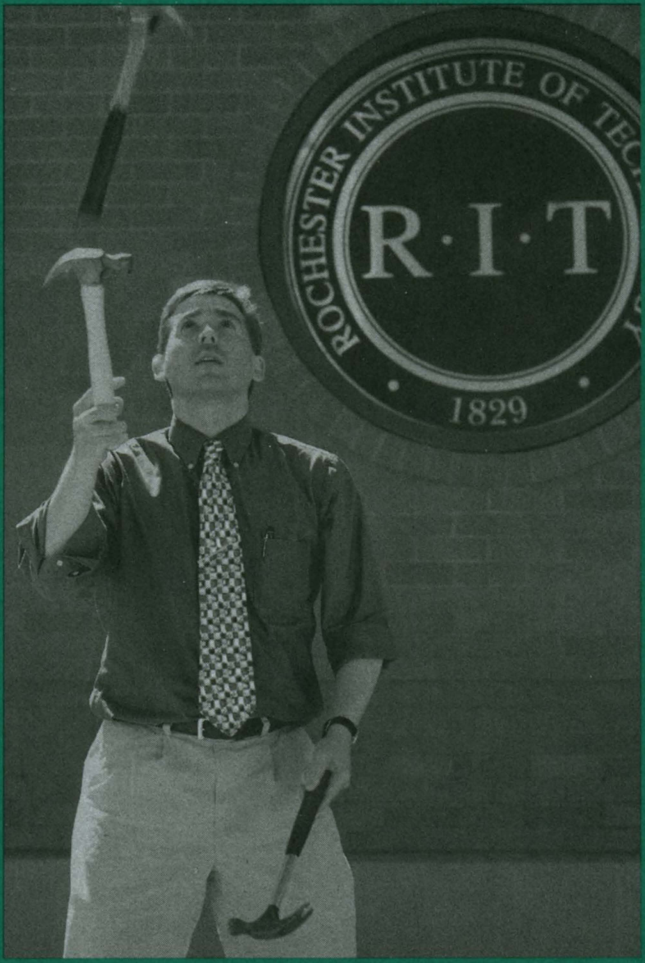
Vol. 79, No. 4, October 2006

MATHEMATICS MAGAZINE



Hammer Juggling

- Hammer Juggling, Rotational Instability, and Eigenvaules
- The Volume Principle
- How Much Does a Matrix of Rank *k* Weigh?
- Folding Optimal Polygons from Squares

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 Allen Schwenk, Editor, Mathematics Magazine, Department of Mathematics, Western Michigan University, Kalamazoo, MI, 49008. 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*.

The cover image shows author Carl Lutzer intently juggling three hammers. See his article for an eigenvalue analysis of rotational instability. The photo has been provided by Dave Londres, a second year photojournalism student at Rochester Institute of Technology. A native of Cherry Hill, NJ, Dave enjoys technology, cars, sports and making photos every day. Feel free to contact him at DLondres@gmail.com.

AUTHORS

Carl Lutzer is an Assistant Professor of Mathematics at the Rochester Institute of Technology. He earned his Ph.D. from the University of Kentucky under the direction of Peter Hislop. His mathematical research interests currently lie in partial differential equations and dynamical systems. In addition to mathematics and teaching, he enjoys writing fiction, and fencing (sabre).

William Dickinson fell in love with geometry in the ninth grade when he attempted to trisect the angle using ruled graph paper. This love continued until the end of his senior year at Cornell University which culminated with a thesis on circle packings on the flat torus. Though he did enjoy much of his thesis work at the University of Pennsylvania in differential geometry, he did not truly rediscover his love of geometry until he started to teach geometry and mentor undergraduate research in this area. These mentoring activities have been (and, he hopes, will continue to be) fruitful and produced the present article. When not at work, he enjoys watching movies with his wife, Andrea, and working in his basement woodshop with his three sons: Andrew, Matthew and Simon.

Kristina Lund graduated from Grand Valley State University utterly obsessed with mathematical research. This infatuation began during an undergraduate summer research project in geometry with a wonderful mentor and some pretty cool results. Kristina is currently a busy graduate student at the University of Nebraska-Lincoln. She wishes to thank her family and friends for their support in her decision to continue her education in mathematics.

Kent E. Morrison is Chair of the Mathematics Department at California Polytechnic State University in San Luis Obispo, where he has taught for over 25 years. He has also taught at Utah State University, Haverford College, and the University of California at Santa Cruz, where he received his Ph.D. in 1977. In recent years his research interests have centered on enumerative and algebraic combinatorics. This article is the result of a summer student research project with Theresa Migler and Mitch Ogle. His 1987 article in this MAGAZINE, Groups generated by perfect shuffles, written with Steve Medvedoff, grew out of an earlier student project.

Theresa Migler is a graduate student at California Polytechnic State University in San Luis Obispo where she received her bachelors degree in 2004. She hopes to pursue a Ph.D. and become a mathematics professor.

Mitchell L. Ogle is currently finishing his undergraduate degree at California Polytechnic State University in San Luis Obispo with a double major in mathematics and electrical engineering. He hopes to attend graduate school to study functional analysis and control theory. This is his first article in this MAGAZINE.

David Dureisseix entered the Ecole Normale Supérieure de Cachan (ENS Cachan, France) in 1988, and received the Agrégation in Mechanics in 1991. He defended his Ph.D. thesis in 1997, and became Assistant Professor at the ENS Cachan. He is now a Professor at the University Montpellier 2 in the Mechanics and Civil Engineering Laboratory. Though his teaching and research interests are related to mechanical engineering, technological design, and computational mechanics, he is also member of several origami societies.

Vol. 79, No. 4, October 2006

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

MAA Advertising
c/o Marketing General, Inc.
209 Madison Street Suite 300
Alexandria VA 22201

Phone: 866-821-1221
Fax: 866-821-1221
E-mail: rhall@marketinggeneral.com

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

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

ARTICLES

Hammer Juggling, Rotational Instability, and Eigenvalues

CARL V. LUTZER
Rochester Institute of Technology
Rochester, NY 14623
Carl.Lutzer@rit.edu

Introduction

Get a hammer. Seriously, get a hammer. As an experiment, hold the hammer in front of you with its head pointing up. Toss it upward (CAREFULLY!), end-over-end, and catch it after one revolution. The orientation of the hammer when you catch it will be the same as when you tossed it.

As a second experiment, hold the hammer in front of you with its head pointing sideways, to the right. Toss the hammer upward, end-over-end, and catch it after one revolution. This time, the orientation changes—the head pointed to the right when you tossed it, but points to the left when you catch it!

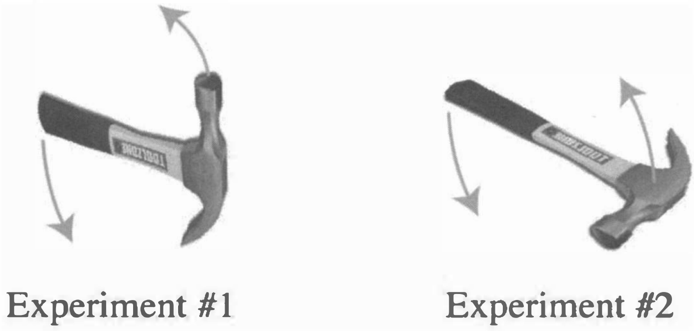


Figure 1 Hammer juggling and unstable rotation

Many people suggest that this strange 1/2-twist in experiment #2 is due to the asymmetry of the hammer's mass distribution, but the same kind of thing will happen with a book, or wallet, or any object with three distinct dimensions. (Try it! Use a rubber-band to keep the wallet or book closed.) We don't always see a *half*-twist (that will depend on the particular orientation of the object when you release it), but we almost always see a twist. Why? The answer is well known to the physics community, but is documented primarily in their parlance. The following exposition explains this phenomenon from a mathematician's point of view. The governing equations will be quickly derived, and the supporting linear algebra will be explored.

We assume that the reader has basic knowledge of multivariate calculus, and is aware that $e^{i\phi} = \cos\phi + i\sin\phi$. We also assume that the reader is familiar with eigenvalues, eigenvectors, linear independence, and understands that a proper choice of basis will diagonalize a symmetric matrix $M \in \mathbb{R}^{3\times 3}$.

The basics

In this section we begin with simple definitions of basic vocabulary, cite of the governing equations of motion, and then proceed with the salient calculations. Proofs of important assertions, and a derivation of the equations of motion are postponed until later sections so that we can focus on answering the question of why the hammer performs a half-revolution in Experiment #2 but not in Experiment #1.

Vocabulary

Angular Velocity Suppose an object is revolving about some particular axis, much like a child's spinning top. The *angular velocity* of the object, denoted by ω, is a vector that points in the direction of that axis. The magnitude of ω is $2\pi\gamma$, where $\gamma \geq 0$ is the number of revolutions per second. As you might infer from the example of the spinning top, the angular velocity vector may change direction and length as time evolves.

Newton's Second Law Most people cite Newton's Second Law as $F = ma$, which isn't quite right. Newton's Second Law says that force is the instantaneous change in momentum. In the case of linear force we write $F = d\rho/dt$ where $\rho = mv$ is the linear momentum of a mass m traveling with velocity v. In the case of angular force and angular momentum we write $\tau = dL/dt$ where τ means torque and L denotes angular momentum (discussed in detail later).

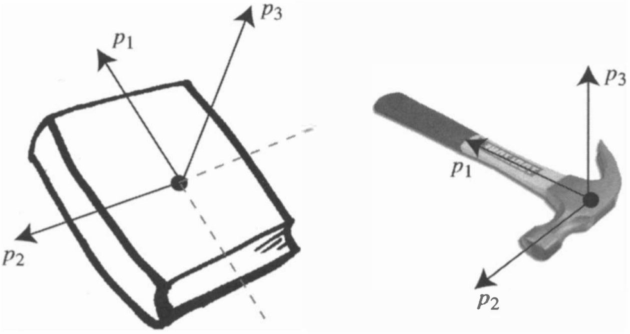
Euler's equation For reasons that will be explained later, the governing equation of motion is

$$\tau = M\dot{\omega} + \omega \times M\omega, \tag{1}$$

where $M \in \mathbb{R}^{3\times3}$ is a symmetric matrix and $\dot{\omega}$ denotes the derivative of ω with respect to time. (This "dot notation" is used throughout the rest of the article to denote differentiation with respect to time.) In later sections we'll see that (1), called *Euler's equation*, is just a fancy restatement of the fact that $\tau = dL/dt$.

Calculations Because the matrix M is symmetric, its eigenvalues are all real, and eigenvectors associated with distinct eigenvalues are orthogonal. In fact, it happens that all the eigenvalues of M are positive! In the case of the hammer, they're also distinct so we label them in increasing order: $0 < \lambda_1 < \lambda_2 < \lambda_3$.

Physicists refer to M as the *moment-of-inertia tensor*, and they often use the letter I (for "inertia") to denote this matrix. (We use M in this exposition to avoid confusion with the identity matrix.) The eigenvalues of M are called the *principal moments of inertia,* and their corresponding unit-eigenvectors are called the *principal axes of rotation.* These unit-eigenvectors, which we'll denote by p_1, p_2, and p_3 respectively, point along "the axes of" the object in question. For example, pull a textbook off of the shelf. It has length, width, and height. The vector p_1 points in the direction of the length, the vector p_2 points in the direction of the width, and the vector p_3 points in the direction of the height (see the figure, below). Notice that, listed in the order prescribed by our indexing, the dimensions of the book are decreasing: length > width > height. If you accepted the earlier invitation to try the experiment with another object (with three distinct dimensions), you found that the rotation was unstable when the axis of rotation was parallel to p_2, which corresponds to the "middle" dimension. This will always be the case, as we'll see in a moment.



Vectors p_1, p_2, p_3 form an orthonormal basis for $\mathbb{R}^3$, so any angular velocity can be expressed as a linear combination of them: $\omega = \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3$. (Recall that ω may change with time, so the scalars α_1, α_2 and α_3 are functions of time.) Moreover, the matrix M is diagonal in the basis $\{p_1, p_2, p_3\}$.

$$M = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}.$$

So when the rotation is free from external torque and we use $\{p_1, p_2, p_3\}$ as our basis, equation (1) becomes

$$\lambda_1 \dot{\alpha}_1 + (\lambda_3 - \lambda_2)\alpha_2\alpha_3 = 0 \tag{2}$$

$$\lambda_2 \dot{\alpha}_2 + (\lambda_1 - \lambda_3)\alpha_1\alpha_3 = 0 \tag{3}$$

$$\lambda_3 \dot{\alpha}_3 + (\lambda_2 - \lambda_1)\alpha_1\alpha_2 = 0 \tag{4}$$

Suppose the object in question (the hammer, in this case) were to rotate about the axis p_1. Then $\alpha_2(0) = 0 = \alpha_3(0)$ and it follows from equations (2)–(4) that α_2 and α_3 *stay* zero. Of course, we see the same behavior whether we rotate about p_1, p_2 or p_3. But rotating about one of the principal axes—*exactly*—is highly unlikely, even if we are meticulous in our efforts to make it happen. So what happens when the object in question rotates about an axis that is very *close* to one of the principal axes?

Stable rotation Suppose ω is initially very close to p_1. Then $\alpha_2(0) \approx \alpha_3(0) \approx \varepsilon \approx 0$, so the second summand on the right-hand side of (2) is order ε^2.

$$\lambda_1 \dot{\alpha}_1 + \underbrace{(\lambda_3 - \lambda_2)\alpha_3\alpha_2}_{O(\varepsilon^2)} = 0. \tag{5}$$

The analogous terms in (3) and (4) are only order ε, so a linear approximation of Euler's equation is

$$\lambda_1 \dot{\alpha}_1 \approx 0 \tag{6}$$

$$\lambda_2 \dot{\alpha}_2 + (\lambda_1 - \lambda_3)\alpha_1\alpha_3 = 0 \tag{7}$$

$$\lambda_3 \dot{\alpha}_3 + (\lambda_2 - \lambda_1)\alpha_1\alpha_2 = 0 \tag{8}$$

Equation (6) indicates that α_1 is constant (or nearly so). This reduces the problem to a system of two equations in two unknowns. Solving (7) and (8) for $\dot{\alpha}_2$ and $\dot{\alpha}_3$, respectively, gives us

$$\begin{pmatrix} \dot{\alpha}_2 \\ \dot{\alpha}_3 \end{pmatrix} = \begin{bmatrix} 0 & \dfrac{(\lambda_3 - \lambda_1)\alpha_1}{\lambda_2} \\ \dfrac{(\lambda_1 - \lambda_2)\alpha_1}{\lambda_3} & 0 \end{bmatrix} \begin{pmatrix} \alpha_2 \\ \alpha_3 \end{pmatrix} \tag{9}$$

which we write as the 2 × 2 system $\dot{x} = Ax$. The eigenvalues of A are

$$\pm i\sqrt{\frac{(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_1)\alpha_1^2}{\lambda_2\lambda_3}},$$

which we will denote by $\pm i\phi$. Suppose the associated eigenvectors are $\vec{a}_1, \vec{a}_2 \in \mathbb{C}^2$. Then, since these vectors are linearly independent, there are scalars $c_1, c_2 \in \mathbb{C}$ such that $c_1\vec{a}_1 + c_2\vec{a}_2 = (\alpha_2(0), \alpha_3(0))^T$. Note that c_1 and c_2 are "small" since $\|\vec{a}_1\| = \|\vec{a}_2\| = |e^{\pm i\phi}| = 1$ and $\alpha_2(0) \approx 0 \approx \alpha_3(0)$. Now by defining $x(t) = c_1 e^{i\phi t}\vec{a}_1 + c_2 e^{-i\phi t}\vec{a}_2$ we have

$$\begin{aligned} \dot{x}(t) &= \frac{d}{dt}\left(c_1 e^{i\phi t}\vec{a}_1 + c_2 e^{-i\phi t}\vec{a}_2\right) = c_1\left(\frac{d}{dt}e^{i\phi t}\right)\vec{a}_1 + c_2\left(\frac{d}{dt}e^{-i\phi t}\right)\vec{a}_2 \\ &= c_1 e^{i\phi t}(i\phi)\vec{a}_1 + c_2 e^{-i\phi t}(-i\phi)\vec{a}_2 = c_1 e^{i\phi t}A\vec{a}_1 + c_2 e^{-i\phi t}A\vec{a}_2 \\ &= A(c_1 e^{i\phi t}\vec{a}_1 + c_2 e^{-i\phi t}\vec{a}_2) = Ax(t) \end{aligned}$$

The function $x(t)$ solves (9) with the correct initial data so, since that solution is unique, $x(t) = (\alpha_2(t), \alpha_3(t))^T$. It follows that α_2 and α_3 not only start small but *stay* small. That is, ω *stays* close to $\alpha_1 p_1$.

In fact, ω revolves around p_1 as the system evolves. It's easy to follow through the same calculations to derive the same behavior when the axis of rotation is close to p_3, but something very different happens when ω is initially near p_2.

Unstable rotation If we begin with ω very near to p_2, $\alpha_1(0) \approx 0 \approx \alpha_3(0)$, so a linear approximation of Euler's equation is

$$\lambda_1\dot{\alpha}_1 + (\lambda_3 - \lambda_2)\alpha_2\alpha_3 = 0 \tag{10}$$

$$\lambda_2\dot{\alpha}_2 \approx 0 \tag{11}$$

$$\lambda_3\dot{\alpha}_3 + (\lambda_2 - \lambda_1)\alpha_1\alpha_2 = 0. \tag{12}$$

Equation (11) indicates that α_2 is constant (or nearly so). This reduces the problem to a system of two equations and two unknowns.

$$\begin{pmatrix} \dot{\alpha}_1 \\ \dot{\alpha}_3 \end{pmatrix} = \begin{bmatrix} 0 & \dfrac{(\lambda_2 - \lambda_3)\alpha_2}{\lambda_1} \\ \dfrac{(\lambda_1 - \lambda_2)\alpha_2}{\lambda_3} & 0 \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_3 \end{pmatrix} \tag{13}$$

The coefficient matrix has eigenvalues

$$\pm\sqrt{\frac{(\lambda_2 - \lambda_3)(\lambda_1 - \lambda_2)\alpha_2^2}{\lambda_1\lambda_3}},$$

which we denote by $\pm\phi$. Suppose the associated eigenvectors are $\vec{a}_1, \vec{a}_2 \in \mathbb{R}^2$. Then the solution to (13) is $x = c_1 e^{\phi t}\vec{a}_1 + c_2 e^{-\phi t}\vec{a}_2$, where c_1 and c_2 are chosen to achieve $x(0) = (\alpha_1(0), \alpha_3(0))^T$. It's important to note that $c_2 e^{-\phi t}\vec{a}_2$ vanishes quickly but that

$c_1 e^{\phi t}\vec{a}_1$ grows exponentially. That is, though α_1 and α_3 started small, they don't stay that way, and it's exactly this instability that makes the hammer change its orientation.

Rolling up our sleeves

Now we undertake the task of supporting the assertions made about the matrix M (that it's symmetric and that all its eigenvalues are positive) and explaining Euler's equation. We begin by defining angular momentum and establishing its relationship to angular velocity.

The relationship between L and ω Suppose a rigid body rotates about the line through its center-of-gravity defined by the vector ω. Taking the center-of-gravity as our origin, an atom at $r_j = (x_j, y_j, z_j)$ has a linear velocity of $v_j = \omega \times r_j$ (see Figure 2). The *angular momentum* of that atom is defined to be $L_j = r_j \times m_j v_j$, where m_j is its mass. That is, $L_j = m_j(r_j \times (\omega \times r_j))$. Grinding through the cross products brings us to

$$L_j = \begin{bmatrix} m_j(y_j^2 + z_j^2) & -m_j x_j y_j & -m_j x_j z_j \\ -m_j x_j y_j & m_j(x_j^2 + z_j^2) & -m_j y_j z_j \\ -m_j x_j z_j & -m_j y_j z_j & m_j(x_j^2 + y_j^2) \end{bmatrix} \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix}. \tag{14}$$

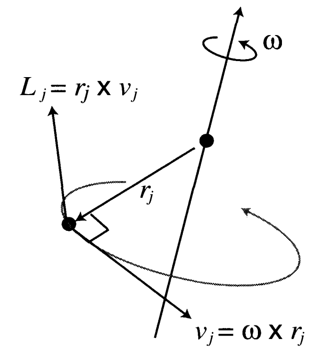


Figure 2 Angular velocity and angular momentum

The angular momentum of the entire object is just the sum of the angular momenta of all its atoms. Summing (14) over all particles gives us

$$L = \underbrace{\begin{bmatrix} \sum_j m_j(y_j^2 + z_j^2) & -\sum_j m_j x_j y_j & -\sum_j m_j x_j z_j \\ -\sum_j m_j x_j y_j & \sum_j m_j(x_j^2 + z_j^2) & -\sum_j m_j y_j z_j \\ -\sum_j m_j x_j z_j & -\sum_j m_j y_j z_j & \sum_j m_j(x_j^2 + y_j^2) \end{bmatrix}}_{\text{This is the matrix } M} \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix}.$$

Defining M to be the coefficient matrix on the right-hand side, we can write $L = M\omega$. We remark that the symmetry of M is now apparent, but why are its eigenvalues always positive and why does it play a role in Euler's equation? These questions are answered in the remaining sections.

The eigenvalues of M We begin our investigation into the eigenvalues of M by writing

$$M = (\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2)I - A^T A, \tag{15}$$

where $\vec{x}_j = \sqrt{m_j}\, x_j$, $\vec{y}$ and $\vec{z}$ are the corresponding vectors of scaled y and z coordinates, and A is the matrix whose columns are $A_{\cdot 1} = \vec{x}$, $A_{\cdot 2} = \vec{y}$, and $A_{\cdot 3} = \vec{z}$. That is, M is a perturbation of the matrix $(\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2)I$, which has a single eigenvalue whose algebraic multiplicity is three. The effect of this perturbation on the set of eigenvalues depends on the "size" of the perturbation. We measure the "size" of a linear function $\mathcal{L} : \mathbb{R}^3 \to \mathbb{R}^3$ with the *operator norm:*

$$\|\mathcal{L}\|_* \overset{\text{def}}{=} \max_{\|u\|=1} \|\mathcal{L}u\|, \tag{16}$$

where $\|v\| = \sqrt{v \cdot v}$ is the standard norm $\mathbb{R}^3$. (The fact that a maximum is always achieved follows from the Heine-Borel Theorem, which is usually taught in a course such as Real Analysis. Its 1-dimensional version is known to calculus students as the Extreme Value Theorem: *A continuous function on a closed interval achieves an absolute maximum value.*) Before continuing, we suggest that the reader verify the following lemma.

LEMMA 1. *Suppose $A, B : \mathbb{R}^3 \to \mathbb{R}^3$ are linear operators. Then*

1. $\|Ax\| \le \|A\|_*\|x\|$
2. $\|AB\|_* \le \|A\|_*\|B\|_*$
3. $\|A\|_* = \|A^T\|_*$

Now let us suppose that u is a unit-eigenvector of M associated with the eigenvalue λ. Then

$$\lambda u = Mu = \big((\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2)I - A^T A\big)\, u$$

from which it follows that $A^T Au = \big(\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2 - \lambda\big)\, u$. That is, u is an eigenvector of $A^T A$. The strategy of our proof is to use this fact to show that

$$\underbrace{\left|\underbrace{\big(\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2\big)}_{\text{anchor value} > 0} - \lambda\right|}_{\text{distance from } \lambda \text{ to anchor value}} < \|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2,$$

from which it follows that $\lambda > 0$. For example, if it was the case that $\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2 = 5$, showing $|5 - \lambda| < 5$ would imply that $\lambda > 0$.

Since $\|u\| = 1$, we have

$$\begin{aligned} \left|\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2 - \lambda\right| &= \left\|\big(\|\vec{x}\|^2 + \|\vec{y}\|^2 + \|\vec{z}\|^2 - \lambda\big)\, u\right\| \\ &= \|A^T Au\| \le \|A^T A\|_* \\ &\le \|A^T\|_*\|A\|_* = \|A^T\|_*^2 \end{aligned} \tag{17}$$

so the proof rests on our estimate of $||A^T||_*$. For any unit vector, v,

$$\begin{aligned} ||A^T v|| &= \sqrt{(\vec{x} \cdot v)^2 + (\vec{y} \cdot v)^2 + (\vec{z} \cdot v)^2} \\ &\leq \sqrt{||\vec{x}||^2 + ||\vec{y}||^2 + ||\vec{z}||^2}. \end{aligned} \tag{18}$$

Note that equality could only occur in (18) if some unit vector v were parallel (or antiparallel) to all three vectors, $\vec{x}$, $\vec{y}$ and $\vec{z}$. But this could only happen if the object in question were 1-dimensional! Restricting ourselves to 3-dimensional objects, we can rewrite (18) as

$$||A^T v|| < \sqrt{||\vec{x}||^2 + ||\vec{y}||^2 + ||\vec{z}||^2}. \tag{19}$$

Since (19) is true for all unit vectors v, it's true when $||A^T v||$ achieves its maximum and, thus, $||A^T||_* < \sqrt{||\vec{x}||^2 + ||\vec{y}||^2 + ||\vec{z}||^2}$. Returning to (17), we have

$$\left| ||\vec{x}||^2 + ||\vec{y}||^2 + ||\vec{z}||^2 - \lambda \right| \leq ||A^T||_*^2 < ||\vec{x}||^2 + ||\vec{y}||^2 + ||\vec{z}||^2,$$

from which it follows that $\lambda \in (0, 2(||\vec{x}||^2 + ||\vec{y}||^2 + ||\vec{z}||^2)]$. That is, the eigenvalues of M are positive.

Euler's equation (explained) The final piece of the puzzle is Euler's equation which, earlier, we asserted was just a fancy way of saying that torque changes angular momentum. When we first introduced the idea of torque we wrote

$$\tau = \frac{dL}{dt}. \tag{20}$$

Equation (20) is correct from the point of view of an observer who is removed from the application of torque and the resulting change in motion—physicists say that such a person is in an *inertial frame*. But we're not dealing with an inertial frame because our coordinate system, $\{p_1, p_2, p_3\}$, depends on M, which depends on the object which is rotating. As the object rotates, so does our basis!

How do we write (20) from our point of view, at the center of the rotating body, with a basis that's rotating? The key is to imagine what an observer in an inertial frame would see if, from our point of view in the rotating basis, we saw no change in the angular momentum. Because our frame of reference is spinning, our observation that L appears to be constant means that L is spinning about the axis of revolution at exactly the same speed as the basis. So an observer in an inertial frame would record the change in angular momentum as $\omega \times L$ (see Figure 3). Using the subscript of 0 to denote the inertial frame and the subscript r to denote the rotating frame, this thought experiement allows us to write (20) from our point of view in the rotating frame:

$$\tau = \left(\frac{dL}{dt}\right)_0 = \underbrace{\left(\frac{dL}{dt}\right)_r}_{\text{our basis}} + \underbrace{\omega \times L_r}_{\text{is spinning}}. \tag{21}$$

Finally, we use the fact that $L_r = M\omega$. Notice that M depends on the physical characteristics of the object but not on time, so we can rewrite (21) as

$$\tau = M\dot{\omega} + \omega \times M\omega, \tag{22}$$

which is Euler's equation.

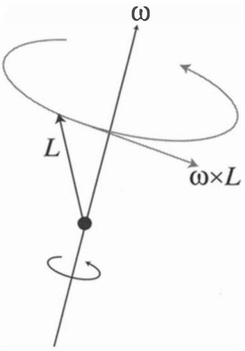


Figure 3 Change of L in an inertial frame

Conclusion

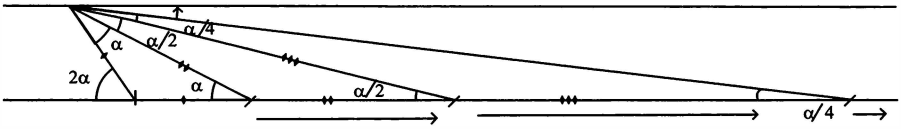
The last implicit supposition in our analysis was that the eigenvalues were distinct. This, at least, is not always true. What would happen if two of the eigenvalues were the same? What if all *three* were the same? What would that imply about the rotating object?

Acknowledgments. The author would like to thank Drs. Scott Franklin and George Thurston of RIT's Department of Physics for their time and conversation regarding this article.

REFERENCES

1. C.H. Edwards & D.E. Penney, *Elementary Differential Equations*, 5e, Prentice Hall, Upper Saddle River, NJ, 2004.
2. D.C. Lay, *Linear Algebra and Its Applications*, 3e, Addison Wesley, Boston, 2003.
3. D. Halliday, R. Resnick, J. Walker, *Fundamentals of Physics*, 6e, John Wiley & Sons, Inc., 2001.

Proof Without Words: Sum of a Geometric Series via Equal Base Angles in Isosceles Triangles



$$\alpha + \frac{\alpha}{2} + \frac{\alpha}{4} + \cdots = \sum_{n=0}^{\infty} \frac{\alpha}{2^n} = 2\alpha$$

—Ángel Plaza
ULPGC, 35017-Las Palmas G.C., Spain

The Volume Principle

WILLIAM C. DICKINSON
Grand Valley State University
Allendale, MI 49401
dickinsw@gvsu.edu

KRISTINA LUND
University of Nebraska–Lincoln
Lincoln, NE 68588-0130
s-klund1@math.unl.edu

Introduction

Beautiful theorems deserve elegant proofs. Among the attractive results in Euclidean geometry are the useful dual theorems of Menelaus and Ceva. Ceva's theorem states that three lines through the vertices of a triangle ABC, that intersect the sides of the triangle at P, Q, and R, are concurrent at a point O if and only if

$$\frac{|AR|}{|RB|}\frac{|BP|}{|PC|}\frac{|CQ|}{|QA|} = 1.$$

See Figure 1. This result is useful in proving the concurrence of many different sets of lines associated with a triangle. For example, the concurrence of the medians, angle bisectors and altitudes of Euclidean triangles can all easily be proved using Ceva's Theorem.

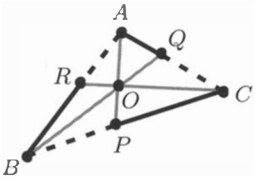


Figure 1 Ceva's Theorem in the Euclidean plane: The product of the dashed lengths equals the product of the solid lengths if and only if $\overleftrightarrow{AP}$, $\overleftrightarrow{BQ}$, and $\overleftrightarrow{CR}$ are concurrent at a point O.

Upon examining the proofs of these theorems as presented in several textbooks (for example [**14**, sect. 4.8] or [**2**, pp. 53–55]), we discovered that, for such beautiful results, the proofs are overly complicated, involved, and not very satisfying. This did not seem befitting for results of this type. Then we came across the work of Grünbaum and Shephard. In their article, *Ceva, Menelaus and the Area Principle* [**10**], they introduce a simple tool, called the Area Principle (which will be explained shortly), and use it to present deservingly elegant proofs of both Ceva's and Menelaus' Theorems. They go on to prove generalizations of these theorems to n-gons which are not necessarily convex or simple.

Grünbaum and Shephard's paper led us to begin thinking about how to generalize the Area Principle to spherical and hyperbolic geometry. Generalizing these theorems to include other geometries is not a new endeavor. In spherical and hyperbolic geometry, the analogs of the theorems of Ceva and Menelaus involve the product of ratios of the sines or hyperbolic sines of lengths. Both of these results have been known for over 100 years. On the sphere, Ceva's and Menelaus' Theorems are discussed and proved using fairly complicated trigonometry in several late 19th century undergraduate level textbooks. (See [**17**, p. 138] or [**15**, chap. IX].) In the hyperbolic plane, these results can be found in early 20th century texts, for example [**6**, p. 105]. More on these theorems and their generalizations can be found in [**4**], [**5**], [**3**], [**8**], and [**9**]. Once again these theorems are useful in proving the concurrence of the medians and other lines associated with hyperbolic and spherical triangles.

While a straightforward attempt to generalize the Area Principle to spherical and hyperbolic geometry fails, we have discovered a tool analogous to the Area Principle that works in spherical and hyperbolic geometry. In this paper, we prove the Volume Principle, which is valid for the Euclidean plane (embedded in $\mathbb{R}^3$), the hyperbolic plane (embedded in $\mathbb{M}^3$, Minkowski space) and the two-dimensional sphere (embedded in $\mathbb{R}^3$).

With appropriate modification, all the theorems from [**10**], except for those generalizations which require linearity, are thereby extended to these other geometries. They are proved using the Volume Principle in exactly the same manner as in Grünbaum and Shephard's article [**10**]. While some of these results may not be new, this technique gives an elegant and unifying proof method for a set of deserving theorems in these three geometries.

Revisiting the area principle

The main tool we want to generalize is the following.

THEOREM 1. (AREA PRINCIPLE) *Let A_1BC and A_2BC be two triangles in the Euclidean plane, where A_1 is distinct from A_2. If P is an intersection point of $\overleftrightarrow{A_1A_2}$ and $\overleftrightarrow{BC}$, then*

$$\left[\frac{A_1P}{A_2P}\right] = \left[\frac{A_1BC}{A_2BC}\right],$$

where $[\text{-}]$ represents the signed ratio of lengths or areas.

It does not matter if the triangles are on the same side of the line containing the common side or not, for if needed we can just extend the line segment $\overline{A_1A_2}$. See Figure 2. Clearly, the equality between the area and the length ratios will only be valid when the denominators are not zero and when $\overleftrightarrow{BC}$ and $\overleftrightarrow{A_1A_2}$ are not parallel. We will assume throughout what follows that the intersections of all necessary lines exist, so that all relevant ratios are defined.

The proof of the Area Principle can be seen in several elementary ways. One is to observe that the ratio of the areas of the triangles is the same as the ratio of their heights, and then to use a pair of similar triangles. A different method involves the law of sines. We choose this latter method of proof and quickly show the details because this method will generalize in other geometries.

In Figure 2, applying the law of sines to the triangles A_1BP and A_2BP and solving for the desired edge lengths, we have that

$$|A_1P| = \frac{|A_1B| \sin(\angle 1)}{\sin(\angle 2)} \quad \text{and} \quad |A_2P| = \frac{|A_2B| \sin(\angle 3)}{\sin(\angle 4)}. \tag{1}$$

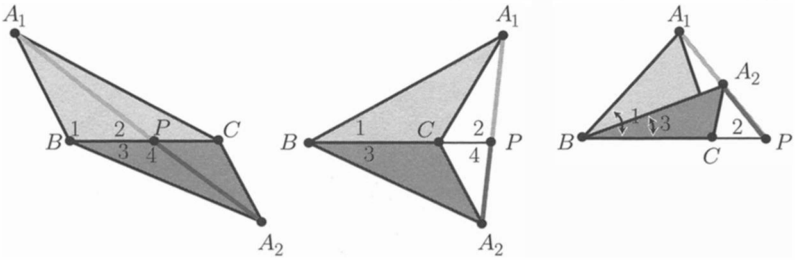


Figure 2 The Area Principle states that $[\frac{A_1P}{A_2P}] = [\frac{A_1BC}{A_2BC}]$, where [-] represents the signed ratio of lengths or areas

Since $\angle 2$ and $\angle 4$ are supplementary (or identical in the case that A_1 and A_2 are on the same side of $\overleftrightarrow{BC}$), it follows that $\sin(\angle 2) = \sin(\angle 4)$. It now follows from (1) that

$$\frac{|A_1P|}{|A_2P|} = \frac{|A_1B|\sin(\angle 1)}{|A_2B|\sin(\angle 3)}. \tag{2}$$

To obtain the areas of the relevant triangles in these ratios, we multiply top and bottom by $|BC|$ to obtain

$$\frac{|A_1P|}{|A_2P|} = \frac{|A_1B|\sin(\angle 1)|BC|}{|A_2B|\sin(\angle 3)|BC|} \tag{3}$$

$$= \frac{|A_1BC|}{|A_2BC|}. \tag{4}$$

As in **[10]**, we will adopt a convention for expressing the ratios of lengths of line segments and the ratio of areas of triangles in the setting of the Euclidean plane. (The same ideas work for the hyperbolic plane and the sphere.) Choosing an orientation on a line l, the sign of the length $\overline{PQ}$ is positive if the orientation of the line segment agrees with the orientation of l, and is negative otherwise. Next, the signed ratio of two line segments on l, denoted $[\frac{AB}{CD}]$, is the ratio of the signed length AB divided by the signed length CD. This ratio is independent of the orientation chosen on the line l. A similar convention is used for the ratio of the areas of triangles on an oriented surface. For two triangles, $[\frac{ABC}{DEF}]$ is the ratio of the signed areas of the triangles.

We will find it convenient to embed the Euclidean plane in $\mathbb{R}^3$ as the plane $z = 1$. In this setting, we choose the sign of the area of a triangle to be positive if the determinant of the matrix whose columns are the coordinate vectors of the vertices of a triangle listed in order is positive. This allows us to extend the Area Principle as is given in Theorem 1.

Notice that if we switch the order of the vertices in either line segment, we must modify the equation involving the ratio of areas and lengths with a minus sign. For example, in the setting of Figure 2, $[\frac{A_1P}{PA_2}] = -[\frac{A_1BC}{A_2BC}]$. If we permute the order of the vertices, the sign of the area ratio changes by the sign of the permutation.

Generalizing the area principle

We begin by choosing appropriate models for the geometries we are interested in, so that the common threads among them can easily be seen. A triangle ABC in any of these geometries is the union of three closed line segments, $\overline{AB}$, $\overline{BC}$ and $\overline{CA}$ where A, B and C are non-collinear. The notation we adopt for the sides and angles of a triangle

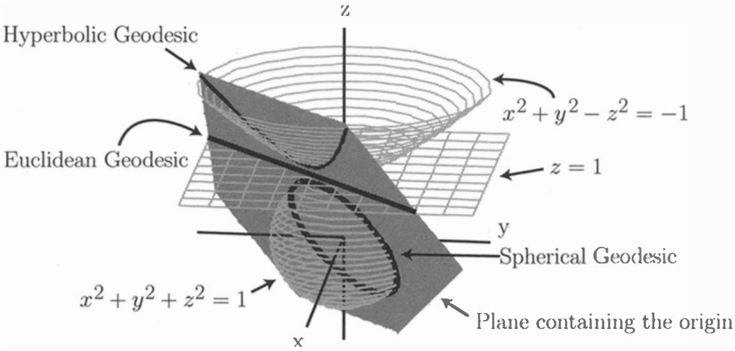


Figure 3 A plane through the origin intersecting the sphere, Euclidean plane, and hyperbolic plane

is the usual one. A triangle with vertices at A, B, and C has angles A, B, and C and the length of the sides opposite them are a, b, and c.

The two-dimensional sphere The model we choose for this surface is the set of points in $\mathbb{R}^3$ such that $x^2 + y^2 + z^2 = 1$. Straight lines on the sphere are *great circles* and are the intersections of the sphere with planes in $\mathbb{R}^3$ that pass through the origin, as shown in Figure 3. More formally, these paths are geodesics because they are the fixed point set of a reflection over the plane that defines them, which is an isometry of the sphere with itself. Unlike Euclidean geometry, there are pairs of points on the sphere that do not determine a unique great circle. Two such points are those formed by the intersection of the sphere with any line that passes through the origin; these are called *antipodal.* Further, a given pair of points does not necessarily determine a unique line segment because we can start at one of them and reach the other by heading in either direction on a great circle connecting them. To help eliminate some of these choices, we will require that all line segments have length less than or equal to π, half the length of a great circle. Hence, two non-antipodal points determine a unique line segment. Under these assumptions, the following statements can be proved:

- By using an isometry of the sphere, we can move any triangle into a standard position. The coordinates of the vertices in standard position are $A = \langle 0, 0, 1\rangle$, $B = \langle \sin(r), 0, \cos(r)\rangle$ and $C = \langle \sin(s)\cos(\theta), \sin(s)\sin(\theta), \cos(s)\rangle$ where $0 < r < \pi, 0 < s < \pi$, and $0 < \theta < \pi$. See Figure 4.
- For any spherical triangle ABC, there is a law of sines which states that

$$\frac{\sin(A)}{\sin(a)} = \frac{\sin(B)}{\sin(b)} = \frac{\sin(C)}{\sin(c)}.$$

For a complete introduction to spherical geometry and trigonometry, see **[17]** and **[15]**.

The Euclidean plane The model we choose for this surface is the plane $z = 1$ in $\mathbb{R}^3$. Analogous to straight lines on the sphere, we can see that the straight lines (geodesics) of the Euclidean plane are also the intersections of the plane $z = 1$ with planes that pass through the origin. See Figure 3. In this situation, the following statements hold:

- By using an isometry, we can move any triangle into a standard position. The coordinates of the vertices in standard position are $A = \langle 0, 0, 1\rangle$, $B = \langle r, 0, 1\rangle$ and $C = \langle s\cos(\theta), s\sin(\theta), 1\rangle$ where $0 < r, 0 < s$, and $0 < \theta < \pi$. See Figure 4.

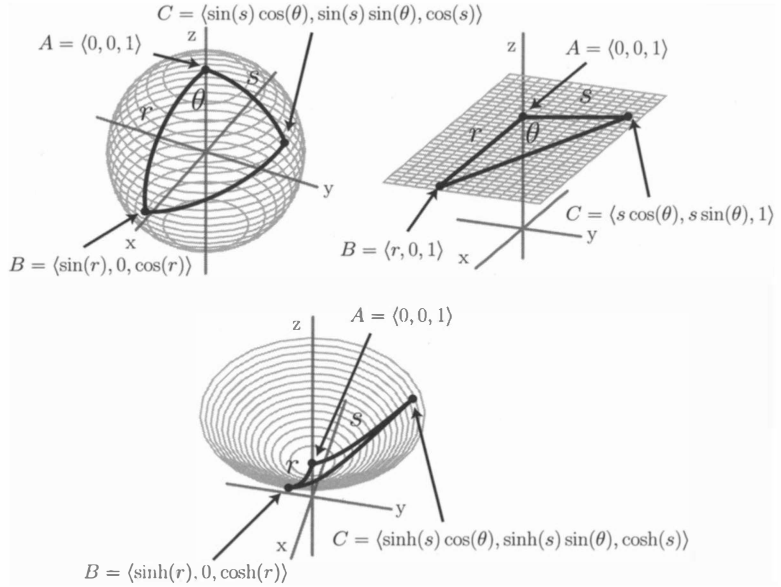

$C = \langle \sin(s)\cos(\theta), \sin(s)\sin(\theta), \cos(s)\rangle$
$A = \langle 0,0,1\rangle$
z
s
r
θ
y
x
$B = \langle \sin(r), 0, \cos(r)\rangle$
z
$A = \langle 0,0,1\rangle$
s
r
θ
$C = \langle s\cos(\theta), s\sin(\theta), 1\rangle$
y
$B = \langle r,0,1\rangle$
x
z
$A = \langle 0,0,1\rangle$
s
r
y
x
$C = \langle \sinh(s)\cos(\theta), \sinh(s)\sin(\theta), \cosh(s)\rangle$
$B = \langle \sinh(r), 0, \cosh(r)\rangle$

Figure 4 Triangles with side lengths r, s, and angle θ between them in standard position in spherical, Euclidean, and hyperbolic geometry

- For any Euclidean triangle ABC, there is a law of sines which states that

$$\frac{\sin(A)}{a} = \frac{\sin(B)}{b} = \frac{\sin(C)}{c}.$$

The hyperbolic plane The model that we choose for this surface is the upper sheet of the hyperboloid of two sheets, $x^2 + y^2 - z^2 = -1$ with $z > 0$, in Minkowski three space $\mathbb{M}^3$. As a set of points, Minkowski three space can be regarded exactly as $\mathbb{R}^3$. However, distances are not measured in the same way. For further details, see **[16]**.

A remarkable feature of this model is that the geodesics are once again the intersections of the hyperboloid and planes which pass through the origin, as seen in Figure 3. They are, just as in the spherical case, fixed point sets of hyperbolic reflections (isometries). In this situation, we can show that the following statements are true.

- By using an isometry (see **[16**, p. 448] for details) we can move any triangle into a standard position. The coordinates of the vertices in standard position are $A = \langle 0, 0, 1\rangle$, $B = \langle \sinh(r), 0, \cosh(r)\rangle$ and $C = \langle \sinh(s)\cos(\theta), \sinh(s)\sin(\theta), \cosh(s)\rangle$ where $0 < r$, $0 < s$, and $0 < \theta < \pi$. See Figure 4.
- For any hyperbolic triangle ABC, there is a law of sines which states that

$$\frac{\sin(A)}{\sinh(a)} = \frac{\sin(B)}{\sinh(b)} = \frac{\sin(C)}{\sinh(c)}.$$

This model is over 100 years old and is isometric to the more familiar Poincaré disk model, as well as the upper half-plane model of the hyperbolic plane. For more details about the hyperboloid model see **[16]**.

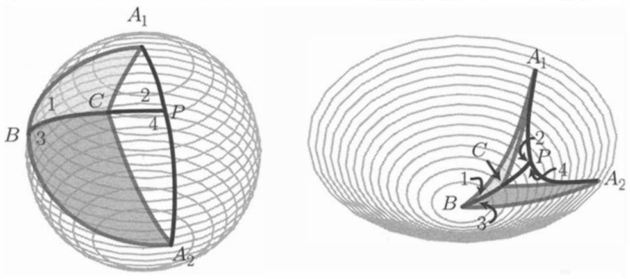


Figure 5 Is $\overline{A_1A_2}$ broken apart by $\overleftrightarrow{BC}$ in any special way?

Equations (1) and (2), we can still use the appropriate law of sines on triangles A_1BP and A_2BP to solve for the sines or hyperbolic sines of the necessary lengths and then divide the expressions to obtain

$$\frac{\sin(A_1P)}{\sin(A_2P)}=\frac{\sin(A_1B)\sin(\angle 1)}{\sin(A_2B)\sin(\angle 3)} \quad \text{or} \quad \frac{\sinh(A_1P)}{\sinh(A_2P)}=\frac{\sinh(A_1B)\sin(\angle 1)}{\sinh(A_2B)\sin(\angle 3)}. \tag{5}$$

Following the proof of the Area Principle and examining the pattern of sines and hyperbolic sines in the law of sines for each geometry, we should multiply by the sine or hyperbolic sine of BC and obtain

$$\frac{\sin(A_1P)}{\sin(A_2P)}=\frac{\sin(A_1B)\sin(\angle 1)\sin(BC)}{\sin(A_2B)\sin(\angle 3)\sin(BC)}$$

or (6)

$$\frac{\sinh(A_1P)}{\sinh(A_2P)}=\frac{\sinh(A_1B)\sin(\angle 1)\sinh(BC)}{\sinh(A_2B)\sin(\angle 3)\sinh(BC)}.$$

The crucial question becomes, what is the geometric interpretation of the quantity in the numerator and denominator of the right-hand sides of these equations? To help investigate this, consider the matrix, M_{ABC}, whose columns are the coordinate vectors of a triangle ABC. By using an isometry of spherical, hyperbolic, or Euclidean geometry, we may assume that the triangle is in standard position. Isometries in all of these geometries have determinant one, so the determinant of M_{ABC} doesn't depend on the location of the triangle in its model. Notice that

$$\det(M_{ABC})=\sin(r)\sin(\theta)\sin(s) \quad \text{or} \quad \det(M_{ABC})=\sinh(r)\sin(\theta)\sinh(s).$$

The volume interpretation of the determinant implies that the right-hand side of the equations in (6) are the ratios of the *Euclidean volumes* of two tetrahedrons with the vertices $O=\langle 0,0,0\rangle$, A_i, B and C ($i=1,2$). (The connection between the volume and the product $\sin(r)\sin(\theta)\sin(s)$ in the spherical case is not new. See [**7**, p. 265].) Hence, we have that

$$\frac{\sin(A_1P)}{\sin(A_2P)}=\frac{\mathrm{Vol}(OA_1BC)}{\mathrm{Vol}(OA_2BC)} \quad \text{or} \quad \frac{\sinh(A_1P)}{\sinh(A_2P)}=\frac{\mathrm{Vol}(OA_1BC)}{\mathrm{Vol}(OA_2BC)}. \tag{7}$$

Observe that in our model of Euclidean geometry ($\{(x,y,1)|x\in\mathbb{R},y\in\mathbb{R}\}\subset\mathbb{R}^3$), the volume of the tetrahedron OA_iBC is numerically one-third the area of triangle A_iBC. Thus, the Area Principle can be stated in a manner that is analogous to the expressions in (7). These three principles are illustrated in Figure 6.

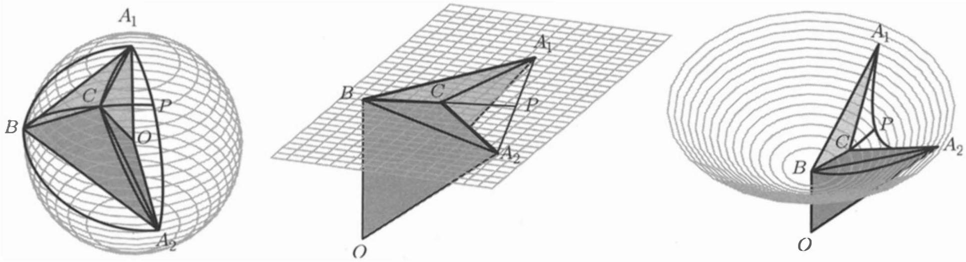


Figure 6 The Volume Principle in spherical, Euclidean and hyperbolic geometry states that

$$\left[\frac{\text{gsin}(A_1P)}{\text{gsin}(A_2P)}\right] = \left[\frac{OA_1BC}{OA_2BC}\right]$$

These results, in all three geometries, are begging to be unified. One way to do so is to use the generalized sine function, gsin(x), which is defined by the power series

$$\text{gsin}(x) = x - \frac{Kx^3}{3!} + \frac{K^2x^5}{5!} - \frac{K^3x^7}{7!} + \cdots,$$

where K is a parameter which we will regard as the constant Gaussian curvature of a complete simply connected two-dimensional space. For our purposes, we are mainly concerned with the sphere (where $K = 1$ and $\text{gsin}(x) = \sin(x)$), the Euclidean plane (where $K = 0$ and $\text{gsin}(x) = x$) and the hyperbolic plane (where $K = -1$ and $\text{gsin}(x) = \sinh(x)$), but these results have obvious generalizations to the other constant curvature spaces corresponding to all the other values of K.

Care must be taken when making the arguments outlined in Equations (5)–(7) in the spherical case for two reasons. Unlike in the Euclidean and hyperbolic cases, if A_1 and A_2 are on the same side of $\overleftrightarrow{BC}$, the point P is not well defined. However, after examining either choice of intersection, and using the fact that the sines of supplementary angles are equal, it is not too difficult to see that this choice doesn't matter. In further contrast to the other geometries, the points A_1 and A_2 may be antipodal. In this case, any choice of the line segment connecting them will work. Just observe that $A_1P + PA_2 = \pi$ and that volumes of the tetrahedrons are equal.

After introducing the signed length and volume in exactly the same way as in the Area Principle, we have the following.

THEOREM 2. (VOLUME PRINCIPLE) *Let A_1BC and A_2BC be two triangles in an appropriate model of a complete simply connected space of constant Gaussian curvature, K, where A_1 is distinct from A_2. If P is an intersection point of $\overleftrightarrow{A_1A_2}$ and $\overleftrightarrow{BC}$, then*

$$\left[\frac{\text{gsin}(A_1P)}{\text{gsin}(A_2P)}\right] = \left[\frac{OA_1BC}{OA_2BC}\right].$$

In the case of $K > 0$, either intersection point of $\overleftrightarrow{A_1A_2} \cap \overleftrightarrow{BC}$ may be chosen and if A_1 and A_2 are antipodal then any line segment that connects them may be chosen.

Ceva and Menelaus type theorems in spherical and hyperbolic geometries

The Volume Principle can be used in exactly the same way that Grünbaum and Shephard used the Area Principle to prove their results. Hence, we expect that most of their results, with appropriate modification, are true in spherical and hyperbolic geometry. Following Grünbaum and Shephard's work in **[10]**, by a *polygon* $P = [V_1, \dots, V_n]$ we mean a cyclic sequence of $n \geq 3$ points V_i (called *vertices*) in a space, together with the closed line segments $\overline{V_i V_{i+1}}$ (called *edges*) with length $V_i V_{i+1}$. Each edge is contained in a line, $\overleftrightarrow{V_i V_{i+1}}$, which we will refer to as a *side* of the polygon. We assume that a polygon is oriented, adjacent vertices are distinct (and in the spherical case, not antipodal) and that V_i, V_{i+1} and V_{i+2} are not collinear for all i. In what follows, all subscripts will be assumed to be reduced modulo n so that $1 \leq i \leq n$.

THEOREM 3. (CEVA'S THEOREM FOR n-GONS) *Let $P = [V_1, \dots, V_n]$ be an arbitrary n-gon in a complete simply connected space of constant Gaussian curvature, C a given point (not on any side of P), and k a positive integer such that $1 \leq k \leq \frac{n}{2}$. For $i = 1, \dots, n$, let W_i be an intersection point of the line $\overleftrightarrow{CV_i}$ and a line $\overleftrightarrow{V_{i-k} V_{i+k}}$. Then,*

$$\prod_{i=1}^{n} \left[\frac{\text{gsin}(V_{i-k} W_i)}{\text{gsin}(W_i V_{i+k})} \right] = 1. \tag{8}$$

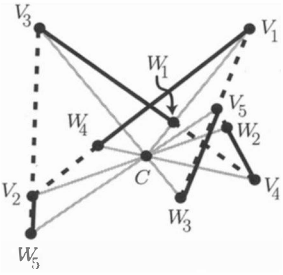


Figure 7 Ceva's Theorem for a 5-gon in the Euclidean plane with $k = 2$ states that

$$\left[\frac{V_4 W_1}{W_1 V_3}\right]\left[\frac{V_5 W_2}{W_2 V_4}\right]\left[\frac{V_1 W_3}{W_3 V_5}\right]\left[\frac{V_2 W_4}{W_4 V_1}\right]\left[\frac{V_3 W_5}{W_5 V_2}\right] = 1.$$

That is, the product of the dashed lengths equals the product of the solid lengths.

The case of $n = 3$ and $k = 1$ in the Euclidean plane is the well-known Ceva's Theorem. In the Euclidean setting, this result appears in **[10]** and elsewhere. The proof of this theorem is almost identical to Grünbaum and Shephard's proof in the Euclidean plane.

Proof. Observe that applying the Volume Principle to the triangles with base $\overline{CV_i}$ we obtain, for $i = 1, \dots n$,

$$\left[\frac{\text{gsin}(V_{i-k} W_i)}{\text{gsin}(W_i V_{i+k})} \right] = \left[\frac{OCV_i V_{i-k}}{OCV_{i+k} V_i} \right].$$

Substituting these terms in the left hand side of (8), we obtain a product of n terms each of which is a quotient of the volumes of certain tetrahedrons. These cancel to yield the value 1 as required. ■

In a similar vein, Menelaus' Theorem for n-gons becomes the following.

THEOREM 4. (MENELAUS' THEOREMS FOR n-GONS) *Let $P = [V_1, \dots , V_n]$ be an arbitrary n-gon in a complete simply connected space of constant Gaussian curvature, K, and suppose that, for $i = 1, \dots , n$, a line, l, cuts the side $\overleftrightarrow{V_i V_{i+1}}$ at W_i and does not pass through any vertex. Then*

$$\prod_{i=1}^{n} \left[\frac{\text{gsin}(V_i W_i)}{\text{gsin}(W_i V_{i+1})}\right] = (-1)^n. \tag{9}$$

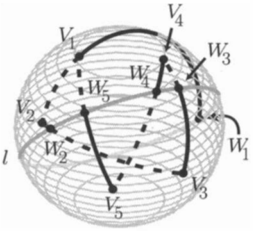


Figure 8 Menelaus' Theorem for a 5-gon on the sphere states that

$$\left[\frac{\sin(V_1 W_1)}{\sin(W_1 V_2)}\right]\left[\frac{\sin(V_2 W_2)}{\sin(W_2 V_3)}\right]\left[\frac{\sin(V_3 W_3)}{\sin(W_3 V_4)}\right]\left[\frac{\sin(V_4 W_4)}{\sin(W_4 V_5)}\right]\left[\frac{\sin(V_5 W_5)}{\sin(W_5 V_1)}\right] = -1.$$

That is, the product of the sines of the dashed lengths equals the opposite of the product of the sines of the solid lengths.

Proof. Observe that applying the Volume Principle to the triangles with base $\overline{W_1 W_2}$ we obtain, for $i = 1, \dots n$,

$$\left[\frac{\text{gsin}(V_i W_i)}{\text{gsin}(W_i V_{i+1})}\right] = -\left[\frac{O V_i W_1 W_2}{O V_{i+1} W_1 W_2}\right].$$

The minus sign comes from the fact that we have switched the order of the vertices in the denominator on the left hand side of the usual Volume Principle. Substituting these terms in the left hand side of (9), we obtain a product of n terms, each of which is a quotient of the volumes of certain tetrahedrons. These cancel to yield the value required. ■

Grünbaum and Shephard's result on the intersections of a polygon with its diagonals (i.e. line segments connecting distinct non-adjacent vertices), which they call selftransversality, becomes the following.

THEOREM 5. (SELFTRANSVERSALITY) *Let j, r and s be integers distinct* (mod n) *and let W_i be an intersection point of a line connecting V_i and V_{i+j} of the polygon $P = [V_1, \dots , V_n]$ with a line connecting V_{i+r} and V_{i+s}. Then a necessary and sufficient condition for*

$$\prod_{i=1}^{n} \left[\frac{\text{gsin}(V_i W_i)}{\text{gsin}(W_i V_{i+j})}\right] = (-1)^n \tag{10}$$

is that either one of the following is true:

Case 1 *$n = 2m$ is even, $j \equiv m \mod n$ and $s \equiv r + m \mod n$*

Case 2 *n is arbitrary and either one of the following is true:*

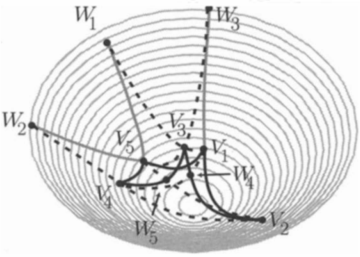


Figure 9 The selftransversality theorem for a 5-gon in the hyperbolic plane with $j = 2$, $r = 3$ and $s = 4$ states that

$$\left[\frac{\sinh(V_1W_1)}{\sinh(W_1V_3)}\right]\left[\frac{\sinh(V_2W_2)}{\sinh(W_2V_4)}\right]\left[\frac{\sinh(V_3W_3)}{\sinh(W_3V_5)}\right]\left[\frac{\sinh(V_4W_4)}{\sinh(W_4V_1)}\right]\left[\frac{\sinh(V_5W_5)}{\sinh(W_5V_2)}\right] = -1.$$

The polygon is shown in solid lines and the dashed lines are the lines in which all the lengths in the product occur.

Sub-case a $s \equiv 2r \mod n$ *and* $j \equiv 3r \mod n$
Sub-case b $r \equiv 2s \mod n$ *and* $j \equiv 3s \mod n$

Proof. Using the Volume Principle for triangles with base $\overline{V_{i+r}V_{i+s}}$ and apices V_i and V_{i+j}, we obtain

$$\left[\frac{\operatorname{gsin}(V_iW_i)}{\operatorname{gsin}(W_iV_{i+j})}\right] = -\left[\frac{OV_iV_{i+r}V_{i+s}}{OV_{i+j}V_{i+r}V_{i+s}}\right].$$

Substituting these terms in the left hand side of (10), we obtain a product of n terms each of which is a quotient of the volumes of certain tetrahedrons. Then we have to determine when the terms will cancel. The analysis is exactly the same as in **[10]** so we will not repeat it here. ■

Hoehn-type theorems in spherical and hyperbolic geometries

The other theorem that Grünbaum and Shephard generalize to n-gons is Hoehn's theorem. Hoehn originally proved his theorem about the product of ratios of side lengths in a convex Euclidean pentagram (see **[13]**) using several applications of Menelaus' theorem. In their proof of the generalization of this result, Grünbaum and Shephard use the Area Principle and the linearity of the gsin function for $K = 0$. Clearly the generalized sine function is not linear for any other values of K and their argument does not directly generalize for use with the Volume Principle. However, as we have proved Menelaus' Theorem in these other geometries, Hoehn's original proof still works. Thus we can prove that Hoehn's theorem, where the length of line segments are replaced with the sine or hyperbolic sine of the line segments, holds for pentagrams in spherical and hyperbolic geometry. See Figure 10.

However, we were unable to find a way around the use of linearity to prove the generalizations of Hoehn's theorems stated in **[10]** using the Volume Principle. Using the program *Spherical Easel* **[1]**, we were able to numerically verify many cases of the generalizations of Hoehn's theorem on the sphere. This evidence supports our conjecture that the generalizations of Hoehn's theorems are true in both spherical and hyperbolic geometry. In addition, we continue to investigate the obvious generalization of the Volume Principle to higher dimensions in an attempt to move all of the theorems from **[11]** and **[12]** into constant curvature spaces of higher dimension.

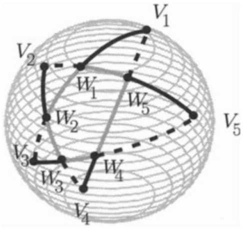


Figure 10 Hoehn's Theorem on the sphere states that

$$\left[\frac{\sin(V_1W_1)}{\sin(W_1V_2)}\right]\left[\frac{\sin(V_2W_2)}{\sin(W_2V_3)}\right]\left[\frac{\sin(V_3W_3)}{\sin(W_3V_4)}\right]\left[\frac{\sin(V_4W_4)}{\sin(W_4V_5)}\right]\left[\frac{\sin(V_5W_5)}{\sin(W_5V_1)}\right]=1.$$

That is, the product of the sines of the solid lengths is the same as the product of the sines of the dashed lengths. Also it states that

$$\left[\frac{\sin(V_1W_2)}{\sin(W_1V_3)}\right]\left[\frac{\sin(V_2W_3)}{\sin(W_2V_4)}\right]\left[\frac{\sin(V_3W_4)}{\sin(W_3V_5)}\right]\left[\frac{\sin(V_4W_5)}{\sin(W_4V_1)}\right]\left[\frac{\sin(V_5W_1)}{\sin(W_5V_2)}\right]=1.$$

Acknowledgments. This work was supported by a Student Summer Scholars grant from Grand Valley State University during the summer of 2003. The authors wish also to thank colleagues Matthew Boelkins and David Austin for their careful reading of and thoughtful suggestions for improving this manuscript.

REFERENCES

1. D. Austin and W. Dickinson, "Spherical Easel," a free spherical drawing program, Grand Valley State University, `http://merganser.math.gvsu.edu/easel/`, 2005.
2. A. Baragar, *A Survey of Classical and Modern Geometries with Computer Activities*, Prentice Hall, Upper Saddle River, New Jersey, 2001.
3. P. Boldescu, "The theorems of Menelaus and Ceva in an n-dimensional affine space," *An. Univ. Craiova Ser. a IV-a* **1** (1970), 101–106. MR MR0333932 (48 #12251)
4. B. Budinský, "Sätze von Menelaos und Ceva für Vielecke im sphärischen n-dimensionalen Raum," *Časopis Pěst. Mat.* **97** (1972), 78–85, 95. MR MR0307021 (46 #6142)
5. B. Budinský and Z. Nádeník, "Mehrdimensionales Analogon zu den Sätzen von Menelaos und Ceva," *Časopis Pěst. Mat.* **97** (1972) 75–77, 95. MR MR0307020 (46 #6141)
6. J. L. Coolidge, *The elements of non-Euclidean geometry*, Clarendon Press, Oxford, 1909, `http://mathbooks.library.cornell.edu:8085/Dienst/UIMATH/1.0/Display/cul.math/04800001`.
7. W. Gellert, H. Kästner, H. Küstner, and M. Hellwich (eds.), *The VNR Concise Encyclopedia of Mathematics*, Van Nostrand Reinhold Co., New York, 1977. With an introduction by Hans Reichardt. MR MR644488 (83c:00011)
8. P. A. Gluškov, "Analogues of the theorems of Menelaus and Ceva in Lobačevskiĭ space," *Barnaul. Gos. Ped. Inst. Učen. Zap.* **9** (1968), 77–81. MR MR0300195 (45 #9242b)
9. ——— "Trigonometric forms of the theory of transversals of Lobačevskiĭ space," *Barnaul. Gos. Ped. Inst. Učen. Zap.* **9** (1968), 73–76. MR MR0300194 (45 #9242a)
10. B. Grünbaum and G. C. Shephard, "Ceva, Menelaus, and the Area Principle," this MAGAZINE **68** (1965), no. 4, 254–268. MR MR1363708 (97f:51028)
11. ———, "Ceva, Menelaus, and Selftransversality," *Geom. Dedicata* **65** (1997), no. 2, 179–192. MR MR1451972 (98d:51012)
12. ———, "Some New Transversality Properties," *Geom. Dedicata* **71** (1998), no. 2, 179–208. MR MR1629791 (99i:51017)
13. L. Hoehn, "A Menelaus-Type Theorem for the Pentagram," this MAGAZINE **66** (1993), no. 2, 121–123.
14. D. Kay, *College Geometry*, second ed., Addison Wesley, New York, 2001.
15. W. J. McClelland and T. Preston, *Spherical Trigonometry with Applications to Spherical Geometry*, fifth ed., vol. I and II, Macmillan, London, 1897.
16. W. F. Reynolds, "Hyperbolic geometry on a hyperboloid," *Amer. Math. Monthly* **100** (1993), no. 5, 442–455. MR MR1215530 (94f:51037)
17. I. Todhunter, *Spherical Trigonometry, for the Use of Colleges and Schools, with Numerous Examples*, fifth ed., Macmillan, London, 1886, `http://mathbooks.library.cornell.edu:8085/Dienst/UIMATH/1.0/Display/cul.math/00640001`.

How Much Does a Matrix of Rank k Weigh?

THERESA MIGLER
California Polytechnic State University
San Luis Obispo, CA 93407
tmigler@calpoly.edu

KENT E. MORRISON
California Polytechnic State University
San Luis Obispo, CA 93407
kmorriso@calpoly.edu

MITCHELL OGLE
California Polytechnic State University
San Luis Obispo, CA 93407
mogle@calpoly.edu

Matrices with very few nonzero entries cannot have large rank. On the other hand matrices without any zero entries can have rank as low as 1. These simple observations lead us to our main question. For matrices over finite fields, what is the relationship between the rank of a matrix and the number of nonzero entries in the matrix? This question motivated a summer research project collaboration among the authors (two undergraduate students and their adviser), and although the question seems natural, we were unable to find any previously published work dealing with it.

We call the number of nonzero entries of a matrix A the *weight* of A and denote it by wt A. For matrices over finite fields, the weight of $A - B$ is a natural way to define the distance between A and B. In coding theory the distance between vectors defined in this way is called the *Hamming distance*, named after Richard Hamming, a pioneer in the field of error correcting codes. The rank of $A - B$, denoted rk $(A - B)$, defines a different distance between the matrices A and B. Thus wt A and rk A give two ways to measure the distance from A to the origin and our fundamental question is about the relationship between them.

The background needed for this paper comes from the undergraduate courses in linear algebra, abstract algebra, and probability. We use the fundamental ideas of linear algebra over finite fields. For each prime power q we let $\mathbf{F}_q$ denote the unique field with q elements. There is no problem, however, in reading the rest of the paper with only the prime fields $\mathbf{F}_p$ (or even $\mathbf{F}_2$) in mind. We use the basic concepts and results of probability up through the central limit theorem.

Having restricted our investigation to matrices over finite fields, we restate the fundamental question in this way: Over $\mathbf{F}_q$ how many $m \times n$ matrices of rank k and weight w are there? In probabilistic terms, we are asking for the distribution of the weight for matrices of rank k. We do not have the complete answer to this question, and it seems we are far from the complete answer, so there is plenty of work left to be done. The main results we offer are the average value of the weight for matrices of fixed rank and the complete description of the weight distribution for rank 1 matrices.

Counting matrices

For matrices of fixed size, the zero matrix is the only matrix of weight 0 and the only matrix of rank 0. Also, any matrix of weight 1 has rank 1, and a matrix of rank k has weight at least k. It is possible to count the matrices of very small rank and weight. For a matrix of rank 1 and weight 1 we choose one of the mn entries in which to place one of the $q-1$ nonzero elements of the field. That gives us $mn(q-1)$ matrices of rank 1 and weight 1. We leave two more results as exercises for the reader.

- The number of rank 1 and weight 2 is
$$\frac{1}{2}mn(m+n-2)(q-1)^2.$$
- The number of rank 2 and weight 2 is
$$\frac{1}{2}mn(m-1)(n-1)(q-1)^2.$$

The number of matrices of weight w (regardless of rank) is easy to count. There are w locations to select and in each location there are $q-1$ nonzero elements to choose. Therefore, the number of $m \times n$ matrices of weight w is

$$\binom{mn}{w}(q-1)^w.$$

The probability that a matrix has weight w, assuming that each matrix is equally probable, is then

$$\frac{1}{q^{mn}}\binom{mn}{w}(q-1)^w = \binom{mn}{w}(1-1/q)^w(1/q)^{mn-w},$$

showing us that the weight follows is a binomial distribution with parameters mn and $1-1/q$.

Next we count the number of $m \times n$ matrices of rank k without regard to weight. Although this is a more difficult problem than counting according to weight, it is an old result with the first derivation of the formula due to Landsberg in 1893 [**5**]. We break the problem into two parts by first counting the matrices with a fixed column space of dimension k and then counting the subspaces of dimension k. The product of these two numbers is the number of $m \times n$ matrices of rank k.

Let V be a fixed k-dimensional subspace of the m-dimensional space $\mathbf{F}_q^m$. The $m \times n$ matrices whose column space is V are bijective with the linear transformations from $\mathbf{F}_q^n$ onto V, which are bijective with the $k \times n$ matrices of rank k.

FORMULA 1. *The number of $k \times n$ matrices of rank k is*

$$\prod_{0 \le i \le k-1} (q^n - q^i) = (q^n-1)(q^n-q)\cdots(q^n-q^{k-1}).$$

Proof. In order for a $k \times n$ matrix to have rank k the rows must be linearly independent. (For a slick proof that row rank equals column rank see the recent note by Wardlaw in this magazine [**6**].) Now the first row can be any nonzero vector with n entries, and there are $q^n - 1$ such vectors. The second row must be independent of the first row. That means it cannot be any of the q scalar multiples of that row, but any other row vector is allowed. There are $q^n - q$ vectors to choose from. The third row can be any vector not in the span of the first two rows. There are q^2 linear combinations of the first two rows, and so there are $q^n - q^2$ possible vectors for row 3. We continue in this way with row $i+1$ not allowed to be any of the q^i linear combinations of the first i rows. ■

FORMULA 2. *The number of k-dimensional subspaces of an m-dimensional vector space over* $\mathbf{F}_q$ *is*

$$\frac{\prod_{0\le i\le k-1}(q^m-q^i)}{\prod_{0\le i\le k-1}(q^k-q^i)}.$$

Proof. To count the number of k-dimensional subspaces of a vector space of dimension m, we count the number of bases of all such subspaces and then divide by the number of bases that each subspace has. A basis is an ordered list of k linearly independent vectors lying in $\mathbf{F}_q^m$. Putting them into a matrix as the rows, we get a matrix of rank k. From Formula 1 (with n replaced by m) we see that there are $\prod_{0\le i\le k-1}(q^m-q^i)$ bases. The number of bases of a k-dimensional space is just the number of $k\times k$ matrices of rank k. Again, we use Formula 1 (with n replaced by k) to see that there are $\prod_{0\le i\le k-1}(q^k-q^i)$ bases of a particular subspace. ■

There is a well-developed analogy in the world of combinatorics between the subsets of a finite set and the subspaces of a finite dimensional vector space over a finite field. The number of k-dimensional subspaces of an m-dimensional vector space is analogous to the number of subsets of size k in a set of size m, which is given by the binomial coefficient $\binom{m}{k}$. So, we let

$$\binom{m}{k}_q := \frac{\prod_{0\le i\le k-1}(q^m-q^i)}{\prod_{0\le i\le k-1}(q^k-q^i)}$$

denote the number of such subspaces, as given in Formula 2. This number is often called a *Gaussian binomial coefficient.* Although the full development of the subset-subspace analogy is not necessary for us, we recommend Kung's introductory survey **[4]** and Cohn's recent discussion of the Gaussian binomial coefficients **[2]**.

Formulas 1 and 2 give the two factors we need for the number of $m\times n$ matrices of rank k. As one should expect the formula is symmetric in m and n.

FORMULA 3. *The number of* $m\times n$ *matrices of rank k is*

$$\frac{\prod_{0\le i\le k-1}(q^n-q^i)\prod_{0\le i\le k-1}(q^m-q^i)}{\prod_{0\le i\le k-1}(q^k-q^i)}.$$

The average weight of rank k matrices

As we have mentioned, the weight of a matrix A is the Hamming distance between A and the zero matrix, and so we expect that in some way increasing weight is correlated with increasing rank. In this section, we determine the average weight of the set of matrices of a fixed rank in terms of the parameters q, m, n, and k. Indeed we find that the average weight grows with k when the other parameters are held fixed.

We consider the weight as a random variable W, which is the sum $\sum_{i,j}W_{ij}$, where W_{ij} is the weight of the i, j entry, meaning that $W_{ij}=1$ for a matrix whose i, j entry is nonzero and $W_{ij}=0$ when the entry is 0. Then the average or expected value of W is the sum of the expected values of the random variables W_{ij}. The expected value of W_{ij} is simply the probability that the i, j entry is nonzero. It is this probability that we will compute. An important observation is that this probability is the same for all i and j. In other words, the W_{ij} are identically distributed.

THEOREM 1. *For* $m\times n$ *matrices of rank k, the probability that the i, j entry is nonzero is the same for all i and j.*

Proof. For a fixed row index i and column index j there is a bijection on the space of $m \times n$ matrices defined by switching row 1 with row i and switching column 1 with column j. This bijection preserves the rank and weight, and so it defines a bijective correspondence between the subset of matrices of rank k with nonzero 1,1 entry and the subset of matrices of rank k with nonzero i, j entry. ■

With this result we know that the expected value of W is mn times the average weight of the 1,1 entry, so that we can focus our attention on the upper left entry. Our analysis depends on what is called the *reduced row echelon form.* Recall that a matrix is in reduced row echelon form if all the nonzero rows are above all the zero rows, if the leftmost nonzero entry of a nonzero row is 1, and if such an entry is the only nonzero one in its column.

The reduced row echelon forms are actually a system of representatives of the *row equivalence classes*, where two matrices are row equivalent if a sequence of elementary row operations changes one into the other. It is also the case that A and B are row equivalent if and only if they have identical row spaces.

For an $m \times n$ matrix A of rank k, the reduced row echelon form of A has k nonzero rows. Let R be the $k \times n$ matrix consisting of those rows. Since the rows of R form a basis of the row space of A, each row of A is a linear combination of the rows of R. That means there is a unique $m \times k$ matrix C such that $A = CR$. Note that the rank of C must also be k.

Using the factorization $A = CR$ we can express the set of rank k matrices as the Cartesian product of the set of $m \times k$ matrices of rank k with the set of $k \times n$ reduced row echelon matrices of rank k. This means that A can be selected randomly by independently choosing the factors C and R. Now the 1,1 entry of A is given by $a_{11} = c_{11}r_{11} + c_{12}r_{21} + \cdots + c_{1k}r_{k1}$. Since R is in reduced row echelon form, r_{11} is 0 or 1 and the rest of the entries in the first column, $r_{21}, r_{31}, \ldots, r_{k1}$, are all 0. Therefore, $a_{11} = c_{11}r_{11}$, and so the probability that the 1,1 entry is nonzero is

$$\mathbf{P}(a_{11} \neq 0) = \mathbf{P}(c_{11} \neq 0)\mathbf{P}(r_{11} \neq 0).$$

Now C is $m \times k$ and has rank k. Thus, the first column of C is any nonzero vector of length m, of which there are $q^m - 1$. There are $q^{m-1} - 1$ of those vectors that have a zero in the top entry, and so there are $q^m - q^{m-1}$ that have a nonzero top entry. Then

$$\mathbf{P}(c_{11} \neq 0) = \frac{q^m - q^{m-1}}{q^m - 1}.$$

The choice of the reduced matrix R is the same as the choice of row space of A. If any of the vectors in the row space has a nonzero first entry, then the first column cannot be the zero column and then r_{11} is not 0. In order that $r_{11} = 0$ the row space of A must be entirely within the $n - 1$ dimensional subspace of vectors of the form $(0, x_2, x_3, \ldots, x_n)$. The probability of that occurring is the ratio of the number of k-dimensional subspaces of a space of dimension $n - 1$ to the number of k-dimensional subspaces of a space of dimension n:

$$\mathbf{P}(r_{11} = 0) = \frac{\binom{n-1}{k}_q}{\binom{n}{k}_q}.$$

Therefore, the complementary probability gives

$$\mathbf{P}(r_{11} \neq 0) = 1 - \frac{\binom{n-1}{k}_q}{\binom{n}{k}_q}.$$

Using Formula 2 we simplify this to get

$$\mathbf{P}(r_{11} \neq 0) = \frac{q^n - q^{n-k}}{q^n - 1}.$$

Putting these results together gives us

$$\mathbf{P}(a_{11} \neq 0) = \left(\frac{q^m - q^{m-1}}{q^m - 1}\right)\left(\frac{q^n - q^{n-k}}{q^n - 1}\right).$$

As a probability this is more easily analyzed in the following form:

$$\mathbf{P}(a_{11} \neq 0) = \frac{(1 - 1/q)(1 - 1/q^k)}{(1 - 1/q^m)(1 - 1/q^n)}.$$

When m, n, and k are large, this probability is close to $1 - 1/q$, which is the probability that an entry is nonzero with no condition on the rank. One case of interest is that of invertible matrices. For $n \times n$ invertible matrices we have $k = m = n$, and so the probability that an entry is nonzero simplifies to

$$\frac{1 - 1/q}{1 - 1/q^n}.$$

We see that it is slightly more likely that an invertible matrix has nonzero entries than an arbitrary matrix.

Having determined the probability that the 1, 1 entry is nonzero and hence that the probability that the i, j entry is nonzero, we have proved the following theorem.

THEOREM 2. *The average weight of an $m \times n$ matrix of rank k over the field of order q is*

$$mn\frac{(1 - 1/q)(1 - 1/q^k)}{(1 - 1/q^m)(1 - 1/q^n)}.$$

We also see that with m and n fixed the average weight increases as k increases. It is this formula that best expresses the intuitive idea that increasing rank is correlated with increasing weight. FIGURE 1 shows a plot of the average weight vs. the rank for matrices of size 10×10 over the field $\mathbf{F}_2$.

The weight of rank 1 matrices

We can analyze the weight distribution more completely for matrices of rank 1. From Theorem 2 with $k = 1$ we see that the average weight of a rank 1 matrix is

$$mn\frac{(1 - 1/q)^2}{(1 - 1/q^m)(1 - 1/q^n)}.$$

For m and n large this average is just about $mn(1 - 1/q)^2$, whereas the average weight for all $m \times n$ matrices is $mn(1 - 1/q)$, and so rank 1 matrices tend to have a lot more zero entries than the average matrix. For $q = 2$ this effect is the most pronounced. An average of one fourth of the entries are 1 in a large rank 1 matrix over $\mathbf{F}_2$, while an average of half the entries are 1 for all matrices.

In the factorization $A = CR$, where $\operatorname{rk} A = 1$, C is a nonzero column vector of length m and R is a nonzero row vector of length n whose leading nonzero entry is 1.

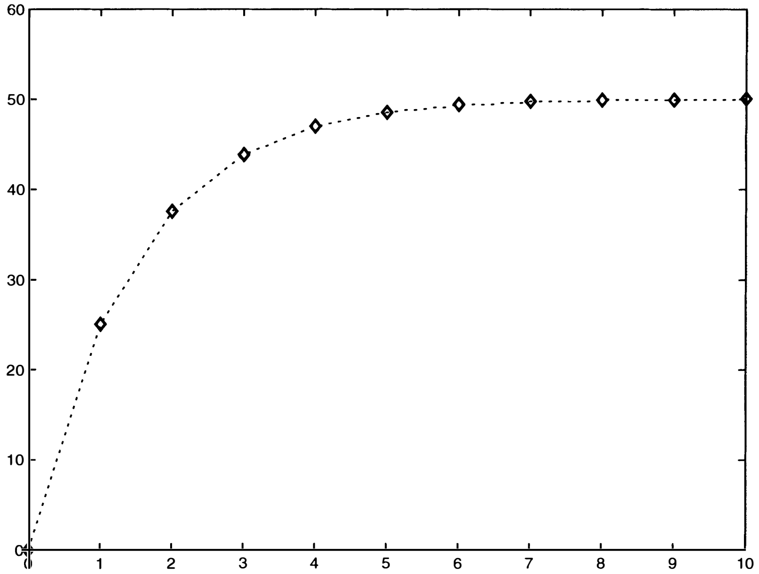


Figure 1 Average weight plotted against rank for 10 × 10 matrices over $\mathbf{F}_2$

The entries of A are given by $a_{ij} = c_i r_j$, and so the weight of A is the product of the weights of C and R. The weight of C has probability distribution given by

$$\mathbf{P}(\operatorname{wt} C = \mu) = \frac{\binom{m}{\mu}(q-1)^{\mu}}{(q^m - 1)},$$

because there are $q^m - 1$ nonzero column vectors of length m, and there are

$$\binom{m}{\mu}(q-1)^{\mu}$$

vectors of weight μ. This is the distribution of a binomial random variable (with parameters m and $1 - 1/q$) conditioned on being positive. Likewise for R the weight distribution is given by

$$\mathbf{P}(\operatorname{wt} R = \nu) = \frac{\binom{n}{\nu}(q-1)^{\nu}}{(q^n - 1)}.$$

To select a random R, choose a random nonzero vector of length n and then scale it to make the leading nonzero entry 1. The scaling does not change the weight.

Immediately we see that there is a restriction on the possible weight of a matrix of rank 1. For example, the weight of a 3 × 4 matrix of rank 1 cannot be 5, 7, 10, or 11 because those numbers are not products $\mu\nu$ with $1 \le \mu \le 3$ and $1 \le \nu \le 4$. All other weights between 1 and 12 are possible.

The weight of rank 1 matrices is the product of these two binomial random variables, each conditioned to be positive.

$$\begin{aligned}\mathbf{P}(\operatorname{wt} A = \omega) &= \sum_{\mu\nu=\omega} \mathbf{P}(\operatorname{wt} C = \mu)\mathbf{P}(\operatorname{wt} R = \nu) \\ &= \sum_{\mu\nu=\omega} \binom{m}{\mu}\binom{n}{\nu}\frac{(q-1)^{\mu+\nu}}{(q^m - 1)(q^n - 1)}\end{aligned}$$

Because not all weights between 1 and mn occur for rank 1 matrices, plots of actual probability densities show spikes and gaps. However, the plots of cumulative distributions are smoother and lead us to expect a limiting normal distribution as the size of the matrices goes to infinity. FIGURES 1 and 2 show this behavior quite well. (In order to plot the approximating normal distribution, we numerically computed the standard deviation of the weight distribution for the given m, n, and q.)

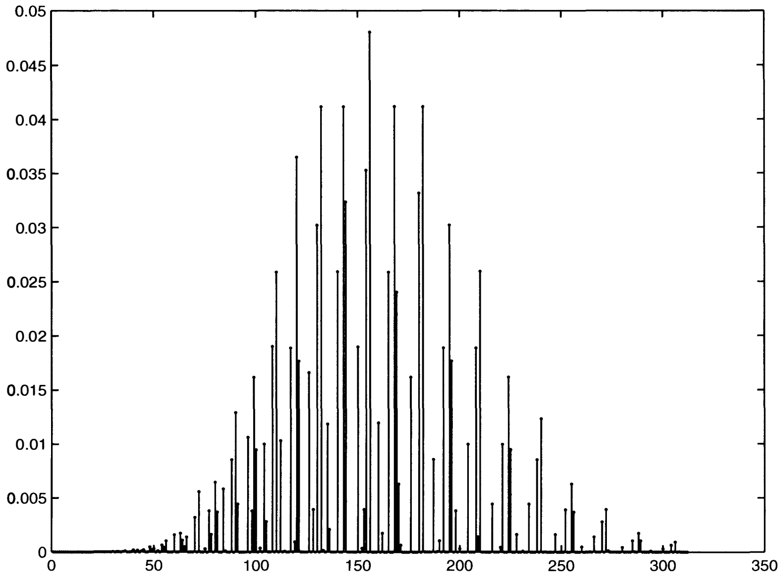


Figure 2 Density for the weight of rank 1 matrices, $m = n = 25$, $q = 2$

THEOREM 3. *As m or n goes to infinity, the weight distribution of rank 1 matrices approaches a normal distribution.*

Proof. The weight random variable for rank 1 matrices of size $m \times n$ is the product of independent binomial random variables conditioned on being positive. Define $W = XY$, where $X = \sum_{1\leq i\leq m} X_i$, $Y = \sum_{1\leq j\leq n} Y_j$, and X_i and Y_j are independent Bernoulli random variables with probability $1/q$ of being 0. Then W is the sum of m independent identically distributed random variables X_iY. Conditioning W on $W > 0$ is the weight of rank 1 matrices. By the central limit theorem the distribution of W converges, as $m \to \infty$, to a normal distribution after suitable scaling. Now conditioning on W being positive does not change this result because the probability that $W > 0$ is $1 - q^{-m}$, which goes to 1 as $m \to \infty$. ■

Therefore, when m and n are large we can use a normal distribution of mean $\mathbf{E}(W)$ and variance var (W) to approximate the weight distribution for rank 1 matrices. Note that this variance is not exactly the variance of the weight of rank 1 matrices because we have not conditioned on W being positive. However, the exact computation of that variance is rather complicated, and because of the theorem, the exact variance is not more informative than the variance of the unconditioned random variable W.

How fast does the variance grow as m and n go to infinity? For simplicity we let $m = n$, but the computation in the general case is similar. The variance of W is given by

$$\operatorname{var}(W) = \mathbf{E}(W^2) - \mathbf{E}(W)^2.$$

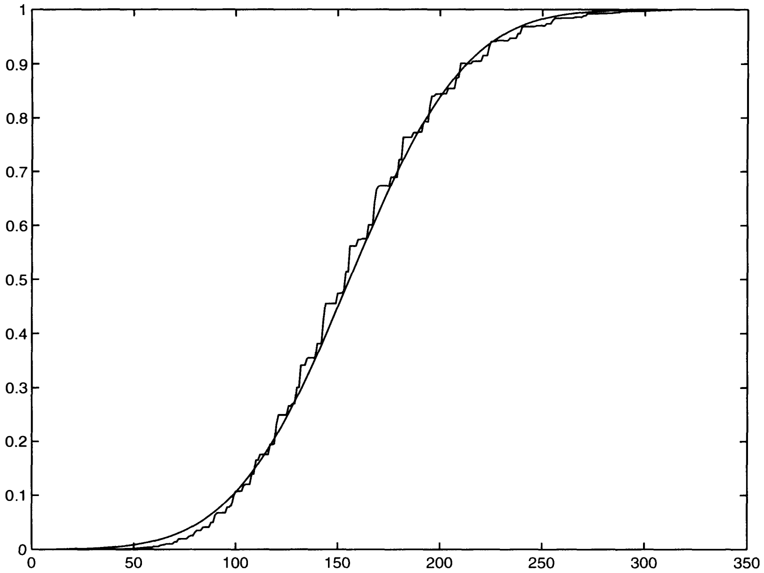


Figure 3 Cumulative frequency distribution for the weight of rank 1 matrices, $m = n = 25$, $q = 2$. The smooth curve is the normal distribution with the same mean ($\approx mn/4 = 156.25$) and standard deviation (≈ 44.63).

We need $\mathbf{E}(X)$ and $\mathbf{E}(X^2)$, which are given by

$$\begin{aligned}\mathbf{E}(X) &= n(1 - 1/q)\\ \mathbf{E}(X^2) &= n(1 - 1/q) + n(n-1)(1 - 1/q)^2.\end{aligned}$$

Because X and Y are independent binomial random variables with the same distribution, it follows that

$$\begin{aligned}\mathbf{E}(W) &= \mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(X)^2\\ \mathbf{E}(W^2) &= \mathbf{E}(X^2Y^2) = \mathbf{E}(X^2)\mathbf{E}(Y^2) = \mathbf{E}(X^2)^2.\end{aligned}$$

After taking care of the algebra we arrive at

$$\mathrm{var}\,(W) = \frac{1}{q^2}\left(1 - \frac{1}{q}\right)^2 n^2 + \frac{2}{q}\left(1 - \frac{1}{q}\right)^3 n^3.$$

From this we can see that for square matrices, the variance grows like n^3, and the standard deviation grows like $n^{3/2}$.

Over the field with two elements, the variance is

$$\frac{n^3}{8} + \frac{n^2}{16},$$

and so the standard deviation is asymptotic to $(n/2)^{3/2}$. In the example shown in FIGURES 1 and 2, the standard deviation of the actual weight distribution of rank 1 matrices is 44.63 (rounded to two places). The value of $(n/2)^{3/2}$ with $n = 25$ is 44.19 (also rounded to two places).

Further questions

Analyzing the CR factorization for rank 2 matrices should conceivably allow us to find the weight distribution for rank 2, but the analysis is considerably more difficult, and for higher ranks the difficulty continues to increase. This suggests gathering some information by simulation. In FIGURE 3, we show a histogram for the weights of 10,000 matrices of rank 2 and size 25 × 25 over $\mathbf{F}_2$. The average weight from Theorem 2 in this case is

$$\frac{25^2(1-1/2)(1-1/2)^2}{(1-1/2^{25})^2} = 234.3750\ldots$$

The sample mean is 234.4434, and the sample standard deviation is 39.5201. The histogram makes it plausible that the weight has a limiting normal distribution. In fact, for any k we expect a limiting normal distribution for the weight (with suitable scaling) of rank k matrices as the size goes to infinity.

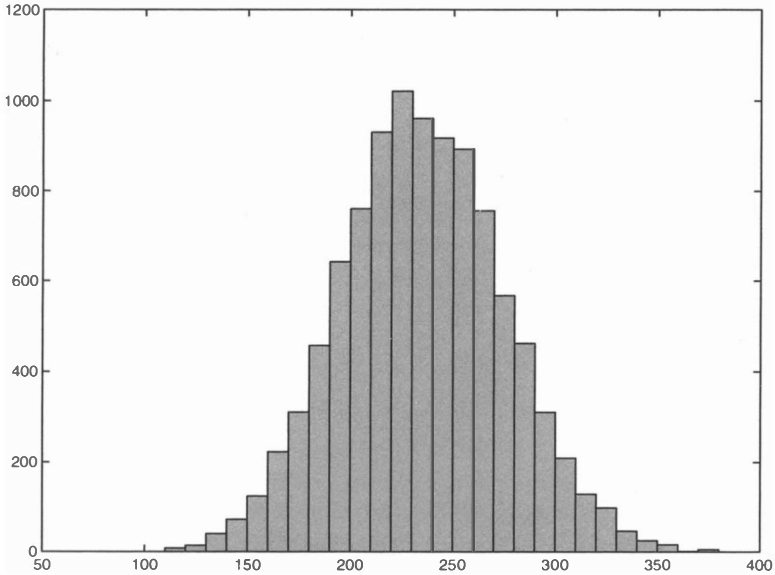


Figure 4 Histogram for the weight of 10,000 rank 2 matrices, $m = n = 25$, $q = 2$; bins have width 10

The question of simulation leads to the question of efficiently generating random matrices of fixed rank. Calabi and Wilf [1] treat the related problem of randomly generating a subspace of fixed dimension over a finite field, and Wilf has suggested to us that a random rank k matrix could be generated by adding together k matrices of rank 1, which are easy to generate, and then keeping those of rank k. Alternatively, one might use the CR factorization. Selecting R is exactly the subspace selection problem just mentioned. Selecting C can be done by generating a random $m \times k$ matrix and keeping those of rank k. Which approach is more efficient we leave as an open question, as well as the question of whether there are even better ways to generate matrices of a fixed rank.

We have focused on the weight of fixed rank matrices, but it would be interesting to look at the rank of fixed weight matrices. As an example, consider the $n \times n$ matrices of weight n. The number of these with rank 1 can be expressed in terms of the divisors of n, using our previous result. Those of rank n are generalizations of permutation matrices and there are $n!(q-1)^n$ of them. What about the other ranks? In particular, how many $n \times n$ matrices of weight n and rank $n-1$ are there over $\mathbf{F}_2$?

Since the weight of A is the Hamming distance from A to 0, it plays a role analogous to the norm of a real or complex matrix. In both cases it is the distance to the only matrix of rank 0. Now we may ask for the distance from A to the subset of matrices of rank 1, that is for the minimal distance from A to some matrix of rank 1. In general we may ask for the distance from A to the matrices of rank k. For real or complex matrices these distances (using the linear map norm) are given by the *singular values* and can easily be computed [**3**, p. 468]. For matrices over finite fields can these distances (defined by the weight) be computed in any other way than by exhaustive search?

REFERENCES

1. E. Calabi and H. S. Wilf, On the sequential and random selection of subspaces over a finite field, *J. Combinatorial Theory (A)*, **22** (1977), 107–109; MR **55** #5649
2. H. Cohn, Projective geometry over $\mathbf{F}_1$ and the Gaussian binomial coefficients, *Amer. Math. Monthly*, **111** (2004), 487–495.
3. D. C. Lay, *Linear Algebra and Its Applications*, second edition, Addison-Wesley, Reading, Massachusetts, 1997.
4. J. P. S. Kung, The subset-subspace analogy, in *Gian-Carlo Rota on combinatorics*, 277–283, Birkhäuser, Boston, Boston, MA, 1995; MR 99b:01027
5. R. Lidl and H. Niederreiter, *Finite Fields*, second edition, Cambridge Univ. Press, Cambridge, UK, 1997; MR 97i:11115
6. W. P. Wardlaw, Row rank equals column rank, this MAGAZINE, **78** (2005), 316–318.

Folding Optimal Polygons from Squares

DAVID DUREISSEIX
University Montpellier 2
Montpellier, F-34095, France
dureisse@lmgc.univ-montp2.fr

What is the largest regular n-gon that fits in a unit square? Can it be folded from a square piece of paper using standard moves from origami? Answering the first question is relatively easy, using simple ideas from geometry. The second is more interesting; our answer illustrates the difference between origami and the standard compass-and-straightedge constructions of the Greeks, where, for instance, the 7-gon cannot be constructed. Not only can we fold a 7-gon, but we can fold the largest one possible from a given square piece of paper.

Origami (from the Japanese *oru*, to fold, and *kami*, paper), is the ancient art of paperfolding. When we fold a paper in half, we create a line segment and bisect a length. These simple moves can be combined to reproduce *any* compass-and-straightedge construction [**1, 14**]. Thus, by origami, as with an unmarked straightedge and compass, we can construct roots of any second-order polynomial from a given unit length.

However, many constructions known to be impossible under the standard Greek rules, such as trisecting a given angle, become possible with origami. For instance, using a construction technique due to Lill and first used for origami by M. P. Beloch [**2**], we can construct roots of cubic polynomials by folding [**1, 5, 7, 12**]. Origami also simplifies certain constructions that are possible, but cumbersome, with compass and straightedege.

Since origami often begins with a square piece of paper, we propose not only to fold a regular n-gon, but to fold the one with the largest area that fits in the square. Such polygons will be called *optimal* polygons. For instance, the side of the largest equilateral triangle that fits in a unit square (shown in FIGURE 4) is known to have length $\sqrt{6} - \sqrt{2} \approx 1.035$. Wetzel [**18**] takes this as the starting point for his article "Fits and Covers," which gives many answers to similar problems, but does not address our question.

Our first step is to determine the proper orientations of optimal polygons with respect to the square. We do this in complete generality and then consider how to construct them by folding. We show how to fold the optimal hexagon and pentagon, which can also be constructed with compass and straightedge. Moving into the realm of techniques that break the Greek rules, we trisect an angle and show how to fold the optimal 7-gon and 9-gon, neither of which can be constructed with straightedge and compass alone. It turns out that in each case we fold a star polygon as an intermediate step.

Are you eager to fold the optimal 11-gon? If so, you will have to invent a folding technique that permits you to construct roots of a quintic polynomial!

Facts about optimal polygons

The goal is to find the largest regular n-gon that can be folded from a square piece of paper, for $n \geq 3$. Of course, the case $n = 4$ is trivial, with no folding required. For the general case, let us review some facts about the regular n-gon.

Let R be the radius of the circumscribed circle and r the radius of the inscribed circle. (Another name for r is the *apothem* of the polygon.) The reader may wish to confirm that $r = R\cos(\pi/n)$. The *diameter* of the n-gon, denoted L, is the maximum distance between any two of its points. A contrasting quantity is the *altitude*, meaning the shortest perpendicular distance from a vertex to an opposite side. When n is even, these quantities are simple, the diameter is just $2R$ and the altitude is $2r$. When n is odd, the altitude is rather easily seen to be $R + r$, which is $2R\cos^2(\pi/(2n))$. The diameter can then be found to be $2R\cos(\pi/(2n))$. (FIGURE 1 will help with this.)

Another useful quantity is l, the side of a star polygon. (Here, by *star polygon*, we mean the figure obtained by joining the vertices in a polygonal path, but skipping over one vertex each time. Experts use the Schläfli symbol $\{n/2\}$ to denote this particular star polygon.) FIGURE 1 shows that l is $2R\sin(2\pi/n)$ and that the side of the polygon is $h = 2R\sin(\pi/n)$.

Before fitting our n-gon into a square, we first fit it into a strip. It turns out to be simplest to consider the n-gon to have a fixed radius R and find out how wide the strip must be to contain it. Depending on the orientation of the n-gon, the necessary width will fall somewhere between the diameter and the altitude. Since the altitude is smaller, we might decide that it is best to orient the polygon with its altitude along one dimension of the square. Unfortunately, the n-gon is always fatter in the perpendicular direction.

Therefore, let us find the narrowest strip of paper that can contain a given polygon when it is tilted with respect to the strip at an angle θ. For a *fixed* rotation angle θ, the minimum strip width is denoted $a(\theta)$. The odd case is shown in FIGURE 1.

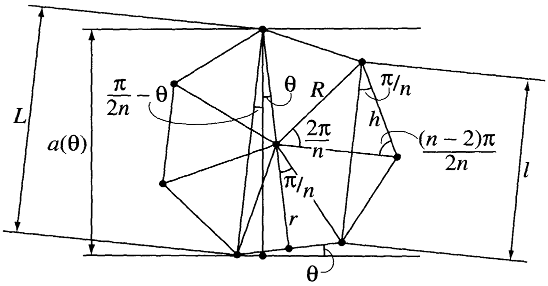


Figure 1 The width of a strip that contains a tilted polygon

Aided by FIGURE 1, which shows the odd case, the reader can verify the following formula for $a(\theta)$, given in terms of the number n of edges (or vertices) of the polygon, and the radius R of the circumscribed circle:

- If n is odd, $a(\theta)$ is (π/n)-periodic and $a(\theta) = L\cos(\theta - \pi/(2n))$ for $\theta \in [0, \pi/n]$. Here, $L = 2R\cos(\pi/(2n))$ is the diameter of the polygon.
- If n is even, $a(\theta)$ is $(2\pi/n)$-periodic and $a(\theta) = L\cos(\theta - \pi/n)$ for $\theta \in [0, 2\pi/n]$. In this case, $L = 2R$.

Now that we know the narrowest strip that contains a tilted regular polygon, we add a second strip of paper, orthogonal to the first one, and require that it too must contain the polygon. Observe from FIGURE 2 that the width of this second strip will be $a(\theta + \pi/2)$. If the polygon is to fit into a square, each strip must have width $A(\theta) = \max\{a(\theta), a(\theta + \pi/2)\}$ as shown in FIGURE 3. Thus, if we minimize $A(\theta)$, we find the smallest square, which is equivalent to find the largest regular polygon within a given square. From the previous expressions for $a(\theta)$ (depicted in FIGURE 3), we derive the following values for the side of the smallest square:

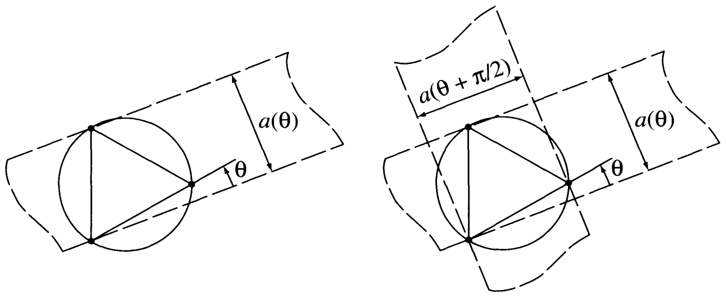


Figure 2 Two perpendicular strips

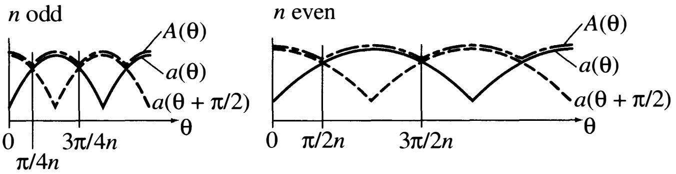


Figure 3 Strip problem solving procedure

- $\theta_{\text{opt}} = \pi/(4n)$ (modulo $\pi/(2n)$) if n is odd,
- $\theta_{\text{opt}} = \pi/(2n)$ (modulo π/n) if n is even.

Note that in each case, $a(\theta_{\text{opt}}) = a(\theta_{\text{opt}} + \pi/2)$ and so one can conclude that each side of the square touches at least one vertex of the optimal polygon. Moreover, using the formulas for these angles θ_{opt}, the reader may prove that each optimal polygon has at least one diagonal of the square as an axis of symmetry. FIGURE 4 shows the optimal polygon placements up to the octagon ($n = 8$ edges).

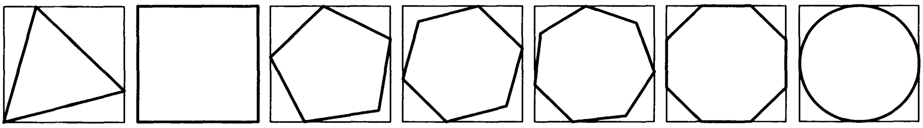


Figure 4 Placement of optimal polygons

We remark that this pattern also gives us the optimal polygons that fit any rectangular piece of paper, and not only square ones.

Folding the optimal hexagon and pentagon

The case of the optimal hexagon ($n = 6$ edges) involves the angle $2\pi/6$, and is very easy to construct. Given a square of paper with side 1, we use the previous formula for θ_{opt} to deduce that $R = 1/(2\cos(\pi/12))$. This gives us the edge length, $h = R = \sqrt{2}(\sqrt{3} - 1)/2$. The side of the the star hexagon is $l = 2h\cos(\pi/6) = \sqrt{2}(3 - \sqrt{3})/2$. Our construction, which first produces the star hexagon, is shown in FIGURE 5.

- Step 1: Fold corner B onto the central vertical line to create line AE, which meets diagonal BD at F. Since $\sin(\angle B'AD) = 1/2$, this is a simple technique to get an angle

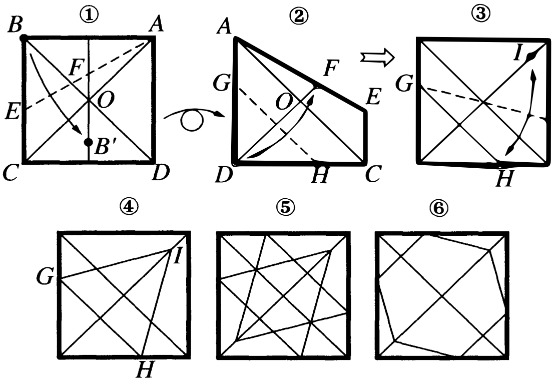


Figure 5 Folding sequence of the optimal hexagon

$\angle B'AD = \pi/6$. Then, $\angle BAE = \angle EAB' = \angle BAB'/2 = (\pi/2 - \angle B'AD)/2 = \pi/6$ again. (We have trisected the angle $\angle BAD = \pi/2$; this is easy for this particular angle, but more difficult for general case, as we will see for the nonagon. Of course, if we have only straightedge and compass we cannot trisect a general angle at all.) Notice that $\angle EAC = \pi/4 - \angle BAE = \pi/12$, so, $DF = DO + OF = OA + OA\tan(\angle EAC) = \sqrt{2}(1 + \tan(\pi/12))/2 = \sqrt{2}(3 - \sqrt{3})/2$. This is the length l of the desired polygon side. The following steps are needed to move the crease of length l in a correct position to obtain an edge of the optimal hexagon.

- Step 2: Turn over the model. Split DF in two by folding, and create the fold GH. Its length is $GH = DF$ and due to symmetry with respect to the diagonal, GH is an edge of the optimal star hexagon.
- Step 3: Unfold. Bring H onto the main diagonal at I, with G as an end point of the crease.
- Step 4: Fold GI and HI.
- Step 5: Complete the star polygon, using symmetries.

Folding a pentagon ($n = 5$ edges) requires the angle $2\pi/5$ and is more difficult than folding the hexagon. As early as 1989, Roberto Morassi [13] designed an origami construction of the optimal pentagon. The technique shown in FIGURE 7 has been developed independently and seems much simpler. As before, the star version of the polygon is used. With an initial square of unit edge length, we get $l = 1/\cos(\pi/20)$, as in FIGURE 6.

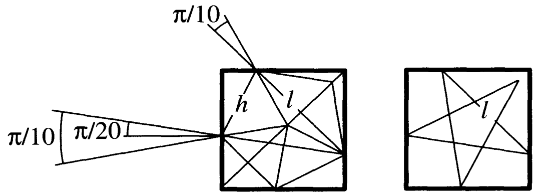


Figure 6 Optimal pentagon and its related star pentagon (or pentagram)

- Step 1: With D as the middle of the edge, bisect the angle $\angle BAD$. The crease is AC, $\tan(\angle DAE) = 1/2$ and $\angle BAC = \angle CAD = (1/2)\angle BAD = (1/2)(\pi/2 - \angle DAE)$. We can compute $\tan(\angle BAC) = (\sqrt{5} - 1)/2 = BC$; this is the reciprocal of the golden ratio.

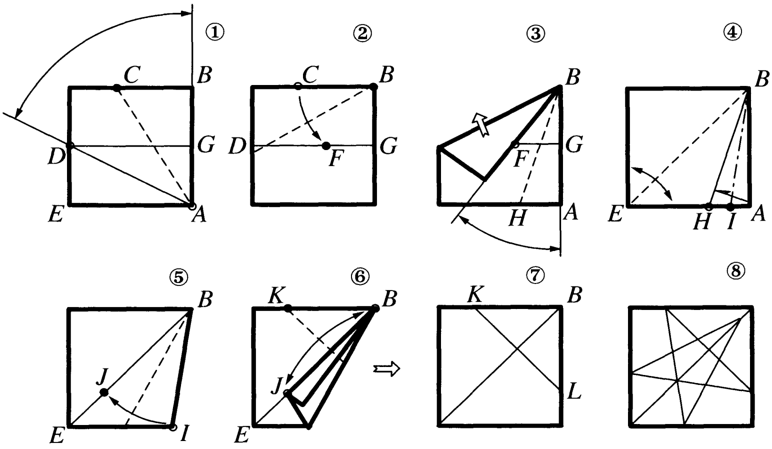


Figure 7 Folding sequence of the optimal pentagon

- Step 2: Fold C on the central horizontal line DG at F, with B as an end point of the crease. Since $\cos(\pi/5) = 1/(\sqrt{5} - 1)$ and $BG = 1/2$, we get $BF = BC = (\sqrt{5} - 1)/2 = 1/(2\cos(\pi/5)) = BG/\cos(\pi/5)$. This allows us to conclude that $\cos(\angle FBG) = BG/BF = \cos(\pi/5)$ and $\angle FBG = \pi/5$.
- Step 3: Bisect $\angle FBG$ to get $\angle HBA = \pi/10$. Unfold.
- Step 4: Bisect $\angle ABH$ (fold behind) to get $\angle ABI = \pi/20$ and $BI = 1/\cos(\pi/20)$. This is the length l of the optimal pentagon edge. As before, the following steps are needed to move the crease in a correct position.
- Step 5: Bring I on BE at J.
- Step 6: Fold in half BJ to get K. Unfold.
- Step 7: $KL = BJ = BI = l$ is the correct edge.
- Step 8: Complete the polygon.

Folding other optimal polygons

If the optimal square, triangle, and octagon, are the easiest regular polygons to design (the reader may begin to try to fold them, except for the square...), the optimal hexagon and pentagon are the next easiest. As soon as a regular polygon can be constructed by folding, the corresponding optimal polygons can be also folded with a technique similar to the one used in this paper [**4**].

In a publication from 1837, P. L. Wantzel [**16**] demonstrated which regular polygons are constructible with straightedge and compass (see also Carrega [**3**]). A necessary condition concerns the number of edges of these polygons: They must have $n = 2^p f_1 f_2 \ldots f_s$ edges, with p an integer and the numbers f_i different *primes* of the form $2^m + 1$, where m is also an integer. This result can be further simplified because a necessary (but not sufficient) condition for these f_i to be primes, is that they be Fermat numbers, that is, numbers of the form $2^{2^m} + 1$, m still being an integer. The only known Fermat primes to date are 3, 5, 17, 257, and 65537 [**17, 9**].

We conclude that all the previously folded polygons ($n = 3$, $n = 2^2 = 4$, $n = 5$, $n = 2^1 \times 3 = 6$, $n = 2^3 = 8$) can also be constructed with straightedge and compass. This will not be the case for heptagon ($n = 7$ edges) and nonagon ($n = 9$ edges), for instance. Using the technique mentioned in the introduction to solve third-order equations by folding, the previous set of constructible polygons can be extended to the set of polygons with $n = 2^m 3^q g_1 g_2 \ldots g_s$ edges, where the g_i are all different primes of the form $2^p 3^r + 1$, with m, p, q, and r being integers [**15, 12, 1**]. Such a construction will be necessary to fold the optimal heptagon and nonagon.

To introduce this new construction, let us recall one of its applications: the trisection of an arbitrary angle θ [**6**], known to be impossible with Euclidean constructions (for a nice discussion on this subject, the reader may refer to the web page `http://www.jimloy.com/geometry/trisect.htm`). To trisect an arbitrary angle, consider the construction of FIGURE 8, in which isosceles triangle $AA'B$ has been folded in half to construct two copies of angle γ. Construct the perpendicular AD to AB to get $\angle A'AD = \gamma$. The main idea is then to look at Δ, the perpendicular bisector of AA', and reflect across Δ to bring B onto B', C onto C', and D onto D'. Note that the angle $\theta = \angle D'A'B$ is 3γ.

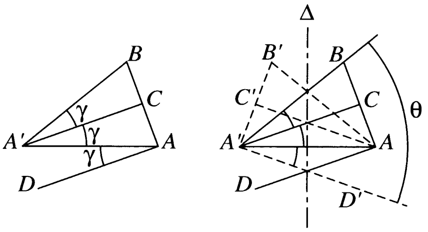


Figure 8 Illustrating the trisection problem

The whole construction can now be reversed to perform the trisection of an angle θ that is shown in FIGURE 9:

- Step 1: Suppose the angle $\angle D'A'E = \theta$ to be trisected is established by the fold $A'E$. Make two horizontal folds, with the only requirement that they be equally spaced ($A'C' = C'B'$). Fold the paper to simultaneously bring B' onto $A'E$ and A' onto the first horizontal fold. Call I the intersection of this new fold with Δ. Unfold.
- Step 2: It is not too hard to show that $\angle IA'E = \angle D'A'E/3$.

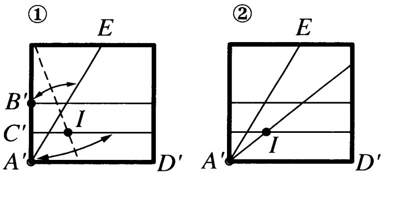


Figure 9 Solving the trisection problem

The trisection of an angle requires solving a third-order polynomial equation, while intersecting circles and lines (basically, the Euclidean constructions) requires only second-order polynomial equations. The new operation in FIGURE 9 allows us to construct additional polygons.

Since folding a nonagon ($n = 9$ edges) is a little easier than folding the heptagon, let us begin with it. FIGURE 10 briefly describes the corresponding sequence. Folding this figure requires precision (and a large square of paper), and proving that an exact optimal nonagon is obtained is a not such an easy task [**4**]. Both are left as challenges to the reader. Here are some guidelines:

- Step 1: Precrease diagonal and central lines. Fold $\pi/3 = \angle BAE$; D should lie on the central horizontal line. Unfold.
- Step 2: With the trisection method, fold $2\pi/9 = \angle BAF$ (then, $\angle FAC = \pi/4 - 2\pi/9 = \pi/36$). The point G is the intersection of AF and the central vertical line. Since $AH = 1/2$, $AG = 1/(2\cos(2\pi/9))$.
- Step 3: Fold the perpendicular to AF (and not to AC!) at G. It intersects diagonal AC at I. Then $AI = AG/\cos(\pi/36)$; this is the length of the edge of our star polygon. (Note that *nonagram* usually refers to a different star polygon, the {9/3} polygon, since edges connect every third vertex. Recall that we are talking about a {9/2} regular star polygon).
- Step 4: Duplicate this distance at JK.
- Step 5: Complete the nonagon (quite challenging, again).

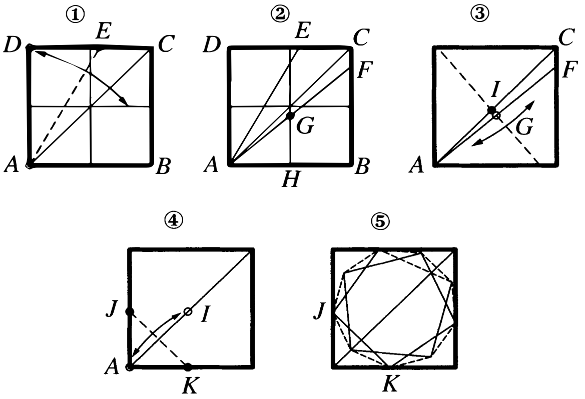


Figure 10 Folding sequence of the optimal nonagon

An even more challenging construction is the one of the heptagon ($n = 7$ edges). This time, the angle $2\pi/7$ is involved, unattainable with Euclidean constructions. It also requires the previous technique, designed to solve any third-order polynomial root [**12**]. (Let us just mention that $2\cos(2\pi/7)$ is a root of $t^3 + t^2 - 2t - 1 = 0$ and that the solving technique is the same as in [**11, 8**].) FIGURE 11 describes briefly the corresponding sequence, with many details left for the reader to verify:

- Step 1: Precrease diagonal and central lines. Fold the left and upper halves in half again and unfold.
- Step 2: Use previous technique, to get a fold HI of slope $2\cos(2\pi/7) = AI/AH$ (G is located halfway between the previous two horizontal folds).
- Step 3: Fold A to I, crease, and unfold.
- Step 4: Fold H onto the new horizontal line through J to obtain $\angle JAH' = 2\pi/7$ (because $AJ/AH' = AI/(2AH) = \cos(2\pi/7)$).
- Step 5: The intersection of the folded edge and the initial central line is K. Note that $\angle CAK = 2\pi/7 - \pi/4 = \pi/28$ and $AK = AE/\cos(2\pi/7) = 1/(2\cos(2\pi/7))$. Fold the perpendicular to AK (and not to AC!) at K: it intersects diagonal AC at L. Unfold (then $AL = AK/\cos(\pi/28)$; this is the length of the edge of our star heptagon, also called the {7/2} regular star polygon, or *heptagram*).
- Step 6: Duplicate this distance at MN.
- Step 7: Complete the heptagon.

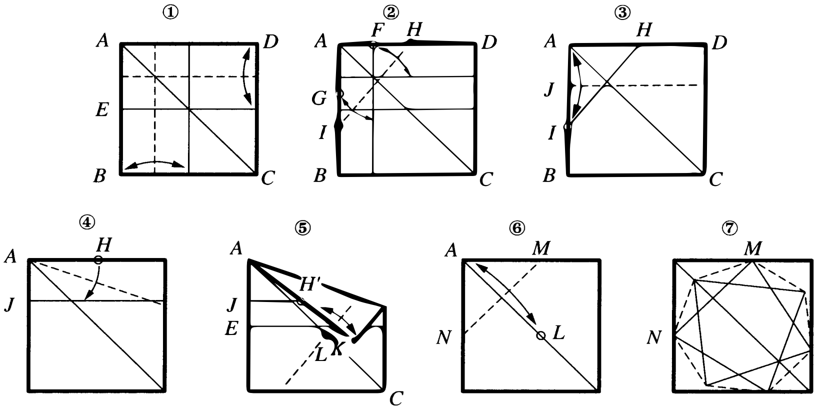


Figure 11 Folding sequence of the optimal heptagon

Prospects

Several regular polygons can be folded with Euclidean constructions. Their optimal versions can be folded also, though in practice they become less and less easy to obtain. With a special basic fold that is recalled in this paper, and that has now widely spread through the community of paperfolders, more can be done. But even with it, not all of the polygons can be folded. For instance, the first unattainable regular polygon is the *hendecagon* ($n = 11$ edges). A construction, simple enough to enter the standard repertoire, and allowing the construction of the regular hendecagon, would require us to solve a 5th-order polynomial equation (the trigonometric functions of $2\pi/11$ are roots of such a polynomial equation). This is still to come.

Acknowledgment. The author wishes to thank the generous referees for their help and fruitful discussions about the topic of this note.

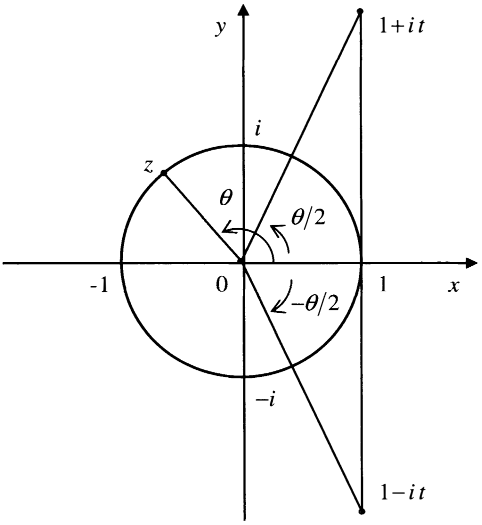
REFERENCES

1. R. C. Alperin, A mathematical theory of origami constructions and numbers, *New York J. Math.* **6** (2000), 119–133.
2. M. P. Beloch, Sul metodo del ripiegamento della carta per la risoluzione dei problemi geometrici, *Periodico di Mathematiche IV*, **16**, :2 (1936), 104–108.
3. J.-C. Carrega, *Théorie des corps : la règle et le compas*, Hermann, Paris (1981, 1989).
4. D. Dureisseix, Searching for optimal polygon—remarks about a general construction and application to heptagon and nonagon, (1997). `http://citeseer.nj.nec.com/article/dureisseix97searching.html`.
5. B. C. Edwards and J. Shurman, Folding quartic roots, this MAGAZINE, **74**:1 (2001), 19–25.
6. K. Fusimi, Trisection of angle by Hisashi Abe, *Science of Origami, a supplement to Saiensu*, (1980), page 8.
7. R. Geretschläger, Euclidean constructions and the geometry of origami, this MAGAZINE, **68**:5 (1995), 357–371.
8. R. Geretschläger, Folding the regular heptagon, *Crux Math.*, **23**:2 (1997), 81–88.
9. R. K. Guy, Mersenne Primes, Repunits, Fermat Numbers, Primes of Shape, in *Unsolved Problems in Number Theory*, Springer-Verlag, second edition (1994), chapter A3, 8–13.
10. H. Huzita, editor, *Proceedings of the First International Meeting of Origami Science and Technology*, Ferrara (1989).
11. H. Huzita, Drawing the regular heptagon and the regular nonagon by origami (paper folding), *Symmetry: Culture and Science*, **5**:1 (1994), 69–84.
12. J. Justin, Résolution par le pliage de l'équation du troisième degré et applications géométriques, in Huzita [**10**] (1989), 251–261.

13. R. Morassi, The elusive pentagon, in Huzita [**10**] (1989), 27–37.
14. T. S. Row, *Geometric Exercises in Paper Folding*, Dover Publishing, New York (1893, 1966).
15. B. Scimeni, Draw of a regular heptagon by the folding, in Huzita [**10**] (1989), 71–77.
16. P.-L. Wantzel, Recherches sur les moyens de reconnaître si un problème de géométrie peut se résoudre à la règle et au compas, *J. Math. Pures Appl.* **2** (1837), 366–372.
17. E. W. Weisstein, *The CRC Concise Encyclopedia of Mathematics*, CRC Press (1998). `http://mathworld.wolfram.com/FermatNumber.html`.
18. J. Wetzel, Fits and covers, this MAGAZINE **76** (2003), 349–363.

Proof Without Words: Complex Numbers with Modulus One

Any complex number z with $|z| = 1$, except $z = -1$, can be expressed as $\frac{1+it}{1-it}$ for some real number t.



$$\arg\frac{1+it}{1-it} = \arg(1+it) - \arg(1-it) = \frac{\theta}{2} - \left(-\frac{\theta}{2}\right) = \theta = \arg z,$$

$$\left|\frac{1+it}{1-it}\right| = \frac{|1+it|}{|1-it|} = 1 = |z|.$$

Observe that $t = \tan\frac{\theta}{2}$.

—Jean Huang (senior)
Technical High School
St. Cloud, MN 56301

NOTES

Eigencircles of 2 × 2 Matrices

M. J. ENGLEFIELD
Honorary Research Associate
School of Mathematical Sciences
Monash University

G. E. FARR
Clayton School of Information Technology
Monash University
Clayton, Victoria 3800 Australia
Graham.Farr@infotech.monash.edu.au

Geometry is a familiar source of insight in linear algebra. Examples include the the geometric effect of linear transformations (rotation, reflection, shearing, etc.) and the interpretation of the determinant of a 2 × 2 matrix as the area of a parallelogram defined by the columns of the matrix (which has higher dimensional analogues). In this paper, we describe a different kind of geometric construction for 2 × 2 matrices, which links properties of the eigenvalues and eigenvectors of the matrix to simple geometric properties of a circle that we associate to the matrix.

If A is a square matrix and λ is a number, then a nonzero vector $\mathbf{w}$ such that $A\mathbf{w} = \lambda\mathbf{w}$ is an *eigenvector* of A, and λ is an *eigenvalue* of A. For a 2 × 2 matrix this is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} \zeta \\ \eta \end{pmatrix} = \lambda\begin{pmatrix} \zeta \\ \eta \end{pmatrix} = \begin{pmatrix} \lambda\zeta \\ \lambda\eta \end{pmatrix}$$

or

$$(a-\lambda)\zeta + b\eta = 0, \quad c\zeta + (d-\lambda)\eta = 0. \tag{1}$$

The eigensystem calculation is to find the values of λ, ζ, and η that satisfy this equation, other than the trivial $\zeta = \eta = 0$. An algebraic procedure is given in any linear algebra textbook.

This article presents a geometric construction for the solution, using a circle we call the *eigencircle.* Real eigenvalues are given by two points on the eigencircle, and certain intersecting chords through these points give corresponding eigenvectors (FIGURE 2). A matrix with complex eigenvalues determines an eigencircle in the same way (FIGURE 1) and this gives the eigenvalues (FIGURE 3b), but the eigenvectors illustration needs a third dimension to show the imaginary parts (FIGURE 4).

As well as these computational aspects, we find that certain properties of the eigensystem correspond to geometric properties of the circle. One example, not commonly given in algebraic treatments, is the angle between two real eigenvectors. The eigencircles of all matrices that share the same eigenvalues form what is called a coaxial system, intersecting in the eigenvalue points if the eigenvalues are real.

The second equation of (1) gives, by inspection,

$$\mathbf{w} = \begin{pmatrix} \zeta \\ \eta \end{pmatrix} = \text{any multiple of } \begin{pmatrix} d-\lambda \\ -c \end{pmatrix}. \tag{2}$$

If $\mathbf{w}$ is nonzero, it is an eigenvector provided the first equation of (1) is also satisfied:

$$(a-\lambda)(d-\lambda)-bc=0. \tag{3}$$

Thus the eigenvalues of A satisfy a quadratic equation (the *characteristic equation* of A).

Define a determinant

$$B(x,y)=\begin{vmatrix} a-x & b-y \\ c+y & d-x \end{vmatrix}.$$

The second degree terms are x^2+y^2, so $B(x,y)=0$ is the equation of a circle, when (x,y) are taken to be Cartesian coordinates. This is the *eigencircle* of $A=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. From (3), if $(\lambda,0)$ lies on the circle, then λ is an eigenvalue of A.

Subsequently, we use the following notation: $\overrightarrow{PQ}$ denotes the vector from the point P to the point Q; $\overline{PQ}$ refers to the line segment; the signed length is denoted by PQ, so that $PQ=-QP$, and when $\overline{PQ}$ is parallel to one of the coordinate axes, the sign of PQ is naturally determined by the relative coordinates of P and Q. For example, in FIGURE 1, $FG>0$, $GE>0$, $EH<0$, and $HF<0$.

Geometric construction of real eigensystems Substituting any of the coordinates $F=(a,b)$, $G=(d,b)$, $E=(d,-c)$, or $H=(a,-c)$ into the matrix whose determinant is $B(x,y)$ produces a zero row or column. These four points, which determine a rectangle, therefore lie on the eigencircle of the matrix A. FIGURE 1 is drawn assuming $a<d$ and $b<-c$, but the other three possibilities merely require the reversal of positive direction on one or both of the coordinate axes. Segments $\overline{FE}$ and $\overline{HG}$ are diameters of the eigencircle, even in the special cases $a=d$ $(F=G,\ H=E)$ or $b=-c$ $(F=H,\ G=E)$. Other special cases, such as both $d=a$ and $c=-b$, are discussed at the end of this section.

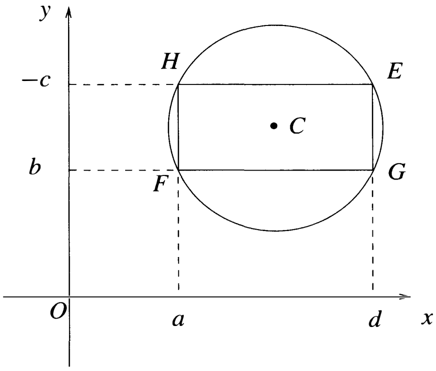


Figure 1 The eigencircle of the 2 × 2 matrix $A=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$

If the x-axis intersects the eigencircle, the eigensystem can be easily determined. If $L=(\lambda,0)$ is a point of intersection, the signed length $OL=\lambda$ is an eigenvalue of A, and the vector $\overrightarrow{LE}$ gives a corresponding eigenvector, where, as before, the point $E(d,-c)$ is the one that makes the second row of the matrix vanish. FIGURE 2 shows a case where $OL_1=\lambda_1<0<\lambda_2=OL_2$. Equation (2) verifies this eigenvector prescription, as

$$\begin{pmatrix} d \\ -c \end{pmatrix} - \begin{pmatrix} \lambda \\ 0 \end{pmatrix} = \overrightarrow{LE}.$$

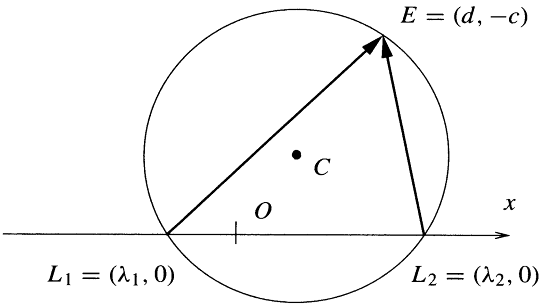


Figure 2 Intersections L_i of the real axis with the eigencircle give eigenvalues λ_i; the vectors $\overrightarrow{L_iE}$ are eigenvectors.

There are some special cases, which we encourage the reader to examine:

(i) If Ox is a tangent to the eigencircle, $L_1 = L_2$ and there is just one eigenvalue.

(ii) If $b = 0$ or $c = 0$, then Ox contains $\overline{FG}$ or $\overline{HE}$, and the eigenvalues are a and d, which includes $b = c = 0$, in which case $\overline{FG} = \overline{HE}$ is a diameter defining the eigencircle.

(iii) The prescription for the eigenvector fails for the eigenvalue d when $c = 0$, because then $L_2 = E$. In that case, the direction of the eigenvector can be shown to be given by the tangent at E.

(iv) When $a = d$ and $c = -b$, the eigencircle becomes a single point $F = G = E = H$. There is a real eigenvalue d only if $c = 0$, so that the single point is $(d, 0)$ on the real axis, and then any nonzero vector is an eigenvector.

Note that the construction can be reversed, to obtain a matrix from its real eigenvalues and eigenvectors. The eigenvalues give the points L_1 and L_2 (on the x-axis) required for FIGURE 2, and E is the intersection of the lines through L_1 and L_2 having the directions of the given eigenvectors. Then triangle L_1L_2E determines the eigencircle. Inscribe in the eigencircle a rectangle with E as one vertex, and sides parallel and perpendicular to the axes, as in FIGURE 1. Then the rows of the matrix can be written down from the coordinates of $E(d, -c)$ and $F(a, b)$.

Complex eigenvalues To include cases when the x-axis does not intersect the eigencircle, the geometry of the previous section may be rewritten using a result proved in Euclid's *Elements* III. 35–36 [**4**]: If a line through any point P meets a circle in points Q and R, the product of the lengths, $PQ \cdot PR$, is the same for any direction of the line. The lengths are signed, so the product is positive (negative) when P is outside (inside) the circle. The value of $PQ \cdot PR$ is called the *power* of the point P with respect to the circle.

Coordinate geometry shows that the power of P may be obtained from the equation of the eigencircle by substituting the coordinates of P into $B(x, y)$. In particular, the power of $O = (0, 0)$ is $\det A$, which is thus positive or negative according as O is inside or outside the circle. For example, in FIGURE 3(b), $\det A = OW^2$.

Our application to eigenvalues, shown in FIGURE 3, takes for P the point Y, which is the projection of the center C of the eigencircle onto the x-axis; the line through

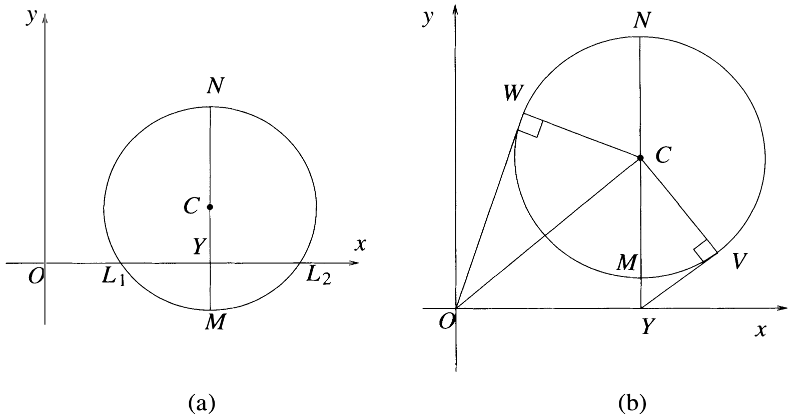


Figure 3 (a) Real eigenvalues $OL_i = OY \pm \sqrt{-YM \cdot YN}$; (b) Complex eigenvalues $OY \pm iYV$; YV (and OW, used later) are tangents

C and Y defines the diameter $\overline{MN}$. In FIGURE 3(a) the power of Y is $YL_1 \cdot YL_2 = YM \cdot YN$. As $YL_1 = -YL_2$, $YL_1^2 = -YM \cdot YN > 0$, and the real eigenvalues $OY + YL_i$ are

$$\lambda = OY \pm \sqrt{-YM \cdot YN}. \qquad (4)$$

This prescription is also valid when the x-axis does not intersect the eigencircle, as in FIGURE 3(b), where $YM \cdot YN > 0$, so (4) gives complex values. If YV is a tangent at V to the circle, then the power of point theorem gives $YM \cdot YN = YV^2$. Thus YV gives the imaginary part of the complex eigenvalue.

To show that (4) satisfies (3), so that λ is an eigenvalue, one uses the coordinate geometry of the eigencircle to obtain the coordinates of Y, M, and N. This is left to the reader.

For a geometric representation of $\begin{pmatrix} d-\lambda \\ -c \end{pmatrix}$ when $\lambda = OY + iYV$, a third dimension may be introduced to show the imaginary part. Put $\lambda = f + ih$. In FIGURE 4, $\overline{YL}$ and $\overline{YK}$, each with length $YV = h$, are perpendicular to the plane of the eigencircle. Then $\overrightarrow{LE}$ and $\overrightarrow{KE}$ represent the complex eigenvectors, the third component giving the imaginary part of the eigenvalue:

$$\overrightarrow{LE} = \overrightarrow{OE} - \overrightarrow{OL} = \begin{pmatrix} d \\ -c \\ 0 \end{pmatrix} - \begin{pmatrix} f \\ 0 \\ h \end{pmatrix}$$

represents the complex eigenvector $\begin{pmatrix} d-f-ih \\ -c \end{pmatrix}$.

Special case (iv) can now be considered. If the radius of the circle in FIGURE 4 shrinks to zero, then all the points on the circle coincide with C, which is $(a, b, 0) = (d, -c, 0)$. Thus $YV = YC = b$ and $OY = a$. The eigenvalues are $f \pm ig = a \pm ib$, and $\overrightarrow{LC} = (0 \;\; b \;\; -b)^T$ and $\overrightarrow{KC} = (0 \;\; b \;\; b)^T$ are eigenvectors.

Geometric derivation of eigensystem properties Our geometric description of eigensystems of 2 × 2 real matrices involves circles, chords, and tangents. These are related by many celebrated theorems of Euclidean geometry. We apply these to give geometric derivations of the properties of eigensystems usually presented in algebraic treatments.

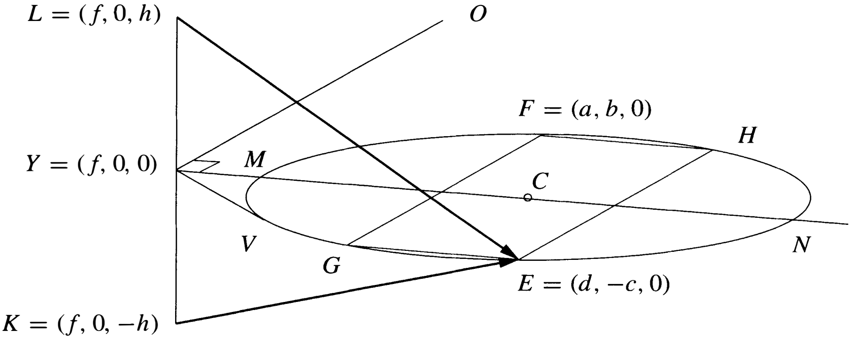


Figure 4 $\overline{YL}$ and $\overline{YK}$ are perpendicular to the eigencircle, with lengths equal to that of the tangent $\overline{YV}$

(a) From (4), the sum of the eigenvalues is $2\,OY$; FIGURE 1 gives the coordinates of the center as $((a+d)/2, (b-c)/2)$, and FIGURE 3 then gives $OY = (a+d)/2$. Thus the sum of the eigenvalues is $a+d$, the trace of A.

(b) From FIGURE 3(a), the product of real eigenvalues is $OL_1 \cdot OL_2 =$ power of $O = \det A$.

(c) From (4), the product of complex eigenvalues is $OY^2 + YM \cdot YN$. From FIGURE 3(b) this is $OY^2 +$ (power of Y), which is

$$OY^2 + YV^2 = OY^2 + YC^2 - CV^2 = OC^2 - CV^2 = OC^2 - CW^2,$$

where W is an eigencircle point defined by a tangent from O, finally giving $OW^2 =$ power of $O = \det A$.

(d) To investigate the angle between eigenvectors, which we will call β, recall the Euclidean theorem that the angle subtended at the center C by a chord (in this case L_1L_2) is twice the angle subtended at any point on the circumference (for instance, E). This quickly leads to $\angle L_1EL_2 \equiv \angle L_1CY$. Drawing segment $\overline{YC}$ into FIGURE 2 shows that $\cos\beta = YC/L_1C$; then, since YC is the y-coordinate of C, given in (a), we have

$$YC = \frac{1}{2}(b-c),$$

and since L_1C, the circle's radius, is $FC = \frac{1}{2}FE$ in FIGURE 1, we have

$$L_1C^2 = \tfrac{1}{4}(d-a)^2 + \tfrac{1}{4}(b+c)^2.$$

Combining these gives

$$\cos^2\beta = \frac{(b-c)^2}{(d-a)^2 + (b+c)^2}.$$

(e) For the special case of a symmetric matrix ($b = c$), the x-axis is a diameter, so the famous result that $\beta = \pi/2$ for this case follows alternatively from the Euclidean theorem that the angle subtended by a diameter (in a semicircle) is a right angle.

(f) From FIGURE 1, the eigencircle diagram for the transposed matrix A^T (which interchanges b and c) is obtained by reflection in the real axis. For real eigenvalues FIGURE 3(a) becomes FIGURE 5, showing that A and A^T have the same eigenvalues. In FIGURE 2 note that $E(d, -c)$ is determined by the second row of A, so the appropriate corresponding point on the eigencircle of A^T is G'. Then $\overrightarrow{L_iG'}$ are eigenvectors of A^T, using the appropriate points in FIGURE 5. These vectors are

$$\begin{pmatrix} d \\ -b \end{pmatrix} - \begin{pmatrix} \lambda_i \\ 0 \end{pmatrix},$$

as expected from (2), and their transposes are row eigenvectors of A. Thus row eigenvectors are determined by the second column of A.

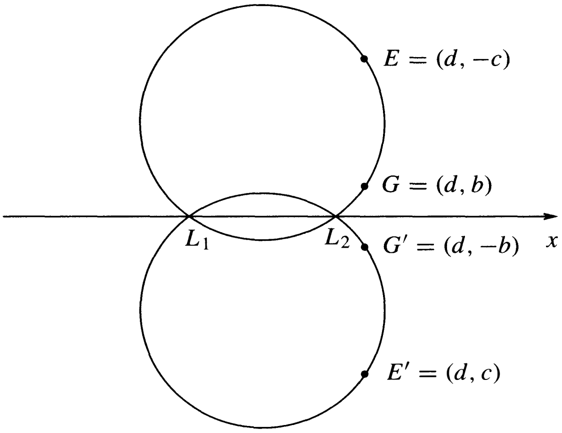


Figure 5 The circle $L_1L_2G'E'$ is the eigencircle of A^T, by reflecting in the real axis the eigencircle of A

For complex eigenvalues similar observations apply. A reflection of FIGURE 3(b) in the x-axis shows the eigenvalues are unchanged ($OY \pm iYV \rightarrow OY \mp iYV'$ and $YV' = YV$). In a reflection of FIGURE 4 the complex eigenvectors will be represented by $\overrightarrow{LG'}$ and $\overrightarrow{KG'}$.

Real matrices with specified eigenvalues Given real eigenvalues, the points L_1 and L_2 (FIGURE 2) are determined. Take any point C on the perpendicular bisector l of L_1L_2 (if $\lambda_1 = \lambda_2$, then l is the line perpendicular to the x-axis through $L_1 = L_2$). Draw the circle with center C through L_1 and L_2 (or with the x-axis tangent if $L_1 = L_2$), and take any point E on this circle. If EF is a diameter, then the information in FIGURE 1 is determined, and we can write down a matrix from the coordinates of E and F.

Given complex eigenvalues $f \pm ih$, draw a circle with center $Y(f, 0)$ and radius h, and the line l through Y perpendicular to Ox. The tangent to this circle q at any point V meets l in a point C (FIGURE 6). Take a circle with center C and radius CV. From FIGURE 3(b) any diameter EF of this eigencircle determines a matrix with the required eigenvalues. As V moves round q, C assumes all points of l outside the diameter UZ, and the resulting eigencircles give all the required matrices. Each eigencircle cuts q orthogonally. The points $Z = (f, h)$ and $U = (f, -h)$ are limiting cases of eigencircles of zero radius, for which $a = d = f$ and $c = -b = \pm h$ (see FIGURE 4).

In the real case, the eigencircles are a coaxial set intersecting in L_1 and L_2 (FIGURE 7(a)). In the complex case the eigencircles are a non-intersecting coaxial system (FIGURE 7(b)). The real axis is always the radical axis of the coaxial system.

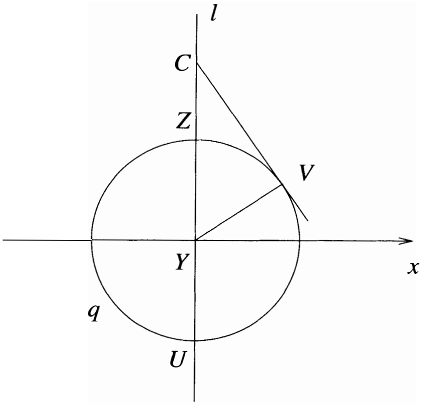


Figure 6 Circle q with center $Y = (f, 0)$ and radius h

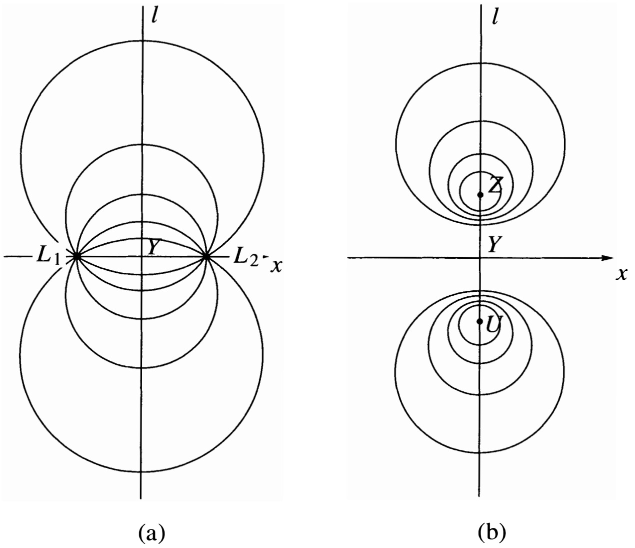


Figure 7 System of coaxial eigencircles giving all eigencircles for the same pair of eigenvalues: (a) Real eigenvalues; (b) Complex eigenvalues

Alternative derivation for a complex eigensystem Define a determinant

$$B'(x, y, z) = \begin{vmatrix} a - x - iz & b - y \\ c + y & d - x - iz \end{vmatrix}$$

where (x, y, z) are Cartesian coordinates. Then $B'(x, y, z) = 0$ can be written as two real equations

$$z(2x - a - d) = 0 \qquad (5)$$

$$x^2 + y^2 - z^2 - x(a + d) - y(b - c) + ad - bc = 0. \qquad (6)$$

Assuming $z \neq 0$ ($z = 0$ gives $B(x, y)$ and previous work) then $x = (a + d)/2$, and

$$\left(y - \tfrac{1}{2}(b - c)\right)^2 - z^2 = \rho^2 \qquad (7)$$

where ρ is the radius of the eigencircle (L_1C in FIGURE 2). This represents a rectangular hyperbola in the plane $x = (a+d)/2$ (FIGURE 8), with center C and axis $\overline{MN}$, which is the eigencircle diameter shown in FIGURES 3(b) and 4.

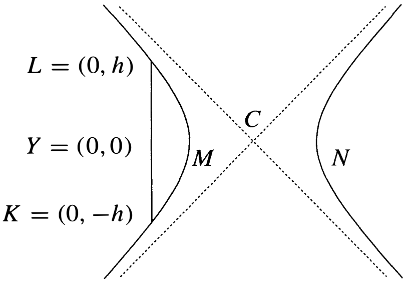


Figure 8 All points have $x = (a+d)/2$; the hyperbola consists of (y, z) coordinates.

Exactly as in the real case, the coordinates of the hyperbola points L and K with $y = 0$ give the complex eigenvalues $(a+d)/2 \pm ih$ of the matrix A. Corresponding eigenvectors have already been shown in FIGURE 4.

Some extensions The eigencircle is

$$0 = B(x, y) = \begin{vmatrix} a - x & b - y \\ c + y & d - x \end{vmatrix}.$$

The determinant B' used in the previous section is B with x allowed to be complex.

One extension we have investigated allows both x and y in B to be complex. Then (5) and (6) have four real variables and represent a two-dimensional surface in four-dimensional space. Eliminating one variable, as in reducing (5)–(6) to (7), leaves a hyperboloid of one sheet in a three-dimensional space. The eigencircle and the hyperbola in FIGURE 8 are recovered as plane sections of the hyperboloid.

Another extension uses the determinant

$$B''(x, y) = \begin{vmatrix} a - x & b - y \\ c + \bar{y} & d - \bar{x} \end{vmatrix}$$

with x and y complex, so the matrix $\begin{pmatrix} x & y \\ -\bar{y} & \bar{x} \end{pmatrix}$ may be regarded as representing a *quaternion* (just as the matrix $\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$ may be regarded as representing a complex number when x and y are real: see [**5**, §41]). This again leads to a two-dimensional surface in a four-dimensional real space, and eliminating one coordinate gives a sphere. We call this the *eigensphere* of the matrix, and the eigencircle is a great circle in the coordinate plane, defined by $\operatorname{Im} x = \operatorname{Im} y = 0$.

Using the real determinant

$$\begin{vmatrix} a - x & b + \alpha y \\ c + y & d - x \end{vmatrix}$$

allows illustration of the eigensystem using any conic; ellipses, hyperbolas, or a parabola, respectively appear with $\alpha < 0$, $\alpha > 0$, or $\alpha = 0$.

Details of these extensions are left for a sequel to this paper.

Finally, we mention a body of research on *multiparameter* eigenvalues such as the points on the eigencircle considered here. The first study of multiparameter eigenvalues of matrices seems to be a paper by Carmichael [3] in 1921. Other examples include [1, 2, 6, 7]. Such work typically considers systems of the form

$$\sum_{j=1}^{k} \lambda_j A_{ij} \mathbf{x}_i = \mathbf{0}, \tag{8}$$

where each A_{ij} is an $m_i \times n_i$ matrix and $\mathbf{x}_i$ is an n_i-element vector. We seek k-tuples $(\lambda_j)_{j=1}^{k}$ such that (8) can be solved with each $\mathbf{x}_i$ nonzero. Much of this work is at quite a general level and there seems to be little explicit discussion of the interesting special case considered here. We have focused on the properties of this special case, which we believe deserves to be better known.

Acknowledgment. We thank the referees for their helpful comments.

REFERENCES

1. F. V. Atkinson, Multiparameter spectral theory, *Bull. Amer. Math. Soc.* **74** (1968), 1–27.
2. P. Binding and P. J. Browne, Two parameter eigenvalue problems for matrices, *Linear Algebra Appl.* **113** (1989), 139–157.
3. R. D. Carmichael, Boundary value and expansion problems: algebraic basis of the theory, *Amer. J. Math.* **43** (1921), 69–101.
4. T. L. Heath, *The Thirteen Books of Euclid's Elements, Vol. 2: Books III-IX* (2nd edn.), Dover, New York, 1956.
5. V. V. Prasolov, *Problems and Theorems in Linear Algebra*, Translations of Mathematical Monographs **134**, Amer. Math. Soc., Providence, RI, 1994.
6. B. D. Sleeman, *Multiparameter Spectral Theory in Hilbert Space*, Research Notes in Mathematics **22**, Pitman, London, 1978.
7. H. Volkmer, *Multiparameter Eigenvalue Problems and Expansion Theorems*, Lecture Notes in Mathematics **1356**, Springer-Verlag, Berlin, 1988.

The Volume Swept Out by a Moving Planar Region

ROBERT L. FOOTE
Wabash College
Crawfordsville, IN 47933
footer@wabash.edu

I would like to call attention to a beautiful theorem about volume. The only place I have found it is in Courant's calculus text [3, p. 295], [4, p. 451]. Given that this book is a classic and that the result is both simple and elegant, it is surprising that it has not appeared in every calculus text since.

The result generalizes both Cavalieri's Principle and the Theorem of Pappus as a means for computing the volume swept out by a moving planar region. Somewhat informally, let S_t, $a \leq t \leq b$, be a planar region of area $A(t)$ moving in space. Let $\mathbf{n}(t)$ be a continuous unit vector normal to the plane of the region, and let $\mathbf{v}(t)$ be the velocity of the centroid of S_t. Then the signed (or oriented) volume swept out by S_t is

$$V = \int_a^b A(t)\, \mathbf{n}(t) \cdot \mathbf{v}(t)\, dt. \tag{1}$$

Intuitively, the volume is signed in the following sense. The vector $\mathbf{n}(t)$ indicates an orientation or forward direction (see FIGURE 1). Volume swept out in this direction is

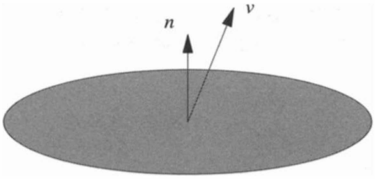


Figure 1 The velocity **v** and forward direction **n** of a moving planar region

taken to be positive; volume swept out in the opposite (or backward) direction is taken to be negative. This is handled in the integrand of (1) by the sign of $\mathbf{n} \cdot \mathbf{v}$. The net signed volume swept out is, in general, the result of both forward and backward motions. The formula also allows for multiplicities: a point in the path of the moving region may be covered more than once and in both forward and backward directions, in which case the volume of some neighborhood of the point is counted in (1) accordingly.

A recent article by England and Miller [5] gives a variation of (1) that is both more general, in that it allows the reference curve $\boldsymbol{\gamma}$ to be one other than the path of the centroid, and more specialized, in that it requires $\boldsymbol{\gamma}'(t)$ to be perpendicular to the plane of S_t. The result is a nice formula for the signed volume swept out in terms of the geometry of $\boldsymbol{\gamma}$ and its relationship to the centroid of S_t. A proof of their result is given below (Theorem 2) as an application of (1).

The familiar Theorem of Pappus and Cavalieri's Principle are easily seen to be special cases of (1).

THEOREM OF PAPPUS. *Suppose S is a bounded planar region of area A that is revolved about a line ℓ lying in the plane of S. If S lies in one of the half-planes bounded by ℓ, then the volume of the solid of revolution swept out by S is $2\pi r A$, where r is the radius of the circle swept out by the centroid of S.*

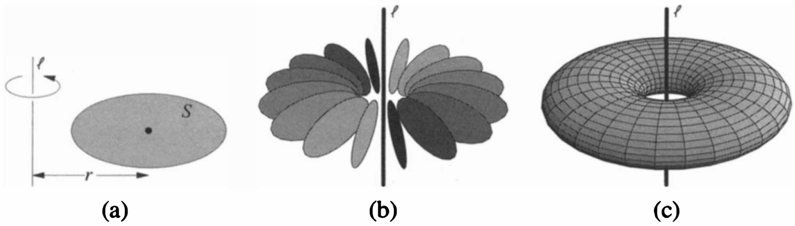


Figure 2 Theorem of Pappus

FIGURE 2a shows a region S and its centroid. FIGURE 2b shows S rotating about a line ℓ, resulting in the solid in FIGURE 2c. Since the area is constant and $\mathbf{n}$ can be taken to be $\mathbf{v}/\,||\mathbf{v}||$, the integral in (1) reduces to A times the distance moved by the centroid. Since every point of S moves in the forward direction, the signed volume is all positive.

The volume formula in the Theorem of Pappus is valid, in fact, even when ℓ passes through S, as in FIGURE 3a, as long as it is interpreted as signed volume. The line ℓ divides S along a chord into two subregions. As S rotates about this chord, the two subregions move in opposite directions, one moving forward and the other backwards. Each subregion generates a solid, shown in FIGURES 3c and 3d. Then (1) implies that $2\pi r A$ is the difference of the volumes of these two solids. This illustrates that a

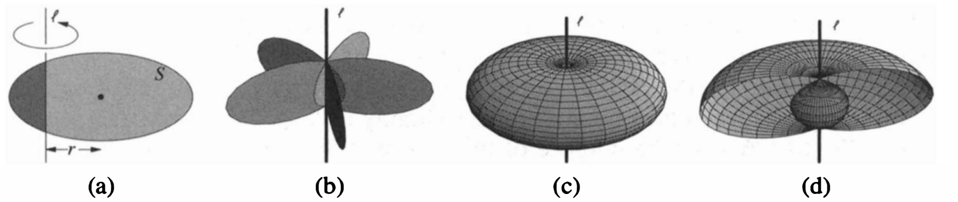


Figure 3 Generalized Theorem of Pappus

moving region can simultaneously generate both positive and negative signed volumes, an observation that is key to interpreting the integrand of (1).

Cavalieri's Principle states that if two solids have equal cross-sectional areas when cut by any plane parallel to a given plane, then the solids have the same volume (FIGURE 4). This is true even if the cross sections of one are stacked up straight and those of the other are skewed. What is important is not the exact path of the cross-sectional centroid, but rather the component of its motion perpendicular to the family of planes, which is computed by the dot product in (1). If the Theorem of Pappus is about rotations, Cavalieri's Principle, in contrast, is about translations. Since the cross sections are parallel, two nearby cross sections are (approximately) translates of each other.

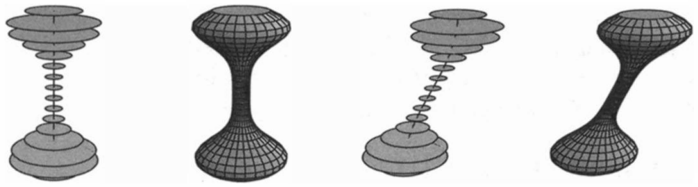


Figure 4 Cavalieri's Principle

Courant states (1) without proof. Instead, he states and proves its analog for the signed area swept out by a line segment moving in $\mathbb{R}^2$,

$$A = \int_a^b L(t)\,\mathbf{n}(t) \cdot \mathbf{v}(t)\,dt, \tag{2}$$

where L is the length of the segment, $\mathbf{n}$ is a forward-pointing unit normal vector, and $\mathbf{v}$ is the velocity of the midpoint, as in FIGURE 5. He goes on to use this to explain how a planimeter works (for more about planimeters, see [**6, 7, 8**]).

One can think of the integrand of (1) as the infinitesimal signed volume swept out due to an infinitesimal motion of the region (with a similar interpretation for the inte-

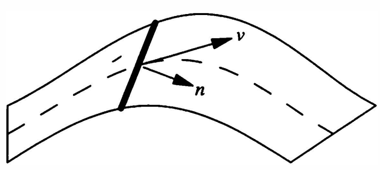


Figure 5 Moving segment sweeping out area in $\mathbb{R}^2$

grand of (2)). The simplicity of this expression hides the fact that even an infinitesimal motion can result in a combination of both positive and negative volumes. This happens, for example, for the region S in FIGURE 3ab as it rotates about the line ℓ. In general it happens when the region S_t rotates slightly to S_{t+dt} about one of its chords, ℓ_t, as in FIGURE 6. (For the purpose of this informal discussion we take a *chord* of S_t to be a line in the plane of S_t such that there are points of S_t on both sides of the line. This agrees with the usual notion of chord when the region is connected.) Part of the content of (1) and its proof is that only the displacement $\mathbf{v}\,dt = d\mathbf{x} = \mathbf{T}\,ds$ of the centroid relative to the direction of the normal $\mathbf{n}$ matters.

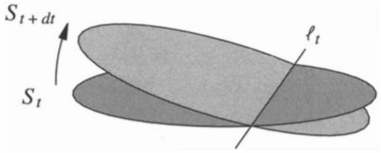


Figure 6 Infinitesimal rotation about a chord

I find the result in $\mathbb{R}^2$ less intriguing—the formula is almost obvious due to the symmetry of a line segment about its midpoint. The "symmetry" of a planar region about its centroid is more subtle. For example, a chord through the centroid generally doesn't bisect the area of the region. (A chord through the centroid of a triangle parallel to one of the sides divides the area in the ratio 4:5.) It follows from (1), however, that if the region rotates about such a chord, the signed volume swept out is zero. Consequently, (1) gives some insight into the geometric significance of the centroid that complements the physical center-of-mass interpretation in many calculus texts.

Definitions and proof To prove (1), we first make the notion of signed volume more precise. Suppose $U \subset \mathbb{R}^3$ is a bounded region, $F : U \to \mathbb{R}^3$ is C^1, and $F(U)$ is bounded. We take the *signed volume covered by* F to be the value of $\iiint_U J_F\,dV$, where $J_F = \det DF$ is the Jacobian determinant. It's clear that if F is one-to-one on $U^+ = \{\mathbf{x} \in U : J_F(\mathbf{x}) > 0\}$ and on $U^- = \{\mathbf{x} \in U : J_F(\mathbf{x}) < 0\}$, then the signed volume covered by F is the volume of $F(U^+)$ minus the volume of $F(U^-)$. (By Sard's Theorem [**11**], the image of $U^0 = \{\mathbf{x} \in U : J_F(\mathbf{x}) = 0\}$ has volume 0 even if the volume of U^0 is positive.) If F is finite-to-one on $U^+ \cup U^-$, the signed volume takes into account the multiplicity of the coverings. The signed volume swept out by the moving planar region in Theorem 1 is to be taken in this sense, as will be clear in the proof.

THEOREM 1. *Let P_t, $a \le t \le b$, be a family of planes. For each t suppose S_t is a region in P_t such that S_t varies continuously with t. Let $A(t)$ be the area of S_t, let $\mathbf{c}(t)$ be the centroid of S_t, and let $\mathbf{n}(t)$ be a unit normal to P_t. Assume $\mathbf{c}$ and $\mathbf{n}$ are C^1, and that $\cup_t S_t$ is bounded. Then the signed volume swept out by S_t is given by (1) (repeated here for emphasis), where $\mathbf{v}(t) = \mathbf{c}'(t)$.*

$$V = \int_a^b A(t)\,\mathbf{n}(t) \cdot \mathbf{v}(t)\,dt. \tag{1}$$

The proof is similar to that given by England & Miller [**5**], but somewhat simpler. Note that much of the notation introduced in the proof is used in the remainder of the paper.

Proof. Let $\mathbf{e}_1(t)$, $\mathbf{e}_2(t)$, and $\mathbf{e}_3(t)$ form a C^1, positively-oriented, orthonormal frame along $\mathbf{c}$ with $\mathbf{e}_1(t) = \mathbf{n}(t)$. (To get this one could, for example, apply the Gram-Schmidt process to $\mathbf{n}(t)$ and $\mathbf{i}$ to obtain $\mathbf{e}_1(t)$ and $\mathbf{e}_2(t)$, assuming $\mathbf{n}(t) \neq \mathbf{i}$ for all t, and then let $\mathbf{e}_3(t) = \mathbf{e}_1(t) \times \mathbf{e}_2(t)$.)

Define $F : [a, b] \times \mathbb{R}^2 \to \mathbb{R}^3$ by

$$F(t, x, y) = \mathbf{c}(t) + x\mathbf{e}_2(t) + y\mathbf{e}_3(t).$$

Note that $F(t, 0, 0) = \mathbf{c}(t)$, that is, F maps the t-axis to the path of the centroid. For $t = t_0$ fixed, note that $\{(t_0, x, y) : (x, y) \in \mathbb{R}^2\}$ is the plane in $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R}^2$ perpendicular to the t-axis at t_0. Similarly, $\{F(t_0, x, y) : (x, y) \in \mathbb{R}^2\}$ is the plane in $\mathbb{R}^3$ passing through $\mathbf{c}(t_0)$ perpendicular to $\mathbf{n}(t_0)$, which is the plane P_{t_0} containing S_{t_0}.

For each t, let $\tilde{S}_t = \{(t, x, y) : F(t, x, y) \in S_t\}$. Since $\mathbf{e}_2(t)$ and $\mathbf{e}_3(t)$ are orthonormal, then $\tilde{S}_t$ and S_t are congruent. Furthermore, the centroid of $\tilde{S}_t$ is $(t, 0, 0)$. Thus the map F achieves a straightening out of the moving region S_t into the moving region $\tilde{S}_t$ that stays perpendicular to a fixed direction (the t-axis), and so that the centroid of $\tilde{S}_t$ moves in a straight line with constant unit speed. The moving region $\tilde{S}_t$ sweeps out a bounded region $\tilde{\Omega}$ in $\mathbb{R} \times \mathbb{R}^2$.

The signed volume swept out by S_t, that is, the signed volume covered by $F|_{\tilde{\Omega}}$, is

$$V = \iiint_{\tilde{\Omega}} J_F \, dt\, dx\, dy = \iiint_{\tilde{\Omega}} \frac{\partial F}{\partial t} \cdot \left(\frac{\partial F}{\partial x} \times \frac{\partial F}{\partial y}\right) dt\, dx\, dy.$$

We have

$$\frac{\partial F}{\partial t} = \mathbf{v}(t) + x\mathbf{e}_2'(t) + y\mathbf{e}_3'(t), \qquad \frac{\partial F}{\partial x} = \mathbf{e}_2(t), \qquad \text{and} \qquad \frac{\partial F}{\partial y} = \mathbf{e}_3(t).$$

Using $\mathbf{e}_2(t) \times \mathbf{e}_3(t) = \mathbf{e}_1(t) = \mathbf{n}(t)$, we have

$$J_F = \frac{\partial F}{\partial t} \cdot \left(\frac{\partial F}{\partial x} \times \frac{\partial F}{\partial y}\right) = \mathbf{v}(t) \cdot \mathbf{n}(t) + x\mathbf{e}_2'(t) \cdot \mathbf{n}(t) + y\mathbf{e}_3'(t) \cdot \mathbf{n}(t). \tag{3}$$

Integrating over $\tilde{\Omega}$, we get

$$\begin{aligned} V &= \iiint_{\tilde{\Omega}} J_F \, dt\, dx\, dy = \int_a^b \left[\iint_{\tilde{S}_t} J_F \, dx\, dy\right] dt \\ &= \int_a^b \left[\mathbf{v}(t) \cdot \mathbf{n}(t) \iint_{\tilde{S}_t} dx\, dy \right. \\ &\qquad \left. + \mathbf{e}_2'(t) \cdot \mathbf{n}(t) \iint_{\tilde{S}_t} x\, dx\, dy + \mathbf{e}_3'(t) \cdot \mathbf{n}(t) \iint_{\tilde{S}_t} y\, dx\, dy\right] dt. \end{aligned}$$

Now, $\iint_{\tilde{S}_t} dx\, dy$ is the area of $\tilde{S}_t$, which is $A(t)$. Furthermore, $\iint_{\tilde{S}_t} x\, dx\, dy$ and $\iint_{\tilde{S}_t} y\, dx\, dy$ are both zero, since the centroid of $\tilde{S}_t$ is $(t, 0, 0)$. Thus the integral reduces to $\int_a^b A(t)\, \mathbf{v}(t) \cdot \mathbf{n}(t)\, dt$, which is the desired result. ■

As an example, consider the undulating torus swept out by a moving disk of varying radius, pictured in FIGURE 7ab. At time $t \in [0, 2\pi]$ the center of the disk is $\mathbf{c}(t) = 4\mathbf{u}(t)$, where $\mathbf{u}(t) = \cos t\, \mathbf{i} + \sin t\, \mathbf{j}$. The radius of the disk is $r(t) = 1 + \frac{1}{2}\cos(3t)$, and the unit normal to the disk is $\mathbf{n}_1(t) = \mathbf{c}'(t)/4$. The volume of the torus is then

$$V_1 = \int_0^{2\pi} A(t)\mathbf{v}(t) \cdot \mathbf{n}_1(t)\, dt = \int_0^{2\pi} \pi\big(2 + \cos(3t)\big)^2 dt = 9\pi^2 \approx 88.8.$$

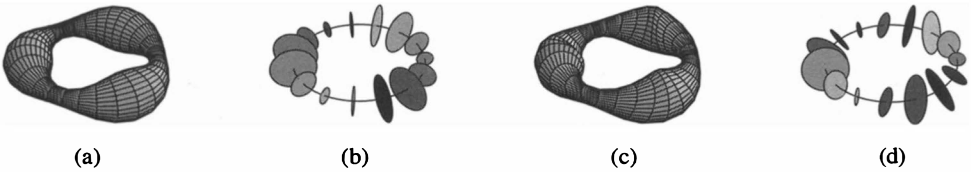


Figure 7 Moving disks sweeping out volume.

We now perturb the disk so that it wobbles as it sweeps out area, pictured in FIGURE 7cd. Let $\mathbf{N}_2(t) = \mathbf{n}_1(t) + \frac{1}{2}\sin(3t)\,\mathbf{k} + \frac{1}{3}\cos(2t)\,\mathbf{u}(t)$ (note that this simply adds to $\mathbf{n}_1$ something in its orthogonal complement) and take $\mathbf{n}_2(t) = \mathbf{N}_2(t)/\,\|\mathbf{N}_2(t)\|$ as the new unit normal. The new volume is

$$V_2 = \int_0^{2\pi} A(t)\mathbf{v}(t)\cdot\mathbf{n}_2(t)\,dt = \int_0^{2\pi} \frac{6\sqrt{2}\pi\left(2+\cos(3t)\right)^2}{\sqrt{85+4\cos(4t)-9\cos(6t)}}\,dt \approx 82.2,$$

which is slightly less than the original, as one might expect.

Moving forward One gets volume, as opposed to signed volume, when $J_F \geq 0$ on $\tilde{\Omega}$. From the proof, the interpretation of $J_F > 0$ is that every point of S_t moves in the forward direction. In many specific examples this is easy to see by inspection, but it's good to know conditions that imply it. Especially useful conditions are ones that can be applied directly to S_t, as opposed to $\tilde{S}_t$ or F.

Note, from (3), that J_F is linear in x and y. Consequently, if J_F is zero at some interior point of $\tilde{S}_t$, but not identically zero on $\tilde{S}_t$, then it is zero along a whole chord $\tilde{C}$ of $\tilde{S}_t$ and takes opposite signs on opposite sides of the chord. This is the infinitesimal version of the fact that if S_{t_1} and S_{t_2} intersect, they do so along a common chord. To see this, let $C = F(\tilde{C})$ be the corresponding chord of S_t. Points of S_t corresponding to points of $\tilde{S}_t$ for which $J_F > 0$ are moving forward. These points are all on one side of C. Points of S_t on the opposite side of C are moving backwards. It follows that C is the chord of intersection of S_t and S_{t+dt} in FIGURE 6, and F fails to be one-to-one on any neighborhood of any point of $\tilde{C}$. Thus, to conclude that (1) computes volume when the S_t are connected, it suffices to assume that S_{t_1} and S_{t_2} are disjoint when $t_1 \neq t_2$, with the possible exception of S_a and S_b, which might be identical. In the standard use of the Theorem of Pappus this is handled by the assumption that the planar region S lies in one of the half-planes of the line of rotation.

The following proposition gives a precise condition on S_t for $J_F > 0$ on $\tilde{S}_t$.

PROPOSITION. *$J_F > 0$ on $\tilde{S}_t$ if and only if $\left(\mathbf{x} - \mathbf{c}(t)\right)\cdot\mathbf{n}'(t) < \mathbf{v}(t)\cdot\mathbf{n}(t)$ for all $\mathbf{x} \in S_t$.*

Proof. The proof involves rewriting the expression for J_F in (3). Since $\mathbf{e}_2$ and $\mathbf{e}_3$ are perpendicular to $\mathbf{n}$, we have

$$\begin{aligned} 0 &= \frac{d}{dt}\Big(\left(x\mathbf{e}_2(t) + y\mathbf{e}_3(t)\right)\cdot\mathbf{n}(t)\Big) \\ &= x\mathbf{e}_2'(t)\cdot\mathbf{n}(t) + y\mathbf{e}_3'(t)\cdot\mathbf{n}(t) + \left(x\mathbf{e}_2(t) + y\mathbf{e}_3(t)\right)\cdot\mathbf{n}'(t). \end{aligned}$$

Now $\mathbf{c}(t)$, $\mathbf{e}_2(t)$ and $\mathbf{e}_3(t)$ determine a Euclidean coordinate system on P_t in which (x, y) are the coordinates of $\mathbf{x} = \mathbf{c}(t) + x\mathbf{e}_2(t) + y\mathbf{e}_3(t)$. By substitution into (3), J_F

can be written as

$$J_F = \mathbf{v}(t) \cdot \mathbf{n}(t) - \big(x\mathbf{e}_2(t) + y\mathbf{e}_3(t)\big) \cdot \mathbf{n}'(t) = \mathbf{v}(t) \cdot \mathbf{n}(t) - \big(\mathbf{x} - \mathbf{c}(t)\big) \cdot \mathbf{n}'(t)$$

for $\mathbf{x} \in S_t$, from which the proposition follows. ■

A few comments will reveal the geometric significance of the proposition. The inequality is a linear condition on points $\mathbf{x}$ in the plane P_t of S_t. If it holds for all $\mathbf{x} \in S_t$, then it holds when $\mathbf{x}$ is the centroid $\mathbf{c}(t)$. In this case we get $\mathbf{v}(t) \cdot \mathbf{n}(t) > 0$, which simply says that the centroid must be moving forward.

The cases when $\mathbf{n}'(t) = \mathbf{0}$ and $\mathbf{n}'(t) \neq \mathbf{0}$ are infinitesimal versions of the hypotheses of Cavalieri's Principle and the Theorem of Pappus. To see this, note that since $\mathbf{n}(t)$ is the unit normal to P_t, the vector $\mathbf{n}'(t)$ is a measure of the rotation of the family of planes. If $\mathbf{n}'(t) = 0$, then having the centroid move forward is sufficient for every point in P_t to do the same. This agrees with intuition—in this case the planes P_t and P_{t+dt} are parallel (at least to first order). On the other hand, if $\mathbf{n}'(t) \neq 0$, then the planes P_t and P_{t+dt} are not parallel, as shown in FIGURE 6. The inequality in the proposition defines a half-plane that contains the centroid. The boundary of the half plane is the line ℓ_t of intersection of P_t and P_{t+dt}. (The details of this are not difficult, and are left to the interested reader.) In order for the region S_t to be moving forward, it must lie in this half plane.

The interested reader may also verify that the inequality in the proposition is satisfied by the wobbling disk example, and so the signed volume computed is the actual volume of the solid swept out.

A variation The hypotheses used by England and Miller [**5**] are somewhat different than those in Theorem 1. Instead of following the centroid of the moving region, they follow another reference curve $\boldsymbol{\gamma}$ that is assumed to be perpendicular to the plane of the moving region. Their integral formula for volume involves the geometry of $\boldsymbol{\gamma}$ and the displacement from $\boldsymbol{\gamma}$ to the centroid.

THEOREM 2. (ENGLAND & MILLER [**5**]) *Suppose that $\boldsymbol{\gamma} : [0, \ell] \to \mathbb{R}^3$ is a C^2 curve parameterized by arc length s. Let P_s be the plane containing $\boldsymbol{\gamma}(s)$ that is perpendicular to $\boldsymbol{\gamma}'(s)$. Let S_s be a region in P_s that varies continuously with s. Let $\mathbf{N}(s)$ be the principal normal vector of $\boldsymbol{\gamma}$ and let $r(s) = \big(\mathbf{c}(s) - \boldsymbol{\gamma}(s)\big) \cdot \mathbf{N}(s)$, where $\mathbf{c}(s)$ is the centroid of S_s. Then the signed volume swept out by S_s is*

$$V = \int_0^\ell A(s)\big(1 - \kappa(s)r(s)\big)\,ds.$$

Note that $r(s)$ is the component of the vector from $\boldsymbol{\gamma}(s)$ to the centroid in the direction of the principal normal.

To prove this we first generalize Theorem 1 and then specialize to the situation of Theorem 2. Under the hypotheses and notation of Theorem 1, suppose that $\boldsymbol{\gamma} : [a, b] \to \mathbb{R}^3$ is a C^1 curve such that $\boldsymbol{\gamma}(t) \in P_t$ for all t. Here we do *not* assume that $\boldsymbol{\gamma}$ is parameterized by arc-length or that P_t is perpendicular to $\boldsymbol{\gamma}'(t)$. From Theorem 1 the signed volume swept out is

$$V = \int_a^b A(t)\Big(\boldsymbol{\gamma}'(t) + \big(\mathbf{c}'(t) - \boldsymbol{\gamma}'(t)\big)\Big) \cdot \mathbf{n}(t)\,dt.$$

Since $\mathbf{c}(t)$ and $\boldsymbol{\gamma}(t)$ are both in P_t, then $\mathbf{c}(t) - \boldsymbol{\gamma}(t)$ is perpendicular to $\mathbf{n}(t)$ and we have

$$0 = \frac{d}{dt}\Big(\big(\mathbf{c}(t) - \boldsymbol{\gamma}(t)\big) \cdot \mathbf{n}(t)\Big) = \big(\mathbf{c}'(t) - \boldsymbol{\gamma}'(t)\big) \cdot \mathbf{n}(t) + \big(\mathbf{c}(t) - \boldsymbol{\gamma}(t)\big) \cdot \mathbf{n}'(t).$$

(This is similar to the first step in the proof of the proposition.) Thus

$$V = \int_a^b A(t)\Big(\gamma'(t)\cdot\mathbf{n}(t) + \big(\gamma(t)-\mathbf{c}(t)\big)\cdot\mathbf{n}'(t)\Big)\,dt. \tag{4}$$

Now assume the hypotheses of Theorem 2. Since P_s is perpendicular to $\gamma'(s)$, we may take $\mathbf{n}(s) = \gamma'(s) = \mathbf{T}(s)$, the unit tangent vector of γ. The second factor in the integrand of (4) becomes

$$\mathbf{T}(s)\cdot\mathbf{T}(s) + \big(\gamma(s)-\mathbf{c}(s)\big)\cdot\mathbf{T}'(s) = 1 - \big(\mathbf{c}(s)-\gamma(s)\big)\cdot\kappa(s)\mathbf{N}(s) = 1-\kappa(s)r(s),$$

which proves Theorem 2. ∎

The integrand of (4) has a nice interpretation. It can be written as

$$dV \quad = \quad A(t)\gamma'(t)\cdot\mathbf{n}(t)\,dt \quad + \quad A(t)\big(\gamma(t)-\mathbf{c}(t)\big)\cdot\mathbf{n}'(t)\,dt. \tag{5}$$

Suppose that the infinitesimal motion of S_t is purely translational. Then $\mathbf{n}'(t) = \mathbf{0}$ and $\gamma'(t) \neq \mathbf{0}$, and the infinitesimal volume swept out is given by the first term of (5). On the other hand, suppose that S_t is infinitesimally rotating about some line in P_t that passes through $\gamma(t)$. In this case $\mathbf{n}'(t) \neq \mathbf{0}$ and $\gamma'(t) = \mathbf{0}$, and the infinitesimal volume swept out is given by the second term of (5). Thus (5) is the decomposition of dV into purely translational and rotational parts from the perspective of the reference curve γ. The distinguishing property of the centroid is that it is the unique choice of $\gamma(t)$ for which only the translational part matters in (5).

Higher dimensions and other geometries Interested readers may enjoy generalizing Theorems 1 and 2 to a moving, codimension-one, flat region sweeping out volume in $\mathbb{R}^n$. The generalizations to spherical and hyperbolic geometries are less obvious, however. For a geodesic segment sweeping out area in S^2 or H^2, the analog of (2) is

$$A = \int_a^b \frac{C\big(L(t)/2\big)}{\pi}\mathbf{n}(t)\cdot\mathbf{v}(t)\,dt, \tag{6}$$

where L, $\mathbf{n}$, and $\mathbf{v}$ are the same as in (2) and $C(r)$ is the circumference of a circle of intrinsic radius r [**6**]. For the unit sphere, in which the Gaussian curvature is 1, we have $C(r) = 2\pi\sin r$, and for the hyperbolic plane with curvature -1 we have $C(r) = 2\pi\sinh r$, but (6) is valid for all constant curvatures. The analogs of (1) and (6) in S^n and H^n for $n \geq 3$, if they have been worked out, are necessarily more complicated because the notion of centroid (and more generally, center of mass) is less clear in these spaces. The definition and equivalence of the various formulations of centroid and center of mass in $\mathbb{R}^n$ depend on the affine structure of $\mathbb{R}^n$, which is absent in S^n and H^n. The first notion of center of mass for regions in symmetric spaces was developed by Cartan [**2**], and has been generalized to other settings (see Berger [**1**]; Galperin [**9**] gives an extrinsic definition of center of mass in S^n and H^n). The notion is more subtle than in $\mathbb{R}^n$ and, in particular, its dynamical properties do not extend (for example, the center of mass of a freely-moving rigid body does not necessarily follow a geodesic), and there are competing notions for a substitute concept in this setting [**10**].

REFERENCES

1. M. Berger, Riemannian Geometry During the Second Half of the Twentieth Century, *Univ. Lecture Series*, Vol. 17, American Mathematical Society, Providence, R.I., 2000.

2. E. Cartan, Groupes simples clos et ouverts et géométrie riemannienne, *J. Math. Pures Appl.* **8**, (1929), 1–33.
3. R. Courant, *Differential and Integral Calculus*, II. Originally published in 1934 as *Vorlesungen über Differential- und Integralrechnung*. Translated by E. J. McShane. Wiley Interscience, New York, 1988.
4. R. Courant and F. John, *Introduction to Calculus and Analysis*, II, Wiley Interscience, New York, 1974.
5. W. T. England and T. L. Miller, Volumes and cross-sectional areas, this MAGAZINE **74**:4 (2001), 288–295.
6. R. L. Foote, Planimeters and isoperimetric inequalities on constant curvature surfaces, preprint.
7. R. L. Foote, Geometry of the Prytz planimeter, *Reports on Mathematical Physics* **42**, no. 1, (1998), 249–271.
8. R. L. Foote, Planimeters, http://persweb.wabash.edu/facstaff/footer/ Planimeter/Planimeter.htm.
9. G. A. Galperin, A concept of the mass center of a system of material points in the constant curvature spaces, *Comm. Math. Phys.* **154**, (1993), 63–84.
10. A. V. Shchepetilov, Two-body problem on two-point homogeneous spaces, invariant differential operators and the mass center concept, *Journal of Geometry and Physics* **48**, (2003), 245–274.
11. M. Spivak, *Calculus on Manifolds*, W. A. Benjamin, New York, 1965.

Equal Sums of Three Fourth Powers or What Ramanujan Could Have Said

RICHARD BLECKSMITH
Northern Illinois University
DeKalb, IL 60115
richard@math.niu.edu

SIMCHA BRUDNO
318 S. Throop St. Apt 204
Chicago, IL 60607

Dedicated to our friend John Selfridge

Math trivia buffs recognize 1729 as the number of the taxi that Hardy took to visit Ramanujan in the hospital. When Hardy complained that it seemed like a typical run of the mill number, Ramanujan countered that 1729 is actually a "very interesting number," because "it is the smallest number expressible as the sum of two cubes in two different ways" ($9^3 + 10^3$ and $1^3 + 12^3$) [**5**, pp. 12]. Ramanujan could have really surprised his friend by saying something like this: "Besides, Hardy, twice its square is the first number which can be written as the sum of three fourth powers in four different ways."

It turns out that 2×1729^2 is the start of an infinite sequence of numbers with the following remarkable property: If $R(n)$ denotes the number of ways that n can be written as the sum of three fourth powers, then $R(n)$ doubles every time we move on to the next number in the list. Consequently, there exist integers with an arbitrarily large number of representations as sums of three fourth powers.

Our investigation utilizes a simple identity:

$$x^4 + y^4 + (x+y)^4 = 2(x^2 + xy + y^2)^2. \tag{1}$$

This equation is easily established by straightforward algebraic manipulation. Dickson referred to (1) as "Proth's identity." First published in 1878 [**3**, p. 657, footnote 227], it is an easy consequence of Candido's identity

$$(x^2 + y^2 + (x+y)^2)^2 = 2(x^4 + y^4 + (x+y)^4), \tag{2}$$

using the equation $x^2 + y^2 + (x + y)^2 = 2(x^2 + xy + y^2)$. A Proof Without Words for (2) appears in the April 2005 issue of the MAGAZINE [**9**].

C. B. Haldeman uses Proth's formula in a 1904 article on biquadrate numbers in *The Mathematical Magazine* [**4**]—not to be mistaken for the journal containing the article you are now reading. (See footnote 196 in Dickson [**3**, p. 650].) One of the first journals of the American mathematical community, *The Mathematical Magazine* was founded, edited, and apparently typeset by the self-taught farmer-turned-mathematician, Dr. Artemas Martin. Sporadically published from 1882 to 1904, it was devoted to "elementary mathematics" and, along with *The Ladies' Diary*, was an excellent source of challenging mathematical problems.

In one of those strange mathematical coincidences, Ramanujan in his third notebook wrote the same identical formula as Haldeman for representing a fourth power as a sum of five fourth powers. Berndt and Bhargava [**1**, p. 647] speculated on the small likelihood that Ramanujan saw Haldeman's article in an obscure journal from America, a continent and an ocean away from his hometown in India. In any case, we may certainly speculate that Ramanujan was well aware of formula (1) and could have easily made the connection between Hardy's taxicab number and sums of fourth powers.

The quadratic polynomial (or form) on the right hand side of (1) is so essential to our argument that we give it a name:

$$f(x, y) = x^2 + xy + y^2. \tag{3}$$

It is clear that every time we represent a given number n as $f(x, y)$, where x and y are integers, then we have a representation of $2n^2$ as a sum of three fourth powers: $x^4 + y^4 + (x + y)^4$. Analyzing $f(x, y)$ leads us into the beautiful world of algebraic integers and binary quadratic forms.

Quadratic forms have been extensively studied over the last three centuries by such prominent mathematicians as Fermat, Lagrange, Legendre, Gauss, and Minkowski. The discriminant of the general quadratic form $ax^2 + bxy + cy^2$ is $d = b^2 - 4ac$; so $f(x, y)$ has discriminant $d = -3$. Much is known about forms having a negative discriminant. For our purposes we need to establish three facts.

The first fact will settle the question: What primes can be written in the form $f(x, y)$? Completing the square transforms $f(x, y)$ into a form without an xy term:

$$x^2 + xy + y^2 = \left(x + \tfrac{1}{2}y\right)^2 + 3\left(\tfrac{1}{2}y\right)^2. \tag{4}$$

It follows from (4) that every integer of the form $f(x, y)$, where x and y are integers, can also be written as $a^2 + 3b^2$, where a and b are half-integers whose numerators are either both even or both odd. The set $\mathcal{O} = \{a + b\sqrt{-3} : a, b \in \frac{1}{2}\mathbb{Z} \text{ and } a - b \in \mathbb{Z}\}$ is the ring of algebraic integers in the field obtained by adjoining $\sqrt{-3}$ to the rationals. (The algebraic integers in the quadratic field $\mathbb{Q}(\sqrt{-3})$ are just those elements of the field that satisfy a *monic* polynomial with integer coefficients. The algebraic numbers $\mathbb{Q}(\sqrt{-3}) = \{r + s\sqrt{-3} : r, s \in \mathbb{Q}\}$ form the field of quotients of the ring of algebraic integers $\mathcal{O}$. For an easy-to-understand definition, see the chapter on algebraic integers in Pollard and Diamond [**10**, pp. 74–79].)

Once we have $\sqrt{-3}$ at our disposal, we are able to factor $a^2 + 3b^2$ as $(a + \sqrt{-3}b)(a - \sqrt{-3}b)$. Thus a prime p can be written in the form $a^2 + 3b^2$ if and only if p factors in the ring $\mathcal{O}$. It is well known that the ordinary primes that factor in this ring are the primes of the form $3k + 1$, together with the prime 3 itself, which factors as $-(\sqrt{-3})^2$. Hence, FACT 1. If p is a prime, then $f(x, y) = p$ has a solution in integers if and only if $p = 3$ or $p \equiv 1 \pmod 3$.

The next easily verified fact shows that if two numbers have representations of the form $f(x, y)$, then so does their product.

FACT 2.

$$f(u, v)f(x, y) = f(ux - vy, uy + vx + vy). \tag{5}$$

It follows from Facts 1 and 2 that a product of primes of the form $3k + 1$ has a representation of the form $f(x, y)$. Our third fact tells us how to count the number of proper representations of such a product if it is squarefree. (To say a representation $f(x, y)$ is proper means that x and y are relatively prime.) A discussion can be found in the chapter on Quadratic Forms in Davenport's interesting book, *The Higher Arithmetic* [**2**, pp. 146–147].

FACT 3. Given t distinct primes $p_1, p_2, \ldots, p_t$, all congruent to 1 mod 3, the number $R(n)$ of proper representations of $n = p_1 p_2 p_3 \cdots p_t$ by the form $f(x, y)$ is the number of automorphs (soon to be defined) of $f(x, y)$ times the number of solutions to the congruence

$$x^2 \equiv -3 \pmod{4n} \tag{6}$$

where x lies in the interval $0 \leq x < 2n$.

An automorph is a unimodular substitution that transforms a form into itself. A unimodular substitution is a linear change of variables $X = ax + by$, $Y = cx + dy$, where a, b, c, and d are integers satisfying $ad - bc = 1$. The form $f(x, y)$ of discriminant -3 has the following six automorphs:

i. $X = x,\ Y = y$ iv. $X = -x - y,\ Y = x$

ii. $X = -x,\ Y = -y$ v. $X = y,\ Y = -x - y$

iii. $X = x + y,\ Y = -x$ vi. $X = -y,\ Y = x + y$

Observe that automorph (iii) generates the other five; a second application of (iii) gives automorph (v), a third application results in (ii), and so on. Possessing six automorphs makes -3 the record holder among all negative discriminants. The discriminant $d = -4$ has four automorphs and all other negative discriminants have just the first two automorphs (i) and (ii) in the above list. It is also known from elementary number theory that the congruence (6) has exactly 2^t solutions in the prescribed interval.

As mentioned earlier, once we have a representation of $f(x, y) = n$, then the triple $(x, y, x + y)$ satisfies $x^4 + y^4 + (x + y)^4 = 2n^2$. These fourth power representations of $2n^2$ fall into groups of twelve: the six automorphs of $f(x, y)$ together with the six automorphs of $f(y, x)$. The corresponding 12 triples $(X, Y, X + Y)$ can be obtained from each other by simply permuting terms or changing signs, and hence they represent equivalent variations of (1). Note that among the 12 equivalent representations, we can always find one pair (x, y) such that $0 < x < y$. It follows that if n is a product of t distinct primes of the form $3k + 1$, then $2n^2$ has 2^{t-1} essentially different representations of the form $x^4 + y^4 + (x + y)^4$. Every time we add another prime factor, we double the number of representations. Since there are infinitely many primes of the form $3k + 1$, we have proved

THEOREM 1. *For every positive integer m, there is an integer n with at least m proper representations as the sum of three fourth powers.*

We can use (5) to generate all 2^{t-1} solutions. Express $p_1 = f(u, v)$ and $p_2 = f(x, y)$. To get two representations of $p_1 p_2$, apply (5) first with $f(u, v)f(x, y)$ and again with $f(u, v)f(y, x)$. Repeat this process to obtain four representations of $p_1 p_2 p_3$, etc.

The 1729 example Let $n = 7 \cdot 13 \cdot 19 = 1729$. By inspection, the primes 7, 13, and 19 have the following representations: $7 = f(1, 2)$, $13 = f(1, 3)$, and $19 = f(2, 3)$. (We will always select the representation $f(x, y)$ where $0 < x < y$.) Using (5) with the representations of 7 and 13, we find

$$f(1, 2) \cdot f(1, 3) = f(-5, 11).$$

Applying automorph (iii) to $x = 11$, $y = -5$ gives $X = 5$, $Y = 6$. To get a second representation of 91, use $f(3, 1)$ instead of $f(1, 3)$ in (5):

$$f(1, 2) \cdot f(3, 1) = f(1, 9).$$

The pair $(x, y) = (1, 9)$ is already in the desired form $0 < x < y$. Next we multiply the two representations $91 = f(5, 6)$ and $91 = f(1, 9)$ by $f(2, 3)$ and $f(3, 2)$, respectively, and apply the appropriate automorphs to generate the four solutions to $f(x, y) = 7 \cdot 13 \cdot 19$ listed in the table below.

#	x	y	$f(x, y)$
1.	8	37	1729
2.	3	40	1729
3.	23	25	1729
4.	15	32	1729

Hence, $2 \cdot 1729^2$ can be written as a sum of three fourth powers in four ways:

$$\begin{aligned} 2 \cdot 1729^2 &= 8^4 + 37^4 + 45^4 = 3^4 + 40^4 + 43^4 \\ &= 23^4 + 25^4 + 48^4 = 15^4 + 32^4 + 47^4. \end{aligned}$$

REMARK 1. In general, if $n = p_1^{e_1} p_2^{e_2} \cdots p_t^{e_t}$, where the distinct primes p_i are all of the form $3k + 1$, then the number of different representations of n by $f(x, y)$ is

$$R(n) = \left\lceil \frac{(e_1 + 1)(e_2 + 1) \cdots (e_t + 1)}{2} \right\rceil, \tag{7}$$

where $\lceil x \rceil$ means the ceiling of x, that is, the least integer $\geq x$. (Allowing 3 as a prime divisor does not change this count, that is, $R(3^e n) = R(n)$, the number of distinct representations of n.) The number of *proper* representations of n, however, still remains 2^{t-1}. Hardy and Wright [**6**, p. 330] utilize this idea by considering powers of the prime 7 in their proof of a version of Theorem 1. The method we present is a generalization, and at the same time, a simplification of their argument.

REMARK 2. In the mid-1960s Lander and Parkin [7] implemented a computer search for identities of the form $a^4 + b^4 = c^4 + d^4$ and $a^4 + b^4 + c^4 = d^4 + e^4 + f^4$. Their search was extended to other powers a year later with the collaboration of Selfridge [**8**]. The first example they found of a number expressible as a sum of two fourth powers in two different ways,

$$133^4 + 134^4 = 59^4 + 158^4,$$

answers Hardy's query when he asked Ramanujan whether he knew "the solution of the corresponding problem for fourth powers" as the 1729 solution for cubes. Ramanujan's reply was "that he knew no obvious example" [**5**, p. 12]. In their original paper [**7**, p. 451], Lander and Parkin found that the first number to have three representations as the sum of three fourth powers is

$$811538 = 29^4 + 17^4 + 12^4 = 28^4 + 21^4 + 7^4 = 27^4 + 23^4 + 4^4.$$

Observe that $811538 = 2(7^2 \cdot 13)^2$ and that with $e_1 = 2$, $e_2 = 1$, the formula (7) predicts a total of 3 essentially different representations of $7^2 \cdot 13$. These three representations are easily obtained by applying (5) to the representations of 7 and 13, as we did in the 1729 example.

REMARK 3. Representations $x^4 + y^4 + z^4$ need not, of course, satisfy $z = x + y$. For example, $3750578 = 2 \cdot 13 \cdot 144253$ can be written as the sum of three fourth powers in three different ways: $1^4 + 24^4 + 43^4$, $3^4 + 7^4 + 44^4$, and $6^4 + 31^4 + 41^4$, and in each expression the sum of the first two numbers does not equal the third. It is the only such example smaller than 2×1729^2.

REMARK 4. Our fundamental equation, (1), gives us infinitely many solutions to the diophantine equation $x^4 + y^4 + z^4 = r^{2m} + s^{2m}$. Just take $n = (p_1 p_2 \cdots p_t)^m$, where each p_i is congruent to 1 mod 3, and let $r = s = p_1 \cdots p_t$.

The calculations discussed in this article were performed by a computer. The extensive calculations of Ramanujan that stoked his prodigious imagination, leading to many volumes of formulas and theorems, were all, of course, done by hand. It is interesting to speculate whether a 21st century computer, together with a symbolic mathematical software system, would have enhanced (or hindered) the discoveries of this most remarkable mathematician.

Simcha Brudno, a Holocaust survivor, philosopher, and mathematician with exceptional intuition, passed away on June 9, 2006, at the age of 82.

REFERENCES

1. B. Berndt and S. Bhargava, Ramanujan for lowbrows, *Amer. Math. Monthly*, **100** (1993), 644–656.
2. H. Davenport, *The Higher Arithmetic*, sixth ed., Cambridge University Press, 1952, Chapter VI.
3. L. E. Dickson, *History of the Theory of Numbers (Volume 2)*, sixth ed., Carnegie Institute of Washington, 1929, p. 657.
4. C. B. Haldeman, On biquadrate numbers, *The Mathematical Magazine*, **2** (1904), 288–296.
5. G. H. Hardy, *Ramanujan: Twelve Lectures suggested by his Life and Work*, third ed., Chelsea, New York, 1999.
6. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, fifth ed., Clarendon Press, Oxford, 1960.
7. L. Lander and T. R. Parkin, Equal sums of biquadrates, *Math. Comp.*, **20** (1966), 450–451.
8. L. Lander, T. R. Parkin, and J. L. Selfridge, A survey of equal sums of like powers *Math. Comp.*, **21** (1967), 446–459.
9. R. B. Nelson, Proof Without Words: Candido's Identity, this MAGAZINE, **78** (2005), 131.
10. H. Pollard and H. Diamond, *The Theory of Alebraic Numbers*, second ed., The Mathematical Association of America, 1975, reprinted by Dover, 1999.

Another Look at an Amazing Identity of Ramanujan

JUNG HUN HAN
Department of Studies in Mathematics, University of Mysore
Manasagangotri, Mysore 570006, India
jhan176@yahoo.com

MICHAEL D. HIRSCHHORN
School of Mathematics
University of New South Wales
Sydney 2052, Australia
m.hirschhorn@unsw.edu.au

Ramanujan [5, p. 341] makes the amazing claim that if integers a_n, b_n, c_n are defined by

$$\begin{aligned}\sum_{n\geq 0} a_n x^n &= \frac{1+53x+9x^2}{1-82x-82x^2+x^3},\\ \sum_{n\geq 0} b_n x^n &= \frac{2-26x-12x^2}{1-82x-82x^2+x^3},\\ \sum_{n\geq 0} c_n x^n &= \frac{2+8x-10x^2}{1-82x-82x^2+x^3},\end{aligned}\tag{1}$$

then they satisfy

$$a_n^3+b_n^3=c_n^3+(-1)^n.\tag{2}$$

Two proofs of this claim and a plausible explanation of how Ramanujan may have been led to it have been given by Hirschhorn [1, 2].

In this note, we give a new and complete explanation in the light of an observation of Maurice Craig and are led to an alternative formulation. Indeed, we shall show that the sequences $\{a_n\}$, $\{b_n\}$, and $\{c_n\}$ are given by

$$\begin{pmatrix} a_n\\ b_n\\ c_n\end{pmatrix}=\begin{pmatrix} 63 & 104 & -68\\ 64 & 104 & -67\\ 80 & 131 & -85\end{pmatrix}^n\begin{pmatrix}1\\2\\2\end{pmatrix}.\tag{3}$$

We begin with Ramanujan's identity [3, p. 286; 4, p. 326],

$$\begin{aligned}&(3a^2+5ab-5b^2)^3+(4a^2-4ab+6b^2)^3+(5a^2-5ab-3b^2)^3\\ &\quad=(6a^2-4ab+4b^2)^3.\end{aligned}\tag{4}$$

As observed by Maurice Craig (in a personal communication), if in (4) we put $a=u+v$, $b=u-2v$, divide by 27, and transpose one term, we obtain

$$\begin{aligned}&(u^2+7uv-9v^2)^3+(2u^2-4uv+12v^2)^3\\ &\quad=(2u^2+10v^2)^3+(u^2-9uv-v^2)^3.\end{aligned}\tag{5}$$

(Previously, Hirschhorn supposed that (5) was Ramanujan's starting point, even though he could not find it in Ramanujan's work.)

Suppose we now, as Ramanujan likely did, recall that the sequence $\{h_n\}$ defined by

$$h_0 = 0, \quad h_1 = 1, \quad h_{n+2} = 9h_{n+1} + h_n \tag{6}$$

has the property (easily proved by induction)

$$h_{n+1}^2 - 9h_{n+1}h_n - h_n^2 = h_{n+1}^2 - h_n h_{n+2} = (-1)^n. \tag{7}$$

Set $u = h_{n+1}$, $v = h_n$ in (5), and define a_n, b_n, c_n by

$$\begin{aligned} a_n &= u^2 + 7uv - 9v^2 = h_{n+1}^2 + 7h_{n+1}h_n - 9h_n^2, \\ b_n &= 2u^2 - 4uv + 12v^2 = 2h_{n+1}^2 - 4h_{n+1}h_n + 12h_n^2, \\ c_n &= 2u^2 + 10v^2 = 2h_{n+1}^2 + 10h_n^2. \end{aligned} \tag{8}$$

Then (5), (7), and (8) give

$$a_n^3 + b_n^3 = c_n^3 + (-1)^n.$$

Our next step is to show that

$$\begin{pmatrix} a_{n+1} \\ b_{n+1} \\ c_{n+1} \end{pmatrix} = \begin{pmatrix} 63 & 104 & -68 \\ 64 & 104 & -67 \\ 80 & 131 & -85 \end{pmatrix} \begin{pmatrix} a_n \\ b_n \\ c_n \end{pmatrix},$$

from which (3) follows.

We have

$$\begin{aligned} a_{n+1} &= h_{n+2}^2 + 7h_{n+2}h_{n+1} - 9h_{n+1}^2 \\ &= (9h_{n+1} + h_n)^2 + 7(9h_{n+1} + h_n)h_{n+1} - 9h_{n+1}^2 \\ &= 135h_{n+1}^2 + 25h_{n+1}h_n + h_n^2 \\ &= 63(h_{n+1}^2 + 7h_{n+1}h_n - 9h_n^2) + 104(2h_{n+1}^2 - 4h_{n+1}h_n + 12h_n^2) \\ &\quad - 68(2h_{n+1}^2 + 10h_n^2) \\ &= 63a_n + 104b_n - 68c_n. \end{aligned}$$

The remaining two relations may be proved in similar fashion.

We now show that the a_n, b_n, c_n satisfy (1).

Set

$$A = \sum_{n\geq 0} a_n x^n, \quad B = \sum_{n\geq 0} b_n x^n, \quad C = \sum_{n\geq 0} c_n x^n.$$

It follows from the above relations that

$$\begin{aligned} A &= 1 + x(63A + 104B - 68C), \\ B &= 2 + x(64A + 104B - 67C), \\ C &= 2 + x(80A + 131B - 85C). \end{aligned} \tag{9}$$

If we solve these equations for A, B, and C, we obtain (1).

Note that if in (3) we allow n to be negative, it is still true that

$$a_n^3 + b_n^3 = c_n^3 + (-1)^n,$$

and these sequences are also given by Ramanujan [**5**, p.341].

REFERENCES

1. M.D. Hirschhorn, An amazing identity of Ramanujan, this MAGAZINE **68** (1995), 199–201.
2. M.D. Hirschhorn, A proof in the spirit of Zeilberger of an amazing identity of Ramanujan, this MAGAZINE **69** (1996), 267–269.
3. S. Ramanujan, *Notebooks*, Vol II, Bombay, Tata Institute of Fundamental Research, 1957.
4. S. Ramanujan, *Collected Papers*, AMS, Chelsea, 2000.
5. S. Ramanujan, *The Lost Notebook and Other Unpublished Papers*, Narosa, New Delhi, 1988.

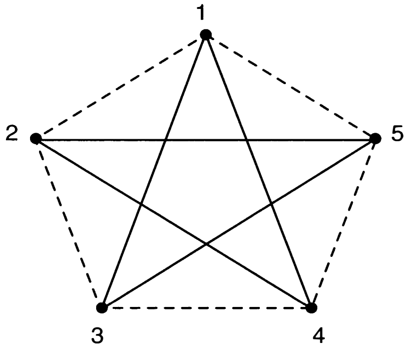
Ramsey's Theorem Is Sharp

SOLOMON W. GOLOMB
Department of Electrical Engineering Systems
University of Southern California
Los Angeles, CA 90089

Ramsey's Theorem [**1**] asserts that if the $\binom{6}{2} = 15$ edges of K_6, the complete graph on six points, are colored using two colors, there will be a *triangle* (a K_3 subgraph of K_6) with all three of its edges having the same color. This is sometimes called "the party problem," because if you select any six people at a party, it is guaranteed that either three of them will all know each other, or there will be three of them no two of whom know each other. To see the equivalence, represent each of the six people by a point, connect two points with a red line if the two people represented know each other, but by a blue line if they don't. Then by Ramsey's Theorem there will either be a solid red triangle (three mutual acquaintances) or a solid blue triangle (three mutual strangers).

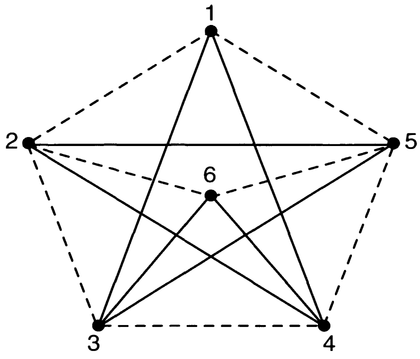
The purpose of this note is to present a visually striking proof that if any one of the 15 edges of K_6 is removed, the resulting graph (with 14 edges connecting the six points) *can* have all its edges colored, using two colors, without creating a solid-color triangle.

We first exhibit a two-coloring of the $\binom{5}{2} = 10$ edges of K_5 that creates no solid-color triangle:



(Our two colors are shown as solid and dotted lines. In this figure, there is a solid pentagon and a dotted pentagon, and clearly no single-color triangle.)

We now adjoin a *sixth* point in the *middle* of the K_5-figure just pictured, connecting it to the previous five points with two solid lines and two dotted lines.



Voila!

In addition to pentagons (5-cycles), there are now also quadrilaterals (4-cycles) in each of the two colors, but still no triangles (3-cycles). The only edge which is missing from the *complete* graph, K_6, on these six points is the edge from 1 to 6.

In terms of the party problem, if only one of the 15 pairs (among 6 people) refuses to acknowledge whether or not they are acquainted, we can no longer promise to exhibit a trio of mutual acquaintances or mutual strangers.

One generalization of Ramsey's original problem is: What is the smallest positive integer $R = R_c$ such that, if the complete graph K_R on R points have all $\binom{R}{2}$ of its edges colored in c colors, a solid color triangle is guaranteed to exist. The exact value of R_c is known for only a few small values of c, such as $R_2 = 6$ (the original Ramsey Theorem) and $R_3 = 17$.

For a context for $R_3 = 17$, suppose that in a certain high school class, each pair of students are either mutual friends, mutual enemies, or mutually indifferent. (While these relationships are symmetric, they are not assumed to carry over to third parties. The friend of a friend can be an enemy; the enemy of an enemy need not be a friend.) Then in any collection of 17 students there is certain to be a trio of either mutual friends, mutual enemies, or people who are mutually indifferent.

The following remarkable generalization of our triangle-free 2-coloring when one edge is removed from K_6 was observed by Herbert Taylor.

THEOREM. *For every $c \geq 2$, when a single edge is removed from the complete graph on R_c points, what remains can be c-colored without forming any solid-color triangles.*

Proof. We do not need to know the actual value of R_c to prove this theorem! By the definition of R_c, the complete graph on $R_c - 1$ points *can* have all its edges colored using c colors in such a way that no solid-color triangle is formed. Start with this coloring of the complete graph on $R_c - 1$ points. Designate any one of these points as P, and introduce a new point P' (the *clone* of P). Connect P', with edges, to each of the original points *except* P, and color the edge from P' to Q with the same color as the edge from P to Q, for every point Q except P and P'. If a solid-color triangle were formed, it would already have existed in the previous graph with P instead of P'.

Since there is no edge between P and P', there can be no triangle using both P and P'; so our new graph with R_c points lacks only the edge from P to P' to be the complete graph on R_c points, and it is edge-colored in c colors with no solid-color triangles. ■

Note that our illustration of a two-coloring of K_6 with one edge missing, having no solid-color triangles, is a special case of this general result, where the new point, 6, is the clone of 1.

Since K_{17} has $\binom{17}{2} = 136$ edges, we can 3-color 135 of these without forming a solid-color triangle!

We can also clone more than one point. For example, it is sufficient to remove only five of the 45 edges of K_{10} so that the remaining 40 edges can be 2-colored without forming a solid-color triangle. To achieve this, start with the triangle-free 2-coloring of the edges of K_5. We then clone each of the five original points of K_5, sequentially, adjoining one at a time, following the procedure in the proof of the Theorem. When we are done, the only edges missing from K_{10} are the five that connect each of the original points to their clones.

If we try to use this procedure when adjoining more than one clone to the same original point, all edges connecting the points in the same clone set must be omitted. For example, we can 2-color all but 15 of the 105 edges of K_{15} without forming a solid-color triangle, by adjoining two clones to each of the five original points of K_5.

When we use only one color, the Ramsey number R_1 is 3. (We can color the single edge of K_2, but not all three edges of K_3, using only one color, without forming a solid-color triangle.) The reader is encouraged to experiment with adjoining clones to the two original points of K_2, using only one color, and avoiding triangles, as just described. (What results are the complete bipartite graphs connecting the two clone sets, and these graphs are all triangle-free.)

Graham, Rothschild, and Spencer [**2**] give an extensive treatment of generalizations of Ramsey's Theorem.

REFERENCES

1. F. P. Ramsey, On a problem of formal logic, *Proc. London Math. Soc.* (2) **30** (1930), 264–286.
2. R.L. Graham, B. Rothschild, and J. Spencer, *Ramsey Theory*, Wiley, New York, 1980.

Where the Camera Was, Take Two

ANNALISA CRANNELL
Franklin & Marshall College
Lancaster, PA 17604
annalisa.crannell@fandm.edu

In the very nice article "Where the camera was" [**1**], Byers and Henle approximated the position of a photographer from geometric clues in an old photograph of John M. Greene Hall at Smith College. Here, we give an approach to the problem that is slightly more geometric.

We will make one simplifying assumption that the original article did not make: that the photo was not cropped, meaning that the center of the photograph was the center of the photographer's aim. Using the diagonals of the rectangle (see FIGURE 1), we can determine where to aim our own camera to best recreate the original photograph.

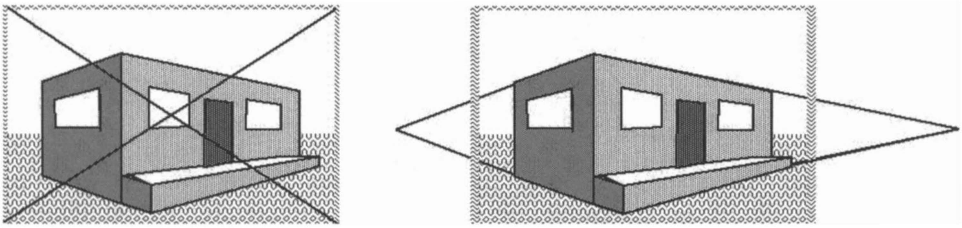


Figure 1 Side-by-side pictures showing the center of the photo and the two vanishing points.

The "x" in Byers and Henle's photograph of Greene Hall appears *above* the doors in the building; this hints that the camera was pointed slightly upward. Nonetheless, we will assume (as did Byers and Henle) that the camera was aimed horizontally, that is, parallel to the ground; the error introduced by this assumption is not very large.

As Byers and Henle noted, the images of lines that are parallel to each other but not parallel to the image plane meet at a single point in the image plane. By looking at the image, we can measure the distance from the center (X) to the left and to the right vanishing points; let us call these distances l and r respectively. If we imagine that the picture plane is correctly positioned in space, so that the image lines up with reality for the viewer, then these points, called the *vanishing points*, occur at the places on the image plane where the viewer's line of sight is parallel to the original lines (see for example Theorem 2 of [**2**]). We illustrate this phenomenon in FIGURE 2.

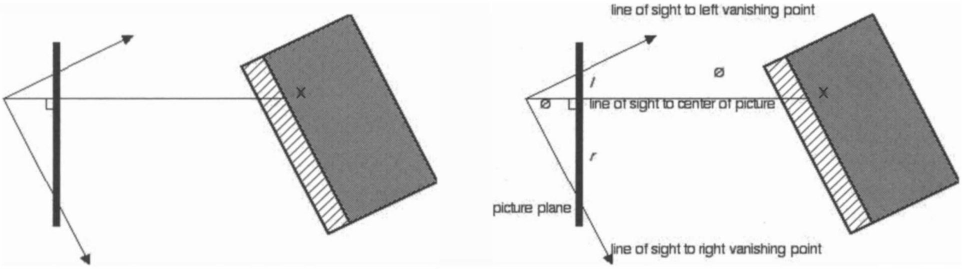

X
line of sight to left vanishing point
ø
l
ø
line of sight to center of picture
X
r
picture plane
line of sight to right vanishing point

Figure 2 The lines of sight to the vanishing points are parallel to the edges of the building.

From FIGURE 2, we can tell that the photographer, pointing at the "X", stands somewhere on a line that makes an angle θ with the front of the building, where $\tan(\theta) = \sqrt{r/l}$. (Note that the altitude of the triangle in FIGURE 2 is $\sqrt{rl}$). One remarkable aspect of this calculation is that it does not depend on the dimensions of the building, but rather on the reasonable assumption that the visible corner of the building is a right angle.

How far back along this line does the photographer stand? At this point, we have to put on our coats and shoes and go outdoors to measure something on the actual building. Here we can, as Byers and Henle did, use similar triangles.

Because we assume that the camera is aimed horizontally, the picture plane is vertical, and so the image of any vertical line is parallel to the original line. Therefore, we can use vertical line segments to construct similar triangles. If the height of a window on the building is W and the height of its image is w, then we get $D/d = W/w$, where d is the distance of the photographer from the picture plane, and D—the quantity we want—is the distance from the photographer to the spot marked "X". Conveniently,

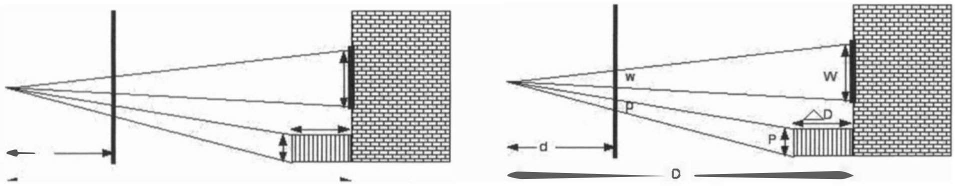


Figure 3 Using similar triangles with one vertex at the camera and opposite edge at the building. Note that *d*, *D*, and ΔD are measured horizontally in the direction of the photographer, not perpendicularly to the building.

my "photograph" has the window aligned vertically with the "X"; it also has a porch, which gives us a second set of similar triangles with a second equation:

$$\frac{D - \Delta D}{d} = \frac{P}{p}.$$

In each of these two formulas, the uppercase letters are measured in the real world; the lowercase letters are measured from the photograph. Solving these two equations yields

$$D = \frac{\Delta D p W}{pW - Pw}.$$

That is, we can determine the distance of the photographer from the "X" by taking three measurements on the building and two more measurements on the picture itself. As Byers and Henle remarked, such a computation is very sensitive to small changes in the measurement of w, the central vertical measurement in the photo.

REFERENCES

1. Katherine Byers and James Henle, Where the Camera Was, this MAGAZINE **77**:4 (2004), 251–259.
2. Marc Frantz, Lesson 3: Vanishing Points and Looking at Art, from *Lessons in Mathematics and Art*, available from `http://mypage.iu.edu/~mathart/viewpoints/lessons/`
3. A. C. Robin, Photomeasurement, *Math. Gaz.*, **62** (1978) 77–85.

C.E. Tripp, Where is the camera? The use of a theorem in projective geometry to find from a photograph the location of a camera, *Math. Gaz.*, **71** (1987), 8–14.

Earlier Citation for Irrational Square Roots

Peter Ungar writes to report that David M. Bradley of the University of Maine in Orono has called to his attention pp. 4–5 of *Conjecture and Proof*, by Miklós Laczkovich, (2001) Mathematical Association of America, where the proof presented in Ungar's note "Irrationality of Square Roots," this MAGAZINE, **79** (2006), 147–148, can already be found.

PROBLEMS

Proposals

To be considered for publication, solutions should be received by March 1, 2007.

1751. *Proposed by Iliya Bluskov, University of Northern British Columbia, Prince George, BC, Canada*

Let $k_1, k_2, \dots, k_n$ be integers with $k_i > 2, i = 1, 2, \dots, n$, and let $N = \sum_{i=1}^{n} \binom{k_i}{2}$. Prove that

$$\sum_{1 \le i < j \le n} \binom{k_i}{2}\binom{k_j}{2} + 3\sum_{i=1}^{n} \binom{k_i + 1}{4} = \binom{N}{2}.$$

1752. *Proposed by John Sternitzky and Robert Calcaterra, University of Wisconsin Platteville, Platteville, WI.*

Let $\mathbb{R}$ be the real line with the standard topology. Prove that every uncountable subset of $\mathbb{R}$ has uncountably many limit points.

1753. *Proposed by John C. George, Eastern New Mexico University, Portales, NM.*

Let n be a positive integer and let S_n be the set of all positive integers whose (base ten) digit sum is n. Determine the convergence or divergence of the series

$$\sum_{k \in S_n} \frac{1}{k}.$$

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 reader's name, full address, and an e-mail address and/or FAX number.

1754. *Proposed by Mihály Bencze, Săcele-Négyfalu, Romania.*

Let $a_1, a_2, \ldots, a_n$ be positive real numbers. Prove that

$$\sqrt[n]{\prod_{k=1}^{n} a_k} \leq \ln\left(1 + \sqrt[n]{\prod_{k=1}^{n}(e^{a_k} - 1)}\right) \leq \frac{1}{n}\sum_{k=1}^{n} a_k.$$

1755. *Proposed by Michel Bataille, Rouen, France.*

Let $a, b, c > 0$ with $b > c$. Prove that, as $n \to \infty$,

$$\frac{(a+b)^{a+b}(2a+b)^{2a+b}\cdots(na+b)^{na+b}}{(a+c)^{a+c}(2a+c)^{2a+c}\cdots(na+c)^{na+c}} \sim \lambda(na)^{n\alpha+\beta},$$

for some positive real numbers λ, α, and β.

Quickies

Answers to the Quickies are on page 316.

Q963. *Proposed by H. A. ShahAli, Tehran, Iran.*

For positive integer $k \geq 3$, determine the range of

$$\frac{\alpha_1}{\alpha_1 + \alpha_2} + \frac{\alpha_2}{\alpha_2 + \alpha_3} + \cdots + \frac{\alpha_k}{\alpha_k + \alpha_1},$$

as $\alpha_1, \alpha_2, \ldots, \alpha_k$ take on positive real values.

Q964. *Proposed by Jung-Jin Lee, University of Illinois, Champaign-Urbana, IL.*

Let $\mathbb{N}$ be the set of natural numbers and let $S_{\mathbb{N}}$ be the set of all permutations on $\mathbb{N}$. Prove that the cardinality of $S_{\mathbb{N}}$ is same as the cardinality of the real numbers.

Solutions

Condition for a congruence October 2005

1726. *Proposed by Jerry Metzger, University of North Dakota, Grand Forks, ND.*

Let a and j be positive integers with $a \geq 2$. Show that there is a positive integer n such that $a^n \equiv -j \pmod{a^j + 1}$ if and only if $j = a^k$ for some $k \geq 0$.

Solution by Northwestern University Math Problem Solving Group, Northwestern University, Evanston, IL.

We first assume that $j = a^k$ for some $k \geq 0$. Note that $a^j \equiv -1 \pmod{a^j + 1}$. Thus, if $j = a^k$, then $a^{j+k} \equiv -j \pmod{a^j + 1}$, so for $n = j + k = a^k + k$ we have the desired congruence.

Next assume that there is a positive integer n for which $a^n \equiv -j \pmod{a^j + 1}$. Write $n = qj + k$ where q and k are nonnegative integers with $0 \leq k < j$. We then have

$$a^n = a^{qj+k} \equiv (-1)^q a^k \equiv \pm a^k \pmod{a^j + 1}.$$

Consequently, $j \equiv \mp a^k \pmod{a^j + 1}$.

If $j \equiv a^k \pmod{a^j+1}$, then because $1 \le j$, $a^k \le a^j$, we must have $j = a^k$.

If $j \equiv -a^k \pmod{a^j+1}$, then $j = a^j - a^k + 1$. However

$$a^j = a \cdot a^{j-1} \ge 2a^{j-1} \ge a^k + a^{j-1} > a^k + j - 1,$$

so $j < a^j - a^k + 1$, that is, we cannot have equality.

Thus, the only possibility is $j = a^k$.

Also solved by Michel Bataille (France), Robert Calcaterra, John Christopher, Chip Curtis, Hugh M. Edgar (Canada), Fejéntaláltuka Szeged Problem Solving Group (Hungary), Dennis Gressis, Enkel Hysnelaj (Australia), Joel Iiams, Tom Leong, Byron Schmuland (Canada), Albert Stadler (Switzerland), Marian Tetiva (Romania), Paul Weisenhorn (Germany), Yan-loi Wong (Singapore), and the proposer.

Chain addition October 2005

1727. *Proposed by Jody M. Lockhart and William P. Wardlaw, U. S. Naval Academy, Annapolis, MD.*

Chain addition is a technique used in cryptography to extend a short sequence of digits, called the seed, to a longer sequence of pseudorandom digits. If the seed sequence of digits is $a_1, a_2, \ldots, a_n$, then for positive integer k, $a_{n+k} = a_k + a_{k+1} \pmod{10}$, that is, a_{n+k} is the units digit in the sum $a_k + a_{k+1}$. Suppose that the seed sequence is 3, 9, 6, 4. Prove that the sequence is periodic and find, without the use of calculator or computer, the number of digits in the sequence before the first repetition of 3, 9, 6, 4.

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

We prove that the sequence has period 1560. Let $\mathbf{X}_1$ be a fixed column vector in $\mathbb{Z}^4$. If for positive integer k we have $\mathbf{X}_k = (a, b, c, d)^t$, then define $\mathbf{X}_{k+1} = (b, c, d, a+b)^t$. Now consider the 4×4 matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Then $\mathbf{X}_{k+1} = A\mathbf{X}_k$ for all positive integer k. Note that the characteristic polynomial of A is $p(x) = x^4 - x - 1$.

Now view p as an element of the ring $\mathbb{Z}_5[x]$. Because p has no zeros in $\mathbb{Z}_5$, it has no linear factors in $\mathbb{Z}_5[x]$. If

$$p(x) = (x^2 + ax + b)(x^2 + cx + d),$$

then $ac + b + d = 0$, $c = -a$, and $ad + bc = -1$. Thus $a(b - d) = 1$, which implies that $a \ne 0$ and $b \ne d$. Because $bd = 4$, we must have $\{b, d\} = \{1, 4\}$. It follows that $b + d = 0$ and then that $-a^2 = 0$, which is impossible. Therefore, p is irreducible in $\mathbb{Z}_5[x]$. Now let F be the field extension $\mathbb{Z}_5[\alpha]$ where α is a zero of p. Then $1, \alpha, \alpha^2, \alpha^3$ is a basis for F over $\mathbb{Z}_5$.

Because $\alpha^4 = \alpha + 1$ and the function $x \to x^5$ is an automorphism of F, it follows that

$$\alpha^{20} = (\alpha + 1)^5 = \alpha^5 + 1 = \alpha^2 + \alpha + 1$$

$$\alpha^{100} = \alpha^{10} + \alpha^5 + 1 = (\alpha^2 + \alpha)^2 + (\alpha^2 + \alpha) + 1 = 2(\alpha^3 + \alpha^2 + \alpha + 1)$$

$$\alpha^{22} = (\alpha^2 + \alpha + 1)\alpha^2 = \alpha^3 + \alpha^2 + \alpha + 1.$$

Hence $\alpha^{78} = 2$ and $\alpha^{312} = 1$. Moreover,

$$\alpha^{24} = (\alpha^2 + \alpha + 1)(\alpha + 1) \neq 1$$

$$\alpha^{104} = 2(\alpha^3 + \alpha^2 + \alpha + 1)(\alpha + 1) \neq 1$$

Therefore $o(\alpha^{78}) = 4$, $o(\alpha^{24}) = 13$, and $o(\alpha^{104}) = 3$, where $o(\omega)$ is the order of ω in the multiplicative group of F. Thus $o(\alpha)$ is a multiple of $3 \cdot 4 \cdot 13 = 156$ and a divisor of 312. Because $\alpha^{156} = 4$, we conclude $o(\alpha) = 312$. If we view the entries of A as elements of $\mathbb{Z}_5$, then α, α^5, α^{25}, and α^{125} are the distinct eigenvalues of A and A is diagonalizable. Because the matrix equation $A^k\mathbf{Y} = \mathbf{Y}$ will have a nonzero solution if and only if 1 is an eigenvalue of A^k, it follows that the sequence defined by $\mathbf{Y}_k = A^k\mathbf{Y}_1$ will have period 312 whenever $\mathbf{Y}_1$ is a nonzero vector in $(\mathbb{Z}_5)^4$.

Now view the coefficients of p and the entries of A as elements of $\mathbb{Z}_2$. Because x, $x + 1$, and $x^2 + x + 1$ are the only linear and irreducible quadratic polynomials over $\mathbb{Z}_2$ and none is a factor of $p(x)$, we conclude that $p(x)$ is irreducible in $Z_2[x]$. If β is a zero of p, then the field $Z_2[\beta]$ has 16 elements generated by the basis $\{1, \beta, \beta^2, \beta^3\}$ over $\mathbb{Z}_2$. Because the multiplicative group of this field has order 15 and $\beta^3 \neq 1$ and $\beta^5 = \beta^2 + \beta \neq 1$, it follows that $o(\beta) = 15$. Therefore, the line of reasoning used earlier shows that the sequence defined by $\mathbf{W}_k = A^k\mathbf{W}_1$ will have period 15 whenever $\mathbf{W}_1$ is a nonzero vector in $(\mathbb{Z}_2)^4$.

Combining the results, we conclude that the period of the given sequence is the least common multiple of 312 and 15, that is 1560.

Also solved by John Christopher, Chip Curtis, G.R.A.20 Problem Solving Group (Italy), Richard F. Mc Coart, Kim McInturff, Albert Stadler (Switzerland), Paul Weisenhorn (Germany), Doug Wilcox, and the proposers. There were two incorrect submissions.

A $3n$-gon inequality October 2005

1728. *Proposed by José Luis Díaz-Barrero, Universitat Politènica de Catalunya, Barcelona, Spain.*

Let $A_1A_2 \dots A_{3n}$ be a regular polygon with $3n$ sides, and let P be a point on the shorter arc A_1A_{3n} of its circumcircle. Prove that

$$\left(\sum_{k=1}^{n} PA_{n+k}\right)\sum_{k=1}^{n}\left(\frac{1}{PA_k} + \frac{1}{PA_{2n+k}}\right) \geq 4n^2.$$

Solution by Tom Leong, Brooklyn, NY.

For $k = 1, 2, \dots, n$, quadrilateral $PA_kA_{n+k}A_{2n+k}$ is cyclic, so by Ptolemy's theorem,

$$PA_{n+k} \cdot A_kA_{2n+k} = PA_k \cdot A_{n+k}A_{2n+k} + PA_{2n+k} \cdot A_kA_{n+k}. \qquad (1)$$

Furthermore, triangle $A_kA_{n+k}A_{2n+k}$ is equilateral, so (1) reduces to

$$PA_{n+k} = PA_k + PA_{2n+k}.$$

Summing over k and then applying the arithmetic-harmonic mean inequality gives

$$\sum_{k=1}^{n} PA_{n+k} = \sum_{k=1}^{n}(PA_k + PA_{2n+k}) \geq 4n^2\left(\sum_{k=1}^{n}\left(\frac{1}{PA_k} + \frac{1}{PA_{2n+k}}\right)\right)^{-1},$$

which is equivalent to the desired result.

Also solved by Michel Bataille (France), J. C. Binz (Switzerland), Brian Bradie, Robert Calcaterra, Enkel Hysnelaj (Australia), Northwestern University Math Problem Solving Group, Albert Stadler (Switzerland), Michel Vowe (Switzerland), Paul Weisenhorn (Germany), and the proposer.

Products of odd numbers October 2005

1729. *Proposed by Brian T. Gill, Seattle Pacific University, Seattle, WA.*

For positive integer k, let c_k denote the product of the first k odd positive integers, and let $c_0 = 1$. Prove that for each nonnegative integer n,

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

Solution by JPV Abad, San Francisco, CA.

Our proof uses the following four identities,

$$c_k = \frac{k!}{2^k}\binom{2k}{k}, \qquad \binom{k-\frac{1}{2}}{k} = \frac{1}{2^{2k}}\binom{2k}{k}, \qquad (-1)^k\binom{-\frac{1}{2}}{k} = \binom{k-\frac{1}{2}}{k},$$

and

$$\sum_k \binom{l}{m+k}\binom{s}{n+k} = \binom{l+s}{l-m+n}.$$

The first three are easily verified and the last follows from Vandermonde's identity. Applying these results in the order presented we find

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

where we applied the first two of the given identities in obtaining the last equality.

Also solved by Michel Bataille (France), J. C. Binz (Switzerland), Con Amore Problem Group (Denmark), Chip Curtis, Knut Dale (Norway), Charles Diminnie and Roger Zarnowski, Michael Goldenberg and Mark Kaplan, G.R.A.20 Problem Solving Group (Italy), Eugene A. Herman, Enkel Hysnelaj (Australia), Tom Leong, John Mangual, Daniel A. Morales (Venezuela), Rob Pratt, Edward Schmeichel, Nicholas C. Singer, Chris Smith, Albert Stadler (Switzerland), Marian Tetiva (Romania), Michael Vowe (Switzerland), Paul Weisenhorn (Germany), Chu Wenchang (Italy), and the proposer.

A bounded transformation October 2005

1730. *Proposed by Steven Butler, University of California San Diego, La Jolla, CA.*

Let A and B be symmetric, positive semi-definite matrices such that $A + B$ is positive definite, and let $\|\mathbf{y}\|$ denote the usual 2-norm of the vector $\mathbf{y}$. Prove that for all $\mathbf{x} \neq \mathbf{0}$,

$$\|(I-A)(I+A)^{-1}(I-B)(I+B)^{-1}\mathbf{x}\| < \|\mathbf{x}\|.$$

Solution by Eugene A. Herman, Grinnell College, Grinnell, IA.

Rather than assuming that $A + B$ is positive definite, we make the weaker assumption that $\mathcal{N}(A) \cap \mathcal{N}(B) = \{\mathbf{0}\}$, where $\mathcal{N}(C)$ denotes the null space of a matrix C.

LEMMA. *If A is symmetric, positive semi-definite, then for all* $\mathbf{x}$

$$\|(I - A)(I + A)^{-1}\mathbf{x}\| \le \|\mathbf{x}\| \tag{1}$$

with equality if and only if $\mathbf{x} \in \mathcal{N}(A)$, *in which case* $(I - A)(I + A)^{-1}\mathbf{x} = \mathbf{x}$.

Proof. The matrix A can be diagonalized and its eigenspaces are mutually orthogonal. Furthermore, the matrices $I - A$ and $(I + A)^{-1}$ have the same eigenspaces as A. Thus, if $\mathbf{x}$ is an eigenvector of A with eigenvalue λ, then

$$(I - A)(I + A)^{-1}\mathbf{x} = \frac{1-\lambda}{1+\lambda}\mathbf{x}. \tag{2}$$

In particular, if $\mathbf{x} \in \mathcal{N}(A)$ (the eigenspace associated with the eigenvalue 0), then $(I - A)(I + A)^{-1}\mathbf{x} = \mathbf{x}$ and equation (1) follows. The orthogonal complement $\mathcal{N}(A)^\perp$ is the sum of the eigenspaces associated with the nonzero eigenvalues of A. We construct an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k\}$ of the orthogonal complement by taking an orthonormal basis of each eigenspace whose associated eigenvalue is nonzero and then forming the union of these bases. Let $\{\lambda_k\}$ be the corresponding (positive) eigenvalues of A. Thus, if $\mathbf{x} \in \mathcal{N}(A)^\perp$, we may write $\mathbf{x} = \sum_{j=1}^k c_j\mathbf{e}_j$, and therefore by equation (2)

$$(I - A)(I + A)^{-1}\mathbf{x} = \sum_{j=1}^k c_j \frac{1-\lambda_j}{1+\lambda_j}\mathbf{e}_j. \tag{3}$$

If $\mathbf{x} \neq \mathbf{0}$ then, since $\left|\frac{1-\lambda_j}{1+\lambda_j}\right| < 1$ for $j = 1, \dots, k$, we conclude that

$$\|(I - A)(I + A)^{-1}\mathbf{x}\|^2 < \sum_{j=1}^k |c_j|^2 = \|\mathbf{x}\|^2,$$

and hence that $\|(I - A)(I + A)^{-1}\mathbf{x}\| < \|\mathbf{x}\|$. From equation (3), note also that $(I - A)(I + A)^{-1}\mathbf{x} \in \mathcal{N}(A)^\perp$. Finally, if $\mathbf{x}$ is any vector not in $\mathcal{N}(A)$, we write $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1 \in \mathcal{N}(A)$ and $\mathbf{x}_2 \in \mathcal{N}(A)^\perp$. Since $\mathbf{x}_2 \neq \mathbf{0}$, we obtain

$$\begin{aligned}\|(I - A)(I + A)^{-1}\mathbf{x}\|^2 &= \|\mathbf{x}_1 + (I - A)(I + A)^{-1}\mathbf{x}_2\|^2\\ &= \|\mathbf{x}_1\|^2 + \|(I - A)(I + A)^{-1}\mathbf{x}_2\|^2\\ &< \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 = \|\mathbf{x}\|^2.\end{aligned}$$

So $\|(I - A)(I + A)^{-1}\mathbf{x}\| < \|\mathbf{x}\|$.

We now use the lemma to prove the main result. If $\mathbf{x} \notin \mathcal{N}(B)$, then

$$\|(I - A)(I + A)^{-1}(I - B)(I + B)^{-1}\mathbf{x}\| \le \|(I - B)(I + B)^{-1}\mathbf{x}\| < \|\mathbf{x}\|.$$

On the other hand, if $\mathbf{x} \in \mathcal{N}(B)$ and $\mathbf{x} \notin \mathcal{N}(A)$, then

$$\|(I - A)(I + A)^{-1}(I - B)(I + B)^{-1}\mathbf{x}\| = \|(I - A)(I + A)^{-1}\mathbf{x}\| < \|\mathbf{x}\|.$$

Further extension. The above result extends easily as follows. Let $\{A_1, A_2, \ldots, A_k\}$ be symmetric, positive semi-definite matrices such that $\bigcap_{j=1}^{k} \mathcal{N}(A_j) = \{\mathbf{0}\}$. Then, for all $\mathbf{x} \neq \mathbf{0}$,

$$\|(I - A_1)(I + A_1)^{-1} \cdots (I - A_k)(I + A_k)^{-1}\mathbf{x}\| < \|\mathbf{x}\|.$$

Also solved by Michel Bataille (France), Brian Bradie, Robert Calcaterra, Gérard Letac (France), Chi-Kwong Li, and the proposer.

A combinatorial identity April 2005

1718. *Proposed by David Callan, Madison, WI.*

Let k, n be integers with $1 \le k \le n$. Prove the identity

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

Solution by I. *William Moser, McGill Univeristy, Montreal, Quebec, Canada.*

We prove the following more general identity: for integers $0 \le k \le n$, $1 \le m \le n$,

$$\sum_{i=0}^{m} \binom{m}{i}\binom{n-m}{k-i} 2^{m-i} = \sum_{j=0}^{m} \binom{m}{j}\binom{n-j}{k} \qquad (1)$$

The case $m = k - 1$ is the identity in the problem statement.

We prove (1) by counting, in two different ways, the cardinality of the set of words of length n using the alphabet $\{A, B, C\}$ and satisfying the condition: precisely k of the letters are A, and all of the letters B must be among the first m letters (reading from left).

First count: in subsets according to the number i of A's among the first m letters of the word. Construct these words as follows. Place m symbols X in a row and following them $n - m$ symbols Y.

$$\underbrace{\underline{X\,X \ldots X\,X}}_{m}\underbrace{\underline{Y\,Y \ldots Y\,Y}}_{n-m} \qquad (2)$$

Choose i of the m X's (this can be done in $\binom{m}{i}$ ways), replace each of these chosen X's by A and replace each of the other $m - i$ X's by B or C (this can be done in 2^{m-i} ways). Then choose $k - i$ of the $n - m$ Y's (in $\binom{n-m}{k-i}$ ways), replace each of these $k - i$ Y's by A, and replace the remaining Y's by C. We have constructed $\binom{m}{i}\binom{n-m}{k-i}2^{m-i}$ words satisfying the conditions. Summing over i we have the left sum in (1).

Second count: in subsets according to the number j of B's. Start with the display (2). Choose j of the X's (in $\binom{m}{j}$ ways), and replace them by B's. Choose k of the $n - j$ remaining symbols (in $\binom{n-j}{k}$ ways), replace them by A's. Any remaining X's or Y's are now replaced by C's. We have constructed $\binom{m}{j}\binom{n-j}{k}$ words satisfying the conditions. Summing over j we have the right sum in (1).

Solution by II. *G.R.A.20 Problem Solving Group, Rome, Italy.*

We consider the polynomial

$$P(x) = (2+x)^{k-1} \cdot (1+x)^{n-(k-1)}$$
$$= \sum_{i=0}^{k-1} \binom{k-1}{i} 2^{(k-1)-i} x^i \cdot \sum_{j=0}^{n-(k-1)} \binom{n-(k-1)}{j} x^j.$$

Then the coefficient of x^k of $P(x)$ is

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

which is the left side of the desired identity.

On the other hand

$$P(x) = (1+(1+x))^{k-1} \cdot \frac{(1+x)^n}{(1+x)^{k-1}} = \left(\frac{1}{1+x}+1\right)^{k-1} \cdot (1+x)^n$$
$$= \sum_{i=0}^{k-1} \binom{k-1}{i} \frac{1}{(1+x)^i} \cdot (1+x)^n = \sum_{i=0}^{k-1} \binom{k-1}{i} (1+x)^{n-i}$$
$$= \sum_{i=0}^{k-1} \binom{k-1}{i} \sum_{j=0}^{n-i} \binom{n-i}{j} x^j.$$

Therefore the coefficient of x^k of $P(x)$ is

$$\sum_{i=0}^{k-1} \binom{k-1}{i} \binom{n-i}{k},$$

which is the right side of the identity.

Also solved by JPV Abad, Tsehaye Andeberhan, Michel Bataille (France), J. C. Binz (Switzerland), Robert Calcaterra, Chip Curtis, Knut Dale (Norway), Daniele Donini (Italy), Tom Leong, Peter W. Lindstrom, Daniel A. Morales (Venezuela), José H. Nieto (Venezuela), Rob Pratt, Muneer Ahmad Rashid (Australia), Henry Ricardo, Edward Schmeichel, Chris Smith, Albert Stadler (Switzerland), Paul Weisenhorn (Germany), Chu Wenchang (Italy), Li Zhou, and the proposer.

Answers

Solutions to the Quickies from page 310.

A963. Let

$$S = \frac{\alpha_1}{\alpha_1+\alpha_2} + \frac{\alpha_2}{\alpha_2+\alpha_3} + \cdots + \frac{\alpha_k}{\alpha_k+\alpha_1}.$$

Because all α_j are positive and $k \geq 3$, we have

$$S > \frac{\alpha_1}{\alpha_1+\alpha_2+\cdots+\alpha_k} + \frac{\alpha_2}{\alpha_1+\alpha_2+\cdots+\alpha_k} + \cdots + \frac{\alpha_k}{\alpha_1+\alpha_2+\cdots+\alpha_k} = 1.$$

If $\alpha_1 = x,\ \alpha_2 = x^2, \ldots, \alpha_k = x^k$, then, with some simple computation,

$$S = \frac{k-1}{x+1} + \frac{x^{k-1}}{x^{k-1}+1}.$$

As $x \to \infty$, this expression approaches 1. Thus the infimum of the set of values of S is 1.

Next set

$$S' = \frac{\alpha_k}{\alpha_k + \alpha_{k-1}} + \frac{\alpha_{k-1}}{\alpha_{k-1} + \alpha_{k-2}} + \cdots + \frac{\alpha_1}{\alpha_1 + \alpha_k}.$$

Then $S + S' = k$, and by the the same argument as above, the infimum of S' is 1. Thus $S < k - 1$, and the supremum of S is $k - 1$. Because S is continuous in the α_j for $\alpha_j > 0$, it follows that the range of S is $(1,\ k - 1)$.

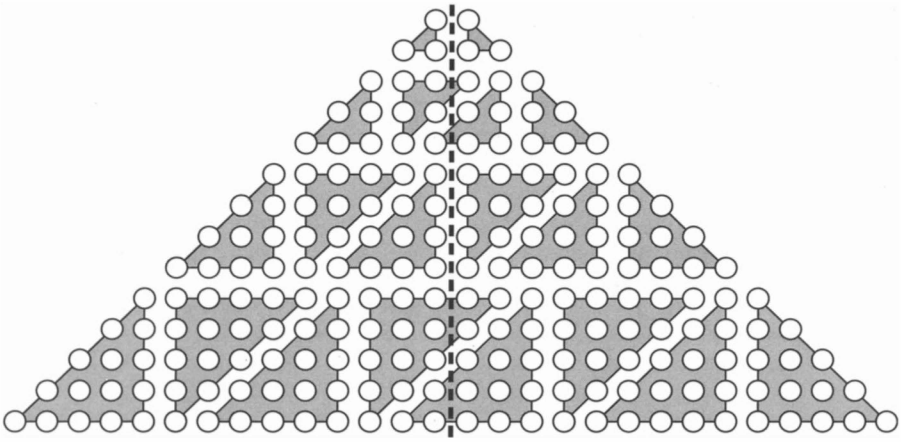
A964. Because $S_{\mathbb{N}} \subseteq \mathbb{N}^{\mathbb{N}}$, it follows that $\operatorname{card}(S_{\mathbb{N}}) \le \aleph_0^{\aleph_0} = c$, the cardinality of $\mathbb{R}$. For the opposite inequality, let $\sum_{k=1}^{\infty} a_k$ be a conditionally convergent series of real numbers. Given any $r \in \mathbb{R}$, select a permutation $\pi_r \in S_{\mathbb{N}}$ with $\sum_{k=1}^{\infty} a_{\pi_r(k)} = r$. The mapping from $\mathbb{R}$ to $S_{\mathbb{N}}$ defined by $r \to \pi_r$ is injective, and it follows that $\operatorname{card}(\mathbb{R}) \le \operatorname{card}(S_{\mathbb{N}})$.

Proof Without Words: A Weighted Sum of Triangular Numbers

$T_n = 1 + 2 + 3 + \cdots + n,\ n \ge 1 \quad \Rightarrow$

$$\sum_{k=1}^{n} kT_{k+1} = T_{T_{n+1}-1}.$$

E.g., $n = 4$:



$$2[T_2 + 2T_3 + 3T_4 + 4T_5] = 2T_{14} = 2T_{T_5-1}.$$

—Roger B. Nelson
Lewis & Clark College

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.

Stern, H.S., In favor of a quantitative boycott of the Bowl Championship Series, *Journal of Quantitative Analysis in Sports* 2 (1) (2006), `www.bepress.com/jqas/vol2/iss1/4/` .

Another football season is under way. The Bowl Championship Series (BCS) is supposed to match the top two college teams in a championship game at the end of the season; it uses polls of coaches combined with computer rankings designed by selected individuals. Stern rejects the computer rankings as having no clear-cut objective (beyond merely "validating" the polls) and being deliberately crippled (they are not allowed to use game scores or locations). He advocates that quantitative analysts have nothing to do with the BCS.

Diaconis, Persi, Susan Holmes, and Richard Montgomery, Dynamical bias in the coin toss, `www-stat.stanford.edu/~susan/papers/headswithJ.pdf` .

This paper proves that "vigorously flipped coins are biased to come up the same way they started," with the bias parametrized by the angle between the normal to the coin and the angular momentum vector. The source of the bias is unavoidable wobbling (precession) of the coin. The authors also offer data (for U.S. half-dollars—seen one lately?) that the bias is about .008 ± .001. OK, but earlier they claim that their data show "a bias of at least .01," and they confusingly conclude that "For tossed coins, the classical assumptions of independence with probability 1/2 are pretty solid."

Ash, Avner, and Robert Gross, *Fearless Symmetry: Exposing the Hidden Patterns of Numbers*, Princeton University Press, 2006; xxix + 272 pp, $24.95. ISBN 0–691–12492–2.

The subtitle is pure fluff, but few books are as ambitious as this one, and even fewer realize their ambitions as well. Recent years have seen an explosion of popularizations of mathematics, such as *Symmetry and the Monster* reviewed below. Where there is still a lack, due partly to the more limited audience, is in explanations of topics in advanced mathematics, *featuring the research motivation and exploring connections and context*, for those with enough background to bear with some notation, equations, and abstraction. This book's authors attempt to explain "cutting-edge mathematics" mainly to an audience that has studied calculus (that subject is not used, just the "mathematical maturity" supposed to be attained thereby). A claim to address readers "who have only studied some algebra" is misguided, except to open the door for bright high school students. The mathematics starts with groups and permutations in modern algebra and number theory, progresses through Galois theory and elliptic curves, investigates reciprocity laws, and culminates in an explanation of the proof of Fermat's Last Theorem. Those likely to benefit the most are mathematics majors and mathematicians (it would be great for a senior seminar); I could learn a lot by studying it carefully instead of just of just sampling it for this review. A useful feature (which more books should adopt) is a "Road Map" paragraph at the start of each chapter, which summarizes the purpose of the chapter and sets it in perspective.

Ronan, Mark, *Symmetry and the Monster: One of the Greatest Quests of Mathematics*, Oxford University Press, 2006; viii + 255 pp, $27. ISBN 0–19–280722–6.

This book traces the classification of simple groups from Lagrange to Richard Borcherds winning the Fields Medal in 1998 for work on the "Monstrous" Lie algebra with its tantalizing connections to number theory, crystals, and string theory. The classification is presented as a search for the "atoms" of symmetry. The author is a mathematician who worked at the edges of the classification theorem; knew all of the principals in it; and cites exact dates, places, and incidents. The work reads briskly and holds interest; apart from the quadratic formula, there is only one equation, near the end, that involves variables.

Packel, Edward, *The Mathematics of Games and Gambling*, 2nd ed., MAA, 2006; xi + 175 pp, $44 (member: $35). ISBN 10: 0–88385–646–8; ISBN 13: 978–0–88385–646–8.

With gambling by students a growing phenomenon, and the likelihood increasing that my town will become host to an Indian casino, popular demand may lead me to teach a course from this fine book. It has been updated to reflect newly popular games (e.g., video poker, Texas Holdem) and expanded gambling opportunities (sports betting on the Internet). Five of the seven chapters—on probability (roulette), dice games (backgammon, craps, chuck-a-luck), permutations/combinations (poker, bridge, Keno), binomial distribution (blackjack), and game theory (bluffing)—have an even dozen exercises each, with answers or hints to about one-third of them. The only background needed is high school algebra.

Fasano, Antonio, and Robert Natalini, Lost beauties of the Acropolis: What mathematics can say, *SIAM News* 39 (6) (July/August 2006) 1, 8. 7.

Air pollution vs. marble monuments: "Mathematics can produce not only elegant theories, but also very concrete answers," such as "cleaning" the marble too often is bad, and halving the damage would require reducing the sulfur dioxide in the air by a factor of four.

Peterson, Ivars, Chaotic chomp: The mathematics of crystal growth sheds light on a tantalizing game, *Science News* 170 (4) (22 July 2006) 58–60.

Classic games remain sources of inspiration for new mathematics. Chomp, reinvented in the 1970s by David Gale, is the latest example. Two players take turns removing a cookie from an initially rectangular layout; all cookies above and to the right of it are removed, too. The loser is the player removes the last cookie. There is a winning strategy for the first player (convince yourself by contradiction and exchanging roles of the players); but for nonsquare layouts with more than two columns, what cookie to take first is largely unknown. A new physics-based approach shows that the location of "winning" cookies and corresponding "losing" cookies can vary greatly with small changes in the size of the layout. Yet there are (broken) patterns, which resemble crystal growth processes and fractals. Among the latest results by Adam S. Landsberg (Claremont Colleges) and Eric J. Friedman (Cornell University), who apply re-normalization techniques from physics, is that each $3 \times n$ rectangle has a unique winning cookie.

Messer, Robert, and Philip Straffin, *Topology Now!*, MAA, 2006; xi + 240 pp, $49.95 (member: $39.95). ISBN 0–88385–744–8.

The exclamation point in the title emphasizes the authors' view that undergraduate "students should see the exciting geometric ideas of topology now (!) rather than later." Most mathematics majors do not take topology, and most topology courses are mainly point-set topology, so the authors have a point—indeed, for most students, there isn't any "later." This textbook emphasizes continuity, convergence, and connectedness, with applications to knots, manifolds, fixed-point theorems, and algebraic topology. It concludes with a chapter summarizing what most other topology texts focus on: metric spaces, topological spaces, and compactness. Recommended prerequisites are linear algebra, vector calculus, and one additional course in proof mathematics; no analysis or advanced calculus is presupposed. [Disclosure: Author Phil Straffin is a close colleague of mine at Beloit.]

NEWS AND LETTERS

Carl B. Allendoerfer Awards – 2006

The Carl B. Allendoerfer Awards, established in 1976, are made to authors of expository articles published in MATHEMATICS MAGAZINE. The Awards are named for Carl B. Allendoerfer, a distinguished mathematician at the University of Washington and President of the Mathematical Association of America, 1959–60.

Jeff Suzuki. The lost calculus (1637–1670): tangency and optimization without limits, MATHEMATICS MAGAZINE, 78 (2005) 339–353.

Citation Every teacher of calculus knows that the need to understand and work with limits is what makes the definition of the derivative hard. Hence we are not surprised when historians tell us that there was much controversy with the original notions of the derivative and that it took several generations of mathematicians to get the meaning and definition of limits right. But the non-historians among us tend not to be aware of the competing definitions of the derivative that were developed in the seventeenth century.

Suzuki's paper takes the reader on a pleasant ride through this little known history, and presents an "alternative universe" in which many of the problems encountered by a calculus student can be solved using an approach that does not need limits.

In his 1637 book *La Géométrie*, Descartes gives an algebraic method for finding tangents to algebraic functions. To find a tangent to a curve at the point P, we first find the equation of a circle tangent to the curve at the point. Then the wanted tangent will be the line through P perpendicular to the radius of the circle. The problem of finding the equation of the circle is reduced to making sure that an algebraic expression has a double root. The work of Jan Hudde (1628–1704) who developed an algorithm for detecting double roots gives this method an important boost.

Using well-chosen examples, Suzuki brings the calculus of Descartes, Hudde, Wallis, and Barrow back to life. He shows us interesting mathematics—with historical import—that can be appreciated by students and teachers of calculus.

Biographical Note Jeff Suzuki, Associate Professor of Mathematics at Brooklyn College, grew up in southern California unable to decide what he really wanted to do, so he studied mathematics, science, and history, eventually earning his Bachelor's in mathematics (with a physics concentration) and history from CSU Fullerton. He went on to earn his M.A. and Ph.D. from Boston University with a dissertation on the history of the Lagrange-Laplace proof of the long-term stability of the solar system. He enjoys trying new things, and introducing them to his children William Z and Dorothy X Suzuki-Burke (yes, their middle names are "Z" and "X"), and his wife Jacqui, who are frequently subjected to his culinary, musical, and linguistic efforts. As a historian of mathematics he is especially interested in the eighteenth century.

Response from Jeff Suzuki I am deeply honored to receive an Allendoerfer Prize for "The Lost Calculus." Just being able to bring to life an interesting episode in the history of mathematics is a great joy, and the added excitement of being nominated for this prize, not to mention actually winning, is immense. I'd like to thank Frank Farris for all his editorial work and patience, as well as Jacqui and the kids for putting up with my idiosyncratic writing and researching habits.

Robb T. Koether and John K. Osoinach Jr. Outwitting the lying oracle, MATHEMATICS MAGAZINE, 78 (2005) 98–109.

Citation A coin is flipped n times, and, each time, you first wager a certain amount and then try to predict the outcome. If you are wrong you lose your money and if you are right your money is doubled. Enter the oracle, and the game becomes much more enticing. Each time, after you announce your wager but before you make the prediction, the oracle will tell you how the coin is going to land. The twist is that the oracle may lie up to k times. What should your strategy be?

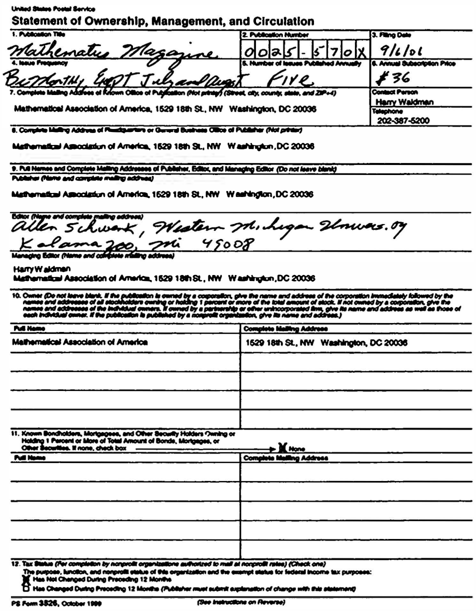
Koether and Osoinach draw the reader in with the case $n = 3$ and $k = 1$, use elementary probability and game theory to investigate the more general case, and then present variants of the problem that entice readers to do investigations of their own. This delightful article is well-written, interesting, and accessible. The "Oracle" metaphor is woven throughout seamlessly, and, before the reader's interest wanes, the authors throw in a variant, a question, or a concrete example. Students will find the article readable and fun.

Biographical Note Robb T. Koether is a professor of mathematics and computer science at Hampden-Sydney College in Virginia, where he has taught for the past 25 years. He earned his bachelor's degree in mathematics at the University of Richmond in 1973 and his Ph.D. in algebra at the University of Oklahoma in 1978 under the direction of Bernard R. McDonald. At Hampden-Sydney College Robb enjoys the opportunity to teach in many different areas of mathematics as well as computer science. He also enjoys solving mathematical contest problems and other puzzles, one of which led to the paper "Outwitting the Lying Oracle," for which this award was given. Outside of teaching and mathematics, he enjoys many outdoor activities, including cycling, camping, and backpacking on the Appalachian Trail. He is active in a number of community organizations, including the local Boy Scout troop and his church.

Response from Robb Koether I am greatly honored and humbled to be awarded the Allendoerfer prize. I would like to thank first my co-author John Osoinach, with whom it was a delight to work. It was largely through his enthusiasm that we pursued the ideas that developed into our paper "Outwitting the Lying Oracle." I would also like to thank Frank Farris and the anonymous reviewers who made many helpful and necessary comments on ways to improve our work. Finally I would like to thank my many graduate school professors who impressed on me, each in his own way, the importance of writing exactly what you mean to say. If winning the Allendoerfer prize is a measure of our success at communicating our enthusiasm for mathematics to others, then we are deeply gratified.

Biographical Note John K. Osoinach, Jr. earned his Ph.D. in 1998 at the University of Texas at Austin under the supervision of Dr. John Luecke. Immediately afterwards, he taught at Eureka College until 2000, when he married and moved to Virginia to take a position at Hampden-Sydney College. While his main area of research is in low-dimensional topology, specifically the geometry and topology of 3-manifolds, his work at small, liberal arts colleges has expanded his range of mathematical curiosity. In addition to his own research, John has supervised several undergraduates in research projects ranging from topology to the mathematics of social choice. He will begin a new position in the fall of 2006 as an Assistant Professor of Mathematics at Millsaps College.

Response from John Osoinach I am honored and deeply humbled in receiving the Carl B. Allendoerfer Award. As an undergraduate at Vanderbilt University I was given a subscription to Mathematics Magazine, and I have always read the articles with keen interest. Many of the ideas in the article originated from an undergraduate contest our department at Hampden-Sydney College sponsored, which itself was intended to spark curiosity among our majors. I am gratified and excited that this article might continue to inspire undergraduates to explore their own ideas in mathematics, so that they too might learn the joys of mathematical discovery. I would like to thank Bud Brown for his reading of the manuscript and his advice in submitting it to the Magazine, and Frank Farris for his numerous helpful comments. Finally, many thanks go to Robb Koether for his shared excitement and enthusiasm in writing the article.



United States Postal Service

Statement of Ownership, Management, and Circulation

1. Publication Title: Mathematics Magazine
2. Publication Number: 0025-570X
3. Filing Date: 9/6/06
4. Issue Frequency: Bi-Monthly, Except July and August
5. Number of Issues Published Annually: Five
6. Annual Subscription Price: $36

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4)
Mathematical Association of America, 1529 18th St., NW Washington, DC 20036

Contact Person
Harry Waldman
Telephone
202-387-5200

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer)
Mathematical Association of America, 1529 18th St., NW Washington, DC 20036

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address)
Mathematical Association of America, 1529 18th St., NW Washington, DC 20036

Editor (Name and complete mailing address)
Allen Schwenk, Western Michigan University
Kalamazoo, MI 49008

Managing Editor (Name and complete mailing address)
Harry Waldman
Mathematical Association of America, 1529 18th St., NW Washington, DC 20036

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

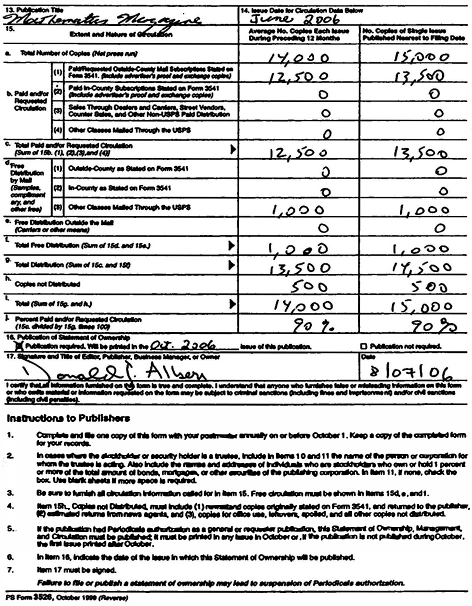
Full Name	Complete Mailing Address
Mathematical Association of America	1529 18th St., NW Washington, DC 20036

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

Full Name	Complete Mailing Address

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, October 1999 (See Instructions on Reverse)



13. Publication Title: Mathematics Magazine
14. Issue Date for Circulation Data Below: June 2006

15. Extent and Nature of Circulation			Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)			14,000	15,000
b. Paid and/or Requested Circulation	(1)	Paid/Requested Outside-County Mail Subscriptions Stated on Form 3541. (Include advertiser's proof and exchange copies)	12,500	13,500
	(2)	Paid In-County Subscriptions Stated on Form 3541 (Include advertiser's proof and exchange copies)	0	0
	(3)	Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution	0	0
	(4)	Other Classes Mailed Through the USPS	0	0
c. Total Paid and/or Requested Circulation (Sum of 15b. (1), (2),(3),and (4)) ▶			12,500	13,500
d. Free Distribution by Mail (Samples, compliment ary, and other free)	(1)	Outside-County as Stated on Form 3541	0	0
	(2)	In-County as Stated on Form 3541	0	0
	(3)	Other Classes Mailed Through the USPS	1,000	1,000
e. Free Distribution Outside the Mail (Carriers or other means)			0	0
f. Total Free Distribution (Sum of 15d. and 15e.) ▶			1,000	1,000
g. Total Distribution (Sum of 15c. and 15f) ▶			13,500	14,500
h. Copies not Distributed			500	500
i. Total (Sum of 15g. and h.) ▶			14,000	15,000
j. Percent Paid and/or Requested Circulation (15c. divided by 15g. times 100)			90 %	90 %

16. Publication of Statement of Ownership
☒ Publication required. Will be printed in the Oct. 2006 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner: Donald J. Albers
Date: 8/07/06

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

Instructions to Publishers

1. Complete and file one copy of this form with your postmaster annually on or before October 1. Keep a copy of the completed form for your records.
2. In cases where the stockholder or security holder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting. Also include the names and addresses of individuals who are stockholders who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing corporation. In item 11, if none, check the box. Use blank sheets if more space is required.
3. Be sure to furnish all circulation information called for in item 15. Free circulation must be shown in items 15d, e, and f.
4. Item 15h., Copies not Distributed, must include (1) newsstand copies originally stated on Form 3541, and returned to the publisher, (2) estimated returns from news agents, and (3), copies for office use, leftovers, spoiled, and all other copies not distributed.
5. If the publication had Periodicals authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published; it must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October.
6. In item 16, indicate the date of the issue in which this Statement of Ownership will be published.
7. Item 17 must be signed.

Failure to file or publish a statement of ownership may lead to suspension of Periodicals authorization.

PS Form 3526, October 1999 (Reverse)

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036