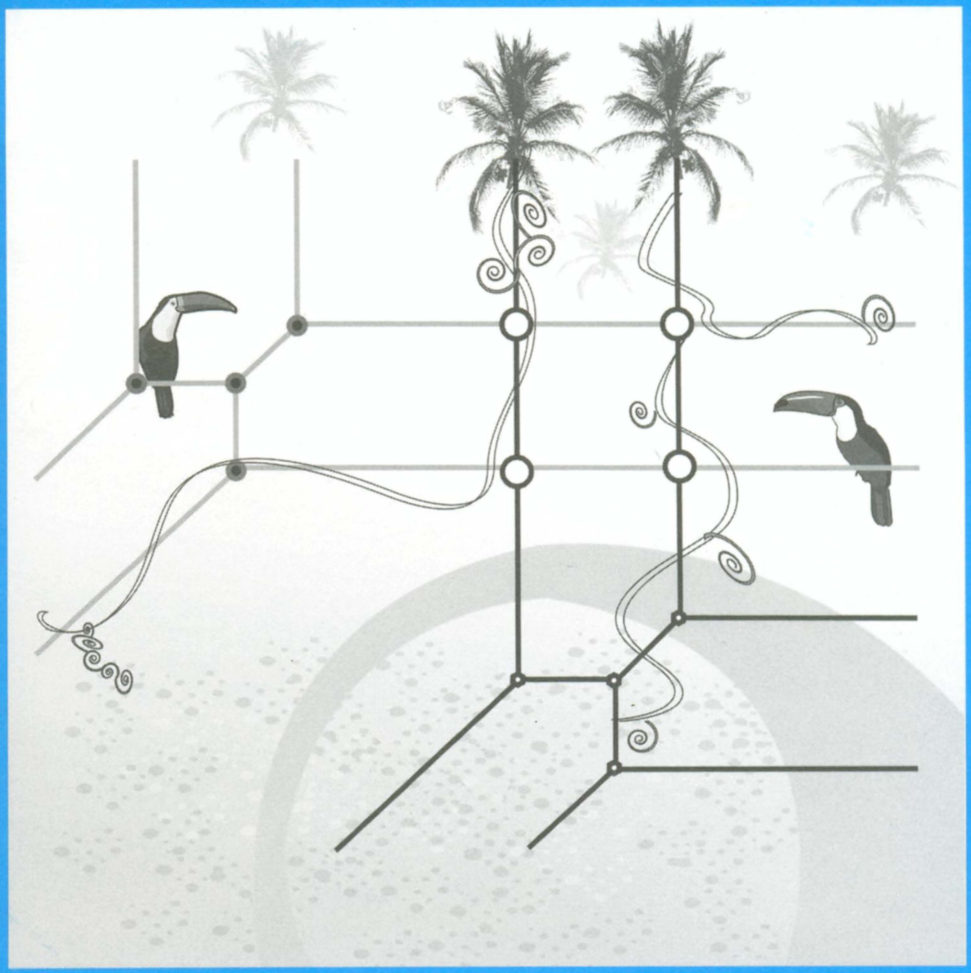
Vol. 82, No. 3, June 2009

MATHEMATICS MAGAZINE



Tropical Bézout

- Tropical Mathematics
- Envelopes and String Art
- Leveling with Lagrange: An Alternate View of Constrained Optimization

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.

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

Cover image: *Tropical Bézout*, by Hunter Cowdery, art student at West Valley College, in transition to San Jose State University, and Jason Challas, who lectures on computer graphics and fine art at West Valley College.

This illustration is an artistic enhancement of a diagram taken from the article "First Steps in Tropical Geometry" by Jürgen Richter-Gebert, Bernd Sturmfels, and Thorsten Theobald in *Idempotent mathematics and mathematical physics, Contemp. Math.*, 377, AMS, 2005, pp. 289–317. It illustrates Bézout's Theorem for two tropical quadrics in the plane: The two piecewise-linear curves intersect in four points, just as any two classical quadrics in the complex projective plane do.

AUTHORS

David Speyer graduated from University of California, Berkeley in 2005 with a Ph.D. dissertation in Tropical Geometry, and he has been a key player in the development of this emerging field. He also works in algebraic combinatorics and its connections to geometry and number theory. A winner of the prestigious Five-Year Research Fellowship awarded by the Clay Mathematics Institute, Speyer is currently based at the Massachusetts Institute of Technology.

Bernd Sturmfels is Professor of Mathematics, Statistics and Computer Science at UC Berkeley. A leading experimentalist among mathematicians, he has authored ten books and about 180 articles, in the areas of combinatorics, algebraic geometry, symbolic computation, and their applications. Sturmfels currently works on algebraic methods in statistics, optimization, and computational biology. His honors include the MAA's Lester R. Ford Award and designation as a George Polya Lecturer.

Gregory Quenell graduated from Harvard College in 1985 and earned his PhD at the University of Southern California in 1992. He has taught at Bucknell University, Oberlin College, Vassar College, Manhattan College, and Mount Holyoke College. He is currently an Associate Professor at the State University of New York College at Plattsburgh.

Dan Kalman received his Ph.D. from the University of Wisconsin in 1980, and has been at American University since 1993. Prior to that he had academic appointments (University of Wisconsin, Green Bay; Augustana College; Sioux Falls) and worked for eight years in the aerospace industry in Southern California. Kalman is a past Associate Executive Director of the MAA, author of a book published by the MAA, and frequent contributor to MAA journals. He delights in puns and word play of all kinds, and is an avid fan of Douglas Adams, J.R.R. Tolkien, and Gilbert and Sullivan.

Vol. 82, No. 3, June 2009

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
1529 Eighteenth St. NW
Washington DC 20036

Phone: (866) 821-1221
Fax: (202) 387-1208
E-mail: advertising@maa.org

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

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

MAA Service Center, *maahq@maa.org*

All at the address:

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

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

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

Printed in the United States of America

ARTICLES

Tropical Mathematics

DAVID SPEYER
Massachusetts Institute of Technology
Cambridge, MA 02139
speyer@math.mit.edu

BERND STURMFELS
University of California at Berkeley
Berkeley, CA 94720
bernd@math.berkeley.edu

This article is based on the Clay Mathematics Senior Scholar Lecture that was delivered by Bernd Sturmfels in Park City, Utah, on July 22, 2004. The topic of this lecture was the *tropical approach* in mathematics. This approach was in its infancy at that time, but it has since matured and is now an integral part of geometric combinatorics and algebraic geometry. It has also expanded into mathematical physics, number theory, symplectic geometry, computational biology, and beyond. We offer an elementary introduction to this subject, touching upon arithmetic, polynomials, curves, phylogenetics, and linear spaces. Each section ends with a suggestion for further research. The proposed problems are particularly well suited for undergraduate students. The bibliography contains numerous references for further reading in this field.

The adjective *tropical* was coined by French mathematicians, including Jean-Eric Pin **[16]**, in honor of their Brazilian colleague Imre Simon **[19]**, who was one of the pioneers in what could also be called *min-plus algebra*. There is no deeper meaning in the adjective *tropical*. It simply stands for the French view of Brazil.

Arithmetic

Our basic object of study is the *tropical semiring* $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$. As a set this is just the real numbers $\mathbb{R}$, together with an extra element ∞ that represents infinity. However, we redefine the basic arithmetic operations of addition and multiplication of real numbers as follows:

$$x \oplus y := \min(x, y) \qquad \text{and} \qquad x \odot y := x + y.$$

In words, the *tropical sum* of two numbers is their minimum, and the *tropical product* of two numbers is their sum. Here are some examples of how to do arithmetic in this strange number system. The tropical sum of 3 and 7 is 3. The tropical product of 3 and 7 equals 10. We write these as

$$3 \oplus 7 = 3 \qquad \text{and} \qquad 3 \odot 7 = 10.$$

Many of the familiar axioms of arithmetic remain valid in tropical mathematics. For instance, both addition and multiplication are *commutative*:

$$x \oplus y = y \oplus x \qquad \text{and} \qquad x \odot y = y \odot x.$$

The *distributive law* holds for tropical multiplication over tropical addition:

$$x \odot (y \oplus z) = x \odot y \oplus x \odot z,$$

where no parentheses are needed on the right, provided we respect the usual order of operations: Tropical products must be completed before tropical sums. Here is a numerical example to illustrate:

$$3 \odot (7 \oplus 11) = 3 \odot 7 = 10,$$
$$3 \odot 7 \oplus 3 \odot 11 = 10 \oplus 14 = 10.$$

Both arithmetic operations have a neutral element. Infinity is the *neutral element* for addition and zero is the *neutral element* for multiplication:

$$x \oplus \infty = x \qquad \text{and} \qquad x \odot 0 = x.$$

Elementary school students tend to prefer tropical arithmetic because the multiplication table is easier to memorize, and even long division becomes easy. Here are the tropical *addition table* and the tropical *multiplication table*:

$\oplus$	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	2
3	1	2	3	3	3	3	3
4	1	2	3	4	4	4	4
5	1	2	3	4	5	5	5
6	1	2	3	4	5	6	6
7	1	2	3	4	5	6	7

$\odot$	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	2	3	4	5	6	7	8
2	3	4	5	6	7	8	9
3	4	5	6	7	8	9	10
4	5	6	7	8	9	10	11
5	6	7	8	9	10	11	12
6	7	8	9	10	11	12	13
7	8	9	10	11	12	13	14

But watch out: tropical arithmetic is tricky when it comes to subtraction. There is no x to call "10 minus 3" because the equation $3 \oplus x = 10$ has no solution x at all. To stay on safe ground, we content ourselves with using addition $\oplus$ and multiplication $\odot$ only.

It is extremely important to remember that 0 is the multiplicative identity element. For instance, the tropical *Pascal's triangle*, whose rows are the coefficients appearing in a binomial expansion, looks like this:

$$\begin{array}{ccccccccccc}
 & & & & 0 & & & & & & \\
 & & & 0 & & 0 & & & & & \\
 & & 0 & & 0 & & 0 & & & & \\
 & 0 & & 0 & & 0 & & 0 & & & \\
0 & & 0 & & 0 & & 0 & & 0 & & \\
\cdots & \cdots & & \cdots & & \cdots & & \cdots & & \cdots &
\end{array}$$

For example, the fourth row in the triangle represents the identity

$$\begin{aligned}(x \oplus y)^3 &= (x \oplus y) \odot (x \oplus y) \odot (x \oplus y) \\ &= 0 \odot x^3 \oplus 0 \odot x^2 y \oplus 0 \odot x y^2 \oplus 0 \odot y^3.\end{aligned}$$

Of course, the zero coefficients can be dropped in this identity:

$$(x \oplus y)^3 = x^3 \oplus x^2 y \oplus x y^2 \oplus y^3.$$

Moreover, the *Freshman's Dream* holds for all powers in tropical arithmetic:

$$(x \oplus y)^3 = x^3 \oplus y^3.$$

The three displayed identities are easily verified by noting that the following equations hold in classical arithmetic for all $x, y \in \mathbb{R}$:

$$3 \cdot \min\{x, y\} = \min\{3x, 2x + y, x + 2y, 3y\} = \min\{3x, 3y\}.$$

Research problem The tropical semiring generalizes to higher dimensions: The set of convex polyhedra in $\mathbb{R}^n$ can be made into a semiring by taking $\odot$ as "Minkowski sum" and $\oplus$ as "convex hull of the union." A natural subalgebra is the set of all polyhedra that have a fixed *recession cone* C. If $n = 1$ and $C = \mathbb{R}_{\geq 0}$, this is the tropical semiring. Develop linear algebra and algebraic geometry over these semirings, and implement efficient software for doing arithmetic with polyhedra when $n \geq 2$.

Polynomials

Let $x_1, \ldots, x_n$ be variables that represent elements in the tropical semiring $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$. A *monomial* is any product of these variables, where repetition is allowed. By commutativity and associativity, we can sort the product and write monomials in the usual notation, with the variables raised to exponents,

$$x_2 \odot x_1 \odot x_3 \odot x_1 \odot x_4 \odot x_2 \odot x_3 \odot x_2 = x_1^2 x_2^3 x_3^2 x_4,$$

as long as we know from context that x_1^2 means $x_1 \odot x_1$ and not $x_1 \cdot x_1$. A monomial represents a function from $\mathbb{R}^n$ to $\mathbb{R}$. When evaluating this function in classical arithmetic, what we get is a linear function:

$$x_2 + x_1 + x_3 + x_1 + x_4 + x_2 + x_3 + x_2 = 2x_1 + 3x_2 + 2x_3 + x_4.$$

Although our examples used positive exponents, there is no need for such a restriction, so we allow negative integer exponents, so that every linear function with integer coefficients arises in this manner.

FACT 1. *Tropical monomials are the linear functions with integer coefficients.*

A *tropical polynomial* is a finite linear combination of tropical monomials:

$$p(x_1, \ldots, x_n) = a \odot x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \oplus b \odot x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n} \oplus \cdots$$

Here the coefficients $a, b, \ldots$ are real numbers and the exponents $i_1, j_1, \ldots$ are integers. Every tropical polynomial represents a function $\mathbb{R}^n \to \mathbb{R}$. When evaluating this function in classical arithmetic, what we get is the minimum of a finite collection of linear functions, namely,

$$p(x_1, \ldots, x_n) = \min\bigl(a + i_1 x_1 + \cdots + i_n x_n,\ b + j_1 x_1 + \cdots + j_n x_n, \ldots\bigr).$$

This function $p : \mathbb{R}^n \to \mathbb{R}$ has the following three important properties:

- p is continuous,
- p is piecewise-linear, where the number of pieces is finite, and
- p is concave, that is, $p(\frac{\mathbf{x}+\mathbf{y}}{2}) \geq \frac{1}{2}(p(\mathbf{x}) + p(\mathbf{y}))$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

It is known that every function that satisfies these three properties can be represented as the minimum of a finite set of linear functions. We conclude:

FACT 2. *The tropical polynomials in n variables $x_1, \ldots, x_n$ are precisely the piecewise-linear concave functions on $\mathbb{R}^n$ with integer coefficients.*

As a first example consider the general cubic polynomial in one variable x,

$$p(x) = a \odot x^3 \oplus b \odot x^2 \oplus c \odot x \oplus d. \tag{1}$$

To graph this function we draw four lines in the (x, y) plane: $y = 3x + a$, $y = 2x + b$, $y = x + c$, and the horizontal line $y = d$. The value of $p(x)$ is the smallest y-value such that (x, y) is on one of these four lines, that is, the graph of $p(x)$ is the lower envelope of the lines. All four lines actually contribute if

$$b - a \leq c - b \leq d - c. \tag{2}$$

These three values of x are the breakpoints where $p(x)$ fails to be linear, and the cubic has a corresponding factorization into three linear factors:

$$p(x) = a \odot (x \oplus (b - a)) \odot (x \oplus (c - b)) \odot (x \oplus (d - c)). \tag{3}$$

See FIGURE 1 for the graph and the roots of the cubic polynomial $p(x)$.

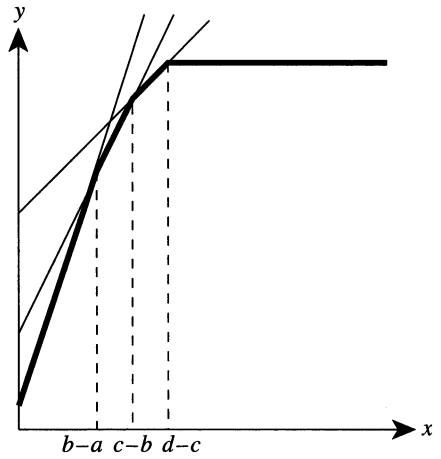


Figure 1 The graph of a cubic polynomial and its roots

Every tropical polynomial function can be written uniquely as a tropical product of tropical linear functions (in other words, the *Fundamental Theorem of Algebra* holds tropically). In this statement we must emphasize the word *function*. Distinct polynomials can represent the same function. We are not claiming that every polynomial factors as a product of linear polynomials. What we are claiming is that every polynomial can be replaced by an equivalent polynomial, representing the same function, that can be factored into linear factors. For example, the following polynomials represent the same function:

$$x^2 \oplus 17 \odot x \oplus 2 = x^2 \oplus 1 \odot x \oplus 2 = (x \oplus 1)^2.$$

Unique factorization of polynomials no longer holds in two or more variables. Here the situation is more interesting. Understanding it is our next problem.

Research problem The factorization of multivariate tropical polynomials into irreducible tropical polynomials is not unique. Here is a simple example:

$$\begin{aligned}&(0 \odot x \oplus 0) \odot (0 \odot y \oplus 0) \odot (0 \odot x \odot y \oplus 0)\\ &\quad = (0 \odot x \odot y \oplus 0 \odot x \oplus 0) \odot (0 \odot x \odot y \oplus 0 \odot y \oplus 0).\end{aligned}$$

Develop an algorithm (with implementation and complexity analysis) for computing all the irreducible factorizations of a given tropical polynomial. Gao and Lauder [**8**] have shown the importance of tropical factorization for the problem of factoring multivariate polynomials in the classical sense.

Curves

A tropical polynomial function $p : \mathbb{R}^n \to \mathbb{R}$ is given as the minimum of a finite set of linear functions. We define the *hypersurface* $\mathcal{H}(p)$ to be the set of all points $\mathbf{x} \in \mathbb{R}^n$ at which this minimum is attained at least twice. Equivalently, a point $\mathbf{x} \in \mathbb{R}^n$ lies in $\mathcal{H}(p)$ if and only if p is not linear at $\mathbf{x}$. For example, if $n = 1$ and p is the cubic in (1) with the assumption (2), then

$$\mathcal{H}(p) = \{b - a, c - b, d - c\}.$$

Thus the hypersurface $\mathcal{H}(p)$ is the set of "roots" of the polynomial $p(x)$.

In this section we consider the case of a polynomial in two variables:

$$p(x, y) = \bigoplus_{(i,j)} c_{ij} \odot x^i \odot y^j.$$

FACT 3. *For a polynomial in two variables, p, the tropical curve $\mathcal{H}(p)$ is a finite graph embedded in the plane $\mathbb{R}^2$. It has both bounded and unbounded edges, all of whose slopes are rational, and the graph satisfies a zero tension condition around each node, as follows:*

Consider any node (x, y) of the graph, which we may as well take to be the origin, $(0, 0)$. Then the edges adjacent to this node lie on lines with rational slopes. On each such ray emanating from the origin consider the smallest nonzero lattice vector. *Zero tension* at (x, y) means that the sum of these vectors is zero.

Our first example is a *line* in the plane. It is defined by a polynomial:

$$p(x, y) = a \odot x \oplus b \odot y \oplus c \qquad \text{where } a, b, c \in \mathbb{R}.$$

The curve $\mathcal{H}(p)$ consists of all points (x, y) where the function

$$p : \mathbb{R}^2 \to \mathbb{R}, \qquad (x, y) \mapsto \min(a + x, b + y, c)$$

is not linear. It consists of three half-rays emanating from the point $(x, y) = (c - a, c - b)$ into northern, eastern, and southwestern directions. The zero tension condition amounts to $(1, 0) + (0, 1) + (-1, -1) = (0, 0)$.

Here is a general method for drawing a tropical curve $\mathcal{H}(p)$ in the plane. Consider any term $\gamma \odot x^i \odot y^j$ appearing in the polynomial p. We represent this term by the point (γ, i, j) in $\mathbb{R}^3$, and we compute the convex hull of these points in $\mathbb{R}^3$. Now project the lower envelope of that convex hull into the plane under the map $\mathbb{R}^3 \to \mathbb{R}^2$, $(\gamma, i, j) \mapsto (i, j)$. The image is a planar convex polygon together with a distinguished subdivision Δ into smaller polygons. The tropical curve $\mathcal{H}(p)$ (actually its negative) is the *dual graph* to this subdivision. Recall that the dual to a planar graph is another planar graph whose vertices are the regions of the primal graph and whose edges represent adjacent regions.

As an example we consider the general quadratic polynomial

$$p(x, y) = a \odot x^2 \oplus b \odot xy \oplus c \odot y^2 \oplus d \odot x \oplus e \odot y \oplus f.$$

Then Δ is a subdivision of the triangle with vertices $(0, 0)$, $(0, 2)$, and $(2, 0)$. The lattice points $(0, 1)$, $(1, 0)$, $(1, 1)$ can be used as vertices in these subdivisions. Assuming that $a, b, c, d, e, f \in \mathbb{R}$ satisfy the conditions

$$2b \leq a + c, 2d \leq a + f, 2e \leq c + f,$$

the subdivision Δ consists of four triangles, three interior edges, and six boundary edges. The curve $\mathcal{H}(p)$ has four vertices, three bounded edges, and six half-rays (two northern, two eastern, and two southwestern). In FIGURE 2, we show the negative of the quadratic curve $\mathcal{H}(p)$ in bold with arrows. It is the dual graph to the subdivision Δ which is shown in thin lines.

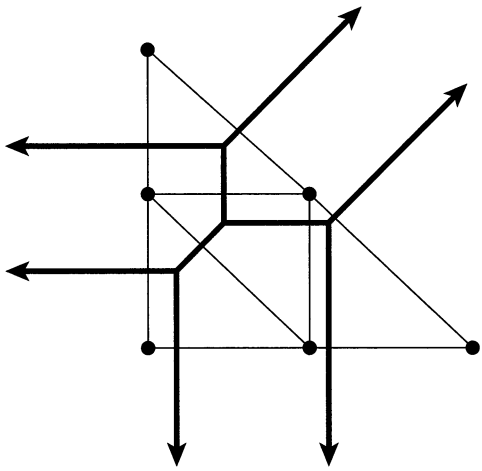


Figure 2 The subdivision Δ and the tropical curve

FACT 4. *Tropical curves intersect and interpolate like algebraic curves do.*

1. *Two general lines meet in one point, a line and a quadric meet in two points, two quadrics meet in four points, etc.*
2. *Two general points lie on a unique line, five general points lie on a unique quadric, etc.*

For a general discussion of *Bézout's Theorem* in tropical algebraic geometry, illustrated on the MAGAZINE cover, we refer to the article [**17**].

Research problem Classify all combinatorial types of *tropical curves in* 3*-space* of degree d. Such a curve is a finite embedded graph of the form

$$C = \mathcal{H}(p_1) \cap \mathcal{H}(p_2) \cap \cdots \cap \mathcal{H}(p_r) \subset \mathbb{R}^3,$$

where the p_i are tropical polynomials, C has d unbounded parallel halfrays in each of the four coordinate directions, and all other edges of C are bounded.

Phylogenetics

An important problem in computational biology is to construct a *phylogenetic tree* from distance data involving n leaves. In the language of biologists, the labels of the leaves are called *taxa*. These taxa might be organisms or genes, each represented by a

DNA sequence. For an introduction to phylogenetics we recommend books by Felsenstein [7] and Semple and Steele [18]. Here is an example, for $n = 4$, to illustrate how such data might arise. Consider an alignment of four genomes:

$$\begin{aligned} \text{Human:} &\quad ACAATGTCATTAGCGAT\ldots \\ \text{Mouse:} &\quad ACGTTGTCAATAGAGAT\ldots \\ \text{Rat:} &\quad ACGTAGTCATTACACAT\ldots \\ \text{Chicken:} &\quad GCACAGTCAGTAGAGCT\ldots \end{aligned}$$

From such sequence data, computational biologists infer the distance between any two taxa. There are various algorithms for carrying out this inference. They are based on statistical models of evolution. For our discussion, we may think of the distance between any two strings as a refined version of the Hamming distance (= the proportion of characters where they differ). In our (Human, Mouse, Rat, Chicken) example, the inferred distance matrix might be the following symmetric 4×4-matrix:

	H	M	R	C
H	0	1.1	1.0	1.4
M	1.1	0	0.3	1.3
R	1.0	0.3	0	1.2
C	1.4	1.3	1.2	0

The problem of phylogenetics is to construct a tree with edge lengths that represent this distance matrix, provided such a tree exists. In our example, a tree does exist, as depicted in FIGURE 3, where the number next to the each edge is its length. The distance between two leaves is the sum of the lengths of the edges on the unique path between the two leaves. For instance, the distance in the tree between "Human" and "Mouse" is $0.6 + 0.3 + 0.2 = 1.1$, which is the corresponding entry in the 4×4-matrix.

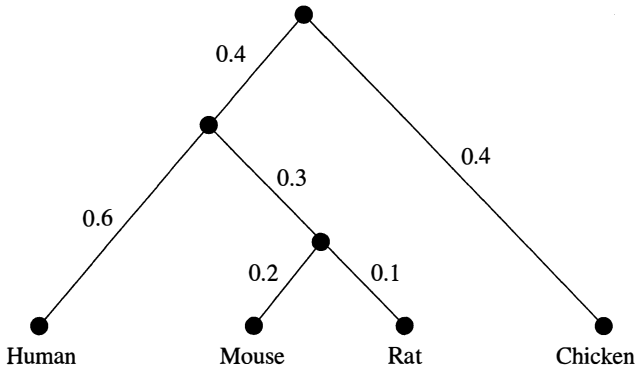


Figure 3 A phylogenetic tree

In general, considering n taxa, the *distance* between taxon i and taxon j is a positive real number d_{ij} which has been determined by some bio-statistical method. So, what we are given is a real symmetric $n \times n$-matrix

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{12} & 0 & d_{23} & \cdots & d_{2n} \\ d_{13} & d_{23} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{1n} & d_{2n} & d_{3n} & \cdots & 0 \end{pmatrix}.$$

We may assume that D is a *metric*, meaning that the triangle inequalities $d_{ik} \leq d_{ij} + d_{jk}$ hold for all i, j, k. This can be expressed by matrix multiplication:

FACT 5. *The matrix D represents a metric if and only if $D \odot D = D$.*

We say that a metric D on $\{1, 2, \ldots, n\}$ is a *tree metric* if there exists a tree T with n leaves, labeled $1, 2, \ldots, n$, and a positive length for each edge of T, such that the distance from leaf i to leaf j is d_{ij} for all i, j. Tree metrics occur naturally in biology because they model an evolutionary process that led to the n taxa.

Most metrics D are not tree metrics. If we are given a metric D that arises from some biological data then it is reasonable to assume that there exists a tree metric D_T that is close to D. Biologists use a variety of algorithms (for example, "neighbor joining") to construct such a nearby tree T from the given data D. In what follows we state a tropical characterization of tree metrics.

Let $X = (X_{ij})$ be a symmetric matrix with zeros on the diagonal whose $\binom{n}{2}$ distinct off-diagonal entries are unknowns. For each quadruple $\{i, j, k, l\} \subset \{1, 2, \ldots, n\}$ we consider the following tropical polynomial of degree two:

$$p_{ijkl} = X_{ij} \odot X_{kl} \oplus X_{ik} \odot X_{jl} \oplus X_{il} \odot X_{jk}. \tag{4}$$

This polynomial is the *tropical Grassmann-Plücker relation*, and it is simply the tropical version of the classical Grassmann-Plücker relation among the 2×2-subdeterminants of a 2×4-matrix [**14**, Theorem 3.20].

It defines a hypersurface $\mathcal{H}(p_{ijkl})$ in the space $\mathbb{R}^{\binom{n}{2}}$. The *tropical Grassmannian* is the intersection of these $\binom{n}{4}$ hypersurfaces. It is denoted

$$Gr_{2,n} = \bigcap_{1 \leq i < j < k < l \leq n} \mathcal{H}(p_{ijkl}).$$

This subset of $\mathbb{R}^{\binom{n}{2}}$ has the structure of a *polyhedral fan*, which means that it is the union of finitely many convex polyhedral cones that fit together nicely.

FACT 6. *A metric D on $\{1, 2, \ldots, n\}$ is a tree metric if and only if its negative $X = -D$ is a point in the tropical Grassmannian $Gr_{2,n}$.*

The statement is a reformulation of the *Four Point Condition* in phylogenetics, which states that D is a tree metric if and only if, for all $1 \leq i < j < k < l \leq n$, the *maximum* of the three numbers $D_{ij} + D_{kl}$, $D_{ik} + D_{jl}$, and $D_{il} + D_{jk}$ is attained at least twice. For $X = -D$, this means that the *minimum* of the three numbers $X_{ij} + X_{kl}$, $X_{ik} + X_{jl}$, and $X_{il} + X_{jk}$ is attained at least twice, or, equivalently, $X \in \mathcal{H}(p_{ijkl})$. The tropical Grassmannian $Gr_{2,n}$ is also known as the *space of phylogenetic trees* [**3, 14, 20**]. The combinatorial structure of this beautiful space is well studied and well understood.

Often, instead of measuring the pairwise distances between the various taxa, it can be statistically more accurate to consider all r-tuples of taxa and jointly measure the dissimilarity within each r-tuple. For example, in the above tree, the joint dissimilarity of the triple {Human, Mouse, Rat} is 1.2, the sum of the lengths of all edges in the subtree containing the mouse, human, and rat. Lior Pachter and the first author showed in [**15**] that it is possible to reconstruct the tree from the data for all r-tuples, as long as $n \geq 2r - 1$.

At this point in the original 2004 lecture notes, we had posed a research problem: to characterize the image of the given embedding of $Gr_{2,n}$ into $\mathbb{R}^{\binom{n}{r}}$, particularly in the case $r = 3$. Since then, Christiano Bocci and Filip Cools [**4**] have solved the problem for $r = 3$, and they made significant progress on the problem for higher r. While there

is still work to be done, we now suggest the following less studied problem, borrowed from the end of [**14**, Chapter 3].

Research problem We say that a metric D has *phylogenetic rank* $\leq k$ if there exist k tree metrics $D^{(1)}, D^{(2)}, \ldots, D^{(k)}$ such that

$$D_{ij} = \max\left(D_{ij}^{(1)}, D_{ij}^{(2)}, \ldots, D_{ij}^{(k)}\right) \qquad \text{for all } 1 \leq i, j \leq n.$$

Equivalently, the matrix $X = -D$ is the sum of the matrices $X^{(i)} = -D^{(i)}$:

$$X = X^{(1)} \oplus X^{(2)} \oplus \cdots \oplus X^{(k)}.$$

The aim of the notion of phylogenetic rank is to model distance data that is a mixture of k different evolutionary histories. The set of metrics of phylogenetic rank $\leq k$ is a polyhedral fan in $\mathbb{R}^{\binom{n}{2}}$. Compute this fan, and explore its combinatorial, geometric, and topological properties, especially for $k = 2$.

Tropical linear spaces

Generalizing our notion of a line, we define a *tropical hyperplane* to be a subset of $\mathbb{R}^n$ of the form $\mathcal{H}(\ell)$, where ℓ is a tropical linear function in n unknowns:

$$\ell(x) = a_1 \odot x_1 \oplus a_2 \odot x_2 \oplus \cdots \oplus a_n \odot x_n.$$

Here $a_1, \ldots, a_n$ are arbitrary real constants. Solving linear equations in tropical mathematics means computing the intersection of finitely many hyperplanes $\mathcal{H}(\ell)$. It is tempting to define tropical linear spaces simply as intersections of tropical hyperplanes. However, this would not be a good definition because such arbitrary intersections can have mixed dimension, and they do not behave the way linear spaces do in classical geometry.

A better notion of tropical linear space is derived by allowing only those intersections of hyperplanes that are "sufficiently complete," in a sense we explain later. The definition we offer directly generalizes our discussion about phylogenetics. The idea is that phylogenetic trees are lines in tropical projective space, whose Plücker coordinates X_{ij} are the negated pairwise distances d_{ij}.

We consider the $\binom{n}{d}$-dimensional space $\mathbb{R}^{\binom{n}{d}}$ whose coordinates $X_{i_1 \cdots i_d}$ are indexed by d-element subsets $\{i_1, \ldots, i_d\}$ of $\{1, 2, \ldots, n\}$. Let S be any $(d-2)$-element subset of $\{1, 2, \ldots, n\}$ and let i, j, k, and l be any four distinct indices in $\{1, \ldots, n\} \backslash S$. The corresponding *three-term Grassmann Plücker relation* $p_{S,ijkl}$ is the following tropical polynomial of degree two:

$$p_{S,ijkl} = X_{Sij} \odot X_{Skl} \oplus X_{Sik} \odot X_{Sjl} \oplus X_{Sil} \odot X_{Sjk}. \tag{5}$$

We define the *Dressian* to be the intersection

$$Dr_{d,n} = \bigcap_{S,i,j,k,l} \mathcal{H}(p_{S,ijkl}) \subset \mathbb{R}^{\binom{n}{d}},$$

where the intersection is over all S, i, j, k, l as above. The term *Dressian* refers to Andreas Dress, an algebraist who now works in computational biology. For relevant references to his work and further details see [**11**].

Note that in the special case $d = 2$ we have $S = \emptyset$, the polynomial (5) is the four point condition in (4). In this special case, $Dr_{2,n} = Gr_{2,n}$, and this is precisely the space of phylogenetic trees discussed previously.

We now fix an arbitrary point X with coordinates $(X_{i_1 \cdots i_d})$ in the Dressian $Dr_{d,n}$. For any $(d+1)$-subset $\{j_0, j_1, \ldots, j_d\}$ of $\{1, 2, \ldots, n\}$ we consider the following tropical linear form in the variables $x_1, \ldots, x_n$:

$$\ell^X_{j_0 j_1 \cdots j_d} = \bigoplus_{r=0}^{d} X_{j_0 \cdots \widehat{j_r} \cdots j_d} \odot x_r, \tag{6}$$

where the $\widehat{\ }$ means to omit j_r. The *tropical linear space* associated with the point X is the following set:

$$L_X = \bigcap \mathcal{H}(\ell^X_{j_0 j_1 \cdots j_n}) \subset \mathbb{R}^n.$$

Here the intersection is over all $(d+1)$-subsets $\{j_0, j_1, \ldots, j_d\}$ of $\{1, 2, \ldots, n\}$.

The tropical linear spaces are precisely the sets L_X where X is any point in $Dr_{d,n} \subset \mathbb{R}^{\binom{n}{d}}$. These objects are studied in detail in [**21**] and [**11**]. The "sufficient completeness" referred to in the first paragraph of this section means that we need to solve linear equations using the above formula for L_X, in order for an intersection of hyperplanes actually to be a linear space. The definition of linear space given here is more inclusive than the one used elsewhere [**6, 17, 20**], where L_X was required to come from ordinary algebraic geometry over a field with a suitable valuation.

For example, a 3-dimensional tropical linear subspace of $\mathbb{R}^n$ (a.k.a. a two-dimensional plane in tropical projective $(n-1)$-space) is the intersection of $\binom{n}{4}$ tropical hyperplanes, each of whose defining linear forms has four terms:

$$\ell^X_{j_0 j_1 j_2 j_3} = X_{j_0 j_1 j_2} \odot x_{j_3} \oplus X_{j_0 j_1 j_3} \odot x_{j_2} \oplus X_{j_0 j_2 j_3} \odot x_{j_1} \oplus X_{j_1 j_2 j_3} \odot x_{j_0}.$$

We note that even the very special case when each coordinate of X is either 0 (the multiplicative unit) or ∞ (the additive unit) is really interesting. Here L_X is a polyhedral fan known as the *Bergman fan* of a matroid [**1**].

Tropical linear spaces have many of the properties of ordinary linear spaces. First, they are pure polyhedral complexes of the correct dimension:

FACT 7. *Each maximal cell of the tropical linear space L_X is d-dimensional.*

Every tropical linear space L_X determines its vector of tropical Plücker coordinates X uniquely up to tropical multiplication (= classical addition) by a common scalar. If L and L' are tropical linear spaces of dimensions d and d' with $d + d' \geq n$, then L and L' meet. It is not quite true that two tropical linear spaces intersect in a tropical linear space but it is almost true. If L and L' are tropical linear spaces of dimensions d and d' with $d + d' \geq n$ and v is a generic small vector then $L \cap (L' + v)$ is a tropical linear space of dimension $d + d' - n$. Following [**17**], it makes sense to define the *stable intersection* of L and L' by taking the limit of $L \cap (L' + v)$ as v goes to zero, and this limit will again be a tropical linear space of dimension $d + d' - n$.

It is not true that a d-dimensional tropical linear space can always be written as the intersection of $n - d$ tropical hyperplanes. The definition shows that $\binom{n}{d+1}$ hyperplanes are always enough. At this point in the original 2004 lecture notes, we had asked: What is the minimum number of tropical hyperplanes needed to cut out any tropical linear space of dimension d in n-space? Are n hyperplanes always enough? These questions were answered by Tristram Bogart in [**2**, Theorem 2.10], and a more refined combinatorial analysis was given by Josephine Yu and Debbie Yuster in [**22**]. Instead of posing a new research problem, we end this article with a question.

Are there any textbooks on tropical geometry? As of June 2009, there seem to be no introductory texts on tropical geometry, despite the elementary nature of the basic

definitions. The only book published so far on tropical algebraic geometry is the volume [**10**] which is based on an Oberwolfach seminar held in 2004 by Ilia Itenberg, Grigory Mikhalkin, and Eugenii Shustin. That book emphasizes connections to topology and real algebraic geometry. Several expository articles offer different points of entry. In addition to [**17**], we especially recommend the expositions by Andreas Gathmann [**9**] and Eric Katz [**12**]. These are aimed at readers who have a background in algebraic geometry. Grigory Mikhalkin is currently writing a research monograph on tropical geometry for the book series of the Clay Mathematical Institute, while Diane Maclagan and the second author have begun a book project titled *Introduction to Tropical Geometry*. Preliminary manuscripts can be downloaded from the authors' homepages. In fall 2009, the Mathematical Sciences Research Institute (MSRI) in Berkeley will hold a special semester on Tropical Geometry.

Acknowledgment. Speyer was supported in this work by a Clay Mathematics Institute Research Fellowship, Sturmfels by the National Science Foundation (DMS-0456960, DMS-0757236).

REFERENCES

F. Ardila and C. Klivans, The Bergman complex of a matroid and phylogenetic trees, *J. Combinatorial Theory Ser. B* **96** (2006) 38–49.

T. Bogart, A. Jensen, D. Speyer, B. Sturmfels, and R. Thomas, Computing tropical varieties, *J. Symbolic Computation* **42** (2007) 54–73.

3. L. Billera, S. Holmes, and K. Vogtman, Geometry of the space of phylogenetic trees, *Advances in Applied Math.* **27** (2001) 733–767.
4. C. Bocci and F. Cools, A tropical interpretation of m-dissimilarity maps, preprint, `arXiv:0803.2184`.

P. Butkovič, Max-algebra: the linear algebra of combinatorics? *Linear Algebra Appl.* **367** (2003) 313–335.

6. M. Develin, F. Santos, and B. Sturmfels, On the rank of a tropical matrix, pages 213–242 in *Combinatorial and Computational Geometry*, Mathematical Sciences Research Institute Publication, Vol. 52, Cambridge Univ. Press, Cambridge, 2005.
7. J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2003.
8. S. Gao and A. Lauder, Decomposition of polytopes and polynomials, *Discrete and Computational Geometry* **26** (2001) 89–104.
9. A. Gathmann, Tropical algebraic geometry, *Jahresbericht der Deutschen Mathematiker-Vereinigung* **108** (2006) 3–32.
10. I. Itenberg, G. Mikhalkin, and E. Shustin, *Tropical Algebraic Geometry*, Oberwolfach Seminars Series, Vol. 35, Birkhäuser, Basel, 2007.
11. S. Hermann, A. Jensen, M. Joswig, and B. Sturmfels, How to draw tropical planes, preprint, `arxiv:0808.2383`.
12. E. Katz, A tropical toolkit, to appear in *Expositiones Mathematicae*, `arXiv:math/0610878`.
13. G. Mikhalkin, Enumerative tropical geometry in $\mathbb{R}^2$, *J. Amer. Math. Soc.* **18** (2005) 313–377.
14. L. Pachter and B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge Univ. Press, Cambridge, 2005.

L. Pachter and D. Speyer, Reconstructing trees from subtree weights, *Appl. Math. Lett.* **17** (2004) 615–621.

16. J.-E. Pin, Tropical semirings, *Idempotency* (Bristol, 1994), 50–69, Publ. Newton Inst., Vol. 11, Cambridge Univ. Press, Cambridge, 1998.
17. J. Richter-Gebert, B. Sturmfels, and T. Theobald, First steps in tropical geometry, pages 289–317 in *Idempotent Mathematics and Mathematical Physics*, Contemporary Mathematics, Vol. 377, American Mathematical Society, Providence, RI, 2005.
18. C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, Oxford, 2003.
19. I. Simon, Recognizable sets with multiplicities in the tropical semiring, pages 107–120 in *Mathematical Foundations of Computer Science* (Carlsbad, 1988), Lecture Notes in Computer Science, Vol. 324, Springer, Berlin, 1988.

D. Speyer and B. Sturmfels, The tropical Grassmannian, *Advances in Geometry* **4** (2004) 389–411.

D. Speyer, Tropical linear spaces, *SIAM J. Disc. Math.* **22** (2008) 1527–1558.

22. J. Yu and D. Yuster, Representing tropical linear spaces by circuits, in *Formal Power Series and Algebraic Combinatorics (FPSAC '07)*, Proceedings, Tianjin, China, 2007.

Envelopes and String Art

GREGORY QUENELL
Plattsburgh State University
Plattsburgh, NY 12901
gregory.quenell@plattsburgh.edu

A familiar problem

For each $\alpha \in [0, 1]$ let ℓ_α be the line segment in $\mathbb{R}^2$ connecting the point $(\alpha, 0)$ with the point $(0, 1 - \alpha)$. FIGURE 1 shows the segments ℓ_α for α equal to integer multiples of $1/20$.

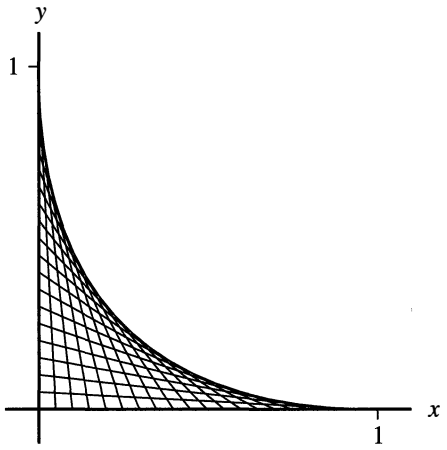


Figure 1 Line segments connecting $(\alpha, 0)$ to $(0, 1 - \alpha)$

The upper right edge of this collection suggests a curve C from $(1, 0)$ to $(0, 1)$. This curve is an *envelope* of the collection $\{\ell_\alpha : 0 \leq \alpha \leq 1\}$, and has the property that each of its tangent lines contains one of the segments ℓ_α. We'd like to find an equation for C.

Although C is characterized in terms of tangent lines, and thus is a solution to a differential equation, it turns out that we can determine C in a quite elementary way. The key is to recognize that when α and β are close together, the intersection point of ℓ_α and ℓ_β is close to C, and as β approaches α, the intersection point of ℓ_α and ℓ_β approaches a point of C, as in FIGURE 2.

To get the calculations going, we parametrize each segment ℓ_α as

$$\begin{aligned} \ell_\alpha(t) &= (1 - t)(\alpha, 0) + t(0, 1 - \alpha) \\ &= ((1 - t)\alpha, t(1 - \alpha)), \quad 0 \leq t \leq 1. \end{aligned}$$

If $\beta \neq \alpha$, then ℓ_α and ℓ_β intersect at the point

$$\ell_\alpha(1 - \beta) = \ell_\beta(1 - \alpha) = (\alpha\beta, (1 - \alpha)(1 - \beta)).$$

To get a point on C, we want to take the limit of this last expression as $\beta \to \alpha$. Even a common-sense notion of limits will suffice here: we simply substitute α for β to find that the point $(\alpha^2, (1 - \alpha)^2)$ lies on C. Just like that, we have a parametrization for C:

$$x = \alpha^2 \quad \text{and} \quad y = (1 - \alpha)^2. \tag{1}$$

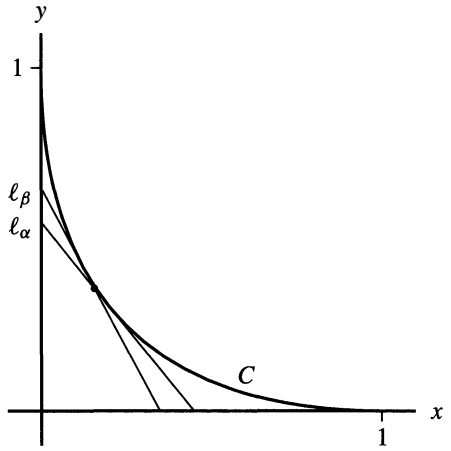


Figure 2 When α is close to β, the segments ℓ_α and ℓ_β intersect near C

With $0 \le \alpha \le 1$, both α and $1 - \alpha$ are nonnegative, so we can write (1) as

$$\sqrt{x} + \sqrt{y} = 1. \tag{2}$$

A popular (in one sense) problem in calculus textbooks asks the student to show that the sum of the x- and y-intercepts of the tangent lines to the curve $\sqrt{x} + \sqrt{y} = \sqrt{c}$ is always equal to c [**5**, p. 234, problem 38]. Here we have solved what appears to be a more difficult problem—finding a curve whose tangent lines have intercepts with a constant sum—and we have done so with only the merest hint of the calculus.

A curve with equation $|ax|^p + |by|^p = c^p$, where $0 < p < 2$, is called a *hypœllipse* [**7**]. Since the equation of our envelope C has this form with $a = b = c = 1$, we can describe C as one quarter of the unit hypocircle with exponent $1/2$.

Eliminating the radicals from (2), we find that the points of C satisfy

$$x^2 + y^2 - 2xy - 2x - 2y + 1 = 0.$$

This is clearly a conic section, and since its discriminant is 0, it is a parabola. If we lift the restriction $0 \le \alpha \le 1$ and consider the ℓ_α as lines rather than just line segments, we find that the envelope of this family of lines is a parabola tangent to the coordinate axes and containing the first-quadrant part of the unit hypocircle. FIGURE 3 shows how all the pieces fit together.

String art

FIGURE 1 calls to mind the craft of string art, in which the artist creates a decorative pattern by driving nails into a board at intervals along a few lines or curves and then connecting selected pairs of nails with stretched strings. One result of such an exercise is shown in FIGURE 4.

A very simple string art recipe calls for spacing nails evenly along the x- and y-axes and running a string between two nails when the sum of their x- and y-coordinates is constant. This gives the pattern in FIGURE 1 with the upper right edge approximating a branch of a hypocircle.

Evenly-spaced nails along perpendicular lines go only so far as an outlet for creative energy, and the ambitious string artist will surely want to experiment with different spacings, nonperpendicular lines, and nonlinear curves.

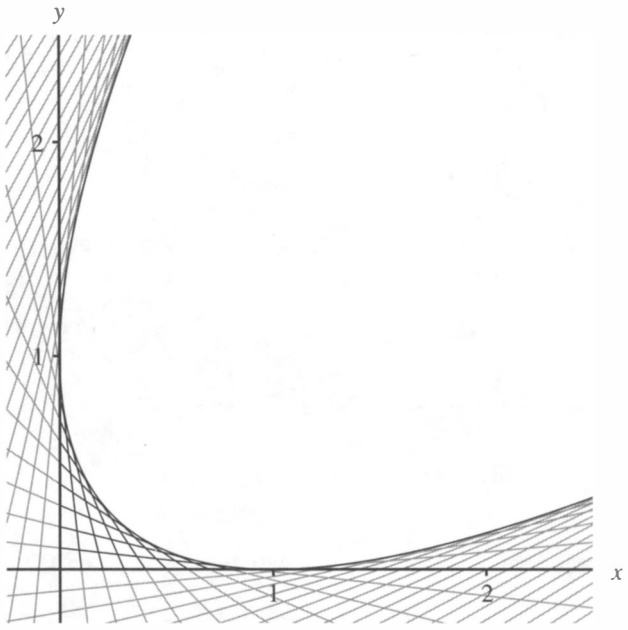


Figure 3 The parabola emerges when we draw the ℓ_α as lines rather than line segments and allow values of α outside $[0, 1]$

Likewise, the mathematician will ask what sort of envelopes arise when we vary the recipe for the parametrized family $\{\ell_\alpha\}$. In the following sections, we experiment with some of these string-art variations, and see what envelope curves we get.

Spacing functions First, let's keep the nails on the x- and y-axes, but change the way we space them. Each line ℓ_α will be determined by two points, $(X(\alpha), 0)$ and $(0, Y(\alpha))$, where X and Y are assumed to be differentiable. We will call X and Y *spacing functions.*

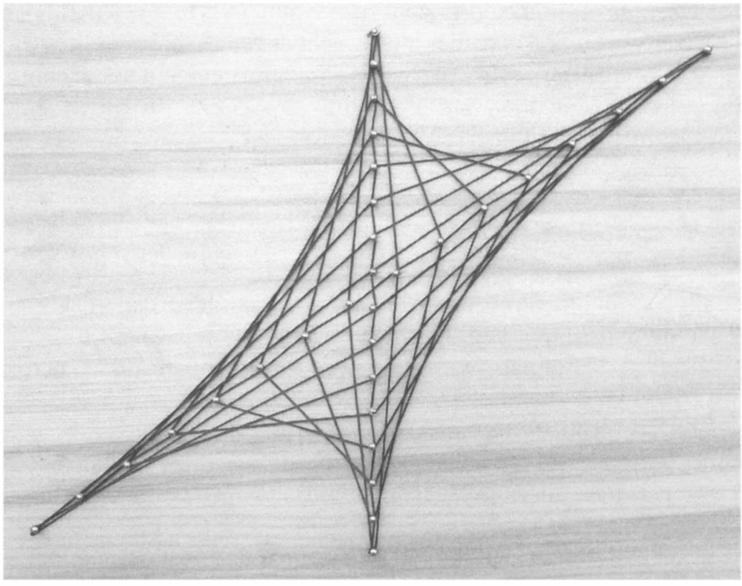


Figure 4 An unpretentious (to say the least) example of string art

Again, we begin by finding the intersection point of two lines in this family. A straightforward calculation shows that ℓ_α and ℓ_β intersect at the point

$$\left(\frac{X(\alpha)X(\beta)(Y(\beta)-Y(\alpha))}{X(\alpha)Y(\beta)-Y(\alpha)X(\beta)}, \frac{Y(\alpha)Y(\beta)(X(\alpha)-X(\beta))}{X(\alpha)Y(\beta)-Y(\alpha)X(\beta)}\right). \tag{3}$$

To find a point on the envelope of the family ℓ_α, we take the limit of (3) as $\beta \to \alpha$. We use the standard trick of adding and subtracting $X(\alpha)Y(\alpha)$ in the denominator and then introducing some $(\beta-\alpha)$s to form difference quotients. For the x-coordinate of the envelope point, we get

$$\begin{aligned} x &= \lim_{\beta\to\alpha} \frac{X(\alpha)X(\beta)(Y(\beta)-Y(\alpha))}{X(\alpha)(Y(\beta)-Y(\alpha))-Y(\alpha)(X(\beta)-X(\alpha))} \\ &= \lim_{\beta\to\alpha} \frac{X(\alpha)X(\beta)\frac{Y(\beta)-Y(\alpha)}{\beta-\alpha}}{X(\alpha)\frac{Y(\beta)-Y(\alpha)}{\beta-\alpha}-Y(\alpha)\frac{X(\beta)-X(\alpha)}{\beta-\alpha}}. \end{aligned} \tag{4}$$

Now X and Y were assumed differentiable, so the mean value theorem says that the limit in (4) is equal to

$$\frac{(X(\alpha))^2Y'(\alpha)}{X(\alpha)Y'(\alpha)-Y(\alpha)X'(\alpha)}. \tag{5}$$

A similar calculation gives the y-coordinate of a point on the envelope as

$$\frac{-(Y(\alpha))^2X'(\alpha)}{X(\alpha)Y'(\alpha)-Y(\alpha)X'(\alpha)}. \tag{6}$$

Expressions (5) and (6) parametrize the envelope of the family ℓ_α determined by any pair of spacing functions X and Y.

There is a standard way to compute envelopes using calculus. (We describe this later in the article; one source for more information is Giblin's article in this MAGAZINE about "zigzags" [**1**].) Readers familiar with that method may enjoy using it to derive equations (5) and (6) to judge whether our way is simpler.

Linear spacing functions: the details In our first example, we used the particularly simple spacing functions

$$X(\alpha)=\alpha \quad \text{and} \quad Y(\alpha)=1-\alpha.$$

Let's look at what we get for an envelope when X and Y are general linear functions

$$X(\alpha)=r\alpha+h \quad \text{and} \quad Y(\alpha)=s\alpha+k,$$

with r and s nonzero. In this case, (5) and (6) yield the parametrization

$$x=\frac{(r\alpha+h)^2s}{sh-rk} \quad \text{and} \quad y=-\frac{(s\alpha+k)^2r}{sh-rk}, \tag{7}$$

provided that $sh \neq rk$. (If $sh=rk$, then the lines ℓ_α are parallel, and there is no envelope.)

What sort of curve is this? Guided by our earlier calculations, we begin experimenting with $\sqrt{|x|}$ and $\sqrt{|y|}$, and eventually find that the x and y in (7) satisfy

$$\sqrt{|sx|}+\sqrt{|ry|}=\frac{|(r\alpha+h)s|}{\sqrt{|sh-rk|}}+\frac{|(s\alpha+k)r|}{\sqrt{|sh-rk|}}. \tag{8}$$

If $(r\alpha + h)s$ and $(s\alpha + k)r$ have different signs, then the right side of (8) reduces to $\sqrt{|sh - rk|}$. Thus the points of our envelope curve for which $(r\alpha + h)s$ and $(s\alpha + k)r$ have different signs lie on the hypoellipse $\sqrt{|sx|} + \sqrt{|ry|} = \sqrt{|sh - rk|}$.

To see that we have a parabola as well, we need only verify (a tedious but straightforward task) that the parametrization in (7) satisfies the equation

$$s^2x^2 + r^2y^2 + 2rsxy - 2(sh - rk)(sx - ry) + (sh - rk)^2 = 0 \tag{9}$$

for every value of α.

As in our introductory example, this parabola is tangent to both the coordinate axes. The points of tangency are $(0, (rk - sh)/r)$ and $((sh - rk)/s, 0)$.

As a simple example, take $X(\alpha) = 2\alpha + 1$ and $Y(\alpha) = \alpha + 2$. The envelope (FIGURE 5) is parametrized by

$$x = -\frac{(2\alpha + 1)^2}{3} \quad \text{and} \quad y = \frac{2(\alpha + 2)^2}{3}.$$

It is tangent to the coordinate axes at the points $(0, 3/2)$ (where $\alpha = -1/2$) and $(-3, 0)$ (where $\alpha = -2$). Between the two points of tangency, the points on the envelope satisfy $\sqrt{|x|} + \sqrt{|2y|} = \sqrt{3}$.

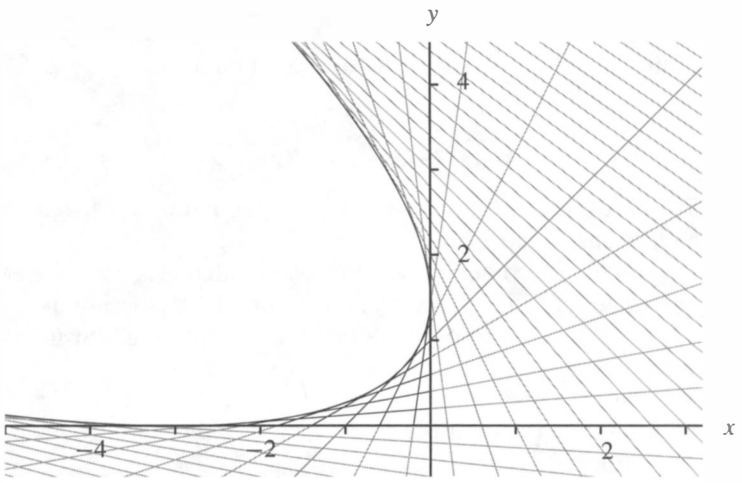


Figure 5 With $X(\alpha) = 2\alpha + 1$ and $Y(\alpha) = 2 + \alpha$, the envelope is a parabola in the second quadrant. The quarter hypoellipse is tangent to the ℓ_α with $-2 \le \alpha \le -1/2$.

The reader with a taste for the classical theory of conic sections might want to verify that the focus and directrix of the parabola given by (9) are

$$\left(\frac{s(sh - rk)}{r^2 + s^2}, -\frac{r(sh - rk)}{r^2 + s^2}\right) \quad \text{and} \quad rx - sy = 0.$$

Off the coordinate axes

Suppose now that the string artist chooses (nonparallel) nailing lines n_1 and n_2 other than the coordinate axes. The mathematician may then straighten out this skewed situation by cooking up an affine transformation $\mathcal{A}$ that takes the x-axis to line n_1 and the y-axis to line n_2.

Assuming that the string artist is still spacing the nails evenly, the mathematician can find spacing functions $X(\alpha) = r\alpha + h$ and $Y(\alpha) = s\alpha + k$ such that $\mathcal{A}(X, 0)$ and $\mathcal{A}(0, Y)$ agree with the artist's nailing pattern. The envelope curve on the string art will then be the image under $\mathcal{A}$ of the parabolic curve that we found earlier. Since a nonsingular affine transformation takes parabolas to parabolas, an envelope curve that arises from evenly-spaced nails along any two lines n_1 and n_2 will also lie on a parabola, tangent to lines n_1 and n_2.

Thus, the envelope curves in FIGURE 4, which might at first glance suggest a pair of hyperbolas, are in fact parts of four parabolas, pairwise tangent at the ends of the nailing lines.

A connection to game theory For a different sort of illustration, we make a short digression into game theory. Consider a two-player, non-zero-sum game in which each player has two pure strategies available. We can represent such a game using a table:

		Player II	
		A	B
Player I	A	(2, 0)	(3, 6)
	B	(4, 2)	(0, 0)

The ordered pair (2, 0) in the upper left corner means that if Player I chooses strategy IA and Player II chooses strategy IIA, then the payoff to Player I is 2 and the payoff to Player II is 0.

Now suppose that the game is played repeatedly. For each play, Player I uses some random device to select a strategy. Suppose she chooses strategy IA with probability α and IB with probability $1 - \alpha$. Similarly, Player II chooses strategy IIA with probability β and IIB with probability $1 - \beta$. For this repeated play, we can calculate an *expected payoff*, since we know the probability with which each of the four payoff pairs will occur. The expected payoff is

$$\alpha\beta(2, 0) + \alpha(1 - \beta)(3, 6) + (1 - \alpha)\beta(4, 2) + (1 - \alpha)(1 - \beta)(0, 0).$$

We factor this to get

$$(1 - \beta)(\alpha(3, 6) + (1 - \alpha)(0, 0)) + \beta(\alpha(2, 0) + (1 - \alpha)(4, 2)). \tag{10}$$

Since $0 \le \beta \le 1$, expression (10) shows that the expected payoff, considered as a point in $\mathbb{R}^2$, lies on the line segment connecting $\alpha(2, 0) + (1 - \alpha)(4, 2)$ with $\alpha(3, 6) + (1 - \alpha)(0, 0)$. We denote this line segment ℓ_α and observe that the set of possible expected payoffs is equal to the union of all the segments ℓ_α, $0 \le \alpha \le 1$.

This is the shaded region in FIGURE 6. The curved part of its boundary is an envelope of exactly the kind we have been discussing, so it must lie on a parabola.

To find an explicit parametrization for this parabola, we straighten things out with the affine map $\mathcal{A} : (x, y) \mapsto (x + y - 2, x + 2y - 4)$. This map takes the portion of the x-axis between $x = 4$ and $x = 6$ to the line segment connecting (2, 0) with (4, 2) and the portion of the y-axis between $y = 2$ and $y = 5$ to the line segment connecting the origin with (3, 6). We take

$$X(\alpha) = 4 + 2\alpha \quad \text{and} \quad Y(\alpha) = 5 - 3\alpha$$

and apply (7) to get the parabola parametrized by

$$x = \frac{3}{22}(4 + 2\alpha)^2 \quad \text{and} \quad y = \frac{1}{11}(5 - 3\alpha)^2.$$

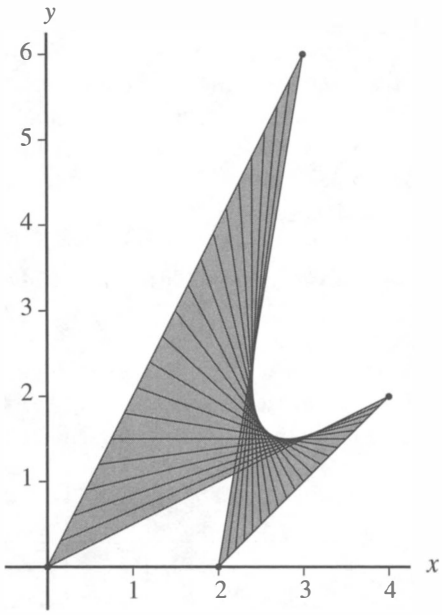


Figure 6 Possible expected payoffs in a non-zero-sum game

The curved part of the boundary of the shaded region in FIGURE 6 is the image of this parabola under $\mathcal{A}$. Its parametrization is

$$x = \frac{3}{11}(5\alpha^2 - 2\alpha + 9) \quad \text{and} \quad y = \frac{6}{11}(4\alpha^2 - 6\alpha + 5),$$

with $0 \leq \alpha \leq 1$.

The shaded region shows the set of expected payoffs that can arise if each player uses a random device to choose a strategy each time the game is played. If the game is played a large number of times and the average payoff converges to a point outside the shaded region, then we have evidence that the players' random devices are *not independent*. In certain circumstances, this might indicate collusion, espionage, or just poor random number generators.

Envelopes from right triangles

We now move our nailing lines back to the coordinate axes and consider some specific nonlinear spacing functions.

Constant area Let $X(\alpha) = ke^{\alpha}$ and $Y(\alpha) = ke^{-\alpha}$ for some nonzero constant k. Then the lines ℓ_α have the property that the product of the x- and y-intercepts of each line is equal to k^2. Put another way, the ℓ_α are the hypotenuses of a family of right triangles with constant area.

Applying formulas (5) and (6), we find that the envelope of this family of lines is parametrized by

$$x = \frac{ke^{\alpha}}{2} \quad \text{and} \quad y = \frac{ke^{-\alpha}}{2}.$$

FIGURE 7 shows the envelope of this family of lines. This time the envelope *is* a branch of a hyperbola; its equation is $xy = k^2/4$.

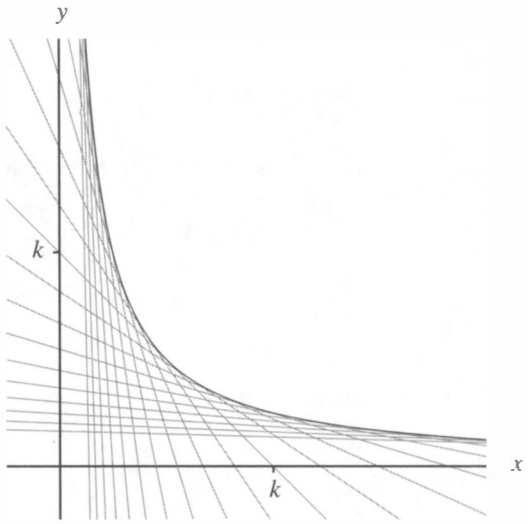


Figure 7 With $X = ke^{\alpha}$ and $Y = ke^{-\alpha}$, the envelope is the hyperbola $xy = k^2/4$

Constant length: the calculus of ladders Consider a ladder of fixed length L sliding down a wall. The ladder describes a family of lines, shown in FIGURE 8, for which the distance between the x-intercept $(X(\alpha), 0)$ and the y-intercept $(0, Y(\alpha))$ is constantly equal to L.

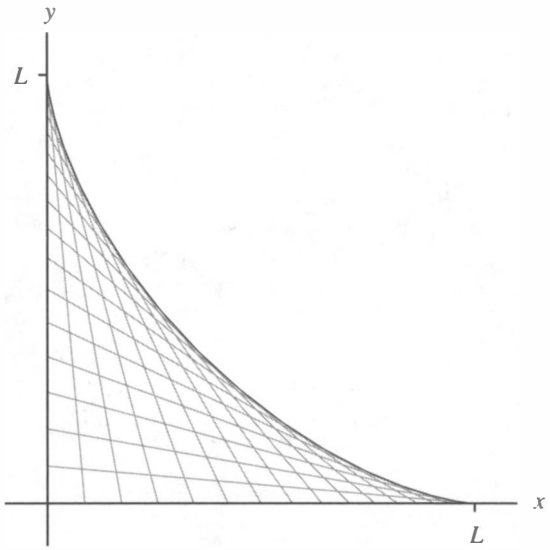


Figure 8 The sliding ladder's envelope lies along the astroid $x^{2/3} + y^{2/3} = L^{2/3}$

We want to choose X and Y so that $X(\alpha)^2 + Y(\alpha)^2 = L^2$. An obvious choice is

$$X(\alpha) = L\cos\alpha \quad \text{and} \quad Y(\alpha) = L\sin\alpha.$$

Formulas (5) and (6) give the envelope of this set of lines as

$$x = L\cos^3\alpha \quad \text{and} \quad y = L\sin^3\alpha. \tag{11}$$

A Cartesian equation for this curve is $x^{2/3} + y^{2/3} = L^{2/3}$; it is called an astroid.

Applying a handful of trig identities, one can rewrite (11) as

$$x = \frac{3L}{4}\cos\alpha + \frac{L}{4}\cos(3\alpha) \quad \text{and} \quad y = \frac{3L}{4}\sin\alpha - \frac{L}{4}\sin(3\alpha),$$

showing that the sliding-ladder envelope also happens to lie on the hypocycloid traced by a point on a circle of radius $L/4$ rolling along the inside of a circle of radius L.

Want to carry your ladder around a corner from one hallway into another? If the widths of the hallways are x and y, then the astroid equation above shows that x, y, and L have to satisfy $x^{2/3} + y^{2/3} \geq L^{2/3}$ in order for the ladder to make it around the corner horizontally.

Constant perimeter Let r be a positive constant. Let

$$X(\alpha) = r - \alpha \quad \text{and} \quad Y(\alpha) = \frac{2r\alpha}{r+\alpha}$$

for $0 \leq \alpha \leq r$. It is not hard to check that

$$X(\alpha) + Y(\alpha) + \sqrt{(X(\alpha))^2 + (Y(\alpha))^2} = 2r.$$

That is, the triangles with vertices at the origin, $(X(\alpha), 0)$, and $(0, Y(\alpha))$ all have perimeter $2r$.

We use (5) and (6) to find a parametrization for the envelope of the hypotenuses of these triangles. We find

$$x = \frac{r(r-\alpha)^2}{r^2+\alpha^2} \quad \text{and} \quad y = \frac{2r\alpha^2}{r^2+\alpha^2}.$$

A little algebra shows that the x and y in this parametrization satisfy $(x-r)^2 + (y-r)^2 = r^2$ for all α. This envelope is part of a circle. Thus if we have a loop of string with length $2r$ and we stretch it into a triangle with a right angle at the origin and the legs along the positive x- and y-axes, the hypotenuse of the triangle will be tangent to the circle with center (r, r) and radius r. FIGURE 9 shows the circle determined by these taut-string triangles.

Free-form nailing

We conclude by considering a generalization in which the nailing lines become arbitrary curves in the plane.

Let $C_1(\alpha) = (X_1(\alpha), Y_1(\alpha))$ and $C_2(\alpha) = (X_2(\alpha), Y_2(\alpha))$ be two differentiable curves. For each value of α, we can parametrize the line ℓ_α determined by $C_1(\alpha)$ and $C_2(\alpha)$ as

$$\ell_\alpha(t) = (1-t)(X_1(\alpha), Y_1(\alpha)) + t(X_2(\alpha), Y_2(\alpha)). \tag{12}$$

Then for $\alpha \neq \beta$, the lines ℓ_α and ℓ_β (assuming they are not parallel) intersect when t has the value

$$\frac{(X_2(\beta) - X_1(\beta))(Y_1(\beta) - Y_1(\alpha)) - (Y_2(\beta) - Y_1(\beta))(X_1(\beta) - X_1(\alpha))}{(X_2(\beta) - X_1(\beta))(Y_2(\alpha) - Y_1(\alpha)) - (Y_2(\beta) - Y_1(\beta))(X_2(\alpha) - X_1(\alpha))}. \tag{13}$$

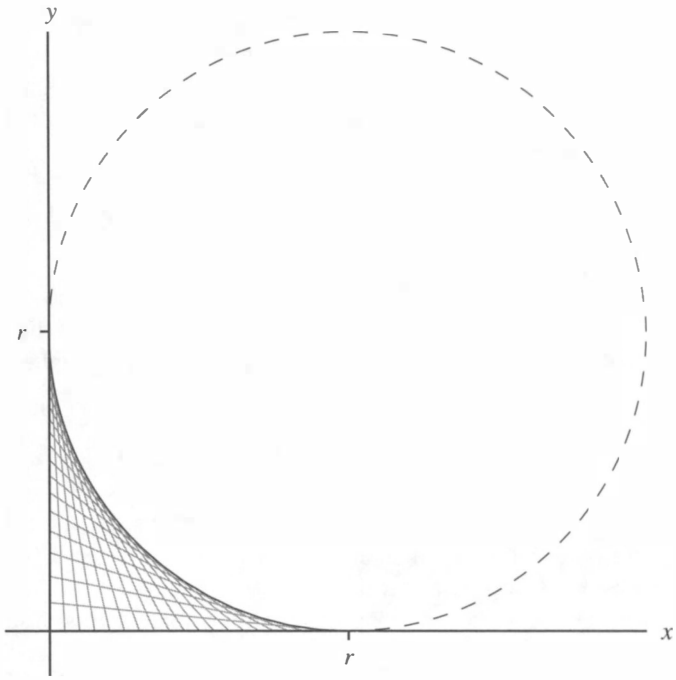


Figure 9 Triangles with a right angle at the origin and perimeter $2r$ are all tangent to the circle $(x-r)^2+(y-r)^2=r^2$

To find the point where ℓ_α is tangent to the envelope of the family $\{\ell_\alpha\}$, we'll need to find the limit of (13) as $\beta \to \alpha$. The factors $(Y_1(\beta)-Y_1(\alpha))$ and $(X_1(\beta)-X_1(\alpha))$ in the numerator of (13) can be rewritten as $(\beta-\alpha)$ times the obvious difference quotients. With a little algebraic sleight of hand, the denominator or (13) can be put in the form

$$[(X_2(\beta)-X_2(\alpha))-(X_1(\beta)-X_1(\alpha))](Y_2(\alpha)-Y_1(\alpha))$$
$$-[(Y_2(\beta)-Y_2(\alpha))-(Y_1(\beta)-Y_1(\alpha))](X_2(\alpha)-X_1(\alpha)).$$

Again, the first factor in each term is $(\beta-\alpha)$ times an appropriate difference quotient. In the limit, the difference quotients become derivatives and we find

$$\lim_{\beta\to\alpha} t = \frac{(X_2-X_1)Y_1'-(Y_2-Y_1)X_1'}{(X_2'-X_1')(Y_2-Y_1)-(Y_2'-Y_1')(X_2-X_1)}, \tag{14}$$

where all the functions on the right are evaluated at α. We evaluate (12) at this value of t to find a point (x, y) on the envelope. We get (evaluating everything at α)

$$x = \frac{(X_1X_2'-X_1'X_2)(Y_2-Y_1)-(X_1Y_2'-Y_1'X_2)(X_2-X_1)}{(X_2'-X_1')(Y_2-Y_1)-(Y_2'-Y_1')(X_2-X_1)} \tag{15}$$

$$y = \frac{(Y_1X_2'-X_1'Y_2)(Y_2-Y_1)-(Y_1Y_2'-Y_1'Y_2)(X_2-X_1)}{(X_2'-X_1')(Y_2-Y_1)-(Y_2'-Y_1')(X_2-X_1)} \tag{16}$$

as the generalization of (5) and (6).

Calling on calculus Here's another route to (15) and (16) that takes a brief detour through 3-space. This may look more familiar to a differential geometer, though the technique uses only multivariable calculus.

For each value of the parameter α, we lift the line ℓ_α up into space as a horizontal line in the plane $z = \alpha$, as if the segments were moving upward as α increases. This makes a (ruled) surface in $\mathbb{R}^3$ and when you look down on a wire-frame version of it, you see pictures just like the ones we have drawn. We can easily write this surface as a level surface of a function of three variables:

$$\begin{aligned} F(x, y, \alpha) &= (x - X_1(\alpha))(Y_2(\alpha) - Y_1(\alpha)) - (y - Y_1(\alpha))(X_2(\alpha) - X_1(\alpha)) \\ &= 0. \end{aligned} \tag{17}$$

It turns out that the envelope of the set of lines is the visual edge of the surface when it is viewed from above and (infinitely) far away. Points on that visual edge satisfy the equation

$$F_\alpha(x, y, \alpha) = 0, \tag{18}$$

because a vertical ray from above is tangent to the surface at such points; equivalently, the gradient vector of F is horizontal at any points where F_α is zero.

The standard way to compute envelopes is to solve (17) and (18) simultaneously, determining x and y as functions of α. Doing that in this case, we recover the parametrization (15) and (16). This is elegant, but our simpler method does not require the use of "It turns out," and might appeal more to those who prefer to keep a planar problem in the plane.

Lines joining two circles Let's try out formulas (15) and (16) on some simple parametrized curves: two circles centered at the origin.

Let $X_1(\alpha) = 2\sin\alpha$, $Y_1(\alpha) = 2\cos\alpha$, $X_2(\alpha) = -\sin\alpha$, and $Y_2(\alpha) = \cos\alpha$. Two points determine each of our lines ℓ_α. The first moves clockwise around a circle of radius 2, starting at the 12 o'clock position; the second moves counter-clockwise around a circle of radius 1, also starting at 12 o'clock. Unlike the hands of an actual clock, our two points move at the same angular rate.

Applying (15) and (16) to these two circles, we get an envelope curve parametrized by

$$x = -4\sin^3\alpha \quad \text{and} \quad y = \frac{4}{3}\cos^3\alpha. \tag{19}$$

The obvious quantity to compute here is $x^{2/3} + (3y)^{2/3}$, and we find that the points along (19) satisfy

$$x^{2/3} + (3y)^{2/3} = 4^{2/3}.$$

Our envelope is another hypoellipse. In fact, since the exponent is $2/3$, we might also describe this envelope as a sort of squashed astroid.

FIGURE 10 shows the two paths, the envelope, and the lines ℓ_α that generate just the first quarter of the envelope. Since the points of tangency along most of this envelope lie outside the segments "between the nails," a traditional string art piece using this recipe would show only a small part of the envelope. To get the whole picture, as FIGURE 10 suggests, we might want to try a kind of augmented string art, in which we stretch the lines all the way to the frame, and anchor them there.

Other envelopes that arise from points moving around circles are the topic of articles by Simoson **[3, 4]**.

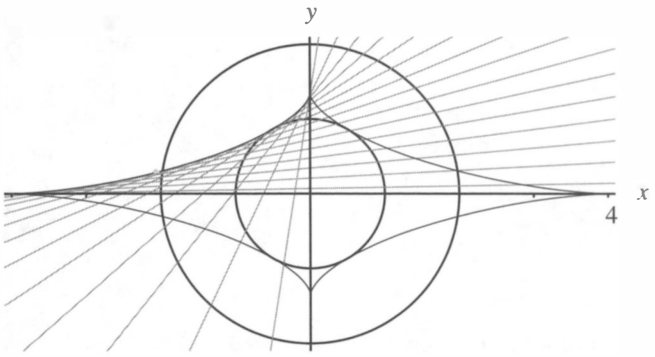


Figure 10 The lines determined by one point moving clockwise at radius 2 and another moving counterclockwise at radius 1 are all tangent to a squashed astroid

Acknowledgment. The author is deeply grateful to his anonymous referees, one of whom helped to identify the proper focus of this article and contributed many of the really good examples. The same referee drew the author's attention to [2, Volume I, Chapter X], where the subject of envelopes is developed more thoroughly.

REFERENCES

1. Peter J. Giblin, Zigzags, this MAGAZINE **74** (2001) 259–272.
2. Édouard Goursat, *A Course in Mathematical Analysis*, Dover, New York, 1959.
3. Andrew J. Simoson, An envelope for a Spirograph, *College Math. J.* **28** (1997) 134–139.
4. Andrew J. Simoson, The trochoid as a tack in a bungee cord, this MAGAZINE **73** (2000) 171–184.
5. James Stewart, *Single Variable Calculus: Early Transcendentals*, Thomson Brooks/Cole, Belmont, CA, 2003.
6. Philip D. Straffin, *Game Theory and Strategy*, volume 36 of New Mathematical Library, Mathematical Association of America, Washington, DC, 1993.
7. David H. Von Seggern, *CRC Standard Curves and Surfaces*, CRC Press, Boca Raton, FL, 1993.

Leveling with Lagrange: An Alternate View of Constrained Optimization

DAN KALMAN
American University
Washington, D.C. 20016
kalman@american.edu

In most calculus books today [**11, 14, 15**], Lagrange multipliers are explained as follows. Say that we wish to find the maximum value of f subject to the condition that $g = 0$. Under certain assumptions about f and g, the Lagrange multipliers theorem asserts that at the solution point, the gradient vectors ∇f and ∇g are parallel. Therefore, either $\nabla f = \lambda \nabla g$ for some real number λ, or $\nabla g = 0$. Combined with the equation $g = 0$, this gives necessary conditions for a solution to the constrained optimization problem. We will refer to this as the standard approach to Lagrange multipliers.

An earlier tradition approaches this subject far differently. It defines a new function, $F = f + \lambda g$, that incorporates both the objective function and the constraint, and in which λ is considered to be an additional variable. Here, F is referred to as a *Lagrangian function*. The conditions for F to achieve an unconstrained extremum are then determined, and these become necessary conditions for a solution to the original problem. This is the Lagrangian function approach to Lagrange multipliers.

Both approaches produce the same necessary conditions, and lead to identical solutions of constrained optimization problems. The second approach is closer to the original spirit of Lagrange's work, and is popular in introductory works on mathematical methods in economics, as well as calculus texts with an applied or business emphasis. Unfortunately, it is often presented with an attractive, but fundamentally incorrect, intuitive justification [**1, 2, 6, 7**]—that the problem of finding a maximum (say) of f subject to a constraint is transformed into a search for a maximum of F without constraint. The problem is, a constrained maximum of f need not correspond to a local maximum of F. The Lagrange theorem asserts the existence of a corresponding *critical point* for F, but says nothing about whether this critical point is actually an extremum. And as we will see, further examination reveals an unexpected result: the critical point of F that satisfies the Lagrange condition is *never* a local extremum. Therefore, the idea that a constrained optimum of f corresponds to an unconstrained optimum of the Lagrangian F is never correct. For ease of reference, this mistaken idea will be termed the *transformation fallacy*, highlighting the intuition that incorporating the constraint into the objective function transforms a constrained optimization problem into an unconstrained optimization problem.

The perpetuation of a flawed intuitive explanation of the Lagrangian approach is a shame, not least because it represents a dilution of the power of a true intuition. At its heart, the idea that the constrained problem is transformed into an unconstrained problem doesn't make any sense. (How could it? It is incorrect!) And yet, on casual consideration, the idea of imposing the constraint implicitly is seductively plausible. There is even an informal proof that seems to justify this idea.

Happily, there is an alternate justification of the Lagrangian function approach to constrained optimization. It provides a memorable geometric intuition and has a catchy name, *Lagrangian leveling*. In contrast to the central idea of the transformation fallacy, which is necessarily vague, Lagrangian leveling directly confronts (and remedies) the true source of difficulty: at a constrained maximum, the partial derivatives of the objective function need not vanish.

The goal of this paper is to draw attention to the transformation fallacy, and to the idea of Lagrangian leveling. A specific example illustrating the fallacy and a critique of the rationale behind the fallacy will be presented. We will also see that leveling leads to both a proof of the Lagrange multiplier theorem and an interpretation of the value of the multiplier λ at the extremum.

The Lagrangian function approach For the sake of concreteness, let us review the standard approach to Lagrange multipliers in a specific context. Suppose that $f(x, y)$ and $g(x, y)$ are differentiable functions on a domain in $\mathbb{R}^2$, and that we wish to find the maximum of f subject to the constraint $g = 0$. As outlined in the introduction, if the solution occurs at a point (x, y) where $\nabla g \neq 0$, then there must exist a real λ such that $g(x, y) = 0$ and $\nabla f(x, y) = \lambda \nabla g(x, y)$, giving necessary conditions for the solution.

In the Lagrangian function approach we arrive at essentially the same conditions but follow a different route. Define $F(x, y, \lambda) = f(x, y) + \lambda g(x, y)$. Now F is regarded as a function of three variables. Under the influence of the transformation fallacy, we observe that a necessary condition for F to achieve a local maximum is that all of its partial derivatives vanish. (The fallacy may be avoided at this stage by asking for critical points, rather than local maxima.) Thus we arrive at the system of equations

$$\begin{aligned} \frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} &= 0 \\ \frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} &= 0 \\ g(x, y) &= 0. \end{aligned} \tag{1}$$

The last equation, which is imposed as a constraint in the standard approach, arises here from the partial derivative $\partial F/\partial \lambda$. It shows that for any critical point (x, y, λ) of F, the components x and y satisfy the constraint equation $g = 0$.

Justifying Lagrange multipliers

In the standard formulation, the key idea is that ∇f and ∇g are parallel at the solution point (x^*, y^*). This can be justified in a variety of ways. For example, suppose $\mathbf{r}(t)$ parameterizes the constraint curve, with $(x^*, y^*) = \mathbf{r}(t^*)$. Then since $g(\mathbf{r}(t))$ is constant and $f(\mathbf{r}(t))$ has a maximum at t^*, both of their derivatives vanish at t^*. Applying the chain rule there, both ∇f and ∇g are orthogonal to $\mathbf{r}'$ and so are parallel. A less formal argument uses the directional derivative. At the point of the constraint curve C where f assumes a maximum, the derivative of f in the direction of C must vanish. This shows that ∇f and the curve's tangent vector are orthogonal. On the other hand, since C is a level curve of g, we also have ∇g orthogonal to C at every point, and the result follows as before.

Our understanding of vectors, and in particular, of ∇f as pointing in the direction of steepest increase of f, provides other justifications. At a maximum, ∇f must be perpendicular to C for otherwise, its nonzero projection along C would point in the direction of increase of f. Or, in the succinct formulation of Farris [**3**],

> We think of ∇f as the "desired direction" and ∇g as the "forbidden direction." If you are at a relative maximum of f on the constraint $g = 0$, then this can only be because the direction you would like to go to get more f, namely ∇f, lines up perfectly with the forbidden direction.

These arguments all depend on properties of the gradient, and there are several others of a similar nature [**8**]. There are other arguments from viewpoints that are decidely different. One involving *penalty functions* is most natural in the context of a minimization problem. So, suppose we wish to minimize $f(x, y)$ subject to the condition $g(x, y) = 0$. We form the function $F(x, y) = f(x, y) + \sigma g(x, y)$, and seek an unconstrained minimum. We think of σ as a penalty imposed for allowing g to assume a positive value. The larger σ is, the larger the penalty. Intuitively, by minimizing F for ever larger values of σ, we should be able to drive the solution toward a point where g is zero and f is minimized. This can be used iteratively to seek numerical estimates for contrained minima. It can also be used to justify the Lagrange multipliers technique, and variants [**5**, pp. 255–261] [**12**].

Our final justification for Lagrange multipliers, like the penalty function approach, takes a dramatically different point of view from the standard gradient-based arguments. Pourciau [**13**] attributes this approach to a 1935 work of Carathéodory, and refers to it as the *Carathéodory Multiplier Rule*. It applies most generally to the case of n variables and $n - 1$ contraints, but we will consider it here in the case $n = 2$. I find it amazingly simple and beautiful.

As before, we wish to maximize $f(x, y)$ subject to $g(x, y) = 0$. However, this time we combine the two functions to define a mapping $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$, given by

$$\Phi(x, y) = (f(x, y), g(x, y)).$$

This situation is illustrated in FIGURE 1. As depicted in the figure, we assume the axes in the image plane are labeled u and v, so that the mapping has the alternate definition

$$u = f(x, y)$$
$$v = g(x, y).$$

Now observe that the constraint set $g = 0$ is characterized precisely as the set of points mapped by Φ to the u axis. Therefore, maximizing f subject to $g = 0$ corresponds to finding the point on the u axis in the range of Φ that is as far to the right as possible. If the constrained optimization problem has a solution at (x^*, y^*), then $(u^*, 0) = \Phi(x^*, y^*)$ must be on the boundary of the range. Otherwise, it would be possible to go even further to the right.

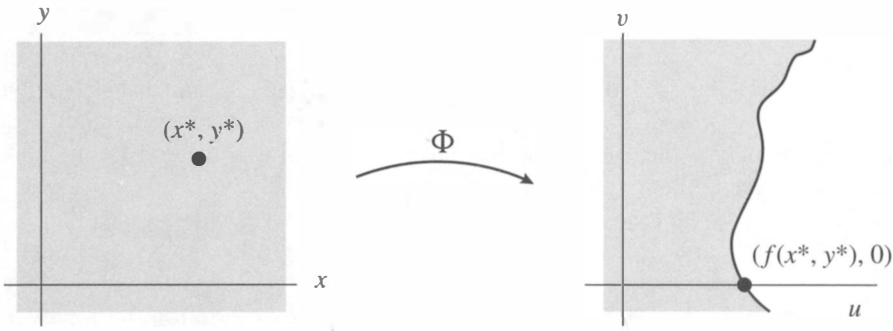


Figure 1 Action of the mapping Φ

Next, consider the derivative of Φ, which can be expressed as the matrix

$$d\Phi(x, y) = \begin{bmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \\ \dfrac{\partial g}{\partial x} & \dfrac{\partial g}{\partial y} \end{bmatrix}.$$

Whenever $d\Phi(x, y)$ is invertible (or nonsingular), it is known from the theory of differentiable functions that Φ maps a neighborhood of (x, y) to a neighborhood of $\Phi(x, y)$. In particular, that would put $\Phi(x, y)$ in the interior of the range. Thus, since $\Phi(x^*, y^*)$ is on the boundary of the range, we conclude that $d\Phi(x^*, y^*)$ is singular. That means the matrix has dependent rows, so $\nabla f(x^*, y^*)$ and $\nabla g(x^*, y^*)$ are parallel.

While this last justification is beautiful and elegant, it depends on a fairly well developed mathematical sophistication, at the level of an advanced calculus course say. The earlier arguments based on properties of the gradient are more accessible, and would be suitable for a multivariable calculus class serving math, science, and engineering students. However, in courses in mathematical methods for economics and applied or business calculus, many of which do not emphasize vector methods, the foregoing justifications for Lagrange multipliers may be out of reach. Perhaps this contributes to the popularity of the Lagrangian function formulation for these classes.

For these students, it is easy to understand the attraction of the transformation fallacy. It provides a memorable rationale for the Lagrange multipliers technique, builds on prior knowledge of unconstrained optimization, and in the end leads to correctly constituted necessary conditions for the solution. Often, it is presented with no justification beyond the vague idea that incorporating g into the objective function F implicitly imposes the constraint on the optimization process. But why should that be? Does the manner of incorporating g into F matter at all? Is there a specific argument to show why a maximum of F has anything to do with a constrained maximum of the original function f?

I have seen a justification in more than one source along the following lines. Because $\partial F/\partial\lambda = g$, a maximum of F will have to occur at a point where $g = 0$. Thus, maximizing F implicitly imposes the constraint condition. At the same time, if $F(x^*, y^*, \lambda^*)$ is a local maximum of F, then it certainly is greater than or equal to the values assumed by F at all nearby points, including those where $g = 0$. So, since F and f are identical for the points where $g = 0$, we have actually found a local maximum of f among such points.

Although this is all correct, it suffers from two major flaws. First, it turns out that trying to maximize F is destined to fail, F will not have any local maxima. Consequently, the strategy of solving the original problem by finding the maximum of F loses much of its appeal.

The second flaw is more subtle, and depends on a level of sophistication that is rare among calculus students. Correctly understood, the Lagrange multipliers technique provides a *necessary* condition for a solution of a constrained optimization problem. This justifies how the technique is usually applied. We find all solutions to the Lagrange conditions, and then choose among them the point that solves our original problem. This is valid because the Lagrange conditions are necessary: They must be satisfied by the solution of the opimization problem.

There is one logically correct way to justify a necessary condition. You must assume that a solution to the original problem is given, and then show that it also satisfies the necessary conditions. In this light, we see that the rationale above for the transformation fallacy is exactly backward. It begins with a point that is assumed to be a local extremum of F (the proposed necessary condition), and then tries to argue that such a point is a local constrained maximum of f. Even if F had local extrema (and it doesn't), there would be no assurance that they include *all* local constrained maxima of f. In particular, there is no assurance that maximizing F will find the global constrained maximum of f, which is what we ultimately wish to find.

An example The preceding argument refutes one proposed justification for the transformation fallacy, but that is not enough to establish that it is indeed a fallacy. There-

fore, let us consider an example. Specifically, we wish to see that an extremum of a function f subject to a constraint need not correspond to an unconstrained local extremum of the Lagrangian function F. Indeed, any example will do, because the Lagrangian function (essentially) *never* possesses any local maxima or minima.

With that in mind, let us consider the following perfectly pedestrian application of Lagrange multipliers. The problem is to find the point of the curve $xy = 1$ that is closest to the origin. In the standard formulation, we must minimize $f(x, y) = x^2 + y^2$ subject to the constraint $g(x, y) = 0$, where $g(x, y) = xy - 1$. By symmetry, we may restrict our attention to (x, y) in the first quadrant. The Lagrange multiplier conditions (1) become

$$\begin{aligned} 2x + \lambda y &= 0 \\ 2y + \lambda x &= 0 \\ xy &= 1. \end{aligned}$$

The only solution is $(x^*, y^*, \lambda^*) = (1, 1, -2)$. Geometric considerations show that $(1, 1)$ is indeed the closest point of the curve $xy = 1$ to the origin.

Now we ask, does $F(x, y, \lambda) = x^2 + y^2 + \lambda(xy - 1)$ have a local extremum at $(1, 1, -2)$? To see that it does not, it will be enough to show that the restriction of F to a particular plane through $(1, 1, -2)$ has a saddle point there. Then any neighborhood of $(1, 1, -2)$ includes points where F assumes both greater values and lesser values than $F(1, 1, -2)$, which can therefore be neither a local minimum nor a local maximum.

Consider the plane generated by unit vectors parallel to $\nabla g(x^*, y^*) = (1, 1)$ and the λ axis. In (x, y, λ) coordinates, the unit vectors are $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2}, 0)$ and $\mathbf{v} = (0, 0, 1)$. Therefore, a point on the plane may be expressed parametrically as

$$\begin{aligned} \mathbf{r}(s, t) &= (1, 1, -2) + s\mathbf{u} + t\mathbf{v} \\ &= (1 + s/\sqrt{2}, 1 + s/\sqrt{2}, t - 2). \end{aligned}$$

Note that $(s, t) = (0, 0)$ corresponds to the point $(1, 1, -2)$, where ∇F is zero.

To restrict F to the plane, we define $h(s, t) = F(\mathbf{r}(s, t))$. We know that h will have a critical point at $(0, 0)$. To see whether this is a local extremum, we use the second derivative test [**14**, p. 794]. First, compute the composition $F(\mathbf{r}(s, t))$ to obtain

$$h(s, t) = (\sqrt{2}s + s^2/2)t + 2.$$

The Hessian matrix of second partial derivatives,

$$\begin{bmatrix} \dfrac{\partial^2 h}{\partial s^2} & \dfrac{\partial^2 h}{\partial s\, \partial t} \\ \dfrac{\partial^2 h}{\partial s\, \partial t} & \dfrac{\partial^2 h}{\partial t^2} \end{bmatrix} = \begin{bmatrix} t & s + \sqrt{2} \\ s + \sqrt{2} & 0 \end{bmatrix},$$

has determinant -2 at $(s, t) = (0, 0)$. Since it is negative, we know that $h(s, t)$ has a saddle point at the critical point $(0, 0)$. This shows that $F(x, y, \lambda)$ cannot have a local maximum or minimum at $(1, 1, -2)$. In fact, $(1, 1, -2)$ and $(-1, -1, -2)$ are the only critical points of F, which therefore has *no* local maxima or minima.

This reveals the flaw in the transformation fallacy, which tells us to solve the constrained optimization problem by finding the unconstrained minimum of F. That is impossible for no minimum exists. Nevertheless, the constrained problem has a solution.

A theorem

What happened in the example always happens. For any objective function and constraint, the restriction of the Lagrangian function to a plane parallel to $\nabla g(x^*, y^*)$ and the λ axis has a saddle point at (x^*, y^*, λ^*), as shown in the following theorem. Although stated for functions of two variables, it extends naturally to functions of n variables.

First, we will establish a few notational conventions. Partial derivatives will be expressed using subscript notation, so f_x is the partial derivative of f with respect to x and f_{xy} is the second partial derivative with respect to x and y. We will abbreviate $\Phi(x^*, y^*)$ by Φ^* for any function Φ. In particular, $f^* = f(x^*, y^*)$, $\nabla f^* = \nabla f(x^*, y^*)$, and $f^*_{xx} = f_{xx}(x^*, y^*)$.

THEOREM. *Let f and g be functions of two variables, with continuous second derivatives. Let $F(x, y, \lambda) = f(x, y) + \lambda g(x, y)$. Then, if (x^*, y^*, λ^*) is a critical point of F at which ∇g^* is not the zero vector, F must have a saddle point at (x^*, y^*, λ^*).*

Proof. At any critical point (x^*, y^*, λ^*), all of the first partial derivatives of F vanish. This shows that $g^* = 0$ and $\nabla f^* + \lambda^* \nabla g^* = 0$. We also assume that ∇g^* is not zero.

Without loss of generality, we may assume $(x^*, y^*) = (0, 0)$ and ∇g^* points in the direction of the x axis: these conditions may be brought about by translating the x-y plane and rotating it about the λ-axis, neither of which alters the character of (x^*, y^*, λ^*) as a saddle point (or not). With these assumptions, $\nabla g^* = (g^*_x, g^*_y) = (a, 0)$ for some $a \neq 0$.

Next we will consider the restriction of F to the plane $y = 0$ in (x, y, λ) space. Let

$$h(x, \lambda) = F(x, 0, \lambda) = f(x, 0) + \lambda g(x, 0).$$

Then

$$\nabla h(x, \lambda) = (f_x(x, 0) + \lambda g_x(x, 0), g(x, 0)).$$

This vanishes at $(x, \lambda) = (0, \lambda^*)$, so $(0, \lambda^*)$ is a critical point of h.

We show that $(0, \lambda^*)$ is a saddle point of h by using the second derivative test. We need the Hessian matrix of second partial derivatives, which is defined by

$$H(x, \lambda) = \begin{bmatrix} h_{xx} & h_{x\lambda} \\ h_{x\lambda} & h_{\lambda\lambda} \end{bmatrix} = \begin{bmatrix} f_{xx}(x, 0) + \lambda g_{xx}(x, 0) & g_x(x, 0) \\ g_x(x, 0) & 0 \end{bmatrix}.$$

At the critical point, the determinant of the Hessian is $\det H(0, \lambda^*) = -(g^*_x)^2$, and this is negative because we know $g^*_x \neq 0$. Therefore, the second derivative test shows that h has a saddle point at $(0, \lambda^*)$. Hence F must have a saddle point at (x^*, y^*, λ^*), as asserted. ∎

The theorem tells us that the Lagrangian function approach has *nothing* to do with finding a local maximum of F, and in practice we know that is true. Rather, *all* the critical points of F are found, and these are further examined to find the constrained maximum of f. Along the way, no one ever checks to see which critical points are local maxima of F. Otherwise, the transformation fallacy could not possibly persist. And indeed, specialists in optimization consider the preceding theorem common knowledge. I've consulted several economics faculty who specialize in mathematical methods, and

they were all well aware of the result. But the result deserves to be better known among mathematicians, particularly the nonspecialists who teach Lagrange multipliers in calculus classes.

The theorem also crystalizes in a dramatic way what is needed in an intuitive justification of the Lagrangian approach. Because F essentially never has local extrema, a proper justification must explain why the critical points of F (and in fact, really why the *saddle* points of F) are significant. And it must do so without assuming that F is maximized somewhere. The transformation fallacy does not provide such a justification. Shortly we will see an intuitive argument that does. First, though, we take a brief look at the history of the Lagrange multiplier technique.

What did Lagrange say?

In the original development of the multiplier method by Lagrange, it is the Lagrangian function approach that appears, although without considering the multipliers to be independent variables of the function. What I have called the standard approach is a later development.

Interestingly, Lagrange did not initially formulate the multiplier method in the context of constrained optimization, but rather in the analysis of equilibria for systems of particles. He reported this application in *Mécanique Analytique*, published in 1788 and available now in English translation [**9**]. Using series expansions to analyze the effects of perturbation at the point of equilibrium, Lagrange derives conditions on the first order differentials for systems under very general assumptions. In the case that the particle motions are subject to external constraints, he points out that the constraints can be used to eliminate some of the variables that appear in his differential equations, before deriving the conditions for equilibrium. He goes on to observe that

> the same results will be obtained if the different equations of [constraint], each multiplied by an undetermined coefficient are simply added to [the general formula of equilibrium]. Then, if the sum of all the terms that are multiplied by the same differential are put equal to zero, as many particular equations as there are differentials will be obtained [**9**, p. 60].

Next, he explains how this procedure can be "treated as an ordinary equation of maxima and minima." However, this last statement must be understood in the context of Lagrange's earlier remarks, where he is careful to point out that "the equation of a differential set equal to zero does not always represent a maximum or minimum [**9**, p. 55]." Thus, Lagrange must have intended to convey that the conditions rigorously derived by his perturbation analysis could be found by formulating a Lagrangian function, and proceeding as if seeking a maximum or minimum. But there is no suggestion that the equilibrium point must actually correspond to a maximum or minimum in general, nor that the existence of such an extremum is necessary to establish the validity of the equilibrium conditions.

Having developed the multiplier technique in the analysis of equilibria, Lagrange proceeded to use it in other settings, notably in the calculus of variations [**4**]. Today's familiar application to constrained optimization problems was presented in two pages in his *Théorie des Fonctions Analytiques* of 1797, nearly a decade after the initial work with statics. As in the earlier work, Lagrange first establishes the conditions for a constrained extremum using series expansions. Then he points out that the conditions so described can be obtained according to a general principle [**10**, p. 198]. In translation, that principle runs as follows:

> When a function of several variables must be a maximum or a minimum, and there are one or more equations relating these variables, it will suffice to add to the proposed function the functions that must be equal to zero, multiplied each one by an indeterminate quantity, and next to find the maximum or minimum as if these variables were independent; the equations that one will find combined with the given equations will serve to determine all the unknowns.

On casual inspection, this appears to be a statement of the transformation fallacy. However, two points are significant. First, in saying to find the extremum as if the variables were independent, Lagrange clearly does not consider these variables to include the multipliers. This is reflected not only by the context in which he uses the phrase *these variables,* but also by his inclusion of the phrase *combined with the given equations.* If the multipliers were considered as variables, there would be no need to separately mention the constraint equations, which would appear as partial derivatives with respect to the multipliers. Second, consistent with my earlier remarks, it seems evident that Lagrange only advised proceeding *as if* seeking a maximum or minimum, and that the key point is the assertion of the final clause: the variables can be found by solving the given set of equations. Remember, too, that the validity of that assertion, established in a separate argument, did not depend on the existence of an unconstrained extremum at the solution point.

Apparently the idea of considering the multipliers as new variables for the optimization problem originated elsewhere, probably when someone recognized that the partial derivatives with respect to the multipliers are exactly the quantities that the constraints hold at zero. To a reader with that observation in mind however, it is easy to imagine that the quoted passage above, taken out of context, might encourage a belief in the transformation fallacy.

Lagrangian leveling We come at last to my proposed rationale for the Lagrangian function approach. The transformation fallacy says that the Lagrangian function is a device by which a constrained problem becomes an unconstrained problem. An equally memorable idea, and one that is correct, is that the Lagrangian is an approach that *levels the playing field.*

To make this idea clear, let us consider once again the fundamental problem, to maximize $f(x, y)$ subject to $g(x, y) = 0$. Suppose that the solution point occurs at (x^*, y^*). If we are very lucky, this will actually be a local maximum of f disregarding the constraint. In this case the graph of f is a surface G with a high point at $(x^*, y^*, f(x^*, y^*))$, so the tangent plane is horizontal there. That is why we can find such points by setting the partial derivatives equal to 0.

Usually, though, this is not what occurs in a constrained problem. As portrayed in FIGURE 2, we can only consider points (x, y) on a curve C in the domain. The graph of the restricted function $f|_C$ is a curve G_C on surface G. The constrained maximum occurs at a high point of G_C but it is not a high point of G, and the tangent plane of G is not horizontal there. Traveling along the plane in the direction of G_C we would experience a slope of 0, because the curve has a high point, but traveling off of the curve we can go even higher. In particular, traveling on the tangent plane perpendicular to G_C the slope is not 0.

In order to rectify this problem, we would like to flatten out the graph of f, so that the tangent plane becomes horizontal. In the process, we do not want to disturb G_C, because that might change the location of the constrained maximum. But it should be perfectly alright to alter the values of the function at points that are not on the constraint curve. In fact, near the high point of G_C, we can imagine pivoting the graph of f around the curve's horizontal tangent line. If we pivot by just the right amount, the

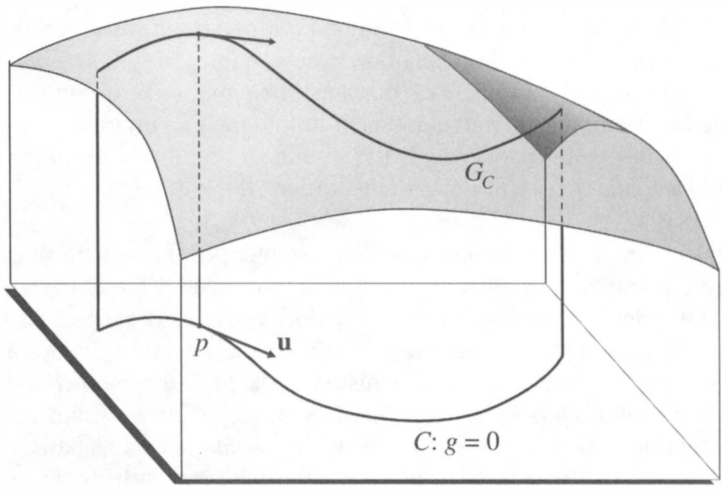


Figure 2 Constraint curve C in the domain of f, the graph G of f, and the graph C_G of $f|_C$

tangent plane will become horizontal, and so detectable by setting partial derivatives to zero. That is the image of leveling the playing field.

How do we put this into effect analytically? The simplest way to alter the values of the function f is to add some perturbation. But we want the amount we add to be 0 along the constraint curve, so as not to alter f for those points. On the other hand, we know that g is 0 along the constraint curve. So, adding g to f is the sort of perturbation we need. It leaves the values of f unchanged along the constraint curve, but modifies the values away from that curve. More generally, we can add any multiple of g to f and achieve the same effect. That is the motivation behind defining for each λ a perturbed function

$$F(x, y) = f(x, y) + \lambda g(x, y).$$

In this conception, we do not think of F as a function of three variables. Rather, we have in mind an entire family of two variable functions, one for each value of the parameter λ. A representation of this situation is shown in FIGURE 3. The surface marked $\lambda = 0$ is the unperturbed graph of f. The other surfaces are graphs of different members of the family of F functions. All of the surfaces intersect in the graph of f over the constraint curve, so the constrained maximum is the high point on the intersection curve.

The figure shows that choosing different values of λ imposes just the sort of pivoting action described earlier. Intuition suggests (and a little further analysis proves) that there must actually be a choice of λ that makes the tangent plane horizontal. That is, if (x^*, y^*) is the location of the maximum value of f subject to the constraint $g = 0$, then for some λ it will be the case that $\nabla F(x^*, y^*) = 0$. In that case, the following equations must hold:

$$\begin{aligned} \frac{\partial f}{\partial x}(x^*, y^*) + \lambda \frac{\partial g}{\partial x}(x^*, y^*) &= 0 \\ \frac{\partial f}{\partial y}(x^*, y^*) + \lambda \frac{\partial g}{\partial y}(x^*, y^*) &= 0 \\ g(x^*, y^*) &= 0. \end{aligned}$$

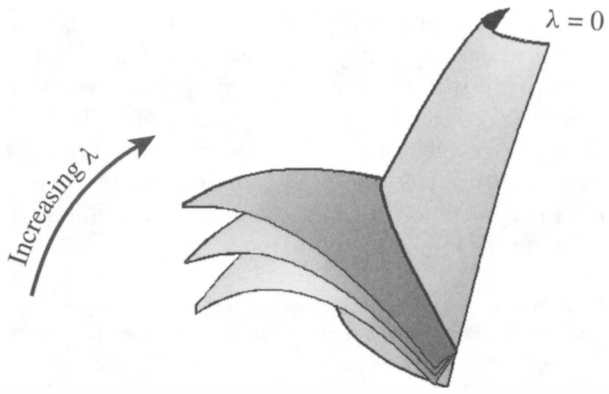


Figure 3 Graphs of several members of the family of F functions

By finding every possible triple (x, y, λ) for which these equations hold, we obtain a candidate set that must contain the solution to the constrained optimization problem. That is what the Lagrange multipliers theorem says.

With the concept of *leveling*, we can also rederive an interpretation of λ at the solution point, popular in the optimization literature for applications in economics [**2**, pp. 376–377], [**6**, p. 611]. At the maximum point over the constraint curve, we know that the directional derivative $D_{\mathbf{u}}f$ in the direction of the curve is zero. And we are assuming that the derivative $D_{\mathbf{u}}f$ in the direction normal to the constraint curve is not 0.

To level the tangent plane, we need to choose λ so that F will have zero directional derivative normal to the curve. That is, for $\mathbf{u}$ normal to the curve, we want $D_{\mathbf{u}}(f + \lambda g) = 0$. Clearly, we need to choose $\lambda = -D_{\mathbf{u}}f/D_{\mathbf{u}}g$. This shows, by the way, when λ exists. We may take for $\mathbf{u}$ the unit vector in the direction of ∇g. Then $D_{\mathbf{u}}g = \nabla g \cdot \mathbf{u} = |\nabla g|$. In particular, $D_{\mathbf{u}}g$ can only be 0 if $\nabla g = 0$. Otherwise, for $\lambda = -D_{\mathbf{u}}f/D_{\mathbf{u}}g$, the modified function F will have a horizontal tangent plane.

But there is also a meaningful interpretation of this choice for λ. The ratio $D_{\mathbf{u}}f/D_{\mathbf{u}}g$ is the rate of change of the objective function f relative to a change in the constraint function g. It shows, for a given perturbation of the point (x^*, y^*) orthogonally away from the constraint curve, how the change in f compares to the change in g. The economists interpret this as the marginal change in the objective function relative to the constraint. It indicates to first order, how the maximum value will change under relaxation or tightening of the constraint.

Concluding remarks There is a vast literature on optimization. It covers in great depth ideas that have been barely touched upon here, and far more. A good general reference most closely related to the topics discussed in this paper is Hestenes [**5**]. This provides a general discussion of Lagrange multipliers, including the transformation of constrained to unconstrained problems through a process called *augmentation*, as well as a detailed account of penalty functions. Although Hestenes does not use the terminology of *leveling*, this idea is implicit in his treatment of *augmentability*. For a general account of multiplier rules, I highly recommend Pourciau [**13**] (winner of a Ford award).

An expanded version of the present paper appears in [**8**]. Among other things, it gives examples of textbooks that encourage a belief in the transformation fallacy, implicitly or explicitly, as well as several geometric arguments in support of the standard approach to Lagrange multipliers.

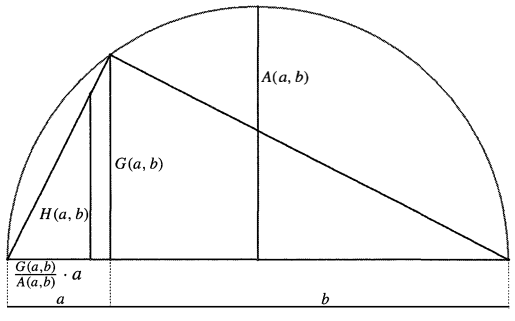
The image of leveling the graph of f using the Lagrangian function F offers an attractive way to think about Lagrange multipliers. In this formulation, we find a clear and convincing intuitive justification of the Lagrangian approach. It also makes the existence of the necessary λ completely transparent, while leading naturally to the interpretation of λ as the marginal rate of change of the objective function relative to the constraint. This has the potential of enriching the insight of students, particularly those interested in the applications of mathematics to economics. And it is certainly superior to perpetuating the transformation fallacy.

Acknowledgment. Bruce Pourciau contributed many ideas that improved this paper; Ed Barbeau shared a wealth of material about Lagrange multipliers; Jennifer Kalman Beal generously provided a translation of Lagrange's description of his multiplier method [**10**].

REFERENCES

1. G. C. Archibald and Richard G. Lipsey, *An Introduction to Mathematical Economics: Methods and Applications,* Harper and Row, New York, 1976.
2. Alpha C. Chiang, *Fundamental Methods of Mathematical Economics,* 3rd ed., McGraw-Hill, New York, 1984.
3. Frank A. Farris, private communication, July 2008.
4. C. G. Fraser, Isoperimetric problems in the variational calculus of Euler and Lagrange, *Historia Mathematica* **19**(1) (1992) 4–23.
5. Magnus R. Hestenes, *Optimization Theory the Finite Dimensional Case,* Wiley, New York, 1975.
6. Michael Hoy, John Livernois, Chris McKenna, Ray Rees, and Thanasis Stengos, *Mathematics for Economics,* 2nd ed., MIT Press, Cambridge, MA, 2001.
7. Deborah Hughes-Hallett, Andrew M. Gleason, William G. McCallum, et al., *Calculus Single and Multivariable,* 2nd ed., Wiley, New York, 1998.
8. Dan Kalman, *Uncommon Mathematical Excursions: Polynomia and Related Realms,* MAA, Washington, DC, 2009.
9. J. L. Lagrange, *Analytical Mechanics,* translated from the Mécanique analytique, nouvelle édition of 1811 by Auguste Boissonnade and Victor N. Vagliente, Kluwer, Dordrecht, 1997.
10. J. L. Lagrange, *Théorie des fonctions analytiques contenant les principes du calcul différentiel, dégagés de toute considération d'infiniment petits ou d'évanouissans, de limites ou de fluxions, et réduits à l'analyse algébrique des quantités finies*, De l'imprimerie de la République, Paris, 1797.
11. Ronald E. Larson, Robert P. Hostetler, and Bruce H. Edwards, *Calculus with Analytic Geometry,* 6th ed., Houghton Mifflin, Boston, 1998.
12. E. J. McShane, The Lagrange multiplier rule, *Amer. Math. Monthly*, **80**(8) (1973) 922–925.
13. B. H. Pourciau, Modern multiplier rules, *Amer. Math. Monthly*, **87**(6) (1980) 433–452.
14. James Stewart, *Calculus Early Transcendentals,* 3rd ed., Brooks/Cole, Pacific Grove, CA, 1995.
15. George B. Thomas, Jr., Maurice D. Weir, Joel D. Hass, and Frank R. Giordano, *Thomas' Calculus Early Transcendentals,* 10th ed., Addison-Wesley, Boston, 2001.

Editor's Note: Unfortunately, an error crept into our April issue (**82**:2, p. 116). In the Proof Without Words: Ordering Arithmetic, Geometric, and Harmonic Means, the vertical line segment labeled $H(a,b)$ should stop at the hypotenuse of the triangle, and not go all the way up to the circle. A corrected image is shown here. We regret the error.



NOTES

Quartic Polynomials and the Golden Ratio

HARALD TOTLAND
Royal Norwegian Naval Academy
P.O. Box 83 Haakonsvern
N-5886 Bergen, Norway
harald.totland@sksk.mil.no

Suppose we have a pentagram in the xy-plane, oriented as in FIGURE 1a, and want to find a quartic polynomial whose graph passes through the three vertices indicated. Out of infinitely many possibilities, there is exactly one quartic polynomial that attains its minimum value at both of the two lower vertices. This graph—shaped like a smooth W with its local maximum at the upper vertex—is shown in FIGURE 1b. Now, how does the graph continue? Will it touch the pentagram again on its way up to infinity? As it turns out, the graph passes through two more vertices, as shown in FIGURE 1c. Furthermore, the two points where the graph crosses the interior of a pentagram edge lie exactly below two other vertices, as shown in FIGURE 1d.

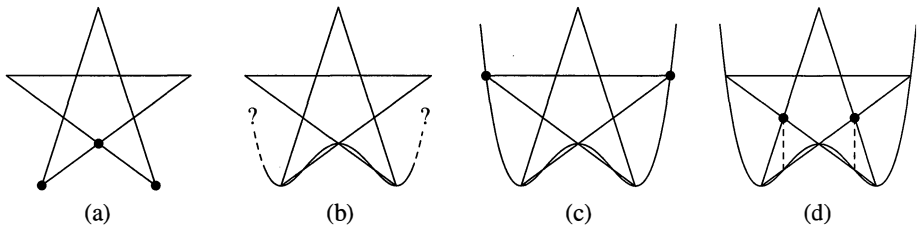


Figure 1 A pentagram and a quartic polynomial

Knowing that length ratios within the pentagram are determined by the golden ratio, we realize that this quartic polynomial has some regularities governed by the same ratio. As we will see, there are many more such regularities. Furthermore, they apply to *all* quartic polynomials with inflection points. This will be clear once we realize that the different quartics are all related by an *affine transformation.*

Symmetric quartic We investigate graphs of quartic polynomials with inflection points by means of certain naturally defined points and length ratios. As an example, we consider the function $f(x) = x^4 - 2x^2$, shown in FIGURE 2. (This quartic's shape differs slightly from the one in FIGURE 1 and is chosen to simplify calculations.) We define $P_0(x_0, y_0)$ as the point where the third derivative vanishes, so that $f'''(x_0) = 0$. The tangent points of the double tangent (the unique line that is tangent to the graph at two points) are called P_1 and P_2. The points where the tangent at P_0 intersects the graph are P_3 and P_4. We number points so that those to the left of P_0 have odd index, while those to the right have even index.

The line through P_0 and P_1 intersects the graph in two additional points, called P_6 and P_7. Similarly, the line through P_0 and P_2 has the additional intersection points P_5

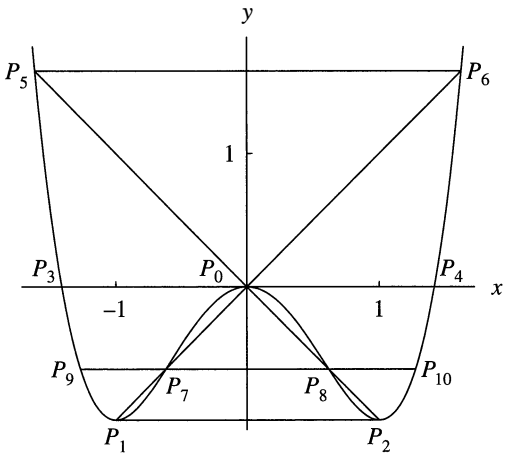


Figure 2 The quartic $x^4 - 2x^2$

and P_8. (The inflection points might appear to be P_7 and P_8, but this is not so.) The line through P_7 and P_8 intersects the graph in P_9 and P_{10}.

For this graph, we easily find the coordinates $P_0(0, 0)$, $P_2(1, -1)$, and $P_4(\sqrt{2}, 0)$. Symmetry guarantees that the points P_1 through P_{10} are located symmetrically about the y-axis. The coordinates of P_5 through P_{10} turn out to involve the golden ratio, $\varphi = (1 + \sqrt{5})/2$. Using the relations $\varphi^2 = \varphi + 1$ and $\varphi^{-2} = 1 - \varphi^{-1}$, we calculate three function values: $f(\varphi) = \varphi$ and $f(\varphi^{-1}) = f(\sqrt{\varphi}) = -\varphi^{-1}$. From these calculations and the fact that the points P_5 through P_8 lie on the lines $y = \pm x$, we find the coordinates $P_6(\varphi, \varphi)$, $P_8(1/\varphi, -1/\varphi)$, and $P_{10}(\sqrt{\varphi}, -1/\varphi)$. From these coordinates, the following relations between line segment lengths follow quickly:

$$P_3P_4 = \sqrt{2}P_1P_2, \quad P_5P_6 = \varphi P_1P_2, \quad P_7P_8 = P_1P_2/\varphi, \quad P_9P_{10} = \sqrt{\varphi}P_1P_2. \tag{1}$$

Our next step is to show that these relations carry over to the general case.

General quartic We will *not* proceed by deriving general expressions for the coordinates of the points P_0 through P_{10}. Instead, we shall see that the graph of every quartic polynomial with inflection points can be obtained as the image of the graph of the symmetric quartic above subject to an appropriate affine transformation. An affine transformation consists of an invertible linear transformation followed by translation along a constant vector. An affine transformation of the plane has the following properties: Straight lines are mapped to straight lines, parallel lines to parallel lines, and tangents to tangents, while length ratios between parallel line segments are preserved [**1**, chapter 2].

Consider the symmetric quartic

$$f(x) = x^4 + wx^2, \quad w < 0, \tag{2}$$

and a general quartic with inflection points,

$$g(x) = ax^4 + bx^3 + cx^2 + dx + e, \quad a \neq 0.$$

Define x_0 by $g'''(x_0) = 0$ (so $x_0 = -b/4a$) and $k = \sqrt{g''(x_0)/2aw}$. (The existence of two inflection points implies that $g''(x_0)$ and a have opposite signs, and since w is

negative, k is real.) The map $(x, y) \mapsto (\bar{x}, \bar{y})$ given by

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ g'(x_0) & 1 \end{pmatrix} \begin{pmatrix} k & 0 \\ 0 & ak^4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x_0 \\ g(x_0) \end{pmatrix} \tag{3}$$

is an affine transformation consisting of a scaling transformation (with a scaling factor k in the x-direction and a scaling factor $|a|k^4$ in the y-direction), a reflection about the x-axis if a is negative, a shear in the y-direction, and a translation. In components, we have the equations

$$\bar{x} = kx + x_0, \quad \bar{y} = ak^4 y + g'(x_0)kx + g(x_0).$$

Suppose (x, y) lies on the graph of f, that is, $y = f(x)$. Substituting $y = x^4 + wx^2$ and $x = (\bar{x} - x_0)/k$ into the expression for $\bar{y}$ yields

$$\bar{y} = a(\bar{x} - x_0)^4 + \frac{1}{2}g''(x_0)(\bar{x} - x_0)^2 + g'(x_0)(\bar{x} - x_0) + g(x_0).$$

The right-hand side is the fourth-degree Taylor polynomial of g at x_0 and is therefore identical to $g(\bar{x})$. Thus, $(\bar{x}, \bar{y})$ lies on the graph of g, so the above transformation indeed maps the graph of f to the graph of g.

Moreover, the origin, where $f'''(x) = 0$, is mapped to $(x_0, g(x_0))$, where $g'''(x) = 0$. From this and the general properties of affine maps it follows that each of the points P_0 through P_{10} on the graph of f is mapped to the analogously defined point on the graph of g. (We will use the same notation P_i for points on both graphs.) The results for the case $w = -2$ then show that the line segments $P_1P_2, P_3P_4, \ldots, P_9P_{10}$ on the graph of g are all parallel, are bisected by the vertical line through P_0, and satisfy the relations (1). FIGURE 3 illustrates this for the quartic $2x^4 - x^3 - 2x^2 + x + 1$. Note that P_0 divides the line segments P_6P_1 and P_5P_2 according to the golden ratio, P_7 divides P_0P_1 according to the golden ratio, and analogously for P_8 and P_0P_2.

Further characteristic ratios We now define some more points on the graph of a quartic, starting with the point P_0 from FIGURE 3 and the inflection points P_{11} and P_{12};

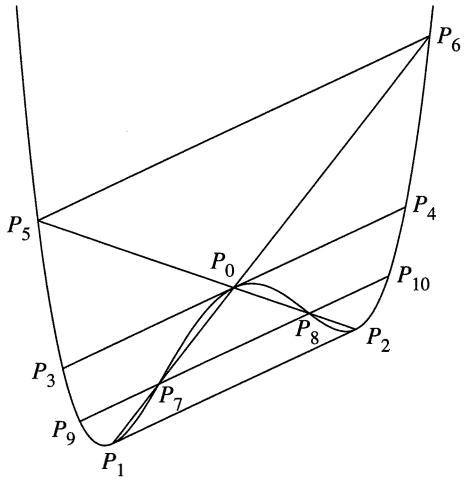


Figure 3 The quartic $2x^4 - x^3 - 2x^2 + x + 1$ and the points P_0 through P_{10} (axes not shown)

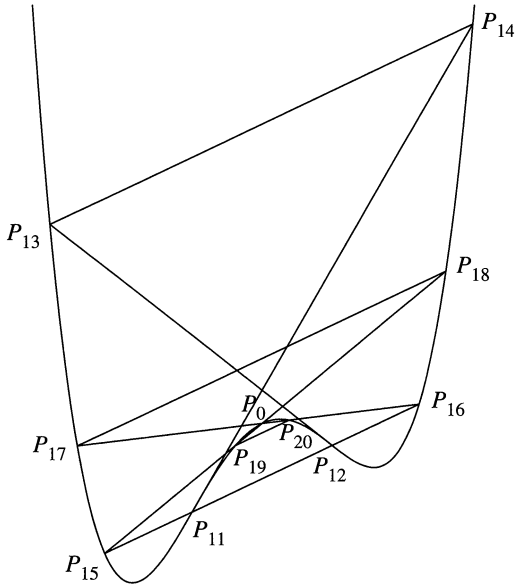


Figure 4 The quartic $2x^4 - x^3 - 2x^2 + x + 1$ and the points P_{11} through P_{20}

see FIGURE 4. The tangents at the inflection points intersect the graph at the points P_{13} and P_{14}. The line through the inflection points intersects the graph at P_{15} and P_{16}. The line through P_0 and P_{15} intersects the graph at P_{18} and P_{19}, while the line through P_0 and P_{16} intersects the graph at P_{17} and P_{20}. The list of statements may now be extended:

THEOREM. *Let $P_0, \ldots, P_{20}$ be points defined as above on the graph of a quartic polynomial with inflection points, and $\varphi = (\sqrt{5}+1)/2$. Then:*

1. *The line segments $P_{2n-1}P_{2n}$ $(n = 1, \ldots, 10)$ are all parallel.*
2. *Intersection points of the graph and a line parallel to the tangent in $P_0(x_0, y_0)$ are symmetrically located about the point on the line with $x = x_0$.*
3. $P_3P_4 = \sqrt{2}P_1P_2$
4. $P_5P_6 = \varphi P_1P_2$
5. $P_7P_8 = P_1P_2/\varphi$
6. $P_9P_{10} = \sqrt{\varphi}P_1P_2$
7. $P_{11}P_{12} = P_1P_2/\sqrt{3}$
8. $P_{13}P_{14} = 3P_{11}P_{12}$
9. $P_{17}P_{18} = \varphi^2 P_{11}P_{12}$
10. $P_{15}P_{11} = P_{12}P_{16} = P_{11}P_{12}/\varphi$
11. $P_{19}P_{20} = P_{11}P_{12}/\varphi^2$

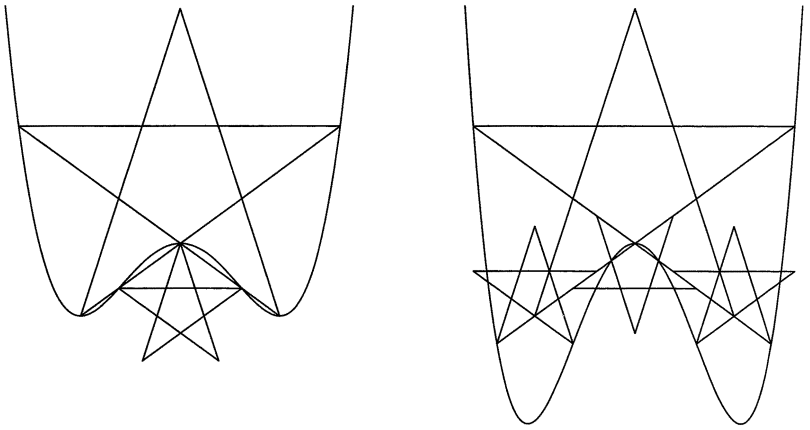
Proof. The new statements (2, 7–11, and part of 1) may be verified relatively easily for the quartic $x^4 - 6x^2$, that is, $f(x)$ from (2) with $w = -6$. (Since then $f''(\pm 1) = 0$, this quartic is simpler to use for P_{11} through P_{20} than $x^4 - 2x^2$.) Now, inflection points are mapped to inflection points by an affine map. (Indeed, $g''(\bar{x}) = ak^2 f''(x)$ in our case.) Then, the arguments made earlier apply here as well. ■

The properties of the line through the inflection points P_{11} and P_{12} (statement 10, and in part 1) have been pointed out earlier [**2**], as has the symmetry property (statement 2) and the fact that P_0 is the point where the tangent is parallel to the double

tangent (a consequence of statement 1) [2, 3]. I have found no reference to the other relations, including, in particular, the five occurrences of the golden ratio.

Our affine transformation (3) shows that the graph of a general quartic function may be regarded as an originally symmetric graph that has been sheared in the y-direction and moved. Considering this, the properties regarding parallelism and symmetry (for instance, that the line segments $P_{15}P_{11}$ and $P_{12}P_{16}$ have equal length) become obvious. The same applies to ratios between areas, since a scaling transformation changes all areas by a constant factor, while a shear preserves areas. For example, it is known that the line through the inflection points of an arbitrary quartic function cuts off three areas that are in the ratio of $1 : 2 : 1$. This is readily verified for $f(x)$ from (2) with $w = -6$ by checking that $\int_0^{\sqrt{5}} (f(x) - (-5))\,dx = 0$, whereby it is immediately proven generally.

Quartic polynomials and pentagrams Returning to FIGURE 1, we see that it is just an example of statements 4 and 5 of the theorem. The same can be illustrated by FIGURE 5a, where the smaller pentagram fits exactly into the inner pentagon of the larger pentagram (meaning the linear size ratio is $1 : \varphi^2$). Similarly, as the reader may check, FIGURE 5b illustrates statements 9, 10, and 11. In each of these graphs, three points are given, two of which are specified as minimum points, (a), or inflection points, (b). This completely determines the graphs; they will automatically pass through four or six more vertices. The possibility of finding such simple constellations of pentagrams and quartic graphs reflects the occurrence of the golden ratio in quartic polynomials.



(a) (b)

Figure 5 Quartic polynomials passing through pentagram vertices

To summarize, we have found simple characteristic length ratios on the graph of a quartic polynomial with inflection points, including several occurrences of the golden ratio. These length ratios are left invariant by an affine transformation that relates a symmetric quartic to a general quartic with inflection points.

REFERENCES

1. D. A. Brannan, M. F. Esplen, and J. J. Gray, *Geometry*, Cambridge University Press, Cambridge, 1999.
2. H. T. R. Aude, Notes on quartic curves, *Amer. Math. Monthly* **56** (1949) 165–170.
3. F. Irwin and H. N. Wright, Some properties of polynomial curves, *Annals Math.* (2nd Ser.) **19** (1917) 152–158.

When Cauchy and Hölder Met Minkowski: A Tour through Well-Known Inequalities

GERHARD J. WOEGINGER
Department of Mathematics
TU Eindhoven
The Netherlands
gwoegi@win.tue.nl

Many classical inequalities are just statements about the convexity or concavity of certain (hidden) underlying functions. This is nicely illustrated by Hardy, Littlewood, and Pólya [**5**] whose Chapter III deals with *"Mean values with an arbitrary function and the theory of convex functions,"* and by Steele [**12**] whose Chapter 6 is called *"Convexity—The third pillar."* Yet another illustration is the following proof of the arithmetic-mean-geometric-mean inequality (which goes back to Pólya). The inequality states that the arithmetic mean of n positive real numbers $a_1, \ldots, a_n$ is always greater or equal to their geometric mean:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} a_i \;\geq\; \left(\prod_{i=1}^{n} a_i \right)^{1/n}. \tag{1}$$

The substitution $x_i = \ln a_i$ shows that (1) is equivalent to the inequality

$$\frac{1}{n} \cdot \sum_{i=1}^{n} e^{x_i} \;\geq\; e^{\frac{1}{n}\sum_{i=1}^{n} x_i}. \tag{2}$$

The correctness of (2) is easily seen from the following two observations. First: $f(x) = e^x$ is a convex function. And second: Jensen's inequality [**7**] states that any convex function $f : \mathbb{R} \to \mathbb{R}$ and any real numbers $x_1, \ldots, x_n$ satisfy

$$\frac{1}{n} \cdot \sum_{i=1}^{n} f(x_i) \geq f\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right). \tag{3}$$

But we do not want to give the impression that this article is centered around *convexity* and that it perhaps deals with Jensen's inequality. No, no, no, quite to the contrary: This article is centered around *concavity*, and it deals with the Cauchy inequality, the Hölder inequality, the Minkowski inequality, and with Milne's inequality. We present simple, concise, and uniform proofs for these four classical inequalities. All our proofs proceed in exactly the same fashion, by exactly the same type of argument, and they all follow from the concavity of a certain underlying function in exactly the same way. Loosely speaking, we shall see that

Cauchy corresponds to the concave function $\sqrt{x}$,
Hölder corresponds to the concave function $x^{1/p}$ with $p > 1$,
Minkowksi to the concave function $(x^{1/p} + 1)^p$ with $p > 1$, and
Milne corresponds to the concave function $x/(1 + x)$.

Interestingly, the cases of equality for all four inequalities fall out from our discussion in a very natural way and come almost for free. Now let us set the stage for concavity and explain the general approach.

Concavity and the master theorem

Here are some very basic definitions. Throughout we use $\mathbb{R}$ and $\mathbb{R}_+$ to denote the set of real numbers and the set of positive real numbers, respectively. A function $g : \mathbb{R}_+ \to \mathbb{R}$ is *concave* if it satisfies

$$\lambda \cdot g(x) + (1 - \lambda) \cdot g(y) \leq g(\lambda x + (1 - \lambda)y) \tag{4}$$

for all $x, y \in \mathbb{R}_+$ and for all real λ with $0 < \lambda < 1$. In other words, for any x and y the line segment connecting point $(x, g(x))$ to the point $(y, g(y))$ must lie below the graph of function g; FIGURE 1 illustrates this. A concave function g is *strictly concave*, if equality in (4) is equivalent to $x = y$. A function g is *convex* (*strictly convex*) if the function $-g$ is concave (strictly concave). For twice-differentiable functions g there are simple criteria for checking these properties: A twice-differentiable function g is concave (strictly concave, convex, strictly convex) if and only if its second derivative is nonpositive (negative, nonnegative, positive) everywhere.

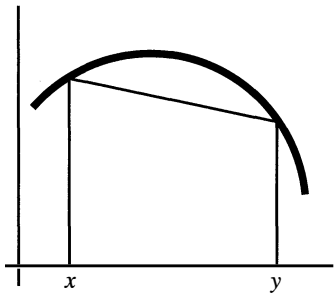


Figure 1 A concave function

Most of our arguments are based on the following theorem which we dub *the master theorem* (although admittedly, it rather is a simple observation on concavity, whose proof is only slightly longer than its statement). We would guess that the statement was known already before the Second World War, but its exact origin is unknown to us. Walther Janous pointed out to us that Godunova [**4**] used the idea in 1967.

MASTER THEOREM. *Let $g : \mathbb{R}_+ \to \mathbb{R}$ be a strictly concave function, and let $f : \mathbb{R}_+^2 \to \mathbb{R}$ be the function defined by*

$$f(x, y) = y \cdot g\left(\frac{x}{y}\right). \tag{5}$$

Then all positive real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ satisfy the inequality

$$\sum_{i=1}^{n} f(x_i, y_i) \leq f\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i\right). \tag{6}$$

Equality holds in (6) *if and only if the two sequences x_i and y_i are proportional* (*that is, if and only if there is a real number t such that $x_i/y_i = t$ for all i*).

Proof. The proof is by induction on n. For $n = 1$, the inequality (6) becomes an equation. Since the two sequences have length one, they are trivially proportional. For $n = 2$, we use the concavity of g: From (4) with $\lambda = y_1/(y_1 + y_2)$, we derive that

$$
\begin{aligned}
f(x_1, y_1) + f(x_2, y_2) &= y_1 \cdot g\left(\frac{x_1}{y_1}\right) + y_2 \cdot g\left(\frac{x_2}{y_2}\right) \\
&= (y_1 + y_2)\left\{\frac{y_1}{y_1 + y_2} \cdot g\left(\frac{x_1}{y_1}\right) + \frac{y_2}{y_1 + y_2} \cdot g\left(\frac{x_2}{y_2}\right)\right\} \\
&\le (y_1 + y_2) \cdot g\left(\frac{x_1 + x_2}{y_1 + y_2}\right) \\
&= f(x_1 + x_2, y_1 + y_2). \qquad (7)
\end{aligned}
$$

Since g is strictly concave, equality holds in this chain if and only if $x_1/y_1 = x_2/y_2$. The inductive step for $n \ge 3$ follows easily from (7), and the proof is complete. ■

Here are two brief remarks before we proceed. First, if the function g in the theorem is just concave (but not strictly concave), then inequality (6) is still valid, but we lose control over the situation where equality holds. The cases of equality are no longer limited to proportional sequences, and can be quite arbitrary. Second, if g is strictly convex (instead of strictly concave), then the inequality (6) follows with a greater-or-equal sign instead of a less-or-equal sign.

Our next goal is to derive four well-known inequalities by four applications of the master theorem with four appropriately chosen strictly concave functions. As a propaedeutic exercise the reader should recall that the functions $\sqrt{x}$ and $x/(1 + x)$ are strictly concave. Furthermore, for any fixed real $p > 1$ the functions $x^{1/p}$ and $(x^{1/p} + 1)^p$ are strictly concave. Throughout $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ will denote sequences of positive real numbers.

Cauchy Augustin-Louis Cauchy [**3**] published his famous inequality in 1821. Then in 1859, Viktor Yakovlevich Bunyakovsky [**2**] derived a corresponding inequality for integrals, and in 1885 Hermann Schwarz [**11**] proved a corresponding version for inner-product spaces. Therefore the Cauchy inequality sometimes also shows up under the name Schwarz inequality, or Cauchy-Schwarz inequality, or Cauchy-Bunyakovsky-Schwarz inequality. In any case, its discrete version states that

$$\sum_{i=1}^{n} a_i b_i \le \sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}. \qquad (8)$$

Cauchy's original proof of (8) rewrites it into the equivalent and obviously true

$$0 \le \sum_{1 \le i < j \le n} (a_i b_j - a_j b_i)^2.$$

We give another very short proof of (8) by deducing it from the master theorem: We use the strictly concave function $g(x) = \sqrt{x}$, which yields $f(x, y) = \sqrt{x}\sqrt{y}$. Then (6) turns into

$$\sum_{i=1}^{n} \sqrt{x_i}\sqrt{y_i} \le \sqrt{\sum_{i=1}^{n} x_i} \cdot \sqrt{\sum_{i=1}^{n} y_i}.$$

Finally, setting $x_i = a_i^2$ and $y_i = b_i^2$ for $1 \le i \le n$ yields the Cauchy inequality (8). Furthermore equality holds in (8) if and only if the a_i and the b_i are proportional.

Hölder We turn to the Hölder inequality, which was first derived in 1888 by Leonard James Rogers [**10**], and then in 1889 in a different way by Otto Ludwig Hölder [**6**].

This inequality is built around two real numbers $p, q > 1$ with $1/p + 1/q = 1$. It states that

$$\sum_{i=1}^{n} a_i b_i \le \left(\sum_{i=1}^{n} {a_i}^p \right)^{1/p} \left(\sum_{i=1}^{n} {b_i}^q \right)^{1/q}. \tag{9}$$

Note that the Cauchy inequality is the special case of the Hölder inequality with $p = q = 2$. One standard proof of (9) is based on Young's inequality, which gives $xy \le x^p/p + y^q/q$ for all real $x, y > 0$ and for all real $p, q > 1$ with $1/p + 1/q = 1$.

But let us deduce the Hölder inequality from the master theorem. We set $g(x) = x^{1/p}$, which is strictly concave if $p > 1$. Then the corresponding function f is given by $f(x, y) = x^{1/p} y^{1/q}$, and inequality (6) becomes

$$\sum_{i=1}^{n} {x_i}^{1/p} {y_i}^{1/q} \le \left(\sum_{i=1}^{n} x_i \right)^{1/p} \left(\sum_{i=1}^{n} y_i \right)^{1/q}.$$

We substitute $x_i = {a_i}^p$ and $y_i = {b_i}^q$ for $1 \le i \le n$, and get the Hölder inequality (9). Clearly, equality holds in (9) if and only if the ${a_i}^p$ and the ${b_i}^q$ are proportional.

Minkowski The Minkowski inequality was established in 1896 by Hermann Minkowski [**9**] in his book *Geometrie der Zahlen* (Geometry of Numbers). It uses a real parameter $p > 1$, and states that

$$\left(\sum_{i=1}^{n} (a_i + b_i)^p \right)^{1/p} \le \left(\sum_{i=1}^{n} {a_i}^p \right)^{1/p} + \left(\sum_{i=1}^{n} {b_i}^p \right)^{1/p}. \tag{10}$$

The special case of (10) with $p = 2$ is the triangle inequality $\|a + b\|_2 \le \|a\|_2 + \|b\|_2$ in Euclidean spaces. Once again we exhibit a very short proof via the master theorem. We choose $g(x) = (x^{1/p} + 1)^p$. Since $p > 1$, this function g is strictly concave. The corresponding function f is given by $f(x, y) = (x^{1/p} + y^{1/p})^p$. Then the inequality in (6) becomes

$$\sum_{i=1}^{n} ({x_i}^{1/p} + {y_i}^{1/p})^p \le \left(\left(\sum_{i=1}^{n} x_i \right)^{1/p} + \left(\sum_{i=1}^{n} y_i \right)^{1/p} \right)^p.$$

By setting $x_i = a_i^p$ and $y_i = b_i^p$ for $1 \le i \le n$ and by taking the pth root on both sides, we produce the Minkowski inequality (10). Furthermore equality holds in (10), if and only if the a_i^p and the b_i^p are proportional, which happens if and only if the a_i and the b_i are proportional.

Milne In 1925 Milne [**8**] used the following inequality (11) to analyze the biases inherent in certain measurements of stellar radiation:

$$\left(\sum_{i=1}^{n} (a_i + b_i) \right) \left(\sum_{i=1}^{n} \frac{a_i b_i}{a_i + b_i} \right) \le \left(\sum_{i=1}^{n} a_i \right) \left(\sum_{i=1}^{n} b_i \right). \tag{11}$$

Milne's inequality is fairly well known, but of course the inequalities of Cauchy, Hölder, and Minkowski are in a completely different league—both in terms of relevance and in terms of publicity. Milne's inequality is also discussed on page 61 of Hardy, Littlewood, and Pólya [**5**]. The problem corner in *Crux Mathematicorum* [**1**] lists three simple proofs that are due to Ardila, to Lau, and to Murty, respectively.

Murty's proof is particularly simple and rewrites (11) into the equivalent

$$0 \leq \sum_{1 \leq i < j \leq n} \frac{(a_i b_j - a_j b_i)^2}{(a_i + b_i)(a_j + b_j)}.$$

And here is our proof: This time we use the strictly concave function $g(x) = x/(1+x)$, which yields $f(x, y) = xy/(x+y)$. The resulting version of (6) yields

$$\sum_{i=1}^{n} \frac{a_i b_i}{a_i + b_i} \leq \left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} b_i\right) \Big/ \left(\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i\right),$$

which is equivalent to (11). Once again equality holds if and only if the a_i and the b_i are proportional.

A generalization of the master theorem

We now generalize the master theorem to higher dimensions. This is a fairly easy enterprise, since all concepts and arguments for the higher-dimensional case run perfectly in parallel to the one-dimensional case. For instance, a function $g : \mathbb{R}_+^d \to \mathbb{R}$ is *concave* if it satisfies

$$\lambda \cdot g(\vec{x}) + (1-\lambda) \cdot g(\vec{y}) \leq g\left(\lambda \vec{x} + (1-\lambda)\vec{y}\right) \tag{12}$$

for all $\vec{x}, \vec{y} \in \mathbb{R}_+^d$ and for all real numbers λ with $0 < \lambda < 1$. A concave function g is *strictly concave*, if equality in (12) is equivalent to $\vec{x} = \vec{y}$. It is known that a twice-differentiable function g is concave (strictly concave) if and only if its Hessian matrix is negative semidefinite (negative definite) for all $\vec{x} \in \mathbb{R}_+^d$.

Here is the higher-dimensional version of the master theorem. Note that by setting $d = 2$ in the new theorem we recover the master theorem.

HIGHER-DIMENSIONAL MASTER THEOREM. *Let $d \geq 2$ be an integer, and let $g : \mathbb{R}_+^{d-1} \to \mathbb{R}$ be a strictly concave function. Let $f : \mathbb{R}_+^d \to \mathbb{R}$ be the function defined by*

$$f(x_1, x_2, \ldots, x_d) = x_d \cdot g\left(\frac{x_1}{x_d}, \frac{x_2}{x_d}, \ldots, \frac{x_{d-1}}{x_d}\right). \tag{13}$$

Then any $n \times d$ matrix $Z = (z_{i,j})$ with positive real entries satisfies the inequality

$$\sum_{i=1}^{n} f(z_{i,1}, z_{i,2}, \ldots, z_{i,d}) \leq f\left(\sum_{i=1}^{n} z_{i,1}, \sum_{i=1}^{n} z_{i,2}, \ldots, \sum_{i=1}^{n} z_{i,d}\right). \tag{14}$$

Equality holds in (14) *if and only if matrix Z has rank* 1 *(that is, if and only if there exist real numbers $s_1, \ldots, s_n$ and $t_1, \ldots, t_d$ such that $z_{i,j} = s_i t_j$ for all i, j).*

Proof. The proof closely follows the proof of the master theorem. As in (7), we observe that all positive real numbers $a_1, \ldots, a_d$ and $b_1, \ldots, b_d$ satisfy

$$f(a_1, \ldots, a_d) + f(b_1, \ldots, b_d) \;\leq\; f(a_1 + b_1, a_2 + b_2, \ldots, a_d + b_d).$$

Equality holds if and only if the a_i and the b_i are proportional. Then an inductive argument based on this observation yields the statement in the theorem, and completes the proof. ∎

We conclude this article by posing two exercises to the reader that both can be settled through the higher-dimensional master theorem. Each exercise deals with inequalities for three sequences $a_1, \ldots, a_n$, $b_1, \ldots, b_n$, and $c_1, \ldots, c_n$ of positive real numbers.

Generalized Hölder The first exercise concerns the generalized Hölder inequality, which is built around three real numbers $p, q, r > 1$ with $1/p + 1/q + 1/r = 1$. It states that

$$\sum_{i=1}^{n} a_i b_i c_i \leq \left(\sum_{i=1}^{n} {a_i}^p\right)^{1/p} \left(\sum_{i=1}^{n} {b_i}^q\right)^{1/q} \left(\sum_{i=1}^{n} {c_i}^r\right)^{1/r}. \tag{15}$$

The reader is asked to deduce inequality (15) from the higher-dimensional master theorem (perhaps by using the function $g(x, y) = x^{1/p} y^{1/q}$), and to identify the cases of equality.

Generalized Milne Problem #68 on page 62 of Hardy, Littlewood, and Pólya [**5**] concerns the following generalization of Milne's inequality (11) to three sequences.

$$\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} b_i\right)\left(\sum_{i=1}^{n} c_i\right) \geq \left(\sum_{i=1}^{n} (a_i + b_i + c_i)\right)\left(\sum_{i=1}^{n} \frac{a_i b_i + b_i c_i + a_i c_i}{a_i + b_i + c_i}\right)\left(\sum_{i=1}^{n} \frac{a_i b_i c_i}{a_i b_i + b_i c_i + a_i c_i}\right)$$

We ask the reader to deduce it from the higher-dimensional master theorem, and to describe the cases of equality. One possible proof goes through two steps, where the first step uses $g(x, y) = xy/(xy + x + y)$, and the second step uses the function $g(x, y) = (xy + x + y)/(x + y + 1)$.

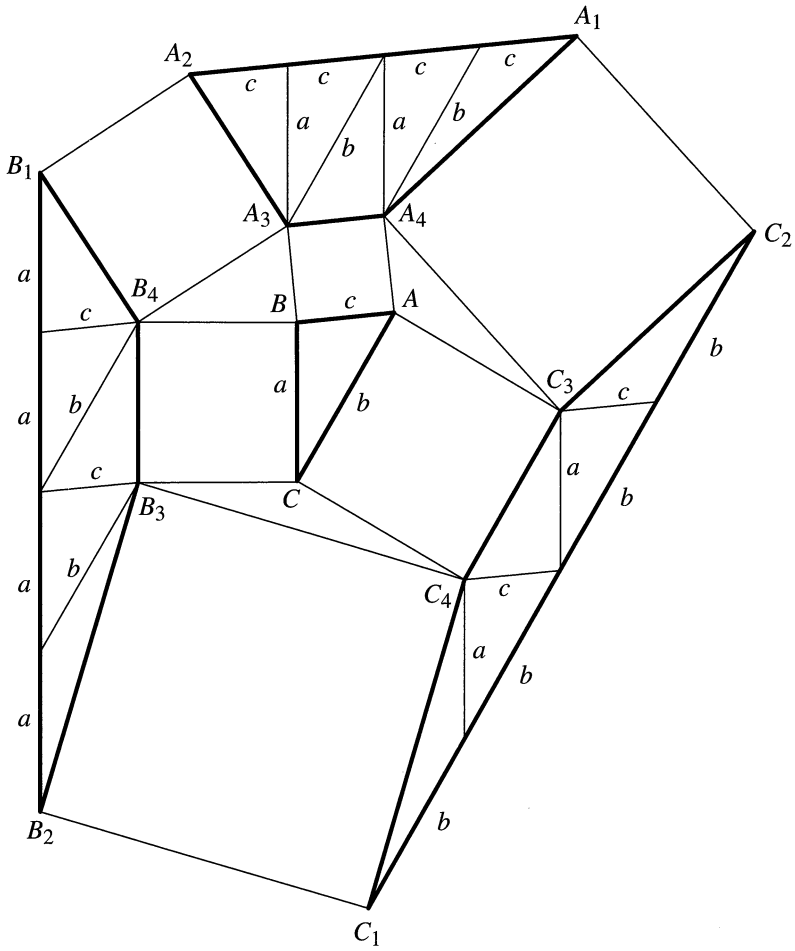
REFERENCES

1. F. Ardila, K. W. Lau, and V. N. Murty, Solution to problem 2113, *Crux Mathematicorum* **23** (1997) 112–114.
2. V. Y. Bunyakovsky, Sur quelques inégalités concernant les intégrales aux différences finies, *Mémoires de l'Academie impériale des sciences de St.-Pétersbourg* **1**(9) (1859), 4.
3. A. L. Cauchy, *Cours d'Analyse de l'Écloe Royale Polytechnique, Premiére Partie, Analyse Algébrique*, Debure frères, Paris, 1821.
4. E. K. Godunova, Convexity of complex functions and its use in proving inequalities (in Russian), *Matematicheskie Zametki* **1** (1967) 495–500. English translation in *Mathematical Notes* **1** (1967) 326–329.
5. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, 1934.
6. O. L. Hölder, Über einen Mittelwerthsatz, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen* (1889) 38–47.
7. J. L. W. V. Jensen, Sur les fonctions convexes et les inégalités entre les valeurs moyennes, *Acta Mathematica* **30** (1906) 175–193.
8. E. A. Milne, Note on Rosseland's integral for the stellar absorption coefficient, *Monthly Notices of the Royal Astronomical Society* **85** (1925) 979–984.
9. H. Minkowski, *Geometrie der Zahlen*, Teubner, Leipzig, 1896.
10. L. J. Rogers, An extension of a certain theorem in inequalities, *Messenger of Mathematics* **17** (1888) 145–150.
11. H. A. Schwarz, Über ein Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung, *Acta Societatis Scientiarum Fennicae* **15** (1885) 315–362.
12. J. M. Steele, *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, Cambridge University Press, Cambridge, 2004.

Proof Without Words: Beyond Extriangles

For any triangle $\triangle ABC$, construct squares on each of the three sides. Connecting adjacent square corners creates three extriangles. Iterating this process once more produces four quadrilaterals whose area is five times the original triangle. In the figure, letting [ ] denote area, we can see

$$[A_1A_2A_3A_4] = [B_1B_2B_3B_4] = [C_1C_2C_3C_4] = 5[ABC].$$



—M. N. Deshpande
Nagpur, India 440 025
dpratap_ngp@sancharnet.in

Varignon's Theorem for Octahedra and Cross-Polytopes

JOHN D. PESEK, JR.
Department of Food & Resource Economics
University of Delaware
Newark, DE 19716
pesek@udel.edu

Varignon's theorem states that the midpoints of a quadrilateral form the vertices of a parallelogram (FIGURE 1). From a historical point of view this simple theorem is noteworthy in that no one seemed to have been aware of it until Varignon discovered it in the 1700s [**7**]. It occurred to us that since an octahedron can be considered to be a 3-dimensional generalization of a quadrilateral and that since a midpoint can be considered as the centroid of the endpoints of a segment, it might be that the centroids of the faces of an octahedron form the vertices of a parallelepiped (FIGURE 4). In this note, we show this and extend the result to cross-polytopes of all dimensions. We also observe that the results remain true if weights are attached to the vertices (FIGURE 5).

At the time this article was submitted, we were not able to find the 3-dimensional and higher results anywhere in print. Since then the unweighted 3-dimensional case has appeared in *Forum Geometricorum* [**5**, p. 127, Proposition 16]. A weighted version for two dimensions can be found in Altshiller-Court [**1**, p. 56, Theorem 162] . A related result is that if a hexagon is derived from an arbitrary hexagon by connecting the centroids of each of six sets of three consecutive vertices, then this hexagon has equal and parallel opposite sides (Weisstein [**9**] and Wells [**10**, p. 53]).

We complete the introduction by reviewing the proof of the classical Varignon result. In the next section we introduce vector methods and define parallelotopes. We define and discuss octahedra and cross-polytopes and move on to state and prove the main result. Finally, we generalize the theorem to include weights and apply the result to find certain plane cross sections of a tetrahedron.

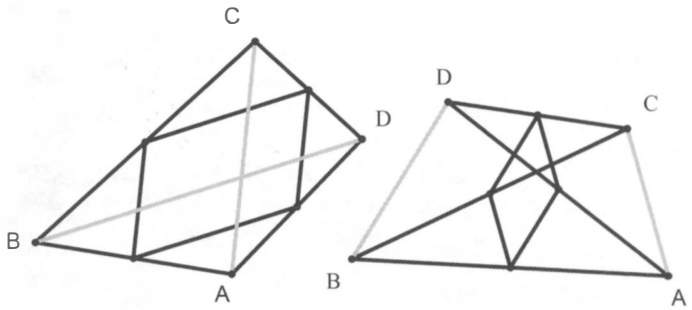


Figure 1 Varignon's theorem for a convex and a self-intersecting quadrilateral

The diagonals of a quadrilateral are the key to proving Varignon's theorem. It is easy to see (for instance, from Euclid's VI.2) that the segments connecting the midpoints are parallel to the respective diagonals and half the length (FIGURE 1) from which the theorem follows. Note that this argument imposes minimal assumptions on the quadrilateral. It does not need to be convex; it can be self-intersecting. And as noted in Coxeter and Greitzer [**3**, p. 56] the quadrilateral can even be skew, that is, the vertices

need not lie in a plane (FIGURE 5). Thus in general the quadrilateral will not have an interior. The critical assumption is that the diagonals of the quadrilateral be neither parallel nor lie on the same line. In either case the parallelogram would collapse to a line segment.

Parallelograms, parallelepipeds, and parallelotopes We have trouble visualizing in dimensions greater than three. However, vectors and linear algebra allow us to work in spaces of any dimension. Let V be a finite-dimensional vector space over the real numbers. A *flat* K is a translate of a subspace W by a vector d that is, $W + d$. The *dimension of a flat* is the dimension of the translated subspace, so a 0-dimensional flat is a point, a 1-dimensional flat is a line, and a 2-dimensional flat is a plane. Two flats are *parallel* if some translate of one flat is contained in the other flat and neither contains the other. If the dimensions are equal, then a translate of one will be equal to the other. We will say that two objects are parallel if the smallest flats containing them are parallel.

We adopt the definition of a parallelotope given by Nash [**6**] and Khosravi and Taylor [**4**] using the concept of coordinates relative to a basis. To motivate this definition, we first consider the 2-dimensional case. We define a *parallelogram* P with *base vector* b and *edge vectors* u_1 and u_2 to be the set of vectors of the form $b + t_1u_1 + t_2u_2$, where t_1 and t_2 are numbers between 0 and 1. The vertices are b, $b + u_1$, $b + u_2$ and $b + u_1 + u_2$ (FIGURE 2).

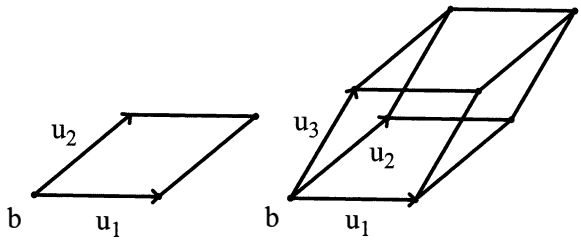


Figure 2 A parallelogram and a parallelepiped showing base and edge vectors

An *n-dimensional parallelotope* P in a vector space V with *base vector* b and (linearly independent) *edge vectors* $\{u_i\}_{i=1,\ldots,n}$ consists of all vectors of the form

$$b + t_1u_1 + t_2u_2 + \cdots + t_nu_n, \quad \text{where } 0 \le t_1, t_2, \ldots, t_n \le 1.$$

Thus a parallelotope can be a point ($n = 0$), segment ($n = 1$), parallelogram ($n = 2$), or parallelepiped ($n = 3$).

It is easy to check that P is convex. We can define *m-dimensional faces* of P by setting $n - m$ of the t_is to be zero or one. It is easy to check that these are also parallelotopes. The 0-dimensional faces are the vertices. They are the 2^n points

$$w_{i_1 i_2 \ldots i_n} = b + \sum_{\{k \mid i_k = 1\}} u_k, \tag{1}$$

where $i_k = 0$ or 1. We need the following observation, which follows quickly by taking $b = w_{00\ldots0}$, to prove our main result.

OBSERVATION 1. If a set of vectors $w_{i_1 i_2 \ldots i_n}$ satisfies (1) for some vector b and independent vectors u_k where $k = 1, \ldots, n$, then they are the vertices of a parallelotope.

Octahedra and cross-polytopes Before we formally define octahedra and cross-polytopes, it will be helpful to discuss them informally. Most likely the figure that

comes to the reader's mind when the word octahedron is mentioned is a regular octahedron, which is one of the five Platonic polyhedra, easily constructed using the six points $(\pm 1, 0, 0)$, $(0, \pm 1, 0)$, and $(0, 0, \pm 1)$. The eight faces are the equilateral triangles $\{(-1, 0, 0), (0, -1, 0), (0, 0, -1)\}$, $\{(-1, 0, 0), (0, -1, 0), (0, 0, 1)\}$, $\{(-1, 0, 0), (0, 1, 0), (0, 0, -1)\}$, and $\{(-1, 0, 0), (0, 1, 0), (0, 0, 1)\}$, as well as 4 others with $(-1, 0, 0)$ replaced by $(1, 0, 0)$. Each vertex lies in four faces and the dihedral angle between adjacent faces is always the same. The three segments $(-1, 0, 0)(1, 0, 0)$, $(0, -1, 0)(0, 1, 0)$ and $(0, 0, -1)(0, 0, 1)$ are called the *diagonals* of the octahedron. They are analogous to the diagonals of a quadrilateral.

For a general octahedron we need six vertices: v_{10}, v_{11}, v_{20}, v_{21}, v_{30}, and v_{31}. There are eight triangular faces, $\{v_{10}, v_{20}, v_{30}\}$, $\{v_{10}, v_{20}, v_{31}\}$, $\{v_{10}, v_{21}, v_{30}\}$, and $\{v_{10}, v_{21}, v_{31}\}$, as well as four others with v_{10} replaced by v_{11}. The notation is illustrated in FIGURE 4. These eight triangles contain all the segments that can be drawn among the six points except for three. These are the *diagonals* $v_{10}v_{11}$, $v_{20}v_{21}$, and $v_{30}v_{31}$ of the octahedron. Like quadrilaterals, these octahedra need be not be convex and may be self-intersecting (FIGURE 3). In general six points will determine a 5-dimensional flat so octahedra may also be skew. Even if the octahedron lies in three dimensions, the diagonals may lie on skew lines. Like quadrilaterals they do not in general have a well-defined interior.

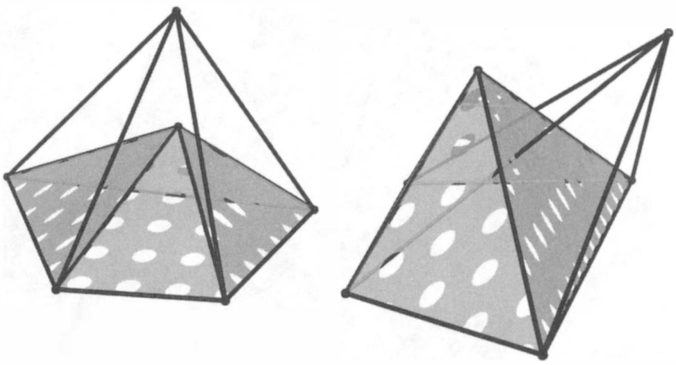


Figure 3 A nonconvex octahedron (left) and a self-intersecting one (right), with half the faces shaded

In order to define cross-polytopes, we first define a simplex. An *n-dimensional simplex* S is the convex hull of a set of $n + 1$ vectors $\{v_0, v_1, v_2, \ldots, v_n\}$, called the *vertices* of S. The convex hull is

$$\left\{ v \mid v = \sum_{i=0}^{n} t_i v_i \text{ where } \sum_{i=0}^{n} t_i = 1 \text{ and } t_i \geq 0 \text{ for } i = 0, \ldots, n \right\}.$$

We require that the points $\{v_0, v_1, v_2, \ldots, v_n\}$ be *affinely independent*. That is, the smallest flat containing them has dimension n. A simplex is a point, segment, triangle, or tetrahedron as n is 0, 1, 2, or 3.

If T is a nonempty subset of $\{v_0, v_1, v_2, \ldots, v_n\}$, then S_T is the *sub-simplex* determined by T. The 0-dimensional sub-simplices of S are the vertices of S. The 1-dimensional sub-simplices are the *edges* of S. The $(n-1)$-dimensional sub-simplices are the *faces* of S. The *centroid* of the simplex S is defined to be $\sum_{i=0}^{n} v_i/n$, which is the center of mass of these points if they are regarded as having equal weights. We can construct a regular $(n-1)$-dimensional simplex by using as vertices

the n n-tuples $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 1)$. All edges have the same length, n faces meet at each vertex, and the angles between each pair of faces are equal [**2, 8**].

A *cross-polytope* is a higher dimensional analog of the quadrilateral and the octahedron. The best known cross-polytope is the regular one, whose vertices are the $2n$ n-tuples $(\pm 1, 0, \ldots, 0)$, $(0, \pm 1, \ldots, 0)$, $(0, 0, \ldots, \pm 1)$. It is regular, because around each vertex there are 2^{n-1} faces (which are regular $(n-1)$-dimensional simplices), all faces are regular and congruent, and the angles between adjacent faces are equal. The regular simplices, hypercubes, and cross-polytopes constitute the only families of regular polytopes that exist in all dimensions [**2, 8**].

A *cross-polytope* C is defined to be a set of $2n$ vectors $\{v_{ij}\}$ where $i = 1, \ldots, n$ and $j = 0, 1$ together with the set of 2^n $(n-1)$-dimensional simplices $\{S_{i_1 i_2 \ldots i_n}\}$ where $i_k = 0, 1$, with vertex sets $\{v_{1i_1}, v_{2i_2}, \ldots, v_{ni_n}\}$. The $\{S_{i_1 i_2 \ldots i_n}\}$ are called the faces of C. For instance, $S_{11\ldots1}$ is a simplex with vertices $\{v_{11}, v_{21}, \ldots, v_{n1}\}$; in FIGURE 4, where $n = 3$, this is the front face on the right.

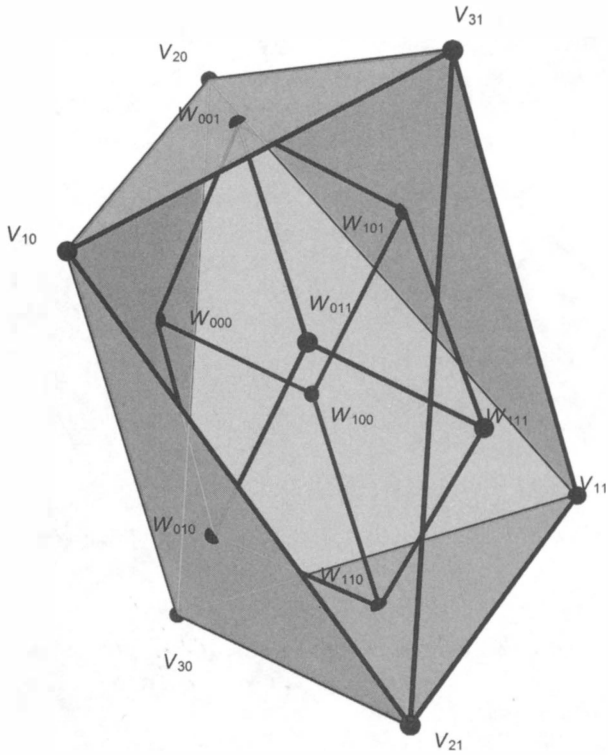


Figure 4 Varignon's theorem for an octahedron

The *diagonals* of a cross-polytope are the segments $v_{i0}v_{i1}$ for $i = 1, \ldots, n$. The n vectors $v_{i1} - v_{i0}$ are required to be linearly independent. If this is true, we say that the diagonals have *independent directions*.

A 2-dimensional cross-polytope is a quadrilateral and a 3-dimensional cross-polytope is an octahedron. Like quadrilaterals and octahedra, a cross-polytope need not be convex and may be self-intersecting. It may be skew and may be contained in anywhere from n to $2n - 1$ dimensions. The diagonals may be skew even if the cross-polytope is contained in n dimensions. In general, it need not have a well-defined interior.

The main result We are now ready to state and prove our main result.

THEOREM 1. *Let C be a cross-polytope with vertex set $\{v_{ij}\}$ and faces $S_{i_1 i_2 \ldots i_n}$. Then the centroids of the faces of C,*

$$w_{i_1 i_2 \ldots i_n} = \left(\sum_{j=1}^{n} v_{j i_j} \right) / n,$$

form the vertex set of an n-dimensional parallelotope P whose edge vectors are $\{(v_{j1} - v_{j0})/n\}$. Each edge of the parallelotope is parallel (or perhaps lies on the same line as) one of the diagonals of the cross-polytope.

The theorem is illustrated by FIGURES 1, 5, and especially 4.

Proof. Computations that prove the theorem are very straightforward and also rather effectively hide what is going on. To gain insight into what is happening, refer to FIGURE 4 and check the following calculation. One set of edges of the proposed parallelotope is

$$w_{1ij} - w_{0ij} = \frac{v_{11} + v_{2i} + v_{3j}}{3} - \frac{v_{10} + v_{2i} + v_{3j}}{3} = \frac{v_{11} - v_{10}}{3},$$

which does not depend on i and j. This shows that these edges are equal and parallel to the diagonal determined by v_{10} and v_{11}. This is true for each diagonal of the octahedron. The same cancellation occurs in general computation, but in a disguised form.

Let $u_j = (v_{j1} - v_{j0})/n$. The u_j are independent since the diagonal vectors are independent by the definition of a cross-polytope. Let $b = w_{00\ldots0}$. By Observation 1 we need to show that

$$w_{i_1 i_2 \ldots i_n} = \left(\sum_{j=1}^{n} v_{j i_j} \right) \Big/ n = b + \sum_{\{j | i_j = 1\}} u_j.$$

This follows from direct computation:

$$\begin{aligned} b + \sum_{\{j|i_j=1\}} u_j &= \left(\sum_{j=1}^{n} v_{j0} \right) \Big/ n + \sum_{\{j|i_j=1\}} (v_{j1} - v_{j0})/n \\ &= \sum_{\{j|i_j=0\}} v_{j0}/n + \sum_{\{j|i_j=1\}} v_{j0}/n + \sum_{\{j|i_j=1\}} v_{j1}/n - \sum_{\{j|i_j=1\}} v_{j0}/n \\ &= \sum_{\{j|i_j=0\}} v_{j0}/n + \sum_{\{j|i_j=1\}} v_{j1}/n = \sum_{j=1}^{n} v_{j i_j}/n \end{aligned}$$

The first equality follows from the definitions. Then we break up the first sum into those terms for which $i_j = 1$ and $i_j = 0$ and distribute the second sum. After we cancel like terms, what remains is the centroid of the face $S_{i_1 i_2 \ldots i_n}$. The proof also shows that the set of vectors $\{(v_{j1} - v_{j0})/n\}$ are a set of edge vectors for the parallelotope. Consider the adjacent vertices of the parallelotope $w_{i_1 \ldots i_{k-1} 1 i_{k+1} \ldots i_n}$ and $w_{i_1 \ldots i_{k-1} 0 i_{k+1} \ldots i_n}$. Calculation shows that the difference between them is $(v_{k1} - v_{k0})/n$. Since the vectors determining the direction of the edge and the kth diagonal are nonzero multiples of each other, the diagonal and the edge must either be parallel or the edge and the diagonal must lie in the same line. ∎

Weights The interested reader can show that the following generalization holds. When we talk about the centroid of a simplex or a set of vectors, we often assume that the weights are equal. If instead we have nonzero weights s_i with $\sum_{i=1}^{n} s_i = 1$ we can define the *weighted centroid* of the vectors v_i to be $\sum_{i=1}^{n} s_i v_i$.

THEOREM 2. *Let C be a cross-polytope with vertex set $\{v_{ij}\}$ and faces $S_{i_1 i_2 \ldots i_n}$. Let $\{s_i\}_{i=1,\ldots,n}$ be scalars such that for $i = 1, \ldots, n$ we have $s_i \neq 0$ and $\sum_{i=1}^{n} s_i = 1$. Then the 2^n vectors*

$$w_{i_1 i_2 \ldots i_n} = \sum_{j=1}^{n} s_j v_{j i_j}$$

are the vertex set of an n-dimensional parallelotope P. The set of vectors $\{s_j(v_{j1} - v_{j0})\}$ are a set of edge vectors for the parallelotope. Each edge of the parallelotope is parallel (or perhaps lies on the same line as) one of the diagonals of the cross-polytope.

The proof is very close to that of the first theorem. FIGURE 5 illustrates the result.

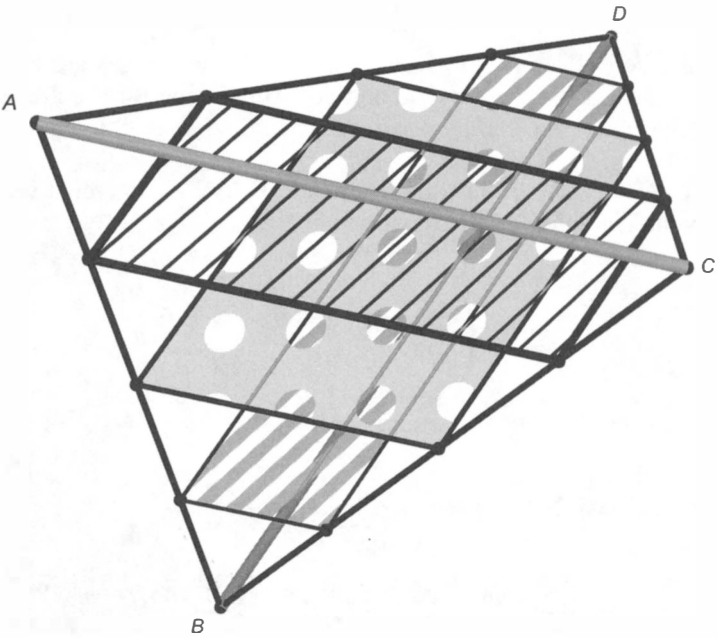


Figure 5 Weighted Varignon's theorem for a skew quadrilateral, using three different sets of weights: {0.25, 0.75}, {0.5, 0.5}, and {0.75, 0.25}

As an application of Theorem 2, consider the tetrahedron $ABCD$ in FIGURE 5. If a plane intersects the tetrahedron and is parallel to the pair of opposite sides AC and BD, then it intersects it in a parallelogram with pairs of sides parallel to AC and BD respectively. This is Theorem 156 of Altshiller-Court [**1**].

Concluding remarks In a future article we will present another generalization of Varignon's theorem and apply it to finding formulas for the volume of a simplex. An article by Mammana *et al.* [**5**] presents a unified approach to theorems about centroids in two and three dimensions. The "centroid hexagon" theorem mentioned in Weisstein

[**9**] and Wells [**10**] is closely related to the 3-dimensional case of the main theorem. It explains what happens when the six vertices of the octahedron are allowed to lie in the same plane. This result can be generalized to polygons with an even number of sides.

Figures in this note and additional figures can be found at the MAGAZINE website, as well as *Geometer's Sketchpad* or *Cabri 3d* files that allow experimentation.

REFERENCES

1. N. Altshiller-Court, *Modern Pure Solid Geometry*, 2nd ed., Chelsea, Bronx, New York, 1964.
2. H. S. M. Coxeter, *Regular Polytopes*, 2nd ed., Macmillan, New York, 1963. Reprinted by Dover Publications, New York, 1973.
3. H. S. M. Coxeter and S. L. Greitzer, *Geometry Revisited*, Mathematical Association of America, Washington, DC, 1967.
4. M. Khosravi and M. D. Taylor, The Wedge Product and Analytic Geometry, *Amer. Math. Monthly* **115** (2008) 623–644.
5. M. F. Mammana, B. Micale, and M. Pennisi, On the centroids of polygons and polyhedra, *Forum Geometricorum* **8** (2008) 121-130.
6. A. Nash, A generalized parallelogram law, *Amer. Math. Monthly* **110** (2003) 52–57.
7. P. N. Oliver, Pierre Varignon and the Parallelogram Theorem, *The Mathematics Teacher* **94** (2001) 316–319.
8. D. M. Y. Sommerville, *An Introduction to the Geometry of N Dimensions*, Dutton, New York, 1929. Reprinted by Dover Publications, New York, 1958.
9. E. W. Weisstein. Centroid Hexagon, *MathWorld–A Wolfram Web Resource*, `http://mathworld.wolfram.com/CentroidHexagon.html`.
10. D. Wells, *The Penguin Dictionary of Curious and Interesting Geometry*, Penguin, London, 1991.

A Curious Way to Test for Primes Explained

DAVID M. BRADLEY
University of Maine
Orono, ME 04469-5752
bradley@math.umaine.edu

In the October 2007 issue of this MAGAZINE [**2**], Walsh presents a curious primality test, attributed to a mysterious taxi-cab driver. Sensing there must be more to the story, I decided to track down Walsh's cab driver. As it turned out, the cabbie was bemused to learn that her off-hand remark became the subject of a journal article, so I showed it to her.

"That's interesting," she said, "but I had a simpler result in mind, and also a simpler proof." She then proceeded to explain. "Walsh's test is based on the Maclaurin series expansion

$$e^{(x^k/k)} = \sum_{j=0}^{\infty} \frac{(x^k/k)^j}{j!} = \sum_{j=0}^{\infty} \frac{x^{kj}}{k^j j!}. \tag{1}$$

Using this, he defined

$$g_n(x) = \sum_{k=1}^{n-1} e^{(x^k/k)} = \sum_{k=1}^{n-1} \sum_{j=0}^{\infty} \frac{x^{kj}}{k^j j!},$$

for real x and integer $n > 1$, and then computed the nth derivative of g_n at 0 as

$$g_n^{(n)}(0) = \sum_{\substack{k=1 \\ k|n}}^{n-1} \frac{n!}{k^{n/k}(n/k)!}, \tag{2}$$

employing the standard abbreviation $k|n$ for the condition that $n/k \in \mathbb{Z}$.

Walsh's test amounts to the observation that an integer $n > 1$ is prime if and only if $g_n^{(n)}(0) = 1$. Now it's not hard to see that the k^j factor in the denominator of the rightmost sum in (1) plays no role other than to reduce the size of $g_n^{(n)}(0)$ when n is composite. What I actually had in mind is the following:

THEOREM. *For each integer $n > 1$, define the function $f_n : \mathbb{R} \to \mathbb{R}$ by*

$$f_n(x) = \sum_{k=1}^{n-1} e^{(x^k)}.$$

An integer $n > 1$ is prime if and only if the nth derivative of f_n satisfies $f_n^{(n)}(0) = 1$.

Proof. In light of the fact that the Maclaurin series expansion

$$e^{(x^k)} = \sum_{j=0}^{\infty} \frac{x^{kj}}{j!}$$

is valid for all real x and all positive integers k, it follows that if $x \in \mathbb{R}$, then

$$f_n(x) = \sum_{k=1}^{n-1} \sum_{j=0}^{\infty} \frac{x^{kj}}{j!}. \tag{3}$$

Now we could calculate $f_n^{(n)}$ following Walsh [**2**], by repeatedly differentiating term by term, but it seems easier to note that by Taylor's theorem, $f_n^{(n)}(0)$ is equal to $n!$ times the coefficient of x^n in $f_n(x)$. Observe that we get a contribution to the coefficient of x^n in (3) if and only if $n = kj$. We conclude that

$$f_n^{(n)}(0) = \sum_{\substack{k=1 \\ k|n}}^{n-1} \frac{n!}{(n/k)!} = 1 + \sum_{\substack{k=2 \\ k|n}}^{n-1} \frac{n!}{(n/k)!}. \tag{4}$$

If n is prime, then the sum on the right is empty; otherwise it is strictly positive." ∎

I then pointed out that the same idea would work if we eliminated not just the k^j in (1), but the $j!$ too. For, if $|x| < 1$ and k is any positive integer, then the formula for the sum of geometric series with ratio x^k gives

$$\frac{1}{1 - x^k} = \sum_{j=0}^{\infty} x^{kj}. \tag{5}$$

If we now define

$$h_n(x) = \frac{1}{n!} \sum_{k=1}^{n-1} \frac{1}{1 - x^k} = \frac{1}{n!} \sum_{k=1}^{n-1} \sum_{j=0}^{\infty} x^{kj} \tag{6}$$

for integer $n > 1$ and real x such that $-1 < x < 1$, then the same reasoning shows that

$$h_n^{(n)}(0) = \sum_{\substack{k=1 \\ k|n}}^{n-1} 1 = \tau(n) - 1, \tag{7}$$

where $\tau(n)$ is the number of positive integer divisors of n. It follows that $h_n^{(n)}(0) = 1$ if and only if n is prime.

The cabbie nodded. "Of course, it would be nice if you could use a single function to test all positive integers n. It's tempting to try something like

$$\sum_{k=1}^{\infty} \frac{1}{1 - x^k},$$

but that diverges if $-1 < x < 1$. But if you look at (5) and (6), you'll see that the $j = 0$ term plays no essential role in the subsequent argument. Dropping it leads us to consider

$$L(x) := \sum_{k=1}^{\infty} \frac{x^k}{1 - x^k} = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} x^{kj},$$

which is valid for all real x such that $-1 < x < 1$. Furthermore, the same reasoning as before shows that

$$\frac{L^{(n)}(0)}{n!} = [\text{coefficient of } x^n \text{ in } L(x)] = \sum_{\substack{k=1 \\ k|n}}^{n} 1 = \tau(n), \tag{8}$$

so a positive integer n is prime if and only if $L^{(n)}(0)/n! = 2$."

"But wait a minute," I said. "What you've actually shown is that if $-1 < x < 1$, then

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

This is nothing other than Lambert's generating series for the divisor function [**1**, p. 280]."

The driver then observed that just as (7) and (8) have obvious combinatorial interpretations, so does (4): It counts the number of ways to partition a set of n distinct objects into ordered tuples of equal length less than n. Obviously, this is equal to 1 if and only if n is prime. The question then arose as to whether Walsh's approach also has a combinatorial interpretation. As Walsh himself confirmed [**3**], his $g_n^{(n)}(0)$ (see (2) above) counts the number of permutations of n distinct objects that can be written as a product of pairwise disjoint cycles of equal length less than n. To see this, note that the number of ways to partition kr distinct objects into r sets of size k is

$$\frac{(kr)!}{r!\,(k!)^r}.$$

For each set, the number of ways to form a cycle of size k is $(k-1)!$. Hence, the number of permutations of kr objects that can be written as a product of pairwise disjoint cycles of length k is equal to

$$\frac{(kr)!}{r!}\left(\frac{(k-1)!}{k!}\right)^r = \frac{(kr)!}{r!\,k^r}.$$

Letting $n = kr$ and summing over $1 \le k \le n-1$ such that $k|n$, we get (2).

Acknowledgment. I am grateful to the anonymous referees for their careful examination of the paper, and for their helpful suggestions and observations, which led to improvements in the exposition.

REFERENCES

1. L. E. Dickson, *History of the Theory of Numbers*, vol. 1, Washington DC, 1919.
2. Dennis P. Walsh, A curious way to test for primes, this MAGAZINE **80**(4) (October 2007) 302–303.
3. Dennis P. Walsh, Counting homocyclic permutations, preprint.

More on the Lost Cousin of the Fundamental Theorem of Algebra

ROMAN SZNAJDER
Bowie State University
Bowie, MD 20715-9465
rsznajder@bowiestate.edu

In his recent note [**2**], Timo Tossavainen proves what he calls "The Lost Cousin of the Fundamental Theorem of Algebra," which we state as:

EXPONENTIAL THEOREM. *For any integer $n \geq 1$, let $0 < \kappa_0 < \kappa_1 < \cdots < \kappa_n$ and a_j (for $j = 0, \ldots, n$) be real numbers with $a_n \neq 0$. Then the function $f : \mathbb{R} \to \mathbb{R}$,*

$$f(t) = \sum_{j=0}^{n} a_j \kappa_j^t$$

has at most n zeros.

Years ago, I was presented by a friend with a copy of a concise monograph [**1**] (112 pages long) on selected topics in polynomial approximation. In this book, apparently unknown to western readers, the following fact and its proof appear:

GENERALIZED POLYNOMIAL THEOREM. *A function g given by the formula*

$$g(x) = a_0 x^{\alpha_0} + a_1 x^{\alpha_1} + \cdots + a_n x^{\alpha_n},$$

where $\alpha_0 < \alpha_1 < \cdots < \alpha_n$ are arbitrary real numbers and $a_n \neq 0$, has no more than n roots.

Proof. We proceed by induction on n, noting that for $n = 1$ the statement is obvious. Assume that for some n the claim is true, but for $n + 1$, it is not. Hence, for some real numbers $\alpha_0 < \alpha_1 < \cdots < \alpha_n < \alpha_{n+1}$ and $a_{n+1} \neq 0$, there is a function

$$g(x) = a_0 x^{\alpha_0} + a_1 x^{\alpha_1} + \cdots + a_n x^{\alpha_n} + a_{n+1} x^{\alpha_{n+1}},$$

whose number of positive roots is larger than $n + 1$. These roots are identical with the roots of the new function

$$g(x)/x^{\alpha_0} = a_0 + a_1 x^{\alpha_1 - \alpha_0} + \cdots + a_n x^{\alpha_n - \alpha_0} + a_{n+1} x^{\alpha_{n+1} - \alpha_0}.$$

By Rolle's theorem, the derivative of the above function, which has the form

$$b_0 x^{\beta_0} + b_1 x^{\beta_1} + \cdots + b_n x^{\beta_n},$$

with $\beta_0 < \beta_1 < \cdots < \beta_n$ and $b_n \neq 0$, has more than n roots. This contradiction to the induction hypothesis concludes the proof. ■

The Exponential Theorem generalizes the fundamental theorem of algebra to exponential functions the way the Generalized Polynomial Theorem does for generalized polynomials. A striking fact is that the two proofs follow the same path. Despite appearances, the theorems are equivalent, as the following argument shows.

Let $f(t) = \sum_{j=0}^{n} a_j \kappa_j^t$ with $0 < \kappa_0 < \kappa_1 < \cdots < \kappa_n$, $a_j \in \mathbb{R}$, and $a_n \neq 0$. Let $\kappa_0 = e^{c_0}, \kappa_1 = e^{c_1}, \ldots, \kappa_n = e^{c_n}$ for some $c_0 < c_1 < \cdots < c_n$. By multiplying $f(t)$ by Δ^t for a suitable $\Delta > 1$, we may assume that $c_0 > 0$ to ascertain that $1 < c_1/c_0 < \cdots < c_n/c_0$. Then

$$\begin{aligned} f(t) &= a_0 e^{c_0 t} + a_1 e^{c_1 t} + \cdots + a_n e^{c_n t} \\ &= a_0 e^{c_0 t} + a_1 \left(e^{c_0 t}\right)^{c_1/c_0} + \cdots + a_n \left(e^{c_0 t}\right)^{c_n/c_0} \\ &= a_0 x + a_1 x^{c_1/c_0} + \cdots + a_n x^{c_n/c_0} = g(x) \end{aligned}$$

with $x = e^{c_0 t}$. By the Generalized Polynomial Theorem, with $\alpha_0 = 1, \alpha_1 = c_1/c_0, \ldots, \alpha_n = c_n/c_0$, there exist at most k positive roots of the corresponding function $g(x)$. Certainly, when x_i is such a root, $t_i := (\ln x_i)/c_0$ becomes a root of $f(t)$, and vice versa. This way, we have shown that the Generalized Polynomial Theorem implies the Exponential Theorem. The opposite implication comes from reversing the argument.

REFERENCES

1. V. N. Malozemov, *Simultaneous Approximation of a Function and Its Derivatives* (in Russian), Leningrad. Gos. Univ., Leningrad, 1973.
2. Timo Tossavainen, The lost cousin of the Fundamental Theorem of Algebra, this MAGAZINE **80** (2007) 290–294.

Closed Knight's Tours with Minimal Square Removal for All Rectangular Boards

JOE DEMAIO
Kennesaw State University
Kennesaw, GA 30144
jdemaio@kennesaw.edu

THOMAS HIPPCHEN
Kennesaw State University
Kennesaw, GA 30144
thippchen@gmail.com

Finding a closed knight's tour of a chessboard is a classic problem: Can a knight use legal moves to visit every square on the board and return to its starting position? [**1, 3**] An open knight's tour is a knight's tour of every square that does not return to its starting position. While originally studied for the standard 8×8 board, the problem is easily generalized to other rectangular boards. In 1991 Schwenk classified all rectangular boards that admit a closed knight's tour [**2**]. He described every board that cannot admit a closed knight's tour and constructed closed knight's tours for all other boards.

SCHWENK'S THEOREM. *An $m \times n$ chessboard with $m \leq n$ has a closed knight's tour unless one or more of the following three conditions hold:*

(a) *m and n are both odd;*

(b) $m \in \{1, 2, 4\}$;

(c) $m = 3$ *and* $n \in \{4, 6, 8\}$.

How close to admitting a closed knight's tour are those boards that satisfy Schwenk's conditions (a), (b), or (c)? Let us call these *obstructed boards*, since they admit no closed knight's tour. The 3×3 board is obstructed; however, once the center square is removed a closed knight's tour does indeed exist as seen in FIGURE 1.

1	6	3
4		8
7	2	5

Figure 1 A closed knight's tour of the 3×3 board with the center square removed

Let the *tour number*, $T(m, n)$, with $m \leq n$ be the minimum number of squares whose removal from an $m \times n$ chessboard will allow a closed knight's tour. Thus, $T(3, 3) = 1$. Note that unless m and n are the dimensions of an obstructed board, $T(m, n) = 0$; no squares need to be removed. Also note that removing $T(m, n)$ squares randomly from an obstructed board does not guarantee the existence of a closed knight's tour. For instance, removing any square other than the center does not allow for a closed knight's tour of the resulting 3×3 board. Furthermore $T(1, n)$ and $T(2, 2)$ are undefined since the knight cannot move from its starting position. Also $T(2, n) = 2n - 2$ for $n \geq 3$ since a knight can move down a $2 \times n$ board but cannot return to its starting position unless only one move has been made.

Parity considerations restrict the number of squares we can remove from an obstructed board if we hope to get a closed knight's tour. Throughout this paper, whenever we color the squares of a chessboard black and white, we will always begin with a black square in the upper left-hand corner. A legal move for a knight whose initial position is a white square will always result in an ending position on a black square and vice versa. Hence, any closed knight's tour must visit an equal number of black squares and white squares. This quickly determines that an odd number of squares must be removed from a board where both m and n are odd and an even number of squares must be removed from all other boards.

Thus, for an obstructed board, the smallest possible tour numbers are 1 and 2 respectively for boards with an odd or even number of squares. Recursive constructions of closed knight's tours will show that tour numbers are actually as small as possible, except in a few special cases. The constructions start with small boards, called *base boards*, and build by tacking on boards with open tours.

We compute all nonzero tour numbers by considering the three cases from Schwenk's Theorem: $m = 3$, $m = 4$, and the case where m and n are both odd.

The case of $m = 3$ For odd n, two base boards are needed for $n \equiv 1, 3 \mod 4$ and one exceptional case exists for $n = 5$. For even n, three boards are examined for $n = 4$, 6, and 8.

To construct a closed knight's tour for the 3×7 board, start with the open 3×4 tour from FIGURE 2, which begins at a and ends at l. Next, take the tour for the 3×3 board

a	*d*	*g*	*j*
l	*i*	*b*	*e*
c	*f*	*k*	*h*

Figure 2 An open knight's tour of the 3 × 4 board

with the center square removed as in FIGURE 1 and delete the 5–6 move. Connect the boards by creating the 5–*a* and 6–*l* moves (these correspond to legal knight moves) as shown in FIGURE 3. This is typical of our approach throughout the paper.

1	6	3	a	d	g	j
4		8	l	i	b	e
7	2	5	c	f	k	h

Figure 3 A closed knight's tour of the 3 × 7 board after minimal square removal

This game can be played an infinite number of times replacing the role of squares 5 and 6 by g and h. Thus, $T(3, n) = 1$ for all $n \equiv 3 \bmod 4$. Note that the lower right hand corner of any board, when used, must contain the g–h move as there are only two legal moves for a knight from that corner square.

For the 3 × 5 board note that the corner squares have only two legal moves, where one move is the center square. At most two of these four squares may be included in a tour. Hence, at least two of these squares must be removed. Furthermore, all four of those corner squares are black and it will also be necessary to remove at least one white square. Thus, $T(3, 5) \geq 3$. The existence of the tour in FIGURE 4 shows that $T(3, 5) = 3$.

1	4	7	10	
	9	12	3	6
	2	5	8	11

Figure 4 A closed knight's tour of the 3 × 5 board after minimal square removal

But the 3 × 5 board is the lone exception for all boards with an odd number of squares. FIGURE 5 shows that $T(3, 9) = 1$. As before, the open tour of FIGURE 2 yields $T(3, n) = 1$ for all $n \equiv 1 \bmod 4$, where $n \neq 5$.

1	4	7	18	21	24	9	12	15
6	19	2	25	8	17	14	23	10
3	26	5	20		22	11	16	13

Figure 5 A closed knight's tour of the 3 × 9 board after minimal square removal

The case of the 3×4 board is a very straightforward one. As shown in FIGURE 6, if two squares are removed, a knight's tour exists. Thus, $T(3, 4) \leq 2$. Since an even number of squares is required, $T(3, 4) = 2$. Furthermore, tacking on the open tour of FIGURE 2 shows that $T(3, 8) = 2$.

1	4	9	6
	7	2	
3	10	5	8

Figure 6 A closed knight's tour of the 3×4 board after minimal square removal

The 3×6 board is the only other obstructed one with $m = 3$. We will analyze FIGURE 7 to show that $T(3, 6) = 4$.

1	4	7	10	13	16
2	5	8	11	14	17
3	6	9	12	15	18

Figure 7 A way to label the 3×6 board

Using all four corners immediately forces the paths 6–1–8–3–4 and 13–18–11–16–15. Since squares 8 and 11 can have no further connections, these paths are necessarily extended to 7–6–1–8–3–4–9 and 12–13–18–11–16–15–10. However none of the four remaining squares (2, 5, 14, and 17) can be included without closing one of these paths before connecting to the other. No tour exists using all four corners that omits exactly two squares. Furthermore, including the 7–12 and 9–10 moves creates a tour when omitting 4 squares and $T(3, 6) \leq 4$.

Next note that squares 5 and 17 cannot both be used without creating a closed cycle with 10 and 12. Similarly, squares 2 and 14 cannot both be used without creating a closed cycle with 7 and 9. Combining this with the previous fact that no tour exists using all four corners that omits exactly two squares, we achieve $T(3, 6) \geq 3$. Since $T(3, 6)$ is even, $T(3, 6) = 4$.

The case of $m = 4$ Any board with $m = 4$ will have an even number of squares; thus, $T(4, n)$ will always be even. The boards of FIGURE 8 show that $T(4, 4) = T(4, 5) = T(4, 6) = 2$.

	5	10	1
13	2	7	4
6	9	14	11
	12	3	8

12	3	16	7	10
17	8	11	2	15
4	13	18	9	6
		5	14	1

21	2	13	8	17	4
12	7	22	3	14	9
1	20	11	16	5	18
		6	19	10	15

Figure 8 Closed knight's tours of the 4×4, 4×5, and 4×6 boards after minimal square removal

Next consider any tour on a $4 \times k$ board that contains the a–b and c–d moves in the lower right-hand corner as in the left-hand side of FIGURE 9. This $4 \times k$ tour can be extended to a $4 \times (k + 3)$ tour by removing moves a–b and c–d and connecting to a 4×3 extension with the moves 1–c, 6–d, a–a_1, and b–a_6. Note that all three base boards and the 4×3 extension contain the a–b and c–d moves in the lower right-hand corner as in FIGURE 9. This proves $T(4, n) = 2$ for all $n \geq 4$.

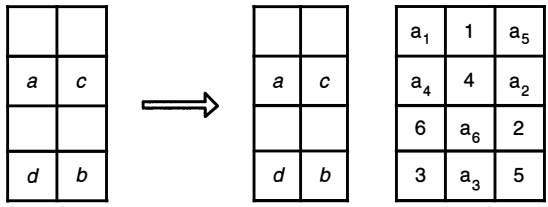


Figure 9 A closed knight's tour of the $4 \times (n + 3)$ board from a closed knight's tour of the $4 \times n$ board

The case of both m and n odd Much like the $m = 3$ and $m = 4$ cases, we use induction with an appropriate base case to analyze all boards with an odd number of squares for $m \geq 5$. In all cases, $T(m, n) = 1$ for both m and n odd with $5 \leq m \leq n$. Four base cases exist, one for each combination of $m, n \equiv 1, 3 \bmod 4$. The boards of FIGURE 10 are used for $m, n \equiv 1 \bmod 4$ and $m \equiv 1, n \equiv 3 \bmod 4$ respectively. For $m, n \equiv 3 \bmod 4$ use the 3×7 board of FIGURE 3 and for $m \equiv 3, n \equiv 1 \bmod 4$ use the 3×9 board of FIGURE 5. The open 3×4 tour of FIGURE 2 and the open 5×4 tour of FIGURE 11 can be used to extend the base boards to any length $n \equiv 1, 3 \bmod 4$ as demonstrated in FIGURE 3. For the 5×5 board delete the 10–11 move and create the 1–10 and 11–20 moves. For the 5×7 board delete the 26–27 move and create the 1–16 and 20–27 moves.

1	18	7	12	
8	13	24	17	22
19	2	21	6	11
14	9	4	23	16
3	20	15	10	5

1	30	17	8	23	28	15
18	9	34	29	16	7	24
31	2		22	25	14	27
10	19	4	33	12	21	6
3	32	11	20	5	26	13

Figure 10 Base boards for $m \equiv 1 \bmod 4$

We have constructed tours for all $3 \times n$ and $5 \times n$ boards for odd $n \geq 7$. Next we need to extend these boards down to an arbitrary odd m. To do so, rotate clockwise the open tours of FIGURE 2 and FIGURE 11 to a 4×3 tour and a 4×5 tour and extend the base $3 \times n$ and $5 \times n$ boards down to any depth $m \equiv 1, 3 \bmod 4$. For $m, n \equiv 1 \bmod 4$ use the $5 \times n$ board (created with FIGURE 10), delete the 14–15 move, and create the 1–14 and 15–20 moves with FIGURE 11 rotated clockwise. For $m \equiv 1, n \equiv 3 \bmod 4$ use the $5 \times n$ board (created with FIGURE 10), delete the 3–4 move and create the 3–a and 4–l moves with FIGURE 2 rotated clockwise. For $m \equiv 3, n \equiv 1 \bmod 4$ use the $3 \times n$ board (created with FIGURE 5), delete 5–6 move, and create the 1–6 and 5–20 moves with FIGURE 11 rotated clockwise. For $m, n \equiv 3 \bmod 4$, use the $3 \times n$ board

12	17	8	3
7	2	13	18
16	11	4	9
1	6	19	14
20	15	10	5

Figure 11 An open tour of the 5 × 4 board

(created with FIGURE 3), delete the 7–8 move, and create the 7–a and 8–l moves with FIGURE 2 rotated clockwise. This process provides us with a closed knight's tour of the top and left side of the board in FIGURE 12.

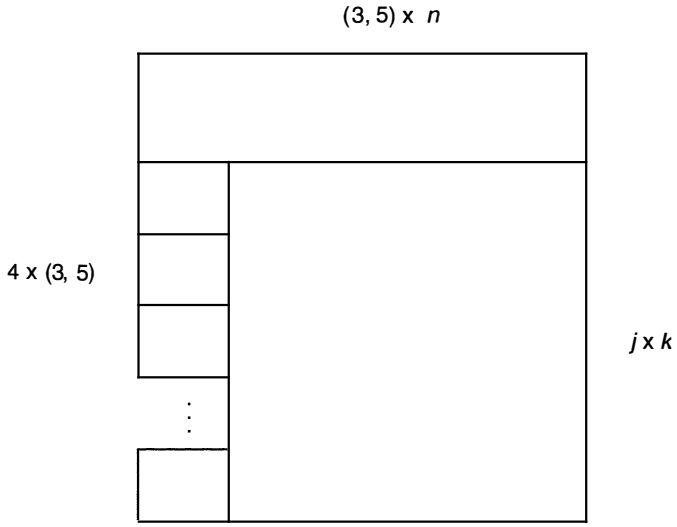


Figure 12 Constructing a closed knight's tour of the $m \times n$ board after minimal square removal for $m, n \equiv 1 \bmod 2$

Now a $j \times k$ gap with $j, k \equiv 0 \bmod 4$ needs to be filled in to complete the $m \times n$ board. Finally, we use the 4 × 4 board of FIGURE 13 to fill in the $j \times k$ gap using the same technique of FIGURE 9.

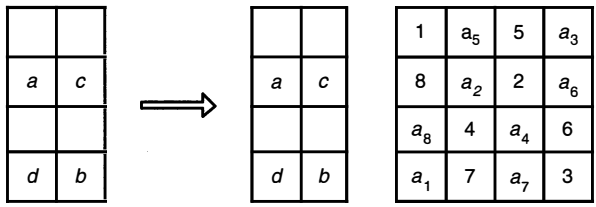


Figure 13 Filling in a $j \times k$ gap for $j, k \equiv 0 \bmod 4$

Conclusion In summary, the tour number for obstructed boards is as small as possible (1 or 2) based on an odd or even number of squares with the few noted exceptions as indicated below.

For the $m \times n$ chessboard with $m \leq n$, either board has a closed knight's tour, so that $T(m, n) = 0$, or else

(a) $T(m, n) = 1$, where m and n are both odd except for $m = 3$ and $n = 5$;
(b) $T(4, n) = 2$ for all $n \geq 4$;
(c) $T(3, 4) = T(3, 8) = 2$, $T(3, 5) = 3$, $T(3, 6) = 4$;
(d) $T(2, n) = 2n - 2$ for $n \geq 3$;
(e) $T(1, n)$ and $T(2, 2)$ are undefined.

Acknowledgment. We thank the anonymous referee whose suggestions significantly improved the clarity and quality of this article.

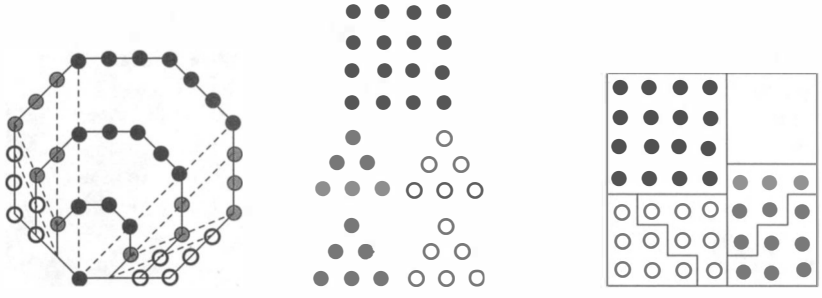
REFERENCES

1. George Jelliss, *Knight's Tour Notes*, `http://www.ktn.freeuk.com/sitemap.htm`.
2. A. J. Schwenk, Which rectangular chessboards have a knight's tour? this MAGAZINE **64** (1991) 325–332.
3. John J. Watkins, *Across the Board: The Mathematics of Chessboard Problems*, Princeton University Press, Princeton, NJ, 2004.

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

$$O_k = 1 + 7 + 13 + 19 + \cdots + (6k - 5) = (2k - 1)^2 - (k - 1)^2$$

For $k = 4$:



$$T_k = 1 + 2 + \cdots + k \Rightarrow O_k = k^2 + 4T_{k-1} \Rightarrow O_k = (2k - 1)^2 - (k - 1)^2.$$

REFERENCE

R. Nelsen, Proof without words: Every octagonal number is the difference of two squares, this MAGAZINE **77**:3 (2004) 200.

ELIZABETH JAKUBOWSKI
Florida State University, School of Teacher Education
Tallahassee, FL 32306

HASAN UNAL
Yildiz Technical University
Istanbul 34210, Turkey

Poem: The Universal Language

For G. G.

> We think we have written the message [on the plaque on the space probe Pioneer 10] in a universal language. The extraterrestrials cannot possibly understand English or Russian or Chinese or Esperanto, but they must share with us common mathematics and physics and astronomy.
>
> —Carl Sagan, *The Cosmic Connection: An Extraterrestrial Perspective*

We send out primes, triangles, digits
of pi—secret alphabets, quirks
and enigmas of my beloved trade
—hoping that some lonely, three-eyed
traveler from another star will find them
a million years hence and think of us.
But I wonder. Even among humans
mathematics is far from universal. Watch
this potter, shaping with her sensitive hands
an inert lump of clay on the wheel
into sensuous, living ripples. She never liked
math—as she's told me ruefully more than
once.
But see the intelligence in those confident
hands, the focused intention adjusting,
adapting, yielding to the feel of the clay,
the delicate progress toward beauty,
the improbable yet harmonious appositions
of unequationed surfaces, convex, concave,
the spontaneous yet watchful groping
toward some new form implicit
in the clay—not *so* different from a
mathematician's work of molding, shaping,
reshaping, polishing equations till they sparkle
with ethereal truth. Imagine the sum
total of all possible states
of awareness that the universe
has to offer (exultation at receiving
a smile from just this woman
whose eyes are just this shade of hazel;
the particular amalgam, in the first
Brahms piano concerto, of tenderness
and troubled majesty . . . not to mention
mental states of the three-eyed);
think of all the different intricate mixtures
of thought and emotion that sentient beings
might conceivably experience; think
of all those states as an immense ocean;
then surely there are waters
where we'll never swim, and yet, here we are,
thriving, more or less, in our harbor,
while the three-eyed do pretty well in their
separate seas. They're potters—I forgot
to tell you that—who never liked math,
nor had the chance. No Newton, not even
a Cardano, has arisen to grace
or trouble their continual state of languid
half-dream, a distant variant of which
we experience perhaps once
a year, when, dozing on a fall day
with sunshine full on our closed eyes, we hear
speech in the scratch and tap of an oak leaf
descending along a trunk. Yet they stay in
touch
across great swaths of space with what we'd
call
radio waves, fashioned as a potter would,
without the tools of my beloved trade,
by dreaming, whirling, prodding,
by shaping space with their gentle, frond-like
wings
not like hands, yet not so unlike, either.

—Robert Gethner
Franklin & Marshall College
Lancaster, PA 17604

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

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

PROPOSALS

To be considered for publication, solutions should be received by November 1, 2009.

1821. *Proposed by Abdullah Al-Sharif and Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.*

Let $ABCD$ be a convex quadrilateral, let X and Y be the midpoints of sides BC and DA respectively, and let O be the point of intersection of diagonals of $ABCD$. Prove that O lies inside of quadrilateral $ABXY$ if and only if

$$\text{Area}\,(AOB) < \text{Area}\,(COD).$$

1822. *Proposed by Pham Van Thuan, Hanoi University of Science, Hanoi, Vietnam.*

Let u and v be positive real numbers. Prove that

$$\frac{1}{8}\left(17 - \frac{2uv}{u^2+v^2}\right) \le \sqrt[3]{\frac{u}{v}} + \sqrt[3]{\frac{v}{u}} \le \sqrt{(u+v)\left(\frac{1}{u}+\frac{1}{v}\right)}.$$

Find conditions under which equality holds.

1823. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.*

Let n and k be positive integers. Find a closed-form expression for the number of permutations of $\{1, 2, \ldots, n\}$ for which the initial k entries have the same parity, but the initial $k+1$ entries do not. (As an example, for the permutation 5712463, the number of initial entries of the same parity is 3, the order of the set $\{5, 7, 1\}$.)

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 on every page.

1824. *Proposed by Cezar Lupu, student, University of Bucharest, Bucharest, Romania.*

Let f be a continuous real-valued function defined on [0, 1] and satisfying

$$\int_0^1 f(x)\,dx = \int_0^1 xf(x)\,dx.$$

Prove that there exists a real number c, $0 < c < 1$, such that

$$cf(c) = \int_0^c xf(x)\,dx.$$

1825. *Proposed by Greg Oman and Kevin Schoenecker, The Ohio State University, Columbus, OH.*

Let R be a ring with more than two elements. Prove that there exist subsets S and T of R, both closed under multiplication, and such that $S \not\subseteq T$ and $T \not\subseteq S$. (Note: We do not assume that R is commutative nor do we assume that R has a multiplicative identity.)

Quickies

Answers to the Quickies are on page 232.

Q991. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let f be a real-valued, differentiable function on $[a, b]$ with $f'(x) \geq f(x) > 0$ for all $x \in [a, b]$. Prove that

$$\int_a^b \frac{1}{f(x)}\,dx \leq \frac{1}{f(a)} - \frac{1}{f(b)}.$$

Q992. *Proposed by Luis H. Gallardo, University of Brest, Brest, France.*

Let n be a perfect number. Prove that if $n - 1$ and $n + 1$ are both prime, then $n = 6$.

Solutions

Staying closer to the center **June 2008**

1796. *Proposed by Matthew McMullen, Otterbein College, Westerville, OH.*

A point is selected at random from the region inside of a regular n-gon. What is the probability that the point is closer to the center of the n-gon than it is to the n-gon itself?

Solution by Houghton College Problem Solving Group, Houghton College, Houghton, NY.

Consider the regular n-gon centered at the origin and with the midpoint of one side at the point $(1, 0)$ in polar coordinates. The desired probability is the same as the probability that a point inside the triangle with vertices $O = (0, 0)$, $A = (1, 0)$ and $B = \left(\sec\frac{\pi}{n}, \frac{\pi}{n}\right)$ is closer to $(0, 0)$ than it is to side $\overline{AB}$. Now let (r, θ) be a point inside of the triangle and equidistant from O and $\overline{AB}$. Then

$$r = 1 - r\cos\theta, \qquad \text{from which} \qquad r = \frac{1}{1 + \cos\theta} = \frac{1}{2}\sec^2\frac{\theta}{2}.$$

The area of the region inside of the triangle and between this curve and the origin is

$$\frac{1}{2}\int_0^{\pi/n} r^2\,d\theta = \frac{1}{8}\int_0^{\pi/n} \sec^4\frac{\theta}{2}\,d\theta = \frac{1}{12}\tan\frac{\pi}{2n}\left(\tan^2\frac{\pi}{2n}+3\right).$$

Dividing this result by $\frac{1}{2}\tan\frac{\pi}{n}$, the area of the triangle, gives the desired probability,

$$\frac{1}{12}\left(4-\sec^4\frac{\pi}{2n}\right).$$

Also solved by Robert A. Agnew, Michael Andreoli, Herb Bailey, Thomas Bass and Kenneth Massey and Alden Starnes, Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Albania), Brian Bradie, Bruce S. Burdick, Robert Calcaterra, John Christopher, Chip Curtis, Jim Delany, Fejéntaláltuka Szeged Problem Group (Hungary), John N. Fitch, Natacha Fontes-Merz and Ramiro Fontes, Leon Gerber, Tom Gearhart, Marvin Glover, G.R.A.20 Problem Solving Group (Italy), Jeffrey M. Groah, Jerrold W. Grossman, Lee O. Hagglund, Eugene A. Herman, Michael Hitchman, Andrew Incognito, Eugen J. Ionascu, Victor Y. Kutsenok, Elias Lampakis (Greece), Charles Lindsey, David Lovit, Bob Mallison, Tim McDevitt, Kim McInturff, Missouri State University Problem Solving Group, Ronald G. Mosier, Erik Murphy and James Bush, Gail Nord, Gary L. Raduns, Edward Schmeichel, Allen Schwenk, Albert Stadler (Switzerland), Britton Stamper, James Swenson, Marian Tetiva (Romania), Bob Tomper, Michael Vowe (Switzerland), John B. Zacharias, and the proposer. There was one solution with no name and two incorrect submissions.

Limit of a radical sum **June 2008**

1797. *Proposed by Ovidiu Furdui, The University of Toledo, Toledo, OH.*

Let a, b, and c be nonnegative real numbers. Find the value of

$$\lim_{n\to\infty}\sum_{k=1}^{n}\frac{\sqrt{n^2+kn+a}}{\sqrt{n^2+kn+b}\sqrt{n^2+kn+c}}.$$

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

The value of the limit is $2(\sqrt{2}-1)$.

The kth term of the sum can be rewritten as

$$t_{n,k} = \frac{\sqrt{n^2+kn+a}}{\sqrt{n^2+kn+b}\sqrt{n^2+kn+c}} = \frac{1}{n}\cdot\frac{\sqrt{1+\frac{k}{n}+\frac{a}{n^2}}}{\sqrt{1+\frac{k}{n}+\frac{b}{n^2}}\sqrt{1+\frac{k}{n}+\frac{c}{n^2}}}.$$

Given $\epsilon > 0$ there exists N so that for $n > N$, each of a/n^2, b/n^2, and c/n^2 is less than ϵ. For such n,

$$\frac{1}{n}\cdot\frac{\sqrt{1+\frac{k}{n}}}{1+\frac{k}{n}+\epsilon} \le t_{n,k} \le \frac{1}{n}\cdot\frac{\sqrt{1+\frac{k}{n}+\epsilon}}{1+\frac{k}{n}}.$$

The sum for $k = 1$ to n for the expressions on both sides of this double inequality are Riemann sums, respectively, for the following two integrals:

$$I_L(\epsilon) = \int_0^1 \frac{\sqrt{1+x}}{1+x+\epsilon}\,dx \longrightarrow \int_0^1 \frac{\sqrt{1+x}}{1+x}\,dx = 2(\sqrt{2}-1),$$

$$I_R(\epsilon) = \int_0^1 \frac{\sqrt{1+x+\epsilon}}{1+x}\,dx \longrightarrow \int_0^1 \frac{\sqrt{1+x}}{1+x}\,dx = 2(\sqrt{2}-1),$$

where the limits are justified because the integrands converge uniformly on [0, 1] as $\epsilon \to 0$. Hence the desired limit is $2(\sqrt{2}-1)$ as claimed.

Also solved by Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Albania), David M. Bradley, Brian Bradie, Robert Calcaterra, Elliot Cohen (France), Knut Dale (Norway), Manuel Fernández-López (Spain), John N. Fitch, G.R.A.20 Problem Solving Group (Italy), Eugene A. Herman, Andrew Incognito, Eugen J. Ionascu, Victor Y. Kutsenok, Elias Lampakis (Greece), David Lovit, Bob Mallison, Reiner Martin, Missouri State University Problem Solving Group, Ronald G. Mosier, Paolo Perfetti (Italy), Éric Pité (France), Gabriel T. Prăjitură, Allen Schwenk, C. R. Selvaraj and Suguna Selvaraj, Nicholas C. Singer, Albert Stadler (Switzerland), Bob Tomper, Michael Vowe (Switzerland), John B. Zacharias, and the proposer.

Minimum of a radical sum **June 2008**

1798. *Proposed by H. A. ShahAli, Tehran, Iran.*

Let x, y, and z be positive real numbers with $x + y + z = xyz$. Find the minimum value of

$$\sqrt{1+x^2} + \sqrt{1+y^2} + \sqrt{1+z^2},$$

and find all (x, y, z) for which the minimum occurs.

Solution by Michael Reid, University of Central Florida, Orlando, FL.

The minimum value is 6, which occurs if and only if $x = y = z = \sqrt{3}$. Let $\alpha = \operatorname{Arctan} x$, $\beta = \operatorname{Arctan} y$, and $\gamma = \operatorname{Arctan} z$. Then

$$\tan\gamma = z = \frac{x+y}{xy-1} = \tan(-(\alpha+\beta)),$$

so γ differs from $-(\alpha+\beta)$ by a multiple of π. Because $\alpha, \beta, \gamma \in (0, \pi/2)$, we have $0 < \alpha + \beta + \gamma < 3\pi/2$, and it follows that $\alpha + \beta + \gamma = \pi$.

In terms of α, β, γ, the quantity to be minimized is $\sec\alpha + \sec\beta + \sec\gamma$. The function $f(t) = \sec t$ is strictly convex on the interval $(0, \pi/2)$, because $f''(t) = \sec t\,(2\tan^2 t + 1)$ is positive on the interval. Therefore we have

$$\sec\alpha + \sec\beta + \sec\gamma \geq 3\sec\left(\frac{\alpha+\beta+\gamma}{3}\right) = 3\sec\left(\frac{\pi}{3}\right) = 6,$$

with equality if and only if $\alpha = \beta = \gamma = \pi/3$. In terms of x, y, and z, this condition is equivalent to $x = y = z = \sqrt{3}$.

Also solved by George Apostolopoulos (Greece), Herb Bailey, Michel Bataille (France), Mihaly Bencze (Romania), D. Bennett and H. To, Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Albania), Brian Bradie, Krista Buchheit, Bruce S. Burdick, Robert Calcaterra, Hongwei Chen, Chip Curtis, Knut Dale (Norway), Charles R. Diminnie, Fejéntaláltuka Szeged Problem Group (Hungary), John Ferdinands, Micheal Goldenberg and Mark Kaplan, Peter Gressis and Dennis Gressis, John G. Heuver, Eugen J. Ionascu, D. Kipp Johnson, Hwan-jin Kim (Korea), Elias Lampakis (Greece), Kee-Wai Lau (China), Jizhou Li, David Lovit, Phil McCartney, Tadele Mengesha, Ronald G. Mosier, Evangelos Mouroukos (Greece), Ken'ichi Nagasaki (Japan), Paolo Perfetti (Italy), Gabriel T. Prăjitură, Toufic Saad, C. R. Selvaraj and Suguna Selvaraj, Nicholas C. Singer, Albert Stadler (Switzerland), Marian Tetiva (Romania), Nora S. Thornber, George Tsapakidis (Greece), Zhexiu Tu, University of Central Oklahoma Problem Solving Group, Michael Vowe (Switzerland), John B. Zacharias, and the proposer. There was one solution with no name and six incorrect submissions.

Matrix matters **June 2008**

1799. *Proposed by Luz DeAlba, Drake University, Des Moines, IA.*

Let $s_1, s_2, \ldots, s_n$ be real numbers with $0 < s_1 < s_2 < \cdots < s_n$. For $1 \leq i \leq j \leq n$ define $a_{ij} = a_{ji} = s_j$, and let A be the $n \times n$ matrix $A = [a_{ij}]_{1\leq i,j\leq n}$.

(a) Calculate $\det A$.

(b) Let $A^{-1} = [b_{ij}]_{1\leq i,j\leq n}$. Find the value of $\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}$.

Solution by Vadim Ponomarenko, San Diego State University, San Diego, CA.

(a) Let T be the matrix with -1 for each entry on the (first) super diagonal and 0 for all other entries. Then

$$(I+T)A(I+T^{\text{tr}}) = \operatorname{diag}(s_1 - s_2, s_2 - s_3, \ldots, s_{n-1} - s_n, s_n).$$

Because $\det(I+T) = 1$, it follows that

$$\det A = \det\left((I+T)A(I+T^{\text{tr}})\right) = (s_1 - s_2)(s_2 - s_3)\cdots(s_{n-1} - s_n)s_n.$$

(b) Let

$$B = \left((I+T)A(I+T^{\text{tr}})\right)^{-1} = \operatorname{diag}\left(\frac{1}{s_1 - s_2}, \frac{1}{s_2 - s_3}, \ldots, \frac{1}{s_{n-1} - s_n}, \frac{1}{s_n}\right).$$

Hence

$$A^{-1} = (I+T^{\text{tr}})B(I+T) = B + BT + T^{\text{tr}}B + T^{\text{tr}}BT.$$

If we let

$$\alpha = \frac{1}{s_1 - s_2} + \frac{1}{s_2 - s_3} + \cdots + \frac{1}{s_{n-1} - s_n},$$

then we find that the sum of the entries of B is $\alpha + 1/s_n$, the sum of the entries of BT is $-\alpha$, the sum of the entries of $T^{\text{tr}}B$ is $-\alpha$, and the sum of the entries of $T^{\text{tr}}BT$ is α. It follows that the sum of the entries of A^{-1} is $1/s_n$.

Also solved by Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Albania), Brian Bradie, Bruce S. Burdick, Kevin Byrnes, Robert Calcaterra, Minh Can, Hongwei Chen, Elliot Cohen (France), Chip Curtis, Knut Dale (Norway), Fejéntaláltuka Szeged Problem Group (Hungary), John Ferdinands, Manuel Fernández-López (Spain), Micheal Goldenberg and Mark Kaplan, Eugene A. Herman, M. Hako and K. Sanderson and H. To, Victor Y. Kutsenok, Reiner Martin, Missouri State University Problem Solving Group, Éric Pité (France), Gabriel T. Prăjitură, Rob Pratt, Michael Reid, Edward Schmeichel, C. R. Selvaraj and Suguna Selvaraj, Raul A. Simon (Chile), Nicholas C. Singer, John H. Smith, Albert Stadler (Switzerland), James Swenson, Marian Tetiva, Dave Trautman, Michael Vowe (Switzerland), John B. Zacharias, and the proposer.

Minimizing a ratio of areas **June 2008**

1800. *Proposed by Michel Bataille, Rouen, France.*

Let ABC be a triangle, let E be a fixed point on the interior of side AC, and let F be a fixed point on the interior of side AB. For P on $\overline{EF}$, define

$$\rho(P) = \frac{[PBC]^2}{[PCA][PAB]}.$$

For which P does $\rho(P)$ take on its minimum value? What is this minimal value?

Solution by Michael Vowe, Therwil, Switzerland.

We determine the normalized barycentric (or *areal*) coordinates for E, F, and P with respect to $A = (1, 0, 0)$, $B = (0, 1, 0)$, and $C = (0, 0, 1)$. Because E is on the interior of AC and F is on the interior of AB, we have

$$E = uA + (1-u)C = (u, 0, 1-u) \quad \text{and} \quad F = vA + (1-v)B = (v, 1-v, 0),$$

where $0 < u, v < 1$. Because P must be on the interior of EF, we have

$$P = xE + (1-x)F = (xu + (1-x)v, (1-x)(1-v), x(1-u)),$$

with $0 < x < 1$. The area of triangle PBC (as a fraction of the area of ABC) is just the first coordinate of P, so $[PBC] = xu + (1-x)v$. Similarly,

$$[PCA] = (1-x)(1-v) \qquad \text{and} \qquad [PAB] = x(1-u).$$

Hence,

$$\rho(P) = \frac{(xu + (1-x)v)^2}{x(1-x)(1-u)(1-v)}.$$

By the arithmetic-geometric mean inequality we obtain

$$\rho(P) \geq \frac{\left(2\sqrt{xu(1-x)v}\right)^2}{x(1-x)(1-u)(1-v)} = \frac{4uv}{(1-u)(1-v)},$$

with equality if and only if $xu = (1-x)v$, that is, if and only if $x = v/(v+u)$. Therefore the minimum value of $\rho(P)$ is $\frac{4uv}{(1-u)(1-v)}$ and occurs at the point

$$P_m = \left(\frac{2uv}{u+v}, \frac{u(1-v)}{u+v}, \frac{v(1-u)}{u+v}\right).$$

In other words, P_m is the point on EF with $\frac{EP}{FP} = \frac{u}{v} = \frac{(AB)(CE)}{(AC)(BF)}$, and the minimum value is $4\frac{(BE)(CF)}{(AE)(AF)}$.

Also solved by Herb Bailey, Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Albania), Robert Calcaterra, Chip Curtis, Michael Goldenberg and Mark Kaplan, Peter Gressis and Dennis Gressis, Eugen J. Ionascu, L. R. King, Victor Y. Kutsenok, Elias Lampakis (Greece), Ken'ichi Nagasaki (Japan), Gabriel T. Prăjiturǎ, Joel Schlosberg, Raul A. Simon (Chile), Albert Stadler (Switzerland), John B. Zacharias, and the proposer.

Answers

Solutions to the Quickies from page 228.

A991. Because $f'(x) \geq f(x)$ on $[a, b]$, it follows that $\left(e^{-x}f(x)\right)' \geq 0$, and hence that $e^{-x}f(x)$ is positive and nondecreasing on $[a, b]$. Thus

$$\frac{f(x)}{e^x} \geq \frac{f(a)}{e^a}, \quad \text{from which,} \quad \frac{1}{f(x)} \leq \frac{e^a}{f(a)e^x}$$

on $[a, b]$. Therefore

$$\int_a^b \frac{dx}{f(x)} \leq \int_a^b \frac{e^a}{f(a)e^x}\,dx = -\frac{e^a}{f(a)}\left(\frac{1}{e^b} - \frac{1}{e^a}\right) = \frac{1}{f(a)}\left(1 - \frac{e^a}{e^b}\right). \tag{1}$$

However

$$\frac{f(b)}{e^b} \geq \frac{f(a)}{e^a}, \quad \text{so it follows that} \quad \frac{f(a)}{f(b)} \leq \frac{e^a}{e^b}.$$

It then follows from (1) that

$$\int_a^b \frac{dx}{f(x)} \leq \frac{1}{f(a)}\left(1 - \frac{e^a}{e^b}\right) \leq \frac{1}{f(a)}\left(1 - \frac{f(a)}{f(b)}\right) = \frac{1}{f(a)} - \frac{1}{f(b)}.$$

This completes the proof.

A992. It is clear that n must be even, so $n = 2^{p-1}(2^p - 1)$, where p and $2^p - 1$ are prime. It is easy to check that if $n > 6$, then $p > 2$ and $n \equiv 1 \pmod 9$. Thus, if $n > 6$ then $n - 1$ is a multiple of 9 and hence is not prime. Note that the conditions of the problem are satisfied for the perfect number $n = 6$.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles, books, and other materials are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of mathematics literature. Readers are invited to suggest items for review to the editors.

Mallion, Roger B., The six (or seven) bridges of Kaliningrad: A personal Eulerian walk, 2006. *MATCH Communications in Mathematical and in Computer Chemistry* #58 (2007) 529–556. A contemporary Eulerian walk over the bridges of Kaliningrad, *BSHM [British Society for the History of Mathematics] Bulletin* 23 (2008) 24–36.

When was the last time you strolled along the bridges of Kaliningrad (formerly Königsberg)? Well, if it has been a while, the bridges may be in different places, and there may even be more of them. Now you can enjoy a vicarious tour, thanks to author Mallion. Euler, who never went there, would appreciate the difference since his time: An Eulerian walk (though not an Eulerian circuit) across the bridges is now possible, as accomplished in under an hour by the author and a companion in 2006. There are some technicalities (are there 6 bridges or 7 now?) and now some phantom bridges (they end in midair), but Mallion has definitively ascertained the Eulerian-ness of the city's bridges throughout the eras from the building in 1286 of its first extant bridge.

Doerfler, Ron, The art of nomography I: Geometric design; II: Designing with determinants; III: Transformations. `http://myreckonings.com/wordpress/wp-content/uploads/nomography.pdf`.

I recently inquired of my class in differential equations who knew what a slide rule was or had ever seen one. No one! Unfortunately, my department's six-foot-long one, with hooks for hanging on the blackboard, disappeared around 1980. So I showed the class the Keuffel & Esser one that my grandfather had bought in 1944 (the company went out of business a few years ago) and a compact circular one that my father bought me. A slide rule is an example of a nomogram, "the graphical representation of mathematical relationships," according to author Doerfler. To distinguish a nomogram from other informative graphics, I would say that it must perform a calculation graphically. Author Doerfler asserts that nomograms originated in 1880; hence, they had a short popular life of a century before the scientific pocket calculator in 1974. Nevertheless, nomograms are still used by doctors in medicine and in some engineering applications. Doerfler shows in his three-part essay how nomograms work and how to construct them. The first part deals with designs with straight scales, the second with curved scales, and the third with other shapes (via determinant transformations). The third part also contains references, including open-source software for creating nomograms using the Python language; you can find the sfotware and a sample nomogram for BMI (body mass index) at `www.pynomo.org`.

Plofker, Kim, *Mathematics in India*, Princeton University Press, 2009; xiii + 357 pp, $39.50. ISBN 978-0-691-12067-6.

The literature on mathematics in India is as scattered (and occasionally inconsistent) as the chronology itself is uncertain. This volume narrates in "condensed form" the "mainstream interpretation" of the history of Indian mathematics, which at its apogee featured not formal deductive proof but algorithms expressed in verse. Author Plofker deals even-handedly with conflicting theories and interpretations, as well as with questions of transmission of mathematical ideas to or from India. There is a useful appendix on Sanskrit and transliterated terms

(quotations in the book appear in translation only), plus another offering an annotated *dramatis personae* of 50 Indian mathematicians.

Emmer, Michele (ed.), *Mathematics and Culture*. 6 vols., Springer, 2003–2009. $59.95–$119. ISBN 978-3-540-: 01770-7, 21368-0, 34259-6, 34254-0, 34277-9, 87568-0.

This series of books consists of English versions of the corresponding volumes of papers *Matematica e cultura* [in Italian] presented at an annual conference. The books feature grandly illustrated essays on the "interplay" between mathematics and various other realms: art, cinema, wine, poetry, theatre, architecture, medicine, cartoons, images, and applications. This is a rich source for ideas and inspiration of how to develop and enhance a mathematical perspective on, and appreciation of, the visual arts.

A Special Issue on Formal Proof, *Notices of the American Mathematical Society* 55 (11) (December 2008) 1370–1414, `http://www.ams.org/notices/200811/index.html`.

This special issue contains four articles on formal proof. The first is by Thomas Hales, author of the proof of the Kepler conjecture about sphere-packing in 3D. The proof, which runs 300 pages supported by 40,000 lines of computer code, was published despite the fact that a team of referees exerting strenuous efforts over several years could not "certify" the proof ("and will not be able to certify it in the future, because they have run out of energy to devote to the problem"). Hales surveys the history of computer-aided proofs and gives an introduction to HOL [Higher Order Logic] Light, which has given formal proofs of various theorems (Jordan curve, Brouwer fixed-point, Cauchy residue, prime number). Georges Gonthier focuses on the four-color theorem, which was completely formalized in 2005. John Harrison considers automated reasoning more generally and welcomes formal verification. Freek Wiedijk surveys current "proof assistants," gives details of a formal proof of the quadratic reciprocity theorem, and considers as excellent the prospects of "formal mathematics."

Demaine, Erik D., Martin L. Demaine, and Tom Rodgers (eds.), *A Lifetime of Puzzles: Honoring Martin Gardner*, A K Peters, 2008; x + 349 pp, $49. ISBN 978-1-56881-245-8.

This book celebrates the 90th birthday of Martin Gardner, long-time popular expositor of mathematics. One section considers Gardner's influence on magic, including a piece by Persi Diaconis and Ron Graham on applications of generalized de Bruijn cycles to card tricks. Another section contains a history of tangrams (earliest known: 1802) and essays on an unpublished 500-year-old recreational mathematics book by Luca Pacioli. Further sections offer essays on all kinds of puzzles (I enjoyed Roger Penrose's on railway mazes) and even one on a geometric aid to scheduling bridge and tennis doubles competitions.

Mitchell, Melanie, *Complexity: A Guided Tour*, Oxford University Press, 2009; xvi + 349 pp, $29.95. ISBN 978-0-19-512441-5.

Author Mitchell identifies common properties of complex systems, whether a rain forest, an insect colony, a brain, an immune system, an economy, or the World Wide Web. Those properties are complex collective behavior, signaling and information processing, adaptation, and no central control. She details the "struggles" to define core concepts (information, computation, order, life) and their connections to measuring, assessing, and explaining complexity. (Curiously, there is no mention of computational complexity nor of $\mathcal{P} = \mathcal{NP}$.) No background in mathematics or science is assumed. Mitchell investigates chaos, networks, automata, and evolution, and wonders if we can invent a "calculus of complexity." She cites as an attempt Wolfram's analysis of complexity via cellular automata; but earlier in the book she confesses to not completely understanding what he is "getting at." Anyway, she begins the Preface by rejecting such "reductionism," in favor of a mystical hope that aspects of complexity will lead to "new ideas for addressing the most difficult problems faced by humans, such as the spread of disease, the unequal distribution of the world's natural and economic resources, the proliferation of weapons and conflict, and the effects of our society on the environment and climate." Despite these worthy causes occurring in her sentence about the "central purpose" of the book, they are not mentioned again. (The index is detialed but nonetheless deficient: Kolmogorov, Chaitin, and other people and topics mentioned in the book are not included.)

49th International Mathematical Olympiad

ZUMING FENG
Phillips Exeter Academy
Exeter, NH 03833-2460
zfeng@exeter.edu

RĂZVAN GELCA
Department of Mathematics and Statistics
Texas Tech University
Lubbock TX 79409
rgelca@gmail.com

IAN LE
Department of Mathematics
Northwestern University
Evanston IL 60208-2730
iantuanle@gmail.com

STEVEN R. DUNBAR
MAA American Mathematics Competitions
University of Nebraska-Lincoln
Lincoln, NE 68588-0658
sdunbar@maa.org

Problems

1. An acute-angled triangle ABC has orthocenter H. The circle passing through H with center the midpoint of BC intersects the line BC at A_1 and A_2. Similarly, the circle passing through H with center the midpoint of CA intersects the line CA at B_1 and B_2, and the circle passing through H with center the midpoint of AB intersects the line AB at C_1 and C_2. Show that A_1, A_2, B_1, B_2, C_1, C_2 lie on a circle.

   Submitted from Russia.

2. (a) Prove that

$$\frac{x^2}{(x-1)^2}+\frac{y^2}{(y-1)^2}+\frac{z^2}{(z-1)^2}\geq 1$$

   for all real numbers x, y, z, each different from 1, and satisfying $xyz = 1$.

   (b) Prove that equality holds above for infinitely many triples of rational numbers x, y, z, each different from 1, and satisfying $xyz = 1$.

   Submitted from Austria.

3. Prove that there exist infinitely many positive integers n such that $n^2 + 1$ has a prime divisor which is greater than $2n + \sqrt{2n}$.

   Submitted from Lithuania.

4. Find all functions $f : (0, \infty) \to (0, \infty)$ (so, f is a function from the positive real numbers to the positive real numbers) such that

$$\frac{(f(p))^2 + (f(q))^2}{f(r^2) + f(s^2)} = \frac{p^2 + q^2}{r^2 + s^2} \qquad (*)$$

for all positive real numbers p, q, r, s, satisfying $pq = rs$.

Submitted from South Korea.

5. Let n and k be positive integers with $k \geq n$ and $k - n$ an even number. Let $2n$ lamps labeled $1, 2, \ldots, 2n$ be given, each of which can be either on or off. Initially all the lamps are off. Consider sequences of steps: at each step one of the lamps is switched (from on to off or from off to on). Let N be the number of such sequences consisting of k steps and resulting in the state where lamps 1 through n are all on, and lamps $n + 1$ through $2n$ are all off. Let M be the number of such sequences consisting of k steps, resulting in the state where lamps 1 through n are all on, and lamps $n + 1$ through $2n$ are all off, but where none of the lamps $n + 1$ through $2n$ is ever switched on. Determine the ratio N/M.

Submitted from France.

6. Let $ABCD$ be a convex quadrilateral with $|BA| \neq |BC|$. Denote the incircles of triangles ABC and ADC by ω_1 and ω_2 respectively. Suppose that there exists a circle ω tangent to the ray BA beyond A and to the ray BC beyond C, which is also tangent to the lines AD and CD. Prove that the common external tangents of ω_1 and ω_2 intersect on ω.

Submitted from Russia.

Solutions We sketch the essential ideas for each problem.

1. Let A_0, B_0, C_0 be the midpoints of the sides $\overline{BC}$, $\overline{CA}$, $\overline{AB}$, respectively. Show B_1, B_2, C_1, C_2 are cyclic. Note that $AC_1 \cdot AC_2 = (AC_0 + C_0H)(AC_0 - C_0H) = AC_0^2 - C_0H^2$ and, likewise, $AB_1 \cdot AB_2 = AB_0^2 - B_0H^2$. It is clear that $\overline{B_0C_0} \perp \overline{AH}$, so $AC_0^2 - C_0H^2 = AB_0^2 - B_0H^2$; that is, $AC_1 \cdot AC_2 = AB_1 \cdot AB_2$. By the Power-of-a-Point Theorem, B_1, B_2, C_1, C_2 are cyclic. The perpendicular bisectors of the segments $\overline{B_1B_2}$, $\overline{C_1C_2}$ are also the perpendicular bisectors of segments $\overline{CA}$, $\overline{AB}$. Hence they meet at O, the circumcenter of $\triangle ABC$. Thus $B_1O = B_2O = C_1O = C_2O$. Likewise, $C_1O = C_2O = A_1O = A_2O$, and A_1, A_2, B_1, B_2, C_1, C_2 all lie on a circle centered at O.

2. Set $a = x/(x-1)$, $b = y/(y-1)$, $c = z/(z-1)$. The inequality becomes $a^2 + b^2 + c^2 \geq 1$. The condition $xyz = 1$ becomes $abc = (a-1)(b-1)(c-1)$ or $(a + b + c) - 1 = ab + bc + ca$. Hence $2(a + b + c) - 2 = 2(ab + bc + ca) = (a + b + c)^2 - (a^2 + b^2 + c^2)$ or $(a^2 + b^2 + c^2) - 1 = (a + b + c)^2 - 2(a + b + c) + 1 = (a + b + c - 1)^2 \geq 0$.

For part (b), since x, y, z are respectively rational in a, b, c, it suffices to show that there are infinitely many triples (a, b, c) of rational numbers satisfying the relation $a^2 + b^2 + c^2 = 1$ and $ab + bc + ca = (a + b + c) - 1 = 0$. Hence $0 = ab + (a + b)c = ab + (a + b)(1 - a - b)$ or $a^2 + (b-1)a + b^2 - b = 0$. The discriminant of this quadratic is $\Delta = (3b + 1)(-b + 1)$. Setting $b = p/(p^2 - p + 1)$ leads to an infinite family of rational solutions.

3. Instead of finding n with relatively large divisor p, search for p with relatively small multiple $n^2 + 1$. For such a prime p, there is a positive integer x such that $x^2 \equiv -1 \pmod{p}$. Since $x^2 \equiv (p - x)^2 \pmod{p}$, we may further assume that $x \leq \frac{p-1}{2}$. Assume that $x = (p - k)/2$ for some integer k with $1 \leq k < p - 1$. Then

$-1 \equiv x^2 \equiv \frac{(p-k)^2}{4} \pmod{p}$. Because p is odd, we have $k^2+4 \equiv 0 \pmod{p}$. In particular, $k^2+4 \geq p$ or $k \geq \sqrt{p-4}$. It follows that $x = \frac{p-k}{2} \leq \frac{p-\sqrt{p-4}}{2}$ or $p \geq 2x + \sqrt{p-4}$. Consequently, we have $\sqrt{p-4} \geq \sqrt{2x+\sqrt{p-4}-4}$. If we further assume that $p > 20$, then $\sqrt{p-4} \geq \sqrt{2x+\sqrt{p-4}-4} > \sqrt{2x}$. Hence for $p > 20$, we have $p \geq 2x + \sqrt{p-4} \geq 2x + \sqrt{2x}$. Then for each of the infinitely many primes $p \geq 20$ congruent to 1 modulo 4, we can find n such that n^2+1 is divisible by p and $p < 2n + \sqrt{2n}$, from which the desired result follows.

4. Setting $p = q = r = s = 1$ in $(*)$ gives $f(1) = 1$. For positive real numbers x, setting $(p, q, r, s) = (1, x, \sqrt{x}, \sqrt{x})$ in $(*)$ yields $2x + 2x(f(x))^2 = 2f(x) + 2x^2 f(x)$. We deduce that $0 = (x - f(x))(1 - xf(x))$. It follows that for positive real numbers x, either $f(x) = x$ or $f(x) = \frac{1}{x}$. Let us assume that $f(x) \neq x$ and $f(x) \neq \frac{1}{x}$. Then there are positive real numbers a and b such that $f(a) \neq a$ and $f(b) \neq \frac{1}{b}$. Deduce that $f(a) = \frac{1}{a}$ and $f(b) = b$. Setting $(p, q, r, s) = (a, b, \sqrt{ab}, \sqrt{ab})$ in $(*)$ gives $f(ab)(a^4 + a^2b^2) = ab(1 + a^2b^2)$. Either $f(ab) = ab$ or $f(ab) = \frac{1}{ab}$. If $f(ab) = ab$, then $f(ab)(a^4 + a^2b^2) = ab(1 + a^2b^2)$ implies that $a = 1$, but then $f(a) = f(1) = 1 = a$, violating our assumption of $f(a) \neq a$. If $f(ab) = \frac{1}{ab}$, then $f(ab)(a^4 + a^2b^2) = ab(1 + a^2b^2)$ implies that $b = 1$, but then $f(b) = f(1) = 1 = \frac{1}{b}$, violating our assumption of $f(b) \neq \frac{1}{b}$. Thus $f(x) = x$ and $f(x) = \frac{1}{x}$ are the only possible solutions of the problem, and it is easy to check they are solutions.

5. Suppose lamp i was switched on or off a_i times. Then in the first situation, a_i is odd for $1 \leq i \leq n$ and a_i is even for $n+1 \leq i \leq 2n$. Then the total number of sequences where lamp i is switched a_i times is then $\binom{k}{a_1, a_2, a_3, \ldots, a_{2n}}$. Thus

$$N = \sum \binom{k}{a_1, a_2, a_3, \ldots, a_{2n}} = \sum \frac{k!}{a_1! a_2! \cdots a_{2n}!},$$

where the sum is taken over all a_i such that $a_1 + a_2 + \cdots + a_{2n} = k$ and a_i is odd for $1 \leq i \leq n$ and a_i is even for $n+1 \leq i \leq 2n$; that is, $\frac{N}{k!}$ is the coefficient of x^k in

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

Similarly, $\frac{M}{k!}$ is the coefficient of x^k in $\left(\frac{e^x - e^{-x}}{2}\right)^n$. Now let $f(x) = \frac{e^x - e^{-x}}{2x}$. Then compare the coefficients of x^k in $k!\, x^n f(x)^n$ and $k!\, x^n f(2x)^n$. It is now evident that the ratio is 2^{k-n}.

6. Start the solution with two interesting geometry facts. (The proofs are exercises for the interested reader.)

   (a) Let BCA be a triangle, and let its incircle ω touch side CA at T_B. Point B_1 is diametrically opposite to T_B on ω. Ray BB_1 meets side CA at B_2. Then $CT_B = AB_2$.

   (b) Let $ABCD$ be a convex quadrilateral and circle ω is a an *excircle of quadrilateral $ABCD$ opposite B*; that is, it is tangent to ray BA (beyond A) at U_A, ray BC (beyond C) at U_C, ray AD (beyond D) at V_A, ray CD (beyond D) at V_C. Then $BA + AD = BC + CD$.

   The result of the problem follows from these facts. Circle ω_2 touches sides AD, DC, CA at S_C, S_A, S_D, respectively. Since ω_2 is the incircle of triangle ADC, we know that $AS_D = (AC + AD - CD)/2$. By facts 6a and 6b conclude that

$AS_D = (AC + AD - CD)/2 = (AC + BC - AB)/2 = CT_B = AB_2$, that is, $B_2 = S_D$. By the fact 6a, points B, B_1, S_D are collinear. Ray BB_1 meets minor arc U_AU_C (which is part of ω) at H_1. Construct points A_4 and C_4 on rays BA and BC, respectively, such that $A_4C_4 \parallel AC$ and H_1 lies on A_4C_4. Since $S_D = B_2$, the excircle of triangle BCA opposite B is tangent to CA at B_2. Consider the homothety $\mathbf{H}_1$ centered at B sending AC to A_4C_4. Since $\mathbf{H}_1(B_2) = H_1$, circle ω is tangent to A_4C_4 at H.

Let segment A_4C_4 intersect segments AA_3 and CC_3 at A_6 and C_6. Consider the homothety $\mathbf{H}_2$ centered at D sending AC to A_6C_6. Let H_2 denote the image of T_2 under $\mathbf{H}_2$. Since $CT_B = AS_D$, applying the fact (a) to triangle DAC implies that the excircle of triangle DAC opposite D is tangent to side CA at T_2. Hence the excircle of triangle DA_6C_6 is tangent to side A_6C_6 at H_2. It is clear that $\mathbf{H}_2(\omega_1) = \omega$. It follows that ω is tangent to lines AC at both H_2 and H_1; that is, we may set $H = H_1 = H_2$ and H lies on both lines T_2D and B_1S_D.

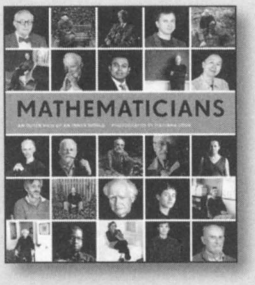
It suffices to show that H is the intersection of the common external tangent lines of ω_1 and ω_2; that is, H is a exterior center of homothety of the 2 circles. Let D_1 be the point on ω_2 diametrically opposite S_D. In the view of the fact (a), D_1 lies on segment DT_B. Therefore, lines B_1S_D and T_BD_1 meet at H. Since B_1T_B and S_DD_1 are two parallel diameters of ω_1 and ω_2, the intersections of lines B_1S_D and T_BD_1 is either the center of the interior homothety or the center of the exterior homothety of the two circles. It is clear that the centers of the ω_1 and ω_2 lie on the same side of H. We conclude H is the center of the exterior homothety, completing our proof.

2008 International Mathematical Olympiad Results The USA team members were chosen according to their combined performance on the 37th annual USAMO and the Team Selection Test. Members of the USA team at the 2008 IMO were Paul Christiano, Shaunak Kishore, Evan O'Dorney, Colin Sandon, Krishanu Sankar, and Alex Zhai. Zuming Feng and Răzvan Gelca served as team leader and deputy leader, respectively accompanied by Ian Le and Steven Dunbar as observers.

The 2008 International Mathematical Olympiad took place in Madrid, Spain July 10-22, 2008 with 537 competitors from 99 countries. Three students had perfect papers: Xiaosheng Mu and Dongyi Wei of China, and Alex Zhai of the USA. The USA team sponsored by the MAA won the following medals:

- Paul Christiano, a graduate of The Harker School, Saratoga, CA won a Silver Medal.
- Shaunak Kishore, a graduate of Unionville High School in Kennet Square, PA received a Gold medal.
- Evan O'Dorney, who attends the Venture School and is from Danville CA, received a Silver medal.
- Colin Sandon who graduated from Essex High School in Essex Junction, VT won a Gold medal.
- Krishanu Roy Sankar who graduated from Horace Mann Hill High School in Hastings-on-Hudson, NY won a Gold medal.
- Alex Zhai, who graduated from University Laboratory High School in Champaign, IL won a Gold medal.

More results and statistics are at the official IMO website: `www.imo-offical.org`.



Mathematicians

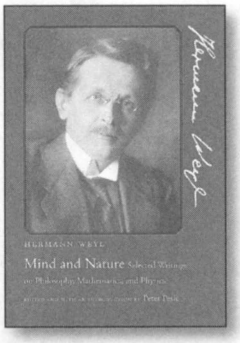
An Outer View of the Inner World

Mariana Cook

With an introduction by R. C. Gunning

Mathematicians is a remarkable collection of ninety-two photographic portraits, featuring some of the most amazing mathematicians of our time. Acclaimed photographer Mariana Cook captures the exuberant and colorful personalities of these brilliant thinkers and the superb images are accompanied by brief autobiographical texts written by each mathematician.

Cloth $35.00 978-0-691-13951-7 July



Mind and Nature

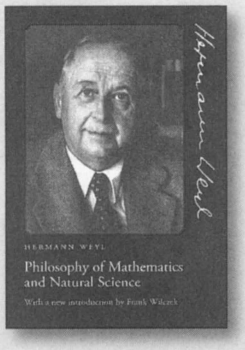
Selected Writings on Philosophy, Mathematics, and Physics

Hermann Weyl

Edited and with an introduction by Peter Pesic

Hermann Weyl was one of the twentieth century's most important mathematicians, as well as a seminal figure in the development of quantum physics and general relativity. He was also an eloquent writer with a lifelong interest in the philosophical implications of the startling new scientific developments with which he was so involved. *Mind and Nature* is a collection of Weyl's most important general writings on philosophy, mathematics, and physics.

Cloth $35.00 978-0-691-13545-8



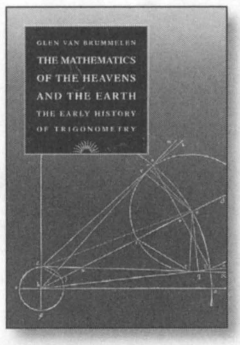
Philosophy of Mathematics and Natural Science

Hermann Weyl

With a new introduction by Frank Wilczek

Drawing on work by Descartes, Galileo, Hume, Kant, Leibniz, and Newton, Weyl provides readers with a guide to understanding science through the lens of philosophy. This is a book that no one but Weyl could have written—and, indeed, no one has written anything quite like it since.

New in Paper $35.00 978-0-691-14120-6



The Mathematics of the Heavens and the Earth

The Early History of Trigonometry

Glen Van Brummelen

This is the first major history in English of the origins and early development of trigonometry. No other book offers the historical breadth, analytical depth, and coverage of non-Western mathematics that readers will find in *The Mathematics of the Heavens and the Earth.*

Cloth $39.50 978-0-691-12973-0

EXPLORE FULBRIGHT

Fulbright Scholar Program 2010–2011

The Fulbright Scholar Program offers **23 awards in lecturing, research or lecturing/research in mathematics**. This includes **two Fulbright Distinguished Chairs**. Faculty and professionals in mathematics also can apply for one of the **144 "All Discipline"** awards open to all fields.

Here are a few awards for mathematicians:

Israel – Award #0024: Fulbright-Israel Distinguished Chair in the Natural Sciences and Engineering

Rwanda – Award #0096: applied mathematics, programming and software, linear programming, modeling, differential equations

Ethiopia – Award #0065: statistics

Portugal – Award #0353: classical theory of partial differential equations and numerical approximation to their solutions or computational algebra or combinatronics – algebraic and/or probabilistic aspects.

Ireland – Award #0285: mathematical economics, mathematical modeling

Deadline: August 1, 2009

Council for International Exchange of Scholars
3007 Tilden Street, NW, Suite 5L
Washington, DC 20008-3009
Tel: 202.686.7877 • E-mail: scholars@cies.iie.org
www.CIES.org

FULBRIGHT

The Fulbright Program is sponsored by the United States Department of State, Bureau of Educational and Cultural Affairs. For more information, visit fulbright.state.gov.

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

The Mathematical Association of America
1529 Eighteenth Street, NW
Washington, DC 20036