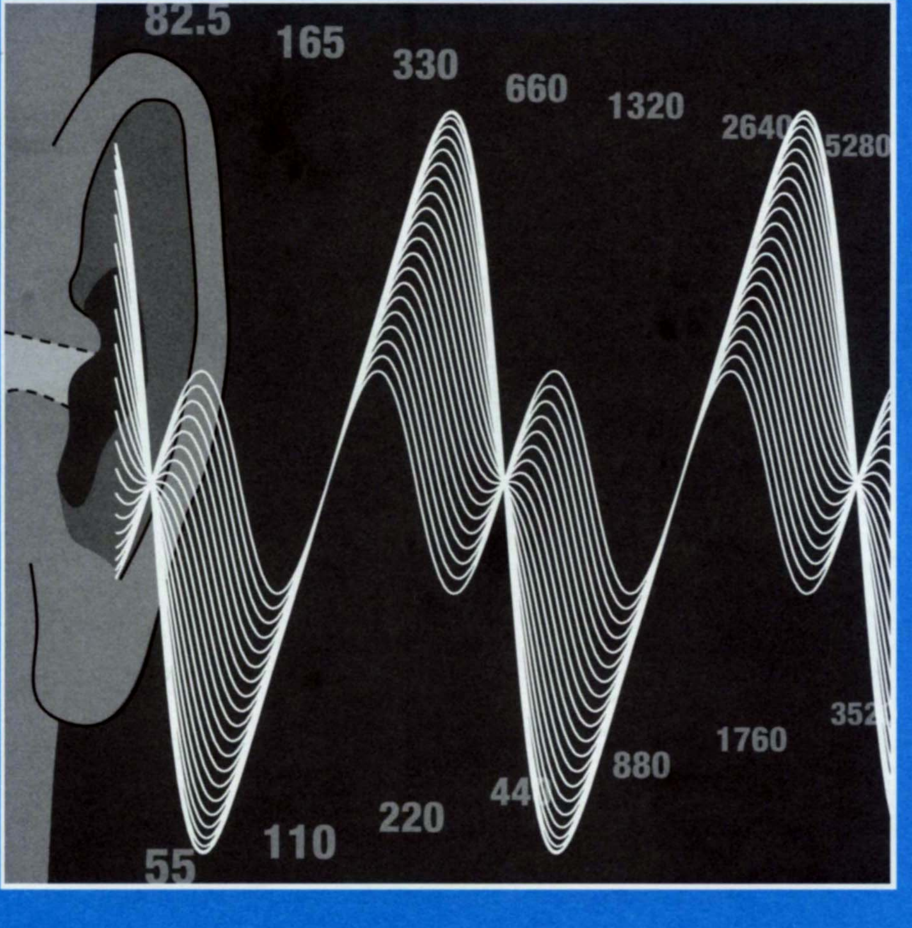
Vol. 74, No. 2, April 2001

MATHEMATICS MAGAZINE



- Designing a Pleasing Sound Mathematically
- A History of Lagrange's Theorem on Groups
- The Euler–Maclaurin and Taylor Formulas: Twin, Elementary Derivations

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

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

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Submit new manuscripts to Frank A. Farris, Editor, Mathematics Magazine, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0373. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*. Copies of figures should be supplied on separate sheets, both with and without lettering added.

Cover image by Jason Challas, who lectures on Computer Art (and making waves) at Santa Clara University.

The waveform depicts Erich Neuwirth's pleasing sound (see p. 96, $q = \frac{1}{3}$), as it moves forward in time.

AUTHORS

Erich Neuwirth holds a Ph.D. in Mathematics and Statistics from the University of Vienna, Austria where he is an associate professor. He teaches statistics, mathematics, and computer science at this university and for many years has been involved in the use of computers in math and science education. He has held visiting positions in Japan (National Institute of Multimedia Education) and the U.S. (College of Computer Science, Northeastern University). For his multimedia project about the mathematical foundations of tuning musical instruments, he received the European Academic Software Award 1996. He is also a hobbyist composer and very proud of the fact that one of his compositions has been premiered at the "Grosser Musikvereinssaal" in Vienna.

Richard L. Roth received his Ph.D. in mathematics in 1963 at the University of California, Berkeley. He has been on the faculty of the University of Colorado, Boulder, since leaving Berkeley. He has also taught and lectured on mathematics in Central America, Colombia, and Sicily. His research interests have varied but have been generally related to groups: collineation groups of finite projective planes, character theory, hypergroups, and symmetry groups such as color groups and the groups of tilings and woven fabrics. He has an ongoing interest in the history of mathematics.

Vito Lampret lives in Slovenia, where he was born. He received his M.M. (master of mathematics) degree in 1980 and his Ph.D. in 1983, both at the University of Ljubljana, Slovenia. His thesis supervisor for both degrees was Professor Ivan Vidav. Since completing his degrees he has taught mathematics on the Faculty of Civil Engineering and Geodesy, Ljubljana. Ever since personal computers first appeared, he has been annoyed by his colleagues' marveling mainly at the speed at which computers calculate. He feels that too many teachers of mathematics (not only in Slovenia) do not recognize the Euler-Maclaurin formula as a fundamental tool for numerical summation. Nowadays he prefers "constructive" mathematics, which actually finds solutions rather than simply establishing their existence.

Vol. 74, No. 2, April 2001

MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $16 included as part of the annual dues. (Annual dues for regular members, exclusive of annual subscription prices for MAA journals, are $64. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 40% dues discount for the first two years of membership.) The nonmember/library subscription price is $68 per year.
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 Dave Riska (*driska@maa.org*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

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

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

Printed in the United States of America

Designing a Pleasing Sound Mathematically

ERICH NEUWIRTH
Department of Statistics and Decision Support Systems
University of Vienna
A-1010 Vienna, Austria

Introduction

If you ask most mathematicians what sound would be best for demonstrating fine distinctions in pitch, they will probably suggest the pure sound of a simple sine wave. In the first version of my book *Mathematical Temperaments* [**2**] (which are the various systems for tuning musical instruments), I used such a wave for all the sound examples, rendered by a computer. That project met with some success, winning the European Academic Software Award in 1996. Even so, while many musicians were interested in the content of the materials, a number of them told me, quite bluntly, that they could not listen to sine waves for more than a few seconds. Could I please reconsider how to implement the sounds? This set in motion the investigation that led to this article.

One of the well-known connections between mathematics and music is the theory of Fourier series. This theory essentially states that every periodic function (in musical terms, every tone) can be represented by an infinite series containing only sine and cosine terms where the frequencies are integer multiples of one fundamental frequency, the frequency of the tone. In this paper, we use the theory of Fourier series and some not too difficult techniques from analysis to construct a tone with certain mathematical characteristics that arose from musical considerations.

Cooperating with a musicologist, I arrived at the following criteria for the sound: Mathematically, the waveform should still be simple: just one relative maximum and one relative minimum, and not too many turning points (changes in concavity). Musically, the waveform should sound, if not pleasing, then at least tolerable for musicians.

In this article, I describe the discovery of a family of sounds that fit these criteria, sounds used in the current version of *Mathematical Temperaments*. On the way, we review the basic facts about Fourier series, in a manner accessible to undergraduates. Also, I describe some simple *Mathematica* programs, and an interactive spreadsheet (running in Microsoft *Excel*) used to experiment with sound. The result is what I hope is an interesting circle of ideas connecting mathematics and music.

Sine waves and Fourier series

Let us briefly summarize a few pieces of mathematical folklore, which are covered in most analysis textbooks covering Fourier analysis [**1**].

A function defined on $\mathbb{R}$ is periodic if there is a (positive) constant c such that $f(x+c) = f(x)$ for all $x \in \mathbb{R}$. We will call c the wavelength of our periodic function if there is no (positive) constant smaller than c with this shift property.

One of the main results of Fourier theory states that every periodic function $f(x)$ with wavelength c, and only finitely many points of discontinuity in the interval $[0, c]$, can be expressed as a series

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n x}{c}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi n x}{c}\right),$$

where equality holds only at points where f is continuous.

At the moment we are only interested in the shape of $f(x)$; therefore we will consider the special case of wavelength $c = 1$ with the series

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(2\pi n x) + \sum_{n=1}^{\infty} b_n \sin(2\pi n x)$$

When all the a_ns in such a Fourier series are zero, the resulting function is symmetric around the origin, $f(-x) = -f(x)$. This is equivalent to the condition that $f(0) = f(1) = 0$, which, in physical terms, corresponds to the waveform representing a vibrating string, secured at its endpoints, a simple model studied in many introductions to Fourier series.

We will restrict our discussion to functions having this symmetry, those constructed with the sine terms only and thus having odd symmetry and period 1. To keep notation simple, when we write an algebraic formula for such a function $f(x)$, we give the formula only in the range [0, 1], intending that this be interpreted to mean the odd, periodic extension of $f(x)$ to the whole line. Typically, we draw just two periods.

Here are some well known examples of basic wave shapes:

$$f_1(x) = 1 - 2x, \qquad \text{sawtooth wave}$$

$$f_2(x) = \begin{cases} 1 & 0 \le x \le \frac{1}{2} \\ -1 & \frac{1}{2} < x < 1 \end{cases}, \qquad \text{square wave}$$

$$f_3(x) = \begin{cases} 4x & 0 \le x \le \frac{1}{4} \\ 2 - 4x & \frac{1}{4} < x \le \frac{3}{4} \\ 4x - 4 & \frac{3}{4} < x \le 1 \end{cases}, \qquad \text{triangle wave}$$

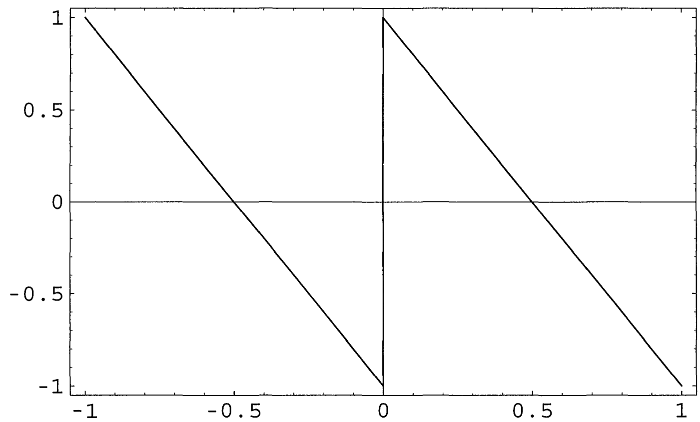


Figure 1 Sawtooth wave

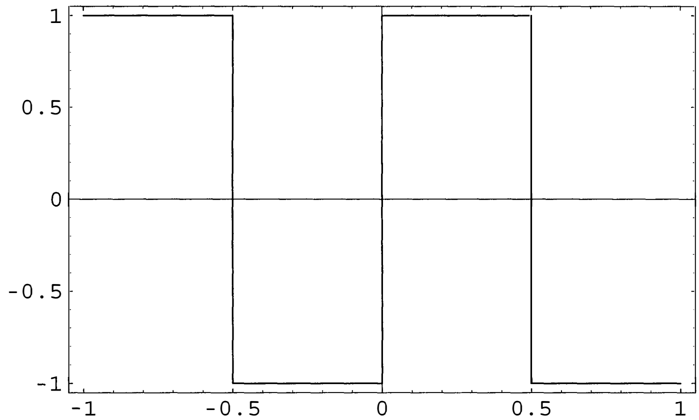


Figure 2 Square wave

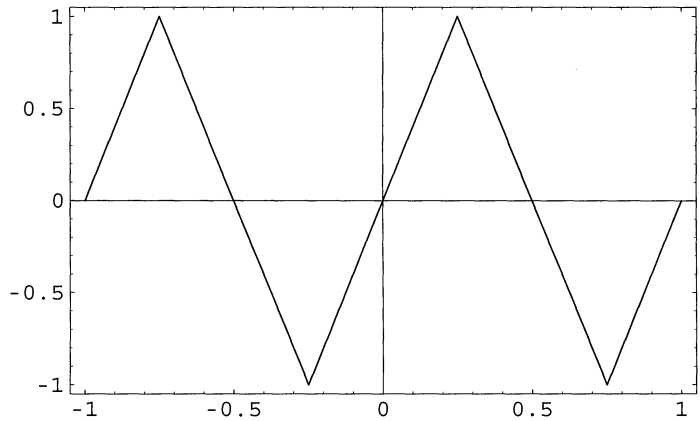


Figure 3 Triangle wave

The standard methods for computing Fourier coefficients show that these functions have the following representations:

$$f_1(x) = \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin(2\pi n x)$$

$$f_2(x) = \frac{4}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}\sin(2\pi(2n-1)x)$$

$$f_3(x) = \frac{8}{\pi^2}\sum_{n=1}^{\infty}(-1)^{n+1}\left(\frac{1}{2n-1}\right)^2\sin(2\pi(2n-1)x)$$

Two helper functions allow us to relate these, which will simplify things when we implement sounds with software. In particular, we use the sign function

$$\sigma(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

and τ defined on $[-1, 1]$ by

$$\tau(x) = \begin{cases} x & |x| \le \frac{1}{2} \\ \sigma(x) - x & |x| > \frac{1}{2} \end{cases}$$

With these in hand, we easily see that $f_2(x) = \sigma(f_1(x))$ and $f_3(x) = 2\tau(f_1(x))$.

These shapes are mathematically attractive, and easy to use with computers. Unfortunately, the sounds they produce sound rather unpleasant, at least to musicians. We offer two ways to hear these waveforms.

An *Excel* spreadsheet can be downloaded from the worldwide web [3]. It allows the user to assign values to the constants b_n, which musicians know as the overtone amplitudes or amplitudes of the harmonics. FIGURE 4 shows a typical view of the spreadsheet.

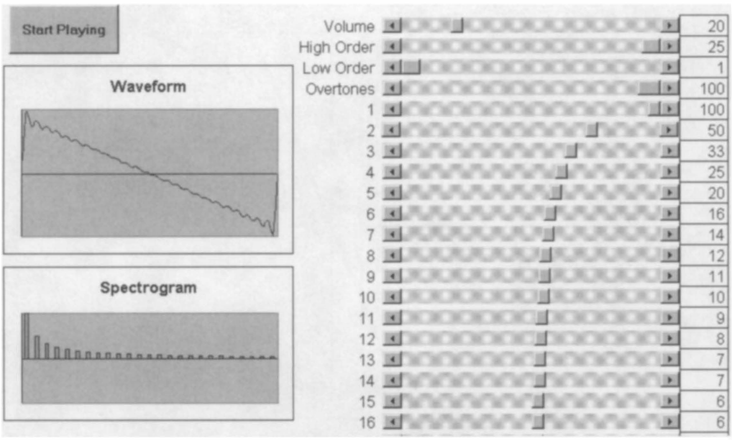


Figure 4 Fourier synthesis spreadsheet

The sliders to the right allow the user to set the values of b_n in the Fourier series. The upper graph displays the shape of the wave, and the button labeled "Start Playing" will play the waveform as a sound. It is possible to change the values of b_n even while the sound is being played, allowing the user to hear changes immediately. More possibilities of interacting with the program are explained in the online document available with the program.

The following *Mathematica* program displays our basic wave shapes, and also "sonifies" the functions we just discussed. The program has been tested in *Mathematica* 3.0 for *Microsoft Windows* 95.

```
SawtoothWave[x_]:=1-2*Mod[x,1]
SquareWave[x_]:=Sign[SawtoothWave[x]]
tau[x_]:=2*If[Abs[x]<1/2,x,Sign[x]-x]
TriangleWave[x_]:=tau[SawtoothWave[x]]
Plot[SawtoothWave[x],{x,-1,1}] Plot[SquareWave[x],{x,-1,1}]
Plot[TriangleWave[x],{x,-1,1}]
Play[SawtoothWave[440*t],{t,0,2},SampleRate->44100]
Play[SquareWave[440*t],{t,0,2},SampleRate->44100]
Play[TriangleWave[440*t],{t,0,2},SampleRate->44100]
```

Setting a high value for SampleRate is quite important, otherwise the sound quality will not be sufficient to demonstrate the effects of the waveforms. The default value of 8192 is too small to get the effects we want to hear.

As I already mentioned, none of these waveforms sound pleasant enough to be used to illustrate musical examples.

In search of the perfect wave

When I was investigating the acoustical and visual properties of waveforms, I implemented them using a spreadsheet program on a Psion, a handheld computer with reasonable graphical capabilities. One entered values for b_n, the amplitudes of the overtones (or harmonics), and saw the graph of $f(x) = \sum_{n=1}^{\infty} b_n \sin(2\pi n x)$.

A young friend, Joseph Proulx, now a mathematics student at the University of Colorado, while playing with this spreadsheet, suddenly produced the following curve:

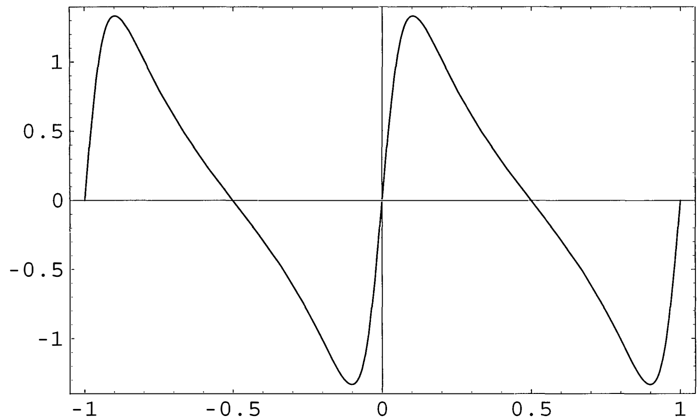


Figure 5 The almost perfect wave

This shape had all the mathematical properties I wanted for a wave: only one local maximum and one local minimum in one cycle, only 3 turning points within one cycle; also, the halfwaves were not symmetric (as in the three examples we discussed before). So I immediately asked the young man how he had constructed this shape. His answer was, essentially, "Well, each overtone amplitude is just half of the previous one." Therefore the function was described by $b_{n+1} = \frac{b_n}{2}$, or $b_n = \left(\frac{1}{2}\right)^{n-1}$.

The coefficients in this Fourier series form a geometric progression with a factor of $\frac{1}{2}$. It is easy to enter these values in the spreadsheet, and immediately see the waveform and hear the corresponding sound.

We also can do this in *Mathematica*:

```
FourierSeries[x_,Coefflist_]:=
   Sum[Sin[2*Pi*n*x]*Coefflist[[n]],
   {n,1,Length[Coefflist]}]
Plot[FourierSeries[t,
   Table[1/(2^(n-1)),{n,1,10}]],
   {t,-1,1},Frame->True]
Play[FourierSeries[440*t,
   Table[1/(2^(n-1)),{n,1,10}]],
   {t,0,2},SampleRate->44100]
```

When I first played a tone with this wave shape to a musicologist, he reacted quite favorably, telling me that this was much better than pure sine waves. And yet, it was not quite perfect, so we considered variations. A natural generalization, of course, would use a geometric progression with a factor q, where $-1 < q < 1$.

So we define

$$f_q(x) = \sum_{n=1}^{\infty} q^{n-1} \sin(2\pi n x).$$

In *Mathematica*, we use:

```
GeomFourierWave[x_,q_,k_]:=
   Sum[Sin[2*Pi*n*x]*q^(n-1),{n,1,k}]
```

This function can be plotted and sonified like the earlier ones.

In discussions with the musicologist we decided that it would be better to make two slight modifications: q will not be constant over the full frequency range, and we will add an envelope so that the amplitude does not go from 0 to 1 at the first wave, but that the volume takes about 0.05 seconds until it reaches "full power."

Before using programs to create the sounds for the production version of *Mathematical Temperaments*, I found that I needed to know more about the properties of $f_q(x)$. I sent a query to an Internet mailing list about sound creation, asking whether anyone had studied it before. With the help of some friendly people in cyberspace, I found that this function has a nice closed-form representation.

To derive this representation it is extremely helpful to consider the function as the imaginary part of a complex function. So, instead of $f_q(x) = \sum_{n=1}^{\infty} q^{n-1} \sin(2\pi n x)$, we study

$$\begin{aligned}
g_q(x) &= \sum_{n=1}^{\infty} q^{n-1} \cos(2\pi n x) + i \sum_{n=1}^{\infty} q^{n-1} \sin(2\pi n x) \\
&= \sum_{n=1}^{\infty} q^{n-1} e^{inx} \\
&= e^{ix} \sum_{n=0}^{\infty} q^n e^{inx} \\
&= e^{ix} \sum_{n=0}^{\infty} (q e^{ix})^n \\
&= \frac{e^{ix}}{1 - q e^{ix}} \\
&= \frac{e^{ix}(1 - q e^{-ix})}{(1 - q e^{ix})(1 - q e^{-ix})} \\
&= \frac{e^{ix} - q}{1 + q^2 - 2q \cos(2\pi x)} \\
&= \frac{(\cos(2\pi x) - q) + i \sin(2\pi x)}{1 + q^2 - 2q \cos(2\pi x)}
\end{aligned}$$

The imaginary part $f_q(x)$ of this function, which we will call $g_q(x)$, is therefore

$$f_q(x) = \frac{\sin(2\pi x)}{1 + q^2 - 2q \cos(2\pi x)}.$$

This representation makes it much easier to study properties of this curve. From the picture, we suspect that the denominator is nonzero for $-1 < q < 1$, but we also can show this directly, as follows:

Clearly, $-|q| \leq q\cos(2\pi x) \leq |q|$, and therefore $1 + q^2 - 2|q| < 1 + q^2 - 2q\cos(2\pi x)$, and $0 < (1 - |q|)^2 \leq 1 + q^2 - 2q\cos(2\pi x)$.

We notice an interesting additional symmetry: $f_q(x) = f_{-q}(\frac{1}{2} - x)$. This is because

$$\begin{aligned} f_{-q}\left(\frac{1}{2} - x\right) &= \frac{\sin\left(2\pi\left(\frac{1}{2} - x\right)\right)}{1 + q^2 + 2q\cos\left(2\pi\left(\frac{1}{2} - x\right)\right)} \\ &= \frac{\sin(\pi - 2\pi x)}{1 + q^2 + 2q\cos(\pi - 2\pi x)} \\ &= \frac{\sin(2\pi x)}{1 + q^2 - 2q\cos(2\pi x)} = f_q(x). \end{aligned}$$

It is also evident that $f_q(x)$ has zeros at $x = \frac{n}{2}$ for all $n \in \mathbb{Z}$.

Next, we study local extrema. The first derivative of $f_q(x)$ is

$$f_q'(x) = \frac{(1 + q^2)2\pi\cos(2\pi x) - 4\pi q}{(1 + q^2 - 2q\cos(2\pi x))^2}.$$

From the equation $f_q'(x) = 0$, we easily derive

$$\cos(2\pi x) = \frac{2q}{1 + q^2}, \quad \text{or } x = \frac{1}{2\pi}\arccos\left(\frac{2q}{1 + q^2}\right).$$

Due to the symmetry of $\cos(x)$ we also have the solution

$$x = 1 - \frac{1}{2\pi}\arccos\left(\frac{2q}{1 + q^2}\right),$$

and the first derivative of $f_q(x)$ does not have any other zeros; therefore, we have just two local extrema on [0, 1]. Since $f_q(x)$ is continuous and periodic with period 1, one of the extremes must be a maximum and the other one has to be a minimum.

In implementation, it is useful to know in advance the highest and lowest points on the curve. Calculating the value of $f_q(x)$ at the extremes yields

$$\frac{(1 + q^2)\sqrt{1 - \frac{4q^2}{(1+q^2)^2}}}{(1 - q^2)^2} \quad \text{and} \quad -\frac{(1 + q^2)\sqrt{1 - \frac{4q^2}{(1+q^2)^2}}}{(1 - q^2)^2}.$$

To study turning points and convexity we need

$$f_q''(x) = \frac{-4\pi^2\sin(2\pi x)(1 - 6q^2 + q^4 + (2q + 2q^3)\cos(2\pi x))}{(1 + q^2 - 2q\cos(2\pi x))^3}.$$

The trivial zeros of this second derivative are the zeros of $\sin(2\pi x)$, which are also zeros of $f_q(x)$. The nontrivial zeros of the second derivative are the zeros of $h(x) = (1 - 6q^2 + q^4 + (2q + 2q^3)\cos(2\pi x))$. As usual, by the symmetry properties of $f_q(x)$, it suffices to study this function for $0 < q < 1$. Since the cosine function ranges from -1 to 1, it is clear that $h(x)$ has zeros iff

$$h_1(q) = (1 - 6q^2 + q^4 + 2q + 2q^3) \quad \text{and} \quad h_2(q) = (1 - 6q^2 + q^4 - 2q - 2q^3)$$

have different signs.

We have $h_1(q) = (1 - 6q^2 + q^4 + 2q + 2q^3) = (q - 1)^2(q^2 + 4q + 1) > 0$ for $0 < q < 1$. Therefore, we must study the sign of

$$h_2(q) = (1 - 6q^2 + q^4 - 2q - 2q^3) = (q + 1)^2(q^2 - 4q + 1).$$

This function has a zero at $q = 2 - \sqrt{3}$ and a zero at $q = 2 + \sqrt{3}$. Since we are only interested in $0 < q \leq 1$, we ignore the larger root, and see that $h_2(q) < 0$ for $q > 2 - \sqrt{3}$. As a consequence, $h(x)$ has zeros between 0 and $\frac{1}{2}$ for $2 - \sqrt{3} < q < 1$. Therefore, $f_q(x)$ has nontrivial turning points not coinciding with its zeros (and therefore is not convex on $[0, \frac{1}{2}]$ for these values of q). On the other hand, for $-(2 - \sqrt{3}) < q < 2 - \sqrt{3}$ our function $f_q(x)$ is convex on $[0, \frac{1}{2}]$ and concave on $[\frac{1}{2}, 1]$.

With these insights about $f_q(x)$, it was not too difficult to use a program for creating sounds based on formulas. In theory, one could even use *Mathematica*, but creating a sound in *Mathematica* (at least the way we did it in our examples) is extremely slow. Since the multimedia project contains more than 400 sound examples, having a fast program to create these files was very important; I used the publicly available program *Csound* [**4**].

It might seem unusual to use mathematics to solve what started out as an aesthetic problem. As digital methods are increasingly used in human environments, there may be many more stories like this one in the future.

REFERENCES

1. H. Dym and H. P. McKean, *Fourier Series and Integrals*, Academic Press, Orlando, FL, 1971.
2. Erich Neuwirth, *Musical Temperaments*, Springer Verlag, Wien, 1997.
3. Erich Neuwirth, "Realtime Fourier Synthesis - Sound generation", `http://sunsite.univie.ac.at/Spreadsite/fourier/fourtone.htm`
4. The Csound Front Page, MIT Press, `http://www.csound.org/`

Proof With Words:

$$2 + 11 - 1 = 12$$

TWo ELeVEn

—Art Benjamin
Harvey Mudd College

A History of Lagrange's Theorem on Groups

RICHARD L. ROTH
University of Colorado
Boulder, CO 80309-0395

Introduction

In group theory, the result known as Lagrange's Theorem states that for a finite group G the order of any subgroup divides the order of G. However, group theory had not yet been invented when Lagrange first gave his result and the theorem took quite a different form. Lagrange's Theorem first appeared in 1770–71 in connection with the problem of solving the general polynomial of degree 5 or higher, and its relation to symmetric functions. It was also anticipated by some results in number theory a few years earlier.

In this article, we explain the historical setting of Lagrange's approach, and follow this train of thought into the twentieth century. Some general references on the history of group theory are Israel Kleiner's article "The Evolution of Group Theory: A Brief Survey" in this MAGAZINE [**22**] and H. Wussing's book, *The Genesis of the Abstract Group Concept* [**31**].

Preliminary Discussion

Let us first review Lagrange's Theorem and its proof, as well as some other results relevant to our discussion. Recall that the *order* of a finite group is the number of elements in the group.

THEOREM A *(Lagrange's Theorem): Let G be a group of order n and H a subgroup of G of order m. Then m is a divisor of n.*

Sketch of Proof. Suppose we list the elements of G in a rectangular array as follows: Let the top row be the list of the m elements of H: $a_1 = e, a_2, \ldots, a_m$. If b is some element of G not in H then let the second row consist of the elements $ba_1, ba_2, \ldots, ba_m$. If there is an element c not in the first two rows then the next row will be $ca_1, ca_2, \ldots ca_m$. This is continued until the elements of G are exhausted. One must then check that the elements in any row are distinct and that no two rows have an element in common. It follows that $n = km$ where k is the number of rows. ■

Early proofs of Lagrange's Theorem generally involved this "rectangular array" explicitly or implicitly. Note that the rows of the rectangular array are simply the left cosets of H in G. In fact, most current texts use the language of cosets to prove this theorem. One must show that the set of left cosets (or right cosets) forms a partition of the group G and that each coset has the same number of elements as H. Note that as an elementary consequence of Lagrange's Theorem we have that the number of cosets of H in G divides the order of G as well. This number is called the index and is denoted $[G : H]$.

Frequently in algebra textbooks, the little theorem of Fermat is proved as a corollary.

FERMAT'S LITTLE THEOREM: *If p is a prime and b is relatively prime to p, then $b^{p-1} \equiv 1 \pmod{p}$.*

Proof. The nonzero elements of $\mathbb{Z}/p\mathbb{Z}$ form a group of order $p-1$ under multiplication, called $(\mathbb{Z}/p\mathbb{Z})^*$. If the congruence class $\overline{b}$ in $(\mathbb{Z}/p\mathbb{Z})^*$ has order m, then it generates a cyclic subgroup of order m. By Lagrange's Theorem, m divides $p-1$; thus $(\overline{b})^{p-1} = (\overline{b}^m)^k = \overline{1}$, where $p-1 = mk$, and the theorem follows. ∎

Also relevant is Theorem B, given below. As we will see, it might be just as appropriate to call this Lagrange's Theorem.

THEOREM B: *Let G be a finite group acting by permutations on a finite set S. Then the size of any orbit is a divisor of the order of G.*

Proof. Let b be an element of some particular orbit and H be the subgroup of G stabilizing b. If c is another element in this orbit then for some τ in G, $\tau b = c$. If $\sigma \in H$ then $(\tau\sigma)b = \tau(\sigma b) = \tau b = c$ (in our notation it is the element σ in the stabilizer of b that acts first) and it is easily seen that the coset τH consists of precisely the elements mapping b onto c. Thus the elements of the orbit are in one-to-one correspondence with the left cosets of H, hence the size of the orbit equals the index $[G : H]$, which, by our consequence of Theorem A above, must divide G. ∎

Lagrange's version of the theorem

In 1770–71, Lagrange published a landmark work on the theory of equations, "*Reflexions sur la résolution algébrique des équations*" [**23**]. (Note: Katz gives a useful discussion of this paper in section 14.2.6 of [**21**].) His concern was the question of finding an algebraic formula for the roots of the general 5th degree polynomial and more generally for the nth degree polynomial for $n > 4$. The quadratic formula had, of course, been known for a very long time and the cubic and quartic equations had been solved in the sixteenth century by algebraists of the Italian school. However, for polynomials of degree greater than four this had remained an open problem for two centuries. Lagrange observed that the solutions for the cubic and quartic equations involved solving supplementary "resolvent" polynomials of lower degree whose coefficients were rational functions of the coefficients of the original polynomial. He found that the roots of these auxiliary equations were in fact "functions" of the roots of the original equation that took on a small number of values when the original roots were permuted in the formulas for these functions.

For example, the quartic was solved using a cubic resolvent polynomial whose roots could be written as $\frac{x_1x_2+x_3x_4}{2}$, $\frac{x_1x_3+x_2x_4}{2}$ and $\frac{x_1x_4+x_2x_3}{2}$ where x_1, x_2, x_3, x_4 were the roots of the original polynomial. If the four roots x_1, x_2, x_3, x_4 are permuted in all 24 possible ways, only these three different "values" typically occur. For convenience, we will omit the denominator 2 in what follows. We list below the result of operating on the function $x_1x_2 + x_3x_4$ by the 24 different permutations of the four variables. The stabilizer of the function is a group of 8 elements and the first row shows the seven other values that are equal to the original one. Underneath each value is the corresponding permutation. The second row shows the set of eight values arising from a different way of combining the four roots, all equal to $x_1x_3 + x_2x_4$; similarly for the third row. Underneath each value is the corresponding permutation. We note that in the second and third lines we have permuted the positions of the variables in the same manner as was done in the first line. This corresponds to the step in the proof of Theorem B above where the stabilizing permutation σ must act before the permutation τ that changes the object. For example, referring to the first two equal functions in

$x_1x_2 + x_3x_4$	$x_2x_1 + x_3x_4$	$x_3x_4 + x_1x_2$	$x_1x_2 + x_4x_3$	$x_4x_3 + x_1x_2$	$x_3x_4 + x_2x_1$	$x_4x_3 + x_2x_1$	$x_2x_1 + x_4x_3$
id	(12)	$(13)(24)$	(34)	(1423)	(1324)	$(14)(23)$	$(12)(34)$
$x_1x_3 + x_2x_4$	$x_3x_1 + x_2x_4$	$x_2x_4 + x_1x_3$	$x_1x_3 + x_4x_2$	$x_4x_2 + x_1x_3$	$x_2x_4 + x_3x_1$	$x_4x_2 + x_3x_1$	$x_3x_1 + x_4x_2$
(23)	$(23)(12) = (132)$	$(23)(13)(24) = (1243)$	$(23)(34) = (234)$	$(23)(1423) = (143)$	$(23)(1324) = (124)$	$(23)(14)(23) = (14)$	$(23)(12)(34) = (1342)$
$x_1x_4 + x_2x_3$	$x_4x_1 + x_2x_3$	$x_2x_3 + x_1x_4$	$x_1x_4 + x_3x_2$	$x_3x_2 + x_1x_4$	$x_2x_3 + x_4x_1$	$x_3x_2 + x_4x_1$	$x_4x_1 + x_3x_2$
(243)	$(243)(12) = (1432)$	$(243)(13)(24) = (123)$	$(243)(34) = (24)$	$(243)(1423) = (13)$	$(243)(1324) = (1234)$	$(243)(14)(23) = (134)$	$(243)(12)(34) = (142)$

rows 1 and 2, we have $\sigma = (1\ 2)$, $\tau = (2\ 3)$ and $\tau\sigma = (2\ 3)(1\ 2) = (1\ 3\ 2)$. We thus have a 3 by 8 rectangular array. The 3 "values" multiplied by 8 gives $24 = 4! = |S_4|$.

The 4th degree polynomial was solvable because there was a "function" of 4 variables which took on 3 "values" when the 4 variables are permuted in all $24 = 4!$ ways. That is, these 3 "values" were the roots of a cubic polynomial (which it was known how to solve); these roots could be used to modify the original 4th degree polynomial so that it would factor into quadratic polynomials. Using the theory of symmetric functions, Lagrange proved that if a rational function of the n roots of a general polynomial of degree n takes on r "values" under the action of all $n!$ permutations, then the function will be a root of a polynomial of degree r whose coefficients are rational functions of the coefficients of the original equation.

Thus Lagrange reasoned that to solve a 5th degree polynomial, one should try to find a function in 5 variables that takes on 3 (or 4) different typical "values" when the variables are permuted in all 5! ways. This would lead to a cubic (or quartic) resolvent that might help to solve the original equation. A similar approach might apply for solving equations of degree n for n greater than 5.

Lagrange was unable to determine if such functions exist. But he did come up with, in essence, the following theorem.

THEOREM C: THEOREM OF LAGRANGE: *If a function $f(x_1, \ldots, x_n)$ of n variables is acted on by all $n!$ possible permutations of the variables and these permuted functions take on only r distinct values, then r is a divisor of $n!$.*

In fact, Lagrange stated his theorem in terms of the degree of the corresponding resolvent equation. Also, we note that if $n = 5$, then 3 and 4 are both divisors of $n!$, so the theorem of Lagrange doesn't answer the previous question as to whether or not a cubic or quartic resolvent exists for the 5th degree equation.

Lagrange's proof of Theorem C consisted essentially of discussing some special cases; it is interesting to note that the treatment that he gave for the first case was partly wrong, although it did give the correct idea for a proof. He said: let us suppose that a function satisfies $f(x', x'', x''', x^{iv} \ldots) = f(x'', x''', x', x^{iv} \ldots)$. Such a function satisfies $f(x^{iv}, x''', x', x'', \ldots) = f(x''', x', x^{iv}, x'' \ldots.)$, because we have permuted the first 3 variables in the same way; hence, he said, all the values will match up in pairs and the possible number of distinct values will be reduced to $\frac{n!}{2}$. However, the permutation involved is actually a cycle of length 3, so in fact in this example the number of values would be divided by 3, and not by 2.

Lagrange then said that if the original function remains the same under 3 or 4 or a larger number of permutations, then the other values will also have that property, and the total number of distinct values will be $\frac{n!}{3}$ or $\frac{n!}{4}$, etc.

Thus Lagrange's original Theorem C might be regarded as a special case of Theorem B, where the group G is the symmetric group S_n, the set S is the set of functions (or formulas) involving n variables formed by all permutations of the n variables and the group action is that which arises from permuting the variables in these functions.

Coincidentally, in 1771 Vandermonde also wrote a paper ([**28**]) on the theory of equations that took an approach similar to Lagrange's. The alternating function $\prod_{1 \le i < j \le n}(x_i - x_j)$ takes on exactly two values when the variables are permuted. In Vandermonde's paper we find this function for the case when $n = 3$. It is used today (for arbitary n) in contemporary abstract algebra books to study even and odd permutations; the set of permutations that stabilize it forms the alternating group. As can be seen from our paper, the use of polynomial functions is historically very much a part of group theory. The alternating function $\prod_{1 \le i < j \le n}(x_i - x_j)$ is actually equal to the Vandermonde determinant, and it is probable that the origin of the name comes from this reference, although Vandermonde (who elsewhere did do important work on determinants) did not express it as a determinant. For the case of $n = 5$, Vandermonde gave a different example of a function taking on two values under permutations, and expressed the opinion that there does not exist such a function of five variables taking on either three or four values.

Some later developments related to Lagrange's work

Several decades later, Paolo Ruffini made further progress in Lagrange's approach to solving polynomial equations. His book of 1799 [**25**] included an informal proof (by example) of Lagrange's Theorem C. Further, Ruffini showed that there does not exist any function of 5 variables taking on three values or four values. Thus the "converse" to Theorem C is false. It seems appropriate to call this "Ruffini's Theorem." In modern terminology this shows that the converse to Lagrange's Theorem (Theorem A) is false because the symmetric group on 5 letters of order 120 has no subgroup of order 40 or 30.

Ruffini also claimed to have proved (as a consequence) that the 5th degree equation (and in general the nth degree for $n \geq 5$) was not solvable. His work drew much criticism and even though he published several more versions, his proof is generally regarded as incomplete. (A more satisfactory proof would be given later by Abel in the 1820s.) A friendly response came from Abbati in 1802 [**1**] who gave some suggestions to improve Ruffini's proof. Abbati's note included an extensive and thoughtful proof of Lagrange's Theorem C. According to Heinrich Burkhardt [**5**] this was the first time a complete proof was given. It resembled the proof of Theorem A given above, and presented a rectangular array of terms (as illustrated in our earlier discussion of the function $x_1x_2 + x_3x_4$). The given function was acted on by the $n!$ permutations. The first row gave the list of functions (arising from permutations) that are equal to the original function. The next row started with a value that was not in the first row and consisted of all subsequent permutations of this function. Because the modern notation used in the proof of Theorem B was not available, a lengthier discussion was used. Abbati took some care to explain that what mattered was the positions of the variables in the function and the way they were combined. Then, as in the proof of Theorem A, each row was shown to have the same number of elements, with no two rows having any element in common.

Further developments were obtained with Cauchy's important 1815 paper [**7**], whose title roughly translated into English reads "*Memoir on the number of values that a function can acquire (when one permutes in all possible ways the quantities it contains).*" This paper launched permutation group theory as an independent topic even though the notion of a group did not appear in it. The paper was not concerned directly with the theory of equations. Cauchy included a proof of Theorem C similar to Abbati's, pointing out that the permutations fixing a function are to be applied to the positions of the variables and not their indices. He went on to generalize Ruffini's theorem as follows: If the number of values of a non-symmetric function of n quantities is less than the largest prime divisor p of n then it must be 2. In the 1820s Abel cited and used this theorem from Cauchy's paper specifically for the special case that $n = 5$ in his work on the unsolvability of the quintic (see [**2**, p. 31] of Oeuvres). Although Ruffini was not explicitly mentioned in Abel's paper, Abel's proof of the unsolvability of the 5th degree polynomial thus relied indirectly on the work of Ruffini.

Galois introduced the term *group* for permutation groups in a paper on solutions of polynomials by radicals in 1831 [**14**, pp. 35–36]. He didn't explicitly mention either form of Lagrange's Theorem in any of his papers, but in the famous letter written the night before his death [**15**], he did include the suggestive equation for the coset decomposition

$$G = H + HS + HS' + \cdots. \quad (1)$$

These works were not widely known until they were published in Liouville's Journal in 1846.

Theory of permutation groups

Almost 30 years after Cauchy's 1815 paper, Cauchy again took up the subject of permutation groups. His paper of 1844 [**8**] did not take the "values of a function" approach; rather it dealt directly with permutation groups. Keeping the older use of the word "permutation" to refer to an ordered arrangement of symbols, he used the term "substitution" to refer to a permutation, and "system of conjugate substitutions" (système de substitutions conjugées) to refer to a permutation group. This was defined here and in later works on the subject merely as a set of permutations closed under composition. This is of course sufficient to define a group: composition is associative, and since we are dealing with finite permutations, the existence of the identity and inverse operations will be necessarily implied. (It is perhaps unfortunate that in many modern books the definition of group only lists three axioms, namely associativity and the existence of the identity and inverse operations. The closure property, which was really the original property used to define permutation groups, is hidden in the term "operation" or "binary operation.") Cauchy then proved the following theorem (translated into English): "The order of a system of conjugate substitutions on n variables is always a divisor of the number N of arrangements which one can form with these variables" [**8**, p. 207]; i.e. the order of a subgroup of the symmetric group S_n is a divisor of $n!$. Thus we now had Theorem A (Lagrange's Theorem) for the case that G is the symmetric group.

Cauchy's long paper of 1844 was followed by a series of shorter papers over the next couple of years that further developed the theory of permutation groups. They included showing the connection between Theorem A and Theorem C. In [**9**], he showed that the set of permutations fixing a function forms a permutation group (i.e. a subgroup of S_n, which today we call the stabilizer of that function) and that, conversely, for any such subgroup H of S_n there is a function whose stabilizer is precisely that subgroup. Given a subgroup H of S_n, we can build a function of n variables whose stabilizer is H in the following way. Let $s = a_1x_1 + \cdots + a_nx_n$, where $a_1, \ldots, a_n$ are distinct numbers. Let $s_1, s_2, \ldots, s_t$ be the images under the t elements of the subgroup H (note we may assume $s = s_1$); then consider the product $s_1s_2 \ldots s_t$. With appropriate assumptions on the coefficients $a_1, a_2, \ldots, a_n$, this will be a function whose stabilizer is precisely H. Cauchy required these coefficients to be nonzero elements whose sum is not zero and his proof is a bit unclear, but if we require instead that the coefficients be a set of nonzero integers whose greatest common divisor is 1, then the result can be shown using the unique factorization property of polynomial rings over a unique factorization domain.

Another example of a function whose stabilizer is H, also given by Cauchy, arises by taking $s = x_1x_2^2x_3^3 \ldots x_n^n$, and defining $s_1 = s, s_2, \ldots, s_t$ to be the images of s under the subgroup H; then the sum $s_1 + \cdots + s_t$ is a function whose stabilizer is H. In a later note [**10**] he showed that if a function takes on m distinct values and the subgroup fixing the function has order M then $mM = n!$. Because every subgroup is the stabilizer of some function, this is a kind of "hybrid form" of Lagrange's Theorem in which Cauchy combined Theorem C and Theorem A for the case of $G = S_n$. In this note he also laid the foundations for the idea of a group acting on a set (as in Theorem B above). In particular he showed that a group of permutations on the n variables $x_1, \dots, x_n$ could also be regarded as acting by permutations on the set of functions that arise from a particular function of those variables under permutations of the variables.

The next step in the development of Lagrange's Theorem was to see that Theorem A holds for any finite permutation group G. This result may be found in Camille Jordan's thesis, published in 1861 [**19**]. This is a rather lengthy and technical paper on permutation groups. In the introduction he cited Lagrange's Theorem, in the "hybrid" form just mentioned, calling it a theorem due to Lagrange and crediting his proof to Cauchy. Forty-five pages later in the midst of a complicated counting argument, he mentioned that he would need a generalization of the Lagrange's Theorem and proved that (in modern language) the order of a subgroup of any permutation group divides the order of the group. Lagrange's Theorem then appeared in this form in the 3rd edition of Serret's important algebra text *Cours d'Algébre supérieure* published in 1866 [**27**] (in a later chapter, the theory is applied to the topic of the number of values of a function of n variables). And it is also found in Jordan's influential 1871 book *Traité des substitutions et des èquations algébriques* [**20**].

Lagrange's Theorem C is essentially Theorem A for the case where $G = S_n$, as we have seen above: each function has its stabilizing subgroup of S_n and each subgroup of S_n may be regarded as the stabilizer of an appropriate function. The extension of Theorem A to the case of G, an arbitrary permutation group, was more general and might seem to have no analogue in terms of functions. However, it is interesting to note that in Netto's book of 1882 [**24**] he did give a translation into the function approach. "Lehrsatz VI" in chapter 3 states: If φ and ψ are two functions of the same n variables and the permutations leaving φ fixed also leave ψ unchanged, then the number of values taken on by φ is a multiple of the number of values taken on by ψ.

Other directions in group theory

While the theory of permutation groups played a major role in the development of general group theory, there were a number of other important sources of group theory including geometry and number theory. One connection to number theory, as we noted in the introduction, is Fermat's Little Theorem. Euler gave several proofs of this theorem. Of interest here is his paper whose title translated into English is "Theorems on residues obtained by division of powers," written in 1758–59 and published in 1761 [**11**]. In it he gave a proof along the lines indicated in the beginning of this article. He proved Lagrange's Theorem in essentially the usual way (the rectangular array) for the case that G is $(\mathbb{Z}/p\mathbb{Z})^*$ (the multiplicative group of the integers relatively prime to p, modulo p) and H is the cyclic subgroup generated by $\overline{b}$. Thus one could argue that in some sense, Lagrange's Theorem appeared 10 years before Lagrange's work. The theorem of Fermat generalizes in two directions; if, instead of $(\mathbb{Z}/p\mathbb{Z})^*$, we consider $(\mathbb{Z}/n\mathbb{Z})^*$, the classes of integers relatively prime to n, then $b^{\phi(n)} \equiv 1 \bmod n$ (where φ is the Euler function). This was shown by Euler in a paper written 1760–61, published in 1763 [**12**]. The other direction is to regard $(\mathbb{Z}/p\mathbb{Z})^*$ as a finite field and to generalize this to Galois fields $GF(p^n)$, in which case we have $x^{p^n-1} = 1$, and this was done by Galois in 1830 [**13**]. Again, in both cases, the proofs implicitly involved a rectangular array approach to a special case of Lagrange's Theorem.

The development of the abstract approach to groups is discussed in [**22**] and [**31**]: abstract groups generalized permutation groups, various groups arising in number theory (including those arising from modular arithmetic, as we've seen in the previous paragraph) and geometrical groups. The abstract approach to groups caught on in the

1880s. Of course Lagrange's Theorem (Theorem A) and its proof were by then easily adapted to abstract groups. It is hard to pinpoint the first abstract version; but, for example, in a paper of Hölder in 1889 [**18**] on Galois theory, an abstract definition for finite groups was given. Lagrange's Theorem was proved by merely displaying the familiar rectangular array. The theorem was not identified by Lagrange's name, but the rectangular array is credited to Cauchy's 1844 paper.

In 1895–96, Weber's *Lehrbuch der Algebra* [**30**] was published in two volumes. This became the standard text for modern algebra for the next few decades. Volume 1 (1895) was the more elementary of the two volumes. As for group theory it treated only permutation groups. It included Theorem A for permutation groups. The theorem was credited to Cauchy. The proof, however, was done using the terminology of cosets ("nebengruppe"). It was shown that two cosets are either equal or disjoint, and stated that the size of any coset equals the order of the subgroup. Following the proof, the symbolic equation

$$P = Q + Q\pi_1 + Q\pi_2 + \cdots + Q\pi_{j-1} \tag{2}$$

was displayed (where P is the group and Q is the subgroup) to give more insight into the proof, and this equation was credited to Galois. Note that Galois's version (equation (1)) did not specify a finite sum, however. This may have been the first time that the theorem was proved using the language of cosets.

Weber's Volume 2 (1896) covered more advanced material than volume 1; it began with abstract groups and proved Lagrange's Theorem again, this time for abstract groups, in section 2 of chapter 1. The proof was essentially the rectangular array approach (without explicitly giving the array). The theorem and its proof were then used as motivation for introducing the concept of coset ("as in the special case of permutation groups") and a coset decomposition equation, similar to equation (2) was displayed.

Theorem C did not appear in Weber's book and the only consideration of the number of values taken on by a function under permutations was the case of symmetric functions and the alternating function (discussed in volume 1). A proof of Fermat's Little Theorem appeared in volume 1 (proved without group theory); in chapter 2 of volume 2 the generalization of the Fermat Theorem for the integers modulo n was proved using group theory and Lagrange's Theorem.

Twentieth century developments

Increasingly in the twentieth century, coset terminology was used in the proof of Lagrange's Theorem. It is not so different from the rectangular array approach, since the rows of the array are in fact the cosets. But it is a different style, and while the rectangular array is usually credited to Cauchy, the coset approach seems to have been inspired by Galois's coset decomposition equation (1). We saw that these two historical threads were brought together in Weber's book.

According to Wussing [**31**], the first monograph devoted to abstract group theory was that of DeSeguier, appearing in 1904 [**26**]. He showed that the double cosets form a partition of a group. Then as a special case, he got the coset decomposition of a group G and then simply mentioned that $[G, A][A, 1] = [G, 1]$. Here $[G, A]$ denotes the index of a subgroup A in G so $[G, 1]$ is the order of G.

Although a number of authors credited the theorem to Lagrange, many did not mention Lagrange's name and it was some time before it became widely known as "Lagrange's Theorem." Van der Waerden's *Moderne Algebra* [**29**] was one of the most influential texts in algebra. It first appeared in 1930. The coset approach was used in proving Lagrange's Theorem. Lagrange's name did not appear; however, in a footnote it was mentioned that the coset decomposition equation (1), which is frequently found in the literature, is due to Galois. In the "second revised edition" of 1937, the fact that any two cosets have the same number of elements was explicitly stated and proved, thus filling a gap in the first edition. In the English translation (of the second revised edition), appearing in 1949, the translator added a footnote stating that this theorem is also known as Lagrange's Theorem.

We might compare two other books which also appeared in 1937. Albert's *Modern Higher Algebra* **[3]** used the language of cosets to prove the theorem; there was no mention of Lagrange or other historical references. On the other hand, in Carmichael's *Introduction to the Theory of Groups of Finite Order* **[6]**, the proof was given using the rectangular array. Carmichael called the theorem "First Fundamental Theorem" (his chapter 2 contained five fundamental theorems) but there was a footnote stating "This has sometimes been called the *theorem of Lagrange*".

In 1941, Birkhoff and MacLane's *A Survey of Modern Algebra* first appeared **[4]**. This book became a model for undergraduate modern algebra textbooks and helped to attach Lagrange's name firmly to the theorem. Section 9 of chapter VI is entitled "Lagrange's Theorem." It started with two lemmas showing that each coset has the same number of elements as the subgroups and any two distinct cosets are disjoint. This led up to "Theorem 18 (Lagrange): *The order of a finite group G is a multiple of the order of every one of its subgroups.*" There were various corollaries including Fermat's Little Theorem.

Some modern books apply the general theory of equivalence relations in connection with cosets and Lagrange's Theorem (for example, see Herstein's *Abstract Algebra* **[17]**). One defines a relation on the group G by letting aRb if $ab^{-1} \epsilon H$. It is proved that this is an equivalence relation and the right cosets are the equivalence classes. The right cosets thus form a partition of the group G. One of the earliest examples of this approach is found in a book of Hasse, published 1926 (see **[16]**).

Acknowledgment. We wish to thank the referees for many helpful suggestions.

REFERENCES

1. P. Abbati, Lettera di Pietro Abbati Modenese al socio Paolo Ruffini da questo presentata il dì 16. Decembre 1802, *Mem. Mat. Fis. Soc. Ital.* **10** (part 2) 1803, 385–409.
2. N. H. Abel, Mémoire sur les équations algébriques, ou l'on démontre l'impossibilité de la résolution de l'equation génerale du cinquième degré. Christiania 1824 (*Oeuvres Complétes de Niels Henrik Abel*, Christiania 1881, Johnson Reprint 1965, vol. 1, pp. 28–33).
3. A. Albert, *Modern Higher Algebra*, University of Chicago Press, Chicago, IL, 1937.
4. G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, Macmillan, New York, 1941.
5. H. Burkhardt, Die Anfänge der Gruppentheorie, Abh. Gesch. **6** (1892), 119–159.
6. R. D. Carmichael, *Introduction to the Theory of Groups of Finite Order*, Dover Publications, New York, 1956 (originally published 1937).
7. A.-L. Cauchy, Sur Le Nombre des Valeurs qu'une Fonction peut acquérir, lorsqu'on y permute de toutes les manières posibles les quantités qu'elle renferme, Journ. Ecole Polyt. **10**, (1815), 1–28. (Oeuvres Complètes, Gauthiers Villars, Paris, 1905, Series 2, vol. 1, pp. 64–90).
8. A.-L. Cauchy, Mémoire sur les arrangements que l'on peut former avec des lettres donneés, et sur les permutations ou substitutions à l'aide desquelles on passe d'un arrangement à un autre. In: Exercises d'analyse et de physique mathématique, 3, Paris, 1844, pp. 151–252. (Oeuvres Complètes, Gauthiers Villars, Paris, 1932, Series 2, vol. 13, pp. 172–282).
9. A.-L. Cauchy, Sur le nombre des valeurs ègales ou inègales que peut acquérir une fonction de n variables indèpendantes, quand on permute ces variables entre elles d'une manière quelconque, Comptes Rendus Paris 21, Oct. 6, 1845, pp. 779–797 (Oeuvres Complètes, Series 1, vol. 9, Gauthiers Villars, Paris, 1896, Extract 303, pp. 323–341).
10. A.-L. Cauchy, Mèmoire sur un nouveau calcul qui permet de simplifier et d'étendre la théorie des permutations, Comptes rendus, Paris 22, Jan. 12, 1846, pp. 53–63 (Oeuvres Complètes, Series 1, vol. 10, Gauthiers-Villars, Paris, 1897, pp. 35–46).

11. L. Euler, Theoremata circa residua ex divisione potestatum relicta, Nov. Com. Acad. Petrop. (1758–59), Petersburg, 1761. (English translation of portion in D. Struik, *Source Book of Mathematics, 1200–1800*, Harvard Univ. Press, Cambridge, MA, 1969, Chapter 1, section 8).
12. L. Euler, Nova Methodo Demonstrata Theoremata Arithmetica, Nov. Comm. Acad. Sci. Petrop. **8** (1760–61, pp. 74–104, publ. 1763), Opera (1), 3, 535–55.
13. E. Galois, Sur La Théorie des Nombres, Bull. Sc. math. Férussac 13, June 1830, (Oeuvres Math., Gauthiers-Villars, Paris, 1897, pp. 15–23).
14. E. Galois, Mémoire dur les conditions de résolubilité des équations par radicaux (Oeuvres math., Gauthiers Villars, Paris, 1897, pp. 33–61).
15. E. Galois, Lettre a Auguste Chevalier, Revue Encyclopedique, 1832 (Oeuvres math., Gauthiers-Villars, Paris, 1897, pp. 25–32).
16. H. Hasse, *Höhere Algebra*, vol 1, Berlin, W. de Gruyter and Co., (Sammlung Göschen 931), 1926.
17. I. N. Herstein, *Abstract Algebra*, Prentice Hall, 1995 (Third edition).
18. O. Hölder, Zurückführung einer beliebigen algebraischen Gleichung auf eine Kette von Gleichungen, Math. Ann. 34 (1889) 26–56.
19. C. Jordan, Mémoire sur le nombre des valeurs des fonctions. Journ. Ecole Polyt. **22** (1861), 113–194. (Oeuvres, Gauthier-Villars, Paris, 1961, vol. 1, item 1, pp. 1–82).
20. C. Jordan, *Traité des Substitutions et des équations algébriques*, Paris, 1870.
21. V. J. Katz, *A History of Mathematics, An Introduction*, Addison Wesley, 1998, Second Edition.
22. I. Kleiner, The Evolution of Group Theory: A Brief Survey, this MAGAZINE **59** (1986) 195–215.
23. J.-L. Lagrange, Réflexions sur la résolution algébrique des équations. Nouv. Mém. Acad. Berlin, pour les anneés 1770/71, Berlin 1772/1773. (See also Oeuvres de Lagrange, edited by J.-A. Serret, Paris, 1869.)
24. E. Netto, *Die Substitutiontheorie und ihre Anwendung auf die Algebra*, Leipzig, 1882.
25. P. Ruffini: *Teoria Generale delle Equazioni, in cui si è dimostrata impossibile la soluzione algebraica delle equazioni generali di grado superiore al quarto.* Bologna, 1799 (Opere Mathematiche, vol I, Ed. E. Bortolotti, Palermo 1915).
26. J.-A. de Séguier, *Théorie des Groupes Finis. Eléments de la Théorie des Groupes Abstraits*, Gauthier Villars, Paris, 1904.
27. J.-A. Serret, *Cours de Algèbre supérieure*, 3rd edition, Paris, 1866.
28. A. Vandermonde, Mémoire sur la résolution des équations, Hist. Acad. Sc. Paris, année 1771: Paris, 1774, pp. 365–416.
29. B. L. van der Waerden, *Moderne Algebra*, Springer, Berlin, First Edition 1930, Second Revised Edition 1937. (English translation of the second revised published by Frederick Ungar, New York, 1949, 1953.)
30. H. Weber, *Lehrbuch der Algebra*, F. Vieweg & Sohn, Braunschweig vol 1, 1895, vol. 2, 1896.
31. H. Wussing, *The Genesis of the Abstract Group Concept*, M.I.T. Press, Cambridge, MA, 1984. (Translation from the German by A. Shenitzer of *Die Genesis des Abstraken Gruppenbegriffes*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1969.)

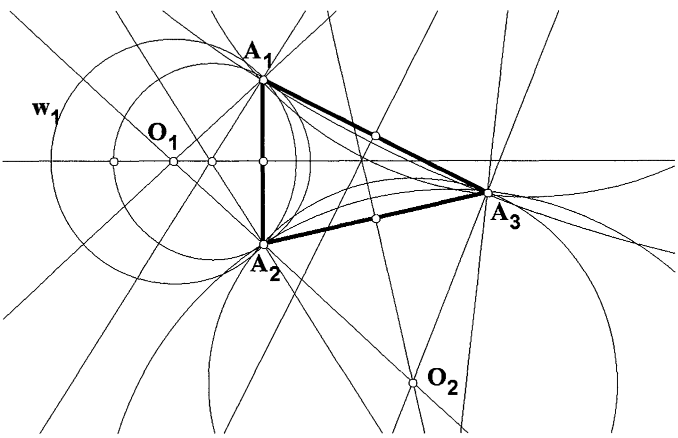


Figure 1 Problem 5 from USA Mathematical Olympiad (p. 167)

The Euler–Maclaurin and Taylor Formulas: Twin, Elementary Derivations

VITO LAMPRET
University of Ljubljana
1000 Ljubljana
386 Slovenia

Introduction

The calculator is perfectly suited to brute force processes. Want to sum a few hundred terms of a series? The calculator can do it almost instantly. The existence of the calculator might suggest that computational leverage provided by calculus is no longer needed. But although the arena in which leverage is required has shifted a bit, our ability to compute is still immeasurably enriched by the power of calculus. Indeed, that power is used in the design of modern software and calculators. For instance, programs like *Maple* and *Mathematica* compute sums like $\sum_{k=1}^{1000^{1000}} 1/k$ and $\sum_{k=1}^{\infty} 1/k^3$ in the blink of an eye. How do they do it?

The *Mathematica* manual [**18**, p. 917] reveals that *Mathematica* actually uses the famous *Euler–Maclaurin* (E–M) *formula*, of which one form states that for $m \le n$,

$$\sum_{k=m}^{n} f(k) - \int_m^n f(x)\,dx = \frac{1}{2}\left[f(m)+f(n)\right] + \frac{1}{12}\left[f'(n)-f'(m)\right] + \rho(f;m,n),$$

where

$$|\rho(f;m,n)| \le \frac{1}{120}\int_m^n \left|f'''(x)\right|\,dx.$$

Let us illustrate how this formula works. By setting $f(x) = 1/x$ we get, for $m \le n$,

$$\sum_{k=1}^{n}\frac{1}{k} = \left(\sum_{k=1}^{m-1}\frac{1}{k} - \ln m + \frac{1}{2m} + \frac{1}{12m^2}\right) + \left(\ln n + \frac{1}{2n} - \frac{1}{12n^2}\right) + \rho(m,n),$$

where

$$|\rho(m,n)| \le \frac{1}{120}\left(\frac{2}{m^3} - \frac{2}{n^3}\right) < \frac{1}{60m^3}.$$

For example, $|\rho(m,n)| \le 1.7 \times 10^{-8}$ for $n \ge m \ge 100$, so

$$\sum_{k=1}^{n}\frac{1}{k} = \left(\sum_{k=1}^{99}\frac{1}{k} - \ln 100 + \frac{1}{200} + \frac{1}{120000}\right) + \left(\ln n + \frac{1}{2n} - \frac{1}{12n^2}\right) + \rho(100,n)$$

$$= 0.577215664 + \ln n + \left(\frac{1}{2n} - \frac{1}{12n^2}\right) + \delta(n).$$

The first expression in parentheses above was truncated to nine places; the resulting error was combined with the error $\rho(100,n)$ to give a new error $\delta(n)$, where

$$|\delta(n)| < 10^{-9} + 1.7 \times 10^{-8} = 1.8 \times 10^{-8}$$

for $n \geq 100$. For $n = 1000^{1000}$ we read

$$S := \sum_{k=1}^{1000^{1000}} \frac{1}{k} = \left(0.577215664 + \ln 1000^{1000}\right) + \delta(1000^{1000})$$

and compute the first parenthesized expression to nine places. The error of this approximation we combine with $\delta(1000^{1000})$ to get a new error Δ with $S = 6908.332494646\ldots + \Delta$. Since $|\Delta| < 1.9 \times 10^{-8}$, $S = 6908.3324946\ldots$ is correct to seven decimal places.

The E–M summation formula is among the most remarkable formulas of mathematics [**15**, p. 11]. In fact, neither Euler nor Maclaurin found this formula with remainder; the first to do so was Poisson, in 1823 ([**14**], see also [**8**, p. 471] or [**11**, p. 521]). Since then the E–M formula has been derived in different ways; one of the earliest derivations (1834) was presented by Jacobi [**10**]. Boas [**3**, p. 246] gave an elegant derivation of the E–M formula using the Stieltjes integral. An elementary derivation of this formula has long been known using the method of integration by parts (see Glaisher as cited in [**4**]; see also [**17**, pp. 125, 127]). Apostol [**1**] presents another nice elementary derivation of the E–M formula by the same means.

Every standard textbook in analysis contains Taylor's formula, but few include the E–M formula, perhaps because of its somewhat cumbersome form and its relatively complicated derivation. In this paper we present a completely elementary derivation of the E–M formula, which also produces Taylor's formula. The main idea stems from the observation that, when integrating a function whose pth derivative is more "controllable" than the function itself, we may apply the method of the integration by parts p times. If we organize this process appropriately, we obtain such interesting formulas as Taylor's formula and the E–M formula.

Preliminaries

For an integer $n \geq 0$ and a closed interval $[a, b]$, let $C^n[a, b]$ denote the set of all n-times continuously differentiable functions defined on $[a, b]$. The *integration by parts* formula asserts that

$$\int_a^b u(t)\, v'(t)\, dt = [u(t)\, v(t)]_a^b - \int_a^b u'(t)\, v(t)\, dt$$

for $u, v \in C^1[a, b]$. This is easily generalized by induction to the following elementary but important lemma on *repeated* integration by parts in closed form:

Lemma. *For $u, v \in C^n[a, b]$,*

$$\int_a^b u(t)\, v^{(n)}(t)\, dt = \left[\sum_{i=0}^{n-1} (-1)^i\, u^{(i)}(t)\, v^{(n-1-i)}(t)\right]_a^b + (-1)^n \int_a^b u^{(n)}(t)\, v(t)\, dt.$$

Example 1. To calculate $\int x^3 e^x dx$, we substitute $a = 0$, $b = x$, $u(t) := t^3$; $v(t) := e^t$; and $n = 4$ into the preceding equation to obtain

$$\int x^3 e^x\, dx = \left[\sum_{i=0}^{3} (-1)^i (t^3)^{(i)}\, (e^t)^{(3-i)}\right]_0^x + 0 + C = \left(x^3 - 3x^2 + 6x - 6\right) e^x + C.$$

If we put $a = 0$, $b = 1$, and $u, v \in C^n[0, 1]$ into the lemma above and then get rid of the factors $(-1)^i$ by replacing $v(t)$ with $v(1-t)$, and then cancel $(-1)^n$, we arrive at the basic equality

$$\int_0^1 u(t)\, v^{(n)}(1-t)\, dt = \sum_{i=0}^{n-1} \left[u^{(i)}(0)\, v^{(n-1-i)}(1) - u^{(i)}(1)\, v^{(n-1-i)}(0)\right] + \int_0^1 v(1-t)\, u^{(n)}(t)\, dt. \tag{1}$$

Both Taylor's formula and the E–M formula can be derived by judicious choices of v in (1).

Taylor's formula

To obtain Taylor's formula it suffices to take for v in formula (1) a function whose derivative vanishes to an appropriate high order. Our derivation is in two steps.

Unit increment. Let p be any nonnegative integer and $u \in C^{p+1}[0, 1]$. In (1) we set $v(t) = t^p/p!$ and $n = p+1$. Since $v^{(p+1)}(t) \equiv 0$ we have $\int_0^1 u(t)v^{(p+1)}(1-t)dt = 0$. Since $v^{(j)}(t) \equiv t^{p-j}/(p-j)!$ for $0 \le j \le p$, we get from (1) the equality

$$u(0) - u(1) + \sum_{i=1}^{p} \left[\frac{u^{(i)}(0)}{i!} - 0\right] + \int_0^1 \frac{(1-t)^p}{p!}\, u^{(p+1)}(t)\, dt = 0.$$

Including $u(0)$ under the summation sign produces

$$u(1) = \sum_{i=0}^{p} \frac{u^{(i)}(0)}{i!} + \frac{1}{p!} \int_0^1 (1-t)^p u^{(p+1)}(t)\, dt, \tag{2}$$

which is Taylor's formula for a unit increment.

Arbitrary increment. For a function $f \in C^{p+1}[a, b]$ and numbers x_o and $x_o + h$ in $[a, b]$, we define the function $u \in C^{p+1}[0, 1]$ by $u(t) = f(x_0 + ht)$. Since $u^{(i)}(t) \equiv h^i f^{(i)}(x_0 + ht)$ for $i = 0, 1, \dots, p+1$, we obtain from (2)

$$f(x_0 + h) = u(1) = \sum_{i=0}^{p} \frac{f^{(i)}(x_0)}{i!} h^i + \frac{h^{p+1}}{p!} \int_0^1 (1-t)^p\, f^{(p+1)}(x_0 + ht)\, dt,$$

which is *Taylor's formula of order p with remainder.*

Euler–Maclaurin formula

To obtain this formula it suffices to take for v in the identity (1) a function whose derivative of an appropriately high order is identically equal to 1.

Connecting integrals and derivatives Let p be a positive integer, $u \in C^p[0, 1]$, and v a function such that $v^{(p)}(t) \equiv 1$. Then we have $\int_0^1 u(t)dt = \int_0^1 u(t)v^{(p)}(1-t)dt$, and the formula (1) can be applied. So, we want a sequence $\{v_k\}$ of polynomials, such that $v_k^{(k)}(t) \equiv 1$ for $k \ge 0$. An easy approach is to put

$$v_0(t) := 1 \quad \text{and} \quad v_k'(t) := v_{k-1}(t) \quad \text{for} \quad k \geq 1. \tag{3}$$

This recursion formula does not uniquely determine the sequence $\{v_k\}$, but (3) implies that $v_p^{(j)} = v_{p-j}$ for $j = 0, 1, \ldots, p$, so $v_p^{(p-1-i)} = v_{i+1}$ for $i = 0, 1, \ldots, p-1$. Putting this into (1) we obtain for $n = p$ the equality

$$\int_0^1 u(t)\,dt = \sum_{j=1}^{p} \left[u^{(j-1)}(0)\,v_j(1) - u^{(j-1)}(1)v_j(0)\right] + \int_0^1 v_p(1-t)\,u^{(p)}(t)\,dt. \tag{4}$$

To simplify the expression in square brackets we require, in addition, that

$$v_j(0) = v_j(1) \quad \text{for} \quad j = 2, 3, \ldots. \tag{5a}$$

By (3), $v_j(1) - v_j(0) = \int_0^1 v_j'(t)dt = \int_0^1 v_{j-1}(t)dt$ for $j \geq 1$, so (5a) is equivalent to

$$\int_0^1 v_k(t)dt = 0 \quad \text{for} \quad k = 1, 2, 3, \ldots. \tag{5b}$$

The polynomial sequence (v_k) is now completely determined by (3) and (5b). For example, (3) implies $v_1(t) \equiv t + C$, and from (5b) we find $C = -1/2$, so $v_1(t) \equiv t - 1/2$. Since $v_1(0) = -1/2 = -v_1(1)$, we can rewrite (4), using (5a), in the form

$$u(1) = \int_0^1 u(t)\,dt + \sum_{j=1}^{p} v_j(1)\left[u^{(j-1)}(t)\right]_0^1 - \int_0^1 v_p(1-t)\,u^{(p)}(t)\,dt. \tag{6}$$

Connecting sums and integrals Our next goal is to connect summation with integration. To this end, let n and p be positive integers and let $\varphi \in C^p[0, n]$. We define functions u_i by $u_i(t) = \varphi(i + t)$, for $t \in [0, 1]$ and integers $i = 0, 1, \ldots, n-1$. Then $u_i \in C^p[0, 1]$ and $\sum_{i=0}^{n-1} u_i(1) = \sum_{i=1}^{n} \varphi(i)$. Now (6) implies the following equalities:

$$\begin{aligned}
\sum_{i=1}^{n} \varphi(i) &= \sum_{i=0}^{n-1} \left\{ \int_0^1 u_i(t)\,dt + \sum_{j=1}^{p} v_j(1)\left[u_i^{(j-1)}(t)\right]_0^1 - \int_0^1 v_p(1-t)\,u_i^{(p)}(t)\,dt \right\} \\
&= \sum_{i=0}^{n-1} \int_i^{i+1} \varphi(\tau)\,d\tau + \sum_{i=0}^{n-1}\sum_{j=1}^{p} v_j(1)\left[\varphi^{(j-1)}(t)\right]_i^{i+1} \\
&\quad - \sum_{i=0}^{n-1} \int_0^1 v_p(1-t)\,\varphi^{(p)}(i+t)\,dt \\
&= \int_0^n \varphi(\tau)\,d\tau + \sum_{j=1}^{p} v_j(1) \sum_{i=0}^{n-1} \left[\varphi^{(j-1)}(t)\right]_i^{i+1} \\
&\quad - \sum_{i=0}^{n-1} \int_0^1 v_p(1-t)\varphi^{(p)}(i+t)\,dt \\
&= \int_0^n \varphi(\tau)\,d\tau + \sum_{j=1}^{p} v_j(1)\left[\varphi^{(j-1)}(t)\right]_0^n - \sum_{i=0}^{n-1} \int_0^1 v_p(1-t)\,\varphi^{(p)}(i+t)\,dt.
\end{aligned}$$

Since $v_1(1) = -v_1(0) = 1/2$, and $v_j(0) = v_j(1)$ for $j \geq 2$, we can rewrite the last equality as

$$\sum_{i=0}^{n-1} \varphi(i) = \int_0^n \varphi(t)\, dt + \sum_{j=1}^{p} v_j(0)\big[\varphi^{(j-1)}(t)\big]_0^n - \sum_{i=0}^{n-1} \int_0^1 v_p(1-t)\, \varphi^{(p)}(i+t)\, dt. \quad (7)$$

In order to simplify the last sum in this formula we introduce periodic functions $w_i(x)$ $(i \geq 0)$, defined by $w_i(x) := v_i(x - \lfloor x \rfloor)$ for any real x. (Here $\lfloor x \rfloor$ denotes the integer part of x.) Then

$$w_i(x) = v_i(x) \text{ for } 0 \leq x < 1 \quad \text{and} \quad w_i(x+1) = w_i(x) \text{ for } x \in \mathbb{R}. \quad (8)$$

As 1 is the period of $w_i(x)$, we have $w_i(x+m) = w_i(x)$ for $x \in \mathbb{R}$ and $m \in \mathbb{Z}$. Substituting $i + t = \tau$ in the integrals we get

$$\begin{aligned}
\sum_{i=0}^{n-1} \int_0^1 v_p(1-t)\, \varphi^{(p)}(i+t)\, dt &= \sum_{i=0}^{n-1} \int_0^1 w_p(-t)\, \varphi^{(p)}(i+t)\, dt \\
&= \sum_{i=0}^{n-1} \int_i^{i+1} w_p(i-\tau)\, \varphi^{(p)}(\tau)\, d\tau \\
&= \sum_{i=0}^{n-1} \int_i^{i+1} w_p(-\tau)\, \varphi^{(p)}(\tau)\, d\tau \\
&= \int_0^n w_p(-\tau)\, \varphi^{(p)}(\tau)\, d\tau.
\end{aligned}$$

Thus, from (7) we conclude

$$\sum_{i=0}^{n-1} \varphi(i) = \int_0^n \varphi(t)\, dt + \sum_{j=1}^{p} v_j(0)\big[\varphi^{(j-1)}(t)\big]_0^n - \int_0^n w_p(-t)\, \varphi^{(p)}(t)\, dt. \quad (9)$$

Connecting Riemann sums and integrals For a function $f \in C^p[a, b]$ and a positive integer n we put $h = (b-a)/n$ and define φ by $\varphi(t) := f(a + ht)$. Hence $\varphi \in C^p[0, n]$, with $\varphi^{(j)}(t) = h^j f^{(j)}(a + ht)$, for $j = 0, 1, \ldots, p$, and $\sum_{i=0}^{n-1} f(a + ih) = \sum_{i=0}^{n-1} \varphi(i)$. Now (9) becomes

$$\begin{aligned}
\sum_{i=0}^{n-1} f(a+ih) &= \int_0^n \varphi(t)\, dt + \sum_{j=1}^{p} v_j(0)\, \big[\varphi^{(j-1)}(t)\big]_0^n - \int_0^n w_p(-t)\, \varphi^{(p)}(t)\, dt \\
&= \frac{1}{h} \int_a^b f(x)\, dx + \sum_{j=1}^{p} v_j(0) h^{j-1} \big[f^{(j-1)}(t)\big]_a^{a+hn} \\
&\quad - h^{p-1} \int_a^b w_p\left(\frac{a-x}{h}\right) f^{(p)}(x)\, dx,
\end{aligned}$$

or

$$\begin{aligned}
h \sum_{i=0}^{n-1} f(a+ih) - \int_a^b f(x)\, dx &= \sum_{j=1}^{p} v_j(0)\, h^j \big[f^{(j-1)}(x)\big]_a^b \\
&\quad - h^p \int_a^b w_p\left(\frac{a-x}{h}\right) f^{(p)}(x)\, dx. \quad (10)
\end{aligned}$$

This is the E–M formula in terms of the chosen functions $v_k(x)$ and $w_k(x)$.

Better known than these functions are the *Bernoulli polynomials* $B_k(x)$ and the *Bernoulli periodic functions* $P_k(x)$, related to $v_k(x)$ and $w_k(x)$ by

$$B_k\,(x) := k!\,v_k(x) \quad \text{and} \quad P_k\,(x) := k!\,w_k(x) \tag{11}$$

for $x \in \mathbb{R}$ and $k \geq 0$. According to (3), (5b), and (8), these functions are uniquely determined by the conditions

$$B_0(x) \equiv 1; \qquad B_k'(x) \equiv kB_{k-1}(x); \qquad \int_0^1 B_k(x)\,dx = 0 \tag{12}$$

for $k \geq 1$, and by

$$P_k(x) \equiv B_k(x) \quad \text{on} \quad [0,1) \quad \text{and} \quad P_k(x+1) \equiv P_k(x) \quad \text{on} \quad \mathbb{R} \tag{13}$$

for all $k \geq 0$.

We can derive from (12) the first eight nonconstant Bernoulli polynomials:

$$\begin{aligned}
B_1(x) &= x - \frac{1}{2}\\
B_2(x) &= x^2 - x + \frac{1}{6} \;=\; -x(1-x) + \frac{1}{6}\\
B_3(x) &= x^3 - \frac{3x^2}{2} + \frac{x}{2} \;=\; -x\,(1-x)\left(x - \frac{1}{2}\right)\\
B_4(x) &= x^4 - 2x^3 + x^2 - \frac{1}{30} \;=\; \left(x - x^2\right)^2 - \frac{1}{30}\\
B_5(x) &= x^5 - \frac{5x^4}{2} + \frac{5x^3}{3} - \frac{x}{6} \;=\; -\left[\frac{1}{3} + x(1-x)\right]B_3(x)\\
B_6(x) &= x^6 - 3x^5 + \frac{5x^4}{2} - \frac{x^2}{2} + \frac{1}{42} = -\left(x - x^2\right)^2\left[\frac{1}{2} + x(1-x)\right] + \frac{1}{42}\\
B_7(x) &= x^7 - \frac{7x^6}{2} + \frac{7x^5}{2} - \frac{7x^3}{6} + \frac{x}{6} = \left[\frac{1}{3} + x(1-x)\,(1 + x(1-x))\right]B_3(x)\\
B_8(x) &= x^8 - 4x^7 + \frac{14x^6}{3} - \frac{7x^4}{3} + \frac{2x^2}{3} - \frac{1}{30}\\
&= \left(x - x^2\right)^2\left[\frac{2}{3} + x(1-x)\left(\frac{4}{3} + x(1-x)\right)\right] - \frac{1}{30}.
\end{aligned} \tag{14}$$

The numbers $B_k := B_k(0)$, $k = 0, 1, 2, \ldots$ are called *Bernoulli coefficients*; from (14) we read

$$B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = B_5 = B_7 = 0, \quad B_4 = B_8 = -\frac{1}{30}, \quad B_6 = \frac{1}{42}. \tag{15}$$

Graphs of several $B_k(t)$ and $P_k(t)$ are shown in FIGURES 1 and 2.

THEOREM 1. (BASIC EULER–MACLAURIN FORMULA OF ORDER p) *For any integers n, $p \geq 1$ and any function $f \in C^p[a,b]$,*

$$\sum_{i=0}^{n-1} f(a+ih)\,h - \int_a^b f(x)\,dx = \sum_{j=1}^{p} h^j \frac{B_j}{j!}\left[f^{(j-1)}(x)\right]_a^b + r_p(a,b,n), \tag{16a}$$

where B_j are Bernoulli coefficients, $h = (b-a)/n$, and $r_p(a,b,n)$ is the remainder of order p given by the formula

$$r_p(a,b,n) = -\frac{h^p}{p!}\int_a^b P_p\left(\frac{a-x}{h}\right) f^{(p)}(x)\,dx, \tag{16b}$$

where P_p is the p-th Bernoulli periodic function.

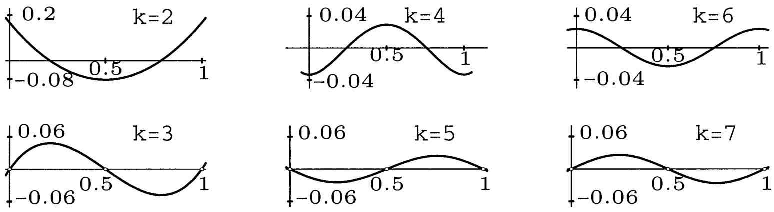


Figure 1 Bernoulli polynomials

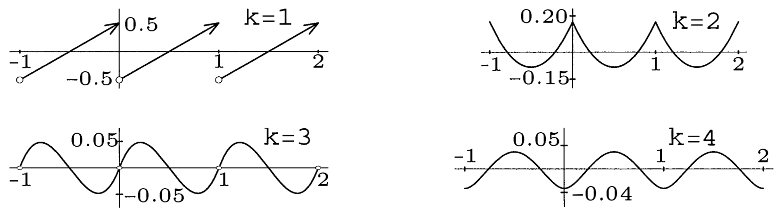


Figure 2 Bernoulli periodic functions

We arrive at the main result by replacing in (10) the functions v_k by B_k and w_k by P_k.

Formula (16a) gives the difference between a Riemann sum and the integral of a function f in terms of its derivatives at the end points of the interval of integration. (We note that it can be shown quite easily that in (16a), all the B_j with odd $j \geq 3$ actually vanish. However, for more general E–M formulas the situation is quite different; see the remarks at the end of this article.) Formulas (16b) and (13) imply a rough estimate for the remainder. If $\mu_p := \max_{0\leq x\leq 1} |B_p(x)|$, then

$$|r_p(a,b,n)| \leq \mu_p \frac{h^p}{p!}\int_a^b |f^{(p)}(x)|\,dx. \tag{17}$$

For elementary applications of the E–M formula, we need some of the numbers μ_p. Equations (14) imply some basic estimates:

$$\mu_1 = \frac{1}{2};\quad \mu_2 = \frac{1}{6};\quad \mu_4 = \mu_8 = \frac{1}{30};\quad \mu_6 = \frac{1}{42};$$
$$\mu_3 < \frac{1}{20};\quad \mu_5 < \frac{1}{35};\quad \mu_7 < \frac{1}{30}. \tag{18}$$

Using μ_6, for instance, we get

$$|r_6(a,b,n)| \leq \frac{h^6}{30240}\int_a^b |f^{(6)}(x)|\,dx \leq \frac{M(b-a)^7}{30240\,n^6}, \tag{19}$$

where $M = \max_{a\leq x\leq b} |f^{(6)}(x)|$.

Applications of the Euler–Maclaurin formula

Numerical integration Formulas (16a) and (16b) constitute a simple and useful method of numerical integration, especially if f has easily computable derivatives through the pth order. From (16a) and (15) we read

$$\int_a^b f(x)\,dx = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{n-1} f(a+ih) - \sum_{j=2}^{p} h^j \frac{B_j}{j!}\left[f^{(j-1)}(x)\right]_a^b - r_p(a,b,n).$$

This formula, together with (17) and (18), represents a good tool for numerical integration. It becomes considerably simpler in the case that $f^{(k)}(a) = f^{(k)}(b)$ for k even and less than p. (Note that $B_j = 0$ for odd $j \geq 3$.) This happens, for instance, when f is periodic with period $b-a$ ([**5**, p. 137]).

Let us take $p = 4$ in the preceding formula. Then (15) and (16b) give

$$\int_a^b f(x)\,dx = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{n-1} f(a+ih) - \frac{h^2}{12}\left[f'(b)-f'(a)\right] - \\ - h^4\frac{B_4}{4!}\left[f^{(3)}(x)\right]_a^b + \frac{h^4}{4!}\int_a^b P_4\left(\frac{a-x}{h}\right) f^{(4)}(x)\,dx.$$

Substituting $x = a + th$ and writing $b = a + nh$, we can combine the last two summands:

$$\begin{aligned}
-h^4\frac{B_4}{4!}&\left[f^{(3)}(x)\right]_a^b + \frac{h^4}{4!}\int_a^b P_4\left(\frac{a-x}{h}\right) f^{(4)}(x)\,dx \\
&= \frac{h^4}{4!}\int_a^b \left[P_4\left(\frac{a-x}{h}\right) - B_4\right] f^{(4)}(x)\,dx \\
&= \frac{h^5}{4!}\int_0^n \left[P_4(-t) - B_4\right] f^{(4)}(a+th)\,dt \\
&= \frac{h^5}{4!} f^{(4)}(\xi)\int_0^n \left[P_4(-t) - B_4\right] dt \\
&= \frac{h^5}{4!} f^{(4)}(\xi)\cdot(-nB_4) \\
&= -\frac{B_4}{4!}\frac{(b-a)^5}{n^4} f^{(4)}(\xi)
\end{aligned}$$

at some $\xi \in [a,b]$. The third equality uses the mean value theorem. Namely, by (14), $B_4(x) - B_4 = \left(x - x^2\right)^2 \geq 0$, so, by (13) the difference $P_4(x) - B_4$ is nonnegative as well. The fourth equation follows from the fact that $\int_0^n P_4(-t)\,dt = 0$, according to (13) and (12). So we get *Hermite's integration formula*:

$$\int_a^b f(x)\,dx = s(a,b,n) + r(a,b,n), \tag{20a}$$

where

$$s(a,b,n) = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{n-1} f(a+ih) - \frac{h^2}{12}\left[f'(b)-f'(a)\right] \tag{20b}$$

and, by (15),

$$r(a,b,n) = \frac{(b-a)^5}{720} f^{(4)}(\xi)\cdot n^{-4} \tag{20c}$$

for some $\xi \in [a,b]$. This remainder is about one-fourth that for Simpson's rule; the advantage is even greater if $f'(b) = f'(a)$.

Example 2. To compute $I = \int_0^2 e^{-x^2}\,dx$ approximately, we put $a = 0$, $b = 2$, $f(x) := e^{-x^2}$, to get $f'(x) \equiv -2xe^{-x^2}$ and $f^{(4)}(x) \equiv 4e^{-x^2}(4x^4 - 12x^2 + 3)$, and evaluate

$$M = \max\left\{\left|f^{(4)}(x)\right| : 0 \le x \le 2\right\} = \max\left\{\left|4e^{-t}\left(4t^2 - 12t + 3\right)\right| : 0 \le t \le 4\right\} = 12.$$

So by (20c) we estimate

$$|r(0,2,n)| \le \frac{2^5}{720}12\cdot n^{-4} = \frac{8}{15}n^{-4};$$

for example, $|r(0,2,20)| \le 4\times 10^{-6}$. We calculate $s(0,2,20) = 0.882081 + \delta$, where $0 < \delta < 10^{-6}$, and by (20a) find

$$I = 0.882081 + \delta + r(0,2,20) = 0.882081 + \Delta,$$

where $-4\times 10^{-6} < \Delta \le 5\times 10^{-6}$. Thus $I = 0.8820\ldots$, correct to four places.

Numerical summation The E–M formula is also a notable tool for numerical summation. Let m and n be integers satisfying $1 \le m \le n$. Setting $a = m$, $b = n$, and $h = 1$ in (16a), we arrive at the basic E–M summation formula for a function $f \in C^p[1,\infty)$:

$$\sum_{k=m}^{n-1} f(k) = \int_m^n f(x)\,dx + \sum_{j=1}^{p}\frac{B_j}{j!}\left[f^{(j-1)}(x)\right]_m^n + \rho_p(m,n). \tag{21a}$$

By (16b) and (13) the remainder is given by

$$\rho_p(m,n) := r_p(m,n,n-m) = -\frac{1}{p!}\int_m^n P_p(-x)\; f^{(p)}(x)\,dx, \tag{21b}$$

and is estimated by

$$\left|\rho_p(m,n)\right| \le \frac{\mu_p}{p!}\int_m^n \left|f^{(p)}(x)\right|dx, \tag{21c}$$

where $\mu_p = \max\limits_{0\le x\le 1}\left|B_p(x)\right|$. Denoting

$$S(k) := \sum_{i=1}^{k} f(i) \quad \text{and} \quad \sigma_p(k) := \sum_{j=1}^{p}\frac{B_j}{j!}f^{(j-1)}(k) \tag{22}$$

for integers k, $p \ge 1$, we can write (21a), with $n \ge m \ge 1$, as

$$S(n) = S(m-1) + f(n) + \big[\sigma_p(n) - \sigma_p(m)\big] + \int_m^n f(x)\,dx + \rho_p(m,n), \qquad (23)$$

where $S(0) = 0$ by definition.

This equality is the basic summation tool derived from the E–M formula. We can use (23) to compute partial sums $S(n)$ if the integral $\int_m^n f(x)\,dx$ is easily computable and if we can adequately estimate the integral $\int_m^n \left|f^{(p)}(x)\right|\,dx$ in (21c) for positive integers m and n.

Let us work out the summation formula with $p = 3$. From (15) and (22), $\sigma_3(k) = -f(k)/2 + f'(k)/12$. Now (23) implies

$$S(n) = S(m-1) + \frac{f(m)+f(n)}{2} + \frac{f'(n)-f'(m)}{12} + \int_m^n f(x)\,dx + \rho_3(m,n), \qquad (23a)$$

where, by (18) and (21c),

$$|\rho_3(m,n)| \le \frac{1}{120}\int_m^n \left|f'''(x)\right|\,dx \qquad (23b)$$

for $m \le n$. (This explains the first formula used in the introduction.)

Euler's constant for a function For a $C^p[1,\infty)$ function f and any positive integer n, we consider the difference $\gamma_n := \sum_{k=1}^n f(k) - \int_1^n f(x)\,dx$. By (23),

$$\gamma_n = f(n) + \big[\sigma_p(n) - \sigma_p(1)\big] + \rho_p(1,n), \quad n \ge 1. \qquad (24a)$$

Let us assume from now on that finite limits $\lambda_0 := \lim_{n\to\infty} f(n)$ and $\lambda_k := \lim_{n\to\infty} f^{(k)}(n)$ exist for every positive integer $k \le p-1$ (the convergence is considered only in the sense of sequences). Let us also suppose that $\int_1^\infty \left|f^{(p)}(x)\right|\,dx < \infty$. By (21b) and (21c), this ensures the existence of the finite limit

$$\rho_p(m,\infty) := \lim_{n\to\infty} \rho_p(m,n) = -\frac{1}{p!}\int_m^\infty P_p(-x)\ f^{(p)}(x)\,dx \qquad (24b)$$

for every integer $m \ge 1$. Our assumptions imply, according to (24a), that the limit $\gamma := \lim_{n\to\infty}\gamma_n$ (called *Euler's constant for the function* f) exists and satisfies the equality

$$\gamma = \lambda_0 + \big[\sigma_p(\infty) - \sigma_p(1)\big] + \rho_p(1,\infty), \qquad (24c)$$

where

$$\sigma_p(\infty) := \sum_{j=1}^p \frac{B_j\lambda_{j-1}}{j!}.$$

(As $B_j = 0$ for odd $j \ge 3$, we could suppose above that limit λ_k exists only for $k = 0$ and all odd $k \le p-1$.) Comparing (24a) and (24c), we obtain, for any integer $n \ge 1$,

$$\gamma = \gamma_n + [\lambda_0 - f(n)] + \big[\sigma_p(\infty) - \sigma_p(n)\big] + \delta_p(n), \qquad (25a)$$

where $\delta_p(n) = -\frac{1}{p!}\int_n^\infty P_p(-x)\ f^{(p)}(x)\,dx$. Now (13) implies that

$$\left|\delta_p(n)\right| \le \frac{\mu_p}{p!}\int_n^\infty \left|f^{(p)}(x)\right|dx, \qquad (25b)$$

when $\mu_p = \max\limits_{0\le x\le 1}\left|B_p(x)\right|$.

Now (25a) and (25b) enable us to compute the Euler's constant γ for a function f; knowing its numerical value, we can compute partial sums $S(n)$. Namely, since $\gamma_n := S(n) - \int_1^n f(x)\,dx$, formula (25a) implies that, for $n \geq 1$,

$$S(n) = \gamma + \int_1^n f(x)\,dx + [f(n) - \lambda_0] + \left[\sigma_p(n) - \sigma_p(\infty)\right] - \delta_p(n). \qquad (26)$$

Example 3. To compute the *Euler–Mascheroni constant* γ^*, and to estimate the harmonic sum $H_n := \sum_{k=1}^n 1/k$ we use in (25a) the sum $\sigma_3(n) = -1/2n - 1/12n^2$ and limits $\lambda_k = 0$ for $k \geq 0$; thus $\sigma_3(\infty) = 0$. Now from (25a) we obtain, for $n \geq 1$,

$$\gamma^* = (H_n - \ln n) - \frac{1}{2n} + \frac{1}{12n^2} + \delta_3(n). \qquad (27a)$$

Using (25b) and (18) we estimate

$$|\delta_3(n)| \leq \frac{\mu_3}{3!}\int_n^\infty \left|f'''(x)\right| dx < -\frac{1/20}{6}\left[f''(x)\right]_n^\infty = \frac{1}{120}\cdot 2n^{-3} = \frac{1}{60n^3}. \qquad (27b)$$

(A more sophisticated approach results in the better estimate $-1/64n^4 \leq \delta_3(n) \leq 0$.) For example, $|\delta_3(100)| < 1.7 \times 10^{-8}$. Calculating $H_n - \ln n - 1/(2n) + 1/(12n^2)$ at $n = 100$ directly to nine places gives $0.577215664\ldots$. Therefore

$$0.577215647 < \gamma^* < 0.577215682; \qquad (27c)$$

that is, $\gamma^* = 0.577\,215\,6\ldots$, correct to seven places. To get more correct places we need only enlarge the parameter n or p in (25a). From (27a) and (27b) we obtain the asymptotic estimates

$$\gamma^* + \ln n + \frac{1}{2n} - \frac{1}{12n^2} - \frac{1}{60n^3} < H_n < \gamma^* + \ln n + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{60n^3}, \qquad (28)$$

for $n \geq 1$, which enable us to compute harmonic sums with high precision. For example, $H_{1000^{1000}} = \gamma^* + 3000\ln 10 + \delta$, where $|\delta| < 10^{-3000}$. According to (28) and (27c) this means $H_{1000^{1000}} = 6908.332\,494\,6\ldots$, correct to seven places.

We remark that $\ln n = \sum_{k=2}^n \ln\frac{k}{k-1}$ by the telescoping method, so

$$\gamma^* = \lim_{n\to\infty}\left(\sum_{k=1}^n \frac{1}{k} + \sum_{k=2}^n \ln\frac{k}{k-1}\right) = 1 + \sum_{k=2}^\infty g(k),$$

where $g(x) = \frac{1}{x} + \ln\frac{x}{x-1}$. So we could also compute γ^* using the theorem below on summation of convergent series. Because the k-th derivative $g^{(k)}(x)$ decreases to 0 faster than $f^{(k)}(x)$ as $x \to \infty$, this method proves better than the previous one.

Formulas (25a) and (25b) are also of theoretical interest. They imply a theorem comparing the convergence of a series $\sum_{k=1}^\infty f(k)$ and an integral $\int_1^\infty f(x)\,dx$. (This theorem, known as *the integral test*, is considered in many analysis textbooks only for *monotone* functions f.) More precisely, the definition of γ_n and formula (26) implies the following result:

THEOREM 2. *If $f \in C^p[1,\infty)$, $\int_1^\infty \left|f^{(p)}(x)\right| dx$ converges, and finite limits $\lambda_0 := \lim_{n\to\infty} f(n)$ and $\lambda_k := \lim_{n\to\infty} f^{(k)}(n)$ exist for all positive integers $k \leq p-1$, then*

(i) The series $\sum_{k=1}^\infty f(k)$ converges if and only if the sequence $n \mapsto \int_1^n f(x)\,dx$ converges.

(ii) If the series $\sum_{k=1}^{\infty} f(k)$ converges, then $\lambda_0 = 0$ and

$$\sum_{k=1}^{\infty} f(k) = \gamma + \lim_{n\to\infty} \int_1^n f(x)\,dx$$

$$= S(m-1) + \int_m^{\infty} f(x)\,dx + \left[\sigma_p(\infty) - \sigma_p(m)\right] + \delta_p(m),$$

where $\left|\delta_p(m)\right| \leq \frac{\mu_p}{p!}\int_m^{\infty}\left|f^{(p)}(x)\right|dx$ for $m \geq 1$.

Example 4. From Theorem 2 we deduce easily that the series $\sum_{k=1}^{\infty}(\sin\sqrt{k})/k$ and $\sum_{k=1}^{\infty}(\sin\sqrt{k})/\sqrt{k}$ converge, setting $p = 1$ for the first series and $p = 2$ for the second. (Here we apply the comparison test for absolute convergence of improper integrals. Note that the numerical computation of sums of these two series requires a larger p, say $p = 6$ or $p = 8$. We leave details to the reader.)

Example 5. To compute $\zeta(3) = \sum_{k=1}^{\infty} 1/k^3$ we put $p = 2$ and $f(x) := x^{-3}$. From (22) and (15) we find $\sigma_2(n) = -1/2n^3 - 1/4n^4$ and $\sigma_2(\infty) = 0$. Now part (ii) of Theorem 2 gives

$$\zeta(3) = \sum_{k=1}^{m-1}\frac{1}{k^3} + \int_m^{\infty}\frac{dx}{x^3} + \frac{1}{2m^3} + \frac{1}{4m^4} + \delta_2(m),$$

where, for $m \geq 1$,

$$|\delta_2(m)| \leq \frac{\mu_2}{2}\int_m^{\infty}\frac{12}{x^5}\,dx = \frac{1}{4m^4}.$$

This means

$$\zeta(3) = \sum_{k=1}^{m-1}\frac{1}{k^3} + \frac{1}{2m^2} + \frac{1}{2m^3} + \Delta(m),$$

where $0 \leq \Delta(m) \leq 1/\left(2m^4\right)$ for $m \geq 2$. Since, for example, $0 \leq \Delta(20) < 4 \times 10^{-6}$, we compute

$$\sum_{k=1}^{19}\frac{1}{k^3} + \frac{1}{2\times 20^2} + \frac{1}{2\times 20^3} = 1.202055\ldots$$

to obtain $\zeta(3) = 1.20205\ldots$, correct to five places. For higher precision we need only use higher values of m or p in (ii) of Theorem 2. (We could also compute $\zeta(3)$ by means of Euler's constant for the function $f(x) = 1/x^3$, applying (ii) of Theorem 2.)

Remark 1 The more general E–M formula

$$h\sum_{i=0}^{n-1} f(a+(i+\omega)h) - \int_a^b f(x)\,dx = \sum_{j=1}^{p} h^j\frac{B_j(\omega)}{j!}\left[f^{(j-1)}(x)\right]_a^b$$

$$-\frac{h^p}{p!}\int_a^b P_p\left(\omega - \frac{x-a}{h}\right) f^{(p)}(x)\,dx,$$

for every $\omega \in [0, 1]$, can be deduced from (1) in the same way as was done for the equality (16a). This E–M formula expresses the difference between the integral and the Riemann integral sum for a uniform partition of $[a, b]$ and the evaluation points $a + (i+\omega)h$, $i = 0, 1, \ldots, n-1$ (see, e.g., [**6**, pp. 51–54]).

Remark 2 The Bernoulli coefficients B_k can be found by formal manipulation of the formula $B_k = (1+B)^k$ for $k \geq 2$. The right side is meant to be expanded by the binomial theorem and then each power B^j is replaced by B_j. The Bernoulli polynomials can be described by applying the same process to the expression $B_k(x) = (x+B)^k$ for $k \geq 1$. Thus, this second formula is to be read

$$B_k(x) = \sum_{j=0}^{k} \binom{k}{j} B_{k-j} x^j,$$

where the recursion

$$B_k = -\frac{1}{k+1} \sum_{j=0}^{k-1} \binom{k+1}{j} B_j$$

for $k \geq 1$ follows formally from $B_{k+1} = (1+B)^{k+1}$. (See [**2**, p. 266], [**16**, p. 87].)

Remark 3 By means of Fourier analysis we obtain the expansion

$$\frac{P_p(x)}{p!} = -2 \sum_{k=1}^{\infty} \frac{\cos\left(2k\pi x - p\frac{\pi}{2}\right)}{(2k\pi)^p},$$

valid for all real x and every integer $p \geq 2$ [**5**, p. 135]. (At $p = 1$ this equality holds for every noninteger x.) From this equality we read an estimate for the so-called *Bernoulli numbers* $b_j := (-1)^{j+1} B_{2j}$:

$$2\frac{(2j)!}{(2\pi)^{2j}} < b_j < 4\frac{(2j)!}{(2\pi)^{2j}},$$

which holds for all $j \geq 1$. (Unfortunately, the name "Bernoulli numbers" is not standard in the literature.) This estimate has three consequences:

(a) Bernoulli coefficients alternate in sign, and are not bounded for even indices (as (15) may have suggested). (See [**8**, p. 452].)

(b) We find $\mu_p = \max\left\{\left|B_p(x)\right| : 0 \leq x \leq 1\right\} < 4\,(p!)\,(2\pi)^{-p}$; by (17), we estimate the remainder in the E–M formula as follows:

$$\left|r_p(a,b,n)\right| \leq 4\left(\frac{h}{2\pi}\right)^p \int_a^b \left|f^{(p)}(x)\right| dx, \qquad p \geq 1.$$

(c) The Bernoulli numbers increase very rapidly, so it is not possible, in general, to set $p = \infty$ in the E–M formula. In fact, the series

$$\sum_{j=1}^{\infty} \frac{B_j}{j!} \left[f^{(j-1)}(b) - f^{(j-1)}(a)\right]$$

turns out to diverge for almost all functions $f(x)$ that occur in applications, regardless of a and b [**11**, p. 525].

REFERENCES

1. T. M. Apostol, An elementary view of Euler's summation formula, *Amer. Math. Monthly* **106** (1999), 409–418.
2. T. M. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, NY, 1976.
3. R. P. Boas, Jr., Partial sums of infinite series, and how they grow, *Amer. Math. Monthly* **84** (1977), 237–258.
4. P. C. Chakravarti, *Integrals and Sums: Some New Formulae for their Numerical Evaluation*, The Athlone Press, London, UK, 1970.
5. P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, Chestnut Hill, MA, 1984.
6. T. Fort, *Finite Differences and Difference Equations in the Real Domain*, Clarendon Press, Oxford, UK, 1948.
7. A. O. Guelfond, *Calcul des différences finies*, Collection Universitaire. de Math., Dunod, Paris, France, 1963.
8. R. L. Graham, D. E. Knuth and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, Reading, MA, 1994.
9. P. Henrici, *Applied and Computational Complex Analysis*, Vol. 2, John Wiley and Sons, New York, NY, 1991.
10. C. G. Jacobi, De usu legitimo formulae summatoriae Maclaurinianae, *Journal für die reine und angew. Math.* **12** (1834), 263–272.
11. K. Knopp, *Theory and Applications of Infinite Series*, Hafner, New York, NY, 1971.
12. D. E. Knuth, *The Art of Computer Programming*, Vol. 1, Addison-Wesley, Reading, MA, 1997.
13. *Mathematica*, version 4.0, Wolfram Research, Inc., 1988–1999.
14. S. D. Poisson, Mémoire sur le calcul numérique des intégrales définies, *Mémoires de l'Académie Royale des Sciences de l'Institut de France*, séries 2, **6** (1823), 571–602.
15. G. C. Rota, Combinatorial snapshots, *Math. Intelligencer* **21** (1999), 8–14.
16. M. R. Spiegel *Schaum's Outline of Theory and Problems of Calculus of Finite Differences and Difference Equations*, McGraw–Hill, Blacklick, OH, 1971.
17. E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, UK, 1935.
18. S. Wolfram, *The Mathematica Book*, 4th ed., Cambridge University Press, Cambridge, UK, 1999.

Nobody alive has done more than Gardner to spread the understanding and appreciation of mathematics, and to dispel superstition. Nobody has worked harder or more steadily to defend and enlarge this little firelit clearing we hold in the dark chittering forest of unreason.

—from John Derbyshire's review of Martin Gardner's *Did Adam and Eve Have Navels?* in *The New Criterion*.

NOTES

Location of Incenters and Fermat Points in Variable Triangles

ANTHONY VÁRILLY
Harvard University
Cambridge, MA 02138

Introduction In 1765, Euler proved that several important centers of a triangle are collinear; the line containing these points is named after him. The incenter I of a triangle, however, is generally not on this line. Less than twenty years ago, Andrew P. Guinand discovered that although I is not necessarily on the Euler line, it is always fairly close to it. Guinand's theorem [**5**] states that for any non-equilateral triangle, the incenter lies inside, and the excenters lie outside, the circle whose diameter joins the centroid to the orthocenter, henceforth the orthocentroidal circle. Furthermore, Guinand constructed a family of locus curves for I which cover the interior of this circle twice, showing that there are no other restrictions on the positions of the incenter with respect to the Euler line.

Here we show that the Fermat point also lies inside the orthocentroidal circle; this suggests that the neighborhood of the Euler line may harbor more secrets than was previously known. We also construct a simpler family of curves for I, covering the interior once only except for the nine-point center N, which corresponds to the limiting case of an equilateral triangle.

Triangle geometry is often believed to be exhausted, although both Davis [**3**] and Oldknow [**6**] have expressed the hope that the use of computers may revive it. New results do appear occasionally, such as Eppstein's recent construction of two new triangle centers [**4**]. This article establishes some relations among special points of the triangle, which were indeed found by using computer software.

The locus of the incenter To any plane triangle ABC there are associated several special points, called *centers.* A few of these, in the standard notation [**1**], are: the centroid G, where the medians intersect; the orthocenter H, where the altitudes meet; the centers of the inscribed and circumscribed circles, called the incenter I and the circumcenter O; and the nine-point center N, halfway between O and H. The radii of the circumcircle and the incircle are called R and r, respectively.

If equilateral triangles BPC, AQC and ARB are constructed externally on the sides of the triangle ABC, then the lines AP, BQ and CR are concurrent and meet at the Fermat point T. This point minimizes the distance $TA + TB + TC$ for triangles whose largest angle is $\leq 120°$ [**2**].

The points O, G, N, and H lie (in that order) on a line called the Euler line, and $OG : GN : NH = 2 : 1 : 3$. They are distinct unless ABC is equilateral. The circle whose diameter is GH is called the *orthocentroidal circle*.

Guinand noticed that Euler's relation $OI^2 = R(R - 2r)$ [**1**, p. 85] and Feuerbach's theorem $IN = \frac{1}{2}R - r$ [**1**, p. 105] together imply that

$$OI^2 - 4\,IN^2 = R(R-2r) - (R-2r)^2 = 2r(R-2r) = \frac{2r}{R}\,OI^2 > 0.$$

Therefore, $OI > 2\,IN$. The locus of points P for which $OP = 2\,PN$ is a circle of Apollonius; since $OG = 2\,GN$ and $OH = 2\,HN$, this is the orthocentroidal circle. The inequality $OI > 2\,IN$ shows that I lies in the interior of the circle [**5**].

Guinand also showed that the angle cosines $\cos A$, $\cos B$, $\cos C$ of the triangle satisfy the following cubic equation:

$$\rho^4(1-2x)^3 + 8\rho^2\sigma^2 x(3-2x) - 16\sigma^4 x - 4\sigma^2\kappa^2(1-x) = 0, \qquad (1)$$

where $OI = \rho$, $IN = \sigma$ and $OH = \kappa$. We exploit this relationship below.

The relation $OI > 2\,IN$ can be observed on a computer with the software *The Geometer's Sketchpad*®, that allows tracking of relative positions of objects as *one* of them is moved around the screen. Let us fix the Euler line by using a Cartesian coordinate system with O at the origin and H at $(3, 0)$. Consequently, $G = (1, 0)$ and $N = (1.5, 0)$. To construct a triangle with this Euler line, we first describe the circumcircle $\odot(O, R)$, centered at O with radius $R > 1$—in order that G lie in the interior—and choose a point A on this circle. If AA' is the median passing through A, we can determine A' from the relation $AG : GA' = 2 : 1$; then BC is the chord of the circumcircle that is bisected perpendicularly by the ray OA'.

It is not always possible to construct ABC given a fixed Euler line and a circumradius R. If $1 < R < 3$ then there is an arc on $\odot(O, R)$ on which an arbitrary point A yields an A' outside the circumcircle, which is absurd. If U and V are the intersections of $\odot(O, R)$ with the orthocentroidal circle, and if UY and VZ are the chords of $\odot(O, R)$ passing through G, then A cannot lie on the arc ZY of $\odot(O, R)$, opposite the orthocentroidal circle (FIGURE 1). OYG and NUG are similar triangles. Indeed, $YO = UO = 2\,UN$ and $OG = 2\,GN$, so $YG : GU = 2 : 1$; in the same way, $ZG : GV = 2 : 1$. Hence, if $A = Y$ or $A = Z$, then $A' = U$ or $A' = V$ respectively, and ABC degenerates into the chord UY or VZ. If A lies on the arc ZY, opposite the orthocentroidal circle, A' will be outside $\odot(O, R)$.

Another formula of Guinand [**5**],

$$OH^2 = R^2(1 - 8\cos A\cos B\cos C), \qquad (2)$$

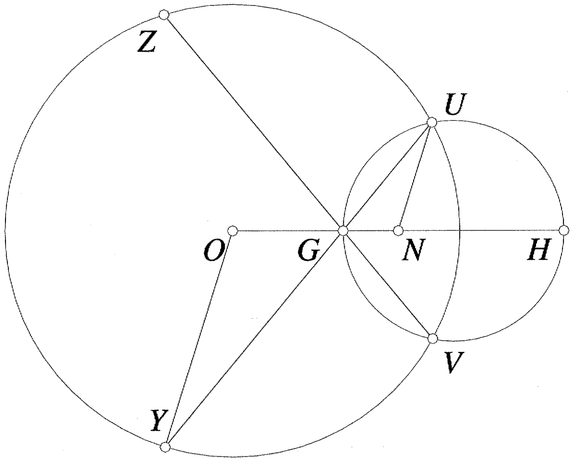


Figure 1 The forbidden arc ZY

shows that the forbidden arc appears if and only if $\cos A \cos B \cos C < 0$, that is, if and only if $1 < R < 3 = OH$. The triangle ABC is obtuse-angled in this case.

Once A is chosen and the triangle constructed, we can find I by drawing the angle bisectors of ABC and marking their intersection. *The Geometer's Sketchpad* will do this automatically, and we can also ask it to draw the locus that I traces as we move the point A around the circumcircle. The idea of parameterizing the loci with R, instead of an angle of the triangle (as Guinand did), was inspired by the drawing tools available in *The Geometer's Sketchpad*. This locus turns out to be a quartic curve, as follows.

PROPOSITION 1. *In the coordinate system described above, the incenter of ABC is on the curve*

$$(x^2 + y^2)^2 = R^2\,[(2x - 3)^2 + 4y^2]. \tag{3}$$

Proof. From equation (2) we see that, once we fix R, the product $\cos A \cos B \cos C$ is fixed. Using Viète's formulas, this product is obtained from the constant term of Guinand's equation (1):

$$\cos A \cos B \cos C = \frac{1}{8}\left(1 - \frac{4\sigma^2\kappa^2}{\rho^4}\right),$$

so that

$$\rho^4(1 - 8\cos A \cos B \cos C) = 4\sigma^2\, OH^2. \tag{4}$$

Now, I is a point of intersection of the two circles

$$\rho^2 = x^2 + y^2 \quad \text{and} \quad \sigma^2 = (x - \tfrac{3}{2})^2 + y^2.$$

We get (3) by substituting these and (2) in (4), and dividing through by the common factor $1 - 8\cos A \cos B \cos C$, which is positive since $O \neq H$ (the equilateral case is excluded). ∎

FIGURE 2 is a *Mathematica* plot of the curves (3) with the orthocentroidal circle.

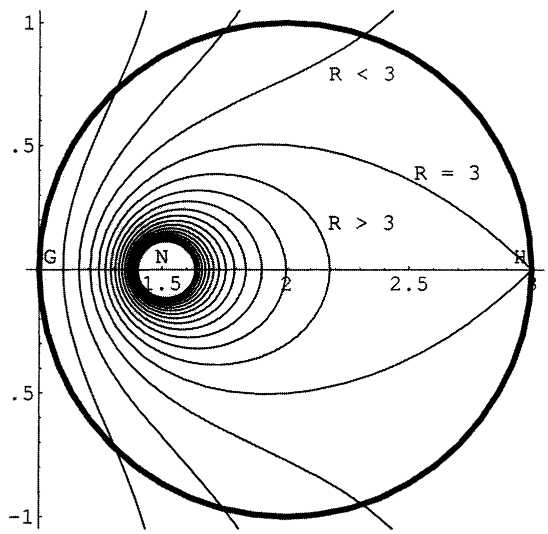


Figure 2 Locus curves for I with fixed R

Every point on a locus inside the orthocentroidal circle is the incenter of a triangle. When $R > 3$, the locus is a lobe entirely inside the circle; if the point A travels around the circumcircle once, then I travels around the lobe three times, since A will pass through the three vertices of each triangle with circumradius R. When $1 < R < 3$, the interior portion of (3) is shaped like a bell (FIGURE 2). Let A travel along the allowable arc from Y to Z, passing through V and U; then I travels along the bell from U to V, back from V to U, and then from U to V again. While A moves from V to U, the orientation of the triangle ABC is reversed. When $R = 3$, the locus closes at H and one vertex B or C also coincides with H; the triangle is right-angled. If A moves once around the circumcircle, starting and ending at H, I travels twice around the lobe.

PROPOSITION 2. *The curves (3), for different values of R, do not cut each other inside the orthocentroidal circle, and they fill the interior except for the point N.*

Proof. Let (a, b) be inside the orthocentroidal circle, that is,

$$(a-2)^2 + b^2 < 1. \tag{5}$$

If (a, b) also lies on one of the curves (3), then

$$R = \sqrt{\frac{(a^2+b^2)^2}{(2a-3)^2+4b^2}}.$$

There is only one positive value of R and thus *at most* one curve of the type (3) on which (a, b) can lie. Now we show (a, b) lies on *at least* one curve of type (3); to do that, we need to show that given (5), $R > 1$. We need only prove

$$(a^2+b^2)^2 > (2a-3)^2 + 4b^2. \tag{6}$$

Indeed, $(2a-3)^2 + 4b^2 = 0$ only if $(a, b) = (\frac{3}{2}, 0)$; this point is N. It cannot lie on a locus of the form (3); in fact, it corresponds to the limiting case of an equilateral triangle as $R \to \infty$.

The inequality (5) can be restated as

$$a^2 + b^2 < 4a - 3, \tag{7}$$

and (6) as

$$\left(a^2+b^2\right)^2 - 4a^2 + 3(4a-3) - 4b^2 > 0. \tag{8}$$

From (7) it follows that

$$\begin{aligned}(a^2+b^2)^2 - 4a^2 + 3(4a-3) - 4b^2 &> (a^2+b^2)^2 - 4a^2 + 3(a^2+b^2) - 4b^2\\ &= (a^2+b^2)(a^2+b^2-1).\end{aligned}$$

But $a^2 + b^2 > 1$ since (a, b) is inside the orthocentroidal circle; therefore, (6) is true. ∎

The whereabouts of the Fermat point The same set-up, a variable triangle with fixed Euler line and circumcircle, allows us to examine the loci of other triangle centers. Further experimentation with *The Geometer's Sketchpad*, as in FIGURE 3, suggests that the Fermat point T also lies inside the orthocentroidal circle in all cases.

THEOREM 1. *The Fermat point of any non-equilateral triangle lies inside the orthocentroidal circle.*

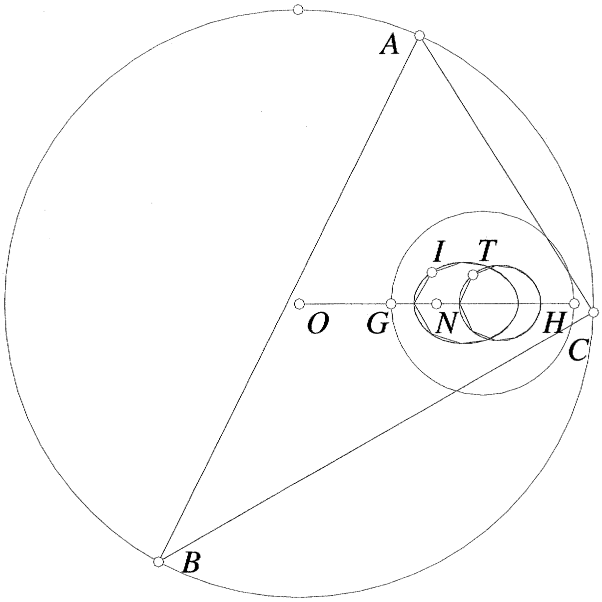


Figure 3 Locus of the Fermat point T

Proof. Given a triangle ABC with the largest angle at A, we can take a coordinate system with BC as the x-axis and A on the y-axis. Then $A = (0, a)$, $B = (-b, 0)$, $C = (c, 0)$ where a, b and c are all positive. Let BPC and AQC be the equilateral triangles constructed externally over BC and AC respectively (FIGURE 4). Then $P = (\frac{1}{2}(c-b), -\frac{1}{2}\sqrt{3}(b+c))$ and $Q = (\frac{1}{2}(\sqrt{3}\,a + c), \frac{1}{2}(a + \sqrt{3}\,c))$.

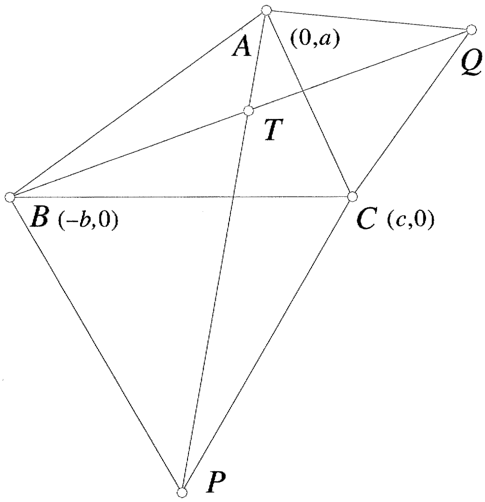


Figure 4 Finding the Fermat point

The coordinates of T can be found by writing down the equations for the lines AP and BQ and solving them simultaneously. After a little work, we get $T = \left(\frac{u}{d}, \frac{v}{d}\right)$, where

$$u = (\sqrt{3}\,bc - \sqrt{3}\,a^2 - ac - ab)(b - c),$$
$$v = (a^2 + \sqrt{3}\,ab + \sqrt{3}\,ac + 3bc)(b + c),$$
$$d = 2\sqrt{3}(a^2 + b^2 + c^2) + 6ac + 6ab + 2\sqrt{3}\,bc. \qquad (9)$$

The perpendicular bisectors of BC and AC intersect at the circumcenter $O = (\frac{1}{2}(c-b), (a^2 - bc)/2a)$. The nine-point center N is the circumcenter of the triangle whose vertices are the midpoints of the sides of ABC; we can deduce that $N = (\frac{1}{4}(c-b), (a^2 + bc)/4a)$.

To show that T lies inside the orthocentroidal circle, we must show that $OT > 2\,NT$, or

$$OT^2 > 4\,NT^2. \qquad (10)$$

In coordinates, this inequality takes the equivalent form:

$$\left[\frac{u}{d} - \left(\frac{c-b}{2}\right)\right]^2 + \left[\frac{v}{d} - \left(\frac{a^2 - bc}{2a}\right)\right]^2$$
$$> 4\left[\frac{u}{d} - \left(\frac{c-b}{4}\right)\right]^2 + 4\left[\frac{v}{d} - \left(\frac{a^2 + bc}{4a}\right)\right]^2,$$

or, multiplying by $(2ad)^2$,

$$[2au - ad(c-b)]^2 + [2av - d(a^2 - bc)]^2$$
$$> [4au - ad(c-b)]^2 + [4av - d(a^2 + bc)]^2.$$

After expanding and canceling terms, this simplifies to

$$4a^2du(c-b) + 4adv[(2a^2 + 2bc) - (a^2 - bc)] - 4a^2d^2bc > 12a^2u^2 + 12a^2v^2,$$

or better,

$$adu(c-b) + dv(a^2 + 3bc) - abcd^2 - 3au^2 - 3av^2 > 0. \qquad (11)$$

One way to verify this inequality is to feed the equations (9) into *Mathematica*, which expands and factors the left hand side of (11) as

$$2(b+c)(\sqrt{3}\,a^2 + \sqrt{3}\,b^2 + \sqrt{3}\,c^2 + \sqrt{3}\,bc + 3ab + 3ac)$$
$$(a^4 + a^2b^2 - 8a^2bc + a^2c^2 + 9b^2c^2).$$

The first three factors are positive. The fourth factor can be expressed as the sum of two squares,

$$(a^2 - 3bc)^2 + a^2(b-c)^2,$$

and could be zero only if $a^2 = 3bc$ and $b = c$, so that $a = \sqrt{3}\,b$. This gives an equilateral triangle with side $2b$. Since the equilateral case is excluded, all the factors are positive, which shows that (11) is true, and therefore (10) holds. ∎

Varying the circumradius R and the position of the vertex A with *The Geometer's Sketchpad* reveals a striking parallel between the behavior of the loci of T and those of I. It appears that the loci of T also foliate the orthocentroidal disc, never cutting each other, in a similar manner to the loci of I (FIGURE 2). The locus of T becomes a lobe when $R = 3$, as is the case with the locus of I. Furthermore, the loci of T close in on the center of the orthocentroidal circle as $R \to \infty$, just as the loci of I close in on N (FIGURE 2).

It is difficult to prove these assertions with the same tools used to characterize the loci of I, because we lack an equation analogous to (1) involving T instead of I. A quick calculation for non-equilateral isosceles triangles, however, shows that T can be anywhere on the segment GH except for its midpoint. This is consistent with the observation that the loci of T close in on the center of the orthocentroidal circle.

Consider a system of coordinates like those of FIGURE 4. Let $b = c$ so that ABC is an isosceles triangle. In this case, by virtue of (9), $T = (0, b/\sqrt{3})$. For this choice of coordinates, $G = (0, a/3)$ and $H = (0, b^2/a)$. T lies on the Euler line, which can be parametrized by $(1 - t)G + tH$ for real t. This requires that

$$(1 - t)\frac{a}{3} + t\,\frac{b^2}{a} = \frac{b}{\sqrt{3}},$$

for some real t. Solving for t, this becomes

$$t = \frac{a^2 - \sqrt{3}\,ab}{a^2 - 3b^2} = \frac{a}{a + \sqrt{3}\,b},$$

unless $a = \sqrt{3}\,b$. This case is excluded since ABC is not equilateral. Note that $t \to \frac{1}{2}$ as $a \to \sqrt{3}\,b$. Thus, t takes real values between 0 and 1 except for $\frac{1}{2}$, so T can be anywhere on the segment GH except for its midpoint.

Acknowledgment. I am grateful to Les Bryant and Mark Villarino for helpful discussions, to Joseph C. Várilly for advice and TEXnical help, and to Julio González Cabillón for good suggestions on the use of computer software. I also thank the referees for helpful comments.

REFERENCES

1. N. A. Court, *College Geometry*, Barnes & Noble, New York, 1952.
2. H. S. M. Coxeter, *Geometry Revisited*, Math. Assoc. of America, Washington, DC, 1967.
3. P. J. Davis, The rise, fall, and possible transfiguration of triangle geometry: A mini-history, *Amer. Math. Monthly* **102** (1995), 204–214.
4. D. Eppstein, Tangent spheres and triangle centers, *Amer. Math. Monthly*, **108** (2001), 63–66.
5. A. P. Guinand, Euler lines, tritangent centers and their triangles, *Amer. Math. Monthly* **91** (1984), 290–300.
6. A. Oldknow, Computer aided research into triangle geometry, *Mathl. Gazette* **79** (1995), 263–274.

Enumerating Row Arrangements of Three Species

MICHAEL AVIDON
Connecticut College
New London, CT 06320

RICHARD MABRY
PAUL SISSON
Louisiana State University in Shreveport
Shreveport, LA 71115

Introduction A classic problem in beginning finite or discrete mathematics courses reads as follows: *A photographer wants to arrange N men and N women in a line so that men and women alternate. In how many ways can this be done?*

This problem nicely illustrates the use of factorials and has a simple solution, though students often neglect the factor of two in the answer $2(N!)^2$. This omission can be instructive, as it leads naturally to generalizations of the problem: How does the answer change if there are N men and $N-1$ women? What if men outnumber women by 2 or more? What if a sexist photographer insists that the lineup start with a man?

These variations are all easily dealt with, and illustrate some of the possible subtleties encountered in counting problems. Another variation of the problem is not, however, so easily handled. The setting (or sitting) must change slightly: *A pet photographer must pose C cats, D dogs, and E emus in a line, with no two animals of the same species adjacent. In how many ways can this be done?*

Skirting the issue of whether emus can in fact be considered pets, we quickly discover many interesting features in this variation of the problem. For example, C, D, and E can differ by more than 1 and still admit nonzero solutions. We also discover that the new problem is quite a bit harder than the original. Dating back to 1966 ([**6**]), it is often called *Smirnov's problem* and it has practical applications in areas such as queuing, transportation flow, sequential analysis, and multivariate order statistics (see, e.g., [**1, 5**]). The problem has been solved in a variety of ways (some of which we mention at the end of this note) but, unlike those methods, what is presented here is entirely elementary and could be discussed in a first course in discrete mathematics or combinatorics.

Two ways of counting Consider any *arrangement* of a row of C cats, D dogs and E emus, distinguishable only by species. Let $\mathcal{A}(C, D, E)$ be the set of all such arrangements. As an example, consider the member of $\mathcal{A}(4, 2, 3)$ represented by the string *cccdeeecd*. Obviously this arrangement is not of the sort we are trying to count, as some adjacent seats are occupied by the same species. On the other hand, the arrangement *cdcecedec* in $\mathcal{A}(4, 2, 3)$ is the kind we want. Let $\mathcal{L}(C, D, E)$ denote the set of all such *legal arrangements* in $\mathcal{A}(C, D, E)$. We need to find the cardinality of $\mathcal{L}(C, D, E)$, which we denote by $L(C, D, E)$. Now since these critters are pets, they should all be considered distinct! In that case we simply multiply $L(C, D, E)$ by $C!D!E!$ to account for the permuting of the animals within each species. But the crux of the problem is to find a formula for $L(C, D, E)$, and so we will henceforth, unless otherwise stated, neglect the individuality of the creatures within each species.

Now let $\mathcal{A}$ and $\mathcal{L}$ denote, respectively, the set of all arrangements and legal arrangements, of any size. Define a map from $\mathcal{A}$ into $\mathcal{L}$ as follows: For a given arrangement in $\mathcal{A}$, replace any run of identical species with a single member of that species (e.g., $cccdeeecd \to cdecd$). Then a member of $\mathcal{A}(C, D, E)$ will be mapped to some member of $\mathcal{L}(C', D', E')$, with $1 \le C' \le C$, $1 \le D' \le D$, and $1 \le E' \le E$. In particular, an arrangement is legal iff it is mapped to itself.

We now seek the number of members of $\mathcal{A}(C, D, E)$ that will be mapped to a fixed member of $\mathcal{L}(C', D', E')$. This is a familiar type of pigeonhole problem, as we are asking for the number of ways that C cats can be distributed among C' compartments, and similarly for the dogs and emus. Many readers will recognize that the answer is

$$\binom{C-1}{C'-1}\binom{D-1}{D'-1}\binom{E-1}{E'-1}. \tag{1}$$

For completeness we include the standard "stars and bars" proof.

LEMMA 1. *The number of ways of distributing N indistinguishable items among N' nonempty compartments is $\binom{N-1}{N'-1}$.*

Proof. Consider a line of N "stars". From among the $N-1$ spaces between the stars, there are $\binom{N-1}{N'-1}$ ways to choose positions for $N'-1$ "bars," resulting in an arrangement such as

$$* * ** \mid ** \mid \cdots * ** \mid *.$$

The $N'-1$ bars delineate N' nonempty "compartments", so the proof is complete. ■

Applying this lemma to each species in turn and invoking the multiplication principle produces formula (1).

Since $L(C', D', E')$ stands for the number of legal arrangements of C' cats, D' dogs, and E' emus, it is clear that

$$\binom{C-1}{C'-1}\binom{D-1}{D'-1}\binom{E-1}{E'-1} L(C', D', E')$$

gives us the number of ways that a chain of $C + D + E$ seats can be filled with C' runs of cats, D' runs of dogs, and E' runs of emus. We thus have two different formulas expressing the number of ways of arranging a chain of C, D, and E indistinguishable cats, dogs, and emus:

$$\sum_{C'=1}^{C}\sum_{D'=1}^{D}\sum_{E'=1}^{E}\binom{C-1}{C'-1}\binom{D-1}{D'-1}\binom{E-1}{E'-1} L(C', D', E') = \frac{(C+D+E)!}{C!\,D!\,E!}. \tag{2}$$

Inverting the formula Our goal is to invert (2) to determine $L(C, D, E)$. Fortunately, the inversion is easily accomplished due to the nature of the much-studied coefficients in front of the function L.

To begin, we define two matrices (of unspecified size, for the moment) $A = (a_{i,j})$ and $B = (b_{i,j})$, where

$$a_{i,j} = \binom{i-1}{j-1} \quad \text{and} \quad b_{i,j} = (-1)^{i-j} a_{i,j}.$$

(As usual, $\binom{0}{0} = 1$ and $\binom{m}{n} = 0$ if $m < n$.) Although A and B are not necessarily square, the following lemma demonstrates that the two matrices are "inverses" of one another, in a sense.

LEMMA 2. *Assume the sizes of A and B (resp. B and A) are such that the matrix product AB (resp. BA) is defined. Then $\sum_j a_{i,j}b_{j,k} = \delta_{i,k}$ (resp. $\sum_j b_{i,j}a_{j,k} = \delta_{i,k}$).*

Proof. Let n equal the number of columns of A (and hence the number of rows of B). We will prove that $\sum_j a_{i,j}b_{j,k} = \delta_{i,k}$. (The proof of the second claim is similar.)

Let $C = AB$, and note that if we let $C = (c_{i,k})$, then $c_{i,k} = 0$ for $i < k$. For $i \geq k$, observe that

$$c_{i,k} = \sum_{j=1}^{n} a_{i,j}b_{j,k} = \sum_{j=1}^{n} \binom{i-1}{j-1}\binom{j-1}{k-1}(-1)^{j-k}$$

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

where we have used the identity

$$\binom{n}{k}\binom{k}{m} = \binom{n}{m}\binom{n-m}{n-k}$$

and the fact that many of the summands are 0. If $i = k$, then the sum in the last line (and hence the entire expression) is 1, while otherwise it follows from the binomial theorem that the sum is $(1-1)^{i-k} = 0$. This completes the proof. ■

Note, then, that if f and g are arbitrary functions defined on the natural numbers, each of the equations $f(i) = \sum_j a_{i,j}\ g(j)$ and $g(i) = \sum_j b_{i,j}\ f(j)$ implies the other. A proof of one implication follows, with the other proved in identical fashion. Assuming the former equation holds, we have

$$\sum_j b_{i,j}\ f(j) = \sum_j b_{i,j}\sum_k a_{j,k}\ g(k) = \sum_k g(k)\sum_j b_{i,j}\ a_{j,k} = \sum_k g(k)\ \delta_{i,k} = g(i).$$

Moreover, we can repeat the process, so that if f and g are two arbitrary functions defined on triples of natural numbers, we have

$$f(i,j,k) = \sum_{i',j',k'} a_{i,i'}a_{j,j'}a_{k,k'}g(i',j',k') \iff g(i,j,k)$$

$$= \sum_{i',j',k'} b_{i,i'}b_{j,j'}b_{k,k'}f(i',j',k'). \tag{3}$$

We are now ready to isolate the function L from equation (2).

Theorem.

$$L(C,D,E) = \sum_{C'=1}^{C}\sum_{D'=1}^{D}\sum_{E'=1}^{E}(-1)^{C+D+E-C'-D'-E'}$$

$$\times\binom{C-1}{C'-1}\binom{D-1}{D'-1}\binom{E-1}{E'-1}\frac{(C'+D'+E')!}{C'!\ D'!\ E'!}.$$

Proof. The formula follows immediately upon applying the observation in equation (3) to equation (2). Note that we are, in effect, "peeling off" the binary coefficients one at a time from the left-hand side of (2), with the result that they appear on the right-hand side along with the appropriate power of -1.

This finally reveals the photographer's conundrum. Once we acknowledge the individuality of the pets, the number of different seatings of the C cats, D dogs, and E emus is $L(C, D, E)C!\ D!\ E!$. ■

Just how bad is the situation for the photographer? We list below some values of special interest, namely, when $C = D = E$. Asymptotic estimates for L are known [**1**, **5**]. In particular, $L(N, N, N)$ is asymptotic to a constant multiple of $\frac{8^N}{N}$.

N	1	2	3	4	5	6	7	8	9
$L(N, N, N)$	6	30	174	1092	7188	48852	339720	2403588	17236524

Some sophisticated methods Having presented an elementary approach to the problem, we offer a few tantalizing glimpses at some methods known to specialists. (Our problem can be viewed in many other frameworks. For example, $L(C, D, E)$ is the number of different ways a labeled chain of $C + D + E$ vertices can be properly colored (in the usual graph-theoretic sense) with three colors 1, 2, and 3, each used C, D, and E times, respectively. Or it can be viewed as the number of *words* that can be formed from the *alphabet* $\{c, d, e\}$, without adjacencies, and with the letter c being used C times, etc.)

1. E. Rodney Canfield alerted the authors to a technique called the "transfer matrix method" (see, e.g., [**3**] or [**7**]), the result of which is that $L(C, D, E)$ is the coefficient of $x^C y^D z^E$ in the matrix product

$$[x\ y\ z]\begin{bmatrix} 0 & y & z \\ x & 0 & z \\ x & y & 0 \end{bmatrix}^{C+D+E-1}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

2. Ira Gessel informed the authors of another approach [**4**, p. 69]: $L(C, D, E)$ is the coefficient of $x^C y^D z^E$ in the power series expansion of

$$\left(1 - \frac{x}{x+1} - \frac{y}{y+1} - \frac{z}{z+1}\right)^{-1}.$$

3. Ira Gessel also related the problem to *rook polynomials* [**2**, pp. 160–162]. To use this approach, begin by defining

$$L_n(x) = \sum_{k=0}^{n-1}(-1)^k \binom{n}{k}\binom{n-1}{k} k!\, x^{n-k}$$

and let Φ be the linear functional (defined on the usual vector space of polynomials) that sends x^m to $m!$. Then

$$\Phi(L_C(x)L_D(x)L_E(x)) = L(C, D, E)\ C!\ D!\ E!.$$

It is also instructive to apply symbolic software, such as *Mathematica*, to explore all four methods of calculating $L(C, D, E)$. Be forewarned, however, that the first two methods above will require significant computing time for all but small values of C, D, and E.

Generalizations The problem can, of course, be generalized. In one direction, each of the formulas presented for $L(C, D, E)$ can easily be extended to handle more species. For example, the number of legal arrangements of C cats, D dogs, E emus, and F frogs is

$$L(C, D, E, F) = \sum_{C'=1}^{C} \sum_{D'=1}^{D} \sum_{E'=1}^{E} \sum_{F'=1}^{F} (-1)^{C+D+E+F-C'-D'-E'-F'} \times \binom{C-1}{C'-1}\binom{D-1}{D'-1}\binom{E-1}{E'-1}\binom{F-1}{F'-1}\frac{(C'+D'+E'+F')!}{C'!\,D'!\,E'!\,F'!}.$$

The formula can also be restricted to the original simple problem. If only cats and dogs are present, the obvious changes lead to a double summation formula for $L(C, D)$. Of course, we already knew $L(C, D)$: it must be either two, one, or zero, depending on whether $C = D$, $|C - D| = 1$, or $|C - D| \geq 2$, respectively. It is an interesting exercise to verify that the formula in the case of two species really does simplify in this manner.

Further generalizations are left to the reader to explore. For example, one of our students asked for the number of arrangements such that no member of a species is trapped between two other members of the same species. Or perhaps the photographer will permit blocks of up to k members of a given species, but no more. Other variations might concern circular arrangements or two-dimensional grid arrangements of seats. (The latter problem belongs to a class of more general problems concerning the enumeration of n-colorings of labeled graphs with specified numbers of each color, already very difficult for $n = 2$.)

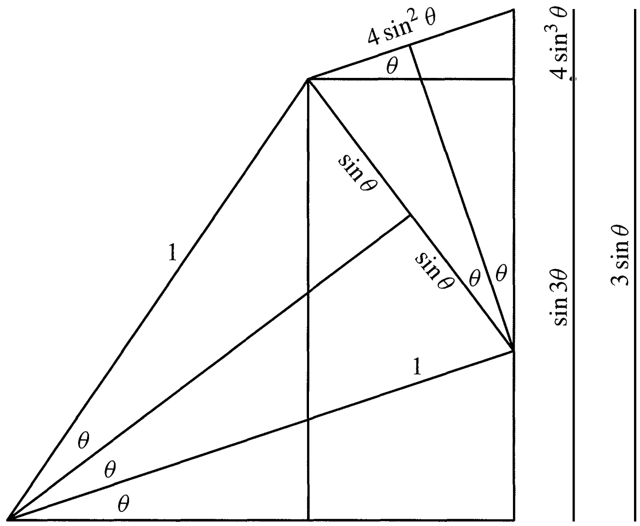
Acknowledgment. The authors thank E. Rodney Canfield, Ira Gessel, Carl Pomerance, and an anonymous reviewer for their help in relating the problem to standard enumeration (emu-neration?) techniques, and for providing references to those methods.

REFERENCES

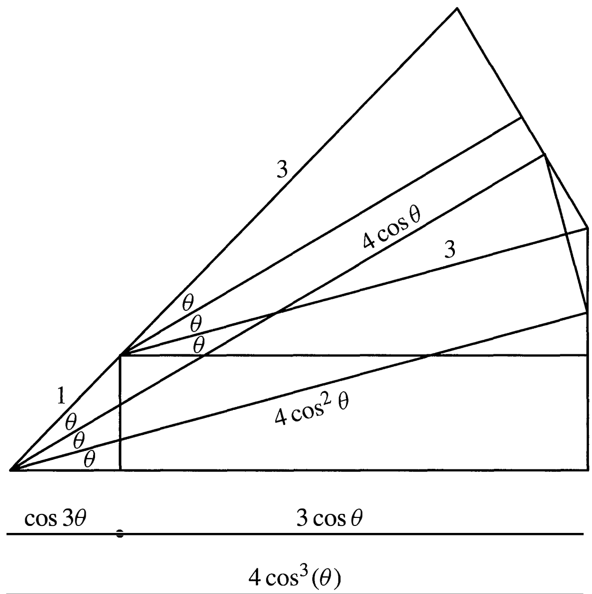
1. R. Alter and B. P. Lientz, Applications of a generalized combinatorial problem of Smirnov, *Naval Res. Logist.* **16** (1969), 543–546.
2. I. M. Gessel, Generalized Rook Polynomials and Orthogonal Polynomials, in *q-Series and Partitions*, IMA Volumes in Mathematics and its Applications 18, Springer-Verlag, New York, 1989, 159–176.
3. I. M. Gessel and R. P. Stanley, Algebraic enumeration, in *Handbook of Combinatorics*, ed. R. L. Graham, M. Grotschel, and L. Lovasz, Elsevier (North-Holland), Amsterdam, Holland, 1995, pp. 1021–1061.
4. I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, Wiley-Interscience, New York, 1983.
5. O. V. Sarmanov and V.K. Zaharov, A combinatorial problem of Smirnov, *Soviet Math. Dokl.* **9** (1967), 1147–1150.
6. N. V. Smirnov, O. V. Sarmanov, and V. K. Zaharov, A combinatorial problem of Smirnov, *Soviet Math. Dokl.* **7** (1966), 563.
7. R. P. Stanley, *Enumerative Combinatorics*, Vol. 1, 2nd ed. (Cambridge Studies in Advanced Mathematics 49), Cambridge University Press, Cambridge, UK, 1997.

Proof Without Words: The Triple-Angle Formulas for Sine and Cosine

SHINGO OKUDA
Tsu High School
Tsu-shi Mie-ken, Japan



$$\sin(3\theta) = 3\sin(\theta) - 4\sin^3(\theta)$$



$$\cos(3\theta) = 4\cos^3(\theta) - 3\cos(\theta)$$

The Disadvantage of Too Much Success

MARTIN L. JONES
College of Charleston
Charleston, SC 29424

REGINALD KOO
University of South Carolina Aiken
Aiken, SC 29801

Introduction Too much success can be a disadvantage. A gambler playing blackjack in a casino may find himself out in the street after enjoying a string of good fortune. Viruses that replicate too rapidly may kill their host before the host is able to infect other members of the population. In this note we consider three coin-tossing models in which "too much success" is defined by the occurrence of success runs of length r, where r is a positive integer. When such a success run occurs, play terminates and the player has no further opportunities to earn rewards. Three models are considered; in each a sequence of independent coin tosses occurs with a fixed success probability p. The objective is to choose the success probability p so as to maximize the expected reward before the stopping criterion applies.

Model 1 In this model the player earns one unit for each success and nothing for failures. The sequence of observations terminates as soon as r consecutive successes occur. Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed (i.i.d.) random variables with $P(X_n = 1) = p = 1 - P(X_n = 0)$, and let r be a positive integer. A stopping time for the sequence is a positive, integer-valued (or possibly infinite) random variable T such that for each positive integer n the event $\{T = n\}$ is independent of $X_{n+1}, X_{n+2}, \ldots$. If we define $T = \min\{n \mid X_n = X_{n-1} = \cdots = X_{n-r+1} = 1\}$ when such an n exists and $T = \infty$ otherwise, then T is a stopping time for the sequence. The reward function R is the sum of the stopped observations: $R = \sum_{n=1}^{T} X_n$. The objective is to choose the success probability p of the coin so as to maximize $E(R)$, the expected value of the reward function.

The expected value of a sum of random variables is the sum of the individual expectations. *Wald's equation* [**2**, p. 105] asserts that this also holds when the number T of i.i.d. random variables in the sum is random, provided that the expectations $E(X_n)$ and $E(T)$ are all finite. We will soon show that $E(T) < \infty$ for $0 < p \leq 1$. Since $E(X_n) < \infty$ for all n, Wald's equation yields

$$E(R) = E\left(\sum_{n=1}^{T} X_n\right) = E(X_1)E(T) = pE(T). \tag{1}$$

For $p = 0$, play never terminates, but the reward is zero. If $p = 1$, then play terminates after exactly r observations and the reward is r. For $0 < p < 1$ we will calculate $E(T)$ by conditioning on L, the length of the initial success string: $L = 0$ if $X_1 = 0$, and for $k \geq 1$, $L = k$ if $X_1 = X_2 = \cdots = X_k = 1$ and $X_{k+1} = 0$. The idea is to express $E(T)$ as a weighted average involving $E(T \mid L = k)$, the expected value of T given that $L = k$. Then, using the independence of the observations, we express $E(T \mid L = k)$ in terms of $E(T)$, thus obtaining an expression for $E(T)$.

The random variable L has a geometric distribution with $P(L = k) = p^k(1 - p)$ and $P(L \geq k) = p^k$ for $k \geq 0$. Conditioning on L, we obtain

$$\begin{aligned} E(T) &= E(E(T \mid L)) = \sum_{k=0}^{\infty} E(T \mid L = k) P(L = k) \\ &= \sum_{k=0}^{r-1} [(E(T) + k + 1) p^k (1 - p)] + r p^r \\ &= (1 - p^r) E(T) + (1 + p + \cdots + p^{r-1}). \end{aligned} \tag{2}$$

Solving (2) for $E(T)$ we obtain

$$E(T) = \frac{1 + p + \cdots + p^{r-1}}{p^r}, \tag{3}$$

and substituting (3) into (1) yields

$$E(R) = \frac{1 + p + \cdots + p^{r-1}}{p^{r-1}}. \tag{4}$$

For $r = 1$, $E(R) \equiv 1$ over the interval $0 < p < 1$. For fixed $r \geq 2$, $E(R)$ is strictly decreasing over the interval $0 < p < 1$. Therefore, for this reward function and stopping criterion, there is no optimal value of p to maximize $E(R)$. However, since for $r \geq 2$,

$$\lim_{p \to 0^+} \frac{1 + p + \cdots + p^{r-1}}{p^{r-1}} = \infty,$$

the expected reward can be made arbitrarily large by choosing p sufficiently close to zero. Intuitively, if there is a positive probability of success then successes will occur infinitely often provided that no stopping criterion is invoked. The objective is to choose p sufficiently small so that play will not terminate too soon. If a restriction is imposed, say that play terminates after N observations if it has not terminated earlier, then selecting p close to zero is no longer advantageous. In fact, if the stopping time is $\min\{T, N\}$, then from Wald's equation we obtain

$$E(R) = pE(\min\{T, N\}) \leq \min\{pE(T), pN\}. \tag{5}$$

For example, if $r = 10$ and $N = 1000$ then the maximum value of $\min\{pE(T), pN\}$ is approximately 540 and occurs with $p \approx .54$. While the bound given in (5) is crude, it does show the effect of terminating play after N observations on $E(R)$ and the choice of p. Exact calculations of $E(\min\{T, N\})$ are difficult, and even good approximations to the distribution of T are unwieldy [**1**, pp. 322–328]. We summarize the analysis of Model 1 in the following theorem.

THEOREM 1. *For Model 1,*

$$E(R) = \frac{1 + p + \cdots + p^{r-1}}{p^{r-1}}.$$

For $r = 1$, $E(R) \equiv 1$, and for $r \geq 2$, $E(R)$ is unbounded and strictly decreasing on $0 < p \leq 1$. No best success probability exists, but $E(R)$ can be made arbitrarily large by choosing p close to zero. If play is limited to N observations, then $E(R)$ is bounded above by $\min\{pN, (1 + p + \cdots + p^{r-1})/p^{r-1}\}$.

Model 2 Here the player earns one unit for each success and loses one unit for each failure. Play terminates as soon as r consecutive successes occur. Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with $P(X_n = 1) = p = 1 - P(X_n = -1)$. As in Model 1, let r be a positive integer, let $T = \min\{n \mid X_n = X_{n-1} = \cdots = X_{n-r+1} = 1\}$ if such an n exists and $T = \infty$ otherwise, and let $R = \sum_{n=1}^{T} X_n$. In Model 1, a wise strategy for maximizing $E(R)$ was to choose p close to zero. In Model 2, since losses occur with each failure, we expect that p should be chosen to be larger than 1/2 to ensure a positive expected reward. In fact, $E(R) \approx -6.2 \times 10^{27}$ for $r = 100$ and $p = .499$.

Since the distribution of T is the same as in Model 1, Wald's equation implies

$$E(R) = E(X_1)E(T) = (2p - 1)\left(\frac{1 + p + \cdots + p^{r-1}}{p^r}\right). \tag{6}$$

For $p < 1/2$ the expected reward is negative, for $p = 1/2$ it is zero, and for $p > 1/2$ the expected reward is positive. As before, for $p = 1$ play terminates in exactly r observations and $E(R) = r$. We wish, therefore, to maximize the expression given in (6) over the range $1/2 \leq p \leq 1$. For $r = 1, 2$ the best success probability is clearly $p = 1$, so let $r \geq 3$ be fixed. Differentiating (6) gives

$$\begin{aligned}\frac{d\,E(R)}{dp} &= \frac{p(1 + p + \cdots + p^{r-1}) + r(1 - 2p)}{p^{r+1}(1 - p)} \\ &= \frac{(p - 1)(p^{r-1} + 2p^{r-2} + \cdots + (r - 1)p - r)}{p^{r+1}(1 - p)}.\end{aligned} \tag{7}$$

For $r \geq 3$, the right side of (7) has a unique positive root in (0, 1]. (This follows from the intermediate value theorem and the fact that $p^{r-1} + 2p^{r-2} + \cdots + (r - 1)p - r$ is increasing in p.) By considering the sign changes of (7) we see that $E(R)$ attains its maximum at these roots, which are therefore the best success probabilities for the coin. For $r = 3$, this root is one. For $r > 3$, the sequence of roots $\{p_r\}$ can be shown by induction to be strictly decreasing with limit 1/2. Numerical approximations for some of these best success probabilities and the corresponding values of $E(R)$ are given in Table 1.

TABLE 1: Best success probabilities and $E(R)$

r	3	4	5	10	20	100	1000
p_r	1	.776	.684	.565	.528	.505	.5005
$E(R)$	3	4.2	6.6	90	41838	9.5×10^{27}	7.9×10^{297}

We summarize the analysis of Model 2 as follows:

THEOREM 2. *In Model 2,*

$$E(R) = (2p - 1)\left(\frac{1 + p + \cdots + p^{r-1}}{p^r}\right).$$

For $r = 1, 2, 3$ *the best success probability is* $p = 1$. *For* $r > 3$ *the best success probability* p_r *is the unique root in* (0, 1] *of* $p^{r-1} + 2p^{r-2} + \cdots + (r - 1)p = r$. *The sequence* $\{p_r \mid r > 3\}$ *of best success probabilities is strictly decreasing with limit* 1/2.

Model 3 The player earns one unit for each success and nothing for failures. Play terminates as soon as r consecutive successes or k consecutive failures occur. Let $X_1, X_2, X_3, \ldots$ be, as in Model 1, i.i.d. random variables with $P(X_n = 1) = p = 1 - P(X_n = 0)$. Let r and k be positive integers, and define $T_s = \min\{n \mid X_n = X_{n-1} = \cdots = X_{n-r+1} = 1\}$ if such an n exists, and $T = \infty$ otherwise, and define $T_f = \min\{n \mid X_n = X_{n-1} = \cdots = X_{n-k+1} = 0\}$ if such an n exists, and $T_f = \infty$ otherwise. Thus T_s is the first time that r consecutive successes are observed; T_f is the first time that k consecutive failures are observed. Let $T = \min\{T_s, T_f\}$ and $R = \sum_{n=1}^{T} X_n$. The cases $p = 0$ and $p = 1$ are easy to analyze, so let $0 < p < 1$. By Wald's equation,

$$E(R) = E(X_1)E(T) = pE(T). \tag{8}$$

We calculate $E(T)$ (using the method appearing in [**2**, pp. 125–128]) by writing

$$E(T_s) = E(T) + E(T_s - T) = E(T) + E(T_s - T \mid T = T_f)P(T = T_f). \tag{9}$$

But $E(T_s - T \mid T = T_f)$ is just the expected additional time until r consecutive successes occur given that we have just observed k consecutive failures. By independence of the observations, this is

$$E(T_s - T \mid T = T_f) = E(T_s). \tag{10}$$

Substituting (10) into (9) and solving for $E(T)$ yields

$$E(T) = E(T_s)(1 - P(T = T_f)). \tag{11}$$

Similarly we can show that

$$E(T) = E(T_f)P(T = T_f). \tag{12}$$

Eliminating $P(T = T_f)$ from (11) and (12) yields

$$E(T) = \frac{E(T_s)E(T_f)}{E(T_s) + E(T_f)}. \tag{13}$$

Since T_s is the T in Model 1, we have $E(T_s) = (1 + p + \cdots + p^{r-1})/p^r$ and similarly $E(T_f) = (1 + q + \cdots + q^{k-1})/q^k$, where $q = 1 - p$. Substituting these two expressions into (13) and then substituting the result into Wald's equation (8) yields

$$E(R) = p\left(\frac{(1 + p + \cdots + p^{r-1})(1 + q + \cdots + q^{k-1})}{q^k(1 + p + \cdots + p^{r-1}) + p^r(1 + q + \cdots + q^{k-1})}\right). \tag{14}$$

The expression in (14) does not yield easily to analytical methods, but graphs of $E(R)$ for various values of r and k are given in FIGURE 1 and the best success probabilities and corresponding $E(R)$ are calculated numerically and displayed in Table 2.

TABLE 2: Best success probabilities and $E(R)$

r	3	3	3	3	4	5	15
k	15	5	4	3	3	3	3
$p_{r,k}$	.242	.477	.546	.650	.664	.683	.801
$E(R)$	16.41	5.74	4.85	4.02	6.19	8.78	58.20

We summarize the analysis of Model 3 as follows:

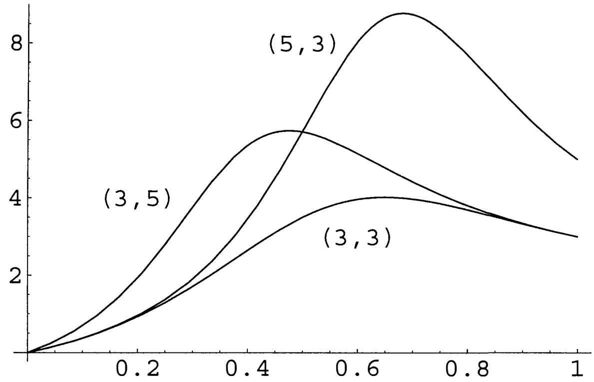


Figure 1 $E(R)$ as a function of p for several values of (r, k)

THEOREM 3. *In Model 3,*

$$E(R) = p\left(\frac{(1+p+\cdots+p^{r-1})(1+q+\cdots+q^{k-1})}{q^k(1+p+\cdots+p^{r-1})+p^r(1+q+\cdots+q^{k-1})}\right),$$

where $q = 1 - p$.

REFERENCES

1. William Feller, *An Introduction to Probability and Its Applications, Volume 1*, Third Edition, John Wiley, New York, 1968.
2. Sheldon M. Ross, *Stochastic Processes*, Second Edition, John Wiley, New York, 1996.

The Wily Hunting of the Proof

This Snark was a Boojum*, no doubt about *that*.
First glimpsed by Fermat in long-ago France.
It eluded pursuers till they fell down face flat,
Or wound up in cells with St. Vitus's Dance.

But one hunter of game, quite shrewd (note his name),
And possessed of persistence exceeding all norms,
Followed its swerves in elliptic curves
Till it finally nested in Modular Forms.

He picked up its spoor and took up the chase
Till he ran it to earth in this Nippon prefecture;
Then he brought it to light, having snared it in place
With a proof of a part of the T–S Conjecture.

**The Hunting of the Snark*, Lewis Carroll

——J. D. Memory
North Carolina State University
Raleigh, NC 27695

Means to an End

RICHARD P. KUBELKA
San Jose State University
San Jose, CA 95192-0103

Introduction It is well known that $\lim_{n\to\infty} \frac{\sqrt[n]{n!}}{n} = \frac{1}{e}$; the simplest proof uses $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$, Stirling's approximation for $n!$ (see, e.g., [**2**, pp. 616–618]). For a one-page derivation of Stirling's formula, see [**4**].

In showing a student a more elementary derivation of this result recently, I noticed that the numerator is the geometric mean of the first n positive integers, and the denominator, while not exactly the corresponding arithmetic mean, $(n+1)/2$, is pretty close to twice that much when n is large. In other words, we have the curious result:

$$\lim_{n\to\infty} \frac{GM(1, 2, \ldots, n)}{AM(1, 2, \ldots, n)} = \frac{2}{e}. \tag{1}$$

In fact, a more general result holds:

Theorem. For any positive real number s,

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

Equation (1) is, of course, the special case $s = 1$. Notice too that the arithmetic–geometric mean inequality [**3**, p. 17] implies that if the limit in question exists, it cannot exceed 1.

Proof. It suffices to show that

$$\ln GM(1^s, 2^s, \ldots, n^s) - \ln AM(1^s, 2^s, \ldots, n^s) \to \ln(s+1) - s \tag{2}$$

as n tends to infinity, and for this we will employ a Riemann sum argument. The discussion of the integral test in a calculus book (e.g., [**1**, pp. 533–536]) gives the basic idea. For the integral test itself one considers functions that are positive and decreasing. The basic sum-vs.-integral inequalities we need,

$$\int_c^n h(x)\,dx + h(c) \le \sum_{k=c}^{n} h(k) \le \int_c^n h(x)\,dx + h(n), \tag{3}$$

can be developed readily for non-negative *increasing* functions and integers c and n.

Now we apply (3), first with $h(x) = \ln x^s = s \ln x$ and then with $h(x) = x^s$ (both functions are non-negative and increasing on $[1, \infty)$), to get

$$\frac{s}{n}\int_1^n \ln x\,dx \le \frac{1}{n}\sum_{k=1}^{n} \ln k^s \le \frac{s}{n}\int_1^n \ln x\,dx + \frac{s}{n}\ln n; \tag{4}$$

$$\frac{1}{n}\int_1^n x^s\,dx + \frac{1}{n} \le \frac{1}{n}\sum_{k=1}^{n} k^s \le \frac{1}{n}\int_1^n x^s\,dx + \frac{1}{n}n^s. \tag{5}$$

Applying the logarithm to (5) gives

$$\ln\left[\frac{1}{n}\int_1^n x^s\,dx + \frac{1}{n}\right] \le \ln\left[\frac{1}{n}\sum_{k=1}^{n} k^s\right] \le \ln\left[\frac{1}{n}\int_1^n x^s\,dx + \frac{1}{n}n^s\right]. \tag{6}$$

Subtracting (6) from (4) gives

$$\begin{aligned} s\ln n - s + \frac{s}{n} - \ln\left[\frac{1}{s+1}n^s - \frac{1}{s+1}n^{-1} + n^{s-1}\right] &\le \frac{1}{n}\sum_{k=1}^{n}\ln k^s - \ln\left[\frac{1}{n}\sum_{k=1}^{n} k^s\right] \\ &\le s\ln n - s + \frac{s}{n} + \frac{s}{n}\ln n - \ln\left[\frac{1}{s+1}n^s - \frac{1}{s+1}n^{-1} + n^{-1}\right], \end{aligned}$$

which the laws of logarithms allow us to rewrite as

$$\begin{aligned} &\ln(s+1) - s + \frac{s}{n} - \ln\left[1 - n^{-s-1} + (s+1)n^{-1}\right] \\ &\le \ln GM(1^s, 2^s, \ldots, n^s) - \ln AM(1^s, 2^s, \ldots, n^s) \\ &\le \ln(s+1) - s + \frac{s}{n} + \frac{s}{n}\ln n - \ln\left[1 + sn^{-s-1}\right]. \end{aligned} \tag{7}$$

Letting n tend to infinity and applying the squeeze theorem and continuity of the logarithm, we get (2). ∎

Remark As mentioned above, Stirling's approximation for $n!$ gives a quick proof of (1). We might ask whether, conversely, our Riemann sum arguments imply Stirling's formula. The answer is "no." Using (3) with $f(x) = \ln x$ we get

$$n\ln n - n + 1 \le \sum_{k=1}^{n}\ln k \le n\ln n - n + 1 + \ln n;$$

exponentiating and rearranging gives

$$e \le n!e^n n^{-n} \le e\,n. \tag{8}$$

But the squeeze play we pulled off to get (2) will not work here—the two ends of (8) do not approach one another. The best we can manage is a rough estimate for $n!$:

$$e\,n^n e^{-n} \le n! \le e\,n^{n+1}e^{-n}.$$

Acknowledgment. The author would like to thank the referees for their careful reading and useful suggestions.

REFERENCES

1. H. Anton, *Calculus*, 5th Ed., John Wiley & Sons, New York, 1995.
2. T. M. Apostol, *Calculus*, Vol. II, 2nd ed., John Wiley & Sons, New York, 1969.
3. G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1983.
4. D. Romik, Stirling's approximation for $n!$: the ultimate short proof? *Amer. Math. Monthly* **107** (2000), 556–557.

Visualizing Leibniz's Rule

MARC FRANTZ
Indiana University
Bloomington, IN 47405

Leibniz's rule for differentiating under the integral sign deals with functions of the form

$$A(x) = \int_0^{\beta(x)} f(x, y)\, dy. \tag{1}$$

If f and $f_x = \partial f/\partial x$ are continuous in a suitable region of the plane, and if β' is continuous over a suitable interval, Leibniz's rule says that A' is continuous, and

$$A'(x) = \int_0^{\beta(x)} f_x(x, y)\, dy + f(x, \beta(x))\beta'(x). \tag{2}$$

(A formal statement of the rule appears at the end of this note.) Most textbooks generalize Leibniz's rule to the case when the lower limit of integration in Equation (1) is also a function $\alpha(x)$. This is easily done, since

$$\int_{\alpha(x)}^{\beta(x)} f(x, y)\, dy = \int_0^{\beta(x)} f(x, y)\, dy - \int_0^{\alpha(x)} f(x, y)\, dy.$$

The reader can find various derivations and applications* of Leibniz's rule in advanced calculus texts (cf. **[2]**, **[3]**, **[4]**). But what interpretation can we give to the two terms—an integral and a product—on the right side of Equation (2)? An integral is the area under a curve, or between two curves, and a product of two factors can be the area of a rectangle. I will show that this is essentially the correct geometric interpretation, and give an informal derivation of Leibniz's rule which may appeal to visually oriented students.

First, we need a geometric interpretation of $A(x)$, and this is given in FIGURE 1.

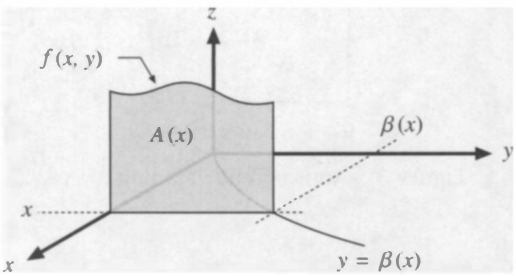


Figure 1 Geometric interpretation of $A(x)$

Here we imagine a surface $z = f(x, y)$ above the xy-plane, and a curve $y = \beta(x)$ lying in the xy-plane itself. If we now consider x fixed, then $A(x)$ as given in Equa-

*A recent application in this MAGAZINE **[1]** prompted me to think about the rule again.

tion (1) is just the area of the lamina (shaded in FIGURE 1) beneath the curve $z(y) = f(x, y)$.

The next step (FIGURE 2) is to consider a corresponding lamina at a new x-coordinate $x + \Delta x$. I have depicted this lamina, whose area is $A(x + \Delta x)$, as being larger than the first, with the area $A(x)$ of the first lamina projected forward onto it for comparison. With x and Δx fixed, the upper boundary is now the curve $z(y) = f(x + \Delta x, y)$, and at any y-coordinate y, the increase in height is approximately $f_x(x, y)\Delta x$. Similarly, the increase in the lamina's width is approximately $\beta'(x)\Delta x$.

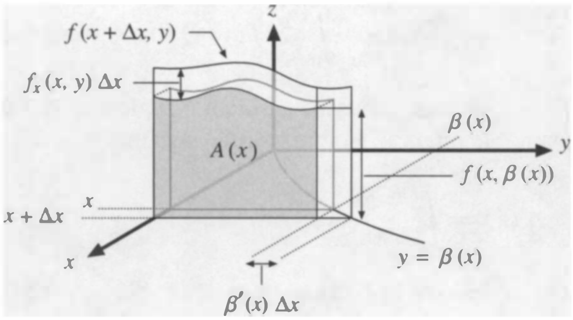


Figure 2 $A(x)$ projected onto $A(x + \Delta x)$

Finally, I've taken the lamina in FIGURE 2 and isolated it in FIGURE 3 for clarity. The area of the small piece in the upper right corner is proportional to $(\Delta x)^2$, so to first order accuracy in Δx we can ignore it. Thus the increase $\Delta A = A(x + \Delta x) - A(x)$ in area is approximately equal to that of the two shaded regions in Figure 3:

$$\Delta A \approx A_1 + A_2. \tag{3}$$

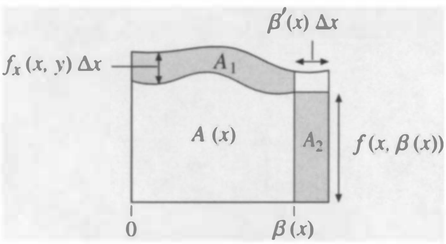


Figure 3 Approximate shading of ΔA

It should be clear from FIGURE 3 that

$$A_1 \approx \int_0^{\beta(x)} f_x(x, y)\, dy\, \Delta x \qquad \text{and} \qquad A_2 \approx f(x, \beta(x))\beta'(x)\Delta x. \tag{4}$$

This is the desired geometric interpretation; the two terms on the right side of Equation (2) are proportional, respectively, to the areas A_1 and A_2 in FIGURE 3. Indeed, we can use Equations (4) to substitute for A_1 and A_2 in Equation (3), and divide by Δx to obtain

$$\frac{\Delta A}{\Delta x} \approx \int_0^{\beta(x)} f_x(x, y)\,dy + f(x, \beta(x))\beta'(x).$$

Letting $\Delta x \to 0$ gives us Leibniz's rule.

Comment This approach seemed so natural to me that I was surprised not to find it in a textbook. However, one of the referees found a closely related exposition in an old classic text [**5**].

Formal statement of Leibniz's rule. Let $A(x)$ be given by (1). If f and $f_x = \partial f/\partial x$ are continuous in a region R, and if $\beta(x) \geq 0$ is continuously differentiable for $a \leq x \leq b$, and if $\{(x, y) \mid a \leq x \leq b,\ 0 \leq y \leq \beta(x)\} \subset R$, then A is continuously differentiable for $a \leq x \leq b$ and $A'(x)$ is given by (2).

Acknowledgment. My thanks to Barbara Nimershiem for pointing out Colin Adams' article [**1**], and for my going-away present—the question that led to this note.

REFERENCES

1. Colin C. Adams, The newest inductee in the number hall of fame, this MAGAZINE **71** (1998), 341–349.
2. F. B. Hildebrand, *Advanced Calculus, Second Ed.*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
3. W. Kaplan, *Advanced Calculus, Second Ed.*, Addison-Wesley, Reading, MA, 1973.
4. M. R. Spiegel, *Advanced Calculus*, Schaum's Outline Series, McGraw-Hill, New York, 1997.
5. E. B. Wilson, *Advanced Calculus*, Ginn & Co., Boston, 1912.

A Celestial Cubic

C. W. GROETSCH
University of Cincinnati
Cincinnati, OH 45221-0025

> ...Nature is pleased with simplicity and affects not
> the pomp of superfluous causes.
> Newton

Introduction The Greek philosopher Epicurus (342?–270 BC) espoused an infinity of worlds like our own. Later thinkers followed suit, sometimes with tragic consequences—Giordano Bruno (1548?–1600) was consumed by the flames of the Inquisition for preaching, among other things, a plurality of worlds. In modern times belief in the existence of extra-solar planets has been nearly universal (leading sometimes to comedy, rather than tragedy—viz. UFOs, alien abductions, etc.), yet the immense inter-stellar distances involved have defeated efforts to observe such planets directly. In fact, direct visual observation of extra-solar planets has been compared to naked-eye viewing, at a distance of many miles, of a moth fluttering around a porch light.

Recently convincing *indirect* evidence of planets orbiting specific stars has emerged. In the past indirect evidence has led to major astronomical discoveries. Indirect visual evidence, in the form of observations of orbital perturbations of Uranus, combined with the solution of a difficult mathematical inverse problem by Adams and Leverrier (independently), led to the discovery of Neptune in 1846 [**5**], and an inverse technique worked out by Bessel resulted in the discovery of a stellar companion of Sirius in 1862 [**2**]. The new indirect evidence for extra-solar planets is spectral data—subtle Doppler shifts resulting from the gravitational tug-of-war between the star and its planet.

Determining the velocity of a planet in a circular orbit of known radius centered on a star of known mass is a straightforward direct problem: simply equate the gravitational acceleration of the planet to its centripetal acceleration. We treat the subtler inverse problem of determining the mass and orbital radius of a planet from (indirect) velocity information (see [**4**] for more examples of elementary inverse problems). This inverse problem leads to an interesting cubic equation that we call the "celestial cubic" and this cubic serves to illustrate convergence of fixed point iteration and to characterize the relative error in a simpler approximation method.

A very simple mathematical model, ultimately involving nothing more than the extraction of a cube root, has been used by astronomers on spectral shift data to estimate the mass and orbital radius of an unseen extra-solar planet. The celestial cubic we discuss is a more exact model and this cubic equation teaches a lesson on iterative methods. No claim is made of a scientific advance—our goal is simply to illustrate how classroom mathematics is related in a meaningful way to a problem that is both old and scientifically *au courant*.

The two body model Adhering to William of Ockham's dictum that hypotheses "must not be unnecessarily multiplied," we treat the simplest model that accounts for the spectral shift data. In this model a far off star of mass M is orbited by a single planet of mass $m < M$. The star and planet interact gravitationally, each orbiting the common center of mass of the star-planet solar system (c.o.m. in the figures) with the result that the star "wobbles" with respect to a terrestrial observer (an Olympic hammer thrower is a helpful mental image; the beefy athlete, wobbling as he spins in the throwing circle, is the star and the ball at the end of the chain is the planet).

The star-planet system is far too distant for such a wobble to be observed directly. However, sensitive spectrometers can analyze light from the star and detect the chromatic Doppler shifts that result as the star moves alternately toward and away from the Earth under the influence of the gravity-induced wobble. As the star moves away from us its light signal shifts toward the red end of the visible spectrum, while as it moves toward us the color of the light shifts towards the blue. The magnitude of these shifts determines the radial velocity of the star relative to the Earth. And the time between successive peaks in the spectral time series of the star light is an important dynamical parameter, the orbital period T of the star-planet about the common center of mass.

Suppose the star orbits the center of mass of the star-planet system in a tight circle of radius r with a uniform speed v. For now we assume that the Earth lies in the orbital plane of the alien star-planet solar system. The form of the radial velocity of the star relative to the Earth can be worked out with the aid of FIGURE 1 (the diameter of the orbit is grossly exaggerated in the FIGURE).

If $D \gg r$ is the distance from the Earth to the center of mass of the star-planet system, then the distance between the Earth and the star is

$$d(t) = D + r\cos\left(\frac{v}{r}\,t\right),$$

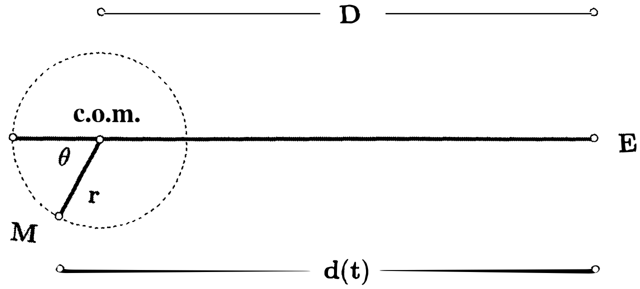


Figure 1 Star-c.o.m.-Earth system (planet not shown)

where $t = 0$ corresponds to that time when the star is most distant from us. The radial velocity, V, of the star relative to the Earth is then

$$V(t) = -v \sin\left(\frac{v}{r}\, t\right) = -v \sin\left(\frac{2\pi}{T}\, t\right),$$

where T is the period of rotation of the star-planet system. In this model the velocity of the star in its orbit about the center of mass is therefore the amplitude of the radial velocity of the star relative to the Earth. This velocity v of the star and the period T of rotation about the center of mass can therefore be obtained from a fit of the radial velocity time series resulting from the observed Doppler shifts (see [**5**, Fig. 3] for real data).

The mass M of the star is the remaining parameter needed to estimate the mass and orbital radius of the alien planet. An estimate of the stellar mass is obtained from luminosity measurements [**3**]. Stars are classified as to *stellar type* by luminosity, and masses (in terms of $M_\odot$, the mass of our sun) are assigned to stellar types. In the next section we show how the data v, T, and M serve to determine the mass and orbital radius of the "moth" that is the unseen extra-solar planet. We note that v and T immediately give the orbital radius r of the *star* about the center of mass of the star-planet system: $2\pi r = vT$.

The celestial cubic Suppose the planet of mass m orbits the star of mass M in a circle of radius R centered on the star. The center of mass of the star-planet solar system lies on the segment connecting the two bodies at a distance r, say, from the star (see FIGURE 2). The acceleration of the star relative to the center of mass is easily calculated if we parameterize the circular orbit as

$$\left(r \cos\left(\frac{v}{r}\, t\right), \quad r \sin\left(\frac{v}{r}\, t\right)\right).$$

The acceleration is then

$$-\frac{v^2}{r}\left(\cos\left(\frac{v}{r}\, t\right), \quad \sin\left(\frac{v}{r}\, t\right)\right)$$

and hence the magnitude of the centripetal force is Mv^2/r. This centripetal force equals the gravitational force as given by Newton's law of gravity (we consider the star-planet solar system in isolation), that is

$$M\frac{v^2}{r} = G\frac{Mm}{R^2}, \tag{1}$$

where G is the universal gravitation constant. The center of mass is determined by the condition

$$rM = m(R - r)$$

and hence

$$R = \frac{M + m}{m} r. \tag{2}$$

Substituting this into (1) results in

$$\frac{G}{rv^2} m^3 = (M + m)^2$$

or, setting $\sigma = \dfrac{GM}{rv^2}$, and denoting the *relative* planetary mass by $x = m/M$, we find that

$$\sigma x^3 = (1 + x)^2. \tag{3}$$

The dimensionless parameter σ is determined by the data v, M, and T; and the solution x, along with an estimate of the stellar mass M, gives the mass m of the planet. The orbital radius R of the planet is then obtained from (2).

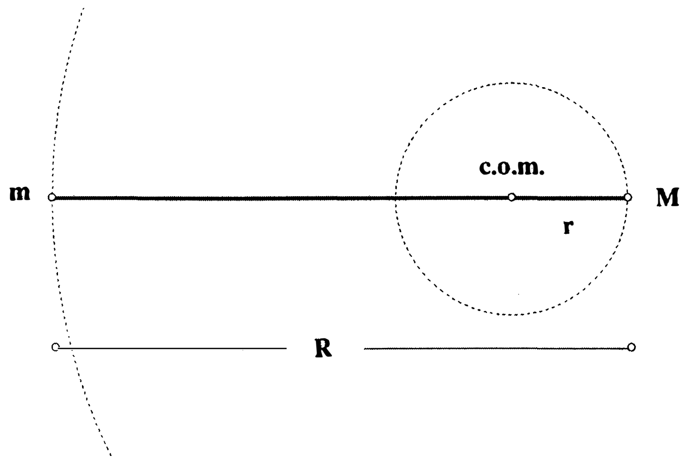


Figure 2 Star-planet solar system

Assuming that the planet is much less massive than the star, one has $x \ll 1$ and hence a good approximation to the relative mass can be obtained by setting $x + 1 \approx 1$, that is, by simply extracting a cube root: $x \approx \sigma^{-1/3}$. For example, observations of the star 51-Pegasi [**6**] provide the data $v = 56m/s$ and $T = 4.231$ days. This, along with the estimate that 51-Pegasi has a mass equal to that of our sun, results in the value $\sigma = 1.38 \times 10^{10}$ and a relative mass estimate of $x \approx 4.17 \times 10^{-4}$. Using the fact that the mass of Jupiter is approximately 9.55×10^{-4} times that of the sun, we find that the the alien planet is about 45% as massive as Jupiter and, from (2), that it orbits 51-Pegasi in a very tight orbit of radius about 0.05 A.U. (an astronomical unit, A.U., is the mean radius of the Earth's orbit, about 150 million kilometers).

If the assumption $m \ll M$ is dispensed with (as would be the case if the orbiting body were a dark companion star instead of a planet), then the relative mass must be obtained from the full cubic (3). Of course, thanks to del Ferro, Tartaglia and Cardano there is an explicit "closed form" solution for a cubic. However, this formula, when applied to (3), results (thanks to *Mathematica*) in the rather uninformative explicit solution

$$x = \frac{1}{3\sigma} - \frac{2^{1/3}(-1-6\sigma)}{3\sigma(2+18\sigma+27\sigma^2+3\sqrt{3}\sigma^{3/2}\sqrt{4+27\sigma})^{1/3}} + \frac{(2+18\sigma+27\sigma^2+3\sqrt{3}\sigma^{3/2}\sqrt{4+27\sigma})^{1/3}}{3\times 2^{1/3}\sigma}$$

The unsavory appearance of this explicit solution suggests that an indirect approach to the solution of (3) might be useful. In the next section we treat an iterative method and a variant.

Fixed point iteration Two simple successive approximation methods arise from algebraic rearrangements of (3). For example, rewriting (3) as

$$x = \sqrt{\sigma}x^{3/2} - 1 \tag{4}$$

suggests the iterative method

$$x_{n+1} = \sqrt{\sigma}x_n^{3/2} - 1, \tag{5}$$

while the expression

$$x = \sigma^{-1/3}(x+1)^{2/3} \tag{6}$$

gives rise to the method

$$x_{n+1} = \sigma^{-1/3}(x_n+1)^{2/3}. \tag{7}$$

While both (4) and (6) are equivalent to the celestial cubic (3), the iterative methods (5) and (7) are far from equivalent. In fact, method (7) is always convergent, while method (5) diverges. The celestial cubic therefore provides a simple classroom illustration of the care that should be taken in preparing an equation for the use of an iterative method.

Both methods (5) and (7) are instances of fixed point iteration (see, e.g., [**1**]), that is, methods of the form $x_{n+1} = f(x_n)$, where $f(x) = f_1(x) = \sigma^{-1/3}(x+1)^{2/3}$ for method (7) and $f(x) = f_1^{-1}(x) = \sqrt{\sigma}x^{3/2} - 1$ for method (5).

Proposition. Equation (3) has a unique solution $x = x(\sigma) \in (0, 1)$ if and only if $\sigma > 4$. If $\sigma > 4$, then method (7) converges to $x(\sigma)$ for any initial approximation $x_0 \in (0, 1)$. On the other hand, $x(\sigma)$ is a point of repulsion of method (5).

Proof. First note that if $x \in (0, 1)$ solves (3), then

$$\sigma > \sigma x = \left(1 + \frac{1}{x}\right)^2 > 4.$$

Suppose then that $\sigma > 4$. Since (3) is equivalent to (6), it suffices to show that $f_1(x) = \sigma^{-1/3}(x+1)^{2/3}$ satisfies the hypotheses of the Contraction Mapping Theorem on [0, 1]. In fact, $f_1(0) > 0$ and $f_1(1) = (\frac{4}{\sigma})^{1/3} < 1$. Furthermore,

$$0 < f_1'(x) = \frac{2}{3}\sigma^{-1/3}(x+1)^{-1/3} < \frac{2}{3}\sigma^{-1/3} < 2/3$$

and therefore $f_1 : [0, 1] \to [0, 1]$ is a contraction and hence has a unique fixed point $x(\sigma) \in (0, 1)$. On the other hand,

$$\left(f_1^{-1}\right)'(x(\sigma)) = \frac{1}{f_1'(x(\sigma))} > \frac{3}{2}$$

and hence $x(\sigma)$ is a point of repulsion for the iteration (5). ■

We note that $\frac{2}{3}\sigma^{-1/3}$ is a contraction constant for f_1 and hence the iteration (7) converges rapidly if σ is large. In particular, this is the case for the 51-Pegasi data where $\sigma = 1.38 \times 10^{10}$. The condition $\sigma > 4$ in the Proposition is not artificial; in fact, this condition *always* holds in the astronomical problem. Indeed, if $M > m$, then, by (2), $R/r > 2$. One then finds, using (1), that

$$\sigma = \frac{GM}{rv^2} = \frac{GMm}{R^2}\frac{R^2}{mrv^2} = \frac{Mv^2}{r}\frac{R^2}{mrv^2} = \left(\frac{R}{r}\right)^2 \frac{M}{m} > 4$$

The Mean Value Theorem can be used to gauge the error committed in replacing the solution $x(\sigma)$ of (3) by the approximation $\sigma^{-1/3}$ suggested in the previous section. Indeed, for $x \in (0, 1)$ we have

$$(1 + x)^{2/3} - 1 = \frac{2}{3}(1 + \xi)^{-1/3}x \quad \text{for some} \quad \xi \in (0, 1)$$

and hence the relative error in the approximation is given by

$$\left|\frac{x(\sigma) - \sigma^{-1/3}}{x(\sigma)}\right| = \sigma^{-1/3}\left|\frac{(1 + x(\sigma))^{2/3} - 1}{x(\sigma)}\right| = \frac{2}{3}\sigma^{-1/3}(1 + \xi)^{-1/3}$$

Therefore, the relative error satisfies

$$\frac{2^{2/3}}{3}\sigma^{-1/3} \le \left|\frac{x(\sigma) - \sigma^{-1/3}}{x(\sigma)}\right| \le \frac{2}{3}\sigma^{-1/3}$$

For the 51-Pegasi data ($\sigma = 1.38 \times 10^{10}$) the error in the approximation is therefore less than three hundreths of one percent—acceptable, to say the least. On the other hand, at the other end of the range of feasible data, $\sigma = 4$, the error exceeds 33%.

Finally, we note that the relative planetary mass, $x = x(\sigma)$, is an increasing function of the observed radial velocity amplitude, v, of the star with respect to the Earth. Indeed, $\frac{d\sigma}{dv} < 0$ and from (3) we have

$$x^3 + 3\sigma x^2 \frac{dx}{d\sigma} = 2(1 + x)\frac{dx}{d\sigma}.$$

However, again by (3), $3\sigma x^2 = \frac{3(1 + x)^2}{x}$, and hence

$$\frac{dx}{d\sigma} = \frac{-x^4}{(x + 1)(x + 3)} < 0.$$

Therefore, $\frac{dx}{dv} = \frac{dx}{d\sigma}\frac{d\sigma}{dv} > 0$. In the next paragraph we show that this result has an important implication related to the interpretation of the computed relative mass.

We have assumed the simplest model in which the Earth is in the orbital plane of a single-planet alien solar system. Of course, in reality the orbital plane may be tilted with respect to the line-of-sight from here and there is no way to estimate this tilt. This means that the observed radial velocity amplitude v is an *underestimate* of the true velocity amplitude of the star (for example, if the orbital plane of the star-planet were orthogonal to the line-of-sight from here, then there would be no detectable "wobble" and the observed radial velocity amplitude of the star would be *zero*). Since $\frac{dx}{dv}$ is positive, the relative planetary mass x calculated using the observed radial velocity v is then actually a *lower bound* for the true relative planetary mass. That is, the spectral wobble method only gives a lower bound for the mass of a single alien planet.

Ockham's Razor dulled A single-planet solar system is the simplest hypothesis that can account for the spectral shift data from a distant star, but there are infinitely many other models consistent with the data. Such is often the case with inverse problems. In this section we treat a simple "in-line" two-planet solar system.

Consider two planets of masses m_1 and m_2 executing circular orbits about a star of mass $M > m_1 + m_2$ in an in-line configuration, so that the segment connecting the planets always intesects the star. Assume that $m_1 < m_2$ and that the radii of the planets' orbits about the star are R_1 and R_2, respectively, with $R_1 < R_2$, as illustrated in FIGURE 3. The center of mass of the entire solar system then lies on the segment joining the star and the more massive planet, say at a distance r from the star. One then has

$$m_1(R_1 + r) + Mr = m_2(R_2 - r) \tag{8}$$

and hence

$$r = \frac{m_2 R_2 - m_1 R_1}{M + m_1 + m_2}.$$

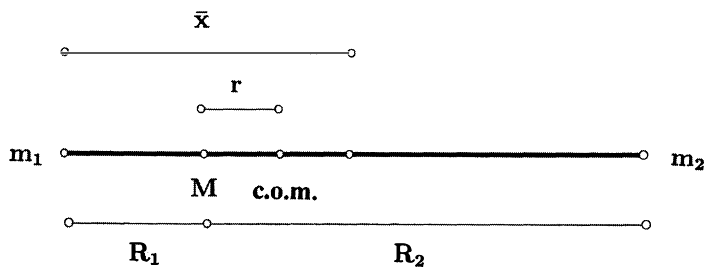


Figure 3 An in-line solar system

Suppose the center of mass of the two planets alone lies on the segment connecting the planets at a distance $\bar{x}$ from the less massive planet, then

$$\bar{x} = \frac{m_2(R_1 + R_2)}{m_1 + m_2}.$$

The distance between the star and the planetary center of mass is then

$$\bar{x} - R_1 = \frac{m_2 R_2 - m_1 R_1}{m_1 + m_2}.$$

Equating the centripetal force on the star as it orbits the center of mass of the solar system to the gravitational force between the star and the planetary system, we get

$$GM\frac{m_1+m_2}{\left(\dfrac{m_2R_2-m_1R_1}{m_1+m_2}\right)^2}=M\frac{v^2}{r},$$

where v is the velocity of the star in its orbit about the center of mass of the solar system. This is equivalent to

$$\frac{GM}{v^2r}\left(\frac{m_1+m_2}{M}\right)^3=\left(1+\frac{m_1}{M}+\frac{m_2}{M}\right)^2.$$

If we set $\alpha=m_1/m_2$, and $x=(1+\alpha)\dfrac{m_2}{M}$, then x satisfies the familiar celestial cubic

$$\sigma x^3=(1+x)^2. \tag{9}$$

The solution $x\in(0,1)$ is then uniquely determined by the Proposition, and may be computed by (7). From (8) we then have

$$R_2=\alpha R_1+(1+\alpha)\left(1+\frac{1}{x}\right)r.$$

So if α and R_1 are chosen, then R_2 and m_2, and hence m_1, are determined. Therefore a given set of spectral data $\{v, T, M\}$ is consistent with a two parameter (α and R_1) family of two planet in-line solar systems. In fact, since the parameter R_1 is free, any given set of spectral data could be explained by a pair of in-line planets with arbitrarily large orbits. So a single planet in a tight orbit is not the only model consistent with the observations. But, in the absence of additional information, it is probably best to eschew "superfluous causes" and take Ockham's advice to keep it simple. And in the case of 51-Pegasi, simple means a surprisingly large planet in a very tight orbit.

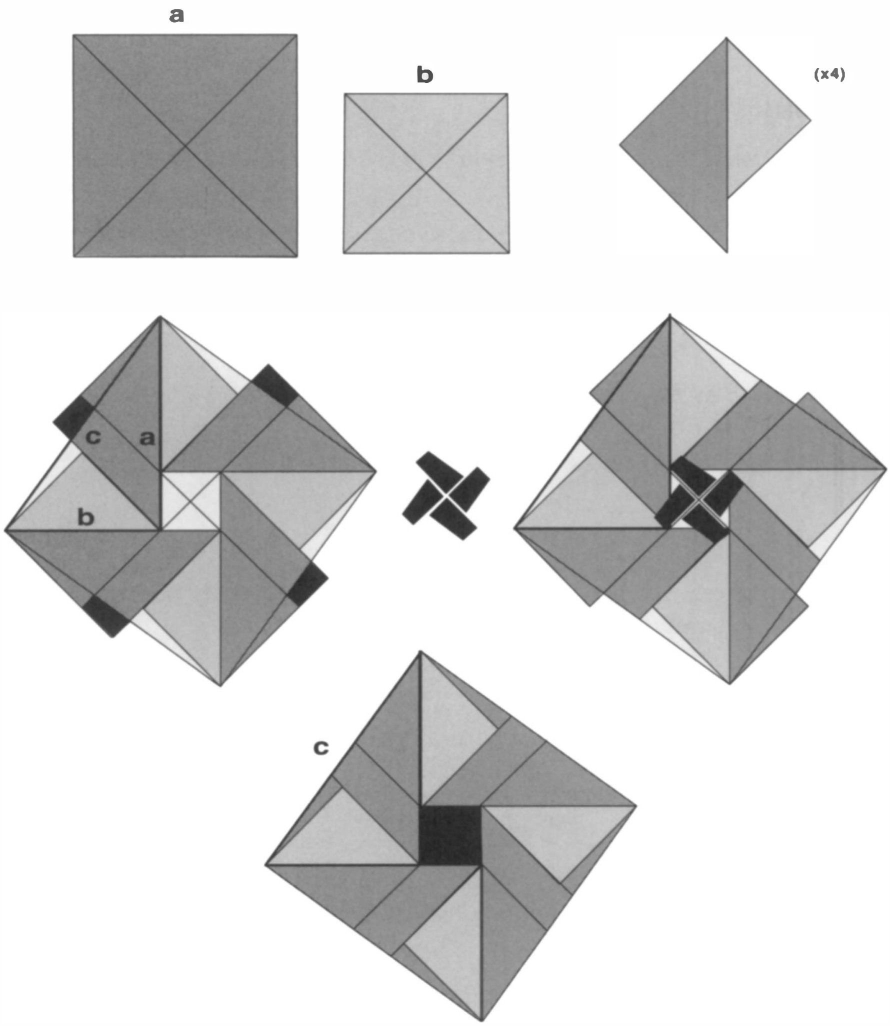
Acknowledgment. I am indebted to Charles Telesco for a very informative conversation. This work was supported in part by the National Science Foundation.

REFERENCES

1. K. E. Atkinson, *Introduction to Numerical Analysis*, Second Edition, Wiley, New York, 1989.
2. A. Chapman, The astronomical revolution, in *Möbius and His Band*, (J. Fauvel, R. Flood, and R. Wilson, Eds.), Oxford University Press, Oxford, 1993, pp. 35–77.
3. K. Crosswell, *Planet Quest, the Epic Discovery of Alien Solar Systems*, Free Press, New York, 1997.
4. C. W. Groetsch, *Inverse Problems*, Mathematical Association of America, Washington, DC, 1999.
5. M. Grosser, *The Discovery of Neptune*, Harvard University Press, Cambridge, MA, 1962.
6. G. W. Marcy, et al., The planet around 51-Pegasi, *Astrophysical Journal* **481** (1997), 926–935.

Proof Without Words: $a^2 + b^2 = c^2$

JOSE A. GOMEZ
South Junior High School
Newburgh, NY 12550



For an interactive version of this diagram, follow the link to Supplements from the MAGAZINE website: `http://www.maa.org/pubs.mathmag.html`.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

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

Proposals

To be considered for publication, solutions should be received by September 1, 2001.

1618. *Proposed by Michael Golomb, Purdue University, West Lafayette, IN.*

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

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

1619. *Proposed by Costas Efthimiou, Department of Physics, University of Central Florida, Orlando, FL.*

Consider the real sequences $(a_k)_{k\geq 0}$ that satisfy the recurrence relation

$$\cos a_{n+m} = \frac{\cos a_n + \cos a_m}{1 + \cos a_n \cos a_m}$$

for all nonnegative integers $n,\ m$. Such a sequence will be called *minimal* if $0 \leq a_k < \pi$ for each k. Determine all minimal sequences.

1620. *Proposed by Mihàly Bencze, Romania.*

In triangle ABC, let $a = BC,\ b = CA$, and $c = AB$. Let $m_a,\ m_b$, and m_c be, respectively, the length of the medians from $A,\ B$, and C, let s be the semiperimeter, and let R be the circumradius of the triangle. Prove that

$$\max\{am_a,\ bm_b,\ cm_c\} \leq sR.$$

1621. *Proposed by Donald Knuth, Stanford University, Stanford, CA.*

Prove that if a, b, and n are arbitrary nonnegative integers, then the sum

$$\sum_{k=-\infty}^{\infty} \left(\binom{n}{a + 5k} - \binom{n}{b + 5k}\right)$$

is a Fibonacci number or the negative of a Fibonacci number.

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

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

1622. *Proposed by Götz Trenkler, Dortmund, Germany.*

On the family $\mathcal{P}_n$ of $n \times n$ positive definite matrices, define the partial order $\leq_L$ by

$$M \leq_L N \qquad \text{if and only if} \qquad N - M \qquad \text{is positive semi-definite.}$$

(This is the Loewner ordering on $\mathcal{P}_n$.) Find the matrix C in $\mathcal{P}_n$ that is the greatest lower bound, with respect to $\leq_L$, for the set

$$\{(A + A^{-1})^2 + (B + B^{-1})^2 \,:\, A \in \mathcal{P}_n,\ B = I_n - A \in \mathcal{P}_n\}.$$

Quickies

Answers to the Quickies are on page 161.

Q909. *Proposed by Joaquin Gómez Rey, Madrid, Spain.*

Let n be a positive integer. Evaluate

$$\frac{2}{n}\sum_{k=1}^{n}[k,\ n] - n\sum_{k=1}^{n}\frac{1}{(k,\ n)}.$$

Here $[k,\ n]$ denotes the least common multiple and $(k,\ n)$ the greatest common divisor of k and n.

Q910. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, AB, Canada.*

Given positive integer k, it is easy to find two base 10 numbers whose product has more than k digits and has all digits the same. As an example, take 9 and $(10^n - 1)/9$ with $n > k$. Give examples for which the two numbers have the same number of digits.

Solutions

A Square in a Triangle in a Quadrilateral — April 2000

1594. *Proposed by Kent Holing, Statoil Research Centre, Trondheim, Norway.*

In quadrilateral $ABCD$, $AB + AD = BC + CD$ and $\angle A$ is a right angle. Square $APQR$ has P, Q, and R on segments AB, BD, and AD, respectively, with $AP = BC$.

(a) Find the number of such quadrilaterals (up to congruence) given the lengths BC and BD.

(b) Show how to construct all such quadrilaterals in terms of well-known straight-edge-and-compass constructions given the lengths BC and BD.

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

(a) Let $a,\ b,\ c,\ s$, and t be the lengths of $AD,\ AB,\ BD,\ BC$, and CD, respectively. See Figure 1. Because $\angle A$ is a right angle, A is on the circle with diameter BD. Since $\triangle PBQ$ is similar to $\triangle ABD$, we have

$$\frac{b-s}{b} = \frac{s}{a} \qquad \text{from which} \qquad s = \frac{ab}{a+b}.$$

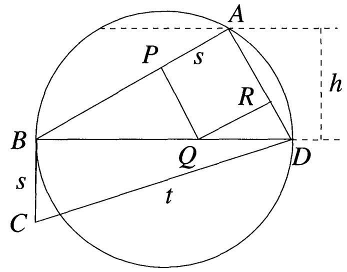


Figure 1

Thus $(a+b)^2 = a^2 + 2ab + b^2 = a^2 + 2s(a+b) + b^2$. Letting $k = a + b = s + t$ and using $a^2 + b^2 = c^2$ we have

$$k^2 - 2sk - c^2 = 0.$$

Solving for k gives $k = s + \sqrt{s^2 + c^2}$, and it follows that $t = \sqrt{s^2 + c^2}$. Thus $\angle CBD$ is a right angle.

Let h be the height of $\triangle ABD$ measured from A. Then the area of $\triangle ABD$ is given by either of $\frac{1}{2}ab$ or $\frac{1}{2}ch$, and it follows that

$$h = \frac{ab}{c} = \frac{s(a+b)}{c} = \frac{s(s+t)}{c} = \frac{s(s+\sqrt{s^2+c^2})}{c}.$$

Thus, given s and c we can find h. Point A is on the circle with diameter BD at a distance h above the diameter as shown in Figure 1. Note that if $0 < h < c/2$, then there are two positions for A and hence two quadrilaterals satisfying the given conditions. If $h = c/2$, then there is one position for A and one quadrilateral, and if $h > c/2$ there are no quadrilaterals.

(b) To construct the quadrilateral, first construct a right triangle with legs of lengths s and c. The hypotenuse has length $t = \sqrt{s^2 + c^2}$. A segment of length h is constructed by constructing a segment OH of length $s + t$ on one side of an angle O and segments OF and OG of length s and c on the other side. See Figure 2. The segment through F parallel to GH intersects OH in point M with $OM = \frac{s}{c}(s+t) = h$. Once h is realized the construction of the quadrilateral(s) is routine.

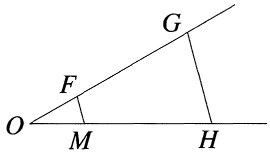


Figure 2

Comment: Some readers noted that allowing A to be below BD in Figure 1 leads to possibly "re-entrant" quadrilaterals that also satisfy the conditions of the problem.

Also solved by Jean Bogaert (Belgium), Daniele Donini (Italy), Marty Getz and Dixon Jones, Hans Kappus (Switzerland), Victor Y. Kutsenok, Raul A. Simon (Chile), Thomas Vanden Eynden, Peter Y. Woo, Li Zhou, and the proposer.

A Divisor Condition **April 2000**

1595. *Proposed by Wu Wei Chao, Guang Zhou Normal University, Guang Zhou City, Guang Dong Province, China.*

Find all pairs of positive integers a and b such that $ab + a + b$ divides $a^2 + b^2 + 1$.

Solution by Walter Stromquist, Berwyn, PA.

Either $a = b = 1$ or a and b are consecutive squares.

The divisibility condition can be written as

$$k(ab + a + b) = a^2 + b^2 + 1, \tag{1}$$

for some positive integer k. If $k = 1$, then (1) is equivalent to

$$(a - b)^2 + (a - 1)^2 + (b - 1)^2 = 0,$$

from which $a = b = 1$. If $k = 2$, then (1) can be written as

$$4a = (b - a - 1)^2,$$

forcing a to be a square, say $a = d^2$. Then $b - d^2 - 1 = \pm 2d$, so $b = (d \pm 1)^2$, and a and b are consecutive squares.

Now assume that there is a solution with $k \geq 3$, and let (a, b) be the solution with a minimal. Because the problem is symmetric in a and b, we have $a \leq b$. Write (1) as a quadratic in b:

$$b^2 - k(a + 1)b + (a^2 - ka + 1) = 0.$$

Because one root, b, is an integer, the other root, call it r, is also an integer. Because (1) must be true with r in place of b, we conclude that $r > 0$. Because $a \leq b$ and the product of the roots, $a^2 - ka + 1$, is less than a^2, we must have $r < a$. But then (r, a) is also a solution to (1), contradicting the minimality of a.

Also solved by Nicolas K. Artemiadis (Greece), The Austrian IMO-Team 2000, Charles Diminnie and Trey Smith, Daniele Donini (Italy), Achilleas Sinefakopoulos (Greece), Li Zhou, Paul Zweir, and the proposer. There were also one solution with no name, four partial solutions, and one incorrect result submitted.

Volumes of Simplexes **April 2000**

1596. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, AB, Canada.*

From the vertices $A_0, A_1, \ldots, A_n$ of simplex S, parallel lines are drawn intersecting the hyperplanes containing the opposite faces in the corresponding points $B_0, B_1, \ldots, B_n$. Determine the ratio of the volume of the simplex determined by $B_0, B_1, \ldots, B_n$ to the volume of S.

Solution by L. R. King, Davidson College, Davidson, NC.

The ratio for n-dimensional simplexes is n. Without loss of generality we may consider the standard simplex S_0 with one vertex at the origin and the others at $e_1, e_2, \ldots, e_n$, where $e_1, \ldots, e_n$ denotes the usual basis for R^n. (The translate $S - A_0$ is the image $L(S_0)$ for some linear transformation $L : R^n \longrightarrow R^n$. Because L has rank n, ratios of volumes are invariant under L.)

Let $B_0 = (s_1, s_2, \ldots, s_n)$ be the intersection point of the line from the origin with the face opposite the origin, so $s_1 + s_2 + \cdots + s_n = 1$. We then find $B_k = e_k - \frac{1}{s_k} B_0$, and $B_k - B_0 = e_k - t_k B_0$, where $t_k = \left(1 + \frac{1}{s_k}\right)$. (Note that $s_k \neq 0$ because otherwise the line through e_k and parallel to B_0 would be parallel to the face opposite e_k.) The volume of the simplex with vertices $B_0, \ldots, B_n$ is

$$\begin{aligned}
&\frac{1}{n!}\,|\det(B_1 - B_0, \ldots, B_k - B_0, \ldots, B_n - B_0)| \\
&\quad = \frac{1}{n!}\,|\det(e_1 - t_1 B_0, \ldots, e_k - t_k B_0, \ldots, e_n - t_n B_0)| \\
&\quad = \frac{1}{n!}\,|\det(e_1, \ldots, e_n) - t_1 \det(B_0, e_2, \ldots, e_n) - t_2 \det(e_1, B_0, \ldots, e_n) \\
&\quad\quad - \cdots - t_n \det(e_1, \ldots, e_{n-1}, B_0)| \\
&\quad = \frac{1}{n!}\,|1 - (t_1 s_1 + t_2 s_2 + \cdots + t_n s_n)| \\
&\quad = \frac{1}{n!}\left|1 - n - \sum_{k=1}^{n} s_k\right| = \frac{1}{n!} n.
\end{aligned}$$

Because the standard simplex has volume $1/n!$, the ratio of the volumes is n.

Comment: Leon Gerber notes that this problem has a long history and that it appeared, with his solution, as a problem in The American Mathematical Monthly 80:10(1973), pp. 1145–1146.

Also solved by Michel Bataille (France), Daniele Donini (Italy), Michael Golomb, Hans Kappus (Switzerland), and the proposer.

Convex Logarithms **April 2000**

1597. *Proposed by Constantin P. Niculescu, University of Craiova, Craiova, Romania.*

For every $x, y \in (0, \sqrt{\pi/2})$ with $x \neq y$, prove that

$$\ln^2 \frac{1 + \sin xy}{1 - \sin xy} < \ln \frac{1 + \sin x^2}{1 - \sin x^2} \cdot \ln \frac{1 + \sin y^2}{1 - \sin y^2}.$$

I. *Solution by Li Zhou, Polk Community College, Winter Haven, FL.*

For $t \in (-\infty, \ln(\pi/2))$, define $f(t) = \ln\left(\ln \dfrac{1 + \sin e^t}{1 - \sin e^t}\right)$. We claim that f is strictly convex. Indeed, direct calculation gives

$$f''(t) = \frac{2e^t}{e^{2f(t)} \cos^2 e^t}\left[(\cos e^t + e^t \sin e^t) \ln\left(\frac{1 + \sin e^t}{1 - \sin e^t}\right) - 2e^t\right]. \tag{1}$$

If we let $g(u) = (\cos u + u \sin u) \ln\left(\dfrac{1 + \sin u}{1 - \sin u}\right) - 2u$, then $g(0) = 0$ and

$$g'(u) = u \cos u \ln\left(\frac{1 + \sin u}{1 - \sin u}\right) + 2u \tan u,$$

which is positive for $0 < u < \pi/2$. Hence $g(u)$ is positive on $(0, \pi/2)$. It follows from (1) that $f''(t) > 0$, so $f(t)$ is strictly convex for $t \in (-\infty, \ln(\pi/2))$. By the definition of (strictly) convex function,

$$\ln^2 \frac{1 + \sin xy}{1 - \sin xy} = e^{2f(\ln x + \ln y)} < e^{f(2\ln x) + f(2\ln y)} = \ln \frac{1 + \sin x^2}{1 - \sin x^2} \cdot \ln \frac{1 + \sin y^2}{1 - \sin y^2}$$

for distinct x and y in $(0, \sqrt{\pi/2})$.

II. *Solution by Daniele Donini, Bertinoro, Italy.*

Fix $x,\ y \in (0,\ \sqrt{\pi/2})$ with $x \neq y$. Because

$$\ln \frac{1+\sin a}{1-\sin a} = 2\int_0^a \sec t\, dt = 2a \int_0^1 \sec at\, dt$$

for $0 \le a < \pi/2$, the desired inequality is equivalent to

$$\left(2xy \int_0^1 \sec(xyt)\, dt\right)^2 < \left(2x^2 \int_0^1 \sec(x^2t)\, dt\right)\left(2y^2 \int_0^1 \sec(y^2t)\, dt\right)$$

and hence to

$$\left(\int_0^1 \sec(xyt)\, dt\right)^2 < \int_0^1 \sec(x^2t)\, dt \cdot \int_0^1 \sec(y^2t)\, dt. \tag{1}$$

By the Cauchy-Schwarz inequality,

$$\left(\int_0^1 \sqrt{\sec(x^2t)\sec(y^2t)}\, dt\right)^2 \le \int_0^1 \sec(x^2t)\, dt \cdot \int_0^1 \sec(y^2t)\, dt.$$

Thus (1) will be established once we show that

$$\int_0^1 \sec(xyt)\, dt < \int_0^1 \sqrt{\sec(x^2t)\sec(y^2t)}\, dt.$$

This last inequality is a consequence of the inequality

$$\sec(xyt) < \sqrt{\sec(x^2t)\sec(y^2t)},$$

or

$$\cos^2(xyt) > \cos(x^2t)\cos(y^2t), \tag{2}$$

for any $0 < t \le 1$. To prove (2) we note that $0 < 2xyt < x^2t + y^2t < \pi$, so

$$\begin{aligned}\cos^2(xyt) &= \frac{1}{2}\,(1+\cos(2xyt)) > \frac{1}{2}\left(\cos(x^2t - y^2t) + \cos(x^2t + y^2t)\right)\\ &= \cos(x^2t)\cos(y^2t).\end{aligned}$$

Also solved by The Austrian IMO-Team 2000, Michel Bataille (France), Kee-Wai Lau (China), Syrous Marivani, Heinz-Jürgen Seiffert (Germany), and the proposer. There was one incorrect submission.

Periodic Transformations of n-tuples **April 2000**

1598. *Proposed by Hoe-Teck Wee, student, Massachusetts Institute of Technology, Cambridge, MA.*

Starting with any n-tuple R_0, $n > 1$, of symbols from $A,\ B,\ C$, we define a sequence $R_0,\ R_1,\ R_2, \ldots$ according to the following relation: If $R_j = (x_1,\ x_2, \ldots, x_n)$, then $R_{j+1} = (y_1,\ y_2, \ldots, y_n)$, where $y_i = x_i$ if $x_i = x_{i+1}$ (taking $x_{n+1} = x_1$) and y_i is the symbol other than x_i and x_{i+1} if $x_i \neq x_{i+1}$. (For example, if $R_0 = (A,\ A,\ B,\ C)$, then $R_1 = (A,\ C,\ A,\ B)$.)

(a) Find all positive integers $n > 1$ for which there exists some integer $m > 0$ such that $R_m = R_0$ for all R_0.

(b) For $n = 3^k$, $k \geq 1$, find the smallest integer $m > 0$ such that $R_m = R_0$ for every R_0.

Solution by Marty Getz and Dixon Jones, University of Alaska Fairbanks, Fairbanks, AK.

(a) Such an m exists if and only if n is odd. Map the n-tuple R_i to $Q_i \in \mathbb{Z}_3^n$ by replacing A, B, C by 0, 1, 2, respectively. The transformation described in the problem statement defines a function $\phi : \mathbb{Z}_3^n \longrightarrow \mathbb{Z}_3^n$ by

$$\phi(Q_i) = \phi(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_n) = Q_{i+1},$$

where $y_i = -x_i - x_{i+1}$, with the addition performed in $\mathbb{Z}_3$ and $x_{n+1} = x_1$.

If n is even, let $Q_0 = (1, 2, 1, 2, \ldots, 1, 2)$. Then for each $m \geq 1$,

$$Q_m = \phi^m(Q_0) = (0, 0, \ldots, 0) \neq Q_0.$$

Next we show that if n is odd, then ϕ is one-to-one on $\mathbb{Z}_3^n$. Suppose that $Q = (x_1, x_2, \ldots, x_n)$, $Q' = (y_1, y_2, \ldots, y_n)$ and that $\phi(Q) = \phi(Q')$. Then, with addition performed in $\mathbb{Z}_3$,

$$-x_j - x_{j+1} = -y_j - y_{j+1}, \tag{1}$$

for $1 \leq j \leq n$. Hence

$$\sum_{j=1}^{n}(-1)^j(-x_j - x_{j+1}) = \sum_{j=1}^{n}(-1)^j(-y_j - y_{j+1}),$$

which reduces in $\mathbb{Z}_3$ to $2x_1 = 2y_1$. Thus $x_1 = y_1$ and repeated substitution in (1) gives $x_j = y_j$ for $j = 2, 3, \ldots n$. Hence $Q = Q'$, proving that ϕ is one-to-one. It now follows that each of the orbits $\{Q, \phi(Q), \phi^2(Q), \ldots\}$, $Q \in \mathbb{Z}_3^n$ is cyclic. If m is the least common multiple of the orders of the orbits, then $Q_m = \phi^m(Q_0) = Q_0$ for each $Q_0 \in \mathbb{Z}_3^n$.

(b) The smallest positive m is 3^k. It can be shown by induction that if $Q_0 = (x_1, x_2, \ldots, x_n)$, then $Q_j = \phi^j(Q_0) = (w_1, w_2, \ldots, w_n)$, where $w_p = (-1)^j \sum_{i=0}^{j} \binom{j}{i} x_{i+p}$, with all operations performed in $\mathbb{Z}_3$ and $x_{r+n} = x_r$. If $j = 3^k$, then the p-th term of Q_{3^k} is

$$w_p = -\sum_{i=0}^{3^k} \binom{3^k}{i} x_{i+p}.$$

Because $\binom{3^k}{i}$ is a multiple of 3 for $1 \leq i \leq 3^k - 1$, it follows that

$$w_p = -(x_p + x_{p+3^k}) = -x_p - x_p = x_p.$$

Thus $Q_{3^k} = \phi^{3^k}(Q_0) = Q_0$ for every $Q_0 \in \mathbb{Z}_3^n$. Finally take $Q_0 = (0, 0, 0, \ldots, 0, 0, 1) \in \mathbb{Z}_3^{3^k}$ and observe that for $0 < m < 3^k$, the $(3^k - m)$-th coordinate of Q_m is 1 or 2. This shows that no $m < 3^k$ has the desired property.

Also solved by J. C. Binz (Switzerland), Daniele Donini (Italy), Robert F. McCoart Jr., and the proposer.

Answers

Solutions to the Quickies from page 155.

A909. Because $[k,\ n](k,\ n) = kn$ and $(k,\ n) = (n-k,\ n)$,

$$\frac{2}{n}\sum_{k=1}^{n}[k,\ n] - n\sum_{k=1}^{n}\frac{1}{(k,\ n)} = 2\sum_{k=1}^{n}\frac{k}{(k,\ n)} - n\sum_{k=1}^{n}\frac{1}{(k,\ n)}$$

$$= \sum_{k=0}^{n-1}\frac{n-k}{(n-k,\ n)} + \sum_{k=1}^{n}\frac{k}{(k,\ n)} - \sum_{k=1}^{n}\frac{n}{(k,\ n)}$$

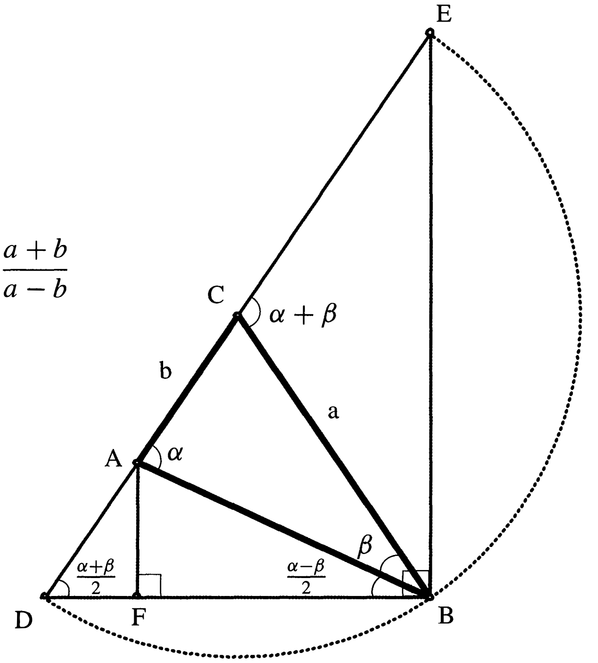
$$= \frac{n}{(n,\ n)} + \sum_{k=1}^{n-1}\frac{n-k}{(n-k,\ n)} - \sum_{k=1}^{n-1}\frac{n-k}{(n-k,\ n)} = 1.$$

A910. First note that $10^{3+6n} + 1 \equiv 0 \pmod 7$ and that $(10^{3+6n} - 1)/9$ and $(10^{3+6n} + 1)/7$ have the same number of digits. We then have

$$\left[(10^{3+6n} + 1)/7\right]\left[7(10^{3+6n} - 1)/9\right] = (10^{6+12n} - 1)/9.$$

Proof Without Words: The Law of Tangents

$$\frac{\tan((\alpha+\beta)/2)}{\tan((\alpha-\beta)/2)} = \frac{a+b}{a-b}$$



Notes: $CD = CB = CE, \alpha > \beta, \dfrac{DE}{DB} = \dfrac{DA}{DF} = \dfrac{AE}{FB}$.

$$\frac{\tan((\alpha+\beta)/2)}{\tan((\alpha-\beta)/2)} = \frac{AF/DF}{AF/BF} = \frac{a+b}{a-b}$$

—REX H. WU
2214 AVENUE O
BROOKLYN, NY 11210

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

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

Cohen, Patricia, What good is math? An answer for jocks, *New York Times* (3 February 2001) (National Ed.) A15, A17. `http://www.nytimes.com/2001/02/03/arts/03MATH.html` .

Readers outside the U.S.—and some inside—can be forgiven for not realizing that the country has a new professional league this spring. The sport is (American) football (played more violently), the commentators and promoters come from professional wrestling (and include the governor of a large Midwestern state), a national TV network puts the games in prime time on Saturday nights—and *mathematicians* made it all feasible. With only 8 teams in 2 divisions, it has a surprising number of constraints about how to schedule the games. Jeff Dinitz and Dalibor Froncek (University of Vermont) realized early last year, when the league was forming, that it would need a schedule; so they called up, offered to devise a good schedule—and found new friends for mathematics.

The Eternity Puzzle. `http://mathpuzzle.com/eternity.html` . Latest News about the Eternity Puzzle. `http://www.eternity-puzzle.co.uk/news.htm` . The Eternity Puzzle Solution. `http://www.msoworld.com/mindzine/news/miscellany/eternity.html` . Selby, Alex, and Oliver Riordan, Description of Method, `http://www.archduke.demon.co.uk/eternity/method/desc.html` .

Dinitz and Froncek got a few thousand dollars and a couple of footballs for their scheduling of the XFL football league, but the big money (£1 million) this past fall went to geometers. The Eternity puzzle consists of 209 all-different asymmetric pieces, and the object is to arrange them into a particular 12-sided figure. The puzzle sold 250,000 copies at £26 each. The first solution submitted was by mathematicians Alex Selby (formerly of Cambridge University), 32, and Oliver Riordan (Cambridge University), 28. This sort of tiling problem is NP-complete, but Selby and Riordan were able to solve it in 7 months; they estimate that there are 10^{95} solutions. Their Web page and its link to their mathematical framework describe their mathematical and computing methodologies rather thoroughly. (Notice—with either elation or despair—that the solution to a puzzle craze can earn you almost as much as solving one of the Clay Institute's $2- Millionennium Problems.)

Seife, Charles, Loopy solution brings infinite relief, *Science* 291 (9 February 2001) 965. Hass, Joel, and Jeffrey C. Lagarias, A Reidemeister move bound for unknotting, `http://www.research.att.com/~jcl/doc/reidemeister.ps` .

The three Reidemeister moves can be used to reduce a knot to a standard form—but how many moves might it take? And is it bounded in terms of the number of crossings of the knot? At last, a little progress has been made on this question, by Jeff Lagarias (AT&T Labs) and Joel Hass (University of California, Davis): The number of moves necessary to reduce an unknot with n crossings is bounded by $2^{100,000,000,000n}$. This bound leaves a little room for further improvement....

Trefethen, Lloyd N., and Lloyd M. Trefethen, How many shuffles to randomize a deck of cards?, *Proceedings of the Royal Society of London* A 456 (2000) 2561–2568. Gloor, Stephanie, and Lorenz Halbeisen, Über das Mischen von Spielkarten, ein mathematisches Menü, *Elemente der Mathematik* 54 (1999) 156–162. Piele, Don, Perfect shuffle, *Quantum* 11 (2) (November/December 2000) 55–57.

The paper by the Trefethens measures the effect of a loose riffle shuffle in terms of entropy of the resulting deck; their results confirm, using this different measure of randomness, the now-classic result that 7 shuffles serve to randomize a 52-card deck. The other two papers treat the mathematics of perfect riffle shuffles (the deck is divided exactly in half and the shuffling perfectly interleaves cards from the two halves). How many shuffles does it take before the deck returns to the original order? Gloor and Halbeisen give theorems to tell that and the positions after any number of shuffles of every card; Piele shows how to explore the same questions using Mathematica and the benefits of color. The results on perfect shuffles are not new but they are nicely done, and the presentations complement each other wonderfully.

Conway, John H., *On Numbers and Games*, 2nd ed., A K Peters, 2001; xi + 242 pp, $39. ISBN 1–56881–127–6. Berlekamp, Elwyn R., John H. Conway, and Richard K. Guy, *Wining Ways for Your Mathematical Plays*, 2nd ed., vol. 1, A K Peters, 2001; xix + 276 pp, $49.95 (P). ISBN 1–56881–130–6.

Here reappear in new editions the two founding classics of combinatorial game theory, after 25 and 20 years respectively, plus the authoritative book on Hex. Conway's book is unchanged except for a prologue and an epilogue, the latter recounting subsequent developments in the understanding of surreal numbers. Berlekamp et al.'s book, originally in two volumes (*Games in General*, *Games in Particular*), is to appear in four volumes over the next year; this first volume, which includes the first 8 chapters, has many additional references, a few corrections, and some new "Extras" in Chapter 6.

Abstract Games, 4 issues/yr, 28 pp/issue, Carpe Diem Publishing; $23/yr (Box 33018, 1583 Marine Dr., West Vancouver, BC, Canada V7V 1H0; `http://www.abstractgamesmagazine.com`). *Board Game Studies*, 1 issue/yr, 127–175 pp/issue, CNWS Publications; 3-issue subscription 68.07 Euros ≈ $62 (c/o Research School CNWS, Leiden University, P.O. Box 9515, 2300 RA, Leiden, The Netherlands, `http://www.BoardGamesStudies.org` , `cnws@Rullet.leidenuniv.nl`). Vol. 1: ISBN 90–5789–005–4; vol. 2: ISBN 90–5789–030–5. Browne, Cameron, *Hex Strategy: Making the Right Connections*, A K Peters, 2001; xi + 242 pp, $338.50 (P). ISBN 1–56881–117–9.

In the past couple of years two new journals have arisen that are devoted to games that may be of interest to mathematicians. *Abstract Games* features "games which have been forgotten or neglected, or games which are relatively new and therefore largely unknown." Games are introduced through a short series of articles featuring the rules, strategy, and tactics; games introduced in the first year included Bashne, Lines of Action, Kyoto Shogi (a chess variant), Hex (by Cameron Browne), and Bao (a form of mancala with four rows). There are also game reviews, book reviews, notice of games available on the Internet, and advertisements for games and books about them. *Board Game Studies* is not particularly concerned with rules or strategy but with the reflections that various disciplines can bring to bear on board games. Topics treated in the first two volumes include Inca dice and board games, classification and distribution of mancala games, the development of English board games, twentieth-century versions of the Maya game patolli, the Egyptian god of board games, and American games; contributions are in English (9), German (2), or French (1), with abstracts in all three languages. In addition, there are research notes and book reviews, plus beautiful color plates. Browne's book on Hex is an example of a thorough treatment of a board game with no associated anthropology (except among mathematicians); he thoroughly incorporates all the literature on the game and offers several levels of strategy.

NEWS AND LETTERS

29th United States of America Mathematical Olympiad
May 2, 2000

edited by Titu Andreescu and Zuming Feng

Problems

1. Call a real-valued function f *very convex* if

$$\frac{f(x)+f(y)}{2} \geq f\left(\frac{x+y}{2}\right) + |x-y|$$

holds for all real numbers x and y. Prove that no very convex function exists.

2. Let S be the set of all triangles ABC for which

$$5\left(\frac{1}{AP}+\frac{1}{BQ}+\frac{1}{CR}\right) - \frac{3}{\min\{AP, BQ, CR\}} = \frac{6}{r},$$

where r is the inradius and P, Q, R are the points of tangency of the incircle with sides AB, BC, CA, respectively. Prove that all triangles in S are isosceles and similar to one another.

3. A game of solitaire is played with R red cards, W white cards, and B blue cards. A player plays all the cards one at a time. With each play he accumulates a penalty. If he plays a blue card, then he is charged a penalty which is the number of white cards still in his hand. If he plays a white card, then he is charged a penalty which is twice the number of red cards still in his hand. If he plays a red card, then he is charged a penalty which is three times the number of blue cards still in his hand. Find, as a function of R, W, and B, the minimal total penalty a player can amass and all the ways in which this minimum can be achieved.

4. Find the smallest positive integer n such that if n unit squares of a 1000×1000 unit-square board are colored, then there will exist three colored unit squares whose centers form a right triangle with legs parallel to the edges of the board.

5. Let $A_1A_2A_3$ be a triangle and let ω_1 be a circle in its plane passing through A_1 and A_2. Suppose there exist circles $\omega_2, \omega_3, \ldots, \omega_7$ such that for $k = 2, 3, \ldots, 7$, ω_k is externally tangent to ω_{k-1} and passes through A_k and A_{k+1}, where $A_{n+3} = A_n$ for all $n \geq 1$. Prove that $\omega_7 = \omega_1$.

6. Let $a_1, b_1, a_2, b_2, \ldots, a_n, b_n$ be nonnegative real numbers. Prove that

$$\sum_{i,j=1}^{n} \min\{a_i a_j, b_i b_j\} \leq \sum_{i,j=1}^{n} \min\{a_i b_j, a_j b_i\}.$$

Comment: If you were not able to make significant progress on Problem 6, you are not alone. There was no complete solution submitted for this problem on the USAMO.

Solutions

Problem 1 The inequality that defines *very convex* functions shows that over the midpoint of any interval, there is a sizeable gap between the graph of the function and its secant line for that interval. There are many ways to show that those gaps add up to a condition that no function could satisfy. One of them follows.

Fix $n \geq 1$. For each integer i, define

$$\Delta_i = f\left(\frac{i+1}{n}\right) - f\left(\frac{i}{n}\right).$$

The given inequality with $x = (i+2)/n$ and $y = i/n$ can be arranged to give

$$f\left(\frac{i+2}{n}\right) - f\left(\frac{i+1}{n}\right) \geq f\left(\frac{i+1}{n}\right) - f\left(\frac{i}{n}\right) + 4/n,$$

that is, $\Delta_{i+1} \geq \Delta_i + 4/n$. Combining this for n consecutive values of i gives $\Delta_{i+n} \geq \Delta_i + 4$. Summing from $i = 0$ to $i = n-1$ and telescoping yields

$$f(2) - f(1) \geq f(1) - f(0) + 4n.$$

This cannot hold for all $n \geq 1$. Therefore, no such function exists.

Problem 2 Let A, B, C be the angles of triangle ABC. Then the formula

$$\tan\frac{A}{2}\tan\frac{B}{2} + \tan\frac{B}{2}\tan\frac{C}{2} + \tan\frac{C}{2}\tan\frac{A}{2} = 1,$$

can be proved by manipulating trigonometric identities, starting from $\tan\frac{A}{2}\tan\frac{B}{2} + \tan\frac{B}{2}\tan\frac{C}{2}$and arriving at $1 - \tan\frac{A}{2}\tan\frac{C}{2}$ without too much difficulty. Alternatively, one can use Heron's identity, as follows

Let a, b, c, r, s denote the side lengths, inradius, semiperimeter of triangle ABC, respectively. Then the area of the triangle, $[ABC]$, is rs. Also, $AP = s - a$, and $\tan(A/2) = r/(s-a)$. Hence $\tan\left(\frac{A}{2}\right) = \frac{[ABC]}{s(s-a)}$. Likewise,

$$\tan\left(\frac{B}{2}\right) = \frac{[ABC]}{s(s-b)} \text{ and } \tan\left(\frac{C}{2}\right) = \frac{[ABC]}{s(s-c)}.$$

Hence

$$\tan\frac{A}{2}\tan\frac{B}{2} + \tan\frac{B}{2}\tan\frac{C}{2} + \tan\frac{C}{2}\tan\frac{A}{2} = \frac{[ABC]}{s^2}\left(\frac{(s-c)+(s-a)+(s-b)}{(s-a)(s-b)(s-c)}\right)$$

$$= \frac{[ABC]}{s(s-a)(s-b)(s-c)} = 1,$$

by **Heron's formula**.

Now, without loss of generality assume that $AP = \min\{AP, BQ, CR\}$. Let $x = \tan(\angle A/2)$, $y = \tan(\angle B/2)$, and $z = \tan(\angle C/2)$. Hence $AP = r/x$, $BQ = r/y$, $CR = r/z$. Then the given equation becomes

$$2x + 5y + 5z = 6. \quad (1)$$

The lemma gives

$$xy + yz + zx = 1. \quad (2)$$

Eliminating x from (1) and (2), and completing squares yields

$$(3y-1)^2+(3z-1)^2=4(y-z)^2.$$

Setting $3y-1=u$, $3z-1=v$ (i.e., $y=(u+1)/3$, $z=(v+1)/3$) gives

$$5u^2+8uv+5v^2=0.$$

Since the discriminant of this quadratic equation is $8^2-4\times 25<0$, the only real solution to the equation is $u=v=0$. Since the tangent ratios of the half-angles of ABC are determined uniquely, namely $x=4/3$, $y=z=1/3$, all triangles in S are isosceles and similar to one another.

In fact, $x=r/AP=4/3$ and $y=z=r/BQ=r/CQ=1/3=4/12$ implies that we can set $r=4$, $AP=AR=3$, $BP=BQ=CQ=CR=12$. Therefore $AB=AC=15$, $BC=24$. By scaling, triangles in S are all similar to the triangle with side lengths 5, 5, 8.

One can also use **half-angle formulas** to calculate the value of

$$\sin B=\sin C=\frac{2\tan\frac{C}{2}}{1+\tan^2\frac{C}{2}}=\frac{3}{5}.$$

Therefore $AQ:QB:BA=3:4:5$ and $AB:AC:BC=5:5:8$.

Problem 3 The minimum achievable penalty is

$$\min\{BW, 2WR, 3RB\}.$$

The three penalties BW, $2WR$, and $3RB$ can clearly be obtained by playing cards in one of the three orders

- **bb** $\cdots$ **brr** $\cdots$ **rww** $\cdots$ **w**,
- **rr** $\cdots$ **rww** $\cdots$ **wbb** $\cdots$ **b**,
- **ww** $\cdots$ **wbb** $\cdots$ **brr** $\cdots$ **r**.

Given an order of play, let a "run" of some color denote a set of cards of that color played consecutively in a row. Then the optimality of one of the three above orders follows immediately from the following lemma, along with the analogous observations for blue and white cards.

Lemma 1 For any given order of play, we may combine any two runs of red cards without increasing the penalty.

Proof: Suppose that there are w white cards and b blue cards between the two red runs. Moving a red card from the first run to the second costs us $2w$ because we now have one more red card after the w white cards. However, we gain $3b$ because this red card is now after the b blues. If the net gain $3b-2w$ is non-negative, then we can move all the red cards in the first run to the second run without increasing the penalty. If the net gain $3b-2w$ is negative, then we can move all the red cards in the second run to the first run without increasing the penalty, as desired. ■

Thus there must be an optimal game where cards are played in one of the three given orders. To determine whether there are other optimal orders, first observe that **wr** can never appear during an optimal game; after all, switching these two cards would decrease the penalty. Similarly, **bw** and **rb** can never appear. Now we prove the following lemma.

Lemma 2 Any optimal order of play must have less than 5 runs.

Proof: Suppose that some optimal order of play had at least five runs. Assume the first card played is red; the proof is similar in the other cases. Say we first play r_1, w_1, b_1, r_2, w_2 cards of each color, where each r_i, w_i, b_i is positive and where we cycle through red, white, and blue runs. From the proof of our first lemma we must have both $3b_1 - 2w_1 = 0$ and $b_1 - 2r_2 = 0$. Hence the game starting with playing $r_1, w_1 + w_2, b_1, r_2, 0$ cards is optimal as well, so we must also have $3b_1 - 2(w_1 + w_2) = 0$, a contradiction. ■

Thus, any optimal game has at most 4 runs. Now from lemma 1 and our initial observations, any order of play of the form

$$\mathbf{rr \cdots rww \cdots wbb \cdots brr \cdots r},$$

is optimal if and only if $2W = 3B$ and $2WR = 3RB \leq WB$; and similar conditions hold for 4-run games that start with **w** or **b**.

Problem 4 We show that $n = 1999$. Indeed, $n \geq 1999$ because we can color 1998 squares without producing a right triangle: fill every square in the first row and the first column, except for the one square at their intersection.

Next we show that with 1999 colored squares, there is no way to avoid creating a triangle as described. Call a row or column *heavy* if it contains at least two colored squares, and *light* otherwise. If all rows are light, then there are at most 1000 colored squares. If all columns are light, then there are at most 1000 colored squares. Neither of these is possible, so the number of light rows and columns together is at most 1998. In a trianglefree arrangement, no colored square lies in both a heavy row and a heavy column, so each of these lies uniquely in a light row or a light column; thus, there are at most 1998 colored squares in a trianglefree arrangment. If you have 1999 squares colored, you must have a triangle.

Problem 5 This can be proved by inverting in any circle centered at A_2, or directly, as follows. Without loss of generality we may assume that A_1, A_2, A_3 occur in that order going counterclockwise around the triangle $A_1A_2A_3$. Let θ_1 be the measure of the arc from A_1 to A_2 along ω_1, taken in the counterclockwise direction. Define $\theta_2, \ldots, \theta_7$ analogously, and let O_i be the center of ω_i. (See the FIGURE on p. 108.)

Adding the angles that meet at A_2 on line lO_1O_2 shows that

$$\frac{\theta_1}{2} + \frac{\theta_2}{2} = 2\angle A_1A_2A_3.$$

By similar reasoning we obtain the system of six equations:

$$\theta_1 + \theta_2 = 2\angle A_1A_2A_3, \quad \theta_2 + \theta_3 = 2\angle A_2A_3A_1,$$
$$\theta_3 + \theta_4 = 2\angle A_3A_1A_2, \quad \theta_4 + \theta_5 = 2\angle A_1A_2A_3$$
$$\theta_5 + \theta_6 = 2\angle A_2A_3A_1, \quad \theta_6 + \theta_7 = 2\angle A_3A_1A_2.$$

Adding the equations on the left column, and subtracting off the equations on the right yields $\theta_1 = \theta_7$.

This last equality implies $\omega_1 = \omega_7$, because we are using signed angles, and ω_1 determines a unique location for O_1 along the perpendicular bisector of A_1A_2. Thus the angle determines the circle.

Problem 6 (Based on work by George Lee) Define

$$L(a_1, b_1, \ldots, a_n, b_n) = \sum_{i,j} \min\{a_i b_j, a_j b_i\} - \min\{a_i a_j, b_i b_j\}.$$

Our goal is to show that

$$L(a_1, b_1, \ldots, a_n, b_n) \geq 0$$

for $a_1, b_1, \ldots, a_n, b_n \geq 0$ by induction on n, the case $n = 1$ being evident. Using the obvious identities

- $L(a_1, 0, a_2, b_2, \ldots) = L(0, b_1, a_2, b_2, \ldots) = L(a_2, b_2, \ldots)$,
- $L(x, x, a_2, b_2, \ldots) = L(a_2, b_2, \ldots)$,

and the less obvious but easily verified identities

- $L(a_1, b_1, a_2, b_2, a_3, b_3, \ldots) = L(a_1 + a_2, b_1 + b_2, a_3, b_3, \ldots)$ if $a_1/b_1 = a_2/b_2$,
- $L(a_1, b_1, a_2, b_2, a_3, b_3, \ldots) = L(a_2 - b_1, b_2 - a_1, a_3, b_3, \ldots)$ if $a_1/b_1 = b_2/a_2$ and $a_1 \leq b_2$,

we may deduce the result from the induction hypothesis unless we are in the following situation:

1. all of the a_i and b_i are nonzero;
2. for $i = 1, \ldots, n$, $a_i \neq b_i$;
3. for $i \neq j$, $a_i/b_i \neq a_j/b_j$ and $a_i/b_i \neq b_j/a_j$.

For $i = 1, \ldots, n$, let $r_i = \max\{a_i/b_i, b_i/a_i\}$. Without loss of generality, we may assume $1 < r_1 < \cdots < r_n$, and that $a_1 < b_1$. Now notice that the function $f(x) = L(a_1, x, a_2, b_2, \ldots, a_n, b_n)$ is *linear* in x over the interval $[a_1, r_2 a_1]$. Explicitly,

$$\begin{aligned} f(x) &= \min\{a_1 x, x a_1\} - \min\{a_1^2, x^2\} + L(a_2, b_2, \ldots, a_n, b_n) \\ &\quad +2\sum_{j=2}^{n} (\min\{a_1 b_j, x a_j\} - \min\{a_1 a_j, x b_j\}) \\ &= (x - a_1)(a_1 + 2\sum_{j=2}^{n} c_j) + L(a_2, b_2, \ldots, a_n, b_n), \end{aligned}$$

where $c_j = -b_j$ if $a_j > b_j$ and $c_j = a_j$ if $a_j < b_j$.

In particular, since f is linear, we have

$$f(x) \geq \min\{f(a_1), f(r_2 a_1)\}.$$

Note that $f(a_1) = L(a_1, a_1, a_2, b_2, \ldots) = L(a_2, b_2, \ldots)$ and

$$\begin{aligned} f(r_2 a_1) &= L(a_1, r_2 a_1, a_2, b_2, \ldots) \\ &= \begin{cases} L(a_1 + a_2, r_2 a_1 + b_2, a_3, b_3, \ldots) & \text{if } r_2 = b_2/a_2, \\ L(a_2 - r_2 a_1, b_2 - a_1, a_3, b_3, \ldots) & \text{if } r_2 = a_2/b_2. \end{cases} \end{aligned}$$

Thus we deduce the desired inequality from the induction hypothesis in all cases.

Acknowledgments We wish to thank Byron Walden of Santa Clara University for improving the text of the solutions.

CONTENTS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036