\ (h/R\sqrt{k})\cos\psi\sin(\sqrt{k}\,t),\ 0)$ clockwise (with respect to standing at B and looking at O) about the y-axis by ψ radians:

$$\boxed{\begin{aligned}&\big(x(t,\ \psi),\ y(t,\ \psi),\ z(t,\ \psi)\big)\\ &\quad=\left(R\cos\psi\cos(\sqrt{k}\,t),\ \frac{h}{R\sqrt{k}}\cos\psi\sin(\sqrt{k}\,t),\ R\sin\psi\cos(\sqrt{k}\,t)\right).\end{aligned}} \tag{22}$$

Observe that when $\psi = \pi/2$ the solution path degenerates to a straight line between the North and South Poles.

The path through the Earth

Although we have determined the path of a pebble P falling through the Earth with respect to a stationary plane, the Earth is rotating. We now change our orientation so that the frame of reference is the rotating Earth and ask, *What is the shape of a hole we must construct in the Earth so that a falling pebble never strikes the walls of the hole?* To answer this question, we find the relationship between θ and t; in particular, we write θ as a function in terms of t, and then graph

$$\boxed{r(t)\left(\cos\left(\theta(t)-\frac{2\pi t}{Q}\right),\ \sin\left(\theta(t)-\frac{2\pi t}{Q}\right)\right),} \tag{23}$$

where Q is the period of the Earth about its axis and $r(t)$ is (13). To derive a relation between θ and t, we start with (3) and (13),

$$\frac{d\theta}{dt} = \frac{h}{R^2\cos^2(\sqrt{k}t) + \frac{h^2}{kR^2}\sin^2(\sqrt{k}t)}, \tag{24}$$

and in the process give an alternate derivation of (18).

To simplify matters let

$$\sigma(t) = \frac{1}{a^2\cos^2(\omega t) + b^2\sin^2(\omega t)},$$

where a, b, ω are constants, and let $\Theta(t)$ be that function for which $d\Theta/dt = \sigma(t)$ with $\Theta(0) = 0$. Since $\Theta(t)$ is simply the area under the always positive, periodic function σ over the interval from 0 to t, $\Theta(t)$ increases without bound.

Observe that

$$\int \sigma(t)\,dt = \int \frac{1}{a^2\cos^2(\omega t) + b^2\sin^2(\omega t)}\,dt = \frac{1}{ab\omega}\tan^{-1}\left(\frac{b}{a}\tan(\omega t)\right) + C, \quad (25)$$

where C is a constant of integration. A disappointing feature of (25) is that the right-hand side is bounded above by $\pi/(2ab\omega)$, making it a poor match with $\Theta(t)$. The remedy is found in $\sigma(t)$'s periodicity. Clearly

$$\Theta(t) = \frac{1}{ab\omega}\tan^{-1}\left(\frac{b}{a}\tan(\omega t)\right), \text{ for } 0 \le t \le \frac{\pi}{2\omega}, \quad \text{and} \quad \Theta\left(\frac{\pi}{2\omega}\right) = \frac{\pi}{2ab\omega}.$$

From FIGURE 4a, for any t with $\pi/(2\omega) < t < \pi/\omega$ note that the area under the curve from $\pi/(2\omega)$ to t is equal to the area from $\pi/\omega - t$ to $\pi/(2\omega)$. Therefore the area from 0 to t is the area from 0 to π/ω minus the area from 0 to $\pi/\omega - t$, which gives

$$\Theta(t) = \frac{\pi}{ab\omega} - \Theta\left(\frac{\pi}{\omega} - t\right), \quad \text{for} \quad \frac{\pi}{2\omega} < t \le \frac{\pi}{\omega}.$$

For $t > \pi/\omega$, let $n(t) = \lfloor \omega t/\pi \rfloor$ and let $q(t) = t - n(t)\pi/\omega$, which are the integral number of times that π/ω divides t and the remainder, respectively. Then

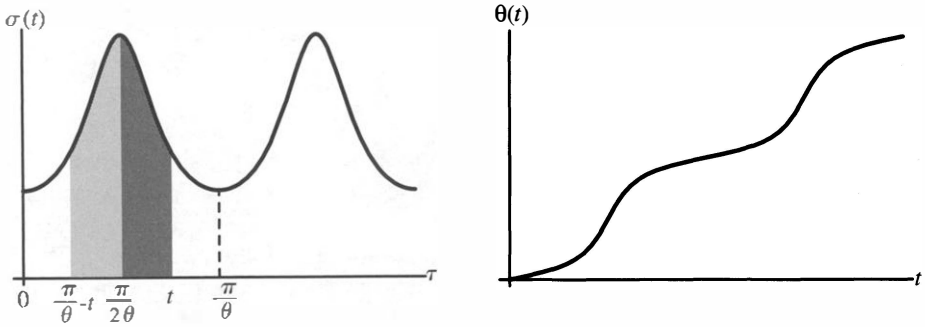
$$\Theta(t) = \frac{n(t)\pi}{ab\omega} + \Theta\big(q(t)\big), \quad \text{for} \quad t > \frac{\pi}{\omega}.$$

Thus $\theta(t) = h\Theta(t)$ for all t, with $a = R$, $b = h/(R\sqrt{k})$ and $\omega = \sqrt{k}$. In particular,

$$\boxed{\theta(t) = \tan^{-1}\left(\frac{h}{R^2\sqrt{k}}\tan(\sqrt{k}t)\right), \text{ for } 0 < t < \frac{\pi}{2\sqrt{k}}.} \quad (26)$$

Note that (26) means that $\tan\theta = h/(R^2\sqrt{k})\tan(\sqrt{k}\,t) = y/x$, corroborating (18). The graph of $\theta(t)$ is shown in FIGURE 4b. Observe that the graph looks like a line modified by a bounded, periodic function. To see whether this is more than an appearance, consider the expression $\tan\alpha = c\tan\beta$ for values α, β and c. Let $\chi = \alpha - \beta$. By the addition identity for the tangent function,

$$\frac{\tan\beta + \tan\chi}{1 - \tan\beta\tan\chi} = \tan(\beta + \chi) = c\tan\beta.$$



a. $\sigma(t)$ where $0 < t < \frac{\pi}{\theta}$ b. θ versus t

Figure 4 The relation between θ and t

Solving for $\tan\chi$ gives

$$\tan\chi = \frac{(c-1)\sin\beta\cos\beta}{1+(c-1)\sin^2\beta},$$

which means that

$$\alpha = \beta + \tan^{-1}\left(\frac{(c-1)\sin\beta\cos\beta}{1+(c-1)\sin^2\beta}\right). \tag{27}$$

Substituting $\alpha = \theta$, $\beta = \sqrt{k}\,t$, and $c = h/(R^2\sqrt{k})$ in (27) gives a nice improvement over (26) for $\theta(t)$:

$$\boxed{\theta(t) = \sqrt{k}\,t + \tan^{-1}\left(\frac{(\frac{h}{R^2\sqrt{k}}-1)\sin(\sqrt{k}\,t)\cos(\sqrt{k}\,t)}{1+(\frac{h}{R^2\sqrt{k}}-1)\sin^2(\sqrt{k}\,t)}\right) \text{ for } t \geq 0.} \tag{28}$$

Note that the inverse tangent term of (27) is bounded and periodic because its argument is periodic, continuous and has positive denominator. FIGURE 5 shows the path that the pebble drills through the Earth (neglecting friction), as given by (23).

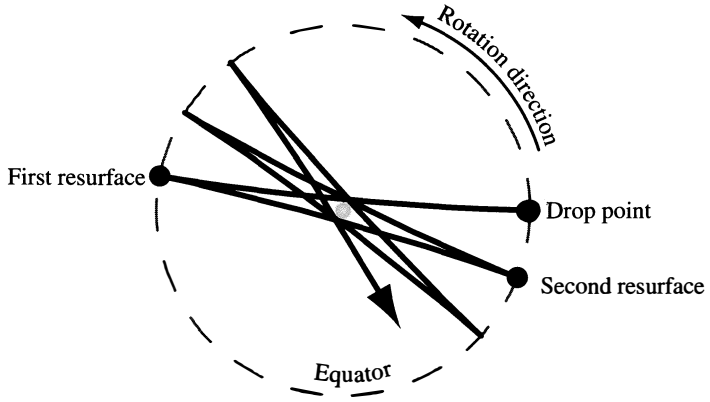


Figure 5 The hole through the Earth

For Earth, a pebble dropped at the Equator will return to its drop point with respect to a nonrotating coordinate system in 84.6 minutes. Because the Earth rotates, the geographical point where the pebble returns is actually about 2360 km due west of the geographical drop point, as illustrated in FIGURE 5.

Measuring the rotation In a lengthy dialogue [**8**, p. 139 etc.], Galileo examined the experiment of dropping a cannon ball from a lofty tower. Data of that day confirmed that the ball fell straight down, seeming evidence that the Earth fails to rotate. To illustrate, let's use the leaning Tower of Pisa in northern Italy, 55 meters tall, atop which Galileo is said to have performed his legendary experiment of dropping cannon balls of disparate masses, confuting Aristotle's premise that more massive objects fall faster than less massive objects [**16**, pp. 19–20]. The Tower in FIGURE 6 is from a painting owned by Gary Feuerstein, who kindly permitted its use, and which can be seen online [**7**].

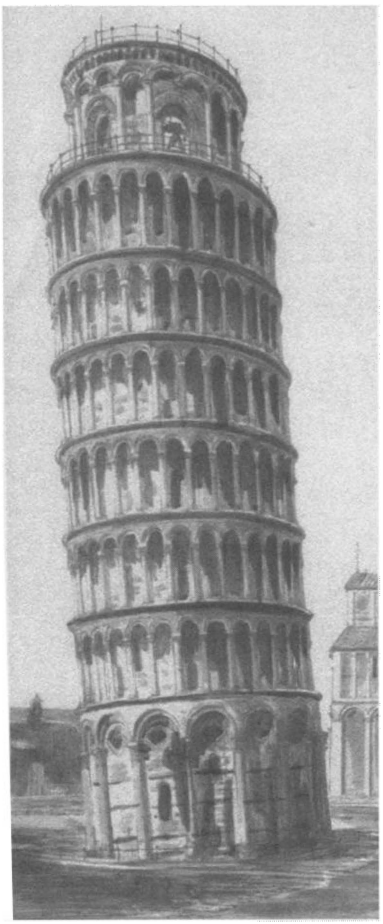


Figure 6 The Tower at Pisa

Let's drop a ball from 50 meters; its flight time (which Galileo could calculate) is about 3.2 seconds. From such a height, and at Pisa's latitude of $\psi = 44°$, a Ptolemaic advocate might reason, *If the Earth rotates once per day, then the ball should fall west of the Tower by about* $(6400\ km)(\cos\psi)(2\pi/Q)(3\ seconds) \approx 1\ km$*; but since no westward displacement occurs, then the Earth must be fixed in space.*

Against such logic, a Copernican advocate would reason, *Just as a ball dropped from atop a tall mast on a moving ship falls straight down, so too on the Earth* (a rationale somewhat anticipating the notion of a Newtonian reference frame). But a ball dropped from the Tower actually falls eastward, the reason being that the tangential velocity of the top of the Tower on a rotating Earth is greater than the tangential velocity at its base, so as the ball falls it continues to go eastward slightly faster than the base of the tower, and thus strikes the ground a bit to the east of straight down. FIGURE 5 illustrates this phenomenon.

In particular, if a pebble falls from the surface at the Equator, then at time t the pebble is at $r(t)$ $(\cos(\theta(t) - 2\pi t/Q), \sin(\theta(t) - 2\pi t/Q))$ with respect to the rotating Equatorial plane. Measured along a circumference of a circle of radius $r(t)$ and center (0, 0), the *eastward deflection* of the pebble away from a vertical line between the Earth's center and the geographical drop point is $r(t)(\theta(t) - 2\pi t/Q)$. By (13) and either (26) or (28), this deflection $E(t)$ is

$$\boxed{E(t) = \left(\tan^{-1}\left(\frac{h}{R^2\sqrt{k}}\tan(\sqrt{k}t)\right) - \frac{2\pi}{Q}t\right)\sqrt{R^2\cos^2(\sqrt{k}t) + \frac{h^2}{kR^2}\sin^2(\sqrt{k}t)},} \tag{29}$$

for $0 < t < \pi(2\sqrt{k})$. The reader may show via (22) that the eastward deflection $E(t, \psi)$ of the falling pebble when dropped at latitude ψ is

$$\boxed{E(t, \psi) = E(t)\cos\psi.} \tag{30}$$

Assuming no air resistance, a little calculation using (29) shows that a drop down a hole of 50 meters at the Equator results in the pebble landing about 8 mm east of straight down. The same result holds if the pebble is dropped from a tower of 50 meters rather than down a hole. At latitude 44°, the same drop gives a deflection of about 6 mm via (30), not much of an eastward deflection for Galileo's cannon ball.

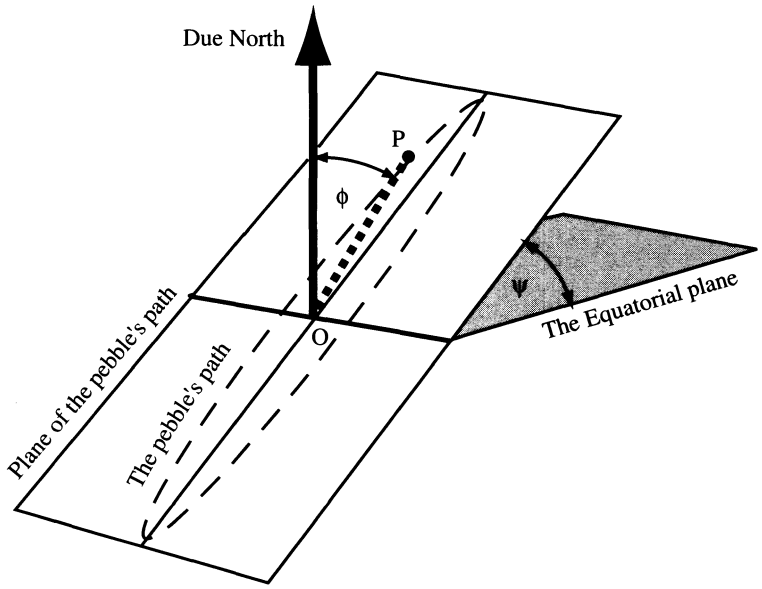


Figure 7 Finding the southward deflection

A pebble dropped at the surface of the Earth away from the Equator falls toward the Equator as well as to the east. To verify this, let's drop a pebble at latitude ψ in the Northern Hemisphere so that the drop point is $(R\cos\psi,\ 0,\ R\sin\psi)$. Let $\phi(t, \psi)$ be the angle between due North and the pebble's position at time t, point P on FIGURE 7. From (22), the pebble's distance from the center of the Earth is

$$r(t, \psi) = \sqrt{R^2\cos^2(\sqrt{k}\,t) + \frac{h^2}{kR^2}\cos^2\psi\sin^2(\sqrt{k}\,t)}. \tag{31}$$

Thus by (22) and (31),

$$\phi(t, \psi) = \cos^{-1}\left(\frac{z(t, \psi)}{r(t, \psi)}\right) = \cos^{-1}\left(\frac{R\sin\psi\cos(\sqrt{k}\,t)}{\sqrt{R^2\cos^2(\sqrt{k}\,t) + \frac{h^2}{kR^2}\cos^2\psi\sin^2(\sqrt{k}\,t)}}\right) \tag{32}$$

While the pebble falls, the point we dropped it from remains at latitude ψ. Measured along the surface of a sphere centered at the origin and of radius $r(t, \psi)$, the pebble's *southward deflection* away from this sphere's circle of latitude ψ is

$$\boxed{S(t, \psi) = \left(\phi(t, \psi) - \left(\frac{\pi}{2} - \psi\right)\right) r(t, \psi).} \tag{33}$$

Evaluating (33) for Galileo's ball, we find that the ball should fall about 8.7 cm south of straight down. Galileo was ironically close to having conclusive evidence that the Earth rotates! However, measuring this deflection is challenging even if one knows to look for it. Although Galileo did not look for eastward or southward deflections, others soon did.

In December of 1679, following Newton's suggestion, Hooke, who as Curator of the Royal Society regularly performed experiments for its weekly meetings, repeated the experiment, this time in a cathedral with its windows and doors closed so as to minimize air currents. Dropping a ball from 9 meters high, Hooke measured the impact as at least a quarter of an inch southeast of what he thought was straight down [**1**, pp. 20–21]. Equations (30) and (33) show that for a fall time of about 1.35 seconds for a fall of 9 meters at latitude 52° in London, the ball's eastward deflection should be about 0.36 mm, whereas the ball's southward deflection should be about 1.5 cm $\approx$ 0.59 inches, values that, at first glance, seem to confirm Hooke's result.

In 1831, F. Reich dropped pellets 188 meters down a mine shaft, measuring the deflections as about 2.8 cm southeast of what he thought was straight down [**10**, p. 395]. If his mine shaft was at latitude 45° with a drop time of 6.2 seconds, (30) and (33) predict an eastward deflection of 4 cm and a southward deflection of 32 cm, predictions which seem far afield, especially the southward deflection. What's going on?

To resolve the disparity between these experiments and the corresponding theoretical southern deflections, define *straight down* as the ray from the drop site to the center of the Earth, and define *apparent down* as the ray along a 10 meter long plumb bob whose upper end is held at the drop site. Except at the poles and at the Equator, straight down and apparent down are not the same direction. At 45°, the discrepancy between them is about 0.1°. An exercise in [**10**, p. 403, exercise 14] shows that for objects dropped 200 meters (no resistance) at the Earth's surface, the southward deflection of the object away from apparent down is about 0.01 mm. So southerly deflections (using a plumb bob) for short drops are basically unmeasurable. As the technology in global positioning satellites improves, our common notion of down may eventually change to mean straight down; if it does, then finding the southerly displacement of a dropped object will be a fun, surprising introductory physics experiment.

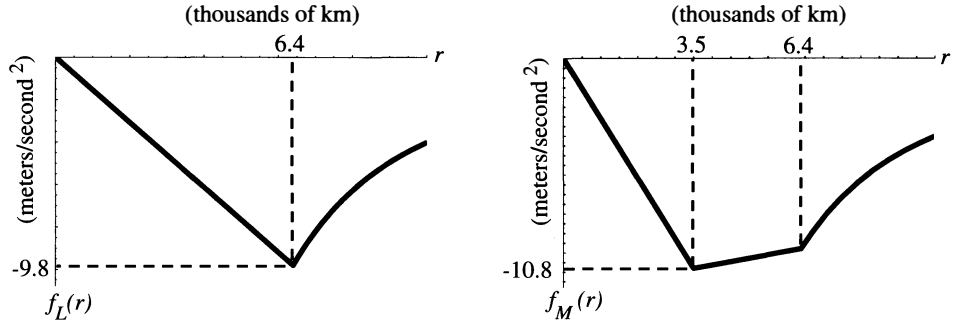
Someday, someone might try the experiment on the Moon, exulting in the turbulence free conditions. In anticipation of this event, here are the vital statistics for the Moon: $g = 1.57$ m/s^2, $R = 1741$ km, $f(r) = -kr$, $k = g/R$, $Q = 2.36$ million seconds. At latitude 45°, dropping a pebble down a deep lunar crevasse of 1 km gives a fall time of 36 seconds, a 4.5 cm eastward deflection and a 4 mm southward deflection from straight down. That is, plumb bobs on the Moon point almost straight down.

More questions

The standard linear gravitational acceleration field strength function

$$f_L(r) = \begin{cases} -\dfrac{g}{R} r, & 0 \le r \le R, \\ -\dfrac{gR^2}{r^2}, & r > R, \end{cases}$$

as depicted in FIGURE 8a is but one of various models for the Earth's gravity, where r is distance from the center of the Earth, $g = 9.8$ m/s^2, and $R = 6400$ km, the radius of the Earth. According to recent geophysical research, the Earth's gravity at the juncture between its outer core and its mantle, 3500 km from the center, is about -10.8 m/s^2. FIGURE 8b is a fairly accurate approximation f_M of Earth's actual gravity field; see Lowry [**9**, p. 155] for today's current best guess for Earth's gravity field. Up to the Earth's radius, f_M is piece-wise linear, connecting the data 0, -10.8 and -9.8 at the center, the core-mantle juncture, and the Earth's surface, respectively; thereafter, f_M follows the inverse square law. Numerically solving (4) using f_M shows that the pebble comes within 302 km of the Earth's center in about 19.1 minutes, which beats f_L's values of 376 km and 21.2 minutes. With this problem in mind, it is natural to solve (4) generally for nonlinear gravity fields $f(r)$. In this section, we first of all formalize what is meant by a gravity field, and then explore several fields, make some observations and definitions, and pose problems, some of which may be good student projects.



a. Linear field b. A geophysical approximation

Figure 8 Gravity fields for the Earth

Let Ω be a planet whose density is spherically symmetric, and let $\delta(r)$ be the density of Ω at radial distance r. We specify that $\delta(r)$ be a nonnegative, real-valued, piece-wise continuous function on the interval $0 < r \leq R$, where R is the radius of Ω, and that δ is identically 0 when $r > R$. As shown in [**15**], the gravitational acceleration, or *gravity field*, f for Ω is

$$\boxed{f(r) = -\frac{4\pi G}{r^2}\int_0^r \rho^2\delta(\rho)\,d\rho, \quad \text{for } r > 0,} \tag{34}$$

where G is the universal gravitational constant and ρ is a dummy variable for radial distance. Observe that for any $r > R$, (34) implies that $f(r) = -f(R)R^2/r^2$. Since we wish to solve (4) for general gravity fields, it would be convenient to have a simple litmus test for detecting them. With this in mind, we observe that any function $f(r)$ that has hopes of being a gravity field must have the following properties:

i. The function $f(r) \leq 0$ for all $r \geq 0$.

ii. The function is continuous on the interval $[0, \infty)$.

iii. The function is piecewise differentiable on the interval $(0, \infty)$.

iv. For each $a \geq 0$ with $f(a) < 0$, $f(r) < 0$ for all $r \geq a$.

Any function f that has these four properties is called *taut*. (We use the descripter *taut* because of property *iv*, that the gravitational effect of a field being nonzero at some

point never completely dissipates out beyond that point.) To find the taut functions that are gravity fields, we first of all use (34) to find the density function $\delta(r)$ for a given gravity field. To do so, rewrite (34) as

$$r^2 f(r) = -4\pi G \int_0^s \rho^2 \delta(\rho)\, d\rho,$$

and differentiate with respect to r, obtaining $r^2 f'(r) + 2rf(r) = -4\pi G r^2 \delta(r)$, which when solved for $\delta(r)$ gives

$$\boxed{\delta(r) = -\frac{2f(r) + rf'(r)}{4\pi G r},} \tag{35}$$

which in turn means that for any gravity field $f(r)$, $rf'(r) + 2f(r)$ is nonpositive whenever f' is defined. Since every gravity field $f(r)$ is nonpositive, this means that

$$\boxed{-\frac{f'(r)}{f(r)} \leq \frac{2}{r},} \tag{36}$$

provided $f(r)$ is nonzero and $f'(r)$ is defined. If f is taut and (36) holds, these steps are reversible. In particular, given that f is negative, (36) means that $-f'(r) \geq 2f(r)/r$, which in turn means that $2f(r) + rf'(r) \leq 0$, which means that the right-hand side of (35) is nonnegative, making it a bona fide density function, and so f as given by (34) is a gravity field. Thus (36) is our litmus test for detecting which taut functions are gravity fields.

For any $a > 0$, we say that f is a gravity field on the interval $[0, a]$ if f is taut and condition (36) is satisfied on the interval $[0, a]$. For example $f(s) = -e^{-s}$ fails to be a gravity field on $[0, a]$ for sufficiently large a. To see this, note that condition (36) for this function becomes $1 \leq 2/s$, which means that this function behaves as a gravity field on $[0, a]$ only when $0 \leq a \leq 2$.

As another example, let $f_1(s) = -2h^2 s^2$ and $f_2(s) = h^2 s(s-3)$, where h is the constant given by (6). Since both f_1 and f_2 are nonpositive, differentiable, and decreasing on the interval $[0, 1]$, condition (36) is trivially satisfied on the interval, which means that both functions behave as gravity fields on the unit interval.

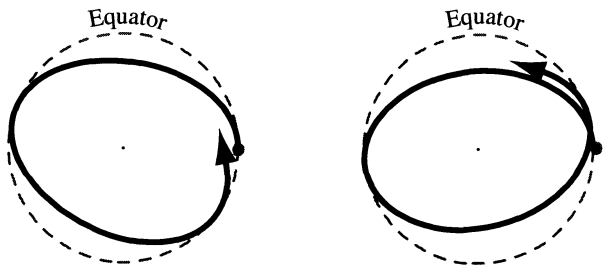
Two fields f and g are said to be gravity fields *modeling* a planet of radius R if $f(R) = g(R)$. When we refer to planet Ω and make reference to its gravity field f we mean the one inherently given by its density. We say that a *pebble-path* for planet Ω is the path generated by (4) and (5) under its inherent gravity field and its rotation rate. The *apogee* and *perigee* of a pebble-path are the greatest and least radial distances, respectively, of the pebble-path. The apogee and perigee associated with a field are the apogee and perigee of the pebble-path associated with the planet. A *proper* field for Ω is one wherein Ω's pebble-path has apogee R, Ω's radius. A *pebble-hole* for Ω is Ω's pebble-path adjusted to Ω's natural rotating coordinate system.

For example, for f_1 and f_2, (7) gives

$$\frac{d^2 z}{d\theta^2} + z = \frac{2}{z^4} \quad \text{and} \quad \frac{d^2 z}{d\theta^2} + z = \frac{3z - 1}{z^4},$$

respectively. For planets of radius 1, the pebble-paths associated with these two gravity fields are FIGURES 9a and 9b, respectively.

To discuss the shapes of these paths with greater facility, we stipulate some additional constraints and definitions. Planet Ω spins at a positive rate counterclockwise from above the polar plane. A standard field for Ω is $f_0(r) = -kr$, where k is some



a. A slow field b. A fast field

Figure 9 Precessing ellipses

positive constant. The *speed* of any proper field for Ω is $1/\pi$ times the angle measured counterclockwise between successive apogee occurrences for the pebble-path. For example, the speed of f_0 is 1, whereas the speed of $f_1 < 1$ and the speed of $f_2 > 1$. Let the *support* of a function be the set of all points in the interval $[0, \infty)$ where the function is nonzero. A function is said to have *positive* support if its support contains an open interval. If f and f_0 model the same planet, field f is *fast* if $f(r) \leq f_0(r)$ for all r between 0 and radius R and if $f - f_0$ has positive support. Similarly, a field f is *slow* if the inequality of the previous sentence is reversed. Note that f_1 is slow and f_2 is fast for a planet of radius 1. For two fields f and g modeling a planet of radius R, f is said to be *faster* than g if $f(r) - g(r)$ is nonpositive and has positive support.

Problem 1: Speed is well-defined With respect to any planet, prove that every proper field has a well-defined speed. That is, if a pebble is dropped, then it returns to the surface.

Problem 2: Speed comparison With $R = 1$ find the speeds of f_1 and f_2. Prove or disprove: a fast proper field's speed exceeds 1, whereas a slow proper field's speed is less than 1.

Problem 3: Perigee comparison Let f and g both model a planet of radius R. Are either of the following intuitive statements true? If f is a faster field than g then the perigee associated with f is less than the perigee associated with g. Furthermore, the speed of f exceeds that of g. For example, the piece-wise linear gravity field f_M is faster than the standard f_L. The speed of field f_M is about 1.012 whereas the speed of f_L is 1. The perigee of f_M is about 302 km, whereas f_L's perigee is about 376 km.

Problem 4: The fastest field Prove that for any planet the greatest speed of any proper field is 2 and occurs when $f(r) = -k/r^2$; this field also generates the least perigee.

Problem 5: Proportionality of speed and perigee For any planet Ω and real number q with $0 < q \leq R$, let Ω_q be the planet of radius q that has the same density function and rotation rate as Ω. For the standard field, by Kepler's modifed third law (21), no matter where the pebble is initially dropped at positive radial length between 0 and R, the speed remains fixed; that is, the speed of the pebble-path for Ω_q is the same for all q with $0 < q \leq R$. Is there a field so that its speed with respect to Ω_q is directly proportional to q?

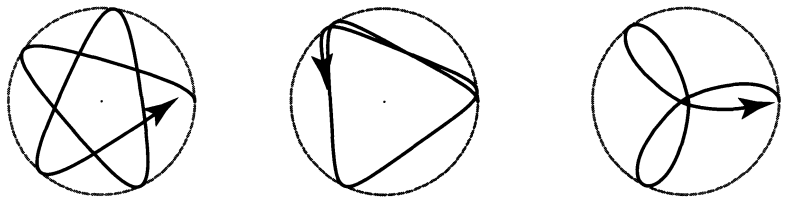
Problem 6: An analytical solution The graphs in FIGURE 9 are like precessing ellipses. The polar parametrizations

$$r(\theta) = \frac{1}{\sqrt{\cos^2(\alpha\theta) + b^2 \sin^2(\alpha\theta)}},$$

where b and α are positive constants with $R = 1$, generate curves resembling pebble-paths associated with fast fields when $\alpha < 1$ and slow fields if $\alpha > 1$. For $\alpha \neq 1$, find a variation of this parametrization and a field so that the variation solves (4) for that field. Perhaps begin by looking at the family of fields $f(r) = -kr^n$ where n is a real number.

Problem 7: Through the antipode Find a proper field for the Earth such that a pebble dropped on the Equator at geographical location X will next surface at X's antipode. That is, find a field for which the pebble-hole directly connects the drop point to its antipode. Note that the standard field does not have this property as FIGURE 5 demonstrates.

Problem 8: Exotic paths Find fields that yield more exotic pebble-paths than the ones shown in FIGURE 10 with $R = 1$ and $h = 1$. Note that these pebble-paths are reminiscent of the familiar trochoids. Find a fast field, other than $f(r) = -k/r^2$, for which its pebble-hole most resembles the polar flower $r = \cos(3\theta)$. Also, for f_c, find the velocity at which one should shoot a cannon ball due west at Ω's surface so that the pebble-path looks like $r = \cos(3\theta)$.



a. $f_a(r) = -7r^3(e^r - 1)$ b. $f_b(r) = -11r^{10}$ c. $f_c(r) = 0.0003r(r-3)^{15}$

Figure 10 A gallery of trochoid-like paths

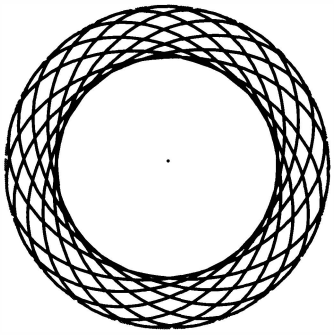
Problem 9: A 3-D extension Set up the three-dimensional equations analogous to (4). To do this, use the standard (ρ, θ, ϕ) spherical coordinate system so that the natural unit directions are

$$\begin{cases} \mathbf{u}_\rho = \sin\phi\cos\theta\,\mathbf{i} + \sin\phi\sin\theta\,\mathbf{j} + \cos\phi\,\mathbf{k} \\ \mathbf{u}_\theta = -\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j} \\ \mathbf{u}_\phi = \cos\phi\cos\theta\,\mathbf{i} + \cos\phi\sin\theta\,\mathbf{j} - \sin\phi\,\mathbf{k}. \end{cases}$$

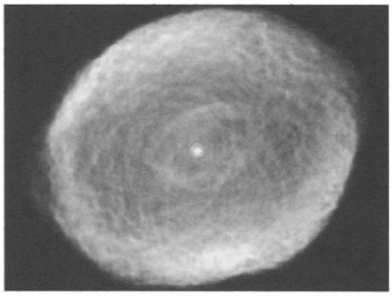
From these relations, find expressions for velocity $\mathbf{v}$ and acceleration $\mathbf{a}$ in terms of $\mathbf{u}_\rho$, $\mathbf{u}_\theta$ and $\mathbf{u}_\phi$. For example, see Spiegel [**17**, p. 163], for details. Then obtain three equations analogous to (2). For the standard gravity field $f(r) = -kr$, generate solutions in a computer algebra system. In terms of an analytic solution, find a formula corroborating (33). Try this twist on an old problem: *For a tunnel along any chord of the Equator not passing through the Earth's center, drop a ball in the eastern end of the tunnel and ascertain how far west it proceeds before doubling back; then generalize to three dimensions.*

Problem 10: Nebula cloud patterns The Spirograph Nebula, 2000 light years away, is a low-density, ellipsoidal cloud of matter encircling a small star; it appears to have structure similar to that of the precessing ellipse of FIGURE 11a. Permission to use the image of FIGURE 11b was very kindly given by R. Sahai of the Hubble Heritage Team out of the Jet Propulsion Laboratory of NASA; a grand color photo is available on the web [**12**]. In this cloud, each bit of matter is drifting in the gravitational field induced by the matter of the cloud as a whole. Therefore the gravity field in the outer reaches of the Nebula little resembles an inverse square law. Each tiny chunk of matter is like a pebble falling through the Nebula. Perhaps resistance to motion is negligible in much of the cloud, causing the paths or currents of matter to trace out spirographic swirls.

Perhaps there are larger chunks of matter falling through the Nebula that behave as bulldozers or vacuum cleaners sweeping out channels in the clouds, somewhat like jets leaving trails in the sky or ocean liners leaving enormously long trails in the sea. Is this gravitational phenomenon part of what helps to create the spirographic structure as captured by the Hubble Telescope in FIGURE 11b, or is it just coincidence?



a. A 2-D spirograph



b. The Spirograph Nebula

Figure 11 Spirographs

Acknowledgment. We thank Raymond Bloomer and Christopher Simoson for help in an attempt to replicate Hooke's experiment.

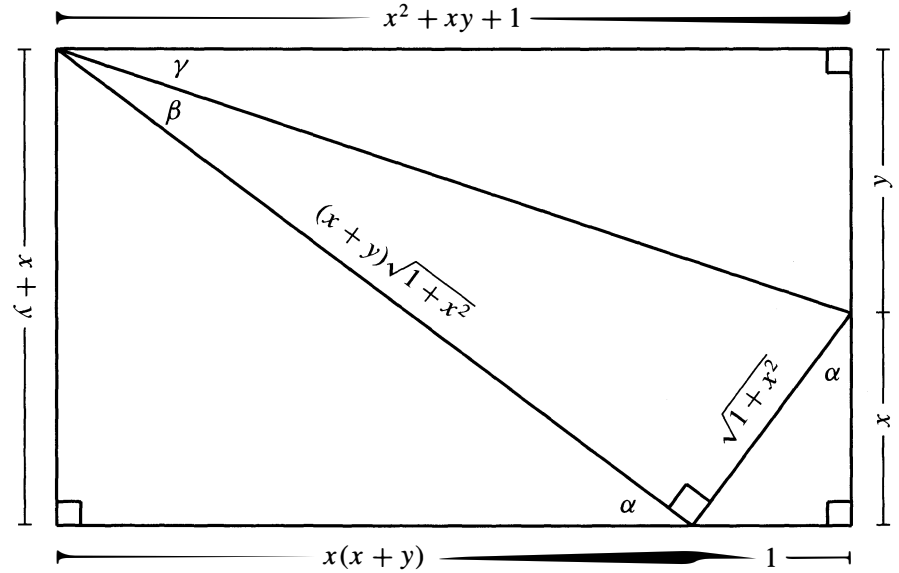
REFERENCES

1. V. I. Arnol'd, *Huygens & Barrow, Newton & Hooke*, Birkhäuser, Boston, 1990.
2. R. B. Banks, *Slicing Pizzas, Racing Turtles, and Further Adventures in Applied Mathematics*, Princeton University Press, Princeton, 1999.
3. Lewis Carroll, *The Annotated Alice*, edited by Martin Gardner, W. W. Norton, New York, 2000.
4. C. H. Edwards, and D. E. Penney, *Calculus: Early Transcendentals Version*, Prentice Hall, Upper Saddle River, NJ, 2003.
5. Camille Flammarion, A hole through the Earth, *The Strand Magazine* **38** (1909), 349–355.
6. G. R. Fowles and G. L. Cassiday, *Analytical Mechanics*, Saunders, Fort Worth, 1999.
7. Gary Feuerstein, painting of the Tower at Pisa, artist unknown, online at www.endex.com/gf/buildings.
8. Galileo Galilei, *Dialogue Concerning the Two Chief World Systems—Ptolemaic & Copernican*, translated by Stillman Drake, University of California Press, Berkeley, 1967.
9. William Lowrie, *Fundamentals of Geophysics*, Cambridge University Press, Cambridge, 1997.
10. J. B. Marion and S. T. Thornton, *Classical Dynamics of Particles and Systems*, Saunders, Fort Worth, 1995.
11. Isaac Newton, *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and his System of the World*, translated by Andrew Motte, and revised by Florian Cajori, University of California Press, Berkeley, 1960.
12. R. Sahai (JPL) et al., Hubble Heritage Team (STScI/AURA), NASA, *Astronomy Picture of the Day, December 14, 2002*, at antwrp.gsfc.nasa.gov/apod.
13. G. Shortley and D. Williams, *Elements of Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

14. G. F. Simmons, *Differential Equations*, McGraw-Hill, New York, 1972.
15. A. J. Simoson, The gravity of Hades, *this* MAGAZINE, **75**:5 (2002) 335–350.
16. Dava Sobel, *Galileo's Daughter*, Penguin, New York, 2000.
17. M. R. Spiegel, *Schaum's Outline of Vector Analysis*, McGraw-Hill, New York, 1959.
18. J. R. R. Tolkien, The Two Towers, *The Lord of the Rings*, Vol. II, Houghton Mifflin, Boston, 1994.

Proof Without Words: Euler's Arctangent Identity

$$\tan^{-1}\left(\frac{1}{x}\right) = \tan^{-1}\left(\frac{1}{x+y}\right) + \tan^{-1}\left(\frac{y}{x^2+xy+1}\right)$$



$$\alpha = \beta + \gamma$$

This is one of the many elegant arctangent identities discovered by Leonhard Euler. He employed them in the computation of π. For $x = y = 1$, we have Euler's Machin-like formula, $\pi/4 = \tan^{-1}(1/2) + \tan^{-1}(1/3)$. For $x = 2$ and $y = 1$, $\tan^{-1}(1/2) = \tan^{-1}(1/3) + \tan^{-1}(1/7)$. Substitute this into the previous identity, we obtain Hutton's formula, $\pi/4 = 2\tan^{-1}(1/3) + \tan^{-1}(1/7)$. In conjunction with the power series for arctangent, Hutton's formula was used as a check by Clausen in 1847 in computing π to 248 decimal places.

—Rex H. Wu
NYU Downtown Hospital
170 William Street
New York, NY 10038
RexHWu@aol.com

REFERENCES

1. P. Beckmann, *A History of* π, Golem Press, New York, 1971, p. 154.
2. R. B. Nelsen, Proof without words: An arctangent identity and series, this MAGAZINE **64** (1991), 241.

Upper Bounds on the Sum of Principal Divisors of an Integer

ROGER B. EGGLETON
Illinois State University
Normal, IL 61790
roger@math.ilstu.edu

WILLIAM P. GALVIN
School of Mathematical and Physical Sciences
University of Newcastle
Callaghan, NSW 2308 Australia

A *prime-power* is any integer of the form p^α, where p is a prime and α is a positive integer. Two prime-powers are *independent* if they are powers of different primes. The *Fundamental Theorem of Arithmetic* amounts to the assertion that every positive integer N is the product of a unique set of independent prime-powers, which we call the *principal divisors* of N. For example, 3 and 4 are the principal divisors of 12, while 2, 5 and 9 are the principal divisors of 90. The case $N = 1$ fits this description, using the convention that an empty set has product equal to 1 (and sum equal to 0).

In a recent invited lecture, Brian Alspach noted [**1**]: *Any odd integer $N > 15$ that is not a prime-power is greater than twice the sum of its principal divisors.* For instance, 21 is more than twice 3 plus 7, and 35 is almost three times 5 plus 7, but 15 falls just short of twice 3 plus 5. Alspach asked for a "nice" (elegant and satisfying) proof of this observation, which he used in his lecture to prove a result about cyclic decomposition of graphs.

Responding to the challenge, we prove Alspach's observation by a very elementary argument. You, the reader, must be the judge of whether our proof qualifies as "nice." We also show how the same line of reasoning leads to several stronger yet equally elegant upper bounds on sums of principal divisors. Perhaps surprisingly, our methods will *not* focus on properties of integers. Rather, we consider properties of finite sequences of positive real numbers, and use a classical elementary inequality between the product and sum of any such sequence. But first, let us put Alspach's observation in its number theoretic context.

Aliquot parts and principal divisors

The positive divisors of a positive integer have fascinated human minds for millennia. The divisors of an integer $N > 1$ that are positive but less than N are the *aliquot parts* of N. It is usual to denote their sum by $s(N)$. Classical Greek mathematicians singled out the aliquot parts of N from among the integers less than N, by noting that N can be "built" additively from multiple copies of any one of the aliquot parts. For Euclid [**4**], a *prime* number is "that which is measured by a unit alone" (Book VII, definition 11), that is, a number which has 1 as its only aliquot part, so N is prime if $s(N) = 1$. Again, a *perfect* number is "that which is equal to its own parts" (Book VII, definition 22), that is, a number that can be "built" additively from a single copy of each of its aliquot parts, so N is perfect if $s(N) = N$. Others, such as Theon, added that N is *deficient* if $s(N) < N$, and *abundant* if $s(N) > N$. The numbers 6, 10, and 12 are examples from the three classes.

Euclid knew that there are infinitely many primes (Book IX, Theorem 20) and that an even number of the form $2^{k-1}(2^k - 1)$ is perfect when $2^k - 1$ is prime (Book IX, Theorem 36). In more modern times it has been proved that every even perfect number *must* have this form, and currently 40 such perfect numbers have been found, corresponding to the known Mersenne primes [**12**], but it is not yet known whether there are infinitely many perfect numbers. Indeed, it is not known whether any odd perfect number exists, though many constraints on the possible form of such a number have been proved. By contrast, infinitely many positive integers are deficient and infinitely many are abundant; there can be no doubt that the Greeks knew easy proofs of these facts.

Interest in such matters underlies sophisticated modern computational studies of *aliquot sequences*, the sequences $a_0, a_1, a_2, \ldots$ beginning at a chosen positive integer $a_0 = N$, with each subsequent term found by computing the sum of the aliquot parts of the current term: $a_{k+1} = s(a_k)$ for $k \geq 0$. (See [**2, 7**] as entry points to current knowledge about aliquot sequences.) The aliquot sequence of a given N behaves in one of three possible ways: (1) after a finite number of terms it arrives at 1; (2) after a finite number of terms it enters a finite cycle, which repeats forever; (3) it continues forever without repetition. Sequences that arrive at a perfect number are of type (2), as are those that arrive at either member of a pair of *amicable* numbers, namely solutions to $s(a) = b$, $s(b) = a$. Pythagoras knew that $a = 220$ and $b = 284$ are the smallest amicable pair. Members of larger cycles are called *sociable numbers,* and several examples have been found in modern times. Intriguingly, it is not yet known whether there are any sequences of type (3); currently there are just five possible candidates with $N < 1000$, the first being $N = 276$.

It can be checked that the ratio $s(N)/N$ is 2 when $N = 120$, and is 3 when $N = 30240$. Indeed, it turns out that $s(N)/N$ has no absolute upper bound, and various simple proofs are known. When we have proved the key inequality we need in this article, we shall show that it also provides an elementary proof of this fact. (A recent Note in the MAGAZINE by Ryan [**10**] concerns the denseness of the set of numbers of the form $s(N)/N$, and of the complementary set, in the positive reals.)

Like the aliquot parts of a positive integer N, the principal divisors are a rather natural subset of the divisors of N. Indeed, if N is not a prime-power, its principal divisors are a proper subset of its aliquot parts. Thus $s^*(N)$, the sum of principal divisors of N, satisfies $s^*(N) < s(N)$ whenever N is not a prime-power. In contrast to $s(N)$, it turns out in fact that $s^*(N)$ never exceeds N. We shall prove this as our first theorem. Following common practice, we write $d|N$ when d is a positive divisor of N, and $p^\alpha \| N$ when p^α is a principal divisor of N. The Fundamental Theorem of Arithmetic implies that any positive integer N can be built multipicatively from a single copy of each of its principal divisors:

$$N = \prod_{p^\alpha \| N} p^\alpha,$$

where the notational convention is that the product ranges over all principal divisors of N. If Π is replaced by Σ, we have the sum of all principal divisors of N. It is simple and instructive to prove

THEOREM 1. *Every positive integer N satisfies*

$$N = \prod_{p^\alpha \| N} p^\alpha \geq \sum_{p^\alpha \| N} p^\alpha = s^*(N), \tag{1}$$

and (1) *holds with equality just when N is a prime-power.*

Proof. When $N = 1$, the set of all principal divisors of N is empty, so by standard conventions for empty sums and products, (1) holds with strict inequality in this case. Clearly (1) holds with equality when N has exactly one principal divisor (so N is a prime-power). Next suppose N has exactly two principal divisors, say $N = p^\alpha q^\beta$. The inequality $p^\alpha q^\beta > p^\alpha + q^\beta$ is equivalent to $(p^\alpha - 1)(q^\beta - 1) > 1$, and the latter is satisfied because $2 \le p^\alpha < q^\beta$ holds without loss of generality. Now suppose for some $k \ge 2$ that (1) holds with strict inequality for every positive integer with k principal divisors. Let N be any integer with exactly $k + 1$ principal divisors, let q^β be one of them, and let $N^* := N/q^\beta$. Then N^* has exactly k principal divisors, so

$$N = q^\beta N^* = q^\beta \prod_{p^\alpha \| N^*} p^\alpha > q^\beta \sum_{p^\alpha \| N^*} p^\alpha = \sum_{p^\alpha \| N^*} p^\alpha q^\beta$$

$$> \sum_{p^\alpha \| N^*} (p^\alpha + q^\beta) = kq^\beta + \sum_{p^\alpha \| N^*} p^\alpha > q^\beta + \sum_{p^\alpha \| N^*} p^\alpha = \sum_{p^\alpha \| N} p^\alpha.$$

Hence (1) again holds with strict inequality. The theorem now follows by induction on k. ∎

From the proof of Theorem 1, we see that the inequality (1) will usually be very weak when N has several principal divisors, especially if any of them is relatively large. So could it be that N is usually at least twice as large as the sum of its principal divisors? It certainly can! This is Alspach's observation, which we mentioned at the outset:

THEOREM 2. *Let N be an odd positive integer with at least two distinct prime factors. If $N > 15$, then*

$$\frac{N-1}{2} \ge \sum_{p^\alpha \| N} p^\alpha = s^*(N). \tag{2}$$

We shall now briefly recall a classic inequality for real numbers, and then use it to prove Theorem 2.

The Bernoulli-Weierstrass inequality

Let $\mathbb{R}^+ := \{x \in \mathbb{R} : x \ge 0\}$ and, for any $n \ge 1$, let $\mathbf{a} := (a_1, a_2, \ldots, a_n) \in (\mathbb{R}^+)^n$ be a sequence of nonnegative real numbers. Weierstrass [**11**] reasoned:

$$(1 + a_1)(1 + a_2) = 1 + a_1 + a_2 + a_1 a_2 \ge 1 + a_1 + a_2,$$

$$(1 + a_1)(1 + a_2)(1 + a_3) \ge (1 + a_1 + a_2)(1 + a_3) \ge 1 + a_1 + a_2 + a_3,$$

and so on. Modulo attention to when equality may hold, this is essentially an inductive proof of the following theorem.

THEOREM 3. (WEIERSTRASS) *If $\mathbf{a} \in (\mathbb{R}^+)^n$ and $n \ge 1$, then*

$$\prod_{i=1}^{n} (1 + a_i) \ge 1 + \sum_{i=1}^{n} a_i, \tag{3}$$

and (3) *holds with equality if and only if at most one of the numbers a_i is nonzero.*

This classical elementary inequality (3) is the key tool underlying our arguments in this article. Some authors, such as Durell and Robson [**3**], call it Weierstrass's inequality but it is not clear whether Weierstrass was the first to establish it. Hardy, Littlewood,

and Pólya [**8**] noted it as Theorem 58 without attribution, though they credited Jacques Bernoulli with the special case in which **a** is a constant sequence with terms greater than -1. We shall refer to (3) as the *Bernoulli-Weierstrass inequality.*

Earlier when discussing aliquot parts we remarked that the ratio $s(N)/N$ is known to have no absolute upper bound. It is of interest here to see how this can be derived from the Bernoulli-Weierstrass inequality.

THEOREM 4. *For any integer $N \geq 2$, the sum of aliquot parts $s(N)$ satisfies*

$$\frac{s(N)}{N} \geq \sum_{p|N} \frac{1}{p}, \tag{4}$$

and (4) *holds with equality if and only if N is prime.*

Proof. The sum of all positive divisors of N is

$$\begin{aligned} N + s(N) &= \prod_{p^\alpha \| N} \sum_{p^\beta | p^\alpha} p^\beta \geq \prod_{p^\alpha \| N} (p^\alpha + p^{\alpha-1}) \\ &= \prod_{p^\alpha \| N} p^\alpha \left(1 + \frac{1}{p}\right) = N \prod_{p|N} \left(1 + \frac{1}{p}\right). \end{aligned}$$

The second step holds with equality if and only if all principal divisors of N are prime, so if and only if N is squarefree. After dividing by N, the Bernoulli-Weierstrass inequality (3) now gives

$$1 + \frac{s(N)}{N} \geq \prod_{p|N} \left(1 + \frac{1}{p}\right) \geq 1 + \sum_{p|N} \frac{1}{p}$$

with equality at the second step just when N has only one prime divisor. The stated result now follows. ∎

Euler proved in 1737 that the sum of reciprocals of all primes is divergent [**9**], so it follows immediately from (4) that $s(N)/N$ has no absolute upper bound.

Deducing Alspach's inequality

The Bernoulli-Weierstrass inequality is really about sums and products of real numbers close to 1. We want to apply it to integers, such as occur in Alspach's inequality (Theorem 2) so we need to *scale* the individual terms to get a version of the Bernoulli-Weierstrass inequality that is about sums and products of real numbers close to some positive real number b, which we shall choose subsequently.

With $\mathbf{a} \in (\mathbb{R}^+)^n$ and $n \geq 1$, and any strictly positive $b \in \mathbb{R}^+$, multiply both sides of the Bernoulli-Weierstrass inequality (3) by b^n. Then

$$\prod_{i=1}^{n} (b + a_i b) \geq b^n + b^{n-1} \sum_{i=1}^{n} a_i b.$$

Put $\mathbf{c} := (c_1, c_2, \ldots, c_n) = (a_1 b, a_2 b, \ldots, a_n b) = b\mathbf{a}$, and choose any strictly positive $d \in \mathbb{R}^+$ such that $dn \leq b^{n-1}$. With this notation, we have

$$\prod_{i=1}^{n} (b + c_i) \geq b^n + b^{n-1} \sum_{i=1}^{n} c_i \geq d \left(bn + n \sum_{i=1}^{n} c_i \right) \geq d \sum_{i=1}^{n} (b + c_i),$$

where the last step holds with strict inequality if $n \geq 2$ and $\Sigma_{i=1}^{n} c_i > 0$. The latter fails only when $\mathbf{c} = \mathbf{0}$, where $\mathbf{0} := (0, 0, \ldots, 0) \in \mathbb{R}^+$. Hence we have a scaled version of the Bernoulli-Weierstrass inequality:

PRODUCT-SUM LEMMA. *For $n \geq 1$, any strictly positive $b \in \mathbb{R}^+$, and any sequence $\mathbf{c} \in (\mathbb{R}^+)^n$, let $N := \Pi_{i=1}^{n}(b + c_i)$. Then for any $d \in \mathbb{R}^+$ satisfying $0 < d \leq b^{n-1}/n$, we have*

$$\frac{N}{d} \geq \sum_{i=1}^{n}(b + c_i), \tag{5}$$

and (5) *holds with equality if and only if $d = b^{n-1}/n$, and $\mathbf{c} = \mathbf{0}$ if $n \geq 2$.*

In (5), note that N is a positive real, not necessarily an integer. To prove Theorem 2, we want an inequality of the form (5) with $d = 2$. But we may take $d = 2$ in the Product-Sum Lemma when $n = 2$ and $b = 4$, or when $n = 3$ and $b = \sqrt{6}$, or when $n \geq 4$ and $b = 2$. Put $\mathbf{d} := (d_1, d_2, \ldots, d_n) = (b + c_1, b + c_2, \ldots, b + c_n)$. Then $\mathbf{d}$ is a sequence of real numbers (not necessarily integers) close to b, and we have

THEOREM 5. *Let $\mathbf{d} \in (\mathbb{R}^+)^n$, with $n \geq 2$.*

(a) *If $4 \leq d_1 \leq d_2$, then $\frac{1}{2}d_1d_2 \geq d_1 + d_2$.*

(b) *If $\sqrt{6} \leq d_1 \leq d_2 \leq d_3$, then $\frac{1}{2}d_1d_2d_3 \geq d_1 + d_2 + d_3$.*

(c) *If $2 \leq d_1 \leq d_2 \leq \cdots \leq d_n$ and $n \geq 4$, then*

$$\frac{1}{2}d_1d_2 \ldots d_n \geq d_1 + d_2 + \cdots + d_n.$$

In each case, the final relation holds with equality if and only if the preceding relations all hold with equality.

Now let us require the d_i of Theorem 5 to be distinct positive integers:

COROLLARY. *Let N be a product of $n \geq 2$ distinct positive integers $d_1 < d_2 < \cdots < d_n$. Then*

$$\frac{N}{2} > \sum_{i=1}^{n} d_i \tag{6}$$

if (a) *$n = 2$ and $d_1 \geq 4$, or* (b) *$n = 3$ and $d_1 \geq 3$, or* (c) *$n \geq 4$ and $d_1 \geq 2$.*

We are now very close to having proved Alspach's inequality, Theorem 2. In fact, we are about to obtain a more comprehensive result that also admits more than half the even integers. Since N and $\Sigma_{i=1}^{n} d_i$ are integers in the Corollary to Theorem 5, the inequality (6) is equivalent to

$$\left\lfloor \frac{N-1}{2} \right\rfloor \geq \sum_{i=1}^{n} d_i.$$

In particular, if N has $n \geq 2$ distinct prime factors, we may take the d_i to be the principal divisors of N, obtaining

$$\left\lfloor \frac{N-1}{2} \right\rfloor \geq \sum_{p^\alpha \| N} p^\alpha = s^*(N)$$

whenever the conditions of the Corollary hold: since (c) certainly holds when $n \geq 4$ and the d_i are principal divisors, the only possible exceptions are when (a) fails, so $n = 2$ and $2^1 \| N$ or $3^1 \| N$, or when (b) fails, so $n = 3$ and $2^1 \| N$. Let us settle the remaining details when $n = 2$ and $n = 3$.

When $n = 2$, let $d_1 < d_2$ be the principal divisors of N. Suppose that $d_1 = 2$, then $\frac{1}{2}N - (2 + d_2) = -2 < 0$, so $\frac{1}{2}N$ cannot exceed $d_1 + d_2$. Again, if $d_1 = 3$, then $\frac{1}{2}N - (3 + d_2) = \frac{1}{2}(d_2 - 6) > 0$ provided $d_2 \geq 7$, so $d_2 = 4$ or 5 are the exceptions. Thus, we have ruled out the cases with $N = 12, 15$ or twice an odd prime-power; in particular, every $N < 20$ with $n = 2$ is ruled out.

When $n = 3$, let $d_1 = 2 < d_2 < d_3$ be the principal divisors of N. In this case, $\frac{1}{2}N - (2 + d_2 + d_3) = (d_2 - 1)(d_3 - 1) - 3 \geq 5$, since $d_2 \geq 3, d_3 \geq 5$. Thus, there are no exceptions to the desired inequality when $n = 3$.

This completes the proof of the following result, which is more comprehensive than Theorem 2, thus achieving our original Alspach objective:

THEOREM 2A. *Any integer $N \geq 20$ with at least two distinct prime factors satisfies*

$$\left\lfloor \frac{N-1}{2} \right\rfloor \geq \sum_{p^\alpha \| N} p^\alpha = s^*(N), \tag{7}$$

except when $N = 2q^\beta$, where q is an odd prime and $\beta \geq 1$.

Note that the exceptions to (7) are actually near-misses: if $N = 2q^\beta$, then

$$\frac{N+4}{2} = \sum_{p^\alpha \| N} p^\alpha = s^*(N).$$

After checking the individual cases with $N < 20$, we deduce:

THEOREM 2B. *If N is any positive integer with at least two distinct prime factors, then*

$$\left\lfloor \frac{N+4}{2} \right\rfloor \geq \sum_{p^\alpha \| N} p^\alpha = s^*(N), \tag{8}$$

and equality holds just when $N = 2q^\beta$, where q is an odd prime and $\beta \geq 1$.

In fact, we can readily establish a stronger upper bound than (7) for $s^*(N)$, an upper bound that depends on the number of distinct prime factors in N. To achieve this, note that $n^2 \leq 3^{n-1}$ for all integers $n \geq 3$, so we may take $b = 3$ and $d = n \geq 3$ in the Product-Sum Lemma. Thus we have an extension of Theorem 5:

THEOREM 5A. *Let $\mathbf{d} \in (\mathbb{R}^+)^n$, with $n \geq 3$. If $3 \leq d_1 \leq \cdots \leq d_n$, then*

$$\frac{1}{n} d_1 d_2 \ldots d_n \geq d_1 + d_2 + \cdots + d_n,$$

and equality holds if and only if $n = 3$ and $d_1 = d_2 = d_3 = 3$.

This leads us to a result that subordinates Theorems 1, 2 and 2A:

THEOREM 6. *Any positive integer N with $n \geq 1$ distinct prime factors satisfies*

$$\frac{N}{n} \geq \sum_{p^\alpha \| N} p^\alpha = s^*(N), \tag{9}$$

except when $N = 12, 15$ or $2q^\beta$, where q is an odd prime and $\beta \geq 1$. Also (9) holds with equality just when $N = 30$ or p^α, where p is any prime and $\alpha \geq 1$.

Proof. The result is obvious when $n = 1$, and follows from Theorem 2A when $n = 2$. Suppose $n \geq 3$. If all principal divisors of N are at least 3, taking the d_i in Theorem 5A to be these principal divisors immediately yields (9) with strict inequality. So suppose $N = 2N^*$, where N^* is a product of $n - 1$ odd principal divisors. If $N = 30$ it is evident that (9) holds with equality. Otherwise we may assume $N^* > 15$, so $N^* \geq 21$ and

$$\frac{N^*}{n-1} > \sum_{p^\alpha \| N^*} p^\alpha = s^*(N^*)$$

follows from Theorem 2A if $n = 3$, and from Theorem 5A if $n \geq 4$. Hence

$$\frac{N}{n} = \frac{2N^*}{n} = \left(1 + \frac{n-2}{n}\right)\frac{N^*}{n-1} > \left(1 + \frac{n-2}{n}\right)\sum_{p^\alpha \| N^*} p^\alpha.$$

But $s^*(N^*) \geq 10$ and $(n-2)/n \geq \frac{1}{3}$, so

$$\frac{N}{n} > \left(1 + \frac{n-2}{n}\right)\sum_{p^\alpha \| N^*} p^\alpha > 3 + \sum_{p^\alpha \| N^*} p^\alpha > \sum_{p^\alpha \| N} p^\alpha = s^*(N).$$

Thus N satisfies (9) with strict inequality, settling all remaining cases. ∎

Reverse arithmetic-geometric mean inequality

The sequence $\mathbf{a} \in (\mathbb{R}^+)^n$ with $n \geq 1$ has *arithmetic mean* $A(\mathbf{a})$ and *geometric mean* $G(\mathbf{a})$ given by

$$A(\mathbf{a}) := \frac{\sum_{i=1}^n a_i}{n} \quad \text{and} \quad G(\mathbf{a}) := \left(\prod_{i=1}^n a_i\right)^{1/n}.$$

The classical inequality comparing products and sums of finite sequences of nonnegative real numbers is the *Arithmetic-Geometric Mean Inequality*:

$$A(\mathbf{a}) \geq G(\mathbf{a}), \tag{10}$$

and (10) holds with equality if and only if $\mathbf{a}$ is a constant sequence.

A constant sequence is a scalar multiple of $\mathbf{1} := (1, 1, \ldots, 1) \in (\mathbb{R}^+)^n$, so $\mathbf{a}$ is constant precisely when $\mathbf{a} = c\mathbf{1}$ for some $c \in \mathbb{R}^+$. Given any $\mathbf{a}, \mathbf{b} \in (\mathbb{R}^+)^n$, we say that $\mathbf{a}$ *dominates* $\mathbf{b}$ if $a_i \geq b_i$ holds for every i in the interval $1 \leq i \leq n$, and that $\mathbf{a}$ *strictly dominates* $\mathbf{b}$ if furthermore the strict inequality $a_i > b_i$ holds for at least one i. Thus (10) holds with strict inequality precisely when $\mathbf{a}$ strictly dominates $c\mathbf{1}$, where $c := \min\{a_i : 1 \leq i \leq n\}$.

We may regard (10) as an inequality in which a multiple of the sum $\Sigma_{i=1}^n a_i$ is at least as large as a power of the product $\Pi_{i=1}^n a_i$. To deduce Alspach's inequality we were concerned with inequalities in the reverse direction, where a multiple of the product is at least as large as the sum. This suggests the unfamiliar novelty of comparing a multiple of $G(\mathbf{a})$ with a power of $A(\mathbf{a})$. Such an inequality does result from the Product-Sum Lemma when we take $b = n, b + c_i = a_i, d = n^{n-2}, N = G(\mathbf{a})^n$ and $\Sigma_{i=1}^n (b + c_i) = nA(\mathbf{a})$. This yields:

THEOREM 7. *For $n \geq 2$, suppose the sequence $\mathbf{a} \in (\mathbb{R}^+)^n$ dominates the constant sequence $n\mathbf{1}$. Then its arithmetic mean $A(\mathbf{a})$ and geometric mean $G(\mathbf{a})$ satisfy*

$$\frac{G(\mathbf{a})}{n} \geq \left(\frac{A(\mathbf{a})}{n}\right)^{1/n} \tag{11}$$

and (11) *holds with equality precisely when $\mathbf{a} = n\mathbf{1}$.*

For instance, $\mathbf{a} = (3, \sqrt{10}, \sqrt{10})$ strictly dominates $(3, 3, 3)$, so Theorem 7 shows that $\mathbf{a}$ satisfies (11) strictly, whence $7/2 > \sqrt{10}$. Theorem 7 yields a further extension of Theorem 5, from which we deduce two corollaries:

THEOREM 5B. *Let $\mathbf{d} \in (\mathbb{R}^+)^n$ with $n \geq 2$. If $n \leq d_1 \leq \cdots \leq d_n$, then*

$$\frac{1}{n^{n-2}} d_1 d_2 \ldots d_n \geq d_1 + d_2 + \cdots + d_n,$$

and equality holds if and only if $n = d_1 = d_2 = \cdots = d_n$.

COROLLARY 1. *If N is the product of $n \geq 2$ distinct positive integers $d_1 < d_2 < \cdots < d_n$, with $d_1 \geq n$, then*

$$\frac{N}{n^{n-2}} > \sum_{i=1}^{n} d_i. \tag{12}$$

COROLLARY 2. *If the positive integer N has $n \geq 2$ principal divisors, and each is at least as large as n, then*

$$\frac{N}{n^{n-2}} > \sum_{p^\alpha \| N} p^\alpha = s^*(N). \tag{13}$$

Extending an upper bound on $s^*(N)$

Let us review our progress. We began with the objective of finding, by elementary means, an upper bound on the sum $s^*(N)$ of principal divisors of a positive integer N. Our target upper bound was N/d with $d = 2$. We began with the modest result that the bound with denominator $d = 1$ holds without exception. Subsequently we found all N for which $d = 2$ holds. Passing from constant denominator to a linear function of n, the number of distinct prime factors of N, in Theorem 6 we found all exceptions to the simple and elegant upper bound with $d = n$. Now Corollary 2 to Theorem 5B has brought to our attention the upper bound with a super-exponential denominator $d = n^{n-2}$. In Corollary 2, that upper bound is subject to quite a strong constraint on the principal divisors of N. In this final section we complete the discussion by showing that the same upper bound holds for a much wider class of principal divisors. We use the following lemma.

MONOTONICITY LEMMA. *With $n \geq 2$, suppose $\mathbf{a} \in (\mathbb{R}^+)^n$ and $c \in \mathbb{R}^+$ satisfy*

$$c \prod_{i=1}^{n} a_i \geq \sum_{i=1}^{n} a_i > 0. \tag{14}$$

If $\mathbf{d} \in (\mathbb{R}^+)^n$ dominates $\mathbf{a}$, then

$$c \prod_{i=1}^{n} d_i \geq \sum_{i=1}^{n} d_i > 0. \tag{15}$$

Moreover, if (14) *holds with strict inequality, or if* **d** *strictly dominates* **a**, *then* (15) *holds with strict inequality.*

Proof. By (14), c and every a_i are strictly positive. Since **d** dominates **a**, every d_i is strictly positive, and the second inequality in (15) follows. Let $\delta := \max\{d_i/a_i : 1 \le i \le n\}$. For some subscript k, we have $d_k = \delta a_k$ and $d_i \ge a_i$ for all $i \ne k$, so $\Pi_{i=1}^n d_i \ge \delta \Pi_{i=1}^n a_i$. Then (15) follows, since

$$c \prod_{i=1}^{n} d_i \ge c\delta \prod_{i=1}^{n} a_i \ge \delta \sum_{i=1}^{n} a_i \ge \sum_{i=1}^{n} d_i. \tag{16}$$

If (14) holds with strict inequality, the second step in (16) is a strict inequality. Also the last step in (16) holds with equality only if $\mathbf{d} = \delta\mathbf{a}$. But then $\Pi_{i=1}^n d_i = \delta^n \Pi_{i=1}^n a_i$, implying strict inequality in the first step of (16) when **d** strictly dominates **a**, for then $\delta > 1$. The lemma follows. ∎

In Corollary 2 to Theorem 5B the upper bound on $s^*(N)$ with super-exponential denominator $d = n^{n-2}$ holds if the principal divisors of N are at least n. We now show that the same upper bound holds if the principal divisors exceed $n/2$, a much milder constraint.

THEOREM 8. *If N is any positive integer with $n \ge 2$ principal divisors, and each is greater than $n/2$, then*

$$\frac{N}{n^{n-2}} \ge \sum_{p^\alpha \| N} p^\alpha = s^*(N), \tag{17}$$

and (17) *holds with equality precisely when $N = 30$.*

Proof. Suppose N has $n \ge 2$ principal divisors, each greater than $n/2$. Let $\mathbf{d} \in (\mathbb{R}^+)^n$ be the sequence of those principal divisors in increasing order. At most one principal divisor of N is even, so **d** dominates the increasing sequence $\mathbf{a}^*(n)$, which we define to comprise the smallest even integer greater than $n/2$ and the smallest $n-1$ consecutive odd integers greater than $n/2$. We claim that

$$\frac{N}{n^{n-2}} = \frac{1}{n^{n-2}} \prod_{i=1}^{n} d_i \ge \sum_{i=1}^{n} d_i = s^*(N).$$

Using the Monotonicity Lemma, this claim will follow if we can show that

$$\frac{1}{n^{n-2}} \prod_{i=1}^{n} a_i^*(n) \ge \sum_{i=1}^{n} a_i^*(n). \tag{18}$$

For brevity we use $P(n)$ and $S(n)$, respectively, to denote the left and right sides of (18). Routine evaluation of (18) for each n in the interval $2 \le n \le 12$ shows that $P(3) = S(3)$, corresponding to $N = 30$, and $P(n) > S(n)$ in all other cases. We now prove inductively that $P(n) > S(n)$ holds for every $n \ge 9$, whence the theorem follows by the Monotonicity Lemma, with $N = 30$ as the sole instance of equality. (The overlap for $9 \le n \le 12$ is needed.)

The fine structure of $\mathbf{a}^*(n)$ depends on the residue class of n modulo 4. The even integer in $\mathbf{a}^*(n+4)$ is 2 greater than the even integer in $\mathbf{a}^*(n)$, and the odd integers in $\mathbf{a}^*(n+4)$ are all but the smallest odd integer in $\mathbf{a}^*(n)$, together with the next 5 odd

integers. In particular, suppose $n = 4k + 1$ for some positive integer k. Then the ratio $P(n+4)/P(n)$ is equal to

$$\frac{2k+4}{2k+2} \cdot \frac{(10k+1)(10k+3)(10k+5)(10k+7)(10k+9)}{2k+1} \cdot \frac{n^{n-2}}{(n+4)^{n+2}}$$
$$= \frac{k+2}{k+1} \cdot \frac{10k+1}{4k+1} \cdot \frac{10k+3}{4k+1} \cdot \frac{10k+7}{4k+5} \cdot \frac{10k+9}{4k+5} \cdot \frac{5}{(1+\frac{4}{n})^n}.$$

As $k \to \infty$ the first, third and sixth factors decrease monotonically, so are always greater than their limits; the other three factors increase monotonically, so if we require $k \geq 2$ they are never less than their values at $k = 2$. Hence when $n = 4k + 1$ and $k \geq 2$ we have

$$\frac{P(n+4)}{P(n)} > 1 \cdot \frac{7}{3} \cdot \frac{5}{2} \cdot \frac{27}{13} \cdot \frac{29}{13} \cdot \frac{5}{e^4} > \frac{7}{3}. \tag{19}$$

Similarly, if $n = 4k + 1$ then

$$\frac{S(n+4)}{S(n)} = \frac{12k^2 + 25k + 14}{12k^2 + k + 1}.$$

As $k \to \infty$ this ratio decreases monotonically, so if we require $k \geq 2$ it never exceeds its value at $k = 2$, and

$$\frac{S(n+4)}{S(n)} \leq \frac{112}{51} < \frac{7}{3}. \tag{20}$$

If $P(n) > S(n)$ when $n = 4k + 1$ for some $k \geq 2$, then (19) and (20) imply

$$P(n+4) > \frac{7}{3}P(n) > \frac{7}{3}S(n) > S(n+4).$$

Since $P(9) > S(9)$, induction now guarantees that (18) holds with strict inequality when $n = 4k + 1$ for all $k \geq 2$.

Similar computations for n in the other residue classes modulo 4 complete the proof. ∎

Closing remarks The Monotonicity Lemma is actually strong enough to yield a number of our earlier results. In particular, once the formulations of the Product-Sum Lemma and Theorem 7 have been discovered, they can be readily proved using the Monotonicity Lemma. It is useful for proving inequalities in which a product is greater than a sum, but is of little help in the initial task of formulating the inequalities. In recent papers [**5, 6**], we studied an inequality between two polynomials related to the sum and product of an arbitrary sequence. Our results generalize the Bernoulli-Weierstrass inequality, so could yield inequalities like those established here. However, in the spirit of Alspach's motivating request, we tried here to keep our arguments as elementary and self-contained as possible.

Acknowledgment. The first author is grateful for the hospitality of the School of Mathematical and Physical Sciences, University of Newcastle, Australia, during completion of this article.

Requiescat. After a long battle with cancer, William P. Galvin passed away on December 12, 2003.

REFERENCES

1. B. Alspach, Group action on graph decompositions, colloquium talk at Illinois State University, Sept. 4, 2002.
2. M. Benito and J. L. Varona, Advances in aliquot sequences, *Math. Comp.* **68** (1999), 389–393.
3. C. V. Durell and A. Robson, *Advanced Algebra*, Vol. III, G. Bell and Sons, London, 1959.
4. Euclid, *The Elements*, trans. Sir Thomas L. Heath, 2nd ed., Dover Publications, New York, 1956.
5. R. B. Eggleton and W. P. Galvin, A polynomial inequality generalising an integer inequality, *J. Inequal. Pure and Appl. Math.* **3** (2002), Art. 52, 9pp. [http://jipam.vu.edu.au/v3n4/043_02.html]
6. ———, Asymptotic expansion of the equipoise curve of a polynomial inequality, *J. Inequal. Pure and Appl. Math.* **3** (2002), Art. 84, 13pp. [http://jipam.vu.edu.au/v3n5/091_02.html]
7. A. W. P. Guy and R. K. Guy, A record aliquot sequence, *Proc. Sympos. Appl. Math.* **48** (1994), 557–559.
8. G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, 2nd ed., Cambridge University Press, 1959.
9. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Oxford University Press, 1979.
10. R. F. Ryan, A simpler dense proof regarding the abundancy index, this MAGAZINE **76** (2003), 299–301.
11. K. Weierstrass, Über die Theorie der analytischen Facultäten, *Crelle's Journal* **51** (1856), 1–60.
12. G. F. Woltman, project founder and website author, Great International Mersenne Prime Search (GIMPS). The 40th Mersenne prime was discovered on Nov. 17, 2003. [http://www.mersenne.org/prime.html]

Proof Without Words: Every Octagonal Number Is the Difference of Two Squares

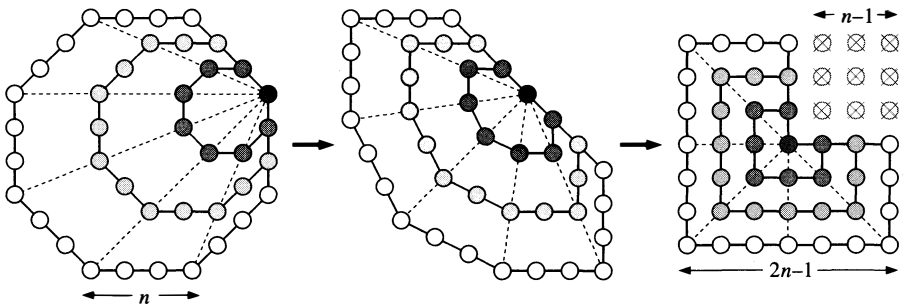
$$1 = 1 = 1^2 - 0^2$$

$$1 + 7 = 8 = 3^2 - 1^2$$

$$1 + 7 + 13 = 21 = 5^2 - 2^2$$

$$1 + 7 + 13 + 19 = 40 = 7^2 - 3^2$$

$$O_n = 1 + 7 + \cdots + (6n - 5) = (2n - 1)^2 - (n - 1)^2$$



—ROGER B. NELSEN
LEWIS & CLARK COLLEGE
PORTLAND, OR 97219

NOTES

Centroids Constructed Graphically

TOM M. APOSTOL
MAMIKON A. MNATSAKANIAN
Project MATHEMATICS!
California Institute of Technology
Pasadena, CA 91125-0001
apostol@caltech.edu
mamikon@caltech.edu

The centroid of a finite set of points Archimedes (287–212 BC), regarded as the greatest mathematician and scientist of ancient times, introduced the concept of center of gravity. He used it in many of his works, including the stability of floating bodies, ship design, and in his discovery that the volume of a sphere is two-thirds that of its circumscribing cylinder. It was also used by Pappus of Alexandria in the 3rd century AD in formulating his famous theorems for calculating volume and surface area of solids of revolution. Today a more general concept, center of mass, plays an important role in Newtonian mechanics. Physicists often treat a large body, such as a planet or sun, as a single point (the center of mass) where all the mass is concentrated. In uniform symmetric bodies it is identified with the center of symmetry.

This note treats the center of mass of a finite number of points, defined as follows. Given n points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$, regarded as position vectors in Euclidean m-space, relative to some origin O, let $w_1, w_2, \ldots, w_n$ be n positive numbers regarded as weights attached to these points. The center of mass is the position vector $\boldsymbol{c}$ defined to be the weighted average given by

$$\boldsymbol{c} = \frac{1}{W_n} \sum_{k=1}^{n} w_k \boldsymbol{r}_k, \tag{1}$$

where W_n is the sum of the weights,

$$W_n = \sum_{k=1}^{n} w_k. \tag{2}$$

When all weights are equal, the center of mass $\boldsymbol{c}$ is called the centroid. If each $w_k = w$, then $W_n = nw$, the common factor w cancels in (1), and we get

$$\boldsymbol{c} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{r}_k, \tag{3}$$

which is equivalent to assigning weight 1 to each point. If the points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ are specified by their coordinates, the coordinates of $\boldsymbol{c}$ can be obtained by equating components in (1).

We describe two different methods for locating the centroid of a finite number of given points in 1-space, 2-space, or 3-space by graphical construction, without using coordinates or numerical calculations. The first involves making a guess and forming a

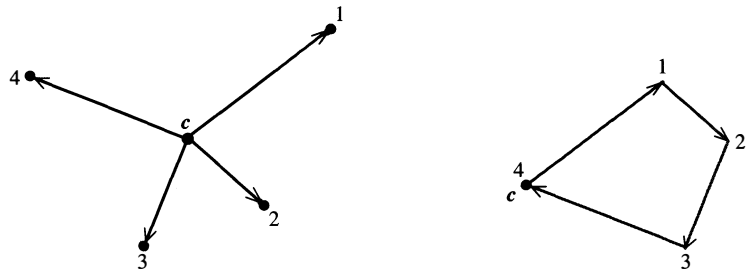
closed polygon. The second is an inductive procedure that combines centroids of two disjoint sets to determine the centroid of their union. For 1-space and 2-space both methods can be applied in practice with drawing instruments or with computer graphics programs. Centroids of points in higher-dimensional spaces can be determined with the help of geometric methods by projecting the points onto lower-dimensional spaces, depending on how the points are given. For example, points in 4-space with coordinates (x, y, z, t) can be projected onto points in the xy plane and in the zt plane where graphic methods apply.

Our geometric methods are best illustrated when the weights are equal (computing centroids), and we will indicate in appropriate places how the methods can be modified to the more general case of unequal weights (computing centers of mass).

Let $\boldsymbol{c}_k = \boldsymbol{r}_k - \boldsymbol{c}$, the geometric vector from centroid $\boldsymbol{c}$ to $\boldsymbol{r}_k$. Then (3) implies that

$$\sum_{k=1}^{n} \boldsymbol{c}_k = \boldsymbol{O}, \tag{4}$$

so $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n$ have centroid $\boldsymbol{O}$. FIGURE 1a shows four points and the geometric vectors represented by arrows emanating from their centroid. FIGURE 1b shows these vectors being added head to tail to form a closed polygon. Of course, the vectors can be added in any order.



(a) Vectors with sum $\boldsymbol{O}$ emanating from the centroid

(b) The vectors form a closed polygon

Figure 1 Vectors from the centroid form a closed polygon when placed successively head to tail

Method 1: Closing a polygon This geometric method for locating the centroid of n given points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ involves choosing a point $\boldsymbol{g}$ as a guess for the centroid. Then we construct a closed polygon and modify the guess once to determine $\boldsymbol{c}$.

We introduce the deficiency vector defined by

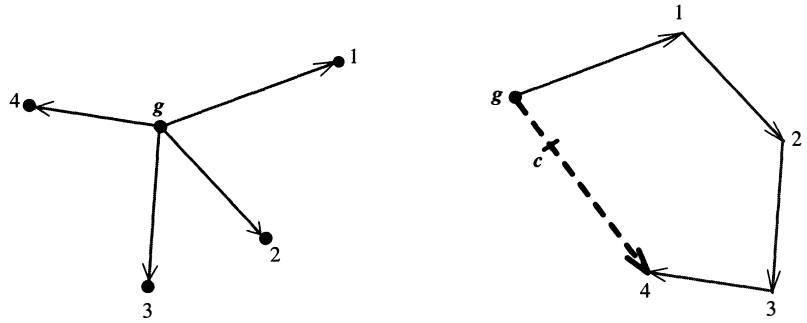
$$\boldsymbol{d} = \sum_{k=1}^{n} (\boldsymbol{r}_k - \boldsymbol{g}) = n\boldsymbol{c} - n\boldsymbol{g},$$

as a measure of the deviation between $\boldsymbol{g}$ and $\boldsymbol{c}$. A knowledge of $\boldsymbol{d}$ gives $\boldsymbol{c}$, because

$$\boldsymbol{c} = \boldsymbol{g} + \frac{1}{n}\boldsymbol{d}. \tag{5}$$

The vector $\boldsymbol{d}/n$ is the error vector. It tells us exactly what should be added to the guess $\boldsymbol{g}$ to obtain the centroid $\boldsymbol{c}$. FIGURE 2 illustrates the method by an example. The

four points used in FIGURE 1a are also shown in FIGURE 2a with a guess $\boldsymbol{g}$ for their centroid, and geometric vectors $\boldsymbol{r}_k - \boldsymbol{g}$ drawn from $\boldsymbol{g}$ to each of the four points. For simplicity, in the figure we use labels $k = 1, 2, 3, 4$ to denote these vectors. In FIGURE 2b the vectors are placed successively head to tail, starting from $\boldsymbol{g}$ to form the sum $\boldsymbol{d}$. If a lucky guess placed $\boldsymbol{g}$ at the centroid, the vectors placed head to tail would form a closed polygon as in FIGURE 1b, and $\boldsymbol{d}$ would be zero. But in FIGURE 2a, $\boldsymbol{g}$ is not the centroid, and the polygon formed by these four vectors in FIGURE 2b is not closed. However, an additional vector joining the tail of $\boldsymbol{r}_1 - \boldsymbol{g}$ to the head of $\boldsymbol{r}_4 - \boldsymbol{g}$ will close the polygon. This vector, shown as a broken line in FIGURE 2b, is the deficiency vector $\boldsymbol{d}$. We find $\boldsymbol{c}$ by simply adding the error vector $\boldsymbol{d}/4$ to $\boldsymbol{g}$. In practice, FIGURES 2a and 2b can be drawn on the same graph. We have separated them here for the sake of clarity.



(a) Vectors from guess $\boldsymbol{g}$ to the given points (b) The sum of the vectors in (a)

Figure 2 If $\boldsymbol{g}$ is not the centroid, the polygon obtained by adding the vectors is not closed. It can be closed by adding $-\boldsymbol{d}$ to the other vectors.

Although this example illustrates the method for four points in a plane, the method works equally well for any number of points in 1-space, 2-space, or 3-space. For n points we add the error vector $\boldsymbol{d}/n$ to $\boldsymbol{g}$, as indicated by (5), to get the centroid $\boldsymbol{c}$.

The error vector $\boldsymbol{d}/n$ is easily constructed geometrically. In fact, to multiply $\boldsymbol{d}$ by any positive scalar λ, plot $1/\lambda$ on a number line drawn in a convenient direction not parallel to $\boldsymbol{d}$, and join $1/\lambda$ and the head of $\boldsymbol{d}$ with a line segment. A parallel line to $\boldsymbol{d}$ through the unit on the number line intersects $\boldsymbol{d}$ at $\lambda\boldsymbol{d}$, as is easily verified by similarity of triangles.

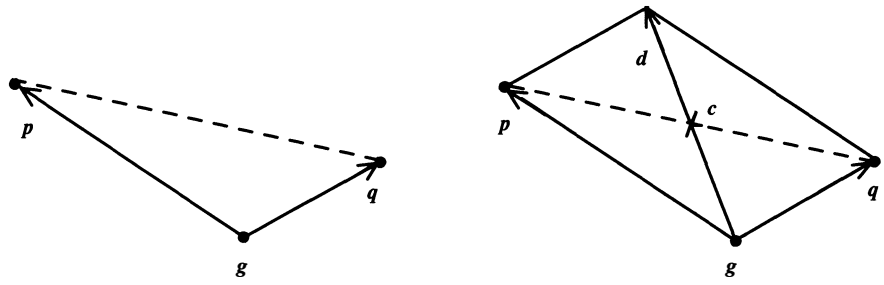
Multiplication by a general positive scalar is needed when the method of guessing once is used to find the weighted average $\boldsymbol{c}$ of n given points as defined by (1). As before, we make a guess $\boldsymbol{g}$ and let $\boldsymbol{d} = \sum_{k=1}^{n} w_k(\boldsymbol{r}_k - \boldsymbol{g}) = W_n(\boldsymbol{c} - \boldsymbol{g})$. This gives us $\boldsymbol{c} = \boldsymbol{g} + \boldsymbol{d}/W_n$, and it is easily verified that the foregoing geometric method can be adapted to find $\boldsymbol{d}$, the error vector $\boldsymbol{d}/W_n$, and hence $\boldsymbol{c}$.

Two simple examples illustrate how the method yields familiar interpretations of the centroid.

Example 1: Centroid of two points The centroid of two points is midway between them, and it is instructive to see how the geometric method works in this simple case. FIGURE 3a shows two points $\boldsymbol{p}$ and $\boldsymbol{q}$ and a guess $\boldsymbol{g}$ for their centroid. This may seem to be an outlandish guess because $\boldsymbol{g}$ is not on the line through $\boldsymbol{p}$ and $\boldsymbol{q}$. Nevertheless, the method works no matter where $\boldsymbol{g}$ is chosen. When we add the vector from $\boldsymbol{g}$ to $\boldsymbol{p}$ to that from $\boldsymbol{g}$ to $\boldsymbol{q}$, the sum $\boldsymbol{d} = (\boldsymbol{p} - \boldsymbol{g}) + (\boldsymbol{q} - \boldsymbol{g})$ is one diagonal of a parallelogram,

as shown in FIGURE 3b and (5) tells us that centroid $\boldsymbol{c} = \boldsymbol{g} + \boldsymbol{d}/2 = (\boldsymbol{p} + \boldsymbol{q})/2$, which obviously lies midway between $\boldsymbol{p}$ and $\boldsymbol{q}$.

Actually, by choosing $\boldsymbol{g}$ outside the line through $\boldsymbol{p}$ and $\boldsymbol{q}$, there is no need to construct $\boldsymbol{d}/2$, because we know that $\boldsymbol{c}$ lies on this line at the point where $\boldsymbol{d}$ intersects it.



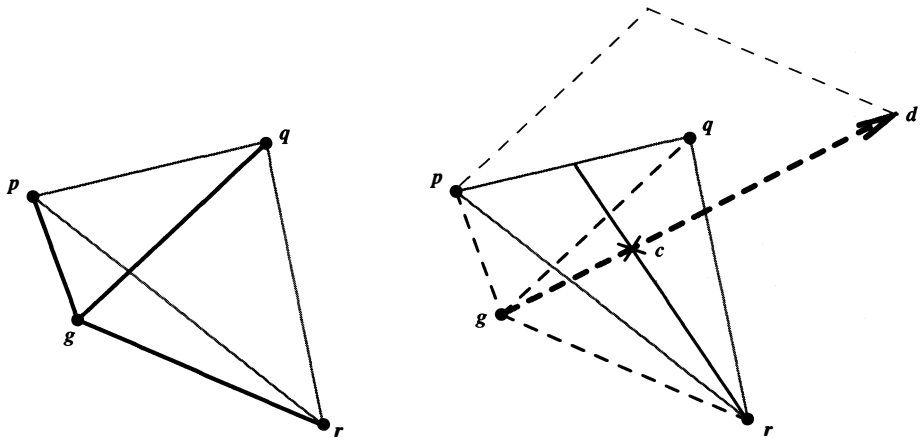
(a) Guess $\boldsymbol{g}$ for the centroid of $\boldsymbol{p}$ and $\boldsymbol{q}$ space (b) Centroid $\boldsymbol{c} = \boldsymbol{g} + \frac{1}{2}\boldsymbol{d} = \frac{1}{2}(\boldsymbol{p} + \boldsymbol{q})$

Figure 3 Determining the centroid of two points

Example 2: Centroid of three points Choose three points $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}$ and make a guess $\boldsymbol{g}$ for their centroid (FIGURE 4a). The given points are vertices of a triangle, possibly degenerate, but the guess $\boldsymbol{g}$ need not be in the plane of this triangle. According to Method 1, we form the sum

$$\boldsymbol{d} = (\boldsymbol{p} - \boldsymbol{g}) + (\boldsymbol{q} - \boldsymbol{g}) + (\boldsymbol{r} - \boldsymbol{g}) = (\boldsymbol{p} + \boldsymbol{q} + \boldsymbol{r}) - 3\boldsymbol{g}$$

and by (5) the centroid is $\boldsymbol{c} = \boldsymbol{g} + \boldsymbol{d}/3 = (\boldsymbol{p} + \boldsymbol{q} + \boldsymbol{r})/3$. It should be noted that the method works even if $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}$ are collinear provided we choose $\boldsymbol{g}$ not on the same line, which turns out to be a wise guess in this case. For such a $\boldsymbol{g}$, we know that $\boldsymbol{c}$ is on the line through $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}$, so vector $\boldsymbol{d}$ will automatically intersect this line at $\boldsymbol{c}$, which means there is no need to divide $\boldsymbol{d}$ by 3.



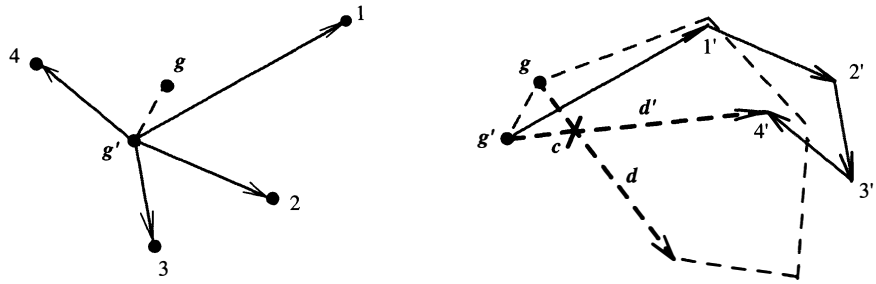
(a) Guess $\boldsymbol{g}$ for the centroid of $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}$ (b) Centroid $\boldsymbol{c} = \boldsymbol{g} + \frac{1}{3}\boldsymbol{d} = \frac{1}{3}(\boldsymbol{p} + \boldsymbol{q} + \boldsymbol{r})$

Figure 4 The three medians of a triangle intersect at the centroid of the vertices

We can also verify that in the nondegenerate case all three medians of the triangle meet at the centroid, and that the centroid divides each median in the ratio 2 : 1. In

FIGURE 4b the centroid is placed at the origin so that $\boldsymbol{p}+\boldsymbol{q}+\boldsymbol{r}=\boldsymbol{O}$. First consider the median from $\boldsymbol{r}$ to the edge joining vertices $\boldsymbol{p}$ and $\boldsymbol{q}$. The vector from the centroid to the midpoint of this edge is $(\boldsymbol{p}+\boldsymbol{q})/2$ whereas $\boldsymbol{r}=-(\boldsymbol{p}+\boldsymbol{q})$. This shows that the median passes through the centroid and that the distance from the centroid to vertex $\boldsymbol{r}$ is twice that from the centroid to the midpoint of the opposite edge. By interchanging symbols, we find the same is true for the other two medians. So all three medians meet at the centroid, and the centroid divides each median in the ratio 2 : 1 as asserted.

Variation of method 1 that avoids dividing FIGURE 5 shows an alternate way to determine the centroid $\boldsymbol{c}$ of the points in FIGURE 2 that avoids constructing the error vector $\boldsymbol{d}/4$. FIGURE 5a shows the same four points with a new guess $\boldsymbol{g}'$ chosen not on the line through $\boldsymbol{d}$. The polygon of FIGURE 2b appears again in FIGURE 5b as dashed lines, together with a new polygon (solid lines) obtained by adding the vectors $\boldsymbol{r}_k-\boldsymbol{g}'$. This produces a new deficiency vector $\boldsymbol{d}'$ joining the tail of $\boldsymbol{r}_1-\boldsymbol{g}'$ to the head of $\boldsymbol{r}_4-\boldsymbol{g}'$. Because $\boldsymbol{g}'$ is not on the line through $\boldsymbol{d}$, the two geometric vectors $\boldsymbol{d}$ and $\boldsymbol{d}'$ intersect at $\boldsymbol{c}$, as shown by the example in FIGURE 5b.



(a) Vectors from guess $\boldsymbol{g}'$ to the given points (b) Closed polygon formed from guess $\boldsymbol{g}'$

Figure 5 The centroid obtained as the intersection of two deficiency vectors $\boldsymbol{d}$ and $\boldsymbol{d}'$

Although the example in FIGURE 5 treats four points, the method also works for any finite number of points. It should be noted that if the given points are collinear a second guess is not needed provided we choose the first guess $\boldsymbol{g}$ not on the line through the given points. The construction used for three collinear points in Example 2 also works for any number of collinear points. The deficiency vector $\boldsymbol{d}$ will intersect the line through these points at the centroid $\boldsymbol{c}$.

The case of unequal weights is treated similarly as described previously. Make two guesses, and the two deficiency vectors so constructed can be shown to intersect at $\boldsymbol{c}$.

Method 2: Inductive process This method regards the given set of points as the union of two disjoint subsets whose centroids are known or can be easily determined. The centroids of the subsets are combined to determine the centroid of the union. The process depends on how the subsets are selected. For example, we can find the centroid of $n+1$ points if we know the centroid of any n of these points. If $\boldsymbol{c}_n$ denotes the centroid of points $\{1, 2, \ldots, n\}$, so that

$$\boldsymbol{c}_n=\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{r}_k, \quad \text{then} \quad \boldsymbol{c}_{n+1}=\frac{1}{n+1}\sum_{k=1}^{n+1}\boldsymbol{r}_k=\frac{1}{n+1}\left(\sum_{k=1}^{n}\boldsymbol{r}_k+\boldsymbol{r}_{n+1}\right),$$

or

$$\boldsymbol{c}_{n+1} = \frac{1}{n+1}(n\boldsymbol{c}_n + \boldsymbol{r}_{n+1}). \tag{6}$$

In other words, $\boldsymbol{c}_{n+1}$ is a weighted average of the two points $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$, with weight n attached to $\boldsymbol{c}_n$ and weight 1 attached to $\boldsymbol{r}_{n+1}$. Because $\boldsymbol{c}_{n+1}$ is a convex combination of $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$ it lies on the line joining $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$. Moreover, from (6) we find

$$\boldsymbol{c}_{n+1} - \boldsymbol{c}_n = \frac{1}{n+1}(\boldsymbol{r}_{n+1} - \boldsymbol{c}_n),$$

which shows that the distance between $\boldsymbol{c}_{n+1}$ and $\boldsymbol{c}_n$ is $1/(n+1)$ times the distance between $\boldsymbol{r}_{n+1}$ and $\boldsymbol{c}_n$. Repeated use of (6) provides a method for determining the centroid of any finite set.

In the case of unequal weights, (6) becomes

$$\boldsymbol{c}_{n+1} = \frac{1}{W_{n+1}}(W_n\boldsymbol{c}_n + w_{n+1}\boldsymbol{r}_{n+1}),$$

a convex combination of $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$ that lies on the line joining $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$. This implies

$$\boldsymbol{c}_{n+1} - \boldsymbol{c}_n = \frac{w_{n+1}}{W_{n+1}}(\boldsymbol{r}_{n+1} - \boldsymbol{c}_n),$$

so the distance between $\boldsymbol{c}_{n+1}$ and $\boldsymbol{c}_n$ is w_{n+1}/W_{n+1} times that between $\boldsymbol{r}_{n+1}$ and $\boldsymbol{c}_n$.

Example 3: Centroid of five points FIGURE 6 shows how this method yields the centroid of five points. Let $\boldsymbol{c}_k$ denote the centroid of the set of points $\{1, 2, \ldots, k\}$. The centroid $\boldsymbol{c}_1$ of point 1 is, of course, the point itself. Using (6) with $n = 1$ we find $\boldsymbol{c}_2$ is midway between 1 and 2. Now connect $\boldsymbol{c}_2$ with point 3 and divide the distance between them by 3 to find the centroid $\boldsymbol{c}_3$. Then connect $\boldsymbol{c}_3$ with point 4 and divide the distance between them by 4 to find $\boldsymbol{c}_4$. Finally, connect $\boldsymbol{c}_4$ with point 5 and divide the distance between them by 5 to get $\boldsymbol{c}_5$. It is clear that this method will work for any number of points.

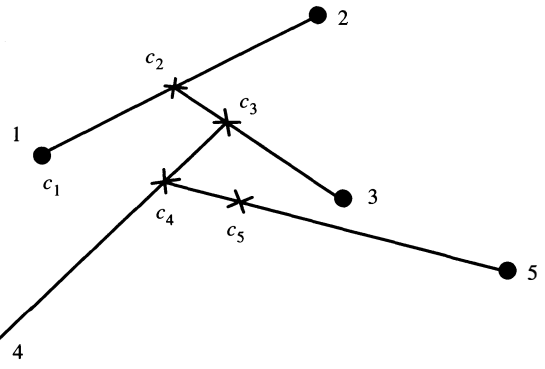


Figure 6 Distance between $\boldsymbol{c}_{k+1}$ and $\boldsymbol{c}_k$ is $1/(k+1)$ times the distance between $\boldsymbol{r}_{k+1}$ and $\boldsymbol{c}_k$

Variation of Method 2 using only bisection and connecting points A variation of Method 2 can be used by those who prefer not to divide vectors into more than

two equal parts. This construction uses only two geometric operations—bisection of segments and connecting points—so it applies only to the case of equal weights. We begin with a special example that constructs the centroid using only repeated bisection of segments.

Example 4: Centroid of four points The centroid of four points p, q, r, s, is given by $c = (p+q+r+s)/4$. By writing this in the form

$$c = \frac{1}{2}\left(\frac{p+q}{2} + \frac{r+s}{2}\right), \tag{7}$$

we see that the centroid is at the midpoint of the segment joining $(p+q)/2$ and $(r+s)/2$ which, in turn, are midpoints of the segments from p to q and from r to s. By permuting symbols in (7), we see that the centroid is also the midpoint of the segment joining $(s+p)/2$ and $(q+r)/2$, and of the segment joining $(p+r)/2$ and $(q+s)/2$. Two quadrilaterals with vertices p, q, r, s are shown in FIGURE 7, one convex and one not convex. In each case the segment joining p and r (shown dotted) is a diagonal of the quadrilateral with vertices p, q, r, s, as is the segment joining q and s. The centroid c lies midway between midpoints of the edges and of the diagonals of the quadrilateral. Any four of the six bisections shown are enough to determine the centroid.

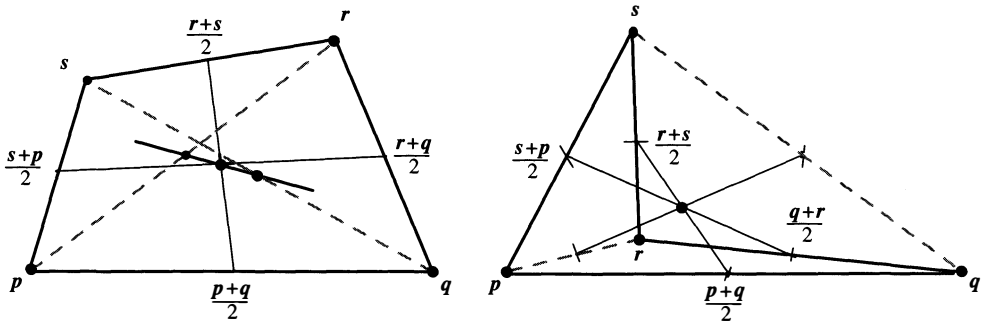


Figure 7 For each set of four points the centroid is midway between midpoints of edges and diagonals

If the four points are collinear, (7) shows that three bisections alone suffice to determine their centroid. An obvious iteration of (7) shows that, if $k \geq 1$, successive bisection suffices to find the centroid of any 2^k points, collinear or not. If the number of points is not a power of 2 we will show that again two operations, bisection of segments and connecting points, suffice to find their centroid. One of the principal tools used in this method is a general property of centroids originally formulated by Archimedes. A general version of this property was used by the authors [**1**] to find centroids of plane laminas. When adapted to finite sets of points, this property can be modified as follows:

ARCHIMEDES'S LEMMA. *If a finite set with centroid* c *is divided into two disjoint sets with centroids* c_1 *and* c_2*, then the three centroids are collinear. Moreover,* c *lies between* c_1 *and* c_2.

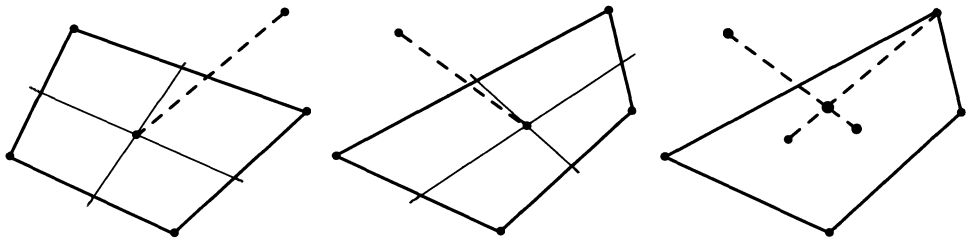
This is easily proved by the same method we used to obtain (6) in Method 2. Instead of (6) we get a formula of the form

$$c = \frac{1}{n_1 + n_2}(n_1 c_1 + n_2 c_2),$$

where c_1 is the centroid of n_1 points and c_2 is the centroid of n_2, points. This shows that c is a convex combination of c_1 and c_2 and hence lies on the line segment joining them.

It may interest the reader to reconsider Example 2 and see this lemma at work there. The next example shows how Archimedes' Lemma can be used to determine centroids of finite sets of points using only bisection of segments and connecting pairs of points.

Example 5: Centroid of five points FIGURES 8a and 8b show five points distributed in two ways as the union of four points and one point. In each case the centroid of the four points is found as in Example 4, so by Archimedes' Lemma the centroid of all five points lies on the dotted segment joining this centroid with the fifth point. FIGURE 8c shows the intersection of the two dotted segments giving the required centroid of the five points.



(a) (b) (c)

Figure 8 Centroid of five points found by using Archimedes' Lemma twice

In the special case when the five points are collinear, the method cannot give the actual centroid c because the intersecting lines are also collinear. But in this case we can adjoin a sixth point not on the common line and find the centroid c' of the six points by using the Archimedes Lemma twice. Three of the five collinear points together with the sixth point form a quadrilateral whose centroid c_1 can be found as in Example 4. The remaining two of the five collinear points have their centroid c_2 at their midpoint. By Archimedes' Lemma the centroid c' lies on the line joining c_1 and c_2. Now repeat the argument, using a different choice of three of the five collinear points to find another line containing c'. Then the line from the sixth point through c', intersects the line through the five points at their centroid c.

Now we have all the ingredients needed to show that centroids can be determined geometrically using only bisection of segments and connecting pairs of points. We state the result as a theorem, whose proof is constructive and outlines a variation of Method 2.

THEOREM. *For $n \geq 2$, the centroid of n points in m-space can be constructed using only bisection of segments and connecting distinct points.*

Proof. The proof is by induction on n. For $n = 2$ bisection suffices. For $n = 3$ we use bisection and drawing lines in the same manner as described in Example 5. For $n = 4$, bisection alone suffices as described in Example 4. Now suppose the theorem is true for n points, and consider any set of $n + 1$ points. Select one of the $n + 1$ points and join it to the centroid of the remaining n, points which, by the induction hypothesis, has been obtained by bisection of segments and connecting points. Repeat the process,

using a different choice for point $n + 1$. If all $n + 1$ points are not collinear, the two lines so obtained will intersect at their centroid. But if all the $n + 1$ points are collinear, choose an additional point outside this line and form, in two ways, a set of n points and a disjoint set of two points (as in Example 5) and apply the inductive procedure twice. This gives two lines whose point of intersection is the centroid, obtained by using only bisection of segments and connecting points. ■

Now we have several procedures at our disposal for finding centroids by an inductive method, two of which have been illustrated for five points. In Example 3 (FIGURE 6) we advanced one point at a time, and in Example 5 (FIGURE 8) we decomposed the set of five points in two different ways as the disjoint union of a single point and a set of four points. In general we can decompose a set of n points into two disjoint subsets in two different ways and use Archimedes' Lemma twice as was done in Example 5. The choice of subsets is a matter of preference, depending on the number of points.

Generalization of a Putnam problem We conclude with an extension of Problem A4 of the 29th William Lowell Putnam Mathematical Competition (1968), which asked to show that the sum of the squares of the $n(n-1)/2$ distances between any n distinct points on the surface of a unit sphere in 3-space is at most n^2. Several solutions of this problem are known, including one by the authors [**2**], using a method that reveals the natural role played by the centroid. The same method is used here to solve a more interesting and more general problem in m-space.

The generalized problem asks for the maximum value of the sum of squares of all distances among n points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ in m-space that lie on concentric spheres of radii $|\boldsymbol{r}_1| \le |\boldsymbol{r}_2| \le \cdots \le |\boldsymbol{r}_n|$. We can imagine a somewhat analogous problem in atomic physics where electrons move on concentric spheres and we ask to minimize the potential energy of the system, which requires minimizing the sum of reciprocals of the distances between charges. Here we wish to maximize the sum of the squares of the distances between points. In [**2**], we showed that this sum is related to their centroid $\boldsymbol{c}$ by the formula

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n \sum_{k=1}^{n} |\boldsymbol{r}_k|^2 - n^2 |\boldsymbol{c}|^2, \tag{8}$$

where $\sum_{k<i}$ is an abbreviation for the double sum $\sum_{i=1}^{n} \sum_{k=1}^{i-1}$. Using (8), we can easily maximize the sum on the left, because the right-hand side has its maximum value if and only if $|\boldsymbol{c}|$ reaches its minimum. In other words, locate the points so the centroid is as close as possible to the common center of the spheres. If the value $|\boldsymbol{c}| = 0$ is possible, then $\boldsymbol{c}$ is at the common center (which is chosen also as the origin $\boldsymbol{O}$, and this maximum value is n times the sum of the squares of the radii of the concentric spheres:

$$\max \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n \sum_{k=1}^{n} |\boldsymbol{r}_k|^2. \tag{9}$$

For example, in the original Putnam problem, each $|\boldsymbol{r}_k| = 1$, and by locating the points so their centroid is at the center, we find that (9) gives the required maximum:

$$\max \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n^2.$$

However, if the points are required to lie on different spheres, the problem of locating them to maximize the sum of squares is more difficult because it is not always possible to place their centroid at the common center. But it can be solved by the graphical method for finding the centroid by closing a polygon. This solution gives a constructive approach as well as a visual interpretation of the results. The solution splits naturally into two cases, depending on how the largest radius $|\boldsymbol{r}_n|$ compares with the sum of all other radii.

Case 1. $|\boldsymbol{r}_n| < \sum_{k=1}^{n-1} |\boldsymbol{r}_k|$, $n \geq 3$. When $n = 3$, this is the triangle inequality, and for $n > 3$ it is a polygonal inequality that makes it possible to choose the vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ to have sum zero. In this case we place the origin at the common center of the spheres, and connect the vectors with hinges, head to tail, to form a closed polygon. Because the vectors joining successive edges have sum zero, they can be translated so all initial points are at the origin. The terminal points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ will lie on concentric spheres of radii $|\boldsymbol{r}_1|, |\boldsymbol{r}_2|, \ldots, |\boldsymbol{r}_n|$, their centroid will be at the common center, and the points will satisfy the maximal sum relation (9).

If $n = 3$ the points are vertices of a rigid triangle. But if $n > 3$, there are infinitely many incongruent solutions represented by closed flexible polygons. Any one of these shapes provides a solution: translate each vector parallel to itself to bring all tails to a common point, the common center of the spheres.

Case 2. $|\boldsymbol{r}_n| \geq \sum_{k=1}^{n-1} |\boldsymbol{r}_k|$, $n \geq 2$. In this case the vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ cannot have sum zero unless $|\boldsymbol{r}_n| = \sum_{k=1}^{n-1} |\boldsymbol{r}_k|$ and the vectors are on a line. In general, we get the largest possible sum of squares of distances from each other by arranging the vectors along a straight line, with the first $n - 1$, vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{n-1}$ pointing in the same direction, and the nth vector $\boldsymbol{r}_n$ (with the largest radius) pointing in the diametrically opposite direction. Unlike in Case 1, this solution is unique; it gives the largest possible sum of squares of distances from each other consistent with (8), but this largest sum will not reach the maximum provided by the right-hand side of (9) because the centroid is not at the origin.

Where is the centroid? Using the guess $\boldsymbol{g} = \boldsymbol{O}$, we find the deficiency vector $\boldsymbol{d}$ in this case is $\boldsymbol{d} = \sum_{k=1}^{n} \boldsymbol{r}_k = n\boldsymbol{c}$, hence $\boldsymbol{c} = \boldsymbol{d}/n$. Therefore (8) implies that the maximum is given by

$$\max \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n \sum_{k=1}^{n} |\boldsymbol{r}_k|^2 - |\boldsymbol{d}|^2, \quad \text{where} \quad |\boldsymbol{d}| = |\boldsymbol{r}_n| - \sum_{k=1}^{n-1} |\boldsymbol{r}_k|,$$

because the vectors are along a line.

As in the original Putnam problem, it is surprising that the maximum in both cases is independent of the dimensionality m of the space if $m \geq 2$. Any solution in one common equatorial plane of the spheres (that is, for $m = 2$) is also a solution in all higher-dimensional spaces.

REFERENCES

1. T. M. Apostol and M. A. Mnatsakanian, Finding centroids the easy way, *Math Horizons*, Sept. 2000, 7–12.
2. ———, Sums of squares of distances in m-space, *Amer. Math. Monthly* **110**, (2003), 516–526.

There Are Only Nine Finite Groups of Fractional Linear Transformations with Integer Coefficients

GREGORY P. DRESDEN
Washington & Lee University
Lexington, VA 24450
dresdeng@wlu.edu

An introduction to fractional linear transformations A *fractional linear transformation* (also called a *Möbius transformation*) over **C** is a function of the form

$$m(x) = \frac{ax + b}{cx + d},$$

with $ad - bc \neq 0$. Most of us first encountered these in our complex analysis class, where we learned that such analytic functions map lines and circles to lines and circles on the complex plane (see, for example, books by Fisher [**3**, p. 187] or Rudin [**6**, p. 280]). In this note, we consider finite groups of fractional linear transformations (where the group operation is composition). We will arrive at the interesting conclusion that, provided we limit ourselves to integer coefficients, there are only nine such groups up to isomorphism. (For real or complex coefficients, there are infinitely many such groups, but we will get into that a little bit later.)

All of this material should be accessible to undergraduates; indeed, I've even had my beginning calculus students play with these functions when learning about compositions and inverses. Students in abstract algebra might appreciate these groups of functions as nice examples of cyclic and dihedral groups, and those interested in finding appropriate research topics will find plenty of material here to explore.

Let's examine our terms and definitions in a little more detail. By *integer coefficients*, of course, we mean that a, b, c, and d are all integers. Thus, we're considering functions of the form

$$m(x) = \frac{2x + 3}{4x + 5} \quad \text{or even} \quad p(x) = \frac{5}{6}x + 7,$$

since this can be written as

$$p(x) = \frac{5x + 42}{0x + 6},$$

but we are not considering functions like $q(x) = \frac{5}{6}$ (as here, $ad - bc = 0$). Let's also observe that the two fractional linear transformations

$$\frac{ax + b}{cx + d} \quad \text{and} \quad \frac{arx + br}{crx + dr} \quad (r \neq 0)$$

are identical. With this in mind, we see that there is no need to concern ourselves with fractional linear transformations with *rational* coefficients, as multiplying top and bottom by a common denominator would give us an identical function with integer coefficients (indeed, relatively-prime integer coefficients, if desired).

As mentioned above, our proposed group operation is function composition, so that, for

$$m(x) = \frac{2x+3}{4x+5}, \quad \text{say, and} \quad p(x) = \frac{6x+7}{8x+9},$$

we will have occasion to form $m \circ p(x) = m(p(x))$. It should be clear that these fractional linear transformations really do form a group under composition; the identity is $e(x) = x$ and the inverse of

$$\frac{ax+b}{cx+d} \quad \text{is} \quad \frac{dx-b}{-cx+a},$$

for which the condition $ad - bc \neq 0$ is required.

The composition of two fractional linear transformations is somewhat tedious to compute. A nice short-cut is provided by the map

$$\phi : \frac{ax+b}{cx+d} \longmapsto \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

It is surprising, but not to hard to check, that composition of functions corresponds exactly to matrix multiplication. In fact, ϕ is an isomorphism from the group of fractional linear transformations with integer coefficients to a group called $PGL(2, \mathbf{Q})$, the projective group of 2×2 matrices with rational entries. The word *projective* simply means that the matrices

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} ar & br \\ cr & dr \end{bmatrix}$$

are considered identical. (This condition is needed if the inverse of

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{is to be} \quad \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.)$$

Thus, as with the fractional linear transformations, we need only consider matrices with integer coefficients if we so desire.

The isomorphism ϕ makes it easy to compose functions; it's much simpler to multiply the matrix $\left[\begin{smallmatrix} 2 & 3 \\ 4 & 5 \end{smallmatrix}\right]$ by $\left[\begin{smallmatrix} 6 & 7 \\ 8 & 9 \end{smallmatrix}\right]$ than to attempt to simplify

$$\frac{2\left(\frac{6x+7}{8x+9}\right)+3}{4\left(\frac{6x+7}{8x+9}\right)+5}.$$

The isomorphism also gives us a shorthand for referring to our group of fractional linear transformations with integer coefficients (which we will henceforth denote as $PGL(2, \mathbf{Q})$.)

As is common in complex analysis, we will occasionally have these fractional linear transformations operate on numbers in the extended complex plane $\widehat{\mathbf{C}} = \mathbf{C} \bigcup \{\infty\}$ (also called the *Riemann sphere*). This is easy to do if we agree that $1/0 = \infty$, that $1/\infty = 0$, and that other standard rules of arithmetic with ∞ apply (see Beardon [**1**, p. 4] or Rudin [**6**, p. 279] for more). In particular, if

$$m(x) = \frac{ax+b}{cx+d},$$

let us agree that $m(\infty) = a/c$ and $m(-d/c) = \infty$ (but note that if $c = 0$ then both these equations read as $m(\infty) = \infty$).

A few special groups Let us define two types of (multiplicative) groups that will prove to be quite important to us. The *cyclic group* C_n of size n can be thought of as the set $C_n = \{e, a, a^2, a^3, \cdots a^{n-1}\}$ such that $a^i \cdot a^j = a^{i+j \bmod n}$ and $a^n = a^0 = e$. The *dihedral group* D_n is a bit more complicated; we write $D_n = \{e, a, a^2, \cdots a^{n-1}, b, ab, a^2b, \cdots a^{n-1}b\}$. The as behave as before, but we now have $b^2 = e$ and $ba = a^{n-1}b$ (equivalently, $b = b^{-1}$ and $b^{-1}ab = a^{-1}$). See Gallian's book [**4**, p. 442] for further discussion. Note that D_1 is simply $\{e, b\}$ and thus is isomorphic to C_2. The group $D_2 = \{e, a, b, ab\}$ is the smallest noncyclic group, and is often referred to as the *Klein four-group*. (The terms D_1 and D_2 are not standard nomenclature for these small groups, but they do help to simplify our notation.)

C_n and D_n can also be thought of as *symmetry groups* for certain geometric objects; indeed, for $n \geq 3$, C_n and D_n are isomorphic to the group of symmetries in $\mathbf{R}^3$ of a regular pyramid and regular prism, respectively, with regular n-gons for bases. Since these pyramids and prisms can each be inscribed within a sphere, we see that C_n and D_n can be thought of as a symmetry group for a sphere as well. A classical result [**8**, p. 40] states that the only finite symmetry groups of the sphere are C_n, D_n, A_4, S_4, and A_5 (where A_n and S_n represent the alternating and the symmetric groups on n letters). See also Toth's book [**10**, chapter 17] for a discussion of A_4, S_4, and A_5 as symmetry groups for the five platonic solids.

The cyclic groups C_n and the dihedral groups D_n also have a natural interpretation as symmetries of regular n-gons in the plane. Here, C_n is the group of pure rotations of an n-gon about its center, and D_n is the group of such rotations, along with the reflections about axes of symmetries (*dihedral* groups refer to the "two sides" of the polygon). Although these interpretations are well known and of interest in their own right, we are mostly concerned here with regarding these groups as symmetries of appropriate polyhedra in three dimensions.

Groups of fractional linear transformations under composition Let us consider examples to see how certain sets of these fractional linear transformations can form groups under composition. First, for $p(x) = 1/(x+1)$, then we have $p(p(x))$ (also written $p \circ p(x)$ or $p^{(2)}(x)$) equals $(x+1)/(x+2)$, and $p^{(3)}$, $p^{(4)}$, and $p^{(5)}$ are

$$\frac{x+2}{2x+3}, \frac{2x+3}{3x+5}, \quad \text{and} \quad \frac{3x+5}{5x+8},$$

respectively. Are you surprised to see the Fibonacci numbers appearing as coefficients? Note that $p(x)$ corresponds to the matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} f_0 & f_1 \\ f_1 & f_2 \end{bmatrix}$$

in $PGL(2, \mathbf{Q})$, where f_i is the ith Fibonacci number. It is well known that this matrix generates the Fibonacci sequence: Its nth power is

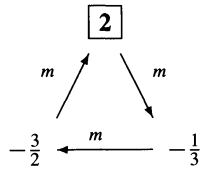
$$\begin{bmatrix} f_{n-1} & f_n \\ f_n & f_{n+1} \end{bmatrix}.$$

Let us define $p^{(0)}$ to be the the identity map $p^{(0)}(x) = x$, and note that the inverse of $p(x)$ is clearly $p^{(-1)}(x) = (-x+1)/x$. Thus, the set $\{p^{(i)}(x) : i \in \mathbf{Z}\}$ under composition forms an (infinite) group, isomorphic to $\mathbf{Z}$. A nice way to visualize the behavior of this group is to observe the orbit of a particular number (say, 1) under repeated iterations of p (and p^{-1}):

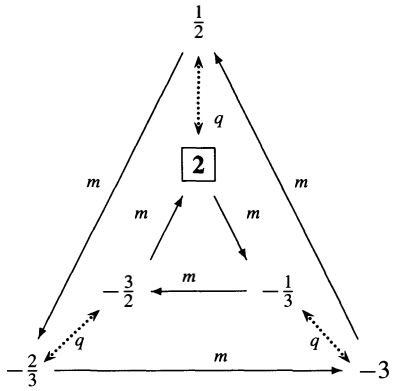
$$\cdots \xleftarrow[p^{-1}]{} -\tfrac{3}{2} \xleftarrow[p^{-1}]{} -2 \xleftarrow[p^{-1}]{} -1 \xleftarrow[p^{-1}]{} \infty \xleftarrow[p^{-1}]{} 0 \xleftarrow[p^{-1}]{} \boxed{1} \xrightarrow{p} \tfrac{1}{2} \xrightarrow{p} \tfrac{2}{3} \xrightarrow{p} \tfrac{3}{5} \xrightarrow{p} \tfrac{5}{8} \xrightarrow{p} \cdots$$

Again, we notice the appearance of the Fibonacci numbers, this time in the numerators and denominators of the terms in the orbit (the set of images of 1).

Next, for the similar function $m(x) = -1/(x+1)$, we find that $m^{(2)}(x)$ equals $(-x-1)/x$ and $m^{(3)}(x) = x$, the identity function. We say that m has order three under composition, and so $m(x)$ generates a finite group (isomorphic to C_3, the cyclic group of three elements). The following diagram shows the orbit of 2 under iteration by m:



Finally, consider $q(x) = 1/x$. Since $q^{(2)}(x)$ is the identity, q generates a group of order two. Of greater interest is the group generated by q and m together; these satisfy the relation $q \circ m = m^2 \circ q$, and so they generate a group of size six, isomorphic to the dihedral group D_3. The following picture illustrates a typical orbit under the group generated by $q(x)$:



This is the structure we are interested in: a set of fractional linear transformations, with integer coefficients, that forms a finite group under composition.

The nine finite groups Recall from our earlier discussion that $PGL(2, \mathbf{Q})$ represents the group of fractional linear transformations $(ax+b)/(cx+d)$ with integer coefficients such that $ad - bc \neq 0$. The following theorem gives our main result:

THEOREM 1. *All finite subgroups of* $PGL(2, \mathbf{Q})$ *are isomorphic to* C_n *or* D_n *for* $n = 1, 2, 3, 4,$ *or* 6.

Since $D_1 = C_2$, we see that there are actually just nine groups, as mentioned in the title. However, if we allow real coefficients in our fractional linear transformations (equivalently, real entries in our projective matrices), we can extend the list of possible finite groups:

THEOREM 2. *Every finite subgroup of* $PGL(2, \mathbf{R})$ *is either cyclic or dihedral, and there exist such subgroups of arbitrary order.*

Our proofs are constructive; see Corollary 1 for an explicit example of an element of $PGL(2, \mathbf{R})$ of arbitrary finite order under composition.

We've generalized from rational to real coefficients; what happens if we allow complex coefficients? Here we refer to a well-known result: any finite group of fractional linear transformations with complex coefficients (that is, a finite subgroup of

$PGL(2, \mathbf{C})$) is isomorphic to a finite group of symmetries of the sphere [**5**], and hence (as mentioned earlier) is isomorphic to a cyclic group C_n, a dihedral group D_n, or one of the symmetry groups A_4, S_4, and A_5 of the tetrahedron, cube, and icosahedron, respectively [**8**]. However, our Theorem 2 indicates that it is impossible to represent A_4, S_4, and A_5 using projective matrices $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ with real coefficients, since these groups are neither cyclic nor dihedral.

Constructing fractional linear transformations of finite order For the moment, let us allow complex coefficients, and let us consider how to construct fractional linear transformations of small order. Clearly, $e(x) = x$ has order 1 and $m(x) = -x$ has order 2 (under composition). It's interesting to note, however, that $-x + b$ also has order 2, for b any complex number. For order 4, we might guess at ix or even $ix + b$, and a moment's work shows that these both work.

The coefficients 1, -1, and i are all *primitive nth roots of unity* for $n = 1$, 2, and 4 respectively. Recall that for n a positive integer, we call ζ a primitive nth root of unity if $\zeta^n = 1$ and $\zeta^k \neq 1$ for every $0 < k < n$. Thus, ζ is a root of $x^n - 1$, and for $n > 1$ it is also a root of the polynomial $(x^n - 1)/(x - 1) = x^{n-1} + \cdots + x^2 + x + 1$. The only rational roots of unity are ± 1; the only quadratic roots are $\pm i$ and $\pm 1/2 \pm i\sqrt{3}/2$ (of orders 3, 4, and 6). All other roots of unity are cubic, quartic, or of even higher degree over the rationals [**4**, chapter 33]. For more information on field extensions and on degrees of algebraic numbers over $\mathbf{Q}$, see Gallian [**4**, chapters 20–21]. We note that one must be careful not to confuse the *order* of a primitive root of unity with its *degree* as an algebraic number over $\mathbf{Q}$, nor should we confuse it with the *order* of a fractional linear transformation under composition. These are all distinct concepts.

Based on our earlier work, we might suspect that the linear transformation $m(x) = \zeta x + b$ (with ζ, b complex) will have order n under composition if and only if ζ is a primitive nth root of unity. This is in fact the case, as can be seen by noticing that $m^{(n)}(x) = \zeta^n x + b(\zeta^{n-1} + \cdots + \zeta^2 + \zeta + 1)$. Thus, if ζ is a primitive nth root of unity, then $m(x)$ has order n, and vice versa.

How do we now proceed to find arbitrary fractional linear transformations of finite order? As it turns out, we can show that any fractional linear transformation $(ax + b)/(cx + d)$ is conjugate to some linear map $Ax + B$, and we now know exactly what is necessary for these linear maps to have finite order. (Recall that one defines two elements α, β of a group G to be *conjugate in* G if there is some $t \in G$ such that $\beta = t^{-1}\alpha t$. It is easy to show that conjugate elements have the same order.) Our first lemma is as follows:

LEMMA 1. *Suppose*

$$m(x) = \frac{ax + b}{cx + d},$$

with rational coefficients, $ad - bc \neq 0$, *has finite order under composition. Then,* $m(x)$ *has order* 1, 2, 3, 4, *or* 6, *and in the last three cases,* $m(x)$ *is conjugate (in* $PGL(2, \mathbf{Q})$*) to*

$$\frac{-1}{x+1}, \frac{x-1}{x+1}, \quad \text{or} \quad \frac{2x-1}{x+1},$$

respectively.

REMARK. The order-2 maps are not all conjugate in $PGL(2, \mathbf{Q})$. In particular, for $p(x) = -4/x$, then $p(x)$ is not conjugate to $q(x) = 1/x$ using only transformations with rational coefficients. This can be seen by trying to find $s(x)$ such that $q = s^{-1} \circ p \circ s$; after much calculation, we find that $s(x)$ can only be $\pm 2ix$.

Proof. It's easy to verify that the three examples given in the lemma have order 3, 4, and 6. Note also that $q(x) = 1/x$ has order 2, and of course the identity map $e(x) = x$ has order 1.

We now show that 1, 2, 3, 4, and 6 are the only orders possible. Suppose that $m(x)$ is as above, with order $n < \infty$. If $c = 0$, then we can write $m(x) = (a/d)x + b/d$, a linear map, and so by our earlier discussion, $a/d = \pm 1$ (the only rational roots of unity) and $m(x)$ has order $n = 1$ or $n = 2$. If $c \neq 0$, we can solve the equation

$$\frac{ax + b}{cx + d} = x$$

to get at least one finite fixed point α for $m(x)$, of degree ≤ 2 over $\mathbf{Q}$. We conjugate $m(x)$ with $s(x) = \alpha + 1/x$ to get $\widehat{m}(x) = s^{-1} \circ m \circ s(x)$, a fractional linear transformation with coefficients in the (possibly quadratic) field $\mathbf{Q}(\alpha)$. Since s takes ∞ to α and m fixes α, then $\widehat{m}(\infty) = \infty$, which implies $\widehat{m}(x)$ is actually strictly linear, of the form $Ax + B$. Thus, since $\widehat{m}$ has the same (finite) order as m, then A is a primitive nth root of unity. Since $A \in \mathbf{Q}(\alpha)$ and since $\mathbf{Q}(\alpha)$ is at worst quadratic over $\mathbf{Q}$, then $A = \pm 1, \pm i$, or $\pm 1/2 \pm i\sqrt{3}/2$. Thus, $n = 1, 2, 3, 4$, or 6.

We finish by showing that, when $n = 3$, 4, or 6, $m(x)$ is conjugate to one of the transformations from our list in the above lemma. For $m(x)$ of order ≥ 3, choose three rational numbers A, B, and C such that $m(x) : A \mapsto B \mapsto C$. Let $s(x)$ be the fractional linear transformation that takes $0 \mapsto A$, $-1 \mapsto B$, and $\infty \mapsto C$ (note that this $s(x)$ will have rational coefficients). Then, for $\widehat{m}(x) = s^{-1} \circ m \circ s(x)$, we have that $\widehat{m}(x) : 0 \mapsto -1 \mapsto \infty$. This implies $\widehat{m}(x)$ has the form $(\widehat{a}x - 1)/(x + 1)$ for some $\widehat{a} = \widehat{m}(\infty)$, and it's now easy to show that the order of $\widehat{m}(x)$ being 3, 4, or 6 forces $\widehat{a}$ to equal 0, 1, or 2, respectively. ∎

In Lemma 1, we can't help but notice the suggestive pattern exhibited by the fractional linear transformations of orders 3, 4, and 6. Each has the form $(ax - 1)/(x + 1)$, with a from 0 to 2, and so we might think that $(3x - 1)/(x + 1)$ would also have finite order. Sadly, that isn't so; the sequence that describes the as is actually $a_n = 1 + 2\cos(2\pi/n)$, as we will see later.

The following lemma states that if we use only *real* coefficients, then we cannot represent A_4, S_4, or A_5 using fractional linear transformations. This is the last step; after this lemma, we can proceed directly to the proof of our two theorems.

LEMMA 2. *If K is a subfield of $\mathbf{R}$, then all finite subgroups of the set of fractional linear transformations with coefficients in K (equivalently, all finite subgroups of $PGL(2, K)$) are either cyclic or dihedral.*

Proof. By results of Lyndon and Ullman [**5**], we need only show that A_4, S_4, and A_5 cannot be realized in $PGL(2, K)$. We will show the impossibility of A_4; since $A_4 \subset S_4$ and $A_4 \subset A_5$, then this implies that S_4 and A_5 cannot be realized either.

Suppose we have a subgroup G isomorphic to A_4. Since A_4 contains elements of order 3, we can assume by Lemma 1 that (after an appropriate conjugation) $-1/(x + 1) \in G$. We note that the product of an element of order 3 (in A_4) with something of order 2 gives us a new element of order 3 (as is easily seen by considering products such as (abc) with $(ab)(cd)$; see Gallian again [**4**, p. 104] for a complete multiplication table for A_4). However, every element of order 2 in $PGL(2, K)$ has either the form $-x + b$ (if it fixes ∞) or $(ax + b)/(x - a)$ (if it takes ∞ to some finite a). Now, $-1/(x + 1)$ composed with $-x + b$ is $1/(x - b - 1)$, which has order 3 only for $b = -1 \pm i$. And, $-1/(x + 1)$ composed with

$$\frac{ax + b}{x - a} \quad \text{is} \quad \frac{-x + a}{(a + 1)x + (b - a)},$$

and a few minutes of algebra will show that this has order 3 only for $b = \frac{1}{2}(1 + 2a \pm \sqrt{-4a^2 - 4a - 3})$, a complex number for all real values of a. Thus, A_4 cannot be realized using only real coefficients. ■

Proof of theorems We are now ready to prove our main theorems, which we restate for convenience:

THEOREM 1. *All finite subgroups of PGL*(2, **Q**) *are isomorphic to* C_n *or* D_n *for* $n = 1, 2, 3, 4,$ *or* 6.

THEOREM 2. *Every finite subgroup of PGL*(2, **R**) *is either cyclic or dihedral, and there exist such subgroups of arbitrary order.*

Proof. By Lemma 2, we only need concern ourselves with cyclic and dihedral subgroups. To show $G = PGL(2, \mathbf{R})$ contains all cyclic and dihedral groups, we will show that for every $n > 0$, there exists an element $m(x)$ of order n such that for $q(x) = 1/x$, then $m(x)$ and $q(x)$ generate the dihedral group D_n (with C_n as a cyclic subgroup).

For $n = 1$, let $m(x) = x$, and for $n = 2$, let $m(x) = -x$. These clearly give us C_1, C_2 and (with $q(x) = 1/x$) D_1 and D_2. Now assume $n \geq 3$. Let ζ_n be a primitive nth root of unity, in particular, $e^{2\pi i/n}$, let $a_n = 1 + 2\cos(2\pi/n) = 1 + \zeta_n + 1/\zeta_n$ (a real number), and define $m(x)$ as $(a_n x - 1)/(x + 1)$. It's not hard to show that

$$m^{-1}(x) = \frac{x + 1}{-x + a_n}$$

and that $q^{-1} \circ m \circ q = m^{-1}$. We need only show that $m(x)$ has order n to exhibit our dihedral group D_n. A clever way to do this is to define $\widehat{m} = s^{-1} \circ m \circ s$ with $s(x) = \zeta_n + 1/x$; since $s(\infty) = \zeta_n$ and $m(\zeta_n) = \zeta_n$, we have that $\widehat{m}(\infty) = \infty$, and so $\widehat{m}(x) = Ax + B$ for some $A, B \in \mathbf{C}$. By comparing $\widehat{m}(x)$ and $s^{-1} \circ m \circ s(x)$ for $x = 0$ and $x = -1/\zeta_n$, we find that $A = \zeta_n$ and $B = \zeta_n/(\zeta_n + 1)$. By a previous discussion, $\widehat{m}(x)$ (and hence $m(x)$ itself) has order n. This proves Theorem 2.

If we now restrict ourselves to rational coefficients, then our discussion above, combined with Lemma 1, proves Theorem 1. (Note that only for $n = 3$, 4, or 6 does the element

$$m(x) = \frac{a_n x - 1}{x + 1}$$

have rational coefficients; these are exactly the elements mentioned in Lemma 1.) ■

The existence of a real fractional linear transformation of any order, as demonstrated in this proof, deserves its own statement:

COROLLARY 1. *For* $n \geq 3$ *and* $a_n = 1 + 2\cos(2\pi/n)$, *then* $(a_n x - 1)/(x + 1)$ *has order n under composition.*

For further study This article just touches the surface of the many fascinating topics associated with groups of fractional linear transformations and groups of matrices with restricted entries. For example, a great deal of attention has been paid to the *modular group* (denoted $\Gamma(1)$), which is the set

$$\left\{ \frac{ax + b}{cx + d} : ad - bc = 1, \quad a, b, c, d \in \mathbf{Z} \right\}.$$

This infinite group is generated by two noncommuting elements: $p(x) = -1/x$ and our old friend $m(x) = -1/(x + 1)$ [**9**]. One could also venture into representation

theory, which (loosely stated) asks which groups can be represented by matrices in $GL(V)$ for V a complex vector space [**7**]. Then there is the fascinating theory of iterations of rational functions. Ours are the quotients of linear polynomials and thus rather simple, but if one considers quotients of polynomials of degree greater than or equal to 2, one begins to venture into the rich area of complex dynamics (see Beardon [**1**], and also Devaney [**2**] for general analytic functions). As a single example, one can show that any rational function of degree 5 in numerator and denominator (even with just integer coefficients!) has periodic points (under iteration) of all orders (see Beardon [**1**, Theorem 6.2.2]).

Finally, there are a few open questions suggested by this article, such as: for which algebraic number fields K will all isomorphic finite groups in $PGL(2, K)$ actually be conjugate in $PGL(2, K)$? This is certainly true for $K = \mathbf{C}$ (see Shurman [**8**], and also Theorem 2.6.1 in the paper by Lyndon and Ullman [**5**]), and certainly not true for $K = \mathbf{Q}$ (as seen in the remark following Lemma 1, above). Most likely, more can be said. One might also ask how many nonconjugate groups (isomorphic to D_3, say) are in $PGL(2, \mathbf{Q})$, and if there is a way to describe or index them all.

REFERENCES

1. Alan F. Beardon, *Iteration of Rational Functions*, vol. **132** of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1991,
2. Robert L. Devaney, *An Introduction to Chaotic Dynamical Systems*, 2nd ed., Addison-Wesley Studies in Nonlinearity, Addison-Wesley Publishing Company Advanced Book Program, Redwood City, CA, 1989.
3. Stephen D. Fisher, *Complex Variables*, Brooks/Cole, Monterey, CA, 1986.
4. Joseph A. Gallian, *Contemporary Abstract Algebra*, 5th ed., Houghton Mifflin, Boston, MA, 2002.
5. R. C. Lyndon and J. L. Ullman, Groups of elliptic linear fractional transformations, *Proc. Amer. Math. Soc.* **18** (1967), 1119–1124.
6. Walter Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1987.
7. Jean-Pierre Serre, *Linear Representations of Finite Groups*, vol. **42** of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1977. Translated from the second French edition by Leonard L. Scott.
8. Jerry Shurman, *Geometry of the Quintic*, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1997.
9. Joseph H. Silverman, *Advanced Topics in the Arithmetic of Elliptic Curves*, vol. **151** of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1994.
10. Gabor Toth, *Glimpses of Algebra and Geometry*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1998.

Path Representation of One-Dimensional Random Walks

OSCAR BOLINA
Departamento de Física-Matemática
Universidade de São Paulo
Caixa Postal 66318
São Paulo 05315-970 BRASIL
bolina@if.usp.br*

Imagine a particle moving on the x axis, starting at some initial position $x = k, k > 0$, and moving back and forth on the line by random steps of unit length. Suppose that the particle moves right with probability p and left with probability $1 - p$, where, of course, $0 < p < 1$.

*Present address: Department of Mathematics, University of California, Davis, CA, 95616-8633.

This describes what is usually called a *random walk*, an example of a *stochastic process*. Random walks are often introduced as motions that mimic a staggering drunk who tries to walk away from a lamp-post on the street [**1**]. This image reminds us that the particle's next step, like the drunk's, is independent of the steps it took before. The same image helps us introduce our stopping condition. We suppose that the origin is like the drunk's home and that the walk stops if it ever arrives there.

In some cases, it is probable that the walk never stops, leading us to ask: What is the probability P_k that the particle eventually reaches the origin, given that it starts at $x = k$?

In this note, we answer the question with a simple formula that involves only P_1, the probability that the particle ends up at the origin given that it starts at $x = 1$:

$$P_k = P_1^k, \tag{1}$$

where

$$P_1 = \begin{cases} \frac{1-p}{p} & \text{if } p > \frac{1}{2} \\ 1 & \text{if } p \le \frac{1}{2}. \end{cases} \tag{2}$$

We will prove this for $k = 1$, 2, and 3 by means of a geometric representation of the one-dimensional random walk in terms of paths on a two-dimensional lattice. Then we will use induction on k to establish (1) for all values of k.

The good thing about path representations is that they transform hard algebraic problems in easier combinatorial ones. They give us a good way to see what's going on.

The one-dimensional random walk has been analyzed in a variety of different ways. There are also many related, equivalent formulations of this problem. The *gambler's ruin* problem is one of them [**2**]: Players A and B have initially $\$x$ dollars and $\$s - x$ dollars, respectively. They play a series of independent games in which at each play A wins \$1 with probability p and loses \$1 to B with probability $1 - p$. What is the probability that A's fortune will be zero before it is s? The one-dimensional random walk is closely related to the gambler's ruin problem when $s \to \infty$.

The path representation Here is how to turn any particular walk into a two-dimensional lattice path: Start at the origin of the standard Cartesian plane, draw a horizontal segment for each step to the right, a vertical unit segment for each step to the left. How can we express the condition that the particle reaches the origin? If the particle started at $x = k$ and never took any steps to the right, our path would go straight up the y-axis to the point $(0, k)$. In general, the path ends the first time it touches the line L with equation $y = x + k$, because then the number of steps to the right is just k more than the number of steps left.

For now, fix positive integers k and n. If a particle takes n steps to the right before reaching the origin, then the total number of steps to the left is $n + k$. A *lattice path of* $2n + k$ *steps*—or, for our purposes, simply a *path*—is a sequence of horizontal and vertical edges that lead from the origin to the point with coordinates $(n, n + k)$ on the line L, with the added requirement that this is the *first* time the path touches L. The set of such paths of $2n + k$ steps is depicted in FIGURE 1. Actually, the drawing shows $k = 1$, but the labels are general, which is not too confusing, since figures for higher values of k are drawn by raising the line L.

To find the probability that the particle ends up at the origin, let us first find the probability that it ends at the origin given that it starts at $x = k$ and takes n steps to the right. The probability of each step to the right is p; since the steps are independent,

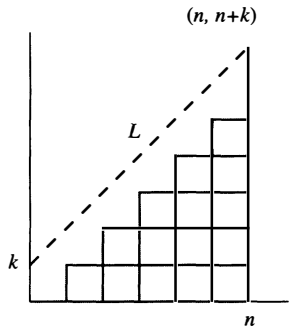


Figure 1 Each lattice path from the origin to the point $(n, n+k)$ represents a way of taking n steps to the right and $n+k$ steps to the left, starting at $x = k$ on the x-axis and ending at the origin. (Note that $k = 1$ in the diagram.)

the probability that the particle takes n steps to the right is p^n. Similarly, the probability that it takes $n+k$ steps to the left is $(1-p)^{n+k}$. Thus, the probability of the walk represented by any particular lattice path of $2n+k$ steps is $p^n(1-p)^{n+k}$. We must multiply this probability times the number of such lattice paths and then add the probabilities for different values of n.

Let $C_k(n)$ be the total number of paths of $2n+k$ steps from the origin to point $(n, n+k)$. You may recognize that $C_1(n)$ has a famous name—the nth Catalan number [**3**]—but we will need to know very little about the numbers $C_k(n)$. Using this notation, the probability we seek can be written as

$$P_k = \sum_{n=0}^{\infty} C_k(n) p^n (1-p)^{n+k}. \tag{3}$$

We will show how to carry out the summation in (3) using the geometry of the lattice paths. We will solve for P_k for $k = 1$, 2, and 3 before showing the general formula in (1) by induction on k. In fact, for the induction, all we need to know is that $C_k(0) = 1$ for any k, which is obvious since there is only one path consisting of one or more edges along the vertical axis in FIGURE 1, and that $C_1(1) = 1$, since there is exactly one path consisting of only one horizontal edge.

The case $k = 1$ Let us use our path representation to compute first the probability that the particle ends up the origin given that it starts at $k = 1$. In this case we write

$$P_1 = \sum_{n=0}^{\infty} C(n) p^n (1-p)^{n+1}, \tag{4}$$

where $C(n)$ is a shorthand for $C_1(n)$. This is consistent with the usual notation, $C(n)$, for the nth Catalan number, which is known [**3**] to be

$$C(n) = \frac{1}{n+1}\binom{2n}{n}.$$

The first few Catalan numbers are $C(0) = 1$, $C(1) = 1$, $C(2) = 2$, $C(3) = 5$, $C(4) = 14$, $C(5) = 42$, $C(6) = 132$, $C(7) = 429$. The Catalan numbers appear frequently in apparently unrelated problems in combinatorics [**3, 6**]. They have some geometric properties associated with lattice paths that we now explore. Our first result is the following.

THEOREM 1. *The Catalan numbers satisfy the following equation*

$$C(n) = \sum_{\alpha=1}^{n} C(\alpha - 1)\, C(n - \alpha). \tag{5}$$

Heuristic proof. We illustrate the theorem with $n = 4$. A walk that starts at $x = 1$ with 4 right steps must return to $x = 1$ before it reaches the origin. The first time this occurs could be after it has taken 1, 2, 3, or 4 right steps. These are the values of the counter α. FIGURE 2 shows the possibilities: For instance, if the particle returns to $x = 1$ after a single step to the left, then it has 3 right steps remaining, which it must use to reach 0. This is exactly the number of possibilities counted by $C(3)$.

In each part of the figure, the dashed paths show the paths that return to $x = 1$ after α steps to the right. The solid paths show the ways of reaching 0 in $4 - \alpha$ steps. There are $C(\alpha)$ dashed paths on the left, and $C(4 - \alpha)$ solid paths on the right. This proves (5) in the special case. The general proof is similar. ■

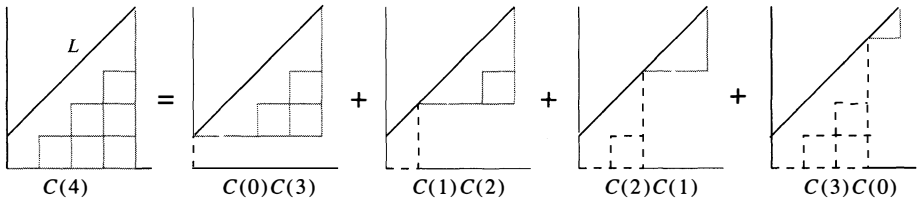


Figure 2 Decomposition of the Catalan numbers

To calculate P_1 we define

$$F(z) = \sum_{\alpha=1}^{\infty} C(\alpha - 1)\, z^{\alpha}, \tag{6}$$

which you may recognize as a generating function, and observe that

$$F^2(z) = \sum_{\alpha,\beta=1}^{\infty} C(\alpha - 1)C(\beta - 1)\, z^{\alpha+\beta} = \sum_{n=2}^{\infty} \sum_{\alpha=1}^{n-1} C(\alpha - 1)C(n - \alpha - 1)\, z^{n},$$

where we have set $n = \alpha + \beta$. Making use of formula (5) we see that the sum over α on the right-hand side is just $C(n - 1)$, and from (6), we have

$$F^2(z) = \sum_{n=2}^{\infty} C(n - 1)\, z^{n} = F(z) - z.$$

Hence $F(z)$ obeys the equation

$$F^2(z) - F(z) + z = 0,$$

the two solutions of which are

$$F(z) = \frac{1 \pm \sqrt{1 - 4z}}{2}. \tag{7}$$

Note now that $F(z)$ is related to our random walk problem in the following way. The sum in (6) for the particular value $z = p(1 - p)$ is given by

$$F(p - p^2) = \sum_{\alpha=1}^{\infty} C(\alpha - 1)p^{\alpha}(1 - p)^{\alpha}.$$

This is—up to a factor—exactly the probability we seek, since

$$P_1 = \sum_{n=0}^{\infty} C(n)p^n(1 - p)^{n+1} = \frac{1}{p}\sum_{\alpha=1}^{\infty} C(\alpha - 1)p^{\alpha}(1 - p)^{\alpha}.$$

From (7) we obtain

$$P_1 = \frac{F(p - p^2)}{p} = \frac{1 \pm \sqrt{1 - 4p + 4p^2}}{2p} = \frac{1 \pm (1 - 2p)}{2p}.$$

This formula gives two possible values for the probability that the particle ends up at the origin, given that it starts at $k = 1$. The correct choice depends on whether $p \leq 1/2$ or $p > 1/2$. For $p \leq 1/2$, we get $P_1 = 1$, because the other root of the quadratic would give a probability greater than 1. For $p > 1/2$, there is some positive probability that the walk never ends, so we get

$$P_1 = \frac{1 - p}{p}. \tag{8}$$

(For an intuitive reason why this is so, observe that the walk cannot return to the origin if $p = 1$.) We have confirmed our claims in (2). For an interpretation of these results and a more detailed discussion of the related problems mentioned in the introduction, see the references [**4, 5**].

The case $k = 2$ Let us compute next the probability that the particle ends up at the origin given that it starts at $k = 2$. The general probability formula becomes

$$P_2 = \sum_{n=0}^{\infty} C_2(n)p^n(1 - p)^{n+2}. \tag{9}$$

This case is simplified by the fact that $C_2(n)$ is the same as $C(n + 1)$. This is almost immediate: Every path that begins at $x = 1$ and takes at least one step to the right must begin with a step to the right. Thus, every path counted by $C(n + 1)$ could be thought of as a path starting at $x = 2$ with one fewer right step. In FIGURE 3 we show how this looks in path representation. Just slide the set of paths on the right one unit to the left and observe that it is a perfect match.

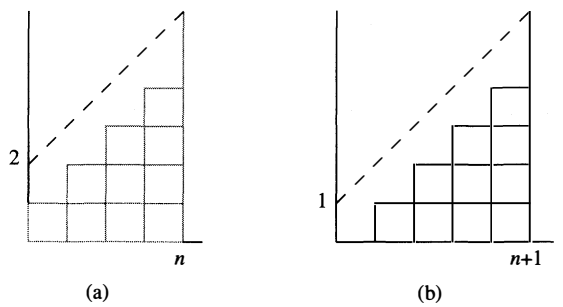


Figure 3 Relationship between $C_2(n)$ and the Catalan numbers

Substituting $C(n+1)$ for $C_2(n)$ in (9) we obtain

$$P_2 = \sum_{n=0}^{\infty} C(n+1)p^n(1-p)^{n+2}.$$

By making the change of variables $m = n+1$, we evaluate the sum as follows.

$$P_2 = \sum_{m=1}^{\infty} C(m)p^{m-1}(1-p)^{m+1} = \frac{1}{p}\sum_{m=0}^{\infty} C(m)p^m(1-p)^{m+1} - \frac{1-p}{p}C(0)$$

Using (4), the known values for P_1, and the fact that $C(0) = 1$, we have

$$P_2 = \frac{P_1}{p} - \frac{1-p}{p} = \begin{cases} 1 & p \le 1/2 \\ \frac{(1-p)^2}{p^2} & p > 1/2. \end{cases}$$

These results verify (1) for $k = 2$.

The case $k = 3$ We now compute the probability that the particle ends up at the origin given that it starts at $k = 3$. We need this case to establish (1) in general and to derive a recurrence relation between the coefficients $C_k(n)$ for different values of n and k.

THEOREM 2. *For $k \ge 3$, the path counts obey the following recurrence formula*

$$C_k(n) = C_{k-1}(n+1) - C_{k-2}(n+1). \tag{10}$$

Heuristic proof. A walk that begins at $x = k-1$ with $n+1$ right steps can start with a step to the left or the right. If it goes to the right first, then the rest of that walk is counted in $C_k(n)$. If it goes to the left, then the rest of the walk is counted in $C_{k-2}(n)$. The difference in (10) is exactly the number of paths counted in $C_{k-1}(n+1)$ that proceed to the right. A geometric representation of (10) is depicted in FIGURE 4. You can slide the drawings over one another to verify the count. ■

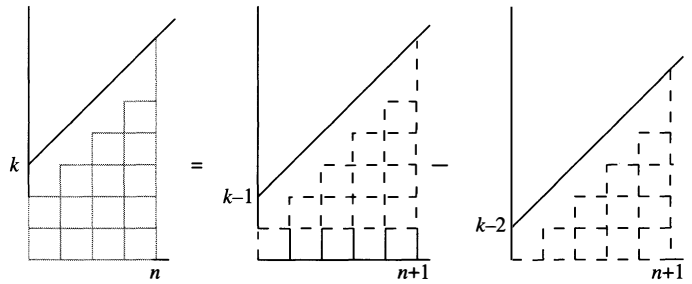


Figure 4 The recurrence relation for $C_k(n)$

The required probability when $k = 3$ is given by

$$P_3 = \sum_{n=0}^{\infty} C_3(n)p^n(1-p)^{n+3}.$$

Use Theorem 2 to write $C_3(n) = C_2(n+1) - C(n+1)$. But we know that $C_2(n) = C(n+1)$, and thus

$$C_3(n) = C(n+2) - C(n+1).$$

We can now write

$$P_3 = \sum_{n=0}^{\infty} C(n+2)p^n(1-p)^{n+3} - \sum_{n=0}^{\infty} C(n+1)p^n(1-p)^{n+3}.$$

Changing the variable to $m = n+2$ in the first sum above and to $m = n+1$ in the second sum, we obtain, after some straightforward computation,

$$P_3 = \sum_{m=0}^{\infty} C(m)p^{m-2}(1-p)^{m+1} - \sum_{m=0}^{\infty} C(m)p^{m-1}(1-p)^{m+2} - \frac{1-p}{p^2},$$

where we have used again that $C(0) = 1$. Factoring out the terms that do not depend on m we obtain

$$P_3 = \frac{1}{p^2}\sum_{m=0}^{\infty} C(m)p^m(1-p)^{m+1} - \frac{1-p}{p}\sum_{m=0}^{\infty} C(m)p^m(1-p)^{m+1} - \frac{1-p}{p^2}.$$

Each sum above is just P_1; hence

$$P_3 = \frac{P_1}{p^2} - \frac{1-p}{p}P_1 - \frac{1-p}{p^2}.$$

Using the values of P_1 we obtain the desired values, as in (1).

The general case As the three cases above suggest, we have the following result.

THEOREM 3. *The probability P_k that the particle ends up at the origin given that it starts at $x = k$ is given by*

$$P_k = P_1^k.$$

Heuristic proof. If a particle starts at $x = k+1$, then its first step is either to the left or to the right. Suppose it starts off to the left, an event with probability $(1-p)$, and so arrives at $x = k$. The probability that it reaches the origin from there is P_k. Repeating the analysis for an initial step to the right, which has probability p, we find that it arrives home from there ($x = k+2$) with probability P_{k+2}. Therefore,

$$P_{k+1} = (1-p)P_k + pP_{k+2},$$

from which we get

$$P_{k+2} = \frac{P_{k+1}}{p} - \frac{(1-p)}{p}P_k.$$

The induction process now follows easily from this equation and the assumption that (1) holds for the probabilities on the right-hand side above. ■

Acknowledgment. Financial support by FAPESP—Fundação de amparo à pesquisa do Estado de São Paulo—under grant 01/08485-6 is greatly appreciated. I thank Pierluigi Contucci for this and *other* path representations.

REFERENCES

1. F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, Inc., San Francisco, 1965, pp. 4–11.
2. Richard Isaacs, *The Pleasures of Probability*, Springer, NY, 1995, pp. 104–108.

3. Ralph P. Grimaldi, *Discrete and Combinatorial Mathematics*, 3rd ed., Addison-Wesley Publishing Company, Reading, MA, 1994, pp. 502–523.
4. G. Grimmett, D. Welsh, *Probability, an Introduction*, Oxford Science Publications, Oxford, 1985, 159–164.
5. F. Mosteller, *Fifty Challenging Problems in Probability*, Dover, New York, 1987, pp. 6–9.
6. R. Stanley, *Enumerative Combinatorics*, vol. 2, Cambridge Univ. Press, Cambridge, UK, 1999, pp. 256–265.

Why Some Elementary Functions Are Not Rational

GABRIELA CHAVES
JOSÉ CARLOS SANTOS
Universidade do Porto
4169–007, Porto
Portugal
gchaves@fc.up.pt
jcsantos@fc.up.pt

A classical exercise for college students is to ask them to prove that the sine function is not a polynomial or, more generally, a rational function. This follows from the fact that the sine function has an infinite number of zeros, which cannot occur for a rational function unless it is identically zero. This, however, does not rule out the possibility that the restriction of the sine function to an open interval is rational. One way to prove that a function defined on some open interval is not a polynomial function is to calculate successive derivatives of the function: Since the degree of the derivative of a nonconstant polynomial function is smaller than the degree of the polynomial, after some time you will get a constant function. This approach, however, does not work for rational functions. Or does it?

In this note we prove, in an elementary way, that the restrictions to an open interval of certain elementary functions are not rational functions, and we do it by using the concept of degree of a rational function.

In what follows, the domain of every function is a fixed nonempty open interval of the real line. When we speak of the exponential function, the sine function, and so on, what we actually mean is the restriction of the function to this interval.

DEFINITION. *If R is a rational function and if $R \not\equiv 0$, then the degree of R (written* $\deg(R)$*) is the difference of the degrees of the numerator and denominators. Specifically, if P_1 and P_2 are polynomials such that $R = P_1/P_2$, then* $\deg(R) = \deg(P_1) - \deg(P_2)$.

It is easy to see that this definition makes sense, that is, that $\deg(R)$ does not depend upon the choice of P_1 and P_2 [**1**, §4.2].

THEOREM. *Suppose f and g are rational functions, neither of which is identically zero.*

1. *If $f + g \not\equiv 0$, then* $\deg(f + g) \leq \max\{\deg(f), \deg(g)\}$.
2. *The degree of a product satisfies* $\deg(f \cdot g) = \deg(f) + \deg(g)$.
3. *If $f' \not\equiv 0$, then* $\deg(f') < \deg(f)$. *More generally, if n is a natural number and if $f^{(n)} \not\equiv 0$, then* $\deg(f^{(n)}) \leq \deg(f) - n$.

Proof. As stated by Bourbaki [**1**, §4.2], the proofs of the first two statements are a direct consequence of the similar formulas for polynomials. Let P_1, P_2, Q_1, and Q_2 be

polynomial functions such that $f = P_1/P_2$ and $g = Q_1/Q_2$. If $f + g \not\equiv 0$, then

$$\begin{aligned}\deg(f+g) &= \\ &= \deg(P_1Q_2 + P_2Q_1) - \deg(Q_1Q_2) \\ &\le \max\{\deg(P_1) + \deg(Q_2), \deg(P_2) + \deg(Q_1)\} - \deg(Q_1) - \deg(Q_2) \\ &= \max\{\deg(P_1) - \deg(Q_1), \deg(P_2) - \deg(Q_2)\} \\ &= \max\{\deg(f), \deg(g)\}.\end{aligned}$$

The second statement follows directly from the definition. Finally, if $f' \not\equiv 0$, then

$$\begin{aligned}\deg(f') &= \deg\left(\frac{P_1'P_2 - P_2'P_1}{{P_2}^2}\right) \\ &= \deg(P_1'P_2 - P_2'P_1) - 2\deg(P_2) \\ &\le \max\{\deg(P_1') + \deg(P_2), \deg(P_2') + \deg(P_1)\} - 2\deg(P_2) \\ &= \max\{\deg(P_1) - 1 + \deg(P_2), \deg(P_2) - 1 + \deg(P_1)\} - 2\deg(P_2) \\ &= \deg(P_1) - \deg(P_2) - 1 = \deg(f) - 1.\end{aligned}$$

It follows by induction that if $n \in \mathbb{N}$ and if $f^{(n)} \not\equiv 0$, then $\deg(f^{(n)}) \le \deg(f) - n$. ■

This theorem already allows us to show that certain functions are not rational. Take, for instance, the function defined by $f(x) = \sqrt[3]{1+x^2}$. If it were rational then it would follow from the second statement of the theorem that $2 = \deg(f^3) = 3\deg(f)$, which is impossible since $\deg(f)$ is an integer. Since $(e^x)' = e^x$, the third statement tells us that the exponential function is not a rational function.

Note that although it is true that, for a nonconstant polynomial function P, we always have $\deg(P') = \deg(P) - 1$, this is not true in general for rational functions: if $n \in \mathbb{N}$ and if $f(x) = (x^n - 1)/(x^n + 1)$, then $\deg(f) = 0$ but $\deg(f') = -n - 1$.

As a consequence of the third statement of the theorem, we have the following

COROLLARY. *If f is a rational function (with $f \not\equiv 0$), $k \in \mathbb{R}$ and $n \in \mathbb{N}$, then $f^{(n)} \not\equiv k \cdot f$, unless f is a polynomial function, $k = 0$, and $n > \deg(f)$.*

This corollary can be used (with $k = 1$ and $n = 1$) to prove again that the exponential function is not rational and also (with $k = -1$ and $n = 2$) to prove that neither the sine function nor the cosine function is rational. With a little extra effort it could also be deduced from the corollary that neither the tangent function nor the cotangent function is a rational function, but it is easier to observe that, since $\tan'(x) = \sec^2(x)$, if the tangent function was rational then $x \mapsto \cos^2(x)$ would also be rational as the reciprocal of a rational function. But

$$\cos^2(x) = \frac{\cos(2x) + 1}{2}$$

and therefore the function $x \mapsto \cos(2x)$ would be rational. That this cannot be the case can be deduced from the corollary (with $k = -4$ and $n = 2$) or from the fact that cosine is not rational.

REFERENCE

1. N. Bourbaki, *Algebra II*, Springer-Verlag, Berlin, 1980.

Another Look at Sylow's Third Theorem

EUGENE SPIEGEL
University of Connecticut
Storrs, Connecticut 06269
spiegel@math.uconn.edu

Among the results that Sylow showed in his famous 1872 paper [**12**] is what is now usually called Sylow's third theorem.

If G is a finite group of order $|G| = p^n m$ where p is a prime, n is a positive integer, and p and m are relatively prime, then the number, N_p, of subgroups of G of order p^n satisfies $N_p \equiv 1$ (modulo p).

This result is among the arsenal of tools that every first year algebra student obtains. A group where the order of every element is a power of p is called a p-group; a p-Sylow subgroup of G is a p-subgroup of G of maximal order p^n. The idea of Sylow's proof, which was originally stated in terms of permutation groups, is to look at the size of the equivalence classes obtained when all p-Sylow subgroups of G are conjugated by the elements of a fixed p-Sylow subgroup of G. The existence of a p-Sylow subgroup was needed for the proof of the third Sylow theorem, although the conclusion of the theorem certainly implies that there are p-Sylow subgroups. Using the Sylow results, Frobenius, in 1895 [**1**], proved a generalization: The number of subgroups of G of order p^s is congruent to 1 modulo p whenever $1 \leq s \leq n$.

Most current texts show the existence of a p-Sylow subgroup and prove Sylow's third theorem using arguments that involve a group acting on a set. This method of proof for the existence of a p-Sylow subgroup was due to Miller between 1910 and 1915 [**8, 9**], but, according to Jacobson [**9**, p. 83], was "forgotten until it was rediscovered" in 1959 by Wielandt [**14**]. Krull [**7**] showed how Wielandt's method could be used to obtain Frobenius' generalization, and Gallagher, in 1967 [**2**], simplified the argument to one that depends upon the order of G rather than the group itself. Illustrating the combinatorics of finite group actions is part of the motivation to use this method of proof both to demonstrate the existence of p-Sylow subgroups in a finite group and to determine their number.

In this note we offer another method to prove these results. Our combinatorial tool will be Möbius inversion on the lattice of subgroups of a finite group. We will see that an application of this method will easily lead to Frobenius' theorem, in fact, a generalization of it. Of course, part of the reason for presenting this proof is to highlight the method.

Möbius inversion In this section, which can be skimmed by those conversant with Möbius inversion, we present all needed facts about incidence algebras and Möbius inversion. Suppose X is a finite partially ordered set. We use standard interval notation in X, for example, $(x, y] = \{z \in X \mid x < z \leq y\}$. If X has a minimum or maximum element, it will be denoted by $\widehat{0}$ or $\widehat{1}$ respectively.

The incidence algebra, $I(X, \mathbb{C})$, of X over $\mathbb{C}$, where $\mathbb{C}$ is the complex field, is defined as $I(X, \mathbb{C}) = \{f : X \times X \to \mathbb{C} \mid f(x, y) = 0 \text{ if } x \not\leq y\}$ with the operations of addition and scalar multiplication defined in the usual way, while multiplication is defined by convolution, $fg(x, y) = \sum_{z \in [x,y]} f(x, z)g(z, y)$, for $f, g \in I(X, \mathbb{C})$. It is straightforward to check that $I(X, \mathbb{C})$ is both a vector space over $\mathbb{C}$ and a ring. This

makes it a $\mathbb{C}$-algebra with identity δ, where $\delta(x, x) = 1$ for $x \in X$, and $\delta(x, y) = 0$, if $x \neq y$. If $\phi \in I(X, \mathbb{C})$ is such that for each $x \in X$, $\phi(x, x)$ is a unit in $\mathbb{C}$, then ϕ is invertible. Its inverse will be seen to be $\phi' \in I(X, \mathbb{C})$, having the following properties: For $x \in X$, $\phi'(x, x) = (\phi(x, x))^{-1}$, while if $\phi'(u, v)$ has been given for any $|[u, v]| < |[x, y]|$, then $\phi'(x, y)$ satisfies

$$\phi'(x, y) = -\left(\sum_{z: x<z\leq y} \phi(x, z)\phi'(z, y)\right)\phi(x, x)^{-1}.$$

It is easy to verify that ϕ' is a right inverse for ϕ. Similarly, ϕ has a left inverse. It follows that an element, $\phi \in I(X, \mathbb{C})$, has an inverse if and only if $\phi(x, x)$ is a unit for each $x \in X$. In particular, the element $\zeta \in I(X, \mathbb{C})$ given by

$$\zeta(x, y) = \begin{cases} 1 & \text{if } x \leq y \\ 0 & \text{otherwise} \end{cases}$$

is a unit whose inverse, μ, is called the Möbius function of the partially ordered set. When we need to specify which partially ordered set we are considering, we will let μ_X denote the Möbius function for the partially ordered set X.

The importance of the Möbius function is seen in the following result, which is known as the Möbius inversion theorem. A more general Möbius inversion theorem appears in the seminal paper of Rota [**10**]. Earlier versions of this theorem can be found, for example, in Weisner [**13**] or Hall [**4**].

THEOREM 1. *Let X be a finite partially ordered set and f a function from X to $\mathbb{C}$. If, for $x \in X$,*

$$g(x) = \sum_{y:\ x\leq y} f(y)$$

then

$$f(x) = \sum_{y:\ x\leq y} \mu(x, y)g(y).$$

Proof.

$$\begin{aligned}
\sum_{y:\ x\leq y} \mu(x, y)g(y) &= \sum_{y:\ x\leq y} \mu(x, y)\Big(\sum_{z:\ y\leq z} f(z)\Big) \\
&= \sum_{y:\ x\leq y} \mu(x, y)\Big(\sum_{z:\ y\leq z} \zeta(y, z)f(z)\Big) \\
&= \sum_{y:\ x\leq y} \sum_{z:\ y\leq z} \mu(x, y)\zeta(y, z)f(z) \\
&= \sum_{z:\ x\leq z} f(z) \sum_{y:\ x\leq y\leq z} \mu(x, y)\zeta(y, z) \\
&= \sum_{z:\ x\leq z} \delta(x, z)f(z) \\
&= f(x).
\end{aligned}$$

■

The following proposition presents three properties of the Möbius function, which are useful in its computation. The first of these, together with the condition $\mu(x, x) = 1$ for every $x \in X$, could be used as the definition of the Möbius function.

The third of these properties is a result of Weisner [**14**]. It applies only when X is a *lattice*, which is a partially ordered set in which every pair of elements, a, b, has a least upper bound, denoted $a \vee b$, and a greatest lower bound. We are particularly interested in the set of all subgroups of a finite group G, partially ordered by inclusion. This set, $\mathcal{L}(G)$, is a lattice with least element $\widehat{0} = \{e\}$ and greatest element $\widehat{1} = G$. If $A, B \in \mathcal{L}(G)$, then $A \vee B$ is the subgroup generated by A and B.

PROPOSITION 1. *Suppose X is a finite partially ordered set with Möbius function μ.*

(i) If $x < y \in X$, then $\sum_{z \in [x,y]} \mu(x, z) = 0$.

(ii) If $\phi : X \longrightarrow X'$ is an isomorphism of partially ordered sets, then $\mu_X(x, y) = \mu_{X'}(\phi(x), \phi(y))$, for any $x, y \in X$.

(iii) If X is a finite lattice and $a \neq \widehat{0} \in X$, then

$$\sum_{x:\ x \vee a = \widehat{1}} \mu(\widehat{0}, x) = 0.$$

Proof.

(i) $0 = \delta(x, y) = \mu\zeta(x, y) = \sum_{z \in [x,y]} \mu(x, z)$.

(ii) One can extend ϕ to a $\mathbb{C}$-isomorphism $\phi' : I(X, \mathbb{C}) \longrightarrow I(X', \mathbb{C})$. Since $\phi'(\zeta_X) = \zeta_{X'}$ then $\phi'(\mu_X) = \mu_{X'}$.

We check (iii) by induction on $|X|$. If $|X| = 2$, then $a = \widehat{1}$, and as $x \vee \widehat{1} = \widehat{1}$ for any $x \in X$, the result follows from (i). Now assume the result for lattices of smaller cardinality than that of X. We have

$$0 = \sum_{x \in X} \mu(\widehat{0}, x) = \sum_{x \in [\widehat{0}, a]} \mu(\widehat{0}, x) + \sum_{b:\ a < b < \widehat{1}} \left(\sum_{x:\ x \vee a = b} \mu(\widehat{0}, x) \right) + \sum_{x:\ x \vee a = \widehat{1}} \mu(\widehat{0}, x).$$

Looking at the right-hand side of this equation, we note that the first sum is 0 by (i) and that each summand in the second sum is 0 by the inductive assumption. The result then follows. ∎

We now use these properties to calculate the Möbius function of the lattice, $\mathcal{L}(G)$, of a finite p-group G. The answer was obtained by Kratzer and Thévenez, [**6**], using a different approach. The computation of $\mu_{\mathcal{L}(G)}(\{e\}, G) = \mu_{\mathcal{L}(G)}(\widehat{0}, \widehat{1})$ is a result of P. Hall [**3**]. For notation, if A is a subgroup of G, let $N(A) = N_G(A)$ denote its normalizer and write Z_p^k for the direct sum of k copies of a cyclic group of order p.

THEOREM 2. *Let G be a finite p-group, $\mathcal{L}(G)$ the lattice of subgroups of G, and μ its Möbius function. If $A, B \in \mathcal{L}(G)$, then*

$$\mu(A, B) = \begin{cases} (-1)^k p^{\binom{k}{2}} & \text{if } A \subseteq B \subseteq N(A) \text{ and } B/A \simeq Z_p^k \\ 0 & \text{otherwise.} \end{cases}$$

Proof. If $|G| = p$ the result is easily verified. Continue by induction on the order of G, and assume the result for groups of smaller order than that of G. Suppose $|G| = p^n$ and let $A \subseteq B$ be subgroups of G. Since the collection of subgroups between A and B is both an interval in $\mathcal{L}(B)$ and in $\mathcal{L}(G)$, by the induction assumption it is sufficient to verify the result when $B = G$. Suppose, first, that A is not a normal subgroup of G. Then $A \subset N(A) \subset G$ and

$$0 = \sum_{X \in [A,G]} \mu(A, X)$$

$$= \sum_{X \in [A,N(A)]} \mu(A, X) + \sum_{X \in [A,G] \setminus [A,N(A)]} \mu(A, X).$$

The first summation in the last line is zero by Proposition 1(i), and each summand with $X \neq G$ in the second summation is zero by our inductive assumption. Hence $\mu(A, G) = 0$, verifying the result in this case.

Suppose now that A is a normal subgroup of G. If $\{e\} \neq A$, by Proposition 1(ii), $\mu_{\mathcal{L}(G)}(A, G) = \mu_{\mathcal{L}(G/A)}(\widehat{0}, \widehat{1})$ and, as $|G/A| < |G|$, the result again follows by the inductive assumption. Hence, to complete the proof, all that remains is to check the result in the case $\mu(\widehat{0}, \widehat{1})$. Let H be a normal subgroup of G of order p. If X is any proper subgroup of G with $G = H \vee X$, then $G = HX$ and $X \simeq G/H$. We conclude that all such subgroups, X, are isomorphic, $G \simeq H \oplus X$, and, from Proposition 1(iii), we have $\mu_{\mathcal{L}(G)}(\widehat{0}, \widehat{1}) = -\lambda \mu_{\mathcal{L}(G/H)}(\widehat{0}, \widehat{1})$, where λ is the number of proper subgroups, X, of G with $H \vee X = G$.

Suppose $\mu_{\mathcal{L}(G)}(\widehat{0}, \widehat{1}) \neq 0$. It follows that $\mu_{\mathcal{L}(G/H)}(\widehat{0}, \widehat{1}) \neq 0$ and, by the inductive assumption, $G/H \simeq Z_p^{n-1}$ and $\mu_{\mathcal{L}(G/H)}(\widehat{0}, \widehat{1}) = (-1)^{n-1} p^{\binom{n-1}{2}}$. Further $G \simeq Z_p^n$. To calculate λ, it is convenient to consider G as an n-dimensional vector space over a field of p elements. We are then looking for subspaces, X, of dimension $n - 1$ which do not contain a fixed 1-dimensional subspace, H. The total number of ordered bases for all such X is $(p^n - p)(p^n - p^2) \cdots (p^n - p^{n-1})$, where, for example, the first factor comes from choosing the first basis vector of X to be any vector of G other than those in H. By a similar argument, each $n - 1$ dimensional space has $(p^{n-1} - 1)$ $(p^{n-1} - p) \cdots (p^{n-1} - p^{n-2})$ ordered bases. Hence $\lambda = p^{n-1}$, it being the quotient of these two numbers. We conclude $\mu_{\mathcal{L}(G)}(\widehat{0}, \widehat{1}) = -p^{n-1}(-1)^{n-1} p^{\binom{n-1}{2}} = (-1)^n p^{\binom{n}{2}}$. This is the desired value. ∎

The Sylow theorem We now obtain the result we sought by applying Möbius inversion. Before we give the main result, it is convenient to handle one special case separately in a lemma. The lemma, like the theorem following it, only requires Cauchy's theorem.

LEMMA 1. *Let G be a finite group and p a prime dividing the order of G. The number, $\mathbf{K}$, of subgroups of G of order p is congruent to 1 modulo p.*

Proof. Suppose $\mathcal{C}(p)$ is the collection of all subgroups of order p in G and P is a p-subgroup of G of maximum order. We let P act on $\mathcal{C}(p)$ by conjugation. For $A \in C(p)$, $\mathrm{Orb}(A) = \{gAg^{-1} \mid g \in P\}$ is the orbit of A. It is not difficult to check that $|\mathrm{Orb}(A)|$ is equal to the index of $\mathrm{Stab}(A) = \{g \in P \mid gAg^{-1} = A\}$ in P. This index, which is a power of p, is greater than 1 unless $P \subseteq N(A)$. Hence

$$|\mathcal{C}(p)| = \mathbf{K} \equiv \overline{\mathbf{K}} \text{ (modulo } p),$$

where $\overline{\mathbf{K}} = |\{A \in \mathcal{C}(p) \mid P \subseteq N(A)\}|$. When $A \in \mathcal{C}(p)$ with $P \subseteq N(A)$, then AP is a p-subgroup of $N(A)$ that contains P. By maximality, $A \subseteq P$. We conclude that $\overline{\mathbf{K}}$ is the number of normal subgroups of order p in P. By having the elements of P act by conjugation on themselves, one can check that any nontrivial normal subgroup of a p-group, in this case P, has a nontrivial intersection with its center, $Z(P)$. In particular, any normal subgroup of P of order p is in $Z(P)$, and in fact in $\mathrm{Soc}_p(Z(P)) = \{x \in Z(P) \mid x^p = 1\} \simeq Z_p^k$, for some integer $k \geq 1$. Conversely, each nonidentity element of $\mathrm{Soc}_p(Z(P))$ generates a normal subgroup of P of order p. Since any subgroup

of order p has $p-1$ nonidentity elements, we have $\mathbf{K} \equiv \overline{\mathbf{K}} = (p^k - 1)/(p-1) = 1 + p + \cdots + p^{k-1} \equiv 1$ (modulo p), establishing the lemma. ∎

We can now give a generalized Sylow's third theorem as an application of Möbius inversion.

THEOREM 3. *Let G be a finite group of order exactly divisible by p^n, where p is a prime and n a positive integer. Suppose m, t are nonnegative integers with $m \leq t \leq n$ and H is a subgroup of G of order p^m. Then the number of subgroups of G of order p^t, each of which contains H, is congruent to 1 modulo p.*

Proof. We prove the result by induction on $s = n - m$. If $s = 0$ the result is immediate. Suppose that we know the result for values smaller than $s = n - m$. Of course, if $t = m$ the result is clear. We can thus assume that $m < t \leq n$. Let $\mathcal{C}_H(t)$ be the collection of subgroups of G of order p^t, each of which contains H. Define the function f on $\mathcal{L}(G)$ by

$$f(S) = \begin{cases} 1 & \text{if } S \in \mathcal{C}_H(t) \\ 0 & \text{otherwise,} \end{cases}$$

and the function g on $\mathcal{L}(G)$ by

$$g(T) = \sum_{S \in \mathcal{L}(G): S \supseteq T} f(S).$$

Then $g(T)$ is the number of elements of $\mathcal{C}_H(t)$ that contain T. We wish to compute $g(H) = |\mathcal{C}_H(t)|$. Using Möbius inversion in $\mathcal{L}(G)$ we obtain

$$f(T) = \sum_{S:\ S \supseteq T} \mu(T, S) g(S).$$

In particular,

$$f(H) = \sum_{S:\ S \supseteq H} \mu(H, S) g(S).$$

Of course, $g(S) = 0$ unless S is a p-group. Furthermore, by Theorem 2, when S is a p-group containing H, $\mu(H, S)$ is divisible by p whenever the index of H in S is divisible by p^2. Hence,

$$0 \equiv g(H) - \sum_{S \subseteq N(H):\ S/H \simeq Z_p} g(S) \text{ (modulo } p).$$

That is

$$|\mathcal{C}_H(t)| \equiv \sum_{S \subseteq N(H):\ S/H \simeq Z_p} g(S) \text{ (modulo } p).$$

Each S in this summand is of order p^{m+1}. As $n - (m+1) < s$, by the induction assumption, $g(S) \equiv 1$ (modulo p). Further, the number of such S coincides with the number of subgroups of order p in $N(H)/H$. By the lemma, this number is congruent to 1 modulo p. We conclude that $|\mathcal{C}_H(t)| \equiv 1$ (modulo p), which establishes the result. ∎

If in the previous theorem we let $m = 0$ and $t = n$ then we obtain the first and third Sylow theorems. The Frobenius theorem is obtained by letting $m = 0$.

Note added in proof: The author thanks Keith Conrad for bringing to my attention another paper of L. Weisner, "Some properties of prime-power groups," *Trans. AMS* **38** (1835), 485–492, in which Weisner proves Theorem 3 in the case that G is a p-group.

REFERENCES

1. G. Frobenius, Verallgemeinerung des Sylow'schen satzes, *Berliner Sitzungsberichte* (1895), 981–993.
2. P. X. Gallagher, On the p-subgroups of a finite group, *Arch. der Math.* **18** (1967), 469.
3. P. Hall, A contribution to the theory of groups of prime-power order, *Proc. London Math. Soc.* **36** (1933), 29–95.
4. ———, The Eulerian functions of a group, *Quart. J. Math.* **7** (1936), 134–151.
5. N. Jacobson, *Basic Algebra I*, W. H. Freeman and Co., New York, 1985.
6. C. Kratzer and J. Thévenez, Fonction de Möbius d'un groupe fini et anneau de Burnside, *Commen. Math. Helvetici* **59** (1984), 425–438.
7. W. Krull, Über die p-untergruppen endlicher gruppen, *Arch. Math.* **12** (1961), 1–6.
8. G. A. Miller, Extensions of two theorems due to Cauchy, *Bull. AMS* **16** (1910), 510–513.
9. ———, A new proof of Sylow's theorem, *Ann. of Math.* **16** (1915), 169–171.
10. G.-C. Rota, On the foundations of combinatorial theory I. Theory of Möbius functions, *Wahr Scheinlichkeits Thoerie und Verw. Gebiete* **2** (1964), 340–368.
11. E. Spiegel and C. O'Donnell. *Incidence Algebras*, Marcel Dekker, Inc., New York, 1997.
12. L. Sylow, Théorèmes sur les groupes de substitutions, *Math. Ann.* **5** (1872), 584–594.
13. L. Weisner, Abstract theory of inversion of finite sets, *Trans. AMS* **38** (1935), 474–484.
14. H. Wielandt, Ein beweiss für die existenz der Sylowgruppen, *Arch. Math.* **10** (1959), 401–402.

Continued from page 246

In terms of total score (out of a maximum of 252), the highest ranking of the 82 participating teams were as follows:

Bulgaria	227	Turkey	133
China	211	Japan	131
USA	188	Hungary	128
Vietnam	172	United Kingdom	128
Russia	167	Canada	119
Korea	157	Kazakhstan	119
Romania	143		

The 2003 USAMO was prepared by Titu Andreescu (Chair), Zuming Feng, Kiran Kedlaya, and Richard Stong. The Team Selection Test was prepared by Titu Andreescu and Zuming Feng. The MOSP was held at the University of Nebraska-Lincoln. Zuming Feng (Academic Director), Gregory Galperin, and Melanie Wood served as instructors, assisted by Po-Shen Loh and Reid Barton as junior instructors, and Ian Le and Ricky Liu as graders. Kiran Kedlaya served as guest instructor.

For more information about the USAMO or the MOSP, contact Steven Dunbar at sdunbar@math.unl.edu.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN GELCA, Texas Tech University; ROBERT GREGORAC, Iowa State University; GERALD HEUER, Concordia College; VANIA MASCIONI, Ball State University; PAUL ZEITZ, The University of San Francisco

Proposals

To be considered for publication, solutions should be received by November 1, 2004.

1696. *Proposed by Albert F. S. Wong, Temasek Polytechnic, Singapore.*

For which positive integers k does the equation

$$x^{2k-1} + y^{2k} = z^{2k+1}$$

have a solution in positive integers x, y, and z?

1697. *Proposed by David Callan, Madison, WI.*

A permutation π on $[n] = \{1, 2, \ldots, n\}$ is *connected* if for each k, $1 \leq k \leq n-1$, there is a j, $1 \leq j \leq k$ with $\pi(j) > k$. Let a_n denote the number of connected permutations on $[n]$. Show that for $n \geq 2$,

$$a_n = \sum_{k=1}^{n-1} k(n-k-1)!a_k.$$

1698. *Proposed by Achilleas Sinefakopoulos, student, Cornell University, Ithaca, NY.*

Let n be an odd positive integer and let r be a positive rational number. Prove that there are positive integers $a_1, a_2, a_3, b_1, b_2, b_3$ such that

$$r = \frac{a_1^n + a_2^{n+1} + a_3^{n+2}}{b_1^n + b_2^{n+1} + b_3^{n+2}}.$$

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. Each solution should begin on a separate sheet.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames IA 50011, or mailed electronically (ideally as a LaTeX file) to ehjohnst@iastate.edu. All communications should include the readers name, full address, and an e-mail address and/or FAX number.

1699. *Proposed by Zhang Yun, First Middle School of Jinchung City, Gan Su, China.*

Let $A_1A_2A_3A_4$ be a nondegenerate tetrahedron, let h_k, $1 \le k \le 4$, be the length of the altitude from A_k, and let r be the radius of the inscribed sphere. Prove that

$$\frac{h_1}{h_1+3r}+\frac{h_2}{h_2+3r}+\frac{h_3}{h_3+3r}+\frac{h_4}{h_4+3r} \ge \frac{16}{7}.$$

1700. *Proposed by Yongge Tian, Queen's University, Kingston, Ontario, Canada.*

Let A and B be $n \times n$ matrices satisfying $A^2 = A$ and $B^2 = B$. Show that $AB = BA$ if and only if range(AB) = range(BA) and range(A^TB^T) = range(B^TA^T), where C^T denotes the transpose of C.

Quickies

Answers to the Quickies are on page 239.

Q941. *Proposed by Elias Lampakis, Kiparissia, Messinia, Greece.*

The sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ of real numbers are defined by

$$a_0 = \frac{1}{2004}, \qquad a_n = \sqrt{\frac{1+a_{n-1}}{2}}, \quad n \ge 1, \qquad \text{and} \qquad b_n = \prod_{k=1}^{n} a_k.$$

Evaluate $\lim_{n\to\infty} b_n$.

Q942. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Given fixed $a > 0$, determine the maximum value of

$$\frac{xyz}{(x+y+a)(y+z+a)(z+x+a)(x+y+z)},$$

where $x, y, z > 0$.

Solutions

Inverses and a Limit **June 2003**

1672. *Proposed by Árpád Bényi, University of Kansas, Lawrence, KS, and Mircea Martin, Baker University, Baldwin City, KS.*

Let f and g be odd functions that are analytic in a neighborhood of 0. Given that $f'(0) = g'(0) \ne 0$, $f^{(3)}(0) = g^{(3)}(0) \ne 0$, and $0 \ne f^{(5)}(0) \ne g^{(5)}(0) \ne 0$, evaluate

$$\lim_{x\to 0} \frac{f(x)-g(x)}{f^{-1}(x)-g^{-1}(x)},$$

where h^{-1} denotes the inverse of the function h.

Solution by Cornelius Stallmann and Gerald Thompson, Augusta State University, Augusta, GA.

Because f and g are analytic in a neighborhood of 0 and $f'(0), g'(0) \ne 0$, it follows that f^{-1} and g^{-1} are also analytic in a neighborhood of 0. We can write

$$f(x) = a_1x + a_3x^3 + a_5x^5 + a_7x^7 + \cdots$$
$$g(x) = a_1x + a_3x^5 + b_5x^5 + b_7x^7 + \cdots,$$

where a_1, a_3, a_5, b_5 are nonzero and $a_5 \neq b_5$. By algebra or Faa di Bruno's formula we find that in a neighborhood of 0,

$$f^{-1}(x) = \frac{1}{a_1}x - \frac{a_3}{a_1^4}x^3 + \left(\frac{3a_3^3 - a_1a_5}{a_1^7}\right)x^5 + \cdots$$
$$g^{-1}(x) = \frac{1}{a_1}x - \frac{a_3}{a_1^4}x^3 + \left(\frac{3a_3^3 - a_1b_5}{a_1^7}\right)x^5 + \cdots.$$

Because $a_5 \neq b_5$ it follows that

$$\lim_{x\to 0}\frac{f(x)-g(x)}{f^{-1}(x)-g^{-1}(x)} = \lim_{x\to 0}\frac{(a_5-b_5)x^5 + (a_7-b_7)x^7 + \cdots}{\left(\frac{b_5-a_5}{a_1^6}\right)x^5 + \cdots} = -a_1^6 = -f'(0)^6.$$

Also solved by Michel Bataille (France), Jean Bogaert (Belgium), Con Amore Problem Group (Denmark), Chip Curtis, Knut Dale (Norway), Jim Delany, Daniele Donini (Italy), Richard Penn, Danrun Huang, Helen Skala, Nicholas C. Singer, Chu Wenchang (Italy), Li Zhou, and the proposers. There was one incomplete submission and one unreadable submission.

Bounding the Sine **June 2003**

1673. *Proposed by P. Ivady, Budapest, Hungary.*

Prove that for $0 < x < \pi$,

$$\frac{\sin^3 x}{x^3} < \left(\frac{\pi^2 - x^2}{\pi^2 + x^2}\right)^2.$$

Solution by Michel Bataille, Rouen, France.

Equivalently, we prove that for $0 < x < 1$,

$$\frac{\sin \pi x}{\pi x} < \left(\frac{1-x^2}{1+x^2}\right)^{2/3}.$$

Because $\sin \pi x = \pi x \prod_{n=1}^{\infty}(1 - x^2/n^2)$, it follows that

$$\frac{\sin \pi x}{\pi x} < (1-x^2)\left(1 - \frac{x^2}{4}\right)\left(1 - \frac{x^2}{9}\right),$$

for $0 < x < 1$. Thus it is sufficient to show that for $0 < x < 1$,

$$(1-x^2)\left(1 - \frac{x^2}{4}\right)\left(1 - \frac{x^2}{9}\right) < \left(\frac{1-x^2}{1+x^2}\right)^{2/3}. \qquad (*)$$

Let

$$f(x) = (1-x^2)^{1/3}(1+x^2)^{2/3}\left(1 - \frac{13}{36}x^2 + \frac{1}{36}x^4\right).$$

Then $(*)$ will follow if $f(x) < 1$ for $0 < x < 1$. Because $f(0) = 1$, we need only prove that f is decreasing on $[0, 1)$. An easy calculation leads to

$$f'(x) = -\frac{x}{54}(1-x^2)^{-2/3}(1+x^2)^{-1/3}\left(9x^6+36x^4+115x^2(1-x^2)+3\right),$$

and this is clearly negative on (0, 1). The conclusion follows.

Note: In Problem 5642 of the *Amer. Math. Monthly* **76** (1969), p. 422, Ray Redheffer posed a related problem asking for a proof of the inequality

$$\frac{\sin x}{x} \geq \frac{\pi^2 - x^2}{\pi^2 + x^2}.$$

In his solution to this MAGAZINE problem, Professor Redheffer obtains a stronger result, proving that

$$\left(\frac{\pi^2-x^2}{\pi^2+x^2}\right)^a < \frac{\sin x}{x} < \left(\frac{\pi^2-x^2}{\pi^2+x^2}\right)^b, \quad 0 < x < \pi$$

holds for constants $a \geq 1$ and $b \leq \pi^2/12$, and that these bounds on the values of a and b are the best possible.

Also solved by Jean Bogaert (Belgium), Paul Bracken, Knut Dale (Norway), Daniele Donini (Italy), Dimitry Fleischman, Julien Grivaux, (France), Elias Lampakis (Greece), Phil McCartney, Ray Redheffer, Albert Stadler (Switzerland), Michael Vowe (Switzerland), Li Zhou, and the proposer.

A Positive Cyclic Sum **June 2003**

1674. *Proposed by H. A. Shah Ali, Tehran, Iran.*

Given that $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n$, and $x_{n+1} = x_1$, prove that

$$\sum_{k=1}^{n} \frac{x_k - x_{k+1}}{1 + x_k x_{k+1}} \geq 0.$$

Solution by Knut Dale, Telemark University College, Bø, Norway.

First observe that if $0 \leq a, b$ and $a + b < \pi/2$, then

$$\tan a + \tan b \leq \frac{\tan a + \tan b}{1 - \tan a \tan b} = \tan(a+b), \tag{1}$$

and that equality holds if and only if $a = 0$ or $b = 0$.

We now prove the desired inequality. If $n = 1$ or $n = 2$, then the sum is 0 and the inequality is true. For $n \geq 3$, write the inequality in the form

$$\sum_{k=1}^{n-1} \frac{x_{k+1} - x_k}{1 + x_k x_{k+1}} \leq \frac{x_n - x_1}{1 + x_1 x_n}. \tag{2}$$

Next find θ_k, $1 \leq k \leq n$, with $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n < \pi/2$ such that $x_k = \tan\theta_k$ for each k. Then (2) is equivalent to

$$\sum_{k=1}^{n-1} \tan(\theta_{k+1} - \theta_k) \leq \tan(\theta_n - \theta_1).$$

This last inequality follows immediately from repeated applications of (1). Furthermore, we have strict inequality if and only if two or more of the differences $\theta_{k+1} - \theta_k$ are nonzero, that is, if and only if the x_ks take on at least three distinct values.

Also solved by Ricardo Alfaro, Michael Andreoli, Roy Barbara (Lebanon), Michel Bataille (France), Jean Bogaert (Belgium), Mark Bowron, Paul Bracken, Carl E. Bredlau, Minh Can, Marco Carone (Canada), Mario Catalani (Italy), Chip Curtis, Jim Delany, Daniele Donini (Italy), Robert L. Doucette, Dmitry Fleishman, Ovidiu Furdui, G.R.A.20 Problems Group (Italy), Jean-Pierre Grivaux (France), Eugene A. Herman, Danrun Huang, Michael A. Jones, Stephen Kaczkowski, Edward Krusling, Victor Y. Kutsenok, Elias Lampakis (Greece), Alan Levine, Kathleen E. Lewis, Justin Marks, W. Marszalek and T. Amdeberhan, Northwestern University Math Problem Solving Group, Rob Pratt, Jawad Sadek, Heinz-Jürgen Seiffert (Germany), Achilleas Sinefakopoulos (Greece), Nicholas C. Singer, Albert Stadler (Switzerland), H. T. Tang, University of Louisiana at Lafayette Math Club, Ian VanderBurgh (Canada), Michael Vowe (Switzerland), Chu Wenchang (Italy), Alfred Witkowski (Poland), Li Zhou, and the proposer.

A Fuhrman Point Fact **June 2003**

1675. *Proposed by Michael Woltermann, Washington and Jefferson College, Washington, PA.*

Let ABC be a triangle, and let $\widehat{AB}$, $\widehat{BC}$, $\widehat{CA}$, respectively, be the arcs of its circumcircle subtended by sides AB, BC, CA. (The arcs are defined so that any two of the three arcs intersect in just one point.) Let X, Y, Z, respectively, be the midpoints of $\widehat{AB}$, $\widehat{BC}$, $\widehat{CA}$, and let X', Y', Z', respectively, be the reflections of X, Y, Z in sides AB, BC, CA. Triangle $X'Y'Z'$ is called the Fuhrman triangle of triangle ABC, and the circumcenter F of triangle $X'Y'Z'$ is the Fuhrman point of ABC. Let I and N be the incenter and nine point center, respectively, of triangle ABC. Prove that N is the midpoint of segment IF.

Solution by Michel Bataille, Rouen, France.

First note that $\angle ZCA$ and $\angle ZBA$ subtend the same arc $\widehat{ZA}$ and that B and C are on the same side of $\overline{AZ}$. Hence $\angle ZCA = \angle ZBA = \angle B/2$. Similarly, $\angle YZC = \angle YAC = \angle A/2$. Thus $\angle ZCX = (\angle B + \angle C)/2 = \pi/2 - \angle YZC$, and it follows that CX is perpendicular to YZ. By similar reasoning applied to AY and BZ, we find that AY, BZ, CX are the altitudes of $\triangle XYZ$.

Now identify the points with complex numbers, with the origin at the center O of the circumcircle Γ of $\triangle ABC$ (and of $\triangle XYZ$.) Without loss of generality, we may assume that Γ is the unit circle. Because I is the orthocenter of $\triangle XYZ$ and N is the midpoint of the segment joining O to the orthocenter of $\triangle ABC$, we have $I = X + Y + Z$ and $2N = A + B + C$.

The condition $\overline{CX} \perp \overline{YZ}$ is equivalent to $\frac{Z-Y}{C-X} = -\frac{\overline{Z}-\overline{Y}}{\overline{C}-\overline{X}}$. Noting that $\overline{P} = 1/P$ for P on the unit circle, we can solve for C to obtain $C = -YZ/X$. The point symmetric to I with respect to N is $F' = (A + B + C) - (X + Y + Z)$. We prove that $F = F'$ by showing that F' is the circumcenter of $\triangle X'Y'Z'$, Indeed,

$$|F' - X'| = |(A + B + C) - (X + Y + Z) - (A + B - X)|$$
$$= \left|-\frac{YZ}{X} - Y - Z\right| = |XY + YZ + ZX|.$$

Because this expression is symmetric in X, Y, and Z, it follows that

$$|F' - X'| = |F' - Y'| = |F' - Z'|.$$

Thus $F' = F$ is the circumcenter of $\triangle X'Y'Z'$, and N is the midpoint of $\overline{IF} = \overline{IF'}$.

Note: R. S. Tiberio and Peter Yff point out that the Fuhrman center is point X(355) in Clark Kimberling's *Encyclopedia of Triangle Centers* at http://faculty.evansville.edu/ck6/encyclopedia/ . It is noted at the site that X(5) (the nine-point center) is the midpoint of X(1) (the incenter) and X(355).

Also solved by Herb Bailey, Daniele Donini (Italy), Jean-Pierre Grivaux (France), John G. Heuver (Canada), L. R. King, Robert L. Young, Li Zhou, and the proposer.

Coupled Congruences **June 2003**

1676. *Proposed by Erwin Just, Emeritus, Bronx Community College of the City University of New York, Bronx, NY.*

Find all pairs of integers m and n such that

$$2m \equiv -1 (\mathrm{mod}\, n) \quad \text{and} \quad n^2 \equiv -2 (\mathrm{mod}\, m).$$

Solution by Nicholas C. Singer, Annandale, VA.

Because $x(\mathrm{mod}\, n) = x(\mathrm{mod}(-n))$ for each integer x and $n^2 = (-n)^2$, every solution (m, n) will have a paired solution $(m, -n)$. Because $n \mid (2m + 1)$ and $m \mid (n^2 + 2)$, m and n must both be odd. Write $2m + 1 = an$ and $n^2 + 2 = bm$, where a and b are odd integers. Then $a^2bm = a^2n^2 + 2a^2 = 4m^2 + 4m + 1 + 2a^2$, so

$$2a^2 + 1 = m(a^2b - 4m - 4).$$

If $a = \pm 1$, then $3 = m(b - 4m - 4)$. There are four sets of paired solutions to this equation: $(m, n, a, b) = (3, \pm 7, \pm 1, 17)$, $(1, \pm 3, \pm 1, 11)$, $(-1, \mp 1, \pm 1, -3)$ and $(-3, \mp 5, \pm 1, -9)$. If $a = \pm 3$, then $19 = m(9b - 4m - 4)$. There are two sets of paired solutions to this equation: $(m, n, a, b) = (19, \pm 13, \pm 3, 9)$ and $(1, \pm 1, \pm 3, 3)$.

Next note that $2n^2 + 4 = 2bm = abn - b$, so $2n^2 - abn + (b + 4) = 0$. Let $k = ab - 2n$. Then k is odd and $ka = 2 + (k^2 + 8)/b$.

If $k = \pm 1$, then $\pm a = 2 + 9/b$. There are six sets of solution pairs, three of which are new: $(m, n, a, b) = (17, \pm 7, \pm 5, 3)$, $(27, \pm 5, \pm 11, 1)$ and $(-11, \pm 3, \mp 7, -1)$. If $k = \pm 3$, then $\pm 3a = 2 + 17/b$. There are two sets of solution pairs, only one of which is new: $(m, n, a, b) = (-3, \pm 1, \mp 5, -1)$. If $k = \pm 5$, then $\pm 5a = 2 + 33/b$. This leads to two sets of paired solutions, and again only one is new: $(m, n, a, b) = (3, \pm 1, \pm 7, 1)$.

For any other solutions we would have $|a| \geq 5$ and $|k| \geq 7$. Because n is a nonzero integer and $n = (k + 4a)/(ak - 2)$, we must have

$$|ak - 2| \leq |k + 4a|. \qquad (*)$$

If a and k are of opposite sign, then because $|a| > 1$ and $|k| > 4$, we have

$$|ak - 2| = |a||k| + 2 > \max\{|k| - 4|a|, 4|a| - |k|\} = |k + 4a|,$$

contradicting $(*)$. If a and k have the same sign, then

$$\begin{aligned} |ak - 2| &= |a||k| - 2 = 4|a| + |k| + (|a| - 1)(|k| - 4) - 6 \\ &\geq |k + 4a| + 6 > |k + 4a|, \end{aligned}$$

again contradicting $(*)$. Hence there are no other solutions.

The list of solutions consists of eleven solution pairs: $(m, n) = (27, \pm 5)$, $(19, \pm 13)$, $(17, \pm 7)$, $(3, \pm 1)$, $(3, \pm 7)$, $(1, \pm 1)$, $(1, \pm 3)$, $(-1, \pm 1)$, $(-3, \pm 1)$, $(-3, \pm 5)$, and $(-11, \pm 3)$.

Also solved by Herb Bailey, Chip Curtis, Daniele Donini (Italy), Dmitry Fleishman, Mikael Lindahl, Achilleas Sinefakopoulos (Greece), Albert Stadler (Switzerland), H. T. Tang, Li Zhou, and the proposer. There were four incorrect submissions.

Answers

Solutions to the Quickies from page 234.

A941. Because $0 < a_0 < 1$, there is a θ, $0 < \theta < \pi/2$ such that $a_0 = \cos\theta$. Noting that

$$\cos\left(\frac{\theta}{2}\right) = \sqrt{\frac{1+\cos\theta}{2}},$$

we conclude that $a_n = \cos(\theta/2^n)$. In addition, because

$$\cos\left(\frac{\theta}{2}\right) = \frac{\sin\theta}{2\sin(\frac{\theta}{2})},$$

it follows that

$$b_n = \frac{\sin\theta}{2^n \sin(\frac{\theta}{2^n})}.$$

Thus

$$\lim_{n\to\infty} b_n = \frac{\sin\theta}{\theta}.$$

Note: There are other values of θ for which $\cos\theta = a_0$. However, if θ is not between $-\pi/2$ and $\pi/2$, then for some n, $\cos(\theta/2^n) < 0$. In such cases we would have to adjust the half angle formula to reflect the recursive definition of a_n. This would lead to some sine expressions in place of some cosine expressions, and would in turn lead to a different trigonometric expression for b_n.

Added in page proofs: In a late communication, the problem poser reports that this problem appears in *Calculus*, Vol. IIa, by Negrepontis, Giotopoulos, and Giannacoulas, published by Aithra Publications, Athens, 1995.

A942. Dividing the numerator into the denominator and rearranging the order of the four expressions in the product, we obtain the equivalent expression

$$\left[\left(1+\frac{y}{z}+\frac{a}{z}\right)^{1/4}\left(1+\frac{a}{x}+\frac{y}{x}\right)^{1/4}\left(1+\frac{x}{y}+\frac{z}{y}\right)^{1/4}(a+z+x)^{1/4}\right]^{-4}.$$

By Hölder's Inequality, this expression is

$$\leq \left(\sqrt[4]{1\cdot 1\cdot 1\cdot a} + \sqrt[4]{\frac{y}{z}\cdot\frac{a}{x}\cdot\frac{x}{y}\cdot z} + \sqrt[4]{\frac{a}{z}\cdot\frac{y}{x}\cdot\frac{z}{y}\cdot x}\right)^{-4} = \frac{1}{81a},$$

and equality holds if and only if $x = y = z = a$.

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.

Devlin, Keith, *The Millennium Problems: The Seven Greatest Unsolved Mathematical Puzzles of Our Time*, Basic Books, 2003; x + 237 pp, $16 (P). ISBN 0–465–01730–4.

The Clay Mathematics Institute offered million-dollar prizes for these problems in 2000. Devlin is forthright about the book's limitations: "I do not aim at a detailed description of the problems. It is just not possible to describe most of them accurately in lay terms—or even in terms familiar to someone with a university degree in mathematics.... Rather, my goal is to provide the background to each problem, to describe how it arose, explain what makes it particularly difficult, and give you some sense of why mathematicians regard it as important." Still, it is disconcerting—no doubt more so for the nonmathematical reader!—to see a section titled "What is calculus" and 16 pages later see the Navier-Stokes equations in all the glory of their vector notation (one of the few equations in the book); and infinite sums and products make an appearance. The problems more difficult to understand come later in the book. This book will attract readers and give them valuable illusion of understanding a little more about mathematics and its driving forces.

Osserman, Robert, Mathematics with a moral, *Chronicle of Higher Education* 50 (33) (23 April 2004) B10, `http://chronicle.com/free/v50/i33/33b1001.htm` .

Robert Osserman (MSRI) claims that the last decade has turned mathematicians from "frogs" into "fascinating royalty," as a consequence of Tom Stoppard's play *Arcadia* and Andrew Wiles's announcement of proving Fermat's Last Theorem, both in 1993. He then notes the announcement by Grigori Perelman of proving the Poincaré and Thurston conjectures and Perelman's practice of posting his work as he went along. Osserman provides background by weaving together Riemann's attempt to understand the geometry of space, Poincaré's conjecture about the characterization of spheres, and Thurston's classification of 3-manifolds (not called that here, of course). Osserman asserts that "Perelman has provided an argument that will resolve the Poincaré conjecture" and proceeds to the double moral: if you can't solve the given problem, try a harder one (the Poincaré conjecture could be proved because Perelman tackled the more difficult Thurston conjecture), and—despite the stereotype—mathematicians don't work in isolation and always depend on groundwork laid by others (Perelman's work has inspired other approaches).

Chin, Gilbert, et al. (eds.), Biology by the numbers, *Science* 303 (6 February 2004) 781–805.

This special section of *Science* focuses on mathematics in biology, offering three news articles (biological structures from simple mechanisms, Bayesian clinical trials, models for fibrillation), two opinion pieces (urging integration of biology into a unified university science curriculum, citing uses and abuses of mathematics in biology), and two topical reviews (evolutionary game theory and cellular networks).

Peterson, Ivars, Heads or tails?, `http://www.maa.org/mathland/mathtrek_03_01_04.html` = *Science News Online* (28 February 2004) `http://www.sciencenews.org/articles/`

20040228/mathtrek.asp . Klarreich, Erica, Toss out the toss-up: Bias in heads-or-tails, *Science News* 165 (9) (28 February 2004) 131, www.sciencenews.org/articles/20040228/fob2.asp . J. Laurie Snell et al., The not so random coin toss, *Chance News* 13.02 (Feb. 10, 2004 to March 9, 2004) 4–6; www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_13.02.html . Gelman, Andrew, and Deborah Nolan, You can load a die but you can't bias a coin, *American Statistician* 56 (4) (November 2002) 308–311.

That humans aren't good at randomizing is a cornerstone of an introductory statistics course, and it is well known that different techniques of "randomizing" a coin—spinning vs. tossing vs. balancing it on edge and whacking the table—can produce differing proportions of heads. But who would have guessed that people can't even toss a coin fairly? According to Gelman and Nolan, you cannot weight a coin to bias its result; but Persi Diaconis (Stanford University) and colleagues have shown from experiments that a coin is more likely (51%) to land with the same face up as before it was tossed. The source of the bias is unavoidable wobbling of the coin.

American Mathematical Society, What's New in Mathematics. Phillips, Tony, Math in the Media: Highlights of math news from science literature and the current media. Breen, Mike, Annette Emerson, Allyn Jackson, and Claudia Clark, Math Digest. Malkevitch, Joseph, Feature Column. http://www.ams.org/new-in-math/ .

The AMS provides several online resources about mathematics at a general level, including two registers of articles in the popular press and a monthly column. Math in the Media provides one-paragraph summaries of articles from such sources as *Nature*, *Science*, *Physical Review Letters*, the Associated Press, *New Yorker*, *Chronicle of Higher Education*, and even *The Guardian* and *China Daily*. The archive goes back to 1997; there are usually three to six entries per month, with topics such as mathematical models for cancer, Bayesianism in athletics and in spam filters, a Möbius-strip hydrocarbon, and the largest prime yet. Math Digest provides usually a dozen one-sentence summaries of articles from *Science*, *New Scientist*, and *Nature*. The Feature Column is a monthly essay on a mathematical topic (diagonals, "colorful" mathematics, oriented matroids, compression codes, apportionment), eminently suitable exposition to interest students in a topic. (The main drawback is that each essay is broken into up to a dozen separate Web pages, making downloading or printing it a chore. When will Webmasters learn that no one spends long hours reading from a screen?)

MATHDI Mathematics Didactics Database http://www.emis.de/MATH/DI.html . Institutional Internet access: €320/year. MATHDI 2004 CD-ROM (for Windows): €200. From Gerhard.Koenig@FIZ-Karlsruhe.de .

The MATHDI database contains 100,000 items (1976–2004), offering abstracts and reviews of materials on research and practice in mathematics education, pedagogy, and instruction at all levels, as well as elementary mathematics and its applications, plus similar materials for computer science and information technology. About half the sources are 500 international journals, with the rest being textbooks, conference papers, dissertations, media, and software. This database is a valuable resource for mathematics educators at all levels, particularly since it abstracts many journals (such as *Mathematics Magazine*!) that feature mathematical exposition but are rarely covered in *Mathematical Reviews*. The online trial mode (with the crisp and clean search engine of *Zentralblatt MATH*, cousin of *Mathematical Reviews*) shows a single result from a query. A separate online CD-ROM "demo" mode (with a different and less smooth search engine) shows all results but with some details suppressed (e.g., publication information).

44th International Mathematical Olympiad

Tokyo, Japan

July 13 and 14, 2003

edited by Zuming Feng

Problems

1. Let A be a 101-element subset of the set $S = \{1, 2, \ldots, 1000000\}$. Prove that there exist numbers $t_1, t_2, \ldots, t_{100}$ in S such that the sets

$$A_j = \{x + t_j \mid x \in A\} \quad j = 1, 2, \ldots, 100$$

are pairwise disjoint.

2. Determine all pairs of positive integers (a, b) such that

$$\frac{a^2}{2ab^2 - b^3 + 1}$$

is a positive integer.

3. A convex hexagon is given in which any two opposite sides have the following property: the distance between their midpoints is $\sqrt{3}/2$ times the sum of their lengths. Prove that all the angles of the hexagon are equal.

4. Let $ABCD$ be a cyclic convex quadrilateral. Let P, Q, and R be the feet of perpendiculars from D to lines BC, CA, and AB, respectively. Show that $PQ = QR$ if and only if the bisectors of angles ABC and ADC meet on segment AC.

5. Let n be a positive integer and $x_1, x_2, \ldots, x_n$ be real numbers with $x_1 \leq x_2 \leq \cdots \leq x_n$.

   (a) Prove that

$$\left(\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|\right)^2 \leq \frac{2(n^2-1)}{3}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

   (b) Show that the equality holds if and only if $x_1, x_2, \ldots, x_n$ form an arithmetic sequence.

6. Let p be a prime number. Prove that there exists a prime number q such that for every integer n, the number $n^p - p$ is not divisible by q.

Note: For interested readers, the editors recommend the *USA and International Mathematical Olympiads 2003*. There many of the problems are presented together with a collection of remarkable solutions developed by the examination committees, contestants, and experts, during or after the contests.

Solutions

1. We construct the set $\{t_j\}$ one element at a time using the following algorithm: Let $t_1 = 1 \in S$. For each j, $1 \le j \le 100$, let t_j be the smallest number in S that has not yet been crossed out, and then cross out t_j and all numbers of the form $t_j + |x - y|$ (with $x, y \in A, x \ne y$) that are in S. At each step, we cross out at most $1 + \binom{101}{2} = 5051$ new numbers. After picking t_1 through t_{99}, we have crossed out at most 500049 numbers, so there are always numbers in S that have not been crossed, so there are always candidates for t_j in S. (In fact, we will never need to pick a t_j bigger than 500050.)

   Now, suppose A_j and A_k are not disjoint for some $1 \le j < k \le 100$. Then $x + t_j = y + t_k$ for some x, y in A. Since we cross out t_j immediately after picking it, $t_k \ne t_j$. Also, if $t_k < t_j$, we would have picked it on step j rather than step k (because $j < k$). Thus $t_k > t_j$, and so $x > y$. But this means that $t_k = t_j + x - y = t_j + |x - y|$, so t_k would have been crossed out on step j. This is a contradiction, so all sets A_j are pairwise disjoint.

2. The answers are $(a, b) = (2t, 1)$, or $(t, 2t)$, or $(8t^4 - t, 2t)$ for all positive integers t. It is routine to check the above are indeed solutions of the problem. We prove they are the only possible solutions. Assume that $a^2/(2ab^2 - b^3 + 1) = k$, where k is a positive integer. Then we have

$$a^2 = 2ab^2k - b^3k + k. \qquad (*)$$

   Because both k and a^2 are positive, $2ab^2 - b^3 + 1 > 0$, or, $2a > b - 1/b^2$. Because a and b are positive integers, we have $2a \ge b$. Because k is a positive integer, $a^2 \ge 2ab^2 - b^3 + 1$, or, $a^2 \ge b^2(2a - b) + 1$. Because $2a - b \ge 0$ and $a^2 > b^2(2a - b) \ge 0$, we have

$$a > b \quad \text{or} \quad 2a = b. \qquad (\dagger)$$

   Viewing equation $(*)$ as a quadratic in a, replace a by x to consider the equation

$$x^2 - 2b^2kx + (b^3 - 1)k = 0 \qquad (*')$$

   for fixed positive integers b and k. We assume that $x_1 = a$ is an integer root of equation $(*')$. Then the other root x_2 is also an integer because $x_1 + x_2 = 2b^2k$. Without loss of generality, we assume that $x_1 \ge x_2$. Then $x_1 \ge b^2k > 0$. Furthermore, because $x_1x_2 = (b^3 - 1)k$, we obtain

$$0 \le x_2 = \frac{(b^3 - 1)k}{x_1} \le \frac{(b^3 - 1)k}{b^2k} < b.$$

   If $x_2 = 0$, then $b^3 - 1 = 0$, and so $x_1 = 2k$ and (a, b) can be written in the form of $(2t, 1)$ for some integers t.

   If $x_2 > 0$, then $(a, b) = (x_2, b)$ is a pair of positive integers satisfying the equations $(\dagger)$ and $(*)$. We conclude that $2x_2 = b$, and so

$$k = \frac{x_2^2}{2x_2b^2 - b^3 + 1} = x_2^2 = \frac{b^2}{4},$$

   and $x_1 = b^4/2 - b/2$. Thus, (a, b) can be written in the form of either $(t, 2t)$ or $(8t^4 - t, 2t)$ for some positive integers t.

3. We list with two geometry facts whose proofs are relatively simple.

LEMMA 1. *In triangle PQS, let M be the midpoint of side QS. If* $2PM \geq \sqrt{3}QS$, *then* $\angle SPQ \leq 60°$. *Equality holds if and only PQS is equilateral.*

LEMMA 2. *Let ABCDEF be a convex hexagon with parallel opposite sides, that is,* $AB \parallel DE$, $BC \parallel EF$, *and* $CD \parallel FA$. *Assume that each pair of three diagonals AD, BE, CF form a* $60°$ *angle and that* $AD = BE = CF$. *Then the hexagon is equal angular. Furthermore, the hexagon can be obtained by cutting three congruent triangles from each corner of a equilateral triangle.*

To solve the given problem, let diagonals AD and BF, CF and DA, and EB and FC meet at X, Y, and Z, respectively. Let M and N be the midpoints of sides AB and DE. We want to "add" the lengths of AB and DE without violating the midpoints constraint. Let G and H be points such that $AMHD$ and $BMGE$ are parallelograms. Thus, $AD = MH$, $BE = MG$, $AD \parallel MH$, and $BE \parallel MG$. We conclude that $\angle AXB = \angle GMH$ and that $MG = MH$ if and only if $AD = BE$.

Because N is the midpoint of segment DE, it is not hard to see that $GEHD$ is also a parallelogram, and so N is the midpoint of segment GH. Note that $AB + ED = GE + ED + DH \geq GH$, and equality holds if and only if $AB \parallel DE$. In triangle MGH, median MN is at least $\sqrt{3}/2$ opposite side GH. By Lemma 1, we conclude that $\angle GMH \geq 60°$, that is, $\angle AXB = \angle ZXY \geq 60°$. Likewise, $\angle XYZ$, $\angle YZX \geq 60°$. Thus, all inequalities hold, implying that all the conditions of Lemma 2 hold, from which our desired follows.

4. The condition that $ABCD$ be cyclic is not necessary. As usual, we set $\angle ABC = \beta$, $\angle BCA = \gamma$, and $\angle CAB = \alpha$. Because $\angle DPC = \angle CQD = 90°$, quadrilateral $CPDQ$ is cyclic with CD a diameter of the circumcircle. By the Extended Law of Sines, we have $PQ = CD \sin \angle PCQ = CD \sin(180° - \gamma) = CD \sin \gamma$. Likewise, by working with cyclic quadrilateral $ARQD$, we find $RQ = AD \sin \alpha$. Hence, $PQ = RQ$ if and only if $CD \sin \gamma = AD \sin \alpha$. Applying the Law of Sines to triangles BAC, we conclude that

$$PQ = RQ \quad \text{if and only if} \quad \frac{AB}{BC} = \frac{AD}{CD}. \tag{$*$}$$

On the other hand, let bisectors of $\angle CBA$ and $\angle ADC$ meet segment AC at X and Y, respectively. By the Angle-bisector Theorem, we have $AX/CX = AB/BC$ and $AY/CY = AD/CD$. Hence, the bisectors of $\angle ABC$ and $\angle ADC$ meet on segment AC if and only if $X = Y$, or,

$$\frac{AB}{BC} = \frac{AX}{CX} = \frac{AY}{CY} = \frac{AD}{CD}. \tag{$**$}$$

Our desired result follows from relations $(*)$ and $(**)$.

5. By the Cauchy–Schwarz Inequality, we have

$$\left(\sum_{i,j=1}^{n}(x_i - x_j)^2\right)\left(\sum_{i,j=1}^{n}(i-j)^2\right) \geq \left(\sum_{i,j=1}^{n}|i-j||x_i - x_j|\right)^2.$$

It suffices to show that

$$\sum_{i,j=1}^{n}(i-j)^2 = \frac{n^2(n^2-1)}{6} \tag{$\dagger$}$$

and

$$\sum_{i,j=1}^{n} |i-j||x_i - x_j| = \frac{n}{2} \sum_{i,j=1}^{n} |x_i - x_j|. \qquad (\ddagger)$$

Identity (†) comes from

$$\sum_{i,j=1}^{n} (i-j)^2 = \sum_{i,j=1}^{n} (i^2 + j^2 - 2ij) = 2\sum_{i,j=1}^{n} i^2 - 2\left(\sum_{i=1}^{n} i\right)\left(\sum_{j=1}^{n} j\right)$$
$$= 2n \cdot \frac{n(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{2} = \frac{n^2(n^2-1)}{6}.$$

To establish identity (‡), we compare the coefficients of x_i, $1 \le i \le n$, on both sides of the identity. The coefficient of x_i on the left-hand side is equal to

$$(i-1) + (i-2) + \cdots + [(i-(i-1)] - [(i+1) - i] - \cdots - (n-i)$$
$$= \frac{i(i-1)}{2} - \frac{(n-i)(n-i+1)}{2} = \frac{n(2i-n-1)}{2}.$$

On the other hand, the coefficient of x_i on the right-and side is equal to

$$\frac{n}{2}\Big[\underbrace{(1+1+\cdots+1)}_{i-1 \text{ times}} - \underbrace{(1+1+\cdots+1)}_{n-i \text{ times}}\Big] = \frac{n}{2}(2i-n-1).$$

Therefore, identity (‡) is true and the proof of part (a) is complete.

In our proof of the desired inequality, the only step where we had an inequality rather than an equality was when we applied the Cauchy-Schwarz Inequality. Equality holds in the problem if and only if we have equality in the Cauchy-Schwarz Inequality, that is, $(x_i - x_j)/(i-j) = d$ is a constant for $1 \le i, j \le n$. In particular, $x_i - x_1 = d(i-1)$, that is, $x_1, x_2, \ldots, x_n$ is an arithmetic sequence.

6. We approach indirectly by assuming that such q does not exist. Then for any fixed prime q, there is a positive integer n such that $n^p - p$ is divisible by q, that is

$$n^p \equiv p \pmod{q}. \qquad (*)$$

If q divides n, then q divides p, and so $q = p$. We further assume that $q \neq p$. Hence q does not divide n. Let d_n be the order of n modulo q. By Fermat's Little Theorem, d_n divides $q-1$. For the positive integer n, because $n^p \equiv p \pmod{q}$, we have $n^{pd_p} \equiv p^{d_p} \equiv 1 \pmod{q}$. Thus, one can show that d_n divides both $q-1$ and pd_p, implying that d_n divides $\gcd(q-1, pd_p)$.

Now we pick a prime q such that (a) q divides $\frac{p^p-1}{p-1} = 1 + p + \cdots + p^{p-1}$, and (b) p^2 does not divide $q-1$. First we show that such a q does exist. Note that $1 + p + \cdots + p^{p-1} \equiv 1 + p \not\equiv 1 \pmod{p^2}$. Hence there is a prime divisor of $1 + p + \cdots + p^{p-1}$ that is not congruent to 1 modulo p^2, and we can choose that prime to be our q.

By (a), $p^p \equiv 1 \pmod{q}$ (and $p \neq q$), implying that d_p divides p, that is, $d_p = p$ or $d_p = 1$. If $d_p = 1$, then $p \equiv 1 \pmod{q}$. If $d_p = p$. Then d_n divides $\gcd(p^2, q-1)$. By (b), the possible values of d_n are 1 and p, implying that $n^p \equiv 1 \pmod{q}$. By relation (*), we conclude $p \equiv 1 \pmod{q}$.

Thus, in any case, we have $p \equiv 1 \pmod{q}$. But then by (a), $0 \equiv 1 + p + \cdots + p^{p-1} \equiv p \pmod{q}$, implying that $p = q$, which is a contradiction. Therefore our

original assumption was wrong, and there is a q such that for every integer n, the number $n^p - p$ is not divisible by q.

2003 Olympiad Results

The top twelve students on the 2003 USAMO were (in alphabetical order):

Boris Alexeev	Cedar Shoals High School	Athens, GA
Jae Bae	Academy of Advancement in Science and Technology	Hackensack, NJ
Daniel Kane	West High School	Madison, WI
Anders Kaseorg	Charlotte Home Educators Association	Charlotte, NC
Mark Lipson	Lexington High School	Lexington, MA
Tiankai Liu	Phillips Exeter Academy	Exeter, NH
Po-Ru Loh	James Madison Memorial High School	Madison, WI
Po-Ling Loh	James Madison Memorial High School	Madison, WI
Aaron Pixton	Vestal Senior High School	Vestal, NY
Kwokfung Tang	Phillips Exeter Academy	Exeter, NH
Tony Zhang	Phillips Exeter Academy	Exeter, NH
Yan Zhang	Thomas Jefferson High School of Science and Technology	Alexandria, VA

Tiankai Liu and Po-Ru Loh, both with perfect scores, were the winners of the Samuel Greitzer-Murray Klamkin award, given to the top scorer(s) on the USAMO. Mark Lipson placed third on the USAMO. They were awarded college scholarships of $5000, $5000, and $2000, respectively, by the Akamai Foundation. The Clay Mathematics Institute (CMI) award, for a solution of outstanding elegance, and carrying a $3000 cash prize, was presented to Tiankai Liu for his solution to USAMO Problem 6. Two additional CMI awards, carrying a $1000 cash prize each, were presented to Anders Kaseorg and Matthew Tang for their solutions to USAMO Problem 5.

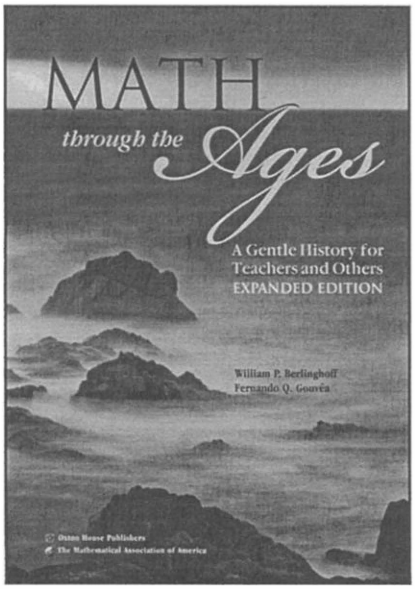
The USA team members were chosen according to their combined performance on the 32nd annual USAMO and the Team Selection Test that took place at the Mathematical Olympiad Summer Program (MOSP) held at the University of Nebraska-Lincoln, June 15–July 5, 2003. Members of the USA team at the 2003 IMO (Tokyo, Japan) were Daniel Kane, Anders Kaseorg, Mark Lipson, Po-Ru Loh, Aaron Pixton, and Yan Zhang. Zuming Feng (Phillips Exeter Academy) and Gregory Galperin (Eastern Illinois University) served as team leader and deputy leader, respectively. The team was also accompanied by Melanie Wood (Princeton University) and Steven Dunbar (University of Nebraska-Lincoln), as the observer of the team leader and deputy leader, respectively.

At the 2003 IMO, gold medals were awarded to students scoring between 29 and 42 points, silver medals to students scoring between 19 and 28 points, and bronze medals to students scoring between 13 and 18 points. There were 37 gold medalists, 69 silver medalists, and 104 bronze medalists. There were three perfect papers (Fu from China, Le and Nguyen from Vietnam) on this very difficult exam. Loh's 36 tied for 12th place overall. The team's individual performances were as follows:

Kane	GOLD Medallist	Loh	GOLD Medallist
Kaseorg	GOLD Medallist	Pixton	GOLD Medallist
Lipson	SILVER Medallist	Y. Zhang	SILVER Medallist

(continued on p. 232)

Don't miss this...



Math Through the Ages

A Gentle History for Teachers & Others

William B. Berlinghoff & Fernando O. Gouvêa

Classroom Resource Materials

Catalog Code: MEX/JR
288 pp., Paperbound, 2004
ISBN: 0-88385-736-7
List: $39.95
Member: $29.95

Expanded edition, jointly published by the MAA & Oxton House Publishing

A marvelous book...very well organized and thus very user-friendly...a wonderful resource for teachers. It provides a superb overview of the history of mathematics and details on the development of the principal ideas of mathematics at the primary, secondary, and beginning college levels.

—Victor Katz,
University of the District of Columbia

This is a beautiful, important book, a pleasure to read, in which the history recounted fully illuminates the mathematical ideas, and the ideas themselves are superbly explained; a wonderful accomplishment.

—Barry Mazur,
Harvard University

Very seldom does one come across a mathematics history text that can be recommended to middle school teachers as well as to those in college and universities. ...The bibliography the authors present is a treasure in itself. I highly recommend this book for every math teacher's personal library.

—Karen Dee Michalowicz,
The Langly School, McLean, VA

from the MAA!

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036