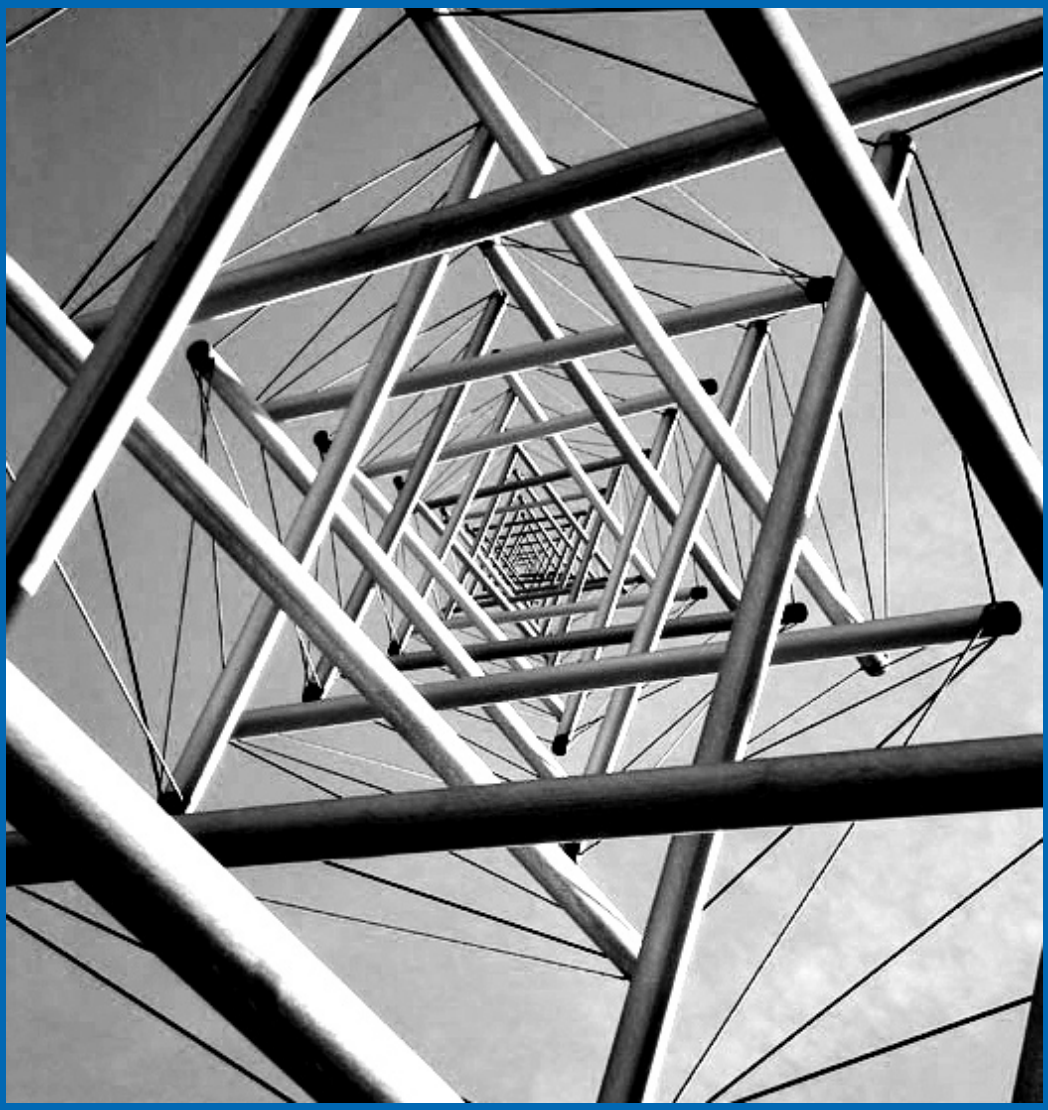
Vol. 84, No. 3, June 2011

MAA

MATHEMATICS MAGAZINE



Kenneth Snelson, "Needle Tower"

- Fibonacci and Budding Yeast
- Branching Blood Vessels
- Sylvester, Stirling, IEEE Floating Point, more . . .

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

EDITORIAL POLICY

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

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

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

Cover: Looking upward from underneath Kenneth Snelson's "Needle Tower" tensegrity sculpture in Washington, D.C., reveals a striking symmetry. Cover photo by I. Peterson.

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

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

Advertising correspondence should be addressed to

MAA Advertising
1529 Eighteenth St. NW
Washington DC 20036

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

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

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

MAA Service Center, *maahq@maa.org*

All at the address:

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

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

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

Printed in the United States of America

Vol. 84, No. 3, June 2011

MAA

MATHEMATICS MAGAZINE

LETTER FROM THE EDITOR

This issue's cover image is from the MAA Found Math Gallery, a feature of the MAA's constantly-improving website. The image is of Kenneth Snelson's sculpture "Needle Tower," near the Hirshhorn Museum in Washington, D.C. (Pop quiz: Would this structure hold its shape in space?) For more mathematically engaging photographs, go to www.maa.org and find the link near the center of the page.

The articles in this issue all involve applications, but in different ways. Perhaps the most direct is John Adam's article on branching blood vessels. Many factors influence the branching angles; in modeling we try to identify the most important factors. When we are successful, we can explain the relationships that we can observe and make predictions about relationships that we cannot observe. The article shows these processes in action.

Olofsson and Daileda apply their model in a slightly different way. They use a Fibonacci-like sequence to explain growth rates of yeast colonies. The model includes a parameter that is related to a hard-to-observe feature of the yeast chromosomes. They connect the parameter to the growth rates. Then by observing the growth rates, they find clues about the chromosomes.

Ray Rosentrater's application is to computer design—specifically, to the representation of floating-point numbers. Here, mathematics can be used to optimize. Is the representation efficient? We learn how to define the redundancy of a representation, and to calculate it. Finally, the application in the article by Eisenberg and Sullivan is within mathematics. Probability is applied to the precise measurement of shapes. The authors introduce a notion of "rotundity." A (round) disk scores highest on this scale; needle-like shapes, not so high.

Do you know the trapezoid rule for estimating integrals? Do you know its error estimate? Paul Levrie states and proves the error estimate in his Note, and then uses it in an unexpected way—to prove Stirling's formula for $n!$. Elsewhere in the Notes we can learn more about disjoint covering systems, continuous and discrete convexity, and what happens when eigenvalues meet fractals.

Are you looking for the usual June feature on the IMO? It was in the October, 2010 issue. You may not have to wait long for the next one. As this issue goes to press, the 2011 USAMO/USAJMO is being graded, and the world's IMO teams are being formed. If the committees can meet some tight deadlines, you will be able to read about these contests in our October, 2011 issue.

Walter Stromquist, Editor

Budding Yeast, Branching Processes, and Generalized Fibonacci Numbers

PETER OLOFSSON
Trinity University
San Antonio, TX 78212
polofsso@trinity.edu

RYAN C. DAILEDA
Trinity University
San Antonio, TX 78212
rdaileda@trinity.edu

The Fibonacci sequence is famous for showing up in nature in many ways, some of them idealized (like Fibonacci's breeding rabbits) and some very practical. In this article we describe an application of a generalization of the Fibonacci sequence, called the "k-nacci numbers." In Olofsson and Bertuch [**6**], a branching process model was used to analyze experiments on growing yeast populations, and it turned out that the k-nacci numbers were crucial to practical results and calculations. That paper mainly addressed the biological problems. Our aim in this article is to elaborate on some of the mathematics involved, and in so doing, to give a nice example of how asymptotic results about simple recurrences can be of significant practical use.

Budding yeast

The yeast *Saccharomyces cerevisiae* is used in baking and brewing, and is also one of the most important model organisms in biology. It is a one-celled organism that reproduces through *budding*, meaning that a new yeast cell starts to grow on the surface of an existing cell, eventually separating from its mother as a newborn daughter cell. This reproduction scheme is different from *binary fission*, common in many bacteria such as *E. Coli*, where the cell divides into two new cells of equal size. Although both reproduction schemes give rise to *clones*, that is, cells that are genetically identical (save for mutations), there are differences that matter to the mathematical modeling of the resulting population growth.

Since yeast has linear chromosomes just like human beings, its genetics can be studied for greater insight into human genetics. One example is the study of the shortening of chromosomal ends known as *telomeres* that occurs also in many of our cells. With each cell division, telomeres become progressively shorter until they reach a point at which the cell stops dividing, to avoid damage to the coding DNA in the interior of the chromosome. A cell that has stopped dividing is said to be *senescent*. To counteract telomere shortening, some cells, for example embryonic stem cells, contain the enzyme *telomerase* which adds telomere sequences so that the chromosomes can maintain a stable telomere length. Some cells manage to keep replicating even without telomerase; such populations were studied in Olofsson and Bertuch, who used branching processes to model populations of yeast cells and to estimate cell population parameters from laboratory data.

Math. Mag. **84** (2011) 163–172. doi:10.4169/math.mag.84.3.163.

Branching processes

A branching process is a stochastic model for a proliferating population. Assumptions are made about individual cell lifetimes and reproduction patterns, and conclusions are drawn about population behavior. In this section we introduce a branching process that applies to cells (such as *E. Coli*) that use binary splitting. We then show how it can be modified to apply to cells (such as yeast) that reproduce by budding.

We begin with a single ancestral cell, which we call *generation* 0. Then, for $n \geq 0$, each cell in generation n divides into two cells in generation $n+1$. Randomness enters via the cell lifetimes (birth to division). These are random variables, independent and identically distributed, with a common cumulative distribution function (cdf) F. That is, if T is a cell lifetime, then

$$\Pr(T \leq t) = F(t) \text{ for all } t \geq 0.$$

In general, if two independent random variables have cdf's F and G, then the cdf of their sum is the *convolution* of F and G, defined by

$$(F * G)(t) = \int_{s=0}^{t} G(t-s)\, dF(t) \text{ for all } t \geq 0.$$

(If F has a derivative, one may read "$dF(t)$" as a synonym for $F'(t)dt$.) Therefore, if F is the cdf for cell lifetimes, the sum of n consecutive cell lifetimes has cdf equal to the n-fold convolution of F with itself, which we denote by F^{*n}.

We would like to derive an expression for $M(t)$, the expected number of individuals present at time t. First note that there are 2^n cells in the nth generation. Such a cell is present at time t if the sum of n cell lifetimes (those of its ancestors) is less than t and the sum of $n+1$ cell lifetimes (its ancestors and itself) is greater than t. Thus, the probability that an nth-generation cell is present at time t equals $F^{*n}(t) - F^{*(n+1)}(t)$ and the expected number of cells in generation n that are present at time t equals

$$2^n \left(F^{*n}(t) - F^{*(n+1)}(t)\right).$$

This formula applies to generation 0 as well, if we follow the usual convention that $F^{*0}(t) = 1$. Now, summing over all generations gives the expression

$$M(t) = \sum_{n=0}^{\infty} 2^n \left(F^{*n}(t) - F^{*(n+1)}(t)\right). \tag{1}$$

So far, we have assumed that all cells survive to reproduce. If we instead assume that each cell survives to reproduce with probability p, and otherwise dies (after its random lifetime) without reproducing, then the mean number of offspring per cell is $m = 2p$. We always assume that $p > 1/2$, so that $m > 1$ and the generations (on average) increase in size. Now, in place of (1) we get the expression

$$M(t) = \sum_{n=0}^{\infty} m^n \left(F^{*n}(t) - F^{*(n+1)}(t)\right). \tag{2}$$

Adopt the standard notation $x(t) \sim y(t)$ if $x(t)/y(t) \to 1$ as $t \to \infty$. It turns out that the asymptotic growth rate of $M(t)$ is given by

$$M(t) \sim be^{\alpha t} \text{ as } t \to \infty \tag{3}$$

where the constants α and b are determined by the lifetime cdf F and the mean number of offspring per individual m. Specifically, let

$$\widehat{F}(s) = \int_0^\infty e^{-st}\, dF(t)$$

the Laplace transform of the probability measure associated with F. The growth rate α, called the *Malthusian parameter*, is defined through the relation

$$m\widehat{F}(\alpha) = 1$$

where $m > 1$ implies that there exists a unique solution $\alpha > 0$. The constant b can be shown to equal

$$b = (m-1)\left(\alpha m^2 \int_0^\infty t e^{-\alpha t}\, dF(t)\right)^{-1}.$$

We refer to Chapter VI in Harris [**4**] or Jagers and Nerman [**5**] for proofs and further details.

Yeast, as we have noted, does not reproduce by binary splitting but rather by budding. Each cell in generation n still gives rise to two cells, and although one is the mother and the other is the daughter, after the division we still refer to both of them as generation $n+1$. In this sense there is no difference from binary splitting; but it is known that a mother cannot give birth to an unlimited number of daughters, so the size of the nth generation is no longer given by m^n.

We can still obtain a formula for $M(t)$, if we let $m(n)$ represent the number of cells in the nth generation. We then have, for yeast,

$$M(t) = \sum_{n=0}^{\infty} m(n)\left(F^{*n}(t) - F^{*(n+1)}(t)\right). \tag{4}$$

Growth rate and other asymptotic properties of $M(t)$ are now determined by F together with $m(n)$. Everything depends on getting a handle on the numbers $m(n)$. We address this task in the next section.

Generalized Fibonacci numbers

By the k-nacci sequence, for $k \geq 2$, we mean the sequence $\{F_j,\ j \geq 0\}$ defined by $F_0 = F_1 = \cdots = F_{k-2} = 0$, $F_{k-1} = 1$, and $F_n = F_{n-1} + F_{n-2} + \cdots + F_{n-k}$ for $n \geq k$. For example, when $k = 4$, each term is the sum of the four previous terms:

$$0, 0, 0, 1, 1, 2, 4, 8, 15, 29, 56, \ldots$$

In the yeast population, suppose that a mother cell can have k daughter cells before she stops reproducing. (In biology, k is known as the *proliferative lifespan.*) In any given generation n, cells can be divided into classes describing how many more daughter cells they can have. Thus, let $N_j^{(n)}$ be the number of cells in generation n that can have exactly j additional daughter cells for $j = 0, 1, \ldots, k$. The class with $j = 0$ is the class of senescent cells, and we assume that they stay in the population indefinitely (although it is easy to model a scenario where they eventually die and disappear). The class with $j = k$ are the newborn cells that have yet to reproduce. As it turns out, the numbers of cells in these classes are precisely described by the k-nacci sequence.

PROPOSITION 1. *Consider the vector* $(N_0^{(n)}, N_1^{(n)}, \ldots, N_k^{(n)})$ *in the* nth *generation for* $n \geq 1$ *of the branching process above. Let* F_i *denote the* ith k*-nacci number and let* $S_n = F_0 + F_1 + \cdots + F_n$. *Then* $(N_0^{(n)}, N_1^{(n)}, \ldots, N_k^{(n)})$ *equals*

$$(S_{n-1}, F_n, F_{n+1}, \ldots, F_{n+k-1}) .$$

Proof. In generation 0 there is one cell that is able to divide k more times which gives the vector $(0, 0, \ldots, 0, 1)$ for generation 0. Each cell with $j \geq 1$ produces a daughter cell that is able to reproduce k times and is then itself able to reproduce another $j - 1$ times. Cells with $j = 0$ remain unchanged. Thus, each class with $j \geq 1$ feeds into the class $j - 1$ immediately below it, and also into the highest class k. The transition from generation $n - 1$ to generation n can be described as follows:

$$\begin{cases} N_0^{(n)} = N_0^{(n-1)} + N_1^{(n-1)} \\ N_{j-1}^{(n)} = N_j^{(n-1)} \text{ for } 2 \leq j \leq k \\ N_k^{(n)} = \displaystyle\sum_{j=1}^{k} N_j^{(n-1)} \end{cases}$$

and the proposition follows. ∎

The total number, $m(n)$, of cells in the nth generation equals

$$\begin{aligned} m(n) &= S_{n-1} + F_n + \cdots + F_{n+k-1} \\ &= S_{n+k-1} \end{aligned}$$

and Proposition 1 provides a recursive scheme that enables us to compute $m(n)$. For example, if $k = 4$, the first terms in the sequence $\{m(n), n \geq 0\}$ are

$$1, 2, 4, 8, 16, 31, 60, 116, 224, \ldots$$

where we recognize the powers of 2 until the 4th generation ($n = 4$) after which the effect of the proliferative lifespan $k = 4$ becomes noticeable and slows down the growth. By (4), we can also compute the expected number $M(t)$ of cells at each time t which enables us to compare the model with laboratory data and estimate unknown parameters. As it turns out, we can even get an explicit expression for $m(n)$, expressed in terms of k-nacci numbers which is crucial to establish asymptotics of the branching process.

In the next section, we study the k-nacci numbers as a special case of linear recurrence.

Linear recurrences

Given a positive integer k and complex numbers $a_0 \neq 0, a_1, \ldots, a_{k-1}$ consider a sequence $\{R_n\}_{n=0}^{\infty}$ that satisfies the k-term linear recurrence

$$R_n = a_{k-1}R_{n-1} + a_{k-2}R_{n-2} + \cdots + a_0 R_{n-k} \tag{5}$$

for all $n \geq k$. Given initial values $R_0, R_1, \ldots, R_{k-1}$, equation (5) defines the sequence $\{R_n\}_{n=0}^{\infty}$ recursively. Define the *characteristic polynomial* of (5) to be

$$p(x) = x^k - a_{k-1}x^{k-1} - a_{k-2}x^{k-2} - \cdots - a_0. \tag{6}$$

If the polynomial (6) has k distinct roots $r_1, r_2, \ldots, r_k$ then there are unique complex coefficients $b_1, b_2, \ldots, b_k$ so that

$$R_n = b_1 r_1^n + b_2 r_2^n + \cdots + b_k r_k^n \tag{7}$$

for all $n \geq 0$. The coefficients $b_1, \ldots, b_k$ can be computed from the initial values $R_0, \ldots, R_{k-1}$. For details, and for the case of repeated roots, see Elaydi [**1**, Section 2.3].

Given a sequence $\{R_n\}_{n=0}^{\infty}$ that satisfies (5), we let $S_n = R_0 + R_1 + \cdots + R_n$ for $n \geq 0$. The following proposition establishes a closed form expression for S_n.

PROPOSITION 2. *Consider the linear recurrence in* (5). *Assume that the characteristic polynomial has k distinct roots, none of which equals* 1. *Then there exist constants* $c_0, \ldots, c_{k-1}$ *such that*

$$S_n = \sum_{l=0}^{k-1} c_l R_{n+l+1} - \sum_{l=0}^{k-1} c_l R_l.$$

Proof. Begin by expressing R_n as given in (7). Then we have

$$\begin{aligned} S_n &= \sum_{i=0}^{n} R_i \\ &= \sum_{i=0}^{n} \sum_{j=1}^{k} b_j r_j^i \\ &= \sum_{j=1}^{k} b_j \sum_{i=0}^{n} r_j^i \\ &= \sum_{j=1}^{k} b_j \frac{r_j^{n+1} - 1}{r_j - 1}. \end{aligned}$$

Since the characteristic polynomial $p(x)$ does not have 1 as a root, $p(x)$ and $x - 1$ are relatively prime so that we can find polynomials $u(x)$ and $v(x)$ which satisfy

$$v(x)p(x) + u(x)(x - 1) = 1. \tag{8}$$

Moreover, by using the division algorithm if necessary, we can assume that the degree of $u(x)$ is less than k. Substituting any of the roots r_j of $p(x)$ into (8) immediately yields

$$\frac{1}{r_j - 1} = u(r_j).$$

It now follows that

$$S_n = \sum_{j=1}^{k} b_j (r_j^{n+1} - 1) u(r_j).$$

If we write $u(x) = c_{k-1}x^{k-1} + \cdots + c_0$, this becomes

$$\begin{aligned} S_n &= \sum_{l=0}^{k-1} c_l \sum_{j=1}^{k} b_j (r_j^{n+l+1} - r_j^l) \\ &= \sum_{l=0}^{k-1} c_l (R_{n+l+1} - R_l) \\ &= \sum_{l=0}^{k-1} c_l R_{n+l+1} - \sum_{l=0}^{k-1} c_l R_l. \end{aligned} \tag{9}$$

which concludes the proof. ∎

In the expression for S_n, note that the first sum includes at most k terms of the sequence $\{R_n\}_{n=0}^{\infty}$, while the second sum depends only on the initial conditions $R_0, R_1, \ldots, R_{k-1}$.

As an example, we apply this result to the Fibonacci numbers, which are simply the $k = 2$ case of the k-nacci numbers. The characteristic polynomial in this case is $p(x) = x^2 - x - 1$, which satisfies $-p(x) + x(x-1) = 1$. Hence, $u(x) = x$ so that Proposition 2 becomes the familiar result $F_1 + F_2 + \cdots + F_n = F_{n+2} - F_1 = F_{n+2} - 1$.

As a corollary we obtain the corresponding result for the k-nacci numbers.

COROLLARY 3. *For the k-nacci sequence* $\{F_j, j \geq 0\}$, *let* $S_n = F_0 + F_1 + \cdots + F_n$. *Then*

$$S_n = \frac{1}{k-1}\left(F_{n+k} - \sum_{l=0}^{k-3}(k-l-2)F_{n+l+1} - 1\right).$$

Proof. To get an expression for the polynomial $u(x)$ which determines the coefficients c_l in Proposition 2, note that for $k \geq 2$ we have

$$\begin{aligned} 1 = &\frac{-1}{k-1}\left(x^k - x^{k-1} - x^{k-2} - \cdots - 1\right) \\ &+ \frac{1}{k-1}\left(x^{k-1} - x^{k-3} - 2x^{k-4} - 3x^{k-5} - \cdots - (k-2)\right)(x-1) \end{aligned}$$

which identifies $u(x)$ as

$$u(x) = \frac{1}{k-1}\left(x^{k-1} - \sum_{l=0}^{k-3}(k-l-2)x^l\right),$$

provided we treat the sum as empty when $k = 2$. Clearly the characteristic polynomial $p(x) = x^k - (x^{k-1} + x^{k-2} + \cdots + 1)$ does not have 1 as a root. Regarding the distinctness of the roots of $p(x)$, observe that

$$\begin{aligned} p(x) &= x^k - \frac{x^k - 1}{x - 1} \\ &= \frac{x^{k+1} - 2x^k + 1}{x - 1} \end{aligned} \tag{10}$$

and the polynomial in the numerator has no repeated roots, as it does not share any roots with its derivative (see Gallian [**3**]). Finally, since $F_0 = F_1 = \cdots = F_{k-2} = 0$ and $F_{k-1} = 1$, the result follows from Proposition 2. ∎

Asymptotics of the branching process

The asymptotic results in this section rely on the fact that k-nacci numbers have asymptotic geometric growth. Following Flores [**2**], there exist numbers r and A such that

$$F_j \sim Ar^j \tag{11}$$

as $j \to \infty$, meaning that $F_j/r^j \to A$ as $j \to \infty$. (This result follows from (7).) The number r is the dominant root of the characteristic equation

$$x^k - x^{k-1} - \cdots - x - 1 = 0$$

(that is, the root with the largest absolute value) and it is known to be real and to lie between the golden ratio $\phi \approx 1.618$ and 2. In fact, for $k = 2$, $r = \phi$ and as $k \to \infty$, $r \uparrow 2$. The constant A equals

$$A = \frac{r-1}{(k+1)r^k - 2kr^{k-1}} \tag{12}$$

By Proposition 1, Corollary 3, and (11), we can establish expressions for the asymptotic composition of cells in the different classes, recalling that class j contains cells that can have j additional daughter cells, $j = 0, 1, \ldots, k$. Of particular practical interest is the proportion of cells in the 0-class and we state its simple asymptotic expression.

PROPOSITION 4. *Let r be as above. The asymptotic proportion of cells in the* 0*-class as $n \to \infty$ equals r^{-k}.*

Proof. By Proposition 1, the proportion equals

$$\frac{S_{n-1}}{S_{n+k-1}}$$

and the proposition follows from Corollary 3 and (11). ∎

Proposition 4 is not just a theoretical limit result; it has important practical implications for the yeast cell population studies. Certain computational expressions become greatly simplified if the finite proliferative lifespan can be neglected, i.e., if we assume that each cell can produce an unlimited number of daughter cells. Since the fraction of senescent cells in a given generation n is roughly r^{-k}, this number can be used to justify such an approximation. For example, for the regular Fibonacci sequence with $k = 2$, we have $r = \phi$ and since $\phi^{-2} \approx 0.38$, as many as 38% of cells have reached the end of their proliferative lifespan and are no longer able to produce daughter cells. In this case, the approximation would not work very well. Note that, since $r \uparrow 2$ as $k \to \infty$, the "∞-nacci" sequence has $r = 2$ and corresponds to a binary splitting branching process where each individual can produce an unlimited number of offspring.

For yeast cells, the proliferative lifespan k has been estimated to be on average 25, Sinclair et al. [**7**], which gives a value of r that for all practical purposes equals 2 and the fraction of senescent cells is less than one in 10 million. For any reasonable duration of a yeast cell experiment, this fraction is negligible although it does of course

matter to the theoretical asymptotic limits. Calculations show that r exceeds 1.99 already for $k = 7$ in which case less than 1% of cells are senescent. For the particular yeast populations studied by Olofsson and Bertuch, k is likely to largely exceed 7 and the approximation works well.

Finally, we obtain the asymptotic growth rate of $M(t)$. As we will soon show, $m(n)$ grows asymptotically as r^n, that is, at the same rate as a binary splitting process with mean number of offspring equal to $m = r$. Since $m = 2p$, in the latter process each cell survives to reproduce with probability $p = r/2$, and dies without reproducing with probability $1 - r/2$. The next result shows that the Malthusian parameter is the same for the budding process with generation sizes $m(n)$ and the binary splitting process with mean r, but that the budding process always has a larger expected value.

PROPOSITION 5. *As $t \to \infty$, $M(t) \sim Cbe^{\alpha t}$ where α and b are as in* (3) *with mean number of offspring $m = r$. The constant C depends on k and satisfies $C > 1$ and $C \to 1$ as $k \to \infty$.*

Proof. Since $m(n) = S_{n+k-1}$, Corollary 3 and (11) yield

$$m(n) \sim Cr^n$$

where

$$C = \frac{A}{k-1}\left(r^{2k-1} - \sum_{l=0}^{k-3}(k-l-2)r^{k+l}\right) \tag{13}$$

A being the constant defined in (12). If we, informally, substitute this expression for $m(n)$ in (4), we get

$$M(t) \sim C\sum_n r^n \left(F^{*n}(t) - F^{*(n+1)}(t)\right) \tag{14}$$

as $t \to \infty$. By (2) and (3), we get

$$M(t) \sim Cbe^{\alpha t}$$

as $t \to \infty$, as desired.

To prove that the substitution leading to (14) is indeed legitimate, let us formally prove that

$$e^{-\alpha t}M(t) \to Cb$$

as $t \to \infty$. Recall that $m(n) \sim Cr^n$, that is, $m(n)/r^n \to C$ as $n \to \infty$. Choose N such that

$$C - \epsilon \le \frac{m(n)}{r^n} \le C + \epsilon$$

for $n > N$. For ease of notation, let

$$P(n,t) = F^{*n}(t) - F^{*(n+1)}(t)$$

for $n = 0, 1, 2, \ldots$, and note that $0 \le P(n,t) \le 1$ for all n and t. We now get

$$\begin{aligned} e^{-\alpha t}M(t) &= e^{-\alpha t}\sum_{n=0}^{\infty} m(n)P(n,t) \\ &= e^{-\alpha t}\sum_{n=0}^{N} m(n)P(n,t) + e^{-\alpha t}\sum_{n=N+1}^{\infty} m(n)P(n,t) \end{aligned}$$

where the first term goes to 0 as $t \to \infty$ and hence the limit of $e^{-\alpha t}M(t)$ is the same as that of the second term, for which we have

$$(C-\epsilon)e^{-\alpha t}\sum_{n=N+1}^{\infty} r^n P(n,t) \le e^{-\alpha t}\sum_{n=N+1}^{\infty} m(n)P(n,t) \le (C+\epsilon)e^{-\alpha t}\sum_{n=N+1}^{\infty} r^n P(n,t)$$

Let $t \to \infty$ and use (3) to obtain

$$(C-\epsilon)b \le \lim_{t\to\infty} e^{-\alpha t}M(t) \le (C+\epsilon)b$$

Since ϵ was arbitrary, we conclude that

$$\lim_{t\to\infty} e^{-\alpha t}M(t) = Cb$$

To prove that $C > 1$, first note that by (12) and some algebra the constant

$$C = \frac{A}{k-1}\left(r^{2k-1} - \sum_{l=0}^{k-3}(k-l-2)r^{k+l}\right)$$

simplifies to

$$C = \frac{r(r^{k+1} - 2r^k + kr - k - r + 2)}{(r-1)(kr+r-2k)(k-1)}$$

From (10) we see that

$$r^{k+1} - 2r^k + 1 = 0$$

which simplifies C further, to

$$C = \frac{r}{k(r-2)+r}$$

and since $k \ge 2$ and $\phi \le r < 2$, we have $C > 1$. Further, (10) yields

$$r^{k+1} - 2x^r + 1 = 0$$

and since $\phi \le r < 2$, we get the inequalities

$$2 - \phi^{-k} \le r < 2$$

which reveals that $C \to 1$ as $k \to \infty$. ∎

Proposition 5 shows that the branching process for budding yeast with proliferative lifespan k, asymptotically grows at the same rate as a binary splitting process with mean number of daughter cells equal to r, in the sense of having the same Malthusian parameter α. However, since the binary splitting population grows as $be^{\alpha t}$ and the budding population as $Cbe^{\alpha t}$ where $C > 1$, the population with the budding process tends to be, on average, larger than the population with the splitting process. Also note that in the binary splitting population, each individual has a positive probability $1 - r/2$ of 0 offspring; hence, this population can go extinct, which is not possible for the budding population. As $k \to \infty$, $C \to 1$ and $r \uparrow 2$ so in the limit, budding and binary splitting are equivalent which makes intuitive sense.

Acknowledgment Peter Olofsson was funded by NIH research grant 1R15GM093957-01.

REFERENCES

1. S. Elaydi, *An Introduction to Difference Equations*, 3rd ed. Springer, New York, 2010.
2. I. Flores, Direct calculation of k-generalized Fibonacci numbers, *The Fibonacci Quarterly* **5** (1967) 259–266.
3. J. Gallian, *Contemporary Abstract Algebra*, 7th ed. Brooks/Cole, Belmont, CA, 2010.
4. T. E. Harris, *The Theory of Branching Processes*, Springer, Berlin, 1963; also, Dover, 2002.
5. P. Jagers and O. Nerman, The growth and composition of branching populations, *Adv. Appl. Prob.* **16** (1984) 221–259. doi:10.2307/1427068
6. P. Olofsson and A. A. Bertuch, Modeling growth and telomere dynamics in Saccharomyces cerevisiae, *J. Theor. Biol.* **263(3)** (2010) 353–359. doi:10.1016/j.jtbi.2009.12.004
7. D. A. Sinclair, K. Mills, and L. Guarente, Aging in Saccharomyces cerevisiae, *Annual Rev. Microbiol.* **52** (1998) 533–560. doi:10.1146/annurev.micro.52.1.533

Summary An application of branching processes to a problem in cell biology is described, in which the generalized Fibonacci numbers known as k-nacci numbers play a crucial role. The k-nacci sequence is used to obtain asymptotics, computational formulas, and to justify certain practical simplifications of the biological model. Along the way, an explicit formula for the sum of k-nacci numbers is established.

PETER OLOFSSON is Professor and Chair of the Mathematics Department at Trinity University in San Antonio, TX. His main area of research is on the theory and biological applications of branching processes. He is the author of the textbook "Probability, Statistics, and Stochastic Processes" and the general audience book "Probabilities: the Little Numbers that Rule Our Lives." He has also written articles criticizing the use of mathematics and statistics by proponents of "intelligent design."

RYAN DAILEDA is currently an Associate Professor in the Mathematics Department at Trinity University. Before moving to south Texas, he earned his B.A. and Ph.D. in southern California, at the University of Redlands and the University of California, Los Angeles, respectively. When he's not thinking about number theory, he enjoys spending time with his wife, Jennifer, and their lovable dog, Rusty.

A Modification of Sylvester's Four Point Problem

BENNETT EISENBERG
Lehigh University
Bethlehem, PA 18015
BE01@lehigh.edu

ROSEMARY SULLIVAN
West Chester University
West Chester, PA 19383
rsullivan@wcupa.edu

Lewis Carroll posed the question "What is the probability that a random triangle is obtuse?" Another classical problem of geometric probability is "What is the probability that a random quadrilateral is convex?" Both problems suffer from the weakness that they have no solution. The meaning of "random triangle" or "random quadrilateral" is ambiguous. If the vertices of the figures are selected at random in the plane, what distribution should be used? The distribution that everyone prefers is the uniform distribution, but there is no uniform probability distribution in the plane.

In place of a uniform distribution in the plane Sylvester in 1865 posed the problem of finding the probability that four points chosen independently from a uniform distribution over a compact convex region K in the plane form the vertices of a convex quadrilateral. Since the answer doesn't change if the region is dilated, the solution to this problem would provide a partial solution to the original question of "What is the probability that a random quadrilateral is convex?" The problem for the compact convex region in turn reduces to that of finding the expected value of the area of a random triangle whose vertices are chosen at random in K. This is because a quadrilateral fails to be convex precisely when one of the four vertices lies in the triangle formed by the other three. Denote the triangle formed by three random points as Δ. It follows that the probability that a random quadrilateral in K is convex is given by

$$P_K(\text{quadrilateral is convex}) = 1 - 4\frac{\text{E}(\text{area}(\Delta))}{\text{area}(K)}.$$

The formula comes from finding the probability that the fourth point lies inside the triangle formed by the first three by first conditioning on the area of the triangle formed by the first three. The factor 4 arises because there are four possible fourth points.

Define

$$\rho_K = \frac{\text{E}(\text{area}(\Delta))}{\text{area}(K)}.$$

The problem of finding ρ_K for different regions K has generated substantial research. Cayley and Sylvester showed that $\rho_K = 1/12 \approx .083$ for a triangle and Woolhouse showed that for a parallelogram $\rho_K = 11/144 \approx .076$ and that $\rho_K = 35/(48\pi^2) \approx .074$ when K is a circle or ellipse. The value of ρ_K is clearly unchanged when K is transformed by an affine transformation since the numerator and denominator are multiplied by the same constant. It was eventually proved that the minimum and maximum values of ρ_K over all K occur for the circle and triangle, respectively. This variational problem along with Sylvester's original question are known as Sylvester's

Math. Mag. **84** (2011) 173–184. doi:10.4169/math.mag.84.3.173.

Four Point Problem. There are more details about the history of this problem along with the historical references in Pfiefer [**2**] and Weisstein [**4**].

In all of this work the random vertices of the triangle are chosen over bounded regions in the plane. This is not necessary if the distribution of the points is not required to be uniform. For a general distribution of the points, solving Sylvester's original problem would require finding the probability that a random point is contained in a random triangle with vertices given by three other random points with the given distribution. We do not tackle this difficult problem here. Instead we consider an easier interesting problem related to the previous work.

The modified problem

We consider the problem of minimizing or maximizing the expected area of a random triangle when the vertices are chosen independently from a probability distribution in the plane. As with the original Sylvester problem, the more dispersed the distribution, the greater the expected area of the triangle, so we have to normalize the expected area of the triangle by something analogous to the area of a convex set. A measure of the spread of the distribution is given by the expected squared distance between two random points from the distribution. Normalizing by this is analogous to normalizing by area in that the units are square units like area. The square of the expected distance between two random points would also be a natural measure to use here. We choose the expected squared distance for this paper simply because it is easier to compute and leads to interesting results.

In mathematical notation we consider the problem of studying the behavior of the ratio

$$\rho_P^* = \frac{\mathrm{E}(\mathrm{area}(\Delta))}{\mathrm{E}(L^2)},$$

where L is the distance between two independently chosen points with distribution P and Δ is a triangle with three independent vertices with distribution P. We call the computation and analysis of ρ_P^* the *Modified Sylvester Four Point Problem*.

Although this does not lead to a formula for the probability of a quadrilateral being convex, it is related to the sizes of random triangles relative to the distances between random points, an interesting geometric question for a probability distribution. We can think of ρ_P^* as the ratio of two different measures of the spread of a probability distribution, a two-dimensional measure and a one-dimensional measure. We see intuitively that a distribution concentrated in a long narrow region will lead to a smaller ratio than a distribution on a rounder region since the areas of random triangles in a long narrow region will be small even though the lengths of random line segments may be large. Thus we will refer to ρ_P^* as the *rotundity* of the distribution P.

We first consider the case where the points are uniformly distributed over bounded convex regions as in the original Sylvester problem. When the distribution P is uniform over a set K, we also denote the ratio ρ_P^* by ρ_K^*. We then consider two additional settings, one where the vertices are uniformly chosen on the circumference of a circle and one where the points are normally distributed in the plane. The normal distribution in the plane is a natural generalization of the uniform distribution on a bounded region.

We begin with an analysis of $\mathrm{E}(L^2)$ for points chosen uniformly over simple convex sets in the plane. These results could very well be exercises in some probability texts. In any case our computational methods are not standard exercises and prepare the reader for calculations in the section on bivariate normal distributions.

Expected squared length

We choose two independent random points $\mathbf{X_1} = (X_1, Y_1)^t$ and $\mathbf{X_2} = (X_2, Y_2)^t$ represented as column vectors and distributed as $\mathbf{X} = (X, Y)^t$. The length $L(\mathbf{X_1}, \mathbf{X_2})$ of the random line segment between the points has the property

$$\mathrm{E}(L^2) = \mathrm{E}\left((X_2 - X_1)^2 + (Y_2 - Y_1)^2\right) = 2\left(\mathrm{Var}(X) + \mathrm{Var}(Y)\right).$$

In the following, K always denotes a convex set described by its boundary curve.

EXAMPLE 1. K A PARALLELOGRAM. Consider an arbitrary parallelogram with side lengths α, β, and included angle θ. We find that $\mathrm{E}(L^2) = (\alpha^2 + \beta^2)/6$.

To derive this first consider points chosen uniformly in the square with vertices $(0, 0)$, $(1, 0)$, $(1, 1)$, $(0, 1)$. In this case X and Y are independent, random variables, each with variance $1/12$, so that $\mathrm{E}(L^2) = 2(1/12 + 1/12) = 1/3$.

We next consider a parallelogram determined by the vectors $\mathbf{V_1} = (a_1, a_2)^t$ and $\mathbf{V_2} = (b_1, b_2)^t$. Just as this parallelogram is a linear transformation of the square with transformation matrix

$$T = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix},$$

the random point $\mathbf{X}' = (X', Y')^t$ chosen in this parallelogram can be written as $T\mathbf{X}$, where $\mathbf{X}$ is the random point in the square. It follows that $X' = a_1X + b_1Y$ and $Y' = a_2X + b_2Y$. A simple computation then shows that

$$\mathrm{E}(L^2) = 2\left(\mathrm{Var}(X') + \mathrm{Var}(Y')\right) = \left(|\mathbf{V_1}|^2 + |\mathbf{V_2}|^2\right)/6 = \left(\alpha^2 + \beta^2\right)/6.$$

Subject to the constraint $\mathrm{area}(K) = \alpha\beta \sin\theta = A$, we see that $\mathrm{E}(L^2)$ takes its minimum value of $A/3$ when the parallelogram is a square. Even when A is fixed there is no upper bound on $\mathrm{E}(L^2)$.

EXAMPLE 2. K A TRIANGLE. Consider an arbitrary triangle with side lengths α, β, included angle θ, and opposite side length γ. We find $\mathrm{E}(L^2) = (\alpha^2 + \beta^2 - \alpha\beta\cos\theta)/9 = (\alpha^2 + \beta^2 + \gamma^2)/18$. To derive this result we first consider points chosen in the triangle with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$. We find that X and Y have variance $1/18$ and covariance $-1/36$. We next consider a triangle determined with vectors $\mathbf{V_1} = (a_1, a_2)^t$ and $\mathbf{V_2} = (b_1, b_2)^t$. As above we find that we can treat the random point $\mathbf{X}'$ chosen in this triangle as $X' = a_1X + b_1Y$ and $Y' = a_2X + b_2Y$. A simple computation then shows that

$$\begin{aligned}\mathrm{E}(L^2) &= 2\left(\mathrm{Var}(X') + \mathrm{Var}(Y')\right) = \frac{1}{18}\left(2|\mathbf{V_1}|^2 + 2|\mathbf{V_2}|^2 - 2(\mathbf{V_2} \cdot \mathbf{V_1})\right) \\ &= \left(\alpha^2 + \beta^2 - \alpha\beta\cos\theta\right)/9 = \left(\alpha^2 + \beta^2 + \gamma^2\right)/18.\end{aligned}$$

Subject to the constraint $\mathrm{area}(K) = \alpha\beta\sin\theta/2 = \mathrm{A}$, the value of $\mathrm{E}(L^2)$ is minimized when $\alpha = \beta$ and $\theta = \pi/3$. That is, for a fixed area, $\mathrm{E}(L^2)$ achieves its minimum value of $2\sqrt{3}\mathrm{A}/9$ over all triangles when the triangle is equilateral. By letting θ approach 0, we see that there is no upper bound on the value of $\mathrm{E}(L^2)$ with this constraint.

EXAMPLE 3. K AN ELLIPSE. Consider an arbitrary ellipse with semi-minor axis α and semi-major axis β. We have $\mathrm{E}(L^2) = (\alpha^2 + \beta^2)/2$. This follows by first considering the unit circle centered at the origin, where $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1/4$. For the

ellipse we have $X' = \alpha X$ and $Y' = \beta Y$. Thus

$$\mathrm{E}(L^2) = 2\left(\frac{\alpha^2}{4} + \frac{\beta^2}{4}\right) = \left(\alpha^2 + \beta^2\right)/2.$$

It follows that $\mathrm{E}(L^2)$ takes its minimum value of A/π over all ellipses with area A when the ellipse is a circle. Once again there is no upper bound.

For K an equilateral triangle, square and circle respectively each of unit area we have the following

$$\mathrm{E}\left(L^2_{\triangle}\right) = \frac{2\sqrt{3}}{9} > \mathrm{E}\left(L^2_{\square}\right) = \frac{1}{3} > \mathrm{E}\left(L^2_{\bigcirc}\right) = \frac{1}{\pi}$$

Observe in FIGURE 1 with all regions centered at the origin how the triangle has regions that protrude well beyond the square and the square has regions extending beyond the circle. This suggests why $\mathrm{E}(L^2_{\triangle}) > \mathrm{E}(L^2_{\bigcirc})$.

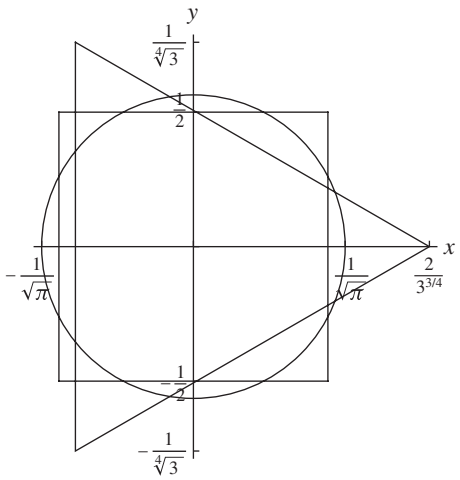


Figure 1 The circle, square, and triangle of area one.

Indeed, for any region of given area A, the minimum possible value of

$$\mathrm{E}\left(L^2(\mathbf{X_1}, \mathbf{X_2})\right)$$

occurs when K is the region inside of a circle. That is, for any region K of area A,

$$\mathrm{E}\left(L^2(\mathbf{X_1}, \mathbf{X_2})\right) \geq \frac{1}{\pi}\mathrm{A}. \tag{1}$$

The following argument shows why this is true. By translating the set K we may assume that $\mathrm{E}(X) = \mathrm{E}(Y) = 0$. We may also assume $\mathrm{A} = 1$. Then

$$2(\mathrm{Var}(X) + \mathrm{Var}(Y)) = 2\int\!\!\int_K (x^2 + y^2)\, dy\, dx.$$

The integrand $x^2 + y^2$ is just the squared distance of the point (x, y) from the origin. Now if there is a subset E of K that is further from the origin than a subset E' of equal

area in K complement, then switching E and E' will produce a set K' with equal area and

$$2\int\int_{K'}(x^2+y^2)\,dy\,dx < 2\int\int_{K}(x^2+y^2)\,dy\,dx.$$

In this fashion, the integral can be reduced unless K is the interior of a circle. An interesting consequence of (1) is that for any K of area A, we have

$$\pi\rho_K = \frac{\pi \mathrm{E}(\mathrm{area}(\Delta))}{A} \geq \frac{\mathrm{E}(\mathrm{area}(\Delta))}{\mathrm{E}(L^2)} = \rho_K^*.$$

It is also interesting to note that if K is centered so $\mathrm{E}(X) = \mathrm{E}(Y) = 0$, and $\mathrm{area}(K) = 1$ then

$$\frac{1}{2}\mathrm{E}(L^2) = \int\int_{K}(x^2+y^2)\,dy\,dx$$

can be interpreted as the polar moment of inertia of a lamina with uniform mass density equal to 1. From our previous examples we have that the minimum polar moment of inertia of a triangular region of area one and centroid the origin occurs when the triangle is equilateral. The minimum for the interior of a parallelogram occurs when the parallelogram is a square, and the minimum for the interior of an ellipse occurs when the ellipse is a circle, with the circle giving the smallest value of all.

Application to the modified Sylvester problem

Uniform distributions on convex sets We next consider the modified Sylvester four point problem when the distribution of points is uniform over a convex region. In this way we can make use of known results for Sylvester's original Four Point Problem. We have

$$\rho_K^* = \frac{\mathrm{E}(\mathrm{area}(\Delta))}{\mathrm{E}(L^2)} = \frac{\mathrm{E}(\mathrm{area}(\Delta))}{\mathrm{area}\,K} \cdot \frac{\mathrm{area}\,K}{\mathrm{E}(L^2)},$$

where Δ is a random triangle with three vertices chosen independently and L^2 is the squared length of a random line segment with end points chosen independently from the uniform distribution on K. Although ρ_K^* is not invariant under affine transformations, it is invariant under dilations, translations, and orthogonal transformations. In other words, if the points are selected in a triangle, ρ_K^* could depend on the shape of the triangle, but not the size or orientation.

Applying the results of the previous section we have the following:

EXAMPLE 1A. It is known that for any parallelogram K of area A, $\mathrm{E}(\mathrm{area}\,\Delta)/\mathrm{A} = 11/144 \approx .076$. It follows from Example 1 that

$$\rho_K^* \leq \frac{11}{144} \cdot 3 = \frac{11}{48} \approx .2292,$$

where equality is achieved when K is bounded by a square.

EXAMPLE 2A. It is known that for any triangle K of area A, $\mathrm{E}(\mathrm{area}\,\Delta)/\mathrm{A} = 1/12 \approx .083$. It follows from Example 2 that

$$\rho_K^* \leq \frac{1}{12} \cdot \frac{9}{2\sqrt{3}} = \frac{\sqrt{3}}{8} \approx .2165,$$

where equality is achieved when K is bounded by an equilateral triangle.

EXAMPLE 3A. It is known that for any ellipse K of area A, $\mathrm{E}(\text{area}\,\Delta)/A = 35/(48\pi^2) \approx .074$. It follows from Example 3 that

$$\rho_K^* \leq \frac{35}{48\pi^2} \cdot \pi = \frac{35}{48\pi} \approx .2321,$$

where equality is achieved when K is bounded by a circle.

It appears from these calculations that in the case of choosing the points from the interior of convex polygons the maximum ρ_K^* is when the polygons are regular. The minimum can be made arbitrarily close to zero as is evident in the case of very thin regions.

It is interesting to note that among the equilateral triangle, square, and circle, ρ_K^* is greatest for the circle and smallest for the equilateral triangle. These observations lead to the following conjectures.

CONJECTURE 1. For K a bounded convex set, the maximum ρ_K^* occurs when K is the interior of a circle.

CONJECTURE 2. For K bounded by a curve in the family of linear transformations of a regular polygon, ρ_K^* is greatest for the interior of the regular polygon itself.

Uniform distribution on the circumference of a circle

We next consider random points chosen on the circumference of a circle. This situation would have been of no interest to Sylvester since all quadrilaterals determined by such points would be convex. However, it is interesting to us because we can ask, which distribution has greater rotundity, the uniform distribution inside the circle, or the uniform distribution on the circumference of the circle? Denoting the uniform distribution on the circumference of the circle by C, this becomes, which is larger, ρ_C^* or $\rho_{\bigcirc}^*$ $(= \frac{35}{48\pi} \approx .2321)$? To answer this, we have to compute ρ_C^*.

We consider the unit circle centered at the origin in the xy plane. We may assume that one point is at $(1, 0)$. The other two points are determined by their polar angles Θ_1 and Θ_2, which are independent and uniformly distributed over $[0, 2\pi]$. We then have that the expected squared distance from a random point on the circumference to $(1, 0)$ is given by

$$\mathrm{E}(L^2) = \frac{1}{2\pi} \int_0^{2\pi} (\cos\theta - 1)^2 + \sin^2\theta \, d\theta = 2.$$

To compute $\mathrm{E}\,(\text{area}(\Delta))$ we use the fact that the area of the triangle Δ with vertices $(1, 0)$, $(\cos\Theta_1, \sin\Theta_1)$, and $(\cos\Theta_2, \sin\Theta_2)$ equals $(1/2)|\mathbf{V_1} \times \mathbf{V_2}|$, where in the standard notation, $\mathbf{V_1} = (\cos\Theta_1 - 1)\mathbf{i} + \sin\Theta_1\,\mathbf{j}$ and $\mathbf{V_2} = (\cos\Theta_2 - 1)\mathbf{i} + \sin\Theta_2\,\mathbf{j}$; see FIGURE 2.

We have $\mathbf{V_1} \times \mathbf{V_2} = Z\mathbf{k}$, where $Z = \sin(\Theta_2 - \Theta_1) + \sin\Theta_1 - \sin\Theta_2$. We thus have $\mathrm{E}\,(\text{area}(\Delta)) = (1/2)\mathrm{E}(|Z|)$. To deal with the problem of the absolute value we break the integral for $\mathrm{E}(|Z|)$ into the regions corresponding to $\Theta_1 < \Theta_2$ and $\Theta_1 > \Theta_2$.

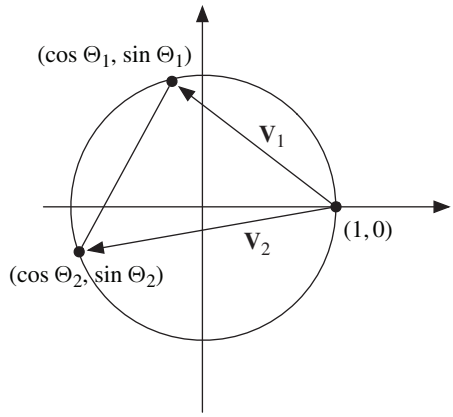


Figure 2 Random triangle with vertices on the unit circle.

In the first case,we see by the right hand rule for cross products that $Z > 0$ and in the second case $Z < 0$.

We find

$$\mathrm{E}(|Z|) = \frac{2}{4\pi^2}\int_0^{2\pi}\int_0^{\theta_2} \sin(\theta_2 - \theta_1) - \sin\theta_2 + \sin\theta_1 \, d\theta_1 \, d\theta_2 = \frac{3}{\pi}.$$

It follows that

$$\mathrm{E}\,(\mathrm{area}(\Delta)) = \frac{3}{2\pi}.$$

Combining this with the fact that $\mathrm{E}(L^2) = 2$, we have

$$\rho_C^* = \frac{3}{4\pi} \approx .2387.$$

Since this value is greater than .2321 we conclude

$$\rho_{\bigcirc}^* < \rho_C^*.$$

This conclusion leads to two more conjectures.

CONJECTURE 3.

$$\max_P \ \rho_P^* = \rho_C^*.$$

Denote the boundary of a set K by ∂K. Assign the uniform distribution with respect to arclength to ∂K. Then we have:

CONJECTURE 4. If K is a bounded convex set, then $\rho_K^* < \rho_{\partial K}^*$.

Behavior of ρ_P^* under linear transformations

In the previous cases of triangles, parallelograms, and ellipses, we have classes of sets which are invariant under linear transformations. The class of bivariate normal distributions is also invariant under linear transformations. Before considering the class

of bivariate normal distributions, we consider in this section the general properties of $\mathrm{E}(L^2)$ and $\mathrm{E}\,(\mathrm{area}(\Delta))$ when the random points or random vectors $\mathbf{X_1}$, $\mathbf{X_2}$, and $\mathbf{X_3}$ are transformed by a linear transformation of the plane.

Let $\mathbf{X} = (X, Y)^t$ be a random point selected from a probability distribution P. The covariance matrix of the coordinate variables X and Y is given by

$$R_{\mathbf{X}} = \begin{pmatrix} \sigma_X^2 & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \sigma_Y^2 \end{pmatrix},$$

This is a symmetric and positive semi-definite matrix. As such it has non-negative eigenvalues λ_1 and λ_2, and its determinant is their product, $\det(R_{\mathbf{X}}) = \lambda_1\lambda_2$. Further, there is an orthogonal matrix A (that is, a matrix with $AA^t = I$) such that

$$AR_{\mathbf{X}}A^t = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = D \tag{2}$$

This transformation preserves the trace, so that

$$\lambda_1 + \lambda_2 = \mathrm{trace}(AR_{\mathbf{X}}A^t) = \mathrm{trace}(R_{\mathbf{X}}) = \sigma_1^2 + \sigma_2^2.$$

Now let the triangle Δ be formed by the three random points $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$ selected independently from the distribution P, and let T be a linear transformation of the plane. Denote the triangle formed by $T\mathbf{X_1}$, $T\mathbf{X_2}$, $T\mathbf{X_3}$ by $T\Delta$. It follows immediately from the linearity of expectation that

$$\mathrm{E}\,(\mathrm{area}(T\Delta)) = |\det(T)|\,\mathrm{E}\,(\mathrm{area}(\Delta))\,.$$

Note that if either eigenvalue λ_1 or λ_2 equals 0, then there is a linear combination of X and Y with variance 0. The linear combination is constant. This means that the three vertices of the random triangle Δ all lie on a straight line and thus $\mathrm{E}\,(\mathrm{area}(\Delta)) = 0$. From here on we avoid this trivial situation by assuming both eigenvalues are positive.

The computation of $\mathrm{E}(L^2)$ requires more attention. We have already seen that

$$\mathrm{E}\left(L^2(\mathbf{X_1}, \mathbf{X_2})\right) = 2\,(\mathrm{Var}(X) + \mathrm{Var}(Y)) = 2(\sigma_1^2 + \sigma_2^2),$$

so we also have

$$\mathrm{E}\left(L^2(\mathbf{X_1}, \mathbf{X_2})\right) = 2(\lambda_1 + \lambda_2). \tag{3}$$

The covariance matrix of the coordinates of $T\mathbf{X_1}$ (or $T\mathbf{X_2}$) is given by $TR_{\mathbf{X}}T^t$. If we write

$$ATA^t = B = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then we have $T = A^tBA$ and we can compute:

$$\begin{aligned} \mathrm{E}\left(L^2(T\mathbf{X}_1, T\mathbf{X}_2)\right) &= 2\,\mathrm{trace}(TR_{\mathbf{X}}T^t) \\ &= 2\,\mathrm{trace}(A^tBAA^tDAA^tB^tA) \\ &= 2\,\mathrm{trace}(BDB^t). \end{aligned}$$

A simple computation now shows that

$$\mathrm{E}\left(L^2(T\mathbf{X}_1, T\mathbf{X}_2)\right) = 2[(a^2 + c^2)\lambda_1 + (b^2 + d^2)\lambda_2].$$

Denoting the probability distribution of $T\mathbf{X}$ by TP, this implies

$$\rho^*_{TP} = \frac{|\det(T)|(\lambda_1+\lambda_2)}{(a^2+c^2)\lambda_1+(b^2+d^2)\lambda_2}\rho^*_P = \frac{|ad-bc|(\lambda_1+\lambda_2)}{(a^2+c^2)\lambda_1+(b^2+d^2)\lambda_2}\rho^*_P. \tag{4}$$

In particular, if B is diagonal with determinant 1, then $b = c = 0$, $d = 1/a$, and (4) gives

$$\rho^*_{TP} = \frac{\lambda_1+\lambda_2}{a^2\lambda_1+a^{-2}\lambda_2}\rho^*_P. \tag{5}$$

Expression (5) is maximized when $a^2 = \sqrt{\lambda_2/\lambda_1}$, in which case

$$\rho^*_{TP} = \frac{\lambda_1+\lambda_2}{2\sqrt{\lambda_1\lambda_2}}\rho^*_P.$$

Indeed, a Lagrange multiplier argument shows that in general,

$$\max_T \rho^*_{TP} = \frac{\lambda_1+\lambda_2}{2\sqrt{\lambda_1\lambda_2}}\rho^*_P = \frac{\sigma_1^2+\sigma_2^2}{\sqrt{4\ \det(R_{\mathbf{X}})}}\rho^*_P \tag{6}$$

In particular, we note that if $R_{\mathbf{X}} = cI$ for some constant c, then $\max_T \rho^*_{TP} = \rho^*_P$. In other words, if the x and y coordinates of the random point are uncorrelated random variables with the same variance, then the rotundity of the distribution is the maximum over all linear transformations of that distribution. We state it as a theorem:

THEOREM 1. *The rotundity of TP is maximized when T is chosen so that the x and y coordinates of a random point from TP are uncorrelated and have the same variance.*

Note that such a transformation T exists if and only if $\lambda_1 > 0$ and $\lambda_2 > 0$.

EXAMPLE 2B. To see how formula (6) might be applied, consider the case of triangles as in Examples 2 and 2A. If P is the uniform distribution on the region A inside the triangle with vertices (0, 0), (1, 0), and (0, 1), then

$$R_{\mathbf{X}} = \begin{pmatrix} 1/18 & -1/36 \\ -1/36 & 1/18 \end{pmatrix}.$$

In this case

$$\rho^*_P = \frac{1}{12}\,\text{area}(A)/\left[2(\sigma_1^2+\sigma_2^2)\right] = 3/16.$$

We see that $(\sigma_1^2+\sigma_2^2)/\sqrt{4\ \det R_{\mathbf{X}}} = 2/\sqrt{3}$, so that from (6), we know that the maximum value of ρ^*_{TP} is given by

$$\max_T \rho^*_{TP} = \frac{2}{\sqrt{3}}\frac{3}{16} = \frac{\sqrt{3}}{8}.$$

This agrees with the conclusion of Example 2A.

One linear transformation that achieves this maximum has matrix

$$T = \begin{pmatrix} 2 & 1 \\ 0 & \sqrt{3} \end{pmatrix}.$$

This moves the vertices to $(0, 0)$, $(2, 0)$, and $(1, \sqrt{3})$. (Any transformation mapping the triangle to an equilateral triangle would do as well). Now the covariance matrix associated with TP is given by

$$TR_{\mathbf{X}}T^t = \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{6} \end{pmatrix},$$

so that the coordinates are uncorrelated and have equal variance as required.

Normal distribution

We next consider points chosen from a mean $\mathbf{0}$ bivariate normal distribution with covariance $R_{\mathbf{X}}$ over the whole plane. We may assume that the mean is zero since the answer is invariant under translation of the distribution. As is seen in (3), $\mathrm{E}(L^2) = 2(\sigma_X^2 + \sigma_Y^2) = 2(\lambda_1 + \lambda_2)$, so our main problem is to compute $\mathrm{E}(\text{area}(\Delta))$. Computation of this expectation relies on a result for quadratic forms in mean 0, jointly normal random variables that is well known to statisticians. It states that if the covariance matrix of $X_1, \ldots, X_n$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, then

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} \lambda_i W_i^2, \tag{7}$$

where $W_1, \ldots, W_n$ are independent, standard normal random variables. For completeness, we derive this result for bivariate normal random vectors in the appendix.

Another fact that we will need is that for W_1 and W_2 independent $N(0, 1)$ or standard normal random variables, $P(W_1^2 + W_2^2 > t) = e^{-t/2}$. That is, $W_1^2 + W_2^2$ has the same distribution as an exponential random variable with mean 2. This result and other basic facts about the multivariate normal distribution are developed in Ross [**3**, pp. 66–76]. We are now ready to prove the main result of this section.

THEOREM 2. *If the triangle Δ has vertices (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) with all components independent $N(0, 1)$ random variables, then* area(Δ) *is an exponential random variable with*

$$\mathrm{E}\,(\text{area}(\Delta)) = \frac{\sqrt{3}}{2}.$$

Proof. Using determinants to compute areas it is easily seen that area(Δ) is given by

$$\frac{1}{2}\left|(X_2 - X_1)(Y_3 - Y_1) - (Y_2 - Y_1)(X_3 - X_1)\right|. \tag{8}$$

The quadratic form inside the absolute value sign is seen to have three double eigenvalues 0, $\sqrt{3}/2$, and $-\sqrt{3}/2$. It follows that the expression in (8) has distribution equal to

$$\frac{\sqrt{3}}{2}\left|\frac{(W_1^2 + W_2^2)}{2} - \frac{(W_3^2 + W_4^2)}{2}\right|$$

where W_1, W_2, W_3, and W_4 are independent $N(0, 1)$ random variables. Now $(W_1^2 + W_2^2)/2$ and $(W_3^2 + W_4^2)/2$ are independent exponential random variables with mean 1.

Thus the area of Δ has the same distribution as $\sqrt{3}\,|T_2 - T_1|/2$, where T_1 and T_2 are independent exponential random variables with mean 1. A simple calculation shows

$$P\left(\frac{\sqrt{3}}{2}|T_2 - T_1| > t\right) = 2P\left(T_2 > T_1 + \frac{2t}{\sqrt{3}}\right) = e^{-2t/\sqrt{3}},$$

so area(Δ) is an exponential random variable with mean $\sqrt{3}/2$. ∎

COROLLARY. If the points in the plane are chosen with coordinates that are independent standard normal random variables, then $\rho_P^* = \frac{\sqrt{3}}{8}$.

Surprisingly, this is the same value for rotundity as for points chosen with a uniform distribution inside an equilateral triangle. It follows from the comments after (6) that if points are chosen with distribution $\mathbf{X}' = T\mathbf{X}$, then

$$\max_T \rho_{TP}^* = \rho_P^* = \frac{\sqrt{3}}{8}.$$

We note here that all bivariate normal random variables in the plane are of the form $T\mathbf{X} + \mu$ for some T and μ, so this shows that among bivariate normal random variables, the one with covariance matrix I has maximum rotundity.

Appendix

In this appendix we prove (7) for the case of bivariate normal random variables. Let $\mathbf{X} = T\mathbf{Z}$, where $\mathbf{Z}$ is a mean $\mathbf{0}$ bivariate normal random vector with covariance matrix $R_{\mathbf{Z}} = I$. Then $R_{\mathbf{X}} = TR_{\mathbf{Z}}T^t = TT^t$. Now write $R_{\mathbf{X}} = A^t DA$ as in (2). Then $\mathbf{X}$ has the same distribution as a random vector of the form $\mathbf{Y} = (Y_1, Y_2)^t = A^t\sqrt{D}A\mathbf{Z}$ since a bivariate normal distribution is determined by its mean and covariance function. (We note that $R_{\mathbf{Y}} = A^t\sqrt{D}AA^t\sqrt{D}A = R_{\mathbf{X}}$.)

Now

$$\sum_{i=1}^{2} Y_i^2 = \mathbf{Y^t Y} = \mathbf{Z}^t A^t\sqrt{D}AA^t\sqrt{D}A\mathbf{Z} = \mathbf{Z}^t A^t DA\mathbf{Z}. \tag{9}$$

Observe that $\mathbf{W} = (W_1, W_2)^t = A\mathbf{Z}$ has covariance matrix $R_{\mathbf{W}} = AA^t = I$, so W_1 and W_2 are independent, standard normal random variables. Substituting $\mathbf{W}$ for $A\mathbf{Z}$ in (9), we see

$$\sum_{i=1}^{2} Y_i^2 = \sum_{i=1}^{2} \lambda_i W_i^2.$$

Since $\sum_{i=1}^{2} Y_i^2$ has the same distribution as $\sum_{i=1}^{2} X_i^2$, the result is proved.

REFERENCES

1. John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Regression Models*, 2nd ed., Irwin, Boston, 1989.
2. Richard E. Pfiefer, The Historical Development of J. J. Sylvester's Four Point Problem, this MAGAZINE **62** (1989), 309–317. doi:10.2307/2689482
3. Sheldon Ross, *Introduction to Probability Models*, 9th ed., Academic Press, New York, 2007.

4. Eric W. Weisstein, "Sylvester's Four-Point Problem," from *MathWorld*—A Wolfram Web Resource, http://mathworld.wolfram.com/SylvestersFour-PointProblem.html

Summary In 1865 Sylvester posed his famous four point problem "What is the probability that a random quadrilateral is convex?" This somewhat ill-defined question led to the problem of finding the minimum and maximum of the expected area of a triangle whose vertices are chosen with a uniform distribution over a convex region of area one. We modify this problem to that of finding the normalized expected area of a triangle whose vertices are chosen at random with an arbitrary probability distribution in the plane. The normalizing constant is the expected squared length of a line segment between two random points with the given distribution. We solve this modified problem in many important cases and conjecture that the maximum value of the normalized expected area occurs when the probability distribution is the uniform distribution on the circumference of a circle.

BENNETT EISENBERG attended Dartmouth College and received his Ph.D. from MIT. He taught at Cornell University and the University of New Mexico before joining the Lehigh University Mathematics Department in 1972. He still enjoys teaching courses in probability theory and likes to draw on classical historical problems for his lectures.

ROSEMARY SULLIVAN received her B.S. from Penn State and her Ph.D. from Lehigh University. Before coming to West Chester University she taught at Widener University and Muhlenberg College. In addition to her interest in geometric probability, she enjoys grilling, swimming and hiking.

Representational Efficiency

C. RAY ROSENTRATER
Westmont College
Santa Barbara, CA 93108
rosentr@westmont.edu

Many people assume that the distribution of initial digits of numbers obtained by measuring random physical objects is essentially uniform. Surprisingly, this is false—an observation typically associated with F. Benford. In fact, numbers arising in applications are not uniformly distributed, and they fail to be uniformly distributed in a predictable way. In this article we ask what this fact implies for representing real numbers in computers.

The IEEE Standard

The most common computer representation for real numbers is defined by IEEE Standard 754. This standard identifies specific bits to represent the sign (S), the exponent (E) and the mantissa (F, for *fraction*). There is always one sign bit, and for a single precision number there are 8 exponent bits and 23 fraction bits, as shown in FIGURE 1.

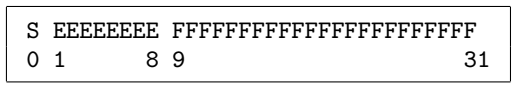


Figure 1 Bit assignment in IEEE Single Precision Format

If we interpret E and F as integers (represented in binary notation) then $0 \le E \le 255$ and $0 \le F \le 2^{23} - 1$. In most cases, the combination of S, E, and F represents the real number

$$x = \pm(1 + 2^{-23}F) \times 2^{E-127},$$

with the sign being negative if $S = 1$. (The cases of $E = 0$ and $E = 255$ are reserved for special purposes. In particular, the case of $E = F = 0$ represents the real number zero, regardless of the sign bit.)

Some reflection shows that (when x is positive) the fraction F consists of the leading binary digits of x (the mantissa)—except that the first binary digit of the mantissa is always 1, and that first binary digit is stripped from F. That is, F consists of the 2nd through 24th most significant digits of the mantissa. If the mantissa M has the digits $1x_1x_2 \ldots x_{23}$, then F has the binary digits $x_1x_2 \ldots x_{23}$.

In this article we are concerned with the efficiency of the representation of F.

Whether the representation is efficient depends on the distribution of the numbers to be represented. To illustrate, we consider a hypothetical example.

EXAMPLE 1. Suppose that we receive a sequence of 64-bit numbers. The most obvious way to store the data is to record the actual bits. Now suppose that, due to the source of the data, we know that an average of one in 32 bits will be nonzero.

Math. Mag. **84** (2011) 185–195. doi:10.4169/math.mag.84.3.185.

With this additional information, we might decide to store each number by recording the number, n, of ones it contains (6 bits will suffice) followed by n 6-bit numbers identifying their positions. This alternate approach requires an average of 18 rather than 64 bits per number ($\approx 72\%$ savings) with no loss of information.

While this example is intentionally extreme, it raises two questions that lie at the heart of this exposition.

QUESTION 1. In a given natural or computational setting where real numbers arise, what is the distribution of F?

QUESTION 2. Is some alternate representation system more efficient?

Note that, while related, the distribution of mantissas is not the same as the distribution of the leading digits. Also, Question 2 is not related to the relevance or usefulness of the digits. Indeed, it takes just as many digits to identify an inaccurate or irrelevant number as one that is accurate and important.

The history

Motivated by a colleague's observation that the earlier portions of logarithmic tables seemed to be more worn than the later sections, Benford [**2**] took counts from various numerical tables such as street addresses from *American Men of Science,* the freezing points in degrees Celsius of chemical compounds, and the areas in square miles of U.S. rivers. (Benford did not say what he meant by "areas" of rivers.) All together, Benford collected over 19,000 observations from 19 different types of sources. In each case he found the fraction of all entries with first (decimal) digit at most p to be approximately $\log_{10}(p+1)$, for $p = 1, 2, \ldots, 9$. In other words, the leading digits follow a *logarithmic* distribution.

Over fifty years earlier, and apparently forgotten by the time Benford published his paper, S. Newcomb [**10**] had published an intuitive explanation. He argued that real numbers found in computational applications are formed by ratios and that successive products of such ratios (after being scaled by integer powers of 10 to fall between 1 and 10) should have logarithms that approximately follow a uniform distribution. A straightforward transformation shows this to be equivalent to Benford's conjecture. While relevant in a computational context, Newcomb's intuitive argument fails to account for Benford's observations in other settings.

In 1966, B. J. Flehinger [**5**] published an explanation from a more general point of view. Taking integers as a proxy for approximate values of physical constants, Flehinger considered the fraction $a_n(p)$ of integers in the set $\{1, 2, \ldots, n\}$ with initial (decimal) digit at most p. For any fixed value of p, the sequence $\{a_n(p)\}_{n=1}^{\infty}$ of these densities satisfies $\liminf_{n\to\infty} a_n(p) = \frac{p}{9}$ and $\limsup_{n\to\infty} a_n(p) = \frac{10p}{9(p+1)}$. By averaging these densities (that is, considering $b_n(p) = \frac{1}{n}\sum_{k=1}^{n} a_k(p)$), she obtained a sequence that oscillates over a smaller range of values. When the averages are averaged repeatedly, a limiting sequence is produced, one that converges to Benford's observed value of $\log_{10}(p+1)$. (These averages are called *Cesàro means*. They are a useful tool for finding "limits" of sequences that oscillate without converging.)

Papers by Pinkhan [**11**], Raimi [**12**], and Hill [**7**], represent another consistent thread in this discussion. Each begins, more or less, by assuming that the distribution of the constants in the universe should not depend upon the choice of units used for measurement. This supposition is expressed mathematically by saying that the distribution should be scale invariant: $\Pr[cS] = \Pr[S]$ for any positive number c and any set S to which one can apply the probability function (that is, any *measurable* set). The

fact that the leading digit follows the logarithmic distribution is a consequence of this assumption.

A significant focus of these papers is the identification of measurable sets. For example, the interval $(0, 1]$ certainly seems like a reasonable set, but a probability distribution that assigns a non-zero probability to $(0, 1]$ cannot exhibit this sort of scale invariance. Scale invariance would imply the impossible conclusion that $\Pr[X \in (0, 1]] = \Pr[X \in (0, c]]$ for all $c \in \mathbb{R}^+$. Therefore neither $(0, 1]$, nor any other bounded set, can usefully be called measurable for this purpose. On the other hand, sets like $\cup_{n=-\infty}^{\infty}[1 \cdot 10^n, p \cdot 10^n)$, the set of real numbers with initial digit less than p, are always measurable.

The generalization

If instead of 10 we choose our base to be $B = 2^k$ (or any other base for that matter) it is a straightforward task to rework any of the proofs mentioned above so that they apply to representations in base B. If D_1 represents the value of the first base-B digit of a random number, then

$$\Pr[D_1 \leq p] = \log_B(p + 1), \tag{1}$$

so that

$$\begin{aligned}\Pr[D_1 = p] &= \log_B(p + 1) - \log_B(p) \\ &= \log_B\left(1 + \frac{1}{p}\right).\end{aligned}$$

For example, when $B = 2$ we have $\Pr[D_1 = 1] = \log_2(1 + 1) = 1$, reflecting the fact that the leading bit of a nonzero binary number is always 1.

Readers interested in a more detailed development of these results should consult the expositions of the distribution of leading digits by Knuth [**9**, section 4.2.4], Hill [**7**], and Raimi [**13**].

The application

R. W. Hamming [**6**] provided additional evidence of the importance of the logarithmic distribution:

1. The product or quotient of two random variables follows the logarithmic distribution if either of the random variables does (regardless of the distribution of the other random variable).
2. The reciprocal of a random variable following the logarithmic distribution also has the same distribution.
3. As long as neither of the original distributions place all the probability on a single value, the distribution of the product of two random variables is more nearly logarithmic than either of the two original distributions.

The distribution of the mantissa M is closely related to the distribution of the first digit in base 2^{24}. In fact, one might guess that M is distributed like D_1 in equation (1) with $B = 2^{24}$. Using the methods in the sections that follow, one can show that, in this case, the *information content* (to be defined later) of this distribution would be about 16.1 bits. This means that—if this guess were true—the usual 23-bit representation of

the mantissa would be using about 50% more space than would be used by a maximally efficient representation of the same information.

Fortunately from the perspectives of both mathematical interest and representational efficiency, M does not have the same distribution as D_1. To see why, consider a simpler setting in which M is four bits long. There are fifteen nonzero "hexits" (base-16 digits), and the logarithmic distribution (first-digit distribution) for these hexits is shown in FIGURE 2.

But M is always shifted (*normalized*)—by single bits, not by hexits—to guarantee that the leading bit is 1. That means that only the last eight hexits in FIGURE 2 are used. Looking at the graph, we see that eliminating the first seven hexits also eliminates much of the bias of the distribution. This, as we shall see, increases the information content, which, in turn, implies that the bits are used more efficiently when representing M than when representing D_1.

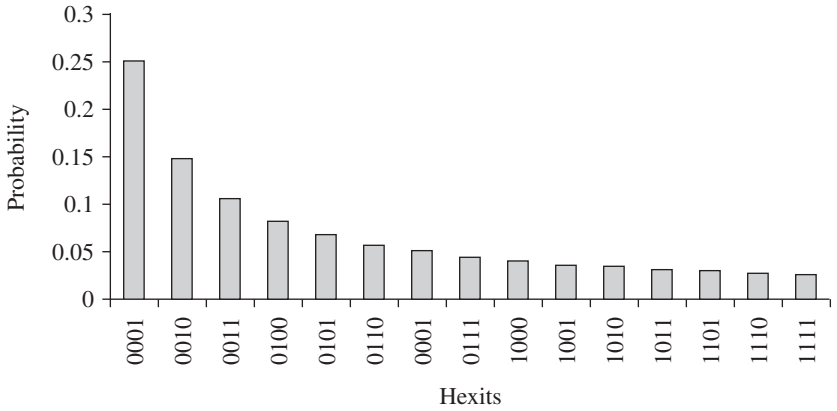


Figure 2 The distribution of base-16 (hexidecimal) hexits

The distribution of the mantissa

To investigate the distribution of the mantissa, consider sets of numbers beginning with specified base-10 digits. For example, the set of real numbers whose first decimal digit is at most 5 is $\cup_{n=-\infty}^{\infty}[1 \cdot 10^n, 6 \cdot 10^n)$ and has probability $\log_{10}(6) = 0.778151$. The set of real numbers whose first decimal digit is equal to 5 is $\cup_{n=-\infty}^{\infty}[5 \cdot 10^n, 6 \cdot 10^n)$ and has probability $\log_{10}(6) - \log_{10}(5) = \log_{10}(6/5) = 0.079181$. We can work more finely and identify the set of numbers whose first two decimal digits are 53 as $\cup_{n=-\infty}^{\infty}[5.3 \cdot 10^n, 5.4 \cdot 10^n) = \cup_{n=-\infty}^{\infty}[53 \cdot 10^n, 54 \cdot 10^n)$. The logarithmic distribution assigns to this set a probability of $\log_{10}(54) - \log_{10}(53) = \log_{10}(54/53) = 0.008118$.

Now consider a more general context. Fix a base B and a leading digit $a \in \{1, 2, \ldots, B-1\}$. Then choose k (base-B) digits, $x_1, x_2, \ldots, x_k \in \{0, 1, 2, \ldots, B-1\}$ to follow a and let $x = \sum_{i=1}^{k} x_i B^{k-i}$. The set of all real numbers whose base-B expressions begin with $ax_1x_2 \ldots x_k$ is

$$S(B, a, k, x) = \cup_{n=-\infty}^{\infty}[(aB^k + x)B^n, (aB^k + x + 1)B^n).$$

Modest tweaking of the derivations cited earlier shows that the probability of $S(B, a, k, x)$ is

$$\begin{aligned} \Pr[S(B, a, k, x)] &= \log_B(aB^k + x + 1) - \log_B(aB^k + x) \\ &= \log_B\left(1 + \frac{1}{aB^k + x}\right). \end{aligned}$$

In the case of our primary interest, namely $B = 2$, the first digit always is $a = 1$. If we further specialize to the case of $k = 23$, the probability of interest is

$$\Pr[S(2, 1, 23, x_1x_2 \ldots x_{23})] =$$
$$\Pr[F = x_1x_2 \ldots x_{23}] = \log_2\left(1 + \frac{1}{2^{23} + x}\right).$$

This is the distribution we expect for values of F in the IEEE representation. It answers our first question and provides an essential ingredient needed to answer the second question.

Entropy and information content

Our ultimate goal is to investigate the representational efficiency of the standard encoding of real numbers in a computer. To clarify the issue, suppose that we decided to duplicate each bit used to represent F. Doing so would require twice as many bits while providing no new information. Our second question amounts to: Does the usual computer representation of real numbers contain a similar but more subtle redundancy?

To address this question we need a way to measure the amount of information required to identify a specific selection from a distribution. The *information content* or *entropy* of a random variable X with a discrete probability distribution is measured by

$$H = -\sum_x p_x \log_2(p_x) = -E\left[\log_2(p_X)\right].$$

Here the sum is taken over the support of the distribution, p_x is the probability of x, and E is the expected value operator. The entropy measures the expected (average) number of bits (or yes/no questions) required to identify which of the possible values has occurred.

To gain a feel for how this works, consider the following examples.

EXAMPLE 2. When $B = 2$, the entropy of the first bit of M is

$$H = -p_1 \log_2(p_1) = -1 \log_2(1) = 0.$$

The leading bit of M carries no information; no bits are required to identify its value. Why? Because the leading bit is always 1.

EXAMPLE 3. Suppose we are flipping a coin. A natural way to record the outcome would be to write 1 for a head and 0 for a tail for an average code length of 1. If the coin is fair, the average code length matches the entropy, which is

$$H_{\text{fair}} = -\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) = 1.$$

Thus, each coin flip requires one full bit to describe, and there is no redundancy in the natural representation.

If the coin is weighted so that a head occurs $\frac{7}{8}$ of the time, the entropy becomes

$$H_{\text{weighted}} = -\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) \approx 0.543\,564.$$

This indicates that an alternate representational scheme (necessarily involving sequences or collections of flips) could reduce the space required to store a sequence of results by nearly half.

EXAMPLE 4. Suppose data is drawn from a distribution with probability mass function

$$p_x = \begin{cases} 1/2 & x = 0 \\ 1/8 & x = \pm 1 \\ 1/16 & x = \pm 2, \pm 3 \\ 0 & \text{otherwise.} \end{cases}$$

Since the possible values range from -3 to $+3$, it is natural to use a 3-bit code in which the first bit indicates the sign and the other two bits determine the magnitude. This results in code 1 of TABLE 1. The entropy of the corresponding distribution is

$$H = 1 \cdot \frac{1}{2} \cdot 1 + 2 \cdot \frac{1}{8} \cdot 3 + 4 \cdot \frac{1}{16} \cdot 4 = \frac{9}{4} = 2.25$$

which suggests a potential 25% savings in average code length. This savings can be achieved for this distribution by using shorter codes for more frequently occurring values and longer codes for values that are less likely. Code 2 in TABLE 1 is one such code. At the cost of using variable-length codes, the average code length exactly matches the entropy. An important feature of code 2 is that, in spite of its varying lengths, it is unambiguous. A sequence beginning with 0 has length 1, all sequences starting with 10 have length 3, and all sequences beginning with 11 have length 4. When receiving or processing a stream of values, there is no uncertainty about how to convert a stream of bits into the corresponding stream of values.

TABLE 1: Average lengths of two codes

value	prob.	code 1	code 2
-3	$1/16$	011	1101
-2	$1/16$	010	1100
-1	$1/8$	001	100
0	$1/2$	000	0
1	$1/8$	101	101
2	$1/16$	110	1110
3	$1/16$	111	1111
Ave. Len.		3	2.25

The difference between the average length of a particular encoding and the entropy is called the *redundancy*. Code 1 of *Example 4* has a redundancy of 0.75 bits or 33%. Code 2 has zero redundancy. One goal of code construction is to find an encoding with minimal redundancy. As is seen in the examples, zero-redundancy encodings of fixed length may not exist. As one might suspect from these examples, a zero-redundancy encoding of the individual outcomes exists when all the probabilities are integer powers of $1/2$. In this case, the length of the code for a value x is $-\log_2(p_x)$. Even when the probabilities are not powers of $1/2$, one still may gain efficiency by encoding collections of values rather than single values. The reader interested in a more complete treatment of information theory is encouraged to read Ash [**1**] or Young [**18**].

The information content of the mantissa

Recall that we are considering a mantissa M represented by a string of $k+1$ bits of the form $M = 1x_1x_2\ldots x_k$ where x_i is 0 or 1. From our previous work, we know that the probability that a real number begins with this sequence of bits is $p_x = \log_2\left(1 + \frac{1}{2^k+x}\right)$ where $x = \sum_{i=1}^{k} x_i 2^{k-i}$. The entropy of the mantissa of the standard real number representation then is

$$H = -\sum_{x=0}^{2^k-1} \log_2\left(1 + \frac{1}{2^k+x}\right) \log_2\left(\log_2\left(1 + \frac{1}{2^k+x}\right)\right).$$

For small values of k, it is easy to write a computer program or to use a spread sheet to evaluate this sum. But as k grows beyond 10, the computation becomes increasingly intractable from both time and error accumulation perspectives. Fortunately, we can use techniques from elementary calculus to reduce the computational load.

For notational convenience, set $D = 2^k$. For $a > 0$, $\ln(1+a) = \int_0^a \frac{1}{1+x}dx > \frac{a}{1+a}$, so

$$p_x = \log_2\left(1 + \frac{1}{D+x}\right) > \frac{\frac{1}{D+x}}{\left(1+\frac{1}{D+x}\right)\ln(2)} = \frac{1}{(D+x+1)\ln(2)}.$$

Since $-x\ln(x)$ is increasing in the interval $\left(0, \frac{1}{(D+1)\ln 2}\right] \subset (0, 1/e]$, we can use the well-known inequality

$$\sum_{x=a}^{b-1} h(x) > \int_a^b h(x)\,dx$$

for positive, decreasing functions h, to see that

$$\begin{aligned} H &= -\sum_{x=0}^{D-1} p_x \log_2(p_x) \\ &> \sum_{x=0}^{D-1} \frac{\log_2((D+x+1)\ln(2))}{(D+x+1)\ln(2)} \\ &> \int_{x=0}^{D} \frac{\log_2((D+x+1)\ln(2))}{(D+x+1)\ln(2)}dx \\ &= \frac{1}{2}\left[\log_2^2((2D+1)\ln(2)) - \log_2^2((D+1)\ln(2))\right]. \end{aligned}$$

Now factor the difference of squares and rewrite the result to obtain

$$\begin{aligned} H &> \frac{1}{2}\log_2\left(\frac{2D+1}{D+1}\right)\log_2\left(\ln^2(2)(2D^2+3D+1)\right) \\ &= \log_2\left(2 - \frac{1}{D+1}\right)\left[\log_2(\ln(2)) + \log_2(D) + \frac{1}{2}\log_2\left(2 + \frac{3D+1}{D^2}\right)\right]. \end{aligned}$$

Since $\log_2(x)$ is concave down and passes through (1,0) and (2,1), we know that $\log_2(x) \geq x - 1$ on the interval $[1, 2]$. In particular, $\log_2\left(2 - \frac{1}{D+1}\right) > 1 - \frac{1}{D+1}$. Hence

$$H > \left(1 - \frac{1}{D+1}\right)\left[k + \frac{1}{2} + \log_2(\ln(2))\right].$$

Asymptotically then, the entropy is at least $k + \frac{1}{2} + \log_2(\ln(2))$. So the asymptotic redundancy of M is at most $k + 1 - \left(k + \frac{1}{2} + \log_2(\ln(2))\right) = 1.028766$. Since increasing the number of bits in M can only increase the redundancy, the redundancy is never more than 1.028766 for *any* choice of k and, as seen in TABLE 2, the redundancy rapidly approaches this limit for relatively small values of k.

TABLE 2: The entropy and redundancy of the mantissa

# bits following 1	entropy	redundancy of M (including first bit)
1	0.979 070	1.020 930
2	1.973 246	1.026 754
3	2.971 741	1.028 259
4	3.971 361	1.028 639
5	4.971 265	1.028 735
6	5.971 242	1.028 758
$k > 6$	$k - 1.028\ 766$	1.028 766

These results indicate that, at least from an information theoretic point of view, the standard real number representation is quite close to the best possible encoding. In fact, since the F in the IEEE standard is formed by dropping M's leading 1, no alternate representation of real number mantissas with the same precision can reduce the average space requirements by even 0.03 bits per number. This is a substantial improvement over the 33% redundancy computed from a naive conflation of D_1 and F for $k = 23$.

Alternate bases

IBM has manufactured several machines in which the real number representation is based on hexits (base-16 digits) rather than on bits. In such machines, F is normalized (scaled) so that the leading hexit rather than the leading bit is nonzero. R. W. Hamming [**6**] speculated that this design was motivated by a desire to reduce the frequency of normalization shifts following multiplication. To illustrate, consider the following binary and decimal multiplications.

$$.11_{\text{two}} \times .11_{\text{two}} = .1001_{\text{two}} \quad \text{and} \quad .11_{\text{two}} \times .10_{\text{two}} = .011_{\text{two}}$$

$$.4 \times .4 = .16 \quad \text{and} \quad .3 \times .3 = .09$$

The result of the second product in each base requires a normalization shift to restore the decimal point to its original position immediately preceding the leading nonzero digit. One quickly sees that the percentage of binary pairs requiring renormalization after multiplication is much higher than the corresponding percentage of decimal pairs. Indeed, Hamming's computations [**6**] show that (assuming a sufficient number of digits) around 39% of binary mantissa pairs have products requiring normalization shifts. This percentage drops to approximately 17% and 13% for decimal and hexadecimal mantissas. So the percentage of mantissa pairs requiring a hexadecimal normalization shift is only a third of the percentage of pairs requiring a binary normalization shift.

But wait! The mantissas are not all equally likely. What happens when the probability of a normalization shift is computed relative to the probability of the mantissa pairs? Hamming [**6**] showed that, as the number of digits grows, the probability of a normalization shift is $\frac{1}{2}$ regardless of the choice of base. So if base-16 normalization was used in order to reduce the frequency of normalization shifts, the effort was wasted. Meanwhile, what is the effect on representational efficiency? More generally, what is the effect of choosing to normalize relative to a base of the form $B = 2^n$ that interacts nicely with the binary representation?

Let $H_n(k)$ represent the entropy of a random base-2^n number of the form $ax_1x_2 \ldots x_k$. This number also corresponds to a binary-normalized number $1x'_1x'_2 \ldots x'_m$ where $nk \le m \le nk + n - 1$, with m determined by the particular value of a. For example, the pattern

$$\underbrace{01\#\#}_{\text{1st hexit}} \mid \underbrace{\#\#\#\# \mid \#\ldots\# \mid}_{k \text{ additional hexits}}$$

represents a hexadecimal number for which $m = 4k + 2$. We note that for $0 \le j \le n - 1$,

$$\begin{aligned} H_2(kn + j) &= -\sum_{x=0}^{2^{kn+j}-1} \log_2\left(1 + \frac{1}{1 \cdot 2^{kn+j} + x}\right) \log_2\left(\log_2\left(1 + \frac{1}{1 \cdot 2^{kn+j} + x}\right)\right) \\ &= -\sum_{a=2^j}^{2^{j+1}-1} \sum_{x=0}^{2^{nk}-1} \log_2\left(1 + \frac{1}{a2^{nk} + x}\right) \log_2\left(\log_2\left(1 + \frac{1}{a2^{nk} + x}\right)\right). \end{aligned}$$

Since $\log_2(\log_{2^n}(x)) = \log_2(\log_2(x)) - \log_2(n)$ and the probabilities sum to one, we see that

$$\begin{aligned} H_n(k) &= -\sum_{a=1}^{2^n-1} \sum_{x=0}^{(2^n)^k-1} \log_{2^n}\left(1 + \frac{1}{a(2^n)^k + x}\right) \log_2\left(\log_{2^n}\left(1 + \frac{1}{a(2^n)^k + x}\right)\right) \\ &= \log_2(n) - \frac{1}{n} \sum_{a=1}^{2^n-1} \sum_{x=0}^{2^{nk}-1} \log_2\left(1 + \frac{1}{a2^{nk} + x}\right) \log_2\left(\log_2\left(1 + \frac{1}{a2^{nk} + x}\right)\right) \\ &= \log_2(n) + \frac{1}{n} \sum_{j=0}^{n-1} H_2(kn + j). \end{aligned}$$

From our previous work, we also know (within 10^{-6}) that when $m > 6$, $H_2(m) = m + \frac{1}{2} + \log_2(\ln(2))$ so that $H_2(m + 1) = H_2(m) + 1$. Thus

$$\begin{aligned} H_n(k) &= \log_2(n) + \frac{1}{n} \sum_{j=0}^{n-1} [H_2(kn) + j] \\ &= \log_2(n) + H_2(kn) + \frac{1}{2}(n - 1) \\ &= kn + \frac{1}{2}n + \log_2(n) + \log_2(\ln(2)). \end{aligned}$$

On the other hand, if we allocate the same number of bits to a binary-normalized number, the entropy is

$$H_2(kn + n - 1) = kn + n - \frac{1}{2} + \log_2(\ln(2)).$$

Hence the difference in information content between n-bit-normalized and binary-normalized numbers is

$$H_n(k) - H_2(kn + n - 1) = \log_2(n) - \frac{1}{2}(n - 1).$$

As one can seen in line 4 of TABLE 3, hexadecimal-normalized mantissas carry $\frac{1}{2}$ bit more information than binary-normalized mantissas of the same length. Indeed, n-bit normalization is preferred over binary for $2 \leq n \leq 6$ with 3-bit normalization being the most efficient.

TABLE 3: Information advantage of n-bit vs. binary normalization

Normalization index (n)	$H_n(k) - H_2(kn + n - 1)$
1	0
2	0.500 000
3	0.584 963
4	0.500 000
5	0.321 928
6	0.084 063
7	−0.192 656

So it turns out that, while the use of hexadecimal normalization is ineffective for the (conjectured) original aim of reducing the frequency of normalization shifts, the choice does result in a slight reduction in redundancy. On the other hand, the IEEE standard takes advantage of the fact that the initial bit of a binary-normalized number must be one and drops it from the representation. When this is done (at the cost of making it more awkward to test that a real value is exactly 0) the redundancy of the binary representation is reduced by 1 and the advantage swings the other way. The IEEE standard has less redundancy than any other normalization option.

REFERENCES

1. R. Ash, *Information Theory*, Interscience Publishers, New York, 1965.
2. F. Benford, The law of anomalous numbers, *Proc. Amer. Philos. Soc.* **78** (1938) 551–572.
3. P. Diaconis, The distribution of leading digits and uniform distribution mod 1, *Annals of Prob.* **5** (1977) 72–81. doi:10.1214/aop/1176995891
4. P. Diaconis and D. Freedman, On rounding percentages, *J. Amer. Stat. Assoc.* (1979) 359–364. doi:10.2307/2286335
5. B. J. Flehinger, On the probability that a random integer has initial digit, *Amer. Math. Monthly* **73** (1966) 1056–1061. doi:10.2307/2314636
6. R. W. Hamming, On the distribution of numbers, *The Bell System Technical Journal* **49**(8) (1970) 1609–1625.
7. T. Hill, Base-invariance implies Benford's law, *Proc. Amer. Math. Soc.* **123**(3) (March 1995) 887–895.
8. ———, A statistical derivation of the significant-digit law, *Statistical Science* **10**(4) (1995) 354–363.

9. D. Knuth, *The Art of Computer Programming*, 2nd ed., vol. 2, Addison-Wesley, Menlo Park, CA, 1981.
10. S. Newcomb, Note on the frequency of the use of different digits in the natural numbers, *Amer. J. Math.* **4** (1881) 39–40. doi:10.2307/2369148
11. R. S. Pinkham, On the distribution of first significant digits, *Ann. Math. Stat.* **32** (1961) 1223–1230. doi:10.1214/aoms/1177704862
12. R. A. Raimi, On the distribution of first significant figures, *Amer. Math. Monthly* **74**(4) (1969) 343–348.
13. ———, The First Digit Problem, *Amer. Math. Monthly* **83**(7) (1976) 521–538. doi:10.2307/2319349
14. P. R. Turner, The distribution of leading significant digits, *IMA Jor. of Num. Anal.* **2** (1982) 407–412. doi:10.1093/imanum/2.4.407
15. ———, Further revelations on leading significant digits, *IMA Jor. of Num. Anal.* **4** (1984) 225–231. doi:10.1093/imanum/4.2.225
16. ———, Will the "real" real arithmetic please stand up? *Notices Amer. Math. Soc.* **38** (1991) 298–304.
17. F. J. Taylor, R. Gill, J. Joseph, and J. Dadke, A 20 bit logarithmic number system processor, *IEEE Trans. on Comp.* **7**(2) (1988) 190-199.
18. J. F. Young, *Information Theory*, Prentice Hall, New York, 1971.

Summary When data with a nonuniform distribution (such as the leading digits of real numbers) is represented using fixed length codes, the representation is inefficient. How is the distribution of leading digits related to the distribution of real number mantissas and what implications does this have for the efficiency of the standard computer representation of real numbers?

Blood Vessel Branching: Beyond the Standard Calculus Problem

JOHN A. ADAM
Department of Mathematics & Statistics
Old Dominion University
jadam@odu.edu

As one who is very thankful to have survived a heart attack in 1996, I have more than a passing interest in the vascular system. Perhaps this is due to the following statement by my cardiologist: "Your left anterior descending coronary artery is 99.9% blocked." While that might have been an exaggeration, the sight of my artery narrowing to a thread on the angiogram was certainly a cause for concern.

The *vascular system* is the collection of all arteries, veins, and capillaries that permit the flow of blood from the heart, around the body, and back to the heart again. A combination of fluid flow and diffusion brings every cell in a healthy body within reach of the nutrients and gases it needs. The combined length of all this plumbing is rather large: Indeed, in 1967 the mathematical biologist Robert Rosen [**5**] made the claim that the total length of the vascular system in large mammals is of the order of 50,000 miles!

The purpose of this article is to describe a significant application of mathematics to the understanding of the vascular system. Over a century ago, in 1878, Wilhelm Roux [**6**] described the relationships among the angles and radii of bifurcating blood vessels, showing how they arise from some simple principles of optimality. Our goals here are to rederive Roux's results for the benefit of a wider audience, to examine their implications, and to establish how relevant they may be to the human vascular system. To accomplish this, we first study the equations for the flow of fluid in a pipe and establish a fundamental result, which is standard fare (or once was) in most engineering-style calculus books. We then present a sequence of increasingly faithful models of blood-vessel branching. In these models, we judge optimality using a succession of so-called *cost functionals*.

Our treatment is based on some rather obscure material, referred to in passing by Thompson [**9**], and in somewhat more detail by Rosen [**5**], but expanded considerably here.

At the outset we acknowledge some simplifying assumptions.

The first concerns the nature of fluids. Newton modeled the simplest kind of fluids by imagining them to consist of thin layers that slide past one another, developing a resistance to the flow. An important factor of proportionality—the ratio of "drag" to the velocity gradient perpendicular to the direction of flow—is called *viscosity*. A fluid with the property that the viscosity is independent of the forces applied to it, is called a *Newtonian fluid*. Naturally enough, if the viscosity does depend on the applied force, then the fluid is called *non-Newtonian*.

In this paper we model blood as a Newtonian fluid. This proves to be a reasonable approximation except when blood cells are forced through the tiniest capillaries, which are barely as large as the blood cells themselves. In these capillaries, it is not reasonable to think of layers of fluid sliding past each other, so our models do not apply.

Furthermore, the pressure driving the whole system is far from constant; there are short time lags between the high pressure and the peak blood flow because of the

Math. Mag. **84** (2011) 196–207. doi:10.4169/math.mag.84.3.196.

inertia of the blood. In addition, the blood vessels are flexible, not rigid (though of course, hardening of the arteries occurs with age). Cumulatively, this can modify the blood dynamics, as do the surface waves along the vessel walls, induced by the changes in pressure and flow generated by the heart. Such waves can be reflected at various points in the arterial tree: at branching junctions, or where the vessel diameter changes significantly. By contrast, we always assume time-independence and a constant shape for the blood vessels.

Having acknowledged these realities, we proceed to build our first model.

Viscous fluid flow

We begin by describing the flow of a fluid, such as blood, through a cylindrical pipe, such as a large blood vessel. We will derive equations for the velocity of the fluid (which varies with location in the pipe) and for the amount of fluid that passes through the pipe per unit time.

Our model is ultimately based on the *Navier-Stokes equations*. These equations arise from applying Newton's second law of motion to viscous fluids, which we imagine to consist of thin layers, each small element of which must obey conservation of momentum. In simplest form, the force on each element of fluid is the sum of a pressure-gradient term and a viscous term. For an incompressible Newtonian fluid, the viscous term, mathematically, is proportional to the Laplacian of the velocity field. This is important because it implies that momentum is diffused in the same way heat diffuses; indeed in the simplest example of unidirectional time-dependent flow, the Navier-Stokes equations reduce to the heat (or diffusion) equation.

Now consider the flow of a fluid in a uniform circular cylindrical pipe of radius a. The fluid is assumed to be incompressible and independent of time, and subject to a fluid pressure that decreases along the pipe in the positive direction. The flow is described by a velocity field $\mathbf{u}$, which is a vector field, giving the velocity of the fluid at each point inside the pipe. Under these circumstances the velocity field $\mathbf{u}$ is in the axial direction only, which means that, in cylindrical coordinates, $\mathbf{u} = \langle 0, 0, u(r)\rangle$, $0 \le r \le a$. In the presence of a uniform pressure gradient P_x along the pipe, the fluid velocity $u(r)$ satisfies the differential equation

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{du}{dr}\right) = \frac{P_x}{\mu}, \tag{1}$$

where μ is the coefficient of viscosity. (You can find a derivation of this equation in any standard text on fluid dynamics [**3**].) Note that the left-hand side of equation (1) is just the radial part of the Laplacian in cylindrical coordinates. Integrating the equation twice, we find that

$$u = u(r) = \frac{P_x r^2}{4\mu} + B\ln r + C,$$

where B and C are constants to be determined. If $u(0)$ is to be well defined then $B \equiv 0$ of course, and because of friction, the velocity is zero at the inner surface of the pipe, so $u(a) = 0$, and this implies that $C = -P_x a^2/4\mu$. Therefore the velocity profile is a parabolic one, moving the direction opposite the pressure gradient:

$$u(r) = \frac{-P_x}{4\mu}(a^2 - r^2). \tag{2}$$

We can use (2) to compute the rate at which volume flows through the vessel. The technical term for volume flow rate is the *volume flux* per unit time through any cross section (as in FIGURE 1). It is given by the integral

$$V(a) = \int_0^a 2\pi r u(r)\, dr,$$

where V is expressed in units of volume/time. The *mass flux* is determined by multiplying this expression by the (constant) fluid density ρ.

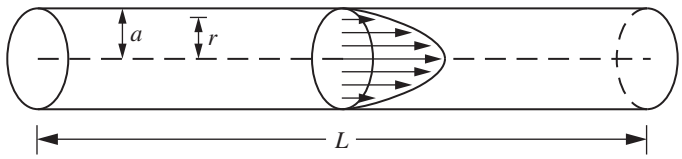


Figure 1 Fluid flow in a cylindrical pipe

Substituting the above expression for u,

$$V(a) = \frac{-\pi P_x}{2\mu} \int_0^a r\left(a^2 - r^2\right) dr = \frac{-\pi P_x a^4}{8\mu}. \tag{3}$$

This result is known as *Poiseuille's law*. Since P_x is the constant negative pressure gradient, we may write it in terms of the pressure difference, ΔP, over the length L of the pipe, that is, $P_x = -\Delta P/L$, so the volume flux (or equivalently, the how much blood flows per unit time) is

$$V(a) = \frac{\pi \Delta P a^4}{8\mu L} \propto a^4 L^{-1}.$$

The fourth power makes volume flux very sensitive to the radius of the pipe, which has significant implications for vascular disease in particular. Halving the radius of the tube results in a volume flux *one sixteenth* of the original, and even a reduction in radius of about 16% will *halve* the volume flow rate!

The reciprocal of V has units of time/volume and can be taken as a measure of the frictional resistance. Given a fixed pressure drop, the frictional resistance is directly proportional to the length of the pipe. The dependence of the resistance on a^{-4} is perhaps harder to visualize, but a standard analogy may prove helpful. Poiseuille's law is effectively a hydraulic metaphor for *Ohm's law* in electrical circuits. This law states that the voltage E in the circuit is equal to the product of the current I and the resistance R: that is, $E = IR$. In fluid flow, the pressure drop ΔP is the analog of the voltage, and the volume flux V is equivalent to the current. Thus the hydraulic resistance can be expressed as

$$\frac{\Delta P}{V} = \frac{8\mu L}{\pi a^4} \propto a^{-4} L. \tag{4}$$

The idea, then, is that the resistance to the flow of a fluid through a pipe, or a system of pipes, measures the amount of work required to pump the fluid through the system. The lower the resistance, the more efficient the system will be. Our first method to find the best angle of branching (and hence the location) of a smaller vessel is to minimize resistance along the branching path. Thus, our first cost functional is based on the total hydraulic resistance of the system to the flow.

The standard calculus model

We consider the case of a smaller vessel branching off from a larger one. For the time being, we consider the larger vessel to be straight. FIGURE 2 shows the basic configuration: The main blood vessel of radius r_1 has a narrower vessel of radius r_2 ($< r_1$) branching at an angle θ to the axis of the former.

We consider the points A, B, and C to be fixed, with BC perpendicular to AC and with distances denoted by $c = |AC|$ and $b = |CB|$. The choice of θ determines the location of the point O along with the distances $L_1 = |AO| = c - b\cot\theta$ and $L_2 = |OB| = b\csc\theta$.

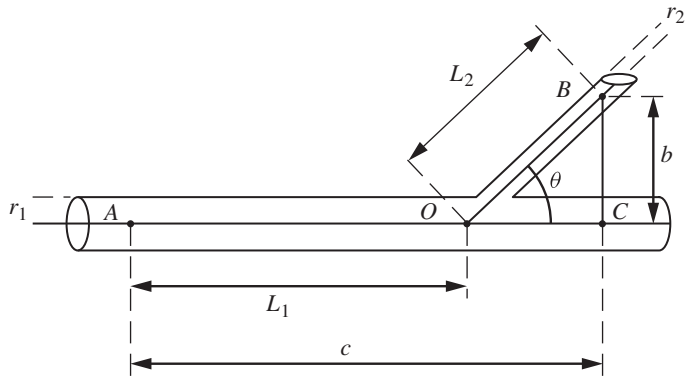


Figure 2 Blood vessel branching from an artery

The first cost functional we examine is one found in many elementary calculus books [**8**]. We use (4) to calculate the minimum total hydraulic resistance along the path AOB.

Since we have established that the hydraulic resistance, according to Poiseuille's law, is proportional to length and inversely proportional to the fourth power of the radius for each component of the system, we may express the total resistance $\mathcal{L}_1$ as

$$\begin{aligned}\mathcal{L}_1 &= k\left[\frac{L_1}{r_1^4} + \frac{L_2}{r_2^4}\right] \\ &= k\left[\frac{c - b\cot\theta}{r_1^4} + \frac{b\csc\theta}{r_2^4}\right].\end{aligned}$$

The constant k will be significant later. To minimize resistance as a function of θ, set

$$\frac{d\mathcal{L}_1}{d\theta} = \frac{kb}{r_2^4}\csc^2\theta\left[\left(\frac{r_2}{r_1}\right)^4 - \cos\theta\right] = 0.$$

Then the only critical value is

$$\theta_m = \cos^{-1}\left(\frac{r_2}{r_1}\right)^4,$$

and a quick verification shows that $\mathcal{L}_1$ assumes an absolute minimum on the interval $(0, \pi/2)$ at θ_m. We substitute some relative values for the radii to test our model: If $r_2 = 0.5r_1$, then $\theta \approx 86°$; and if $r_2 = 0.8r_1$, then $\theta \approx 66°$.

The first angle is almost a right angle. Is our model sound enough to justify such an extreme prediction? One concern is that the cost functional relates only to the path *AOB*; the implied assumption is that the effect of θ on the path *AOC* can be neglected. We will address that issue soon; but first, we consider variations of the cost functional.

Varying the cost functional

In addition to minimizing the total resistance to flow, Roux and Rosen consider minimizing the cost of *maintaining* the structure. It seems reasonable to suppose that the maintenance cost should be proportional to the *volume* of the system. The volume along the path *AOB* is given by

$$\mathcal{L}_2 = K\left(L_1 r_1^2 + L_2 r_2^2\right)$$

where K is another constant. This is minimized when

$$\theta = \cos^{-1}\left(\frac{r_1}{r_2}\right)^2,$$

which is *not* the same as the value θ_m that minimizes the previous functional $\mathcal{L}_1$.

We can also combine these functionals. Our next cost functional includes a term proportional to resistance and a term proportional to volume:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_1 + \mathcal{L}_2 \\ &= k\left(\frac{L_1}{r_1^4} + \frac{L_2}{r_2^4}\right) + K\left(L_1 r_1^2 + L_2 r_2^2\right) \\ &= L_1\left(\frac{k}{r_1^4} + K r_1^2\right) + L_2\left(\frac{k}{r_2^4} + K r_2^2\right) \\ &\equiv L_1 A_1 + L_2 A_2, \end{aligned} \tag{5}$$

where k and K are constants of proportionality. Proceeding in the same manner as before, we find that

$$\frac{d\mathcal{L}}{d\theta} = A_2 b \csc^2\theta\left[\frac{A_1}{A_2} - \cos\theta\right]$$

which is zero when

$$\theta = \cos^{-1}\left[\frac{A_1}{A_2}\right] = \cos^{-1}\left[\left(\frac{r_2}{r_1}\right)^4\left(\frac{k + K r_1^6}{k + K r_2^6}\right)\right].$$

This result depends on the relative values of k and K. For any positive values of the constants, it agrees with the original minimizer θ_m only in the two limiting situations: (i) $r_2 \to r_1$ and (ii) $r_1 \to 0$.

Observations

How can we choose a cost functional? More generally, how can we choose a good model? One test is how well its predictions match observations.

Roux gave a set of empirical observations about about vascular branching (cited separately by Thompson [**9**] and Rosen [**5**]). In paraphrased form, they are:

(1) *If an artery bifurcates into two equal branches, they deviate at equal angles from the main stem.*

(2) *If the branches do not have the same radius, then the larger branch (i.e., the continuation of the main artery) makes a smaller angle with the original direction than does the smaller branch.*

(3) *Branches that are so narrow that they do not significantly diminish the flow in the main stem branch off at large angles (typically between* 70° *and* 90°*).*

Considering units

One way to understand the combined functional (5) is as a measure of the *power dissipated* by the flow. Power is defined as work done per unit time, in this case by the flow, and work has the dimensions of force times distance. To make this clear, we need to understand the units of the constants k and K.

In *dimensional analysis*, use is often made of the units of mass, length, and time (denoted by $[M]$, $[L]$, and $[T]$, respectively). These define the dimensions of particular physical quantities, and are called *fundamental units*. They are analogous to basis vectors for a vector space, such as the unit vectors $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ for the space $\mathcal{R}^3$. Thus, the dimension of force being mass $\times$ length / time^2 is denoted $[M][L][T]^{-2}$. (Other choices of fundamental units are possible; for example, mass, length, and force could have been used just as well. The author has written elsewhere about this and the valuable Buckingham Pi Theorem [**1**].)

Considering our combined functional (5), we see that the first parenthetical expression has units $[L]^{-3}$. But from the definition above, power has units

$$\left([M][L][T]^{-2}\right)[L][T]^{-1} = [M][L]^2[T]^{-3}.$$

Therefore, for the term itself to be in units of power, the constant k must have units of

$$[M][L]^5[T]^{-3}.$$

By the same token, the second parenthetical expression has units $[L]^{+3}$, so the constant K must have units

$$[M][L]^{-1}[T]^{-3}.$$

These considerations would be the starting point for any serious inquiry into what determines the values of the constants k and K.

Optimizing a single vessel

Next, we consider the contribution to $\mathcal{L}$ from a single unbranched vessel of radius r and length L:

$$\mathcal{L} = kLr^{-4} + KLr^2. \tag{6}$$

Biologically, this means that the vessel's contribution to the total cost is determined by two competing factors: the power required to maintain the volume of blood needed to fill the vessel, and the power required to pump blood through that same vessel. The first is proportional to the inverse fourth power of the radius, and for optimality this requires that the radius be as large as possible; the second is proportional to the square of the radius, and so the radius should be as small as possible [**11**].

If we expect that the branching configuration minimizes the functional, then we should expect that the radius of each single vessel should minimize it as well. From equation (6),

$$\frac{d\mathcal{L}}{dr} = (-4kr^{-5} + 2Kr)L$$

which is zero when $k = Kr^6/2$. Since $\mathcal{L}''(r) > 0$, this is a minimum. For the optimal value of r we have

$$\mathcal{L} = \beta L r^2, \tag{7}$$

where $\beta = 3K/2$ is another constant.

Branching versus bifurcation

Thus far the cost functionals used have referred to a smaller vessel branching off a larger one, not bifurcation. This will now be amended.

From this point forward we consider all three branches in FIGURE 3. We consider the points A, B, and C fixed. The choice of the point O (in the same plane) determines the branching angles ϕ and θ.

The radii and lengths of the vessels AO, OB, and OC are denoted respectively by r_0, L_0, r_1, L_1, and r_2, L_2.

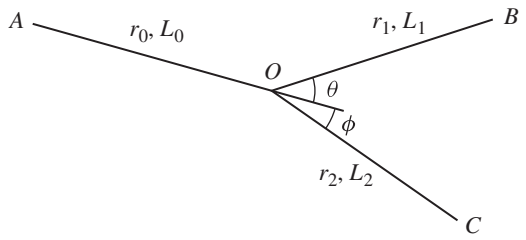


Figure 3 Bifurcation

Optimizing the configuration

The important relation (7) will be used in connection with the three diagrams in FIGURE 4. In each diagram we consider a small change in the location of the bifurcation point O to a new point O'. In each case, the distance from O to O' is δ, which is small compared with the distances AO, OB, and OC.

Now from the theorem from the calculus of variations known as the Principle of Minimum Work [**7**], we know that a small *first-order* change in the minimal configuration results in a *second-order* change in the cost functional, so to the first order, the value of the functional is unchanged. In each part of FIGURE 4, this means that the change in $\mathcal{L}$, summed over the three vessels in the configuration, must be zero (to first order in δ). We will see what this implies in each diagram.

Using the law of cosines for FIGURE 4(a), we have

$$(O'B)^2 = L_1^2 + \delta^2 - 2\delta L_1 \cos\theta \approx L_1^2 \left(1 - \frac{2\delta}{L_1}\cos\theta\right).$$

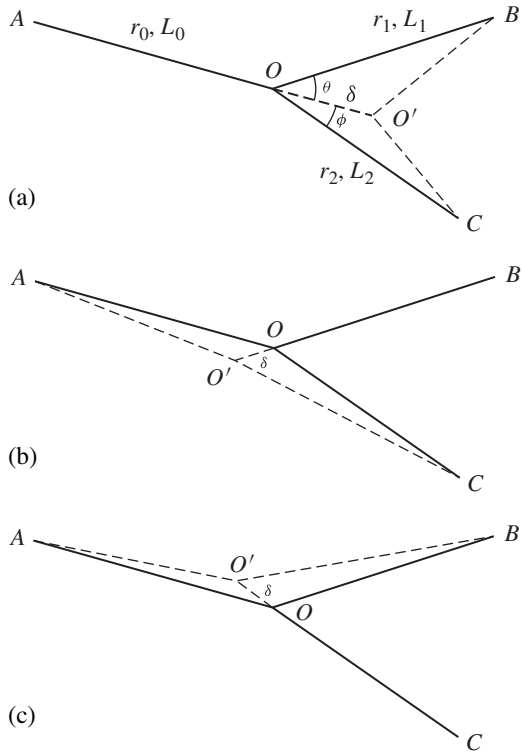


Figure 4 First-order changes to the configuration

Therefore

$$O'B \approx L_1\left(1 - \frac{\delta}{L_1}\cos\theta\right) = L_1 - \delta\cos\theta,$$

so the path OB has been shortened by approximately $\delta\cos\theta$. Similarly, the path OC is shortened by $\delta\cos\phi$, and of course, the path AO has been increased by the amount δ. Now we apply the result $\mathcal{L} = \beta L r^2$ for each branch. The total increment in cost is zero, and this is equal to the sum of these cost increments. After canceling the common factors β and δ, we find that

$$r_0^2 = r_1^2\cos\theta + r_2^2\cos\phi.$$

In FIGURE 4(b), the length of OB is increased by δ. It is readily shown using the law of cosines that AO is reduced by approximately $\delta\cos\theta$, OC is increased by approximately $\delta\cos(\theta+\phi)$, and once again, the cost increments sum to zero, so that

$$r_1^2 = r_0^2\cos\theta - r_2^2\cos(\theta+\phi). \tag{8}$$

The result corresponding to FIGURE 4(c) is most easily accomplished by interchanging r_1 and r_2 and θ and ϕ in the result (8):

$$r_2^2 = r_0^2\cos\phi - r_1^2\cos(\theta+\phi).$$

From these three equations we can solve for the three angles to obtain

$$\cos\theta = \frac{r_0^4 + r_1^4 - r_2^4}{2r_0^2r_1^2};\ \cos\phi = \frac{r_0^4 + r_2^4 - r_1^4}{2r_0^2r_2^2};\ \cos(\theta+\phi) = \frac{r_0^4 - r_1^4 - r_2^4}{2r_1^2r_2^2}. \tag{9}$$

These results can be taken further. For optimal flow in a single vessel, the flow rate f (for a given value of L) turns out to be proportional to the cube of the radius. When vessels branch the flow must split into two parts, which means that $f_0 = f_1 + f_2$ and

$$r_0^3 = r_1^3 + r_2^3. \tag{10}$$

(This equation might appear to contradict Poiseuille's law (3), which might suggest that $r_0^4 = r_1^4 + r_2^4$. But applying Poiseuille's law in this case would require identical pressure gradients P_x in all three branches, which is not the case in general.)

Eliminating the superfluous radius in each case we obtain

$$\cos\theta = \frac{r_0^4 + r_1^4 - \left(r_0^3 - r_1^3\right)^{4/3}}{2r_0^2 r_1^2};\ \cos\phi = \frac{r_0^4 + r_2^4 - \left(r_0^3 - r_2^3\right)^{4/3}}{2r_0^2 r_2^2}, \tag{11}$$

and a similar equation for $\cos(\theta + \phi)$.

To examine some consequences of this newest cost functional, let $r_1 = r_2$, corresponding to equal branchings from the primary vessel. According to (11), this implies that $\theta = \phi$: *Vessels with equal radii branch off at equal angles to the main stem.* This is exactly the first empirical law of Roux.

Next let $r_2 = \alpha r_1$ with $0 < \alpha < 1$. The formulas in (9) imply that

$$\cos\theta = \frac{r_0^4 + r_1^4(1-\alpha^4)}{2r_0^2 r_1^2};\ \cos\phi = \frac{r_0^4 - r_1^4(1-\alpha^4)}{2\alpha^2 r_0^2 r_1^2}.$$

Is it possible to determine from these equations which of the two angles θ or ϕ is the larger? For angles in $(0, \pi/2)$, the equation $\cos(\theta) > \cos(\phi)$ is equivalent to

$$\left(1+\alpha^2\right)^2 r_1^4 > r_0^4.$$

But $r_0^3 = r_1^3 + r_2^3$ leads to $r_0^4 = (1+\alpha^3)^{4/3} r_1^4$. This and the easily verified inequality

$$\left(1+\alpha^2\right)^2 > \left(1+\alpha^3\right)^{4/3}$$

show that for $\theta, \phi \in (0, \pi/2)$, $\theta < \phi$. Thus *the larger vessel branches off at a smaller angle than the narrower vessel.* This is the second empirical law of Roux.

Now suppose that $r_2 \ll r_1$ (so that $\alpha \ll 1$). This also implies that $r_1 \lesssim r_0$. Then using the results

$$\cos\phi = \frac{r_0^4 - r_1^4(1-\alpha^4)}{2\alpha^2 r_0^2 r_1^2} \text{ and } r_0^4 = \left(1+\alpha^3\right)^{4/3} r_1^4,$$

we may write

$$\cos\phi = \frac{\left(1+\alpha^3\right)^{4/3} - 1 + \alpha^4}{2\alpha^2\left(1+\alpha^3\right)^{2/3}}.$$

For a branching angle $\phi = 70°$, this gives $\alpha \approx 0.34$; and if $\phi = 80°$, then $\alpha \approx 0.17$, clearly indicating that *the branching angle steepens as the vessel radius narrows.* Retaining only the leading terms in α we obtain

$$\alpha \to 0 \text{ as } \alpha \to 0,$$

so $\phi \to \pi/2$ as $\alpha \to 0$, that is, the branching angle approaches 90° as the vessel radius tends to zero. These statements are the third empirical law of Roux. Thus with the cost

functional (5) all three empirical laws have been verified, leading one to conclude that the model may well be fairly realistic for many animal vasculatures, particularly for the smaller vessels.

Calculating total length

We now apply these ideas to examine the statement of Rosen, quoted at the beginning of the article, regarding the total length of the (large) mammalian vascular system. Suppose for simplicity that the primary branch (aorta) bifurcates into two vessels with equal radii: if $r_1 = r_2$ the statement $r_0^3 = r_1^3 + r_2^3$ is equivalent to $r_1 = 2^{-1/3} r_0 \approx 0.794 r_0$. Suppose further that every vessel in a generic vascular system arises from an equal bifurcation of the immediately preceding larger vessel. In mammals, a typical capillary radius is about 5 microns, and it is of interest to calculate how many bifurcations from a given primary vessel are required to reach this radius. Therefore the number n of bifurcations from an aorta of radius r_0 that will result in such a capillary must satisfy the equation

$$(0.794)^n r_0 = 5 \times 10^{-6} \text{ m}.$$

For an animal such as a dog, $r_0 \approx 0.5$ cm $= 5 \times 10^{-3}$ m, so

$$(0.794)^n = 10^{-3}, \text{ and } n = \frac{-3}{\log_{10}(0.794)} \approx 30. \tag{12}$$

It follows from this that the number of vessels in the system resulting from the final bifurcation is $2^{30} \approx 1.07 \times 10^9$, which is close to the estimate of 1.2×10^9 given by Rashevsky [**4**] and Zamir [**11**]. As pointed out by Rosen [**5**], the agreement becomes even closer when one realizes that observationally, it is probably difficult to distinguish vessels in the final bifurcation from those arising from several preceding ones (the 29th or even 28th bifurcation).

Now we try to estimate of the total *length* of the vascular system. In the absence of much physiological information (but see below), and in the spirit of a back-of-the-envelope calculation, we suppose that the primary vessel, the aorta, has length L_0, and that each bifurcation k produces twice as many vessels, each of length $L_k = \eta L_{k-1}$, where it is assumed that $0 < \eta < 1$. Then the total length of the system after n bifurcations is

$$\begin{aligned} L_n &= L_0 + 2\eta L_0 + 2^2 \eta L_1 + 2^3 \eta L_2 + \cdots + 2^n \eta L_{n-1} \\ &= L_0 \left(1 + 2\eta + (2\eta)^2 + (2\eta)^3 + \cdots + (2\eta)^n\right) \\ &= L_0 \frac{\left[(2\eta)^{n+1} - 1\right]}{2\eta - 1}, \eta \neq \frac{1}{2}. \end{aligned} \tag{13}$$

For discussion purposes, based on the table below, we suppose that $L_0 = 40$ cm, and take various values for $\eta > 1/2$. For $\eta = 2/3$,

$$L_{30} = 0.4 \frac{\left[(4/3)^{31} - 1\right]}{1/3} \text{ m} \approx 9 \times 10^3 \text{ m } \approx 10 \text{ km};$$

whereas if $\eta = 7/8$, then

$$L_{30} = 0.4 \frac{\left[(7/4^{31} - 1\right]}{3/4} \text{ m} \approx 2 \times 10^7 \text{ m } = 2 \times 10^4 \text{ km}.$$

If $\eta > 1$ (a given bifurcation is longer than the immediately preceding one), the length explosion is even more rapid: $\eta = 6/5$ results in $L_{30} \approx 6 \times 10^7$ km! If Rosen's estimate of 50,000 km for the total length of the vascular system in large mammals is accurate, then we would on average expect η to be slightly less than $15/16$ on the basis of this approach. It would appear that since real vascular systems are unlikely to have this degree of predictability, our version of Rosen's estimate is at best an educated guess!

The table below, adapted from that in Zamir [**11**], is based on the so-called "classical model," wherein the arterial tree was thought of as being highly ordered on a global scale. It remains essentially unmodified today apart from some numerical changes, but the main difficulty, according to Zamir, is not its numerical accuracy, but its central premise, namely that all these vessels have a well defined beginning, end, length, and diameter. Such conceptual considerations, while interesting, would take us too far afield in this paper. Nevertheless, the approximate length of the arterial tree (of the dog) is obtained from level nine of the classical model, yielding about 10^8 cm or 1000 km, again, considerably lower than Rosen's estimate.

TABLE 1: Hierarchical structure of the arterial tree

Level	Vessel Description	No. of Vessels	Length (cm)	Diameter (mm)
1	aorta	1	40	10
2	large arteries	40	20	3
3	main branches	600	10	1
4	secondary branches	1.8×10^3	4	0.6
5	tertiary branches	7.6×10^4	1.4	0.14
6	terminal arteries	10^6	0.1	0.05
7	terminal branches	1.3×10^7	0.15	0.03
8	arterioles	4×10^7	0.2	0.02
9	capillaries	1.2×10^9	0.1	0.008

Finally, we summarize the appropriateness of the fundamental result (10), which is really the basis for both the estimate for the number of bifurcations in the vascular system of large mammals, and the derivation of Roux's empirical laws. According to Zamir, in his discussion of the branching structure of arterial trees, "A consideral amount of data from the cardiovascular systems of man and animals has been shown to support these results, though with a considerable scatter in some cases." [**11**] Such a cube law is also consistent with there being a constant shear force throughout the arterial network, a conclusion reached by several authors. (Zamir's papers give details.[**10, 11**])

Regarding the number of bifurcations in the arterial tree, Zamir also writes "If the global structure of the tree were purely ad hoc, the number of branching sites along certain paths could, simply by chance, be very large, perhaps of the order of thousands. Yet it is usually found to be surprisingly small, of the order of 30 or so." Furthermore, the average increase in cross-sectional area at an arterial bifurcation, that is, the ratio

$$A = \frac{r_1^2 + r_2^2}{r_0^2},$$

is found to be about 1.26 for symmetrical bifurcations, and somewhat lower for non-symmetrical ones. As is easily shown from equation (10), used in conjunction with the additional statement (used above) that $r_2 = \alpha r_1$, $0 < \alpha < 1$,

$$A = \frac{1 + \alpha^2}{(1 + \alpha^3)^{2/3}}.$$

For $\alpha = 1$, $A = 2^{1/3} \approx 1.26$; for $\alpha = 0.7$, $A \approx 1.22$, and for $\alpha = 0.5$, $A \approx 1.16$. Indeed, the function $A(\alpha)$ is readily seen to possess a single maximum at $(1, 2^{1/3})$ on any interval $[0, b]$, where $b > 1$. Recall that the reciprocal of this maximum value, $2^{-1/3}$, arises naturally in the corresponding bifurcation problem discussed above. It is also noteworthy that the accepted value for the increase in cross-section area from the aorta to the capillary beds is about 1000 [**2**]. Since we have identified the likely bifurcation number in this system as $n = 30$, the definition of A requires that

$$A^{30} = 1000, \quad \text{so that} \quad A = 10^{0.1} \approx 1.26.$$

Acknowledgment I am grateful to Dr. Thomas Klevan and Professor M. Zamir for their kind advice.

REFERENCES

1. J. A. Adam, *Mathematics in Nature: Modeling Patterns in the Natural World*, Princeton University Press, Princeton, 2006.
2. A. C. Burton, *Physiology and Biophysics of the Circulation*, Year Book Medical Publishers, Inc., 1965.
3. S. N. Curle and H. J. Davies, *Modern Fluid Dynamics*, Vol. 1, Van Nostrand Reinhold, Wokingham, England, 1968.
4. N. Rashevsky, *Mathematical Biophysics*, Vol. 2, Chapter XXVII, Dover, New York, 1960.
5. R. Rosen, *Optimality Principles in Biology*, Butterworth, London, 1967.
6. W. Roux, doctoral thesis on the bifurcation of blood vessels, cited by Thompson [**9**].
7. H. Sagan, *Introduction to the Calculus of Variations*, Dover, New York, 1992.
8. J. Stewart, *Calculus, Early Transcendentals*, 5th ed., Brooks/Cole, Belmont, CA, 2003.
9. D'Arcy W. Thompson, *On Growth and Form*, Dover, New York, 1992.
10. M. Zamir and H. Chee, Segment Analysis of Human Coronary Arteries, *Blood Vessels* **24** (1987) 76–84.
11. M. Zamir, The Branching Structure of Arterial Trees, *Comments on Theoretical Biology* **1** (1988) 15–37.

Summary Calculating the optimal angle for blood vessel branching is a standard calculus problem. However, optimality in that setting is judged by a *cost functional* that turns out not to give realistic results. We study a sequence of improvements to the cost functional, finally arriving at one that passes an important modeling test: From this last functional, we derive three empirical laws of blood vessel branching, originally proposed by German zoologist Wilhelm Roux.

JOHN A. ADAM is Professor of Mathematics at Old Dominion University in Norfolk, VA. His research interests are currently associated with theoretical problems in meteorological optics. He enjoys nature photography (of the point-and-shoot kind), and is a frequent contributor to the Earth Science Picture of the Day website (http://epod.usra.edu/). He is author of *Mathematics in Nature: Modeling Patterns in the Natural World* and coauthor of *Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin* (both Princeton University Press). His most recent book, *A Mathematical Nature Walk* (also Princeton) was published in June 2009. He enjoys being a grandparent of four small children: the geometric mean of their ages is 1.19 years (at the time of writing).

NOTES

Stirred, Not Shaken, by Stirling's Formula

PAUL LEVRIE
Karel de Grote University College, Hoboken, Belgium
and K. U. Leuven, Heverlee, Belgium
paul.levrie@cs.kuleuven.be

In this note we present a smooth and easy derivation of Stirling's asymptotic formula for $n!$,

$$n! \sim \frac{n^n\sqrt{2\pi n}}{e^n}, \quad n \to \infty. \tag{1}$$

We use the notation for asymptotic equivalence, so that

$$f(n) \sim g(n), \quad n \to \infty \qquad \text{means} \qquad \lim_{n\to\infty} \frac{f(n)}{g(n)} = 1.$$

Thus (1) is equivalent to

$$\lim_{n\to\infty} \frac{n^n\sqrt{2\pi n}}{n!\,e^n} = 1.$$

Over the years many proofs of this result have been published. Quite a number of them use tools such as the gamma function (see [**5**] for a recent one), or the Euler-Maclaurin summation formula (see the historic note in [**2**]). For an extensive bibliography, see [**3**], [**6**].

Our method is based on Wallis's product formula for π (most published proofs use this formula, see for instance [**1**]) and the trapezoidal rule for approximating an integral.

The trapezoidal rule

The trapezoidal rule for approximating a definite integral can be found in most calculus textbooks (see [**1**]). We will need an error estimate, and the one given in the following theorem will suffice:

THEOREM. *If for the function f the second derivative exists for all $x \in [a, b]$, then*

$$\int_a^b f(t)\,dt - \frac{f(a)+f(b)}{2}\cdot(b-a) = -\frac{f''(c)}{12}\cdot(b-a)^3$$

for some $c \in [a, b]$.

Math. Mag. **84** (2011) 208–211. doi:10.4169/math.mag.84.3.208.

Proof. The proof uses Rolle's mean value theorem. Define the constant K by:

$$\int_a^b f(t)\,dt - \frac{f(a)+f(b)}{2}\cdot(b-a) = K\cdot(b-a)^3.$$

As a consequence we have that the function

$$F(x) = \int_a^x f(t)\,dt - \frac{f(a)+f(x)}{2}\cdot(x-a) - K\cdot(x-a)^3$$

satisfies $F(a) = F(b) = 0$. Hence we may apply Rolle's theorem to F and so we know that for some $t \in (a,b)$ the derivative

$$F'(x) = f(x) - \frac{f'(x)}{2}\cdot(x-a) - \frac{f(a)+f(x)}{2} - 3K\cdot(x-a)^2 \tag{2}$$

is equal to zero: $F'(t) = 0$. Furthermore by evaluating F' for $x = a$ we find that $F'(a) = 0$. Applying Rolle's theorem a second time, but now to F', we find a value $c \in (a,t)$ with $F''(c) = 0$. From (2) we get

$$F''(x) = -\frac{f''(x)}{2}\cdot(x-a) - 6K\cdot(x-a).$$

So $F''(c) = 0$ implies that

$$K = -\frac{f''(c)}{12}.$$ ∎

Wallis's formula for π

The well known Wallis product formula for π states:

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

It follows immediately from Euler's product formula for the sine function [**1**]

$$\frac{\sin x}{x} = \left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\cdots\left(1-\frac{x^2}{n^2\pi^2}\right)\cdots$$

by taking $x = \frac{\pi}{2}$. (See also [**7**] for a nice proof without calculus or [**4**] for a proof without Euler's product formula.)

We need a concise form of Wallis's formula. To find it, we take the square root in (3) and rearrange the result:

$$\sqrt{\frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdots\frac{2n\cdot 2n}{(2n-1)\cdot(2n+1)}} = \frac{2\cdot 4\cdots(2n)}{1\cdot 3\cdots(2n-1)}\cdot\frac{1}{\sqrt{2n+1}}.$$

We insert some matching factors into the denominator and the numerator:

$$\frac{2\cdot 4\cdot 6\cdots(2n)}{1\cdot 3\cdot 5\cdots(2n-1)}\cdot\frac{1}{\sqrt{2n+1}} = \frac{2^2\cdot 4^2\cdot 6^2\cdots(2n)^2}{1\cdot 2\cdot 3\cdot 4\cdots(2n-1)\cdot 2n}\cdot\frac{1}{\sqrt{2n+1}}.$$

Note that the numerator is equal to $2^{2n}\cdot(n!)^2$. Hence Wallis's formula reduces to:

$$\lim_{n\to\infty}\frac{(n!)^2}{(2n)!}\cdot\frac{2^{2n}}{\sqrt{2n+1}} = \sqrt{\frac{\pi}{2}}.$$

From this it follows that

$$\frac{(n!)^2}{(2n)!} \sim \frac{\sqrt{\pi n}}{2^{2n}}, \quad n \to \infty. \tag{4}$$

Stirling's formula

If we multiply the left hand side of (4) with $\frac{(2n)!}{n!}$, we get $n!$. We now use the trapezoidal rule to find an asymptotic estimate for $\frac{(2n)!}{n!}$. We start by rewriting this expression in the following way:

$$\frac{(2n)!}{n!} = (2n)(2n-1)\cdots(n+1) = n^n \left(1+\frac{n}{n}\right)\left(1+\frac{n-1}{n}\right)\cdots\left(1+\frac{1}{n}\right) \tag{5}$$

and we take a closer look at the factors in parentheses. If we take the logarithm of this part of the previous equation, we get a sum that reminds us of a Riemann sum:

$$\ln\left(1+\frac{n}{n}\right) + \ln\left(1+\frac{n-1}{n}\right) + \cdots + \ln\left(1+\frac{1}{n}\right).$$

Indeed, if we write it like this:

$$n \cdot \left[\frac{1}{n}\ln\left(1+\frac{1}{n}\right) + \frac{1}{n}\ln\left(1+\frac{2}{n}\right) + \cdots + \frac{1}{n}\ln\left(1+\frac{n}{n}\right)\right] \tag{6}$$

we have a right Riemann sum for $\ln x$ in the interval $[1, 2]$. Since

$$\int_1^2 \ln x \, dx = 2\ln 2 - 1$$

we find that

$$\ln\left[\left(1+\frac{n}{n}\right)\left(1+\frac{n-1}{n}\right)\cdots\left(1+\frac{1}{n}\right)\right] \sim n\cdot(2\ln 2 - 1), \quad n \to \infty.$$

This asymptotic estimate isn't quite good enough; we need to do better. Using the trapezoidal rule instead of the right Riemann sum makes the difference in this case.

The trapezoidal sum is the average of the left and right Riemann sum, and hence equals

$$\frac{1}{2n}\ln 1 + \frac{1}{n}\ln\left(1+\frac{1}{n}\right) + \cdots + \frac{1}{n}\ln\left(1+\frac{n-1}{n}\right) + \frac{1}{2n}\ln 2.$$

Since the first term is zero, this expression differs only in one term from the factor between brackets in (6). Its limit is still $2\ln 2 - 1$, and the theorem above provides us with an error estimate. We use the theorem with $f(x) = \ln x$ for the intervals in our Riemann sum, and sum the results. Since $-f''(x) = \frac{1}{x^2}$ is bounded above by 1 in the interval $[1, 2]$, we get the inequality:

$$0 \le 2\ln 2 - 1 - \left[\frac{1}{n}\ln\left(1+\frac{1}{n}\right) + \cdots + \frac{1}{n}\ln\left(1+\frac{n}{n}\right) - \frac{1}{2n}\ln 2\right] \le \frac{1}{12n^2}.$$

We now multiply by n and rearrange the result:

$$0 \le n(2\ln 2 - 1) + \frac{1}{2}\ln 2 - \ln\left[\left(1+\frac{1}{n}\right)\cdots\left(1+\frac{n}{n}\right)\right] \le \frac{1}{12n}.$$

Applying the exponential function and taking the limit for $n \to \infty$ leads to

$$\left(1+\frac{n}{n}\right)\left(1+\frac{n-1}{n}\right)\cdots\left(1+\frac{1}{n}\right) \sim e^{n\cdot(2\ln 2-1)+\frac{\ln 2}{2}}, \quad n \to \infty.$$

If we multiply this result by n^n and rewrite the right hand side, we get the following asymptotic estimate for (5):

$$\frac{(2n)!}{n!} \sim \frac{2^{2n}n^n\sqrt{2}}{e^n}, \quad n \to \infty. \tag{7}$$

By multiplying the estimates in (4) and (7), we get Stirling's formula (1).

REFERENCES

1. Richard Courant and Fritz John, *Introduction to Calculus and Analysis*, Vol. I, Springer-Verlag, New York, 1989.
2. Persi Diaconis and David Freedman, An elementary proof of Stirling's formula, *Amer. Math. Monthly* **93** (1986) 123–125. doi:10.2307/2322709
3. Diego Dominici, Variations on a theme by James Stirling, *Note Mat.* **28** (2008) 1–13.
4. Paul Levrie and Walter Daems, Evaluating the probability integral using Wallis's product formula for π, *Amer. Math. Monthly* **116** (2009) 538–541. doi:10.4169/193009709X470443
5. Reinhard Michel, The $(n+1)$th proof of Stirling's formula, *Amer. Math. Monthly* **115** (2008) 844–845.
6. Thomas J. Osler, Another intuitive approach to Stirling's formula, *Int. J. Math. Educ. Sci. Technol.* **35** (2004) 111–118. doi:10.1080/00207390310001639925
7. Johan Wästlund, An elementary proof of the Wallis product formula for pi, *Amer. Math. Monthly* **114** (2007) 914–916.

Summary In this note an elementary proof of Stirling's asymptotic formula for $n!$ is given. The proof uses the Wallis formula for π and the trapezoidal rule for the calculation of a definite integral, with error estimate.

A Note on Disjoint Covering Systems—Variations on a 2002 AIME Problem

JOHN W. HOFFMAN
Kent State University
Kent, OH 44242
jhoffma9@kent.edu

W. RYAN LIVINGSTON
University of Georgia
Athens, GA 30602
rlivings@math.uga.edu

JARED RUIZ
University of Nebraska
Lincoln, NE 68588
s-jruiz8@math.unl.edu

We will consider two questions about covering systems, both suggested by Problem 9 from the 2002 American Invitational Mathematics Examination (AIME) [**1**]. The AIME problem, in effect, is the following:

Math. Mag. **84** (2011) 211–215. doi:10.4169/math.mag.84.3.211.

AMIE PROBLEM. Harold, Tanya, and Ulysses paint a very long fence. Harold starts with the first picket and paints every hth picket; Tanya starts with the second picket and paints every tth picket; and Ulysses starts with the third picket and paints every uth picket. If every picket gets painted exactly once, find all possible triples (h, t, u).

After some analysis the only such triples can be seen to be $(3, 3, 3)$ and $(4, 2, 4)$. Hint: if (h, t, u) is such a triple then $h^{-1} + t^{-1} + u^{-1} = 1$.

We use this problem to introduce the topic of covering systems. Our work on this subject led us to the following two questions that are inspired by Problem 9, and for which we will use the same metaphor.

QUESTION 1. Six students, Wesley, Lindsay, Ryan, Nicole, Jared and Kelly are painting a very long fence. Wesley paints every wth picket, Lindsay every lth picket, Ryan every rth picket, Nicole every nth picket, Jared every jth picket and Kelly every kth picket. While painting, they notice that pickets $11, 12, \ldots, 34$ are all painted exactly once. When they are finished the entire fence is painted. Did they paint each picket exactly once?

QUESTION 2. The six students paint as before. After a while, they notice that each of the pickets starting from the 9th picket to the 72nd picket is painted exactly once. If they continue in this fashion, will they paint each picket? If so, will they paint each picket exactly once?

(To be clear, we mean for each student to paint an entire congruence class. For example, Wesley may start at any picket, but then must paint every wth picket in each direction.)

We will answer these questions affirmatively. We start by introducing basic definitions and useful known facts. As is customary for integers a and $m > 0$, by the congruence class $a \pmod{m}$, we mean the set $\{a + jm : j \in \mathbb{Z}\}$.

We need some definitions. Let $\mathcal{A} = \{x \equiv a_i \pmod{m_i} : 1 \leq i \leq k\}$ be a system of congruences. Then

- $\mathcal{A}$ is called a *covering system* (CS) if every integer x satisfies at least one of the congruences in $\mathcal{A}$.
- If $\mathcal{A}$ is a CS and deletion of any of the congruences causes $\mathcal{A}$ to no longer be covering then $\mathcal{A}$ is said to be a *regular covering system* (RCS). Some authors call this type of covering system *irredundant.*
- If every integer satisfies exactly one congruence in $\mathcal{A}$ then it is said to be a *disjoint covering system* (DCS). This is also referred to as an *exact covering system.*

Let $c(x)$ be the number of congruences that $x \in \mathbb{Z}$ satisfies. This function is called the *covering function.* If a system is covering then $c(x) \geq 1$ for all $x \in \mathbb{Z}$, and if a system is a DCS, then $c(x) = 1$ for all $x \in \mathbb{Z}$.

We introduce two lemmas which will prove to be useful throughout this paper. Both were proved by R. J. Simpson [**6**]:

LEMMA 1. (SIMPSON'S LEMMA) *If* $\{x \equiv a_i \pmod{m_i} : 1 \leq i \leq k\}$ *is an RCS, and a prime* p *divides some modulus, then the set* $\{a_i : p \mid m_i\}$ *contains a full set of residues modulo* p.

LEMMA 2. (SIMPSON'S INEQUALITY) *If* $\{x \equiv a_i \pmod{m_i} : 1 \leq i \leq k\}$ *is an RCS and* $\prod_{i=1}^{t} p_i^{\alpha_i}$ *is the canonical prime factorization of* $\operatorname{lcm}(m_1, \ldots, m_k)$*, then* $k - 1 \geq \sum_{i=1}^{t} \alpha_i (p_i - 1)$.

(Simpson's inequality is corollary 2 of [**6**].)

The affirmative answer to Question 1 follows from the following theorem. Observe that six students are told that the entire fence is painted, and they painted $24 = 3 \cdot 2^{6-3}$ consecutive pickets exactly once.

THEOREM 1. *If* $\mathcal{A} = \{x \equiv a_i \pmod{m_i} : 1 \le i \le k\}$ *is an RCS such that* $c(x) = 1$ *for* $\max\left\{2, 3 \cdot 2^{k-3}\right\}$ *consecutive integers, then* $\mathcal{A}$ *is a DCS.*

Proof. We proceed by induction on k, the number of congruences in the system. If $k = 2$ then an RCS must have the form $\{x \equiv 0 \pmod 2, x \equiv 1 \pmod 2\}$. So $c(0) = 1 = c(1)$, thus it is a DCS satisfying the theorem. If $k = 3$, there are only two RCS (up to shifting): $\{x \equiv 0 \pmod 3, x \equiv 1 \pmod 3, x \equiv 2 \pmod 3\}$ and $\{x \equiv 0 \pmod 2, x \equiv 1 \pmod 4, x \equiv 3 \pmod 4\}$. Both are DCS which satisfy the theorem.

Now let $k_0 \ge 4$ and assume that the theorem holds for all $k < k_0$. Suppose, for the sake of contradiction, that there exists an RCS, $\mathcal{A} = \{a_i \pmod{m_i} : 1 \le i \le k_0\}$, which has the property that $c(x) = 1$ for $3 \cdot 2^{k_0-3}$ consecutive integers but which is not a DCS.

Clearly there is some integer x for which $c(x) > 1$. By shifting the system, we can assume without loss of generality that $c(0) > 1$ and $c(1) = c(2) = \cdots = c(3 \cdot 2^{k_0-3}) = 1$.

We shall prove that 0 cannot be covered by two congruences with distinct prime moduli. Suppose that $\mathcal{A}$ contains the congruences $x \equiv 0 \pmod p$ and $x \equiv 0 \pmod q$ where p and q are distinct primes. Clearly $c(pq) \ge 2$. If $pq \le 3 \cdot 2^{k_0-3}$ then this contradicts the fact that $c(x) = 1$ for $1 \le x \le 3 \cdot 2^{k_0-3}$. Let us assume that $pq > 3 \cdot 2^{k_0-3}$. By Simpson's inequality we see that $k_0 \ge p + q - 1$ which implies $pq > 3 \cdot 2^{p+q-4}$ which is equivalent to $16/3 > 2^p/p \cdot 2^q/q$. The function $2^x/x$ is increasing for $x \ge 2$ so the right-hand product has its smallest value, $16/3$ when $p = 2$ and $q = 3$ which contradicts the assumption that $pq > 3 \cdot 2^{k_0-3}$.

It follows that there is at least one congruence covering 0 that involves a modulus, say m, that is composite. Let p be a prime divisor of m. By Simpson's Lemma $p \le k_0$.

We now partition $\mathcal{A}$ into three collections:

$$\mathcal{A}_1 = \{x \equiv a_i \pmod{m_i} \in \mathcal{A} : p \nmid m_i\}$$
$$\mathcal{A}_2 = \{x \equiv a_i \pmod{m_i} \in \mathcal{A} : p \mid m_i \text{ and } p \mid a_i\}$$
$$\mathcal{A}_3 = \{x \equiv a_i \pmod{m_i} \in \mathcal{A} : p \mid m_i \text{ and } p \nmid a_i\}.$$

By Simpson's lemma, $|\mathcal{A}_3| \ge p - 1$. Thus, $|\mathcal{A}_1 \cup \mathcal{A}_2| \le k_0 - (p - 1)$. Now, let us form a new collection $\mathcal{A}^* = \mathcal{A}_1^* \cup \mathcal{A}_2^*$ where

$$\mathcal{A}_1^* = \{x \equiv a_i p^{-1} \pmod{m_i} : x \equiv a_i \pmod{m_i} \in \mathcal{A}_1\}$$
$$\mathcal{A}_2^* = \{x \equiv a_i/p \pmod{m_i/p} : x \equiv a_i \pmod{m_i} \in \mathcal{A}_2\}$$

and p^{-1} is the multiplicative inverse of p modulo m_i.

Consider integers pn, where n is any integer. Each of these must belong to at least one congruence in $\mathcal{A}_1 \cup \mathcal{A}_2$ so n belongs to the corresponding congruence in $\mathcal{A}^*$. In particular, 0 belongs to at least two congruences and $1, 2, \ldots, \lfloor 3 \cdot 2^{k_0-3}/p \rfloor$ each belong to exactly one. It is easily checked that $\lfloor 3 \cdot 2^{k_0-3}/p \rfloor \ge \max\{2, 3 \cdot 2^{k_0-3-(p-1)}\}$ which means that $\mathcal{A}^*$ is a counterexample to the inductive hypothesis. The theorem follows. ∎

The bound in Theorem 1 cannot be improved, as the following example demonstrates. For $k \ge 4$, consider the RCS

$$x \equiv \begin{cases} 2^{j-1} & (\text{mod } 2^j), \text{ for } 1 \le j \le k-3 \\ 2^{k-3} & (\text{mod } 3 \cdot 2^{k-3}) \\ 2^{k-2} & (\text{mod } 3 \cdot 2^{k-3}) \\ 0 & (\text{mod } 3 \cdot 2^{k-4}) \end{cases}$$

In this system $c(x) = 2$ where $x = 3 \cdot 2^{k-4} + n \cdot 3 \cdot 2^{k-3}$ for $n \in \mathbb{Z}$, and $c(x) = 1$ for other integers. So $c(x) = 1$ for $3 \cdot 2^{k-3} - 1$ consecutive integers, but the system is not disjoint.

To answer the second question, we must appeal to another fact about covering systems. In 1962 Erdős [**3**] proved that a system of congruences will cover every integer if it covers all integers in the interval $[1, k2^k]$ where k is the number of congruences. This was not improved until 1970 when R. B. Crittenden and C. L. Vanden Eynden in [**2**] proved the following:

LEMMA 3. *Let $a_1, a_2, \ldots, a_n, m_1, m_2, \ldots, m_k$ be given, with m's positive. Suppose there exists an integer x_0 satisfying none of the congruences*

$$x \equiv a_i \pmod{m_i}, \quad i = 1, 2, \ldots, k$$

Then there is such an x_0 among $1, 2, \ldots, 2^k$.

A shorter sequence will not suffice in Lemma 3, as the following example shows

$$\{x \equiv 2^{j-1} \pmod{2^j}, 1 \le j \le k\}$$

This system covers $2^k - 1$ consecutive integers exactly once, but 0 is not covered by any congruence so the system is not covering. Now the answer to Question 2 is provided by the following Corollary:

COROLLARY 1. *If a system with k congruences covers 2^k consecutive integers exactly once, then it is a DCS.*

Proof. This follows immediately from Theorem 1 and the Theorem of Crittenden and Vanden Eynden. ∎

Finally we provide an upper bound for the least common multiple of the moduli in an RCS.

THEOREM 2. *Given any regular covering system $\{x \equiv a_i \pmod{m_i} : 1 \le i \le k\}$, it follows that $\text{lcm}(m_1, \ldots, m_k) \le 2^{k-1}$.*

Proof. Let $\text{lcm}(m_1, \ldots, m_k) = \prod_{i=1}^{t} p_i^{\alpha_i}$ be the canonical prime factorization. We wish to maximize this value subject to Simpson's inequality [**6**, corollary 2], $k - 1 \ge \sum_{i=1}^{t} \alpha_i (p_i - 1)$. For every prime p we have $p \le 2^{p-1}$ therefore

$$\log_2 \prod_{i=1}^{t} p_i^{\alpha_i} = \sum_{i=1}^{t} \alpha_i \log_2 p_i \le \sum_{i=1}^{t} \alpha_i (p_i - 1) \le k - 1$$

by Simpson's inequality. The result follows. ∎

The interested reader should refer to [**5**] and [**7**] for surveys of the work done on covering systems of congruences. Both sources also contain extensive bibliographies on the topic of covering systems.

Acknowledgment Funding for research was provided by the Dr. J. Douglas Faires and Dr. Barbara T. Faires endowment and the McNair Scholar Program, under the supervision of Dr. J. Fabrykowski and Youngstown State

University CURMath. We express our sincere gratitude to our advisor, Jacek Fabrykowski, and to J. Douglas Faires for his meticulous reading of this paper.

REFERENCES

1. American Invitational Mathematics Examination, 2002, http://www.artofproblemsolving.com/Forum/resources/files/usa/45/USA-AIME-2002-45.pdf.
2. R. B. Crittenden and C. L. Vanden Eynden, Any n arithmetic progressions covering the first 2^n integers cover all integers, *Proc. Amer. Math. Soc.* **24** (1970) 475–481.
3. P. Erdős, Remarks on number theory IV (in Hungarian), *Mat. Lapok* **13** (1962) 228–255.
4. J. Fabrykowski and T. Smotzer, Covering systems of congruences, *Math. Mag.* **78** (2005) 228–231. doi:10.2307/30044163
5. Š. Porubský and J. Schönheim, Covering systems of Paul Erdős, Past, present and future, pp. 581–627 in *Paul Erdős and His Mathematics I*, János Bolyai Mathematical Society and Springer-Verlag, 2002.
6. R. J. Simpson, Regular coverings of the integers by arithmetic progressions, *Acta Arithmetica* **45** (1985) 145–152.
7. Štefan Znám, A survey of covering systems of congruences, *Acta Math. Univ. Comenian.* **40–41** (1982) 59–71.

Summary A *covering system* is a system of k arithmetic progressions whose union includes all integers. It is a *disjoint covering system* (or *exact covering system*) if the progressions are also pairwise disjoint, so that each integer is covered exactly once. This paper presents upper bounds on the number of consecutive integers which need to be checked to determine whether a covering system is a disjoint covering system. The bounds depend only on the number of congruences in the system. The results provide an analog of a theorem by R. B. Crittenden and C. L. Vanden Eynden from 1969 and are presented as solutions to some variations of a 2002 AIME Problem about painting a picket fence.

Convexity and Center of Mass

ZSOLT LENGVÁRSZKY
Louisiana State University in Shreveport
Shreveport, LA 71115
zlengvar@lsus.edu

In differential calculus, we encounter the notion of convexity as one of the main characteristics of the graph of certain functions; and in integral calculus, centers of mass (or centroids) are discussed as an important practical application. In this note we explore a link between these two concepts—a link that appears to have gone unnoticed despite its simplicity and plausibility.

A function is called "convex" (on an interval) if its graph is "concave up." We show that centers of mass preserve convexity in the following sense:

THEOREM 1. *Suppose that the functions f and g are continuous and convex on an interval $[a, b]$ and satisfy $f(x) \geq g(x)$ for all x in $[a, b]$. Then, for any partition $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, the centroids of the regions bounded by the lines $x = x_i$ and by the graphs of f and g are the vertices of a convex polygonal curve.*

From elementary geometry, we know that the centroid of a trapezoid lies on the median that connects the midpoints of the two parallel sides. Trapezoids can be treated as a special case of Theorem 1 in which both f and g are linear. In that case the centroids of the regions all fall on the line $y = \frac{1}{2}(f(x) + g(x))$, and the polygonal curve formed by them can be viewed as both convex and concave.

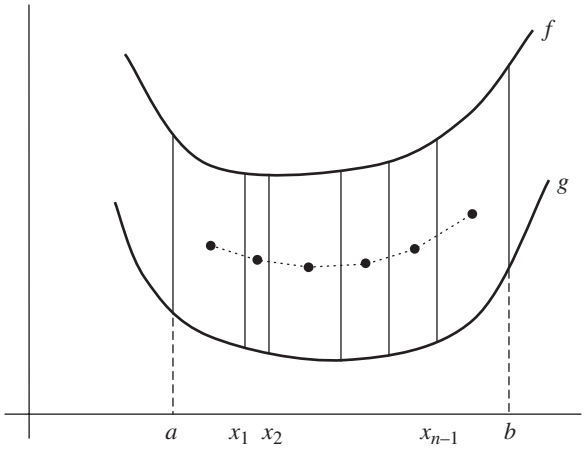


Figure 1 Convexity is preserved by centers of mass

We assume familiarity with the basic center-of-mass and centroid formulas. For example, the coordinates of the centroid of a region bounded by the lines $x = a, x = b$, and by the graphs of continuous functions $f(x) \geq g(x)$, are given by

$$\bar{x} = \frac{\int_a^b x(f(x) - g(x))\,dx}{\int_a^b (f(x) - g(x))\,dx}, \quad \text{and} \quad \bar{y} = \frac{\int_a^b \frac{1}{2}((f(x))^2 - (g(x))^2)\,dx}{\int_a^b (f(x) - g(x))\,dx}.$$

Similarly, given a curve $y = h(x)$, $a \leq x \leq b$, with mass distribution having density $\mu(x) > 0$ at $(x, h(x))$, the coordinates of the center of mass of the arc between $(a, h(a))$ and $(b, h(b))$ are given by

$$\bar{x} = \frac{\int_a^b x\mu(x)\,dx}{\int_a^b \mu(x)\,dx}, \quad \text{and} \quad \bar{y} = \frac{\int_a^b h(x)\mu(x)\,dx}{\int_a^b \mu(x)\,dx}. \tag{1}$$

These well-known formulas (together with some other more general ones) can be found in most calculus texts. Some books also mention the fact that if a region is partitioned into non-overlapping parts then the center of mass of the whole is the weighted average of the centers of mass of the parts, where the weights are the masses of the parts. In the special case when the number of parts equals two, the center of mass of the region lies on the line segment connecting the centers of mass of the two parts.

An article by Steven Krantz [**2**] provides a development of some of the fundamental properties of centroids and explores how the geometry of a planar domain affects its center of mass.

Proof and generalization

It is not very difficult to prove our theorem using an elementary argument. We will however proceed by quoting a classical result from the theory of convex sets which will do two things for us. First, it makes the proof simpler; second, it allows us to make a more general statement. Later in this section, we will use more familiar calculus techniques to provide an independent proof of Theorem 5, which implies Theorem 3, which implies Theorem 1.

The following can be found on page 9 in the English translation [**1**] of T. Bonnesen and W. Fenchel's book that was first published in 1937:

THEOREM 2. *The convex hull of a closed and bounded set S is exactly the set of all centers of mass of mass distributions over S.*

The *convex hull* of a set S is the minimal convex set containing S. If S is convex, then its convex hull is itself. If S is a piece of a convex curve with endpoints A and B, then the convex hull is the closed region bounded by S and by the line segment connecting A and B.

We make the connection to our convex functions by considering $h(x) = \frac{1}{2}(f(x) + g(x))$. It is easily seen that h is a convex function. Now assign a mass distribution on the curve $y = h(x)$, giving to each point $(x, h(x))$ the mass density $\mu(x) = f(x) - g(x)$. We think of the convex curve $y = h(x)$ as a nonhomogeneous material wire whose mass density at each point $(x, h(x))$ is $\mu(x)$. Note that the centroid of the region between the vertical lines $x = x_{i-1}$ and $x = x_i$ and bounded above by f and below by g is the center of mass of the piece of wire between $x = x_{i-1}$ and $x = x_i$. But according to Theorem 2 this center of mass belongs to the convex hull of the arc between $x = x_{i-1}$ and $x = x_i$; see FIGURE 2.

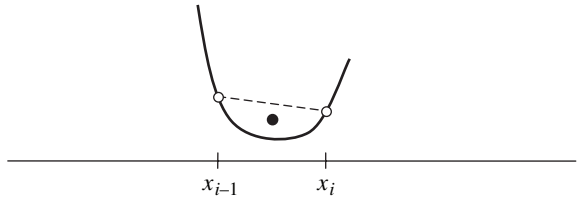


Figure 2 The center of mass is in the convex hull.

The proof of Theorem 1 is obvious now. Because of the convexity of h, the centers of mass of the arcs of wire from P_{i-1} to P_i and from P_{i+1} to P_{i+2} both lie above the line $\overleftrightarrow{P_i P_{i+1}}$, and the center of mass of the arc of wire from P_i to P_{i+1} lies below the line $\overleftrightarrow{P_i P_{i+1}}$.

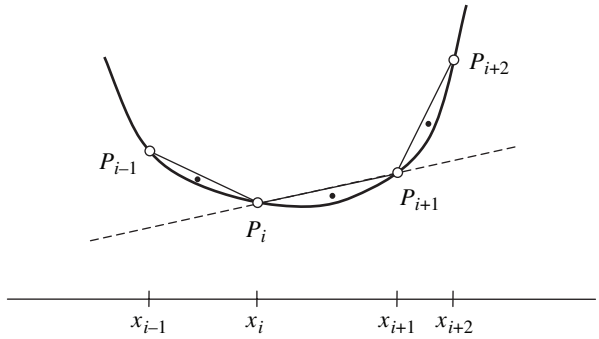


Figure 3 Proof of Theorem 1

It now becomes clear that Theorem 1 is a special case of the following:

THEOREM 3. *Let $y = h(x)$ be any convex function and let $\mu(x)$ be any positive mass distribution on the graph with density $\mu(x)$ at $(x, h(x))$. Then, if the wire is partitioned into subarcs, the centers of mass of the pieces are vertices of a convex polygon.*

Next we consider a continuous version of Theorem 1.

THEOREM 4. *Suppose that the functions f and g are continuous and convex on an interval $[a, b]$ and satisfy $f(x) \geq g(x)$ for all x in $[a, b]$. Let $C(t) = (\bar{x}(t), \bar{y}(t))$ be the center of mass of the region bounded by the lines $x = a$, $x = t$, and the graphs of f and g. Then the curve $C(t)$, $a \leq t \leq b$, is convex as well.*

This result can also be treated as a special case of a more general theorem using loaded wires:

THEOREM 5. *Let $y = h(x)$ be any convex function and let $\mu(x)$ be any positive mass distribution on the graph with density $\mu(x)$ at $(x, h(x))$. Let $C(t) = (\bar{x}(t), \bar{y}(t))$ be the center of mass of the piece of wire between $x = a$ and $x = t$. Then the curve $C(t)$, $a \leq t \leq b$, is convex.*

Proof. Let $a = t_0 < t_1 < t_2 < t_3 \leq b$. We need to show that $C(t_2)$ lies below the line segment $C(t_1)C(t_3)$. Define W_i for $i = 1, 2$, and 3 to be the pieces of wire between $x = t_{i-1}$, and $x = t_i$, and let C_i be the center of mass of W_i (see FIGURE 4). Note that $C_1 = C(t_1)$. By Theorem 3, C_2 lies below the line segment C_1C_3. Since $C(t_2)$ lies on the line segment C_1C_2, and $C(t_3)$ lies on the line segment $C(t_2)C_3$, it follows that $C(t_2)$ is below $C(t_1)C(t_3)$. ∎

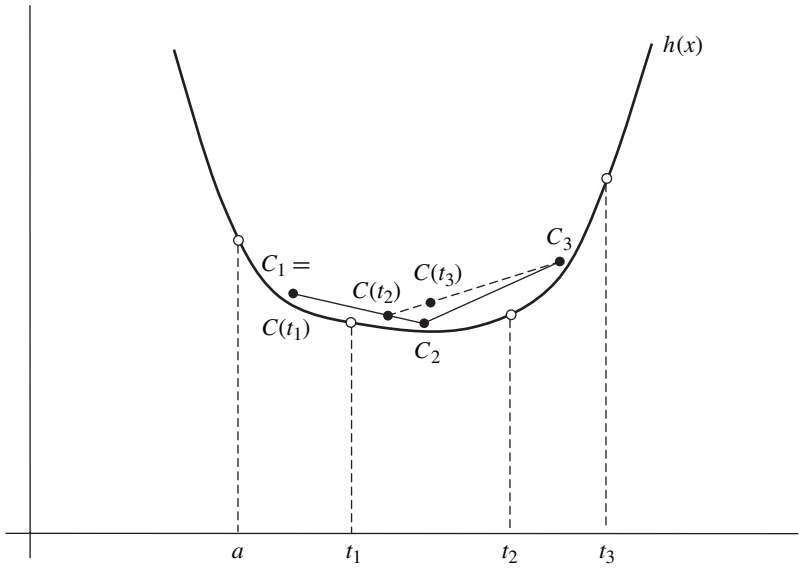


Figure 4 $C(t_1)C(t_2) \cup C(t_2)C(t_3)$ is the graph of a convex function.

Note that Theorem 3 could be used to prove Theorem 5; hence, the two statements are equivalent. In addition, Theorem 5 can be proved using an analytical argument; in this case however, we assume that h is twice differentiable having nonnegative second derivative (or at least once differentiable with increasing first derivative). Viewing $\bar{y}$ as a function of $\bar{x}$, and assuming $h'(x)$ is increasing on $[a, b]$, we need to show $d^2\bar{y}/d\bar{x}^2 \geq 0$ for all $\bar{x}$, that is, for all t in (a, b). Changing the upper limit b to t in formula 1, and taking the derivative with respect to t we have

$$\frac{d\bar{x}}{dt} = \frac{t\mu(t)\left(\int_a^t \mu(x)\,dx\right) - \mu(t)\int_a^t x\mu(x)\,dx}{\left(\int_a^t \mu(x)\,dx\right)^2} = \frac{\mu(t)\left(\int_a^t (t-x)\mu(x)\,dx\right)}{\left(\int_a^t \mu(x)\,dx\right)^2}$$

and a similar expression for $d\bar{y}/dt$. The parametric differentiation formula

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} \tag{2}$$

now gives

$$\frac{d\bar{y}}{d\bar{x}} = \frac{\int_a^t ((h(t) - h(x))\,\mu(x)\,dx}{\int_a^t (t-x)\mu(x)\,dx}. \tag{3}$$

A further application of (2) to (3) then yields

$$\frac{d^2\bar{y}}{d\bar{x}^2} = \frac{\left(\int_a^t \mu(x)\,dx\right)^3 \int_a^t (h'(t)(t-x) - ((h(t) - h(x)))\,\mu(x)\,dx}{\int_a^t \mu(t)\left(\int_a^t (t-x)\mu(x)\,dx\right)^3}.$$

Observe that the differences and integrands above are all nonnegative. The only nontrivial case is the first factor of the second integrand in the numerator which can be handled by applying the Mean Value Theorem: There is a number $x < \tau < t$ such that

$$h'(\tau)(t-x) = h(t) - h(x),$$

and since the first derivative $h'(x)$ is increasing, the left-hand side becomes larger when we replace τ by t. It follows that the integrand in question is also nonnegative.

We note that treating $\bar{y}$ as a function of $\bar{x}$ is indeed valid, since under the assumptions, $\bar{x}$ is a strictly increasing, hence invertible function of t.

A variant of Theorem 5 can be obtained by letting both of the endpoints of a subinterval of $[a, b]$ vary, instead of just one.

THEOREM 6. *Let $y = h(x)$ be any convex function and let $\mu(x)$ be any continuous positive mass distribution on the graph with density $\mu(x)$ at $(x, h(x))$. Then the locus of the centers of mass of subarcs with a fixed mass is itself a convex curve.*

This is an immediate consequence of a result given on page 11 in Bonnesen and Fenchel [**1**]; their theorem applies to sets in spaces of any dimension:

THEOREM 7. *If one cuts subbodies of fixed total mass m from a body with total mass M by hyperplanes in all possible ways, then the centroids of these subbodies form a convex surface.*

A corollary of Theorem 6 then is

THEOREM 8. *Given convex functions $f(x) \geq g(x)$, $a \leq x \leq b$, consider vertical lines ℓ_1 and ℓ_2 such that the region between ℓ_1 and ℓ_2 and below f and above g, has a given fixed area. Then the locus of the centroids of all such regions is itself a convex curve.*

It is worth noting that, in either Theorem 6 or Theorem 8, we cannot replace the fixed-mass or fixed-area requirement by a fixed-distance or "moving-window" condition. That is, if one only assumes that the distance between ℓ_1 and ℓ_2 be fixed, the locus of centroids is no longer guaranteed to be convex. The reader is encouraged to find counterexamples.

Derivatives of other orders

Concavity is associated with the sign of the second derivative. Can our Theorems be extended to other derivatives?

Consider the case of first derivatives. Assuming $h'(x) \geq 0$, thus, h is increasing, is it true that the polygonal curve defined by the centers of mass of pieces of wires is increasing as well? The answer is "yes," and the "picture-proof" of the result analogous to Theorem 3 is given by FIGURE 5. It is sufficient to consider two pieces only: W_1, the piece between $x = x_1$, and $x = x_2$, and W_2, the piece between $x = x_2$, and $x = x_3$. Drawing horizontal and vertical lines across P_1, P_2, and P_3, we see W_1 is entirely inside the (convex) rectangle $P_1Q_2P_2Q_1$, and W_2 is entirely inside the (convex) rectangle $P_2Q_3P_3Q_2'$. It follows that C_1, the center of mass of W_1, is below, and C_2, the center of mass of W_2, is above the horizontal line across P_2.

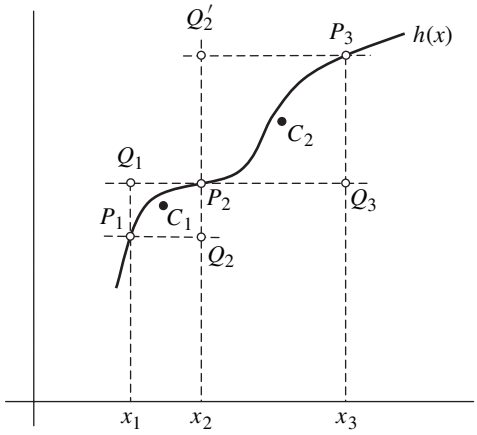


Figure 5 Centers of mass C_1 and C_2

For a calculus proof of the first derivative analogue of Theorem 5 just observe that (3) is nonnegative.

The case of the 0th derivatives is trivial: the assumption $h(x) \geq 0$ means that the entire curve is above the x-axis and the same will be true for the centers of mass.

We might then conjecture that the sign of the nth derivative in general is preserved by centers of mass. However, the statement breaks down with $n = 3$, as witnessed by the following counterexample.

Let $f(x) = 1 + x(x-1)(x-2)$ and $g(x) = 0$ (or $h(x) = \mu(x) = 1 + x(x-1)(x-2)$), and let $[a, b] = [0, 3]$. Then we have $f(x) \geq g(x)$ on $[a, b]$, and $f'''(x) = 6 > 0$ for all x. Yet, for

$$\begin{aligned}\frac{d^3\bar{y}}{d\bar{x}^3} &= 10125(4 + 4t - 4t^2 + t^3)^4 \\ &\quad \cdot (21000 + 6720t + 55580t^2 - 127460t^3 + 39268t^4 + 74604t^5 \\ &\quad - 55029t^6 - 8122t^7 + 22460t^8 - 10662t^9 + 2544t^{10} - 324t^{11} + 18t^{12}) \\ &\quad \big/(224(1 + 2t - 3t^2 + t^3)^3(30 + 20t - 15t^2 + 3t^3)^5),\end{aligned}$$

we find

$$\left.\frac{d^3\bar{y}}{d\bar{x}^3}\right|_{t=2} = -\frac{6905250}{9938999} < 0.$$

REFERENCES

1. T. Bonnesen and W. Fenchel, *Theory of Convex Bodies*, BCS Associates, 1987.
2. S. Krantz, A matter of gravity, *Amer. Math. Monthly* **110** (2003) 465–481. doi:10.2307/3647903

Summary We show that convexity is preserved by centers of mass by considering regions between two convex functions both in a discrete and continuous sense. The proofs are partly based on theorems in the classical literature, however, elementary arguments are given as well. An example is presented to demonstrate that while the results are also true for 0th and 1st derivatives (convexity being tied to second derivatives), a simple generalization to higher derivatives is not possible.

Eigenvalues in Filled Julia Sets

JONATHON E. FASSETT
Central Washington University
Ellensburg, WA 98926-7424
fassett@cwu.edu

The study of Julia sets, the Mandelbrot set, and complex dynamics in general is an active area of research that students find fascinating. Exposing our students to these contemporary topics as early as possible is a worthwhile endeavor. Many wonderful materials have already been developed along these lines. Among my favorite resources are the books by Devaney [**4, 5**] and Peitgen et al. [**9**]. In this paper, we see how Julia sets can be introduced very naturally in a junior-level linear algebra course. We do not discuss the complicated (actually chaotic) dynamics or the fractal geometry involving Julia sets, but limit ourselves to using standard topics from linear algebra to prove that Julia sets are lurking in the background disguised as eigenvalues. In fact, with the exception of the final theorem (whose proof uses Cauchy's integral formula) our discussion does not rely upon any mathematics beyond the topics typically covered in a junior-level linear algebra course.

What are Julia sets?

The study of the dynamics of complex analytic functions began during the First World War. At this time, Gaston Julia (1893–1978) and Pierre Fatou (1878–1929) independently published many remarkable results that laid the groundwork for the field. Julia sets and Fatou sets are named in their honor. Their results were mostly forgotten until Benoit Mandelbrot in the late 1970s produced the first of the computer images that have now become so popular.

Let $p : \mathbb{C} \to \mathbb{C}$ be a polynomial and define

$$p^n(z) = \underbrace{(p \circ \cdots \circ p)}_{n \text{ times}}(z).$$

The *basin (of attraction)* of ∞ is defined as

$$A_p(\infty) = \{z \in \mathbb{C} : |p^n(z)| \to \infty\}.$$

Math. Mag. **84** (2011) 221–227. doi:10.4169/math.mag.84.3.221.

The *Julia set* of p, denoted by J_p, is defined as the boundary of the basin of ∞ and the *filled Julia set of* p is defined as

$$K_p = \{z \in \mathbb{C} : |p^n(z)| \not\to \infty\} = \mathbb{C} \setminus A(\infty).$$

The set $A(\infty)$ is always an open set, so the Julia set J_p is always contained in the filled Julia set K_p.

Although our discussion focuses on the filled Julia set, we should mention that the *Fatou set* is defined as the complement of the Julia set.

EXAMPLE 1. Let $p(z) = z^2$. Then for $z = re^{i\theta}$, we have $p^n(z) = r^{2^n} e^{2n\theta i}$ and $|p^n(z)| = r^{2^n}$. Therefore, $|p^n(z)| \to \infty$ if and only if $|z| = r > 1$. It follows that

$$A_p(\infty) = \{z \in \mathbb{C} : |z| > 1\},$$
$$J_p = \{z \in \mathbb{C} : |z| = 1\},$$

and

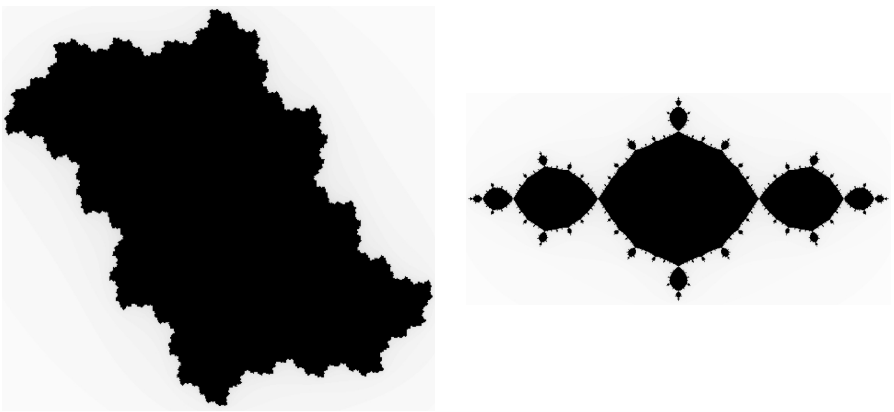
$$K_p = \{z \in \mathbb{C} : |z| \leq 1\}.$$

A similar argument shows that the Julia set for $p(z) = z^m$ with $m \geq 2$ is the unit circle.

Most Julia sets are more interesting than a circle and analytically describing the set as we did in the previous example is often too difficult. However, several algorithms for visualizing Julia sets have become very popular [**4, 9**]. We note that the images in this article were produced using the algorithm found in Devaney's book [**4**, p. 233].

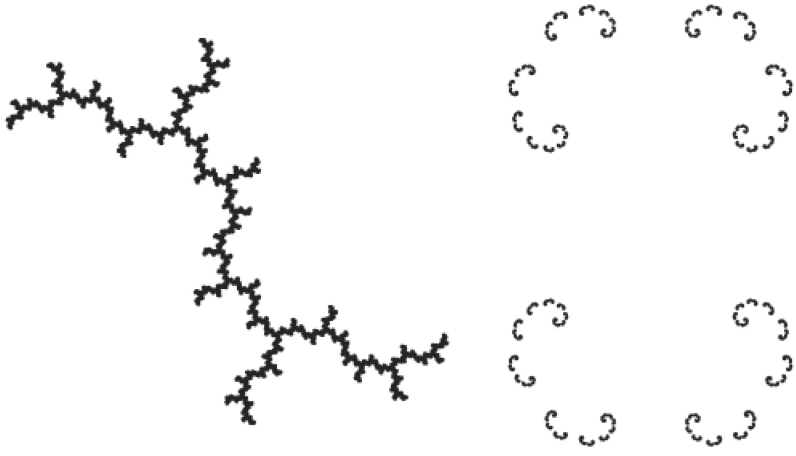
FIGURES 1 and 2 show the filled Julia set for selected quadratic functions of the form $p(z) = z^2 + c$ with $c \in \mathbb{C}$. In each of the examples in FIGURE 1, the filled Julia set has positive area. In each of the examples in FIGURE 2 the Julia set is equal to the filled Julia set, so the filled Julia set has empty interior and zero area. The first example in FIGURE 2 is called a *dendrite*, and the second example is a *Cantor set.* (The figures appear to have more substance than they do because dots are printed with finite size.)

Much could be said about Julia sets (and it is fun to learn). For now we are content to see how Julia sets appear in the setting of linear algebra.



$p(z) = z^2 + .5i$ $\qquad$ $p(z) = z^2 - 1$

Figure 1 Two filled Julia sets that have positive area



$p(z) = z^2 + i$ $p(z) = z^2 + .5$

Figure 2 Two Julia sets that have zero area

Linear algebra basics

Before beginning our hunt for Julia sets we collect some necessary tools. Let $\mathbb{C}^{k\times k}$ be the space of all $k \times k$ matrices over $\mathbb{C}$. For $Z \in \mathbb{C}^{k\times k}$, we define the *spectrum* of Z as

$$\sigma(Z) = \{\lambda \in \mathbb{C} : \lambda \text{ is an eigenvalue of } Z\}.$$

The *spectral radius* of Z is defined by

$$\rho(Z) = \max\{|\lambda| : \lambda \in \sigma(Z)\}.$$

For a matrix norm on $\mathbb{C}^{k\times k}$ we use the *Frobenius norm* defined by

$$\|Z\| = \left(\sum_{j=1}^{k}\sum_{i=1}^{k} |z_{i,j}|^2\right)^{1/2}.$$

By calling it a *matrix norm*, we mean that in addition to the properties satisfied by general norms, this norm satisfies the identity $\|kA\| = k\,\|A\|$ and has the sub-multiplicative property

$$\|ZW\| \le \|Z\|\,\|W\| \text{ for all } Z, W \in \mathbb{C}^{k\times k}. \tag{1}$$

We recommend Leon's book [**8**] as a good introduction to matrix norms.

Note that if λ is an eigenvalue for Z with corresponding eigenvector $\mathbf{x}$ then, with $X \in \mathbb{C}^{k\times k}$ having all columns equal to $\mathbf{x}$, we have $ZX = \lambda X$ and

$$|\lambda|\,\|X\| = \|\lambda X\| = \|ZX\| \le \|Z\|\,\|X\|.$$

Dividing through by $\|X\|$ (how do we know $\|X\| \neq 0$?) we have $|\lambda| \le \|Z\|$ and we get the following nice bound on the spectral radius.

$$\rho(Z) \le \|Z\|. \tag{2}$$

Julia sets and eigenvalues

We are now ready to find out where Julia sets are hiding. Suppose $p(z) = a_m z^m + a_{m-1}z^{m-1} + \cdots + a_1 z + a_0$ is a polynomial with complex coefficients. If $Z \in \mathbb{C}^{k\times k}$ then it makes sense to define

$$p(Z) = a_m Z^m + a_{m-1}Z^{m-1} + \cdots + a_1 Z + a_0 I$$

where I is the $k \times k$ identity matrix. It is also very natural to consider the set

$$\mathcal{K}_p = \{Z \in \mathbb{C}^{k\times k} : \|p^n(Z)\| \not\to \infty\}$$

which is analogous to the filled Julia set of p. In fact, if $k = 1$ then $\mathcal{K}_p = K_p$. Hence, $\mathcal{K}_p$ generalizes the filled Julia set from $\mathbb{C}$ to $\mathbb{C}^{k\times k}$. If we need to emphasize the dependence on k, which is usually clear from context, we will write $\mathcal{K}_p(k)$.

THEOREM 1. *If $Z \in \mathcal{K}_p$ then $\sigma(Z) \subseteq K_p$. That is, if Z is bounded under iteration by p, then the eigenvalues of Z are contained in the filled Julia set of p.*

Proof. Suppose $Z \in \mathcal{K}_p$ and $\lambda \in \sigma(Z)$. A well known result called the *spectral mapping theorem* (which is not hard to prove) is that $p^n(\lambda) \in \sigma(p^n(Z))$. Using (2) we have

$$|p^n(\lambda)| \le \rho(p^n(Z)) \le \|p^n(Z)\| \not\to \infty.$$

It follows that $\lambda \in K_p$. Therefore $\sigma(Z) \subseteq K_p$. ∎

The converse of Theorem 1 is not true as the following example shows.

EXAMPLE 2. Let $p(z) = z^2$ and

$$Z = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

Clearly (why is this clear?) -1 is a repeated eigenvalue for Z. Since $p^n(-1) = 1$ for all $n \ge 1$, we conclude $\sigma(Z) \subseteq K_p$. However, direct computation shows

$$p^n(Z) = \begin{bmatrix} 1 & -2^n \\ 0 & 1 \end{bmatrix}$$

and $\|p^n(Z)\| \ge 2^n \to \infty$. Therefore, $Z \notin \mathcal{K}_p$.

In this example note that the eigenvalues for Z lie on the boundary of K_p (which is equal to the Julia set J) and that Z is not diagonalizable. However, if Z is diagonalizable then $\sigma(Z) \subseteq K_p$ does imply that $Z \in \mathcal{K}_p$ regardless of whether or not any of the eigenvalues lie on the boundary of K_p. The argument is as follows. Suppose that $Z \in \mathbb{C}^{k\times k}$ is diagonalizable and $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the eigenvalues of Z (repetitions allowed). Then there exists a nonsingular matrix $S \in \mathbb{C}^{k\times k}$ such that $Z = SDS^{-1}$ where

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & & 0 \\ 0 & \lambda_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \\ & & & \ddots & 0 \\ 0 & \cdots & & 0 & \lambda_k \end{bmatrix} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k).$$

In addition, $p^n(Z) = SD^nS^{-1}$ where $D^n = \operatorname{diag}(p^n(\lambda_1), p^n(\lambda_2), \ldots, p^n(\lambda_k))$. Now

$$\|D^n\| = \left(\sum_{i=1}^{k} |p^n(\lambda_i)|^2\right)^{1/2} \leq \sqrt{k} \cdot \rho(p^n(Z))$$

and by (1)

$$\|p^n(Z)\| \leq \|S\| \, \|D^n\| \, \left\|S^{-1}\right\| \leq \|S\| \, \left\|S^{-1}\right\| \sqrt{k} \cdot \rho(p^n(Z)). \tag{3}$$

Combining (2) and (3) we obtain

$$\rho(p^n(Z)) \leq \|p^n(Z)\| \leq \|S\| \, \left\|S^{-1}\right\| \sqrt{k} \cdot \rho(p^n(Z)).$$

Thus, $\|p^n(Z)\| \not\to \infty$ if and only if $\rho(p^n(Z)) \not\to \infty$. We now have a partial converse to Theorem 1.

THEOREM 2. *Suppose Z is diagonalizable. If $\sigma(Z) \subseteq K_p$ then $Z \in \mathcal{K}_p$. That is, if the eigenvalues of Z are contained in the filled Julia set of p, then Z is bounded under iteration by p.*

EXAMPLE 3. Let $p(z) = z^3 - (.5 + .1i)$ and

$$Z = \begin{bmatrix} 0 & -1 & -1 \\ 0 & 1 & 1-i \\ 0 & 0 & -i \end{bmatrix}.$$

The eigenvalues of Z are $\lambda_1 = 0$, $\lambda_2 = 1$, and $\lambda_3 = i$ with corresponding eigenvectors $\mathbf{x}_1 = [1, 0, 0]^T$, $\mathbf{x}_2 = [-1, 1, 0]^T$, and $\mathbf{x}_3 = [0, -1, 1]^T$. It follows that Z is similar to the diagonal matrix

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -i \end{bmatrix}.$$

It is easy to show that the eigenvalues of Z are contained in the filled Julia set of p. Thus, Z is bounded under iteration by p. The filled Julia set of p along with the eigenvalues of Z (highlighted) are shown in FIGURE 3. It is interesting to note that the filled Julia set of p is not convex. For example, both 0 and i are in the filled Julia set but $-.6i$, which is on the line segment connecting 0 and i, is no longer in the filled Julia set. In fact, if we replace the diagonal entry $-i$ with $-.6i$ in the matrix Z, then Z is still diagonalizable, $-.6i$ is not in the filled Julia set of p, and Z is no longer bounded under iteration by p. The number $-.6i$ is also highlighted in FIGURE 3.

The situation is more complicated if $Z \in \mathbb{C}^{k\times k}$ is not diagonalizable. As we saw in Example 2, it is possible for $\sigma(Z) \subseteq K_p$ and $Z \notin \mathcal{K}_p$. This will not happen however if the eigenvalues of Z are contained in the *interior* of K_p (which we denote by K_p°). Proof of this result requires *Cauchy's integral formula* for derivatives. But first we consider the *Jordan canonical form* of Z. Suppose $S^{-1}ZS = \operatorname{diag}(J_1, J_2, \ldots, J_r)$ is the *Jordan canonical form* of Z with

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \cdots & 0 \\ 0 & \lambda_i & 1 & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & \lambda_i \end{bmatrix}$$

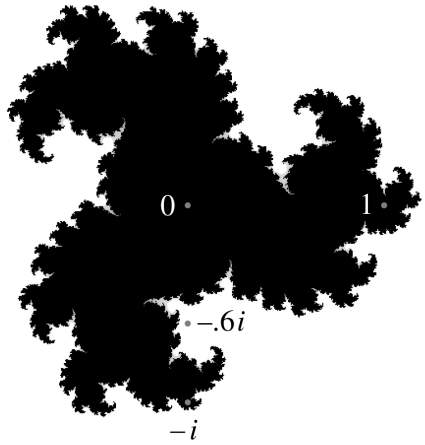


Figure 3 The filled Julia set for $p(z) = z^3 - (.5 + .1\,i)$

being an $m_i \times m_i$ *Jordan block where* $m_1 + \cdots + m_r = k$. Note that if Z is diagonalizable then each Jordan block is 1×1.

As is shown in [**6**, Theorem 11.1.1]

$$p^n(Z) = S\,\mathrm{diag}(p^n(J_1),\, p^n(J_2), \ldots,\, p^n(J_r))\; S^{-1} \tag{4}$$

where

$$p^n(J_i) = \begin{bmatrix} p^n(\lambda_i) & (p^n)^{(1)}(\lambda_i) & & \cdots & \frac{(p^n)^{(m_i-1)}(\lambda_i)}{(m_i-1)!} \\ 0 & \ddots & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & (p^n)^{(1)}(\lambda_i) \\ 0 & \cdots & & 0 & p^n(\lambda_i) \end{bmatrix}. \tag{5}$$

We can see from (4) that $Z \in \mathcal{K}_p$ if and only if $p^n(J_i) \in \mathcal{K}_p$ for each i. (Technically, we should write $Z \in \mathcal{K}_p$ and $p^n(J_i) \in \mathcal{K}_p(i)$ as, from context, $\mathcal{K}_p = \mathcal{K}_p(k)$ contains $k \times k$ matrices and $\mathcal{K}_p(i)$ contains $m_i \times m_i$ matrices.) From (5), we see that determining whether $p^n(J_i) \in \mathcal{K}_p$ requires that we show not only that the sequence $\{p^n(\lambda_i)\}$ is bounded as $n \to \infty$, but also that each of the sequences

$$\left\{(p^n)^{(1)}\,(\lambda_i)\right\}, \ldots, \left\{\frac{(p^n)^{(m_i-1)}\,(\lambda_i)}{(m_i-1)!}\right\}$$

is bounded as $n \to \infty$. One approach would be to consider the more general (and traditional) definition of Julia sets using *normal families*. However, this would lead us into a more advanced discussion involving *compact sets* and the *Arzelà-Ascoli theorem* [**3**]. We instead use the concept of a *locally bounded family*. The family of iterates $\{p^n\}$ is said to be a *locally bounded family* at w if there exist $\delta > 0$ and $M > 0$ such that $|p^n(z)| \le M$ for all z with $|z - w| < \delta$ and for all $n \ge 0$. We require the following result which is stated without proof [**3**, §III.4].

PROPOSITION 1. *If* $w \in K_p^\circ$ *then* $\{p^n\}$ *is a locally bounded family at* w.

We are now ready to prove our final result, which can be considered a generalization of Theorem 2 and a weak converse of Theorem 1.

THEOREM 3. *If $\sigma(Z) \subseteq K_p^\circ$ then $Z \in \mathcal{K}_p$. That is, if the eigenvalues of Z are contained in the interior of the filled Julia set, then Z is bounded under iteration by p.*

Proof. Suppose $\sigma(Z) \subseteq K_p^\circ$. It suffices to show $p^n(J_i) \in \mathcal{K}_p$, where J_i is a corresponding Jordan block with eigenvalue λ_i. Since $\lambda_i \in \sigma(Z)$, by Proposition 1 $\{p^n\}$ is a locally bounded family at λ_i. Therefore, there exist $\delta > 0$ and $M > 0$ such that $|p^n(z)| \leq M$ for all z with $|z - \lambda_i| < \delta$. Recall the Cauchy integral formula for the derivative as applied to p^n,

$$(p^n)^{(1)}(w) = \frac{1}{2\pi i} \int_C \frac{p^n(\xi) d\xi}{(\xi - w)^2}$$

where $C = \{z : |z - \lambda_i| = r\}$ with $r < \delta$. Hence, for all $w \in \{z : |z - \lambda_i| \leq \frac{r}{2}\}$,

$$\left|(p^n)^{(1)}(w)\right| \leq \frac{1}{2\pi} \int_C \frac{|p^n(\xi)|\, d\xi}{\left|(\xi - w)^2\right|} \leq \frac{1}{2\pi} \int_C \frac{4M d\xi}{r^2} \leq \frac{4M}{r}.$$

Therefore, $\{(p^n)^{(1)}\}$ is a locally bounded family at λ_i. In particular, the sequence $\{(p^n)^{(1)}(\lambda_i)\}$ is bounded. An induction argument can be used to conclude that for all $j \geq 1$ the sequence $\{(p^n)^{(j)}(\lambda_i)\}$ is bounded. It follows that $\|p^n(J_i)\| \not\to \infty$. ∎

It is the dynamical properties of the Julia set that require us to limit the eigenvalues to the interior of the filled Julia set in the nondiagonalizable case. In particular, if the eigenvalues are in the interior of the filled Julia set then their behavior under iteration is much more predictable than if they were on the boundary. There are many more comparisons between eigenvalues, Julia sets, and dynamical systems in general that the reader is invited to explore.

REFERENCES

1. L. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, 1979.
2. A. Beardon, *Iteration of Rational Functions*, Springer-Verlag, New York, 1991.
3. L. Carleson and T. Gamelin, *Complex Dynamics*, Springer-Verlag, New York, 1993.
4. R. Devaney, *A First Course in Chaotic Dynamical Systems: Theory and Experiment*, Westview Press, 1992.
5. R. Devaney, *The Mandelbrot and Julia Sets: A Tool Kit of Dynamic Activities*, Key Curriculum Press, 2000.
6. G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., John Hopkins University Press, 1989.
7. R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
8. S. Leon, *Linear Algebra: With Applications*, 7th ed., Pearson Prentice Hall, 2006.
9. H.-O. Peitgen, H. Jürgens, D. Saupe, and C. Hösselbarth, *Fractals for the Classroom*, Springer-Verlag, New York, 1991.

Summary We show how Julia sets can be introduced very naturally in a junior-level linear algebra course, as a way of exposing students to the contemporary area of complex dynamics. The standard definition of the filled Julia set of a polynomial is generalized to the setting of polynomial iteration of matrices. We prove that the eigenvalues of any matrix bounded under iteration by a polynomial must lie in the corresponding filled Julia set. A partial converse is obtained if the matrix is assumed to be diagonalizable. Still another partial converse is proven by assuming the spectrum of the matrix is contained in the interior of corresponding filled Julia set.

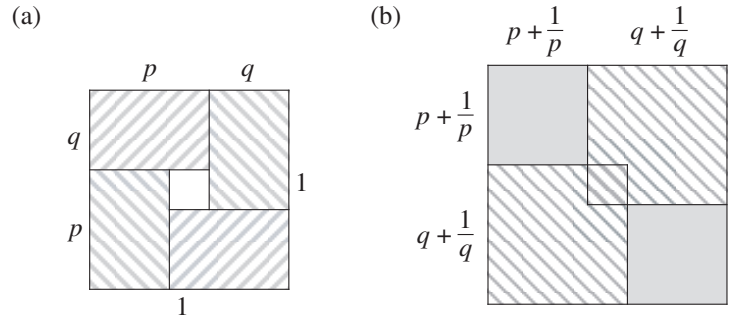
Proof Without Words: Inequalities for Two Numbers Whose Sum Is One

CLAUDI ALSINA
Universitat Politècnica de Catalunya

ROGER B. NELSEN
Lewis & Clark College

$$p, q > 0,\ p + q = 1 \implies \frac{1}{p} + \frac{1}{q} \geq 4 \quad \text{and} \quad \left(p + \frac{1}{p}\right)^2 + \left(q + \frac{1}{q}\right)^2 \geq \frac{25}{2}$$

Proof.



(a) $1 \geq 4pq \implies \dfrac{1}{p} + \dfrac{1}{q} \geq 4$,

(b) $2\left(p + \dfrac{1}{p}\right)^2 + 2\left(q + \dfrac{1}{q}\right)^2 \geq \left(p + \dfrac{1}{p} + q + \dfrac{1}{q}\right)^2 \geq (1 + 4)^2 = 25.$ ∎

REFERENCE

1. Problema 20/3(c), XX Olympiada Matemática Española (1983–84), *Olympiada Matemática Española en el Siglo XX,* Real Sociedad Matemática Española, Madrid, 2000.

PROBLEMS

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

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

PROPOSALS

To be considered for publication, solutions should be received by November 1, 2011.

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

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

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

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

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

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

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

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

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

Math. Mag. **84** (2011) 229–237. doi:10.4169/math.mag.84.3.229.

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. Submitted problems should not be under consideration for publication elsewhere.

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

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

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

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

1875. *Proposed by Éric Pité, Paris, France.*

Let $a_0 = a_1 = 1$ and for $n \geq 2$ define a_n as the number of permutations σ of $\{1, 2, \ldots, n\}$ such that

$$\sigma(1) < \sigma(2), \sigma(3) < \sigma(4), \ldots, \sigma(2j-1) < \sigma(2j) \text{ with } j = \lfloor n/2 \rfloor, \text{ and}$$

$$\sigma(2) > \sigma(3), \sigma(4) > \sigma(5), \ldots, \sigma(2k) > \sigma(2k+1) \text{ with } k = \lfloor (n-1)/2 \rfloor.$$

Prove that for every $z \in \mathbb{C}$ such that $|z| < 1$,

$$\sum_{n=0}^{\infty} \frac{a_n}{n!} z^n = \frac{\sin z + 1}{\cos z}.$$

Quickies

Answers to the Quickies are on page 236.

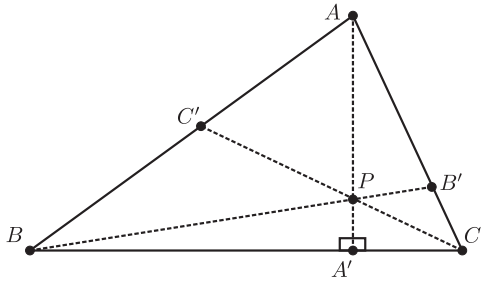
Q1011. *Proposed by Matthew McMullen, Otterbein University, Westerville, OH.*

Find the value of

$$\lim_{k\to\infty} \sum_{n=1}^{\infty} \frac{\sqrt{k}}{(n+k)\sqrt{n} + n\sqrt{n+k}}.$$

Q1012. *Proposed by Tahani Fraiwan and Mowaffaq Hajja, Mathematics Department, Yarmouk University, Irbid, Jordan.*

Let ABC be a triangle such that its altitude $\overline{AA'}$ by A lies inside the triangle. Prove that there exists a point P on the segment AA' such that the cevians $\overline{BB'}$ and $\overline{CC'}$ through P have the property that $AB' = AC'$.



Solutions

Permutations covering all possible segment sums **June 2010**

1846. *Proposed by Eddie Cheng and Jerrold W. Grossman, Department of Mathematics and Statistics, Oakland University, Rochester, MI.*

For which $n \geq 1$ is it possible to place the numbers $1, 2, \ldots, n$ in some order (a) on a line segment, or (b) on a circle, so that for every s from 1 to $\frac{1}{2}n(n+1)$ there is a connected subset of the segment or circle such that the sum of the numbers on that subset is s?

Solution by Peter McPolin, St. Mary's University College, Northern Ireland.

(a) It is only possible for $n = 1, 2, 3$. It is clear for $n = 1$ or $n = 2$ and the order $(1, 3, 2)$ works for $n = 3$. We now show that if $n \geq 4$, any ordering σ of the numbers $1, 2, \ldots, n$ on a line segment fails to have the stated property. Let $T = \frac{1}{2}n(n+1)$ so that $T \geq 10$. It is clear that if a connected subset of σ is to sum to $T - 1$ then 1 must occur either at the start or at the end of σ. Similarly, in order to have sums equal to $T - 2$ and $T - 4$ the ordering σ must have one of the forms $(13 \ldots 2)$ or $(2 \ldots 31)$, where the number 4 occurs between 2 and 3. However it is now impossible to find a connected subset with sum $T - 5$ because the complement of such a subset is either $\{5\}$, $\{2, 3\}$, or $\{1, 4\}$ and these cannot occur as complements of connected subsets of the orderings $(13 \ldots 2)$ or $(2 \ldots 31)$.

(b) It is possible for all $n \geq 1$ on a circle. We prove this by giving explicit arrangements for n even and for n odd.

When n is even, pair the numbers $1, 2, \ldots, n$ into $\frac{n}{2}$ pairs with the numbers in each pair adding to $n + 1$. Any circular arrangement of the numbers $1, 2, \ldots, n$ that keeps the numbers in each pair adjacent to one another gives all possible sums. To see this consider any value s such that $1 \leq s \leq \frac{1}{2}n(n+1)$. By the division algorithm, there are integers q and r such that $s = q(n+1) + r$, where $0 \leq r \leq n$ and $0 \leq q \leq n/2$. If $r = 0$, then we simply choose any q consecutive pairs as their sum is $s = q(n+1)$. If $r > 0$ (and hence $q < n/2$), we choose the connected subset beginning with r and having q consecutive pairs, reading clockwise or counterclockwise as appropriate, thus obtaining the sum $s = q(n+1) + r$.

If n is odd, then we form $\frac{1}{2}(n+1)$ number pairs each with sum n, where we think of the number n as a "degenerate pair." Any arrangement of $1, 2, \ldots, n$ that keeps the numbers in each pair adjacent to one another gives all possible sums. The justification for this is similar to the even case by writing $s = qn + r$ with $0 \leq r \leq n - 1$.

Also solved by Chris Hill, L. R. King, Emanuele Natale (Italy), Marian Tetiva (Romania), University of Louisiana at Lafayette Math Club, and the proposers. Part (a) solved by George Apostolopoulos (Greece), and Elton Bojaxhiu (Germany) and Enkel Hysnelaj (Australia).

An inequality of heights, bisectors, medians, and exradii **June 2010**

1847. *Proposed by Panagiote Ligouras, "Leonardo da Vinci" High School, Noci, Italy.*

Let ABC be a scalene triangle. Let h_a, l_a, and m_a be the respective lengths of the height, bisector, and median, of $\triangle ABC$ with respect to A, and let r_a be the exradius of the excircle of $\triangle ABC$ opposite to A. Similarly, define h_b, l_b, m_b, and r_b, with respect to B, and h_c, l_c, m_c, and r_c with respect to C. Prove that

$$\frac{l_a^4(m_a^2 - h_a^2)}{h_a^3 r_a(l_a^2 - h_a^2)} + \frac{l_b^4(m_b^2 - h_b^2)}{h_b^3 r_b(l_b^2 - h_b^2)} + \frac{l_c^4(m_c^2 - h_c^2)}{h_c^3 r_c(l_c^2 - h_c^2)} > \frac{16}{3}.$$

Solution by Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Germany)
Let Δ and $s = \frac{1}{2}(a+b+c)$ denote the area and the semiperimeter of $\triangle ABC$, respectively. We use the following formulas for the area of $\triangle ABC$,

$$\Delta = rs = r_a(s-a) = r_b(s-b) = r_c(s-c) = \sqrt{s(s-a)(s-b)(s-c)} = \frac{ah_a}{2},$$

to deduce that

$$\frac{1}{r_b} + \frac{1}{r_c} = \frac{(s-b)}{r_b(s-b)} + \frac{(s-c)}{r_c(s-c)} = \frac{(s-b)+(s-c)}{\Delta} = \frac{a}{\Delta} = \frac{2}{h_a} \tag{1}$$

and

$$\begin{aligned} r_a r_b + r_b r_c + r_c r_a &= \frac{\Delta^2((s-c)+(s-b)+(s-a))}{(s-a)(s-b)(s-c)} \\ &= \frac{\Delta^2 s}{(s-a)(s-b)(s-c)} = s^2. \end{aligned} \tag{2}$$

We also use the following well-known relations:

$$l_a^2 = bc\left(1 - \left(\frac{a}{b+c}\right)^2\right),$$

$$h_a^2 = \frac{4}{a^2}(s(s-a)(s-b)(s-c)), \text{ and}$$

$$m_a^2 = \frac{2b^2+2c^2-a^2}{4}.$$

These identities imply that

$$\frac{l_a^4(m_a^2-h_a^2)}{l_a^2-h_a^2} = b^2c^2.$$

In addition $h_a = bc/(2R)$, where R denotes the circumradius of $\triangle ABC$; so the left-hand side of the required inequality becomes

$$\begin{aligned} \frac{l_a^4(m_a^2-h_a^2)}{h_a^3 r_a(l_a^2-h_a^2)} + \frac{l_b^4(m_b^2-h_b^2)}{h_b^3 r_b(l_b^2-h_b^2)} + \frac{l_c^4(m_c^2-h_c^2)}{h_c^3 r_c(l_c^2-h_c^2)} &= \frac{4R^2}{h_a r_a} + \frac{4R^2}{h_b r_b} + \frac{4R^2}{h_c r_c} \\ &= 4R^2\left(\frac{1}{h_a r_a} + \frac{1}{h_b r_b} + \frac{1}{h_c r_c}\right). \end{aligned}$$

Identity (1) implies that

$$\frac{2}{h_a r_a} + \frac{2}{h_b r_b} + \frac{2}{h_c r_c} = \sum_{\text{cyclic}} \left(\frac{1}{r_b} + \frac{1}{r_c}\right)\frac{1}{r_a} = 2\sum_{\text{cyclic}} \frac{1}{r_a r_b},$$

so the initial inequality is equivalent to

$$\frac{1}{r_a r_b} + \frac{1}{r_b r_c} + \frac{1}{r_c r_a} > \frac{4}{3R^2}.$$

The Arithmetic Mean–Harmonic Mean Inequality and Identity (2) imply that

$$\frac{1}{r_a r_b} + \frac{1}{r_b r_c} + \frac{1}{r_c r_a} \geq \frac{9}{r_a r_b + r_b r_c + r_c r_a} = \frac{36}{(a+b+c)^2}.$$

Finally, it is well-known that $a + b + c \leq 3\sqrt{3}R$ with equality if and only if $\triangle ABC$ is equilateral. Therefore

$$\frac{1}{r_a r_b} + \frac{1}{r_b r_c} + \frac{1}{r_c r_a} \geq \frac{36}{(a+b+c)^2} > \frac{4}{3R^2}$$

as required.

Also solved by George Apostolopoulos (Greece), Michel Bataille (France), Minh Can, Chip Curtis, Omran Kouba (Syria), Elias Lampakis (Greece), Peter Nüesch (Switzerland), and the proposer.

Last will be first **June 2010**

1848. *Proposed by Herb Bailey, Rose–Hulman Institute of Technology, Terre Haute, IN.*

Let N be a base ten positive integer with nonzero last digit. Let N^* be the integer formed by moving the last digit of N to the front. For example, if $N = 867053$ then $N^* = 386705$. Find all N for which N is divisible by N^*.

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

We show that N is divisible by N^* if and only if the digits in base 10 representation of N are (a) all the same, (b) consist of the string 428 571 concatenated with itself one or more times, or (c) consist of the string 857 142 concatenated with itself one or more times. To verify this claim, let $n = N/N^*$, $N = 10a + b$ for some integers a and b with $a \geq 0$ and $1 \leq b \leq 9$, and let k be the number of digits in the base 10 representation of a. Then $N^* = 10^k b + a$ and $(10^k b + a)n = 10a + b$. Therefore

$$b(10^k n - 1) = a(10 - n). \tag{1}$$

If $n = 1$, then Identity (1) implies $b(10^k - 1) = 9a$. But $(10^k - 1)/9$ in base 10 is just a string of ones. Therefore, all the digits of N must be the same in this case. Next consider the case $n > 1$. Since $a < 10^k$, Identity (1) implies $(bn - 10 + n)10^k < b < 10$. Hence $bn - 10 + n \leq 0$ and $n(b + 1) \leq 10$. Moreover, since $10^k n - 1$ is not divisible by 2 or by 5, Identity (1) further implies that 8 divides b when $n = 2$, 2 divides b when $n = 4$, and 5 divides b when $n = 5$. These restrictions together with $n(b + 1) \leq 10$ imply that $n = 3$ and b is either 1 or 2. If we now consider Identity (1) in the field of integers modulo 7, we see that $3^k \cdot 3 - 1 \equiv 0 \pmod 7$ and so k must be congruent to 5 modulo 6. Substituting $n = 3$, $b = 1$, and $k = 5$ into Identity (1) yields $a = 42\,857$ and $N = 428\,571$. Using $k = 11$ instead of $k = 5$ yields $N = 428\,571\,428\,571$. In general, each higher value of k simply concatenates an additional 428 571 to N. Finally, if $b = 1$ is replaced by $b = 2$, the only difference in the analysis is that the string 428 571 is replaced by 857 142. This completes the verification of the claim.

Also solved by JPV Abad; George Apostolopoulos (Greece); Michel Bataille (France); Elton Bojaxhiu(Germany) and Enkel Hysnelaj(Australia); Eddie Cheng, László Lipták, David Sherman, and Michael Dimattia; John Christopher; Calvin A. Curtindolph; Chip Curtis; Dmitry Fleischman; Natacha Fontes-Merz; Fullerton College Math Association; Eugene A. Herman; Chris Hill; Omran Kouba (Syria); Victor Y. Kutsenok; Graham Lord; Joel Schlosberg; H. T. Tang; Texas State Problem Solvers Group; Michael Vowe (Switzerland), and the proposer. There were 11 incorrect submissions.

An alternating double series **June 2010**

1849. *Proposed by Ovidiu Furdui, Campia Turzii, Cluj, Romania.*

Find the sum

$$\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\frac{(-1)^{n+m}}{(\lfloor\sqrt{n+m}\rfloor)^3},$$

where $\lfloor a\rfloor$ denotes the greatest integer less than or equal to a.

Solution by Omran Kouba, Higher Institute for Applied sciences and Technology, Damascus, Syria.

First, we note that the considered double series is not absolutely convergent and consequently the order of summation is important. The desired sum S is defined by

$$S=\sum_{m=1}^{\infty}\left(\sum_{n=1}^{\infty}\frac{(-1)^{m+n}}{\left\lfloor\sqrt{m+n}\right\rfloor^3}\right)=\lim_{M\to\infty}\sum_{m=1}^{M}\left(\lim_{N\to\infty}\sum_{n=1}^{N}\frac{(-1)^{m+n}}{\left\lfloor\sqrt{m+n}\right\rfloor^3}\right).$$

We prove that S does exist and that $S=\frac{3}{4}\zeta(3)-\ln 2-\frac{\pi^2}{12}$, where ζ is the well-known Riemann zeta function.

Because the series $\{1/\left\lfloor\sqrt{m+n}\right\rfloor^3\}_{n=1}^{\infty}$ is decreasing and its limit is zero, the alternating series criterion implies that the series $\sum_{n=1}^{\infty}(-1)^{m+n}/\left\lfloor\sqrt{m+n}\right\rfloor^3$ is convergent, let us denote its sum by R_{m+1}. In addition,

$$\begin{aligned}|R_{m+1}|=\left|\sum_{k=m+1}^{\infty}\frac{(-1)^k}{\lfloor\sqrt{k}\rfloor^3}\right|&\le\frac{1}{\left\lfloor\sqrt{m+1}\right\rfloor^3}\\&\le\left(\frac{1}{\sqrt{m+1}-1}\right)^3=\left(\frac{\sqrt{m+1}+1}{m}\right)^3\le\frac{27}{m^{3/2}},\quad(1)\end{aligned}$$

since $\sqrt{m+1}+1<3\sqrt{m}$ for each $m\ge 1$. This proves that the series $\sum_{m=2}^{\infty}R_m$ is absolutely convergent, and that its sum S does exist.

Noting that $R_m-R_{m+1}=(-1)^m/\lfloor\sqrt{m}\rfloor^3$ for $m\ge 1$, it follows that $(m+1)R_{m+1}-mR_m=R_{m+1}-m(-1)^m/\lfloor\sqrt{m}\rfloor^3$ and thus

$$\sum_{m=1}^{M}((m+1)R_{m+1}-mR_m)=\sum_{m=1}^{M}R_{m+1}-\sum_{m=1}^{M}\frac{(-1)^m m}{\lfloor\sqrt{m}\rfloor^3}.$$

By Inequality (1), $\lim_{M\to\infty}MR_M=0$. Hence, taking limits when $M\to\infty$ gives $-R_1=S-\sum_{m=1}^{\infty}(-1)^m m/\lfloor\sqrt{m}\rfloor^3$. That is,

$$S=\sum_{m=1}^{\infty}\frac{(-1)^m(m-1)}{\lfloor\sqrt{m}\rfloor^3}.\quad(2)$$

Noting that $\lfloor\sqrt{m}\rfloor=j$ for $j^2\le m<(j+1)^2$ implies that

$$S=\sum_{j=1}^{\infty}\left(\frac{1}{j^3}\sum_{m=j^2}^{(j+1)^2-1}(-1)^m(m-1)\right).$$

But,

$$\sum_{m=1}^{p-1}(-1)^m(m-1) = \left(\frac{d}{dx}\sum_{m=1}^{p-1}x^{m-1}\right)\Bigg|_{x=-1} = \left(\frac{d}{dx}\left(\frac{x^{p-1}-1}{x-1}\right)\right)\Bigg|_{x=-1}$$
$$= \frac{(p-1)x^{p-2}(x-1)+1-x^{p-1}}{(x-1)^2}\Bigg|_{x=-1} = \frac{1-(2p-3)(-1)^p}{4}.$$

Hence,

$$\sum_{m=j^2}^{(j+1)^2-1}(-1)^m(m-1) = (-1)^j(j^2+j-1).$$

Therefore,

$$S = \sum_{j=1}^{\infty}\frac{(-1)^j(j^2+j-1)}{j^3} = \sum_{j=1}^{\infty}\frac{(-1)^j}{j} + \sum_{j=1}^{\infty}\frac{(-1)^j}{j^2} - \sum_{j=1}^{\infty}\frac{(-1)^j}{j^3}.$$

Finally, the well-known identities

$$\sum_{j=1}^{\infty}\frac{(-1)^j}{j} = -\ln 2, \quad \sum_{j=1}^{\infty}\frac{(-1)^j}{j^2} = -\frac{\pi^2}{12}, \text{ and } \sum_{j=1}^{\infty}\frac{(-1)^j}{j^3} = -\frac{3}{4}\zeta(3)$$

imply that $S = \frac{3}{4}\zeta(3) - \ln 2 - \frac{\pi^2}{12}$, which is the desired conclusion.

Editor's Note. Most potential solvers found the right value of S assuming that the original series could be evaluated by adding over the diagonals $m+n=k$. In this way one would immediately obtain Identity (2). However, this assumption needs proper justification as shown by the series $\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}(-1)^{m+n}/(m+n-1)$ which is convergent, but results in the divergent series $\sum_{k=2}^{\infty}(-1)^k$ when added over the diagonals $m+n=k$.

Also solved by Eugene A. Herman, Peter McPolin (Northern Ireland), and Tiberiu Trif (Romania). There were 15 incomplete or incorrect submissions.

Regular topologies on finite sets **June 2010**

1850. *Proposed by Richard Stephens, Department of Mathematics, Columbus State University, Columbus, GA.*

Let τ be a topology on a finite set X. Define a topology on X to be *regular* if for any nonempty closed $E \subseteq X$ and $x \in X \setminus E$, there exist disjoint open sets U and V in τ such that $E \subseteq V$ and $x \in U$. Prove or disprove that the topological space (X, τ) is regular if and only if τ has a base B which is a partition of X.

Solution by Michael W. Botsko, Saint Vincent College, Latrobe, PA.

We prove that the statement is true.

First assume that (X, τ) has a base $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ which is a partition of X. Since X is finite, $\mathcal{B}$ is finite as well. Let $E \subseteq X$ be any nonempty closed set and let $x \in X \setminus E$. Since $X \setminus E$ is an open set, there is a B_i in $\mathcal{B}$ such that $x \in B_i \subseteq X \setminus E$. Thus $E \subseteq X \setminus B_i = \bigcup_{k\neq i} B_k$. In addition, $B_i \cap \bigcup_{k\neq i} B_k = \emptyset$. Now let $U = B_i$ and $V = \bigcup_{k\neq i} B_k$. Since U and V are disjoint open sets such that $x \in U$ and $E \subseteq V$, it follows that (X, τ) is regular.

Next assume (X, τ) is a regular topological space. For each $x \in X$, let $B_x = \bigcap\{O \in \tau : x \in O\}$. Then B_x is an open set containing x since τ is finite. Let $\mathcal{B} = \{B_x : x \in X\}$. Then $\mathcal{B}$ is a base for X. To see this let $x \in O$ where O is an open set. Then clearly $x \in B_x \subseteq O$. We claim that $\mathcal{B}$ is a partition of X. Since $X = \bigcup_{x\in X} B_x$, we need only show that the sets in $\mathcal{B}$ are pairwise disjoint. We already know that each B_x is an open set. We now show that each B_x is also a closed set. Since $x \in B_x$ and (X, τ) is regular, there exists U an open set such that $x \in U \subseteq \overline{U} \subseteq B_x$, where $\overline{U}$ denotes the closure of U. Now $B_x \subseteq U$ so that $B_x \subseteq U \subseteq \overline{U} \subseteq B_x$. Therefore $B_x = \overline{U}$, so B_x is a closed set.

Finally we note that each B_x is a connected set. To see this suppose $B_x = A \cup C$ where A and C are nonempty open sets and $A \cap C = \emptyset$. Suppose that $x \in A$ so that $B_x \subseteq A$. Thus $B_x = A$ and then C is the empty set which is a contradiction. Thus each B_x is open, closed, and connected which imply that each B_x is a connected component of X. Therefore $\mathcal{B} = \{B_x : x \in X\}$ is a partition of X since all connected components are pairwise disjoint.

Editor's Note. For the proof of the necessity, instead of considering for each $x \in X$ the minimal open set containing x, Mark Bowron used a condition equivalent to regularity, namely that any topological space X is regular if and only if for every $x \in X$ the family of closed neighborhoods of x is a local base for x. Texas State University Problem Solvers Group proved a slightly stronger result for the necessity: If τ is a finite topology on X such that for every E closed set and for every $x \notin E$ there exists an open set W such that $E \subseteq W$ with $x \notin W$, then τ has a basis which is a partition of X. David P. Lang notes that the result is false in general if X is infinite and David A. Rose points out that the result is true when X is infinite assuming that arbitrary intersections of open sets are open, for instance when X is an Alexandroff space.

Also solved by John Atkins, Michel Bataille (France), Tom Beatty, Mark Bowron, Paul Budney, Andrew Burruss, Robert Calcaterra, Calvin A. Curtindolph, Dave Feil, Fullerton College Math Association, Jaime Gutierrez (Panama), Eugene A. Herman, Elias Lampakis (Greece), David P. Lang, Mathramz Problem Solving Group, Richard P. Millspaugh, Northwestern University Math Problem Solving Group, Bee Tammatada Pongsriiam (Thailand), Robert C. Rhoades, Marian Tetiva (Romania), Texas State University Problem Solvers Group, David A. Rose, and the proposer. There were two incomplete submissions.

Answers

Solutions to the Quickies from page 230.

A1011. Note that

$$\frac{\sqrt{k}}{(n+k)\sqrt{n}+n\sqrt{n+k}} = \frac{1}{\sqrt{k}}\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+k}}\right).$$

Thus the required limit equals

$$\lim_{k\to\infty} \frac{1}{\sqrt{k}} \sum_{n=1}^{\infty} \left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+k}}\right) = \lim_{k\to\infty} \frac{1}{\sqrt{k}} \sum_{n=1}^{k} \frac{1}{\sqrt{n}}$$

$$= \lim_{k\to\infty} \frac{1}{k} \sum_{n=1}^{k} \sqrt{\frac{k}{n}} = \int_0^1 \frac{1}{\sqrt{x}}\, dx = 2.$$

A1012. Let $a = BC$, $b = AC$, and $c = AB$. If $b = c$, then any point P on AA' would do. So assume $c > b$, i.e., $\angle C > \angle B$. Let P move from A to A' along segment AA',

and let $x = AC'$ and $y = AB'$. By Ceva's theorem,

$$\frac{x}{y} = \frac{AC'}{AB'} = \frac{BC'}{B'C} \cdot \frac{A'C}{BA'} = \frac{c-x}{b-y} \cdot \frac{b\cos\angle C}{c\cos\angle B}$$

If $AP \to 0$, then $x \to 0$, $y \to 0$, and

$$\frac{x}{y} \to \frac{c-0}{b-0} \cdot \frac{b\cos\angle C}{c\cos\angle B} = \frac{\cos\angle C}{\cos\angle B} < 1.$$

If $AP \to AA'$, then $x \to c$, $y \to b$, and

$$\frac{x}{y} \to \frac{c}{b} > 1.$$

By the intermediate value property, there is a point P on line segment AA' for which $x = y$ as desired.

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.

Clark, Carol, New theories reveal the nature of numbers, http://esciencecommons.blogspot.com/2011/01/new-theories-reveal-nature-of-numbers.html.

Ono, Ken, New theories reveal the nature of numbers, video (71 mins.), http://www.youtube.com/watch?feature=player_embedded&v=aj4FozCSg8g.

Bruinier, Jan Hendrik, and Ken Ono, An algebraic formula for the partition function, http://www.aimath.org/news/partition/brunier-ono.

Amanda Folsom, Zachary A. Kent, and Ken Ono, ℓ-adic properties of the partition function, http://www.aimath.org/news/partition/folsom-kent-ono.pdf.

Bruinier and Ono have made an astonishing breakthrough by finding an explicit finite algebraic formula for the number of partitions $p(n)$ of an integer n—the number of different sums of positive integers that add up to the integer. Moreover, *partition numbers are fractal*: Folsom et al. show that $p(n)$ is a fractal for every prime, in the sense that it can be reduced to multiples of the partition numbers of smaller integers modulo that prime. They give the examples

$$p(13^3n + 1007) \equiv 6p(13n + 6) \pmod{13},$$

where 1007 is the multiplicative inverse of 24 mod 13^2 and 6 is the inverse of 24 mod 13; and

$$p(13^4n + 27371) \equiv 45p(13^2n + 162) \pmod{13^2},$$

where 27371 is the inverse of 24 mod 13^4 and 162 is the inverse of 24 mod 13^2. The fractal theory also explains the famous Ramanujan congruences

$$p(5n + 4) \equiv 0 \pmod{5}, \qquad p(7n + 5) \equiv 0 \pmod{7}, \qquad p(11n + 6) \equiv 0 \pmod{11}.$$

In the video, the formula for partition numbers is at 25:00 and discussion of fractalness at 47:00.

Hua, Jasmin Malik, Continuum: An app that turns any drawing into a made-to-measure dress, http://www.ecouterre.com/continuum-an-app-that-turns-any-drawing-into-a-bespoke-dress/.

The prototype Continuum software converts a customer's drawing of a dress into a three-dimensional model and provides cutting patterns suitable for manufacture or for sewing by hand. The software uses Delaunay triangulations to envision the dress in terms of triangular planes, with adjustments implemented by adding or subtracting triangles. (A Delaunay triangulation of a set of points is the dual graph of the points' Voronoi tessellation; in the latter, the tessellation tile of a point in the set consists of all points in the plane that are nearest to that point.) A Delaunay triangulation and has the property that none of the points is inside the circumcircle of any triangle; so such a triangulation avoids triangles with small angles ("skinny" triangles), which may be a desirable feature for dress design or manufacture. The Website features a video preview of how the software works, plus the claim: "Standardized sizing would be a thing of the past, since garments can be made to order in your exact measurements." Well, given the economic interests of the vast channels of mass marketing, maybe not.

Math. Mag. **84** (2011) 238–239. doi:10.4169/math.mag.84.3.238.

Ceci, Stephen J., and Wendy M. Williams, *The Mathematics of Sex: How Biology and Society Conspire to Limit Talented Women and Girls*, Oxford University Press, 2010; xv + 270 pp, $34.95. ISBN 978-0-19-538939-5.

Ceci, Stephen J., and Wendy M. Williams, Understanding current causes of women's underrepresentation in science, *Proceedings of the National Academy of Sciences*, published online 7 February 2011, doi:10.1073/pnas, http://www.pnas.org/content/early/2011/02/02/1014871108.full.pdf.

Ceci, Stephen J., and Wendy M. Williams, Sex differences in math-intensive fields, *Current Directions in Psychological Science* 19 (5) 275–279.

This book and accompanying papers tackle the question of underrepresentation of women in "math-intensive" fields. If the authors are lucky, their balanced presentation, based on data, will avoid the vituperation that greeted Larry Summers (then president of Harvard) when he speculated about the causes of such underrepresentation. From their abstract in the *PNAS* article: "[W]e reprise claims of discrimination and their evidentiary bases. Based on a review of the past 20 y of data, we suggest that some of these claims are no longer valid and, if uncritically accepted as current causes of women's lack of progress, can delay or prevent understanding of contemporary determinants of women's underrepresentation. We conclude that differential gendered outcomes in the real world result from differences in resources attributable to choices, whether free or constrained, and that such choices could be influenced and better informed through education if resources were so directed. Thus, the ongoing focus on sex discrimination in reviewing, interviewing, and hiring represents costly, misplaced effort. . . ."

Diefenderfer, Caren L., and Roger B. Nelsen (eds.), *The Calculus Collection: A Resource for AP* and Beyond*, MAA, 2010; xx + 507 pp, $74.95 ($59.95 to MAA members). ISBN 978-0-88385-761-8.

If you teach calculus, you should get, use, and enjoy this book! MAA periodicals are a rich source of vitalizing materials for teachers, and republishing in themed "resource volumes" is useful for those who did not see the articles originally and for those who did but have forgotten about them. This book is the third MAA collection about calculus; the previous two, Parts I and II of *A Century of Calculus* (1991), are out of print. This volume is particularly targeted to teachers of Advanced Placement calculus, a wise choice since more students study calculus in high school than in college; of students entering college, one-third have studied calculus and one-fifth took an AP calculus exam. The 123 articles, from the 1980s through 2007, are grouped into General and Historical; Derivatives; Integrals; and Polynomial Approximations and Series. Most are short—some just a page—and only a handful exceed 10 pages. Appendices give the syllabi for the AP calculus exams and (valuably) suggested uses for each article (classroom examples, projects, and/or background reading). There is an author index but no subject index. (Titles and names of the authors have been set in new type, with a few resulting infelicities: For the next printing, let's put a space between Arctan and its argument x on p. 279 and in the table of contents on p. xi, and spell both "Riemann" and "exponential" correctly on p. 413; for consistency, I would capitalize "slope" and "leads" on p. 360 and in the table of contents.)

*The College Board, holder of the trademark "AP," was not involved in the book's production.

Loyer, Milton W., and Gene D. Sprechini, Can the probability of an event be larger or smaller than each of its component conditional probabilities?, *Chance* 24 (1) (2011) 44–53.

The well-known Simpson's paradox points out that baseball player A can have higher batting averages than player B separately against left-handed pitchers and against right-handed pitchers but have a lower average overall than B; the explanation lies in differences in the number of at-bats in each category. Authors Loyer and Sprechini investigate and explain what they assert is a new paradox: That if A or B is chosen to bat with probability one-half, the unconditional probability $P(H)$ of a hit can be greater than both $P(H|R)$ and $P(H|L)$. Again, the weightings in the averages provide the explanation. The authors provide an illustration from their experience with analyzing data on fruit, point out potential pitfalls in using past data to predict future performance (e.g., how do we know what $P(L)$ will be in the future?), and suggest a compromise analysis for the fruit data.

To appear in *College Mathematics Journal,* September 2011

Articles

Student Research Project

Classroom Capsules

Problems and Solutions

Book Reviews

MAA
MATHEMATICAL ASSOCIATION OF AMERICA 1529 Eighteenth St., NW • Washington, DC 20036

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS