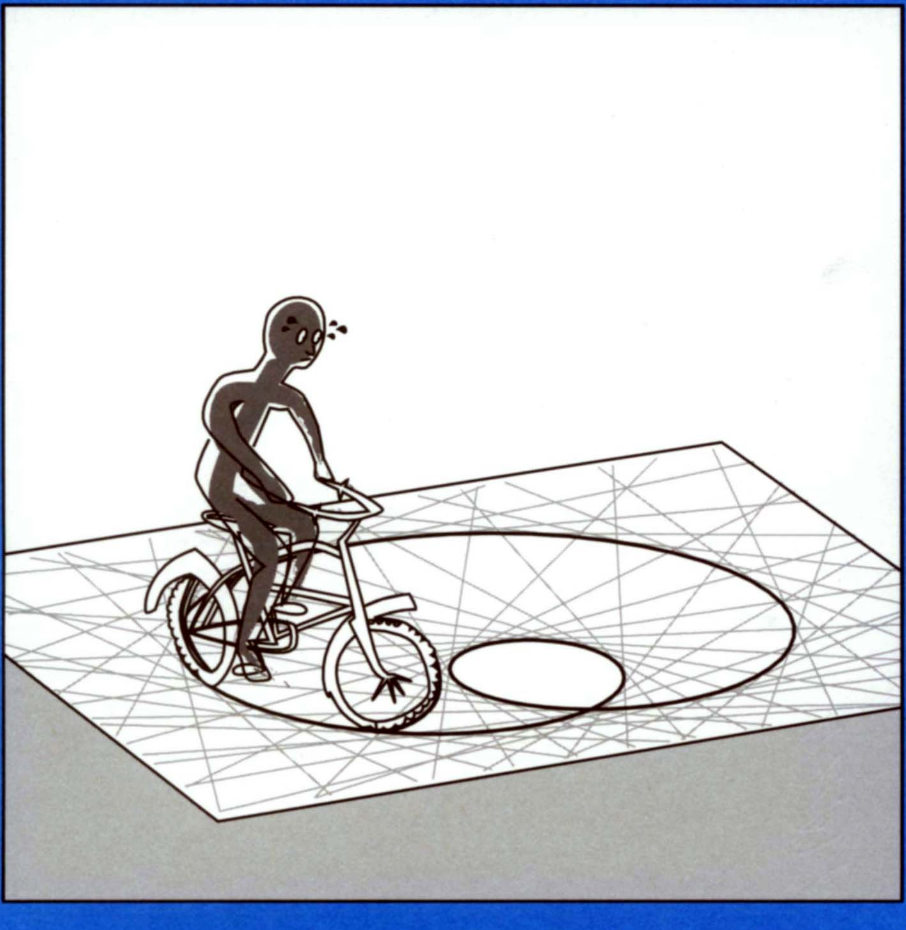
Vol. 74, No. 4, October 2001

MATHEMATICS MAGAZINE



- Zigzags
- The Track of a Bicycle Back Tire

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, *Riding a Bike around a Whirligig?*, by Jason Challas, who zigs and zags and teaches Computer Art at Santa Clara University.

AUTHORS

Peter Giblin is professor of mathematics and head of the Department of Mathematical Sciences at the University of Liverpool, England. His research interests are in singularity theory, applied to differential geometry and computer vision. He is a regular visitor to the United States and has been a visiting professor at the University of North Carolina at Chapel Hill, the University of Massachusetts at Amherst, and Brown University. He has been intrigued and puzzled by zigzags for several years, having drawn them on computer screens using everything from Basic to Java. Besides mathematics and his family and friends, he enjoys music (including playing piano duets), theatre, films, cycling to work, and the work that awaits him when he gets there.

Steven R. Dunbar received a bachelor's degree in mathematics at the University of Nebraska, and doctorate at the University of Minnesota in 1981, and returned to the University of Nebraska-Lincoln in 1985. His research interests are in nonlinear differential equations, and applications of mathematics in all areas, especially physics, which led to collaboration with Reinier Bosman and Sander Nooij. When not solving differential equations, he enjoys gardening and, of course, riding his bicycle.

Reinier Bosman was born in 1975 on Curacao, an island located in the Caribbean. In 1977 his family returned to Holland where he attended preliminary school and high school in The Hague area. After graduation in 1995, he studied physics at the University of Amsterdam. On a study trip with Sander Nooij to Nebraska for a bike project, he met Professor Dunbar. There the idea was born to do some research about the mathematical connection of the front and back wheels of a bicycle. Last August he received his Master's degree in theoretical physics with a thesis on biological neural networks. After graduation, he is planning to start a career in business.

Sander Nooij was born in Haarlem, the Netherlands in 1977. After school in Haarlem, he continued his education at the University of Amsterdam studying physics. Further study took him to the University of Nebraska, Lincoln to work on the bike project. In 1998, as part of this project, he and Reinier Bosman made a 300-mile bike trip across northern Nebraska. One year later he went to the Particle Accelerator CERN in Geneva, Switzerland, where he worked on the detection of supersymmetry in physics. Last June he received his Master's degree in theoretical physics. Sander will continue his studies by reading towards a Ph.D. degree in theoretical physics at the University of Oxford.

Vol. 74, No. 4, October 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

Zigzags

PETER J. GIBLIN
The University of Liverpool
L69 3BX, England
pjgiblin@liv.ac.uk

Look at the four diagrams in the figure. Each of them is produced by the same procedure, which we describe shortly. In the upper left is a family of lines (360 of them, in fact), which has a very clearly defined *envelope*, that is a curve tangent to all of the lines. Though not explicitly drawn in the figure, the envelope, which has 10 cusps or sharp points and 10 self-crossings, is immediately evident to the eye. In the upper right the lines produce not one connected envelope but three, each one of which has four cusps and no self-crossings, though the components cross one another. The lower left figure is rather curious: some lines are drawn and also a very loopy curve which, you can verify, is tangent to all of the lines. However it's a very poor excuse for an envelope of the lines, which should evidently have four cusps and two crossings. Finally, in the lower right the lines appear to have a number of circles for their envelope—how many do you see?

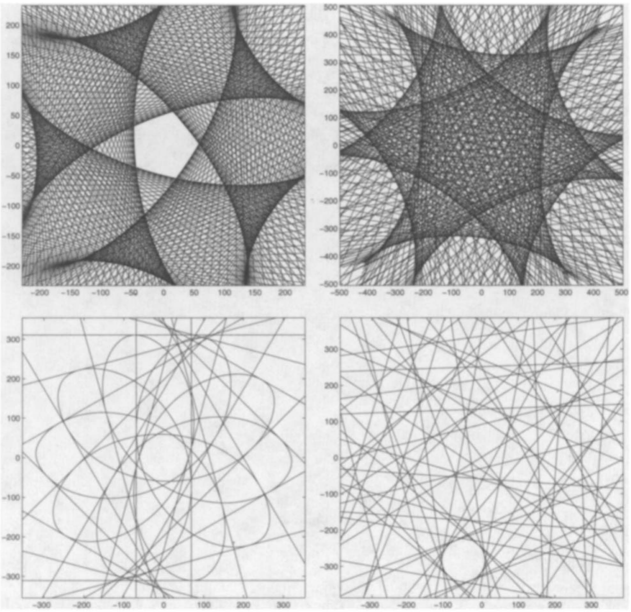


Figure 1 Some families of lines created by the zigzag construction, and, lower left, one incorrect attempt to draw their envelope

In this article I shall describe the way in which these finite sets of lines are generated: the *zigzag construction*. It is very striking that the lines often form visually evident envelopes; indeed that is what prompted this investigation in the first place—the challenge is to find a curve, or several curves, that form precisely this visually evident envelope of the lines. As the lower left example in the figure illustrates, an arbitrary curve tangent to all the lines may well be wrong from a visual standpoint.

If we are given a family of lines, say $a(t)x + b(t)y = c(t)$, parametrized by a continuous parameter t, then there is a standard method for finding the envelope curve: solve for x and y in the system consisting of this equation and its derivative with respect to t, namely $a'(t)x + b'(t)y = c'(t)$. (See for example [**2**, p. 57].) But for a *finite* family of lines such as those being considered here, we will have a very wide choice for an envelope curve tangent to all of them—how do we choose the visually correct one? I shall present one method, which works quite often—the *whirligig construction*—but I do not know the complete answer.

Two Java applets demonstrating these constructions are available. See http://www.liv.ac.uk/~tobyhall/Zigzag/ for the background leading up to the topic of this article, and http://www.liv.ac.uk/~pjgiblin/Zigzag/ for the specific envelopes considered here. Alternatively, visit the MAGAZINE website: http://www.maa.org/pubs/mathmag.html.

In this article I describe the zigzag and whirligig constructions, and give two ways of marrying the two, with numerous examples along the way. Finally, there is a discussion of the special case (such as FIGURE 1, lower right) where the envelope consists of a number of circles.

Zigzags

The basic idea of a zigzag is illustrated in FIGURE 2; the original idea comes from *Turtle Geometry* [**1**, p. 114]. A straight horizontal line—the zeroth *zig*—is drawn to the right from the origin, of length 100. At the end of this, another straight horizontal line—the zeroth *zag*—is drawn, of length l. If $l < 0$ then the line is drawn to the left and otherwise to the right; in either case it terminates at $(100 + l, 0)$. The vector $(100, 0)$ is denoted by $\mathbf{v}_1$ and $(l, 0)$ by $\mathbf{v}_2$, as in the figure. At this stage we say that *step zero*—a zig and a zag—has been completed. So far there is not much zigzagging in evidence.

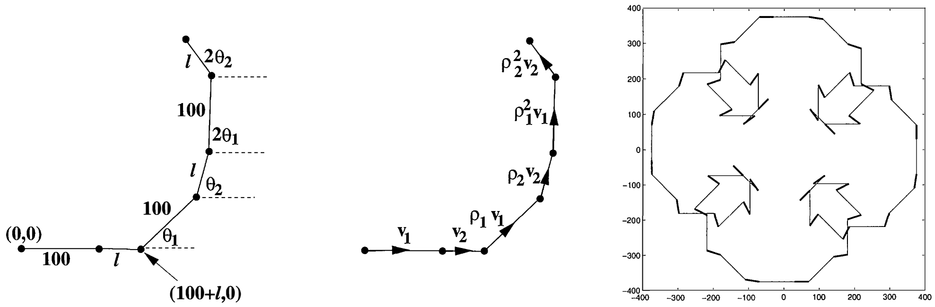


Figure 2 The basic zigzag, defined by lengths of 100 and l, and angles θ_1, θ_2. In the middle figure, ρ_1 and ρ_2 are rotations through θ_1 and θ_2 respectively. Right: a simple completed zigzag with $l = 40$, $\theta_1 = 45°$, $\theta_2 = 9°$, with the zags drawn heavily. Note that the origin here has been moved to the center of the zigzag. In all the remaining figures in this article, *only* the zags are drawn, and they are extended right across the viewing area.

But now the true zigzagging begins. We have two angles θ_1, θ_2 given to us, which will always be whole numbers of *degrees*. We draw a straight line—the first zig—of length 100 from $(100 + l, 0)$, at an angle θ_1 with the positive x-axis (so this angle is measured counterclockwise from this axis). The termination of this line is at $(100 +$

$l + 100\cos\theta_1, 100\sin\theta_1)$. From this point we draw a line—the first zag—of length l at an angle θ_2 with the horizontal, thereby arriving at the point

$$(100 + l + 100\cos\theta_1 + l\cos\theta_2, 100\sin\theta_1 + l\sin\theta_2).$$

At this stage, step one has been completed. In FIGURE 2, ρ_1 and ρ_2 are counterclockwise rotations through θ_1 and θ_2.

The lengths of the added lines are always alternately 100 and l. However, the *angles* between the added lines and the horizontal go up by θ_1 and θ_2 respectively at every step. Thus step two consists of drawing two lines at angles of $2\theta_1, 2\theta_2$ to the horizontal, step three of drawing two lines at angles of $3\theta_1, 3\theta_2$ to the horizontal, etc.

There is some resemblance between the above construction and one described by Maurer [**4**], but we use a pair of angles and he uses one angle.

The figures in this article are examples of sets of *zags only*, which are extended across the page to give the envelopes a chance to form. Of course one could also consider the zigs alone and obtain analogous pictures and results. Note that in all the figures, the origin has been translated to the center of the zigzag, as in the analysis that follows.

We shall need the equation of the j^{th} zag, where $j = 0$ indicates the original horizontal zag of length l. This (oriented) zag makes an angle of $j\theta_2$ with the horizontal drawn to the right. It is convenient to translate the axes parallel to themselves so that they pass through the center $\mathbf{c}$ of the zigzag. For a full account, see *Mathematical Explorations with MATLAB* [**2**, Ch. 11]; the following equation, as well as an expression for $\mathbf{c}$ is derived below:

$$x\sin(j\theta_2) - y\cos(j\theta_2) = \frac{50}{\sin\left(\frac{1}{2}\theta_1\right)}\cos(j(\theta_1 - \theta_2) + \tfrac{1}{2}\theta_1) + \tfrac{1}{2}l\cot(\tfrac{1}{2}\theta_2). \tag{1}$$

As well as considering all the zags for given θ_1, θ_2 and l, it is interesting to select just a subset by starting from $j = k_0$ and increasing j by $\delta > 0$ at each step, that is to consider only the j^{th} zags for $j = k_0 + n\delta$, $n = 0, 1, 2, \ldots$ in (1). In practice we shall usually take $k_0 = 0$, taking "every δ^{th} zag." See, for example, FIGURE 3 where showing every eighth zag (b) gives a very different result from showing every fifth zag (c). Of course, if there were 40 zags, then drawing every zag ($\delta = 1$) is the same as drawing every third zag ($\delta = 3$), since 40 is not a multiple of 3, though the way in which these zags *step round* the envelope curve may well be different. We shall expand on the latter idea when we relate zigzags to whirligigs. For the present, here is a formula ([**2**, p. 132]) for the total number $s(\delta)$ of zags that occur before the figure closes and repeats. It is assumed that θ_1, θ_2 are integers fixed in advance, so we do not include them in the notation for s.

$$s(\delta) = \frac{360}{(360, \delta\theta_1, \delta\theta_2)}, \tag{2}$$

the round brackets denoting the greatest common divisor. Two values of δ (both with $k_0 = 0$) will give the same *set* of zags precisely when they give the same *number* of zags (we are dealing here essentially with an additive cyclic group of order $s(1)$, generated by 1, and the subgroup generated by δ).

Derivation of the equation of a zag For the time being the origin remains at the beginning of the zigzag. Then define $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}$ from the initial zig and zag vectors, using the rotations ρ_1 and ρ_2 as follows:

$$\mathbf{v}_1 = \mathbf{c}_1 - \rho_1\mathbf{c}_1, \quad \mathbf{v}_2 = \mathbf{c}_2 - \rho_2\mathbf{c}_2, \quad \mathbf{c} = \mathbf{c}_1 + \mathbf{c}_2.$$

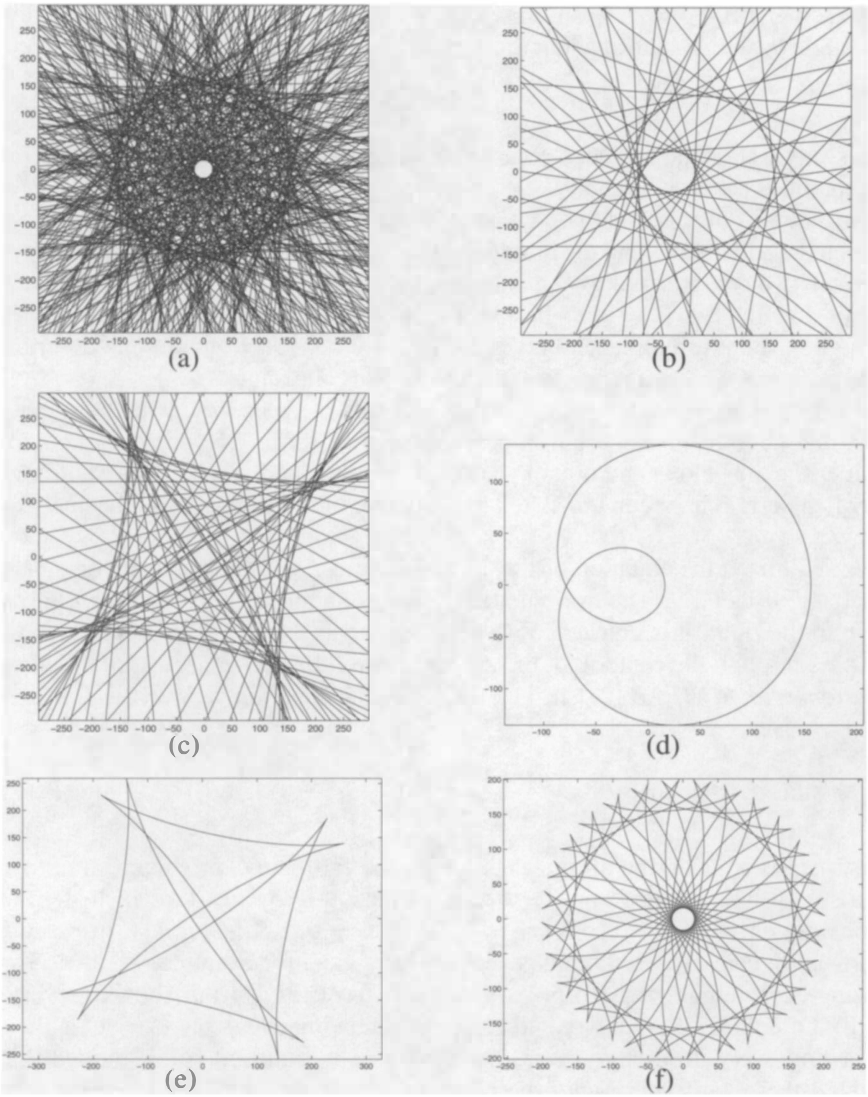


Figure 3 $l = 75$, $\theta_1 = 91°$, $\theta_2 = 47°$. (a) The whole set of 360 zags, making a mess. (b) With $\delta = 8$, $k_0 = 0$ we pick out one of eight envelope components, with $360/8 = 45$ zags tangent to it. The other seven components (obtained from this one by rotation) are given by $k_0 = 1, \dots, 7$. (c) With $\delta = 5$, $k_0 = 0$ we pick out one of five components, with $360/5 = 72$ zags tangent to it. (d) The envelope of (b), generated by a continuous family of lines. (e) The envelope of (c), generated by a continuous family of lines. The whirligig curves here are determined by the method of direct comparison. (f) A whirligig tangent to *all* the zags, that is to all the lines in (a); this is hardly a visually striking envelope!

Operating on the first of these equations by $\rho_1, \rho_1^2, \dots, \rho_1^j$ and adding, we get

$$\mathbf{v}_1 + \rho_1\mathbf{v}_1 + \rho_1^2\mathbf{v}_1 + \dots \rho_1^j\mathbf{v}_1 = \mathbf{c}_1 - \rho_1^{j+1}\mathbf{c}_1,$$

with a similar equation having subscript 2 throughout.

Now the point which the zigzag reaches after j steps—the end of the j^{th} zag—can be seen in FIGURE 2 to be

$$\mathbf{v}_1 + \mathbf{v}_2 + \rho_1\mathbf{v}_1 + \rho_2\mathbf{v}_2 + \dots + \rho_1^j\mathbf{v}_1 + \rho_2^j\mathbf{v}_2,$$

where as before $j = 0$ means the starting zig and zag, both of which are horizontal. Using the above formulae this becomes

$$\mathbf{c}_1 + \mathbf{c}_2 - \rho_1^{j+1}\mathbf{c}_1 - \rho_2^{j+1}\mathbf{c}_2.$$

Translating the origin to the point $\mathbf{c} = \mathbf{c}_1 + \mathbf{c}_2$, called the *center* of the zigzag, we can drop the first two terms. For an explicit formula, replace $\mathbf{c}_1$ by $(I - \rho_1)^{-1}\mathbf{v}_1$ (where I is the identity), and similarly for $\mathbf{c}_2$; after applying some trigonometric identities, we find that the two ends of the j^{th} zag are

$$\left(\frac{100}{2\sin\frac{1}{2}\theta_1}\sin\left(j+\tfrac{1}{2}\right)\theta_1 + \frac{l}{2\sin\frac{1}{2}\theta_2}\sin\left(j\pm\tfrac{1}{2}\right)\theta_2,\right.$$
$$\left. -\frac{100}{2\sin\frac{1}{2}\theta_1}\cos\left(j+\tfrac{1}{2}\right)\theta_1 - \frac{l}{2\sin\frac{1}{2}\theta_2}\cos\left(j\pm\tfrac{1}{2}\right)\theta_2\right),$$

where the lower sign is the beginning of the zag and the upper sign is the end.

We can now check that the line (1) has the correct slope $j\theta_2$ and passes through one of the above points (or alternatively passes through both points), and is hence the line along the j^{th} zag. This completes the proof that (1) gives the equation of this zag relative to axes parallel to the original axes but translated to the center $\mathbf{c}$ of the zigzag.

Whirligigs

Imagine a circle that spins at the same time as its center travels around another circle. Now attach a tangent line to the spinning circle. The envelope of this set of lines is a whirligig curve. These are the most general kind of envelope that we will use to compare with the zag-envelopes.

For a precise definition, consider a circle of radius R centered at the origin (which by assumption is now the center $\mathbf{c}$ of the zigzag above). As in FIGURE 4, locate a point on the circumference, and call $\phi(t)$ the radial angle with the downward vertical; using this point as the center, draw another circle, of radius r. Orienting this circle counterclockwise, consider the (oriented) tangent to this circle making an angle $\psi(t)$ with the positive x-axis. We shall take ϕ, ψ to be *linear* functions of t, so that the speeds of rotation are constant:

$$\phi(t) = at + b, \quad \psi(t) = ct + d, \quad a, b, c, d \text{ constants.} \tag{3}$$

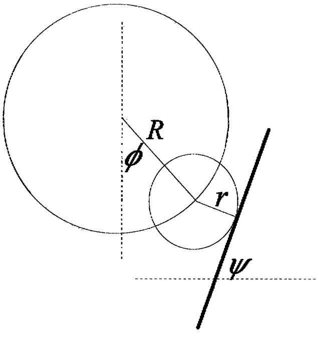


Figure 4 A *whirligig* is the envelope of lines tangent to a rotating circle of radius r whose center moves on a circle of radius R. The angles ϕ, ψ are then functions of 'time' t.

A brief analysis shows that the point of contact of the tangent line with the circle of radius r is

$$(R\sin\phi + r\sin\psi, \quad -R\cos\phi - r\cos\psi),$$

and that the equation of the tangent line is

$$x\sin\psi - y\cos\psi = R\cos(\phi-\psi) + r. \tag{4}$$

The whirligig determined by R, r, a, b, c and d is, then, the envelope of these tangent lines as the spinning circle of radius r moves round the circle of radius R. Whenever we draw a whirligig we shall draw simply the curve itself which is tangent to all these lines. Mathematically, this is obtained by solving for the variables x and y in the system comprising the equation (4) and the derivative of the same equation with respect to t.

For the record, here is the resulting parametrization:

$$cx = Rc\sin\phi + rc\sin\psi - Ra\sin(\phi-\psi)\cos\psi,$$
$$cy = -Rc\cos\phi - rc\cos\psi - Ra\sin(\phi-\psi)\sin\psi.$$

As a simple example, if $a = c \neq 0$ then the whirligig is a circle with center the origin. If $a = 0$, $c \neq 0$, it is a circle with center at some point of the circle radius R.

Remark. It is worth noting that what is here called a *whirligig* appears also in the literature as a *line-roulette* or more specifically a *line trochoid*; see, for example, Lockwood [**3**, Ch. 17]. The connection is not immediate since it is usual to require a circle to *roll* on a fixed circle, the rolling circle carrying with it a point, giving a *point-trochoid*, or a line, giving as envelope a *line-trochoid*. Note that it is not assumed that the moving point (or line) is on the circumference of the rolling circle (or tangent to the rolling circle).

In fact, it is not hard to see that, in FIGURE 4, we can always find a circle concentric with our fixed circle and a circle concentric with the spinning circle that *do* roll on one another. Taking $a > 0$ there are three cases according to whether $c < 0$, $0 < c < a$, or $c > a$ and the reader may enjoy finding the radii of the fixed and rolling circles when the rolling condition is imposed. For example, when $0 < c < a$ the radii are $(a-c)R/c$ and aR/c. Of course r now plays the role of telling us the location of the line rigidly attached to the rolling circle whose envelope produces the line-trochoid.

The purpose of introducing whirligigs here is to compare (4) with the equation of the j^{th} zag. If every zag is one of these lines then the envelope of the lines—that is, the whirligig—will be tangent to all the zags and so may serve as an envelope of the zags. On the other hand the whirligig may turn out to be much more complicated than the visually evident envelope of the zags; see FIGURE 1, lower left, for an example of a whirligig that is, to be sure, tangent to all the zags, but is visually wrong. The correct whirligig is the one shown in FIGURE 5, left.

Remark. It is clear that the whirligigs in the figures often have cusps. Here is a formula for the number of cusps, when a and c are integers; verification is left as a pleasant exercise for the reader:

$$\frac{2|a-c|}{(a,c)}.$$

We often take a and c relatively prime, so the number is then $2|a-c|$.

Zigzags and whirligigs: direct comparison method

First, a direct comparison between (1) and (4) shows that it makes good sense to take

$$R = \frac{50}{\sin \frac{1}{2}\theta_1}, \quad r = \tfrac{1}{2} l \cot \tfrac{1}{2}\theta_2, \tag{5}$$

and we shall always do this. It remains to choose a and c.

Let θ_1, θ_2 and δ be given integers. We consider the zags with $j = 0, \delta, 2\delta, \ldots$. Let

$$\delta\theta_1 \equiv k_1 \quad \text{and} \quad \delta\theta_2 \equiv k_2 \mod 360. \tag{6}$$

We shall usually take the k_i to be the smallest positive residues mod 360, or the residues that are smallest in absolute value.

Proposition. *Suppose that a and c are integers and that there exists τ with $a\tau$, $b\tau$ integers and*

$$a\tau \equiv k_1, \quad c\tau \equiv k_2 \mod 360. \tag{7}$$

Then all the zags ($j = 0, \delta, 2\delta, \ldots$ as above) are lines of the form (4) with the above a, c and $b = \frac{1}{2}\theta_1$, $d = 0$; the zags are therefore tangent to the whirligig given by these values (with R, r as in (5), as usual).

Remark. We can use the zags with $j = k_0, k_0 + \delta, k_0 + 2\delta, \ldots$ by adjusting the values of b and d to $(k_0 + \frac{1}{2})\theta_1$, $k_0\theta_2$ respectively.

Proof. Take $\phi = at + \frac{1}{2}\theta_1$, $\psi = ct$, $t = n\tau$ in (4), where a, c, τ satisfy (7). Then the line (4) clearly coincides with the zag

$$x \sin(n\delta\theta_2) - y \cos(n\delta\theta_2) = R \cos(n\delta(\theta_1 - \theta_2) + \tfrac{1}{2}\theta_1) + r$$

for $n = 0, 1, 2, \ldots$. ■

We shall use this simple proposition to propose whirligigs as possible envelopes of zags. Experiment suggests that a more visually plausible result is obtained if a and c are reasonably small, but the total number of zags drawn (given by $s(\delta)$ as in (2)) is reasonably large. However, *small* here must be taken with a grain of salt; for instance, FIGURE 7 shows an example where $a = 17$, $c = 8$ and the whirligig is clearly right.

An immediate solution to (7) is $a = k_1$, $c = k_2$, $\tau = 1$. Note that we can take out a common factor from a and c in (7) by multiplying τ by the same factor. So we can take out all common factors and assume that a and c are *relatively prime*. Thus

$$a = \frac{k_1}{(k_1, k_2)}, \quad c = \frac{k_2}{(k_1, k_2)}, \quad \tau = (a, c) \tag{8}$$

is a simpler solution. Determining a and c this way will be called the *direct comparison* method. We shall consider other solutions in a later section.

Before giving a number of examples, it is worth introducing the notion of *stepping round* the whirligig. Suppose that we have found a and c that give a visually correct envelope, as in FIGURE 3(c) and (e). The family of lines in (c) can be generated by taking $\delta = 5$ or 15 or 30 or 85, or any other δ with $(360, 91\delta, 47\delta) = 5$, which here amounts simply to δ being a multiple of 5. (To obtain $a = 5$ and $c = 1$ by the method of direct comparison, we can take $\delta = 85$ or 115; see Example 1 below. We can, in fact, obtain $a = 5$, $c = 1$ from $\delta = 5$ by an alternative method, developed in what follows.) If we take one of these values of δ and draw every δ^{th} zag, starting with the 0^{th} zag, then these will eventually fill up all of FIGURE 3(c) but will in general dance about over the whirligig curve in (e) rather than stepping along it with the points of contact covering the whirligig just once. When they *do* cover it just once in order we say that this value of δ makes the zags *step round* the whirligig.

Suppose that a and c, with $(a, c) = 1$, are determined by the method of direct comparison, from a particular value of δ. For the whirligig, the small spinning circle turns $|c|$ times before returning to its starting place. On the other hand consider the zags given by $j = \delta n$, $n = 0, 1, 2, \ldots$. The number of zags before the whole zigzag repeats is given by (2), that is,

$$\frac{360}{(360, \delta\theta_1, \delta\theta_2)} = \frac{360}{(360, k_1, k_2)}.$$

We shall take k_1, k_2 to be the least residues of $\delta\theta_1, \delta\theta_2 \bmod 360$, in the sense of absolute value. A negative value indicates that the selected zags (multiples of δ) turn clockwise instead of counterclockwise. The total number of turns of the zag before returning to the start is therefore the above number times k_2, divided by 360. If drawing every δ^{th} zag is to step round the whirligig defined as in the proposition, we require that $(k_1, k_2) = (360, k_1, k_2)$, which is the same as saying that (k_1, k_2) is a factor of 360:

> *The stepping round criterion for the direct comparison method is that* (k_1, k_2) *divides exactly into* 360.

Examples are given in the next section.

Stepping round is hard to demonstrate with a still picture, but the second Java applet mentioned previously allows a delay between the drawing of successive zags; this makes the idea immediately attractive.

Examples

Example 1. $\theta_1 = 91$, $\theta_2 = 47$. Table 1 shows all the values of $a = k_1/(k_1, k_2)$, $c = k_2/(k_1, k_2)$ which are both ≤ 10, for values of δ from 1 to 180. The number s is the number of steps (here the number of zags) in a complete cycle; as above, this equals $360/(360, k_1, k_2)$. Note that the outlandish $a = 17$, $c = -11$ of FIGURE 3(f) is not in the table because of the cutoff value of 10 for a and c.

The first entry in the table, $\delta = 8$, gives FIGURE 3(b), (d). The second entry, $\delta = 15$, has a and c which are merely the negatives of those in the more interesting entry $\delta = 115$, and the latter has three times as many zags tangent to it. It is shown in FIGURE 3(c),(e). The third entry, $\delta = 16$, is like $\delta = 8$, but without the feature that the zags step round the whirligig. (This feature is indicated in the table by a 1 in the column headed ? and its absence by a 0.) When $\delta = 18$, the number of zags (20) is quite small, and a and c are relatively large, so this is a complicated whirligig with the zags spaced very far apart (in terms of arclength) along it. From the zags one would never pick out this whirligig as their envelope. For another entry in the table, $\delta = 90$ gives $a = c = -1$, a circle, with just 4 zags tangent to it. Note that there are no values of $\delta \leq 180$ for which the number of steps s is > 72. Thus (at least for this range of δ) it is to be expected that the 72-zag whirligigs are not part of larger ones; in fact that there are $360/72 = 5$ of these which are obtained by choosing different starting points, that is, different values of k_0. In the above, k_0 is always chosen to be 0.

TABLE 1: $\theta_1 = 91, \theta_2 = 47$. This table shows, for δ up to 180, values of a and c, computed by the direct comparison method, and both ≤ 10 in modulus, giving whirligigs that are tangent to all the s resulting zags. Low values of a, c and high values of s tend to give recognizable envelopes. The column headed ? contains a 1 if the zags *step round* the whirligig and a 0 otherwise.

δ	k_1	k_2	a	c	?	s
8	8	16	1	2	1	45
15	−75	−15	−5	−1	1	24
16	16	32	1	2	0	45
18	−162	126	−9	7	1	20
20	20	−140	1	−7	1	18
24	24	48	1	2	1	15
30	−150	−30	−5	−1	1	12
32	32	64	1	2	0	45
36	36	−108	1	−3	1	10
40	40	80	1	2	1	9
45	135	−45	3	−1	1	8
48	48	96	1	2	0	15
54	−126	18	−7	1	1	20
56	56	112	1	2	0	45
60	60	−60	1	−1	1	6
63	−27	81	−1	3	0	40
64	64	128	1	2	0	45
72	72	144	1	2	1	5
75	−15	−75	−1	−5	1	24
80	80	160	1	2	0	9
84	84	−12	7	−1	1	30
85	175	35	5	1	0	72
88	88	176	1	2	0	45
90	−90	−90	−1	−1	1	4
96	96	−168	4	−7	1	15
99	9	−27	1	−3	1	40
100	100	20	5	1	1	18
108	108	36	3	1	1	10
115	25	5	5	1	1	72
120	120	−120	1	−1	1	3
126	−54	162	−1	3	0	20
130	−50	−10	−5	−1	1	36
135	45	−135	1	−3	1	8
138	−42	6	−7	1	1	60
140	140	100	7	5	1	18
144	144	−72	2	−1	1	5
145	−125	−25	−5	−1	0	72
150	−30	−150	−1	−5	1	12
160	160	−40	4	−1	1	9
162	−18	54	−1	3	1	20
168	168	−24	7	−1	1	15
170	−10	70	−1	7	1	36
180	180	180	1	1	1	2

For the remaining examples, we shall not give so much detail.

Example 2. $\theta_1 = 45°, \theta_2 = 15°$. (Take $l = 50$.)

δ	k_1	k_2	a	c	?	s	Comment
1	45	15	3	1	1	24	Correct visually: FIGURE 5, left
7	−45	105	−3	7	1	24	Wrong visually: FIGURE 1, lower left

Example 3. $\theta_1 = 77°, \theta_2 = 22°$. (Take $l = 50$.)

δ	k_1	k_2	a	c	?	s	Comment
131	7	2	7	2	1	360	See FIGURE 1, upper left
72	144	144	1	1	0	5	Regular pentagon and circle: k_0 changes radius

Example 4. $\theta_1 = 21°$, $\theta_2 = 47°$ (take $l = 50$). Here $\delta = 1$ gives all of FIGURE 1, upper right, and $\delta = 3$ gives just one-third of the zags, which are tangent to one of the three curves visible in this figure. The value $\delta = 39$ gives the same zags as $\delta = 3$,

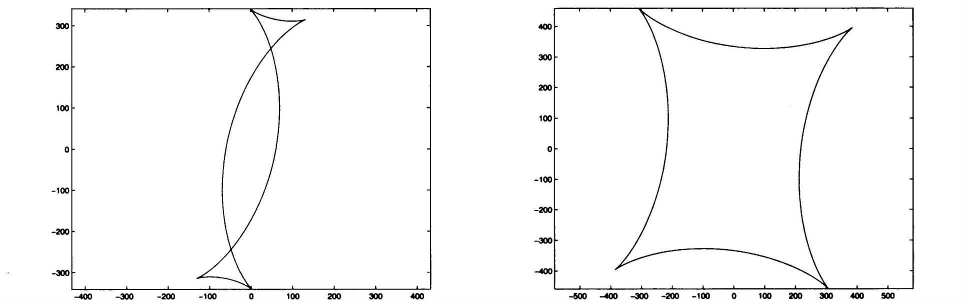


Figure 5 Left: the visually correct envelope for the family of lines in FIGURE 1, lower left (see Example 2). Here $\theta_1 = 45°$, $\theta_2 = 15°$, $l = 50$. Right: one of the three components of the visually correct envelope for the family in FIGURE 1, upper right (see Example 4). Here $\theta_1 = 21°$, $\theta_2 = 47°$, $l = 50$.

but the minimization method of the next section shows that $a = 3, c = 1$ provides an appropriate whirligig as in FIGURE 5, right. To get the other components with $\delta = 39$ we take $k_0 = 1, 2$ and adjust b and d as in the Remark above. The value $\delta = 69$ gives the same zags again as $\delta = 3$ but in addition steps round the figure.

Example 5. $\theta_1 = 21°$, $\theta_2 = 49°$ (take $l = 50$). Here $\delta = 1$ gives $a = 3$, $c = 7$ by direct comparison. See FIGURE 6. In order to step round the envelope we can take $\delta = 103$.

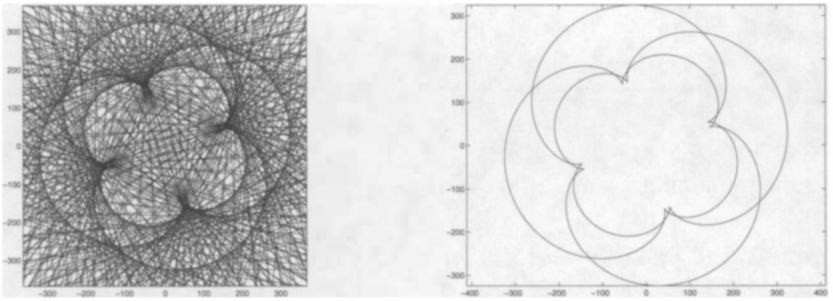


Figure 6 Left: $\theta_1 = 21°$, $\theta_2 = 49°$, $l = 50$. The envelope has one piece, though it is hard to be sure what it looks like. Right: the envelope produced as the envelope of a continuous family of lines. See Example 5.

Example 6. $\theta_1 = 23°$, $\theta_2 = 32°$ (take $l = 50$). Here, as in FIGURE 7, we need to take the larger values $a = 17$, $c = 8$, given by $\delta = 79$ (direct method). This value of δ also steps round the resulting whirligig.

The minimization method

The formula (8) is not the only solution to the equations (7) for finding a whirligig that is tangent to all the zags. For example, with $\theta_1 = 91°$, $\theta_2 = 47°$, $\delta = 5$ it misses the good solution $a = 5,\ c = 1$, which gives the same picture as FIGURE 3(c),(e).

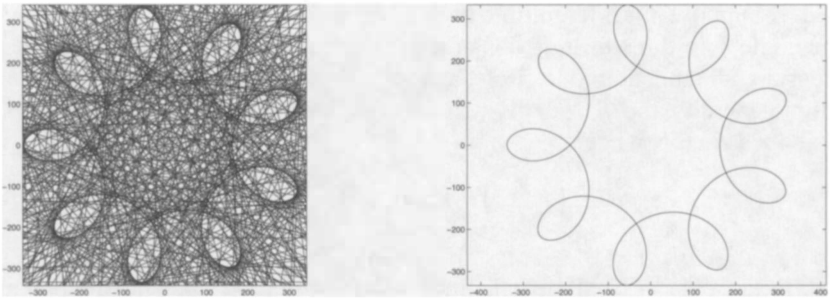


Figure 7 The case $\theta_1 = 23°$, $\theta_2 = 32°$, $l = 50$ where we need $a = 17$, $c = 8$ to generate the envelope (right). See Example 6.

Here is a sketch of another possible method, called the *minimization method.* To simplify notation we rewrite (7) as

$$a\tau \equiv u, \quad c\tau \equiv v \quad \text{mod } w. \tag{9}$$

We shall assume in what follows that τ is an *integer* and that $(a, c) = 1$. We seek to *minimize the value of a*. Let $(\tau, w) = h$; then (9) implies $h|u$, $h|v$. Write $u = u_1 h$, $v = v_1 h$, $w = w_1 h$, $\tau = \tau_1 h$ so that $(\tau, w_1) = 1$ and (9) can be replaced by

$$a\tau_1 \equiv u_1, \quad c\tau_1 \equiv v_1 \quad \text{mod } w_1.$$

It now follows, using $(a, c) = 1$, that u_1, v_1, w_1 cannot all have a common factor so that in fact $h = (u, v, w)$.

Since $(\tau_1, w_1) = 1$ we can find the inverse $s_1 \equiv \tau_1^{-1} \text{ mod } w_1$ and then $a \equiv s_1 u_1$, $c \equiv s_1 v_1 \text{ mod } w_1$ is a solution. So we proceed as follows:

Let $g = (u_1, w_1)$, $u_1/g = u_2$, $w_1/g = w_2$ and, to make a as small as possible, choose for s_1 the number $u_2^{-1} \text{ mod } w_2$. Then $a \equiv g \text{ mod } w_1$ (so take $a = g$) and $c \equiv s_1 v_1 \text{ mod } w_1$. Note that there is a possibility here that $(s_1, w_1) > 1$, even though $(s_1, w_2) = 1$. This would prevent us from deducing that a and c satisfy (9), since we could not choose $\tau_1 \equiv s_1^{-1} \text{ mod } w_1$. There is also the possibility that a and c are not, in fact, relatively prime. ($\theta_1 = 73°$, $\theta_2 = 26°$, $\delta = 36$ makes $(a, c) = 2$ by this method.) So when using this method we need to check both of these conditions en route.

Of course, we can minimize c by the same method. As an example, let $\theta_1 = 91°$, $\theta_2 = 47°$, as in Table 1 where the direct method was used to find a and c. Then $\delta = 5, 25, 35, 65, \ldots$ all give $s = 72$ steps, as in FIGURE 3(c), and the new method correctly predicts $a = 5$, $c = 1$ is a solution here.

Note that the *stepping round* criterion is slightly different now. We need $\tau s(\delta) = w = 360$ and since $s(\delta) = 360/(u, v, w)$ and $\tau = \tau_1(u, v, w)$ we need $\tau_1 = 1$ (or -1).

The stepping round criterion for the minimization method is that $\tau_1 = \pm 1$.

The special case when the zags are tangent to circles

A glance at FIGURES 8, 9 and 10 shows that there are a number of situations where the zags are all tangent to one or more circles. We consider some of these here.

Case 1. Suppose $\delta\theta_1$ is a multiple of 360. Then clearly we can take $a = 0$ in (7), that is, the angle ϕ in the whirligig construction (FIGURE 4) is *constant*. This means that the zags are all tangent to one circle, as in FIGURE 8, left.

More generally, the j^{th} zag (1) will coincide with the line (4), for the usual values of R and r as in (5), when

$$\phi \equiv \left(j + \tfrac{1}{2}\right)\theta_1, \quad \psi \equiv j\theta_2 \mod 360.$$

Write $j = k_0 + n\delta, n = 0, 1, 2, \ldots$. If $m\delta\theta_1$ is a multiple of 360 for an integer m then $n = 0, 1, 2, \ldots, m-1$ will give distinct ϕ and the zags will be tangent to m circles in turn which will then repeat. See FIGURE 8, right, where $\theta_1 = 20°$, $\theta_2 = 49°$, $l = 50$.

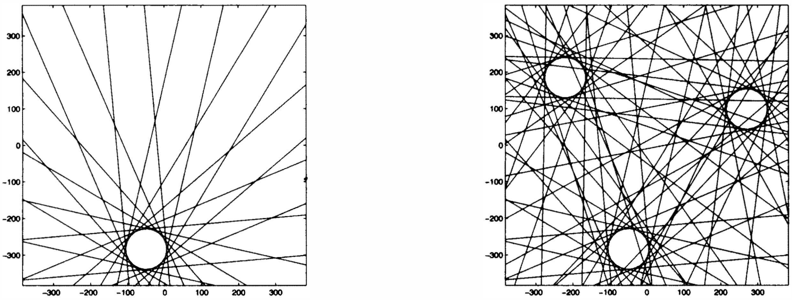


Figure 8 Two examples where the zags are all tangent to one or more circles. Here, θ_1 is a factor of 360, in fact 20°, and $\delta = 18$ on the left, $\delta = 6$ on the right. Here $\theta_2 = 49°$, $l = 50$.

Case 2. A *different* way of identifying (1) and (4) is to take

$$\phi \equiv j(2\theta_2 - \theta_1) - \tfrac{1}{2}\theta_1, \quad \psi \equiv j\theta_2 \mod 360.$$

This makes use of the evenness of the cosine function. If now $\delta(2\theta_2 - \theta_1)$ is a multiple of 360, then $j = k_0 + n\delta$ will give a constant ϕ (mod 360), namely $\phi \equiv k_0(2\theta_2 - \theta_1) - \frac{1}{2}\theta_1$. Thus all the zags will be tangent to one circle. For example, FIGURE 9, right, we have $\theta_1 = 34°$, $\theta_2 = 37°$ (and $l = 30$), so that $2\theta_2 - \theta_1 = 40 = 360/9$. Then $\delta = 9$ gives one circle and $\delta = 1$, by an argument similar to Case 1, gives nine circles, as shown in the figure. (If $m\delta(2\theta_2 - \theta_1) = 360N$ where $(m, N) = 1$, then we get m circles.)

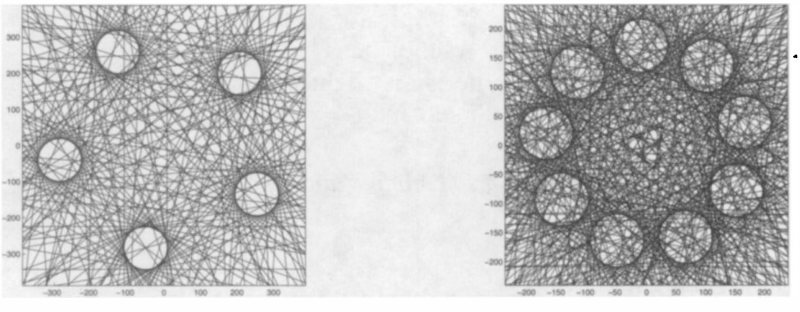


Figure 9 Left: five circles produced by $\theta_1 = 20°$, $\theta_2 = 46°$, $\delta = 1$, making $2\theta_2 - \theta_1 = 2\pi/5$. Taking $\delta = 5$ reduces to a single circle. Right: nine circles produced by $\theta_1 = 34°$, $\theta_2 = 37°$, $\delta = 1$ making $2\theta_2 - \theta_1 = 2\pi/9$. Taking $\delta = 9$ reduces to a single circle.

The enigmatic FIGURE 1, lower right, is a strange hybrid: here $\theta_1 = 20°$, $\theta_2 = 46°$ (and $l = 50$) so that $2\theta_2 - \theta_1 = 72 = 360/5$, but also, as in Case 1, $\theta_1 = 360/18$. The case $\delta = 1$ is shown in FIGURE 9, left, and we see the expected five circles. In FIGURE 1, lower right, we have $\delta = 3$. Since $20 \times 3 = 360/6$ we might expect six circles as in Case 1, but $72 \times 3 = 3 \times 360/5$ so perhaps there are also five Case 2 circles present! What do you think?

Case 3. There is another case where the zags are all tangent to one or more circles. If $\delta(\theta_1 - \theta_2)$ is a multiple of 2π then the cosine term on the right hand side of (1) is *constant*. We can then make (1) match (4) with $R = 0$ and

$$r = \frac{50}{\sin\frac{1}{2}\theta_1}\cos\left(k_0(\theta_1 - \theta_2) + \tfrac{1}{2}\theta_1\right) + \tfrac{1}{2}l\cot\tfrac{1}{2}\theta_2.$$

Thus all the zags are tangent to one circle, centered at $\mathbf{c}$. If $k_0 = 0$ then the radius of the circle is

$$50\cot\tfrac{1}{2}\theta_1 + \tfrac{1}{2}l\cot\tfrac{1}{2}\theta_2.$$

If, in fact, $m\delta(\theta_1 - \theta_2)$ is a multiple of 2π with the integer m as small as possible, then there will be m *concentric* circles. An example is shown in FIGURE 10, where $l = 50$, $\theta_1 = 19°$, $\theta_2 = 73°$. Here $\theta_1 - \theta_2 = -54°$ and $\delta = 20$ is the smallest number making $\delta(\theta_1 - \theta_2)$ a multiple of 360, so we obtain (left) one circle with all zags tangent to it. The right hand figure shows $\delta = 10$ giving two circles.

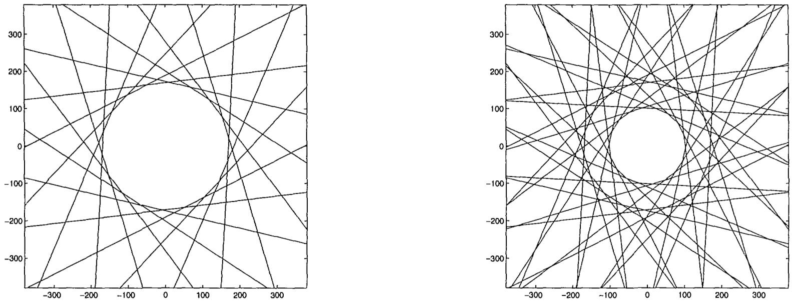


Figure 10 An example where all zags are tangent to (left) one circle and (right) two *concentric* circles.

A concluding problem

What exactly is it about θ_1, θ_2 and δ which allows the existence of a reasonably simple whirligig tangent to the visible envelope? We want, in rough terms, a and c to be small but the number $s(\delta)$ of zags to be large. I do not know the full answer to this.

Acknowledgments. The applets referred to in the Introduction were written by Toby Hall with input from Andy Stubbs and Matthew Trout. I am also grateful to Simon Butler for working on some aspects of this article and finding some nice examples, and to Victor Flynn for mathematical advice. All the figures showing actual examples were produced using MATLAB. I am grateful to one of the referees for pointing out that whirligigs are thinly disguised line-trochoids and to the other for drawing my attention to [**4**].

REFERENCES

1. H. Abelson and A. A. diSessa, *Turtle Geometry*, M.I.T. Press, 1980.
2. Ke Chen, Peter Giblin and Alan Irving, *Mathematical Explorations with MATLAB*, Cambridge University Press, 1999.
3. E. H. Lockwood, *A Book of Curves*, Cambridge University Press, 1961.

→ P. M. Maurer, ‘A rose is a rose…’, *Amer. Math. Monthly* **94** (1987), 631–645.

Math Bite: Four Constants in Four 4s

The well-known problem of representing numbers with four 4s appeared in 1881 in a letter to the editor of *Knowledge: An Illustrated Magazine of Science, Plainly Worded—Exactly Described*, an informal London science journal. The letter mentions that all of the first 20 integers except 19 can be represented using four 4s and the operations $+$, $-$, $\times$, and $\div$. The problem was extended by Martin Gardner [**1**]: Allowing the use of square roots, decimals, factorials, concatenations of 4s (.4, 444!, etc.), and $.\dot{4} = .4444\ldots$, all positive integers less than 113 can be represented. Gardner also claimed that 113 could not be represented.

Our variant of the extended problem is to use four 4s to approximate accurately and elegantly four notable constants. These are the mathematical constants $e = 2.71828\ldots$ and $\pi = 3.1415926\ldots$, the acceleration of gravity $g = 9.8067\ldots$ (in meters/second^2), and Avogadro's number $N_A \approx 6.0221 \times 10^{23}$.

$$e \approx (4!!)\sqrt[4!!]{\frac{\sqrt{4!!}}{4!!!}}. \qquad \pi \approx \sqrt{\sqrt{4!\cdot 4 + \sqrt{\sqrt{\sqrt[4]{\sqrt{\sqrt{\sqrt{\sqrt{.4}}}}}}}}} \approx 3.1415932.$$

$$g \approx \sqrt{\sqrt{\sqrt{\sqrt{4^{4!}}\cdot\sqrt{4!+\sqrt{4}}}}} \approx 9.8068.$$

$$N_A \approx (4!!)\,\frac{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{\sqrt{.4}}}}}}}}}}}{\sqrt[4!]{\sqrt{4}}} \approx 6.0225 \times 10^{23}.$$

The expression for e was derived from Stirling's approximation for $n!$, and is accurate to the 21st decimal place. It can be made arbitrarily accurate by repeatedly replacing 4 with 4!. The other expressions were found by trial and error using calculators.

Similar expressions for e, π, and g could be derived using three 3s or five 5s, but finding an expression for N_A could be difficult without using 4!!.

REFERENCE

1. Martin Gardner, Mathematical Games, *Scientific American* **207** (1964), 120–126.

—A. Bliss, S. Haas, J. Rouse, and G. Thatte
Harvey Mudd College
Claremont, CA 91711

The Track of a Bicycle Back Tire

STEVEN R. DUNBAR
University of Nebraska-Lincoln
Lincoln, NE, 68588-0323
sdunbar@math.unl.edu

REINIER J. C. BOSMAN AND SANDER E. M. NOOIJ
Institute for Theoretical Physics
University of Amsterdam
The Netherlands
bosman@science.uva.nl
semnooij@science.uva.nl

A rider on a bicycle goes down a road, steering a path with the front tire, or perhaps just weaving back and forth. If the tires pass through a puddle, we can see from the tracks that the back tire follows a path similar to that of the front tire. Suppose we know the path of the front tire precisely; what is the path of the back tire? A related question is: If the front tire travels some distance, how far does the back tire travel? It seems obvious from experience that the back tire travels a shorter path than the front tire. Bicycle folklore says that after a long trip the back tire will show about 10% less wear than the front tire. Is it possible to verify the folklore?

In this article we derive and solve differential equations for the path of the back tire, given the path of the front tire. For a parametric form of the front-tire path the differential equations for the path of the back tire are a pair of coupled nonlinear differential equations.

These equations for the back-tire path are a simple example of vector Riccati equations. In a few idealized cases, we can solve them directly, with geometrical arguments or by "guess-and-check." More realistic cases require more sophisticated methods. We use both regular perturbation (matching terms of the solution's Taylor expansion in terms of a small parameter) and iteration (successive approximation) techniques. We'll then derive some quantitative rules for the distance the back tire travels compared to the front tire.

The formulation and solution of the differential equations for the back-tire path given the front-tire path is an example of a *forward* or *direct problem*. In contrast, an *inverse problem* would be: Given the paths of both the front tire and the back tire, determine which was the path of the front tire. This inverse problem is explored in the well-known article, "Which way did the bicycle go?" [**9**], which leads to an elementary calculus problem relating the tangent lines of the two paths. We use this same relation of the tangent lines in formulating the differential equations for the direct problem. The solution of the direct problem will also give another way to characterize the front- and back-tire paths in terms of magnitude of the oscillations. This provides another elementary way to solve the inverse problem in special cases when the paths are sinusoidal.

Investigating the dynamics of bicycles has been a favorite topic in physics for a long time, and there are many references in the physics education journals [**1, 2, 4, 5, 3, 6, 7**], particularly with reference to the stability of a bicycle. There are fewer references to bicycles in the mathematics literature [**9, 10**], generally dealing with the paths of the tires. This article provides an elementary application of coupled nonlinear differential equations to a familiar situation. The application and methods are suitable for classes on methods of applied mathematics and differential equations. A *Maple* worksheet showing the computations and figures is available at the MAGAZINE website: http://www.maa.org/pubs/mathmag.html.

Model equations

Assume that the path of the contact point of the front tire of a bicycle is given parametrically as a function of time as $(x_f(t), y_f(t))$. The contact point is where the tire touches the road. The problem is to determine the path of the contact point of the rear tire, represented parametrically by $(x_b(t), y_b(t))$. Let a be the wheel-base of the bike, that is, the distance between the front and back contact points. As the handlebars are turned, the front tire more or less swivels on the contact point below. Actually, depending on the design of the bicycle, the front contact point moves slightly relative to the back, but we will treat a as a constant. Analyzing how this assumption affects the results would be a good exercise in modeling. The dimensions of a, $x_b(t)$, $y_b(t)$, $x_f(t)$ and $y_f(t)$ are in the units of distance.

We draw on a primary physical fact (the main ingredient in [**9**]): the tangent vector to the path of the rear tire always points to the contact point on the path of the front tire. Since the wheel-base is constant, we may express this fact as a pair of differential equations for the path of the rear-tire contact point:

$$x_b(t) + a\frac{x_b'(t)}{\sqrt{(x_b')^2 + (y_b')^2}} = x_f(t)$$

$$y_b(t) + a\frac{y_b'(t)}{\sqrt{(x_b')^2 + (y_b')^2}} = y_f(t).$$

This pair of equations can be rearranged as

$$\frac{x_b'(t)}{\sqrt{(x_b')^2 + (y_b')^2}} = \frac{x_f(t) - x_b(t)}{a} \tag{1a}$$

$$\frac{y_b'(t)}{\sqrt{(x_b')^2 + (y_b')^2}} = \frac{y_f(t) - y_b(t)}{a}. \tag{1b}$$

This says that the unit velocity vector for the back tire points in the direction (unit vector!) of the wheel-base of the bicycle. Taking a closer look, equation (1) is a relation between two unit vectors. There is no way to determine the magnitude of the velocity vector $(x_b'(t), y_b'(t))^T$. With this in view, call the speed of the back tire point $v(t) = \sqrt{(x_b')^2 + (y_b')^2}$. Rewriting (1) gives

$$x_b'(t) = v(t)\frac{x_f(t) - x_b(t)}{a} \tag{2a}$$

$$y_b'(t) = v(t)\frac{y_f(t) - y_b(t)}{a}. \tag{2b}$$

If we can express the speed of the back-tire point in terms of the known front-tire speed, then we will be able to express the differential equation for the back-tire path in terms of $(x_b(t), y_b(t))$ and known quantities.

Think of the back-tire velocity as though the velocity vector of the front tire is dragging the back tire along. The velocity of the back tire will be the component of the front-tire velocity in the direction of the back-tire motion. The back-tire speed is the magnitude of this projection, found using the dot product:

$$v(t) = \begin{pmatrix} x_f'(t) \\ y_f'(t) \end{pmatrix} \cdot \begin{pmatrix} \frac{x_f(t)-x_b(t)}{a} \\ \frac{y_f(t)-y_b(t)}{a} \end{pmatrix}. \tag{3}$$

With this, the right sides of equations (2) are expressed in terms of the given quantities $(x_f(t), y_f(t))$ and $(x_f'(t), y_f'(t))$ and the unknown track $(x_b(t), y_b(t))$. This puts our equations (2) into a nice form: derivatives of the unknown quantities are expressed in terms of the unknown quantities and various known quantities.

Expression as a Riccati equation

We rewrite the equations for the bicycle back tire in an alternative form, relating them to the standard theory for matrix Riccati equations. Although we will not pursue solving the equations in this form, the matrix form suggests some other approaches to analyzing the equations theoretically and solving them efficiently with numerical methods.

This application of matrix Riccati equations to bicycle tire tracks is a very elementary and easily derived example; other less elementary applications of Riccati equations arise in transmission line theory, random noise theory, variational equations, and control theory [**8**].

Insert expression (3) into (2) and collect terms

$$x_b'(t) = \frac{1}{a^2}\left[x_f'(t)\left\{x_f(t) - x_b(t)\right\}^2 + y_f'(t)\left\{x_f(t) - x_b(t)\right\}\left\{y_f(t) - y_b(t)\right\}\right] \tag{4a}$$

$$y_b'(t) = \frac{1}{a^2}\left[x_f'(t)\left\{x_f(t) - x_b(t)\right\}\left\{y_f(t) - y_b(t)\right\} + y_f'(t)\left\{y_f(t) - y_b(t)\right\}^2\right]. \tag{4b}$$

Now the quadratic form characteristic of Riccati equations is explicit, but cumbersome. For a more elegant presentation, we write the equations in terms of the quantities $w_1(t) = x_b(t) - x_f(t)$, and $w_2(t) = y_b(t) - y_f(t))$. A little work leads to:

$$w_1'(t) = \frac{1}{a^2}\left[x_f'(t)w_1^2(t) + y_f'(t)w_1(t)w_2(t)\right] - x_f'(t) \tag{5a}$$

$$w_2'(t) = \frac{1}{a^2}\left[x_f'(t)w_1(t)w_2(t) + y_f'(t)w_2^2(t)\right] - y_f'(t). \tag{5b}$$

Let $W(t) = [w_1(t), w_2(t)]^T$ be a 2×1 matrix (or vector). Then equations (5) can be expressed as

$$W'(t) - \frac{1}{a^2}\left[W(t)\, F^T(t)\, W(t)\right] + F(t) = \mathbf{0},$$

where $F(t) = [x_f'(t), y_f'(t)]^T$. This is a special case of the general Riccati matrix differential equation treated by Reid [**8**, equation (2.1), page 11], where linear terms are allowed in addition to the quadratic term in our equation. This demonstrates the remarkable likeness between the equations for such simple things as bicycle paths and complicated things like random noise theory. Matrix Riccati equations can arise even from familiar physics.

Some simple paths for the front tire

We check our modeling by seeing that the differential equations predict the right behavior in cases where we already know what happens.

Simple case I: A straight path, no turns, constant front-tire speed Suppose the path of the front tire is a straight line, say along the x-axis. Our experience tells us that if the front-tire path is a straight line, then the back tire should follow behind on the same straight line. We leave it to the reader to verify this directly from the differential equations, which are easy to solve in this case. This is recommended as a way to get to know the equations better. Set

$$x_f(t) = v_f\, t, \;\; y_f(t) = 0.$$

After writing down the differential equations for $x_b(t)$ and $y_b(t)$, it is easy to check that the solutions are

$$x_b(t) = -a + v_f\, t, \, y_b(t) = 0.$$

Thus, the back tire follows a straight line, lagging by only the wheel-base of the bicycle.

Simple case II: A large circular path Take the path of the front tire to be a large circle. From our experience riding a bike around a circle, the back-tire contact point trails behind and inside the circular front-tire path. In steady state, the angle between the circle and the bike body (wheel-base) should be constant, as shown by the symmetry of the system. Since the bike is of constant length and makes a constant angle, the back-tire contact point follows a concentric circular path. The situation is depicted in FIGURE 1. The question now becomes: “What is the radius of the inner circle of the back tire?”.

Set some notation: Let c be the radius of the large circle traced by the front-tire contact point. Let a be the wheel-base of the bike. Let b be the (unknown) radius of the inner circle traced by the back tire. Remember that the velocity vector of the back tire contact point will point at the contact point of the front tire. But the velocity vector of the back-tire contact point is perpendicular to the radius vector of the back-tire contact point. As seen in the diagram, a right triangle is formed by the back-tire

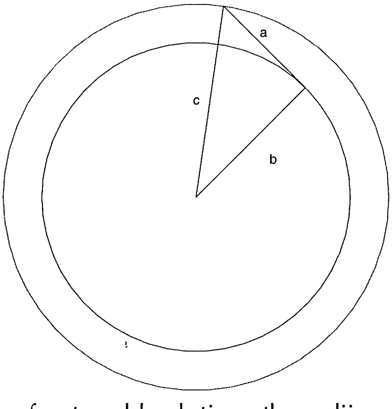


Figure 1 The tracks of the front and back tires, the radii, and the bicycle wheel-base when the front tire follows a circular path

contact point radius vector (of length b), the bike wheel-base vector (of length a), and the front-tire contact point radius vector (of length c). Therefore, $a^2 + b^2 = c^2$ and so $b = \sqrt{c^2 - a^2}$.

How much less does the back tire travel? The ratio of the length of the paths of the small back-tire circle to the large front-tire circle is $\sqrt{c^2 - a^2}/c = \sqrt{1 - a^2/c^2}$.

In this situation, a limiting case occurs as the radius of the large circle goes to infinity. Then we can consider the front path to be a straight line, for instance, the x-axis. According to our results, the radius of the back-tire path also goes to infinity, which we should expect, since its path is also a straight line.

We now formulate this in terms of the differential equations; say that the front tire starts at $(c, 0)$. Then, convenient parametric equations are

$$x_f(t) = c\cos(\omega t),\ y_f(t) = c\sin(\omega t).$$

After computing $v(t)$ from formula (3) and simplifying (2), the differential equations for the coordinates of the back contact point become

$$\frac{dx_b(t)}{dt} = (c\omega/a^2)(c\cos(\omega t) - x_b(t))(\sin(\omega t)x_b(t) - \cos(\omega t)y_b(t)) \qquad (6a)$$

$$\frac{dy_b(t)}{dt} = (c\omega/a^2)(c\sin(\omega t) - y_b(t))(\sin(\omega t)x_b(t) - \cos(\omega t)y_b(t)) \qquad (6b)$$

Although equations (6) look difficult to solve, a solution by inspection is readily possible. The previous geometric analysis suggests that the solution should have the form

$$x_b(t) = b\cos(\omega t - \psi) \quad \text{and} \quad y_b(t) = b\sin(\omega t - \psi),$$

where b is the radius, and $\psi = \arcsin(a/c)$ is the phase shift indicating the back tire trails behind the front tire. Insert this trial solution into the differential equation; using standard trigonometric identities you will see that it works, with $b = \sqrt{c^2 - a^2}$.

Simple case III: A stunt circular turn with rear tire as pivot Another limiting case occurs when the radius of the front-tire circle shrinks to a, the wheel-base of the bicycle. In this case, the front tire is turned at a right angle to the bike body and moves in a circle of radius a. According to our formula, the radius of the back-tire path should be 0. This is a stunt turn, a spin or pivot on the back tire. Physically, the back-tire contact point remains motionless.

For this special case, solving the differential equations is easy: Say that the front tire starts at $(a, 0)$, the back tire is positioned at the origin; as in the previous example, write parametric equations for the path of the front tire, and write out the differential equations, with initial conditions $x_b(0) = 0$ and $y_b(0) = 0$. It will then be clear that $x_b(t) = 0$, and $y_b(t) = 0$ satisfy the initial conditions and the differential equations. Therefore, by the uniqueness theorem for first-order ordinary differential equations, this is the unique solution.

When the front path is a sine curve

Set-up and numerical analysis Assume that the front tire follows a sine curve. Experience suggests that the back-tire track should also follow a sine curve with a phase shift and a smaller amplitude.

Start by introducing the path of the front tire, written parametrically as

$$x_f(t) = st,\ y_f(t) = A_f \sin(\xi st). \tag{7}$$

In this set-up, s is a speed parameter that converts time to distance. Let ξ be a spatial frequency, so that one oscillation of the front tire takes $2\pi/\xi$ units of horizontal distance or $2\pi/\xi s$ units of time. Let A_f be the amplitude of the front-tire oscillation. As usual, call a the wheel-base of the bicycle. When we incorporate the parametric equations (7) into the differential equations (2), with an appropriate substitution of the velocity (3), we arrive at the following equations for the x- and y-coordinates of the back tire:

$$\frac{dx_b(t)}{dt} = \left[\frac{s\{st - x_b(t)\}}{a} + \frac{A_f \cos(\xi st)\xi s\{A_f \sin(\xi st) - y_b(t)\}}{a}\right]\frac{st - x_b(t)}{a}. \tag{8}$$

$$\frac{dy_b(t)}{dt} = \left[\frac{s\{st - x_b(t)\}}{a} + \frac{A_f \cos(\xi st)\xi s\{A_f \sin(\xi st) - y_b(t)\}}{a}\right]\frac{A_f \sin(\xi st) - y_b(t)}{a}. \tag{9}$$

We choose $x_b(0) = -a$, $y_b(0) = 0$ as reasonable initial conditions, but other initial conditions are possible.

These equations are certainly difficult to solve analytically; we will first find approximate numerical solutions, for which we need to choose specific values for the parameters. We will take $a = 1$, which, in SI units, means that the wheel-base is 1 meter long; this is reasonable (although just a little short for most adult bicycles, which have a measured wheel-base of slightly more than one meter). Take the amplitude of oscillation A_f to be 0.3 meters, the speed s to be 5 meters per second and the frequency ξ to be 1. The graph of the numerical solution, rescaled to show the details, is shown in FIGURE 2. Of course, the actual path looks far less oscillatory.

The path of the back tire has the general form of a sine curve. Note that the amplitude of oscillation of the back tire is less than the amplitude of oscillation of the

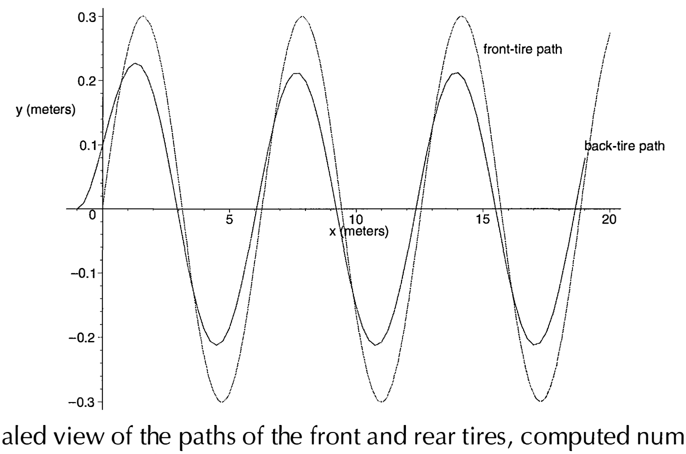


Figure 2 Scaled view of the paths of the front and rear tires, computed numerically with $a = 1$, $A_f = 0.3$, $s = 5$, and $\xi = 1$

front tire, and the back tire is slightly phase-shifted behind the front tire. This seems reasonable.

A good guess at an approximate solution Based on the numerical solution, we can make a good guess at the approximate solution. Since we expect a path similar to the front-tire path, we guess a parametrization, $x_b(t) = st - a$, and $y_b(t) = A_b \sin(\xi st - \psi)$, where A_b is the amplitude of the back-tire path and ψ is a phase shift. Note that this cannot be the exact solution, because here the horizontal distance between the tires' contact points is a in these formulas; in reality, it must be shorter.

From the differential equation (2) (in a slightly different form) we know

$$\frac{y_f(t) - y_b(t)}{x_f(t) - x_b(t)} = \frac{dy_b(t)}{dx_b(t)}.$$

Our assumptions make the left-hand denominator equal to the wheel-base a. Using the chain rule and rearranging the equation, we get

$$y_b(t) = y_f(t) - a\frac{dy_b(t)}{dt}\frac{dt}{dx_b(t)}. \tag{10}$$

Inserting the guessed solution and the known front-tire formula into equation (10) yields

$$A_b \sin(\xi st - \psi) = A_f \sin(\xi st) - a\xi s A_b \cos(\xi st - \psi)\frac{1}{s}.$$

Since this equation must hold for any time, choosing $t = 0$ we get the phase shift

$$\psi = \arctan(\xi a) \tag{11}$$

and using $t = \pi/(2s\xi)$, we get the amplitude

$$A_b = \frac{A_f}{\sqrt{1 + \xi^2 a^2}}.$$

We can plot the guess along with the numerical solution to the equations on scaled axes to compare them in FIGURE 3. Except for a short transient, the guess seems to be identical with the numerical solution. The difference between the numerically

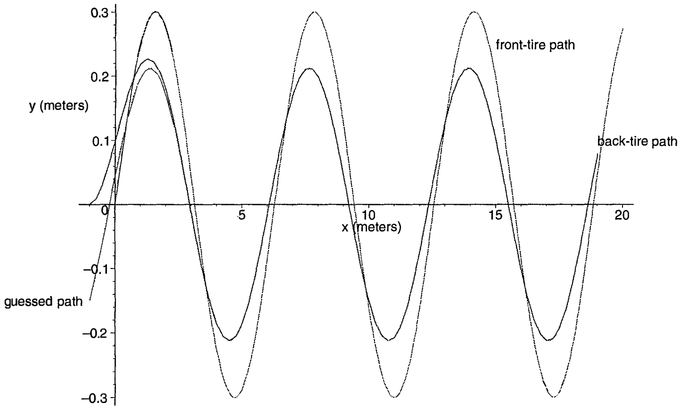


Figure 3 A scaled plot of the front-tire path, the numerically computed back-tire path and the guessed solution

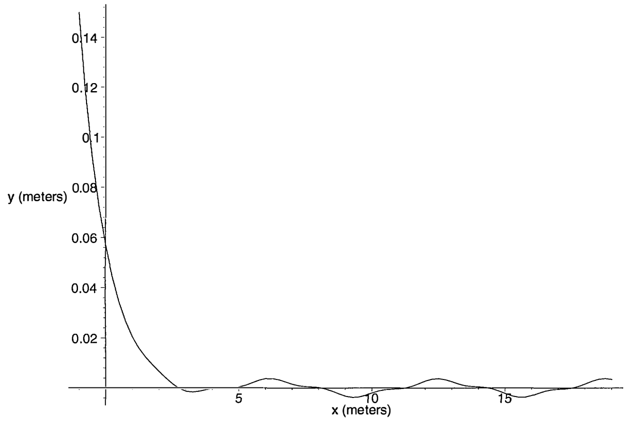


Figure 4 The difference between the numerically computed back-tire path and the sinusoidal guess

computed back-tire path and the guessed solution is plotted in FIGURE 4; except for the transient, the difference is about a centimeter, less than 5% of the magnitude of the back-tire oscillation. The solution $A_b \sin(\xi st - \psi)$ was a very good guess, and the back-tire path is nearly a sine curve. Is there another way to justify the guess?

A solution based on linearized equations The amplitude of the front-tire oscillation is small compared to the other physical parameters. Furthermore, as the amplitude of the front-tire path goes to zero, that is, the path approaches a straight line, the back-tire path approaches the same straight line. We can see that the back-tire path depends on the front-tire amplitude. This suggests that we take the amplitude of the front-tire oscillation to be a small parameter in the differential equations.

We will assume that the back-tire path can be expressed as a power series based on this parameter. Inserting the power series expansion into the differential equations and gathering like terms gives a sequence of linear differential equations that we can solve. In applied mathematics, this method is called a regular perturbation expansion. Regular perturbation is routinely used in all applications where we need to solve a nonlinear equation, at least approximately.

To work out the details, assume

$$x_b(t) = x_{b0}(t) + A_f x_{b1}(t) + \mathcal{O}(A_f^2), \tag{12a}$$

$$y_b(t) = y_{b0}(t) + A_f y_{b1}(t) + \mathcal{O}(A_f^2), \tag{12b}$$

and insert these expansions into equations (8) and (9) to get a perturbation expansion. It is not necessary, but using a symbolic computation system simplifies matters.

Zeroth order Inserting the posited form of the solution, expanding and comparing the terms with no coefficient of A_f, we find an equation for the leading order term

$$\frac{dx_{b0}(t)}{dt} = \frac{s^3t^2}{a^2} - 2\frac{s^2tx_{b0}(t)}{a^2} + \frac{s\,\{x_{b0}(t)\}^2}{a^2} \tag{13}$$

with initial condition $x_{b0}(0) = -a$. Note that this is a Riccati equation because of the term $x_{b0}(t)^2$. However, the right-hand side can be easily factored and the equation solved as either a separable equation or by inspection (or checked!) to yield

$$x_{b0}(t) = st - a. \tag{14}$$

Likewise we can find the leading order equation for $y_{b0}(t)$, using the new information from (14)

$$\frac{dy_{b0}(t)}{dt} = -\frac{sy_{b0}(t)}{a}, \tag{15}$$

with initial condition $y_{b0}(0) = 0$. Because this equation is linear and homogeneous with 0 as initial condition, the solution must be identically zero:

$$y_{b0}(t) = 0. \tag{16}$$

This proves the zeroth order perturbation solution agrees with the formulas in Simple Case I. To lowest order of approximation, the motion of the back tire following a front tire weaving back and forth with small amplitude is a straight line.

First order We equate the terms of the x_b equation with coefficient A_f, and insert the now known solutions (14) and (16), to find

$$\frac{dx_{b1}(t)}{dt} = -\frac{2sx_{b1}(t)}{a}, \tag{17}$$

with initial condition $x_{b1}(0) = 0$. The result is a homogeneous linear differential equation, and is therefore easy to solve:

$$x_{b1}(t) = 0. \tag{18}$$

The equation for $y_{b1}(t)$ is more interesting. Comparing terms with one power of A_f, and using all of the previous information about x_{b0}, y_{b0} and x_{b1}, we find

$$\frac{dy_{b1}(t)}{dt} = \frac{s}{a}\left[\sin(\xi st) - y_{b1}(t)\right]. \tag{19}$$

The solution with initial condition $y_{b1}(0) = 0$ is

$$y_{b1}(t) = \frac{-a\xi\cos(\xi st) + a\xi e^{(-\frac{st}{a})} + \sin(\xi st)}{1 + \xi^2 a^2}.$$

It is easy to use standard identities to write this solution as

$$y_{b1}(t) = \frac{\sin(\xi st - \psi)}{\sqrt{1 + \xi^2 a^2}} + \frac{\sin(\psi)e^{-\frac{st}{a}}}{\sqrt{1 + \xi^2 a^2}},$$

with ψ as in (11).

We now assemble the power series expansion of the solution by substituting $x_{b0}(t)$, $x_{b1}(t)$, $y_{b1}(t)$, and $y_{b2}(t)$ into (12). Ignoring quadratic and higher orders of the front-tire oscillation amplitude, we have the following parametric equations for the motion of the back tire:

$$x_b(t) = st - a \tag{20a}$$

$$y_b(t) = A_f\left[\frac{\sin(\xi st - \psi)}{\sqrt{1 + \xi^2 a^2}} + \frac{\sin(\psi)e^{-\frac{st}{a}}}{\sqrt{1 + \xi^2 a^2}}\right] \tag{20b}$$

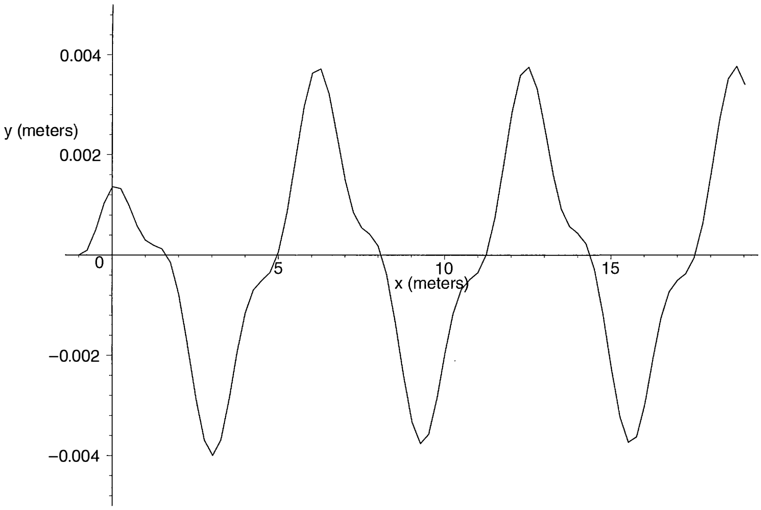


Figure 5 The difference between the numerically computed and the perturbation solutions

We could do some additional work to compare terms with coefficients A_f^2, using the known $x_{b0}(t)$, $x_{b1}(t)$, $y_{b0}(t)$, and $y_{b1}(t)$ to derive linear equations for $x_{b2}(t)$ and $y_{b2}(t)$. However, before we do that let's stop and examine graphically what we have so far.

Plotting the perturbation solutions (20) parametrically, we discover that the back path appears identical with the numerical solution, even including the short transient. In fact, superimposing the perturbation solution with the numerical solution simply yields another copy of FIGURE 2. The difference of the numerically computed back-tire path and the perturbation solution is plotted in FIGURE 5. The transient difference is nearly zero, and the difference is less than 2% of the amplitude of the back-tire oscillation. The back-tire path is very nearly a sine curve with an exponential transient. This explains why the guess was a good approximation except for the transient.

As mentioned above, with more work we could derive additional terms in the perturbation expansion to get an even better approximation. However, it appears that we now have a solution that sufficiently explains and confirms our intuition about the back tire.

A general front-tire path

Now assume that the front-tire path is given by $x_f(t) = st$ and $y_f(t) = A_f\ \mathrm{f}(st)$, where s is a speed parameter and $\mathrm{f}(\cdot)$ is a bounded continuously differentiable function whose maximum value is scaled to be 1. Then the amplitude of the motion of the front tire is a small parameter A_f. Additionally we assume that $\mathrm{f}(0) = 0$ so the motion of the front tire starts at the origin. Again we will discover an approximation for the motion of the back tire by means of regular perturbation expansion.

Looking back at the work in the previous section, we see that the expansions for $x_{b0}(t)$, $y_{b0}(t)$ and $x_{b1}(t)$ don't involve A_f or the front-tire function f. Therefore, the equations and their solutions will be the same, yielding again $x_{b0}(t) = s\,t - a$, $y_{b0}(t) = 0$, and $x_{b1}(t) = 0$. The equation corresponding to (19) for the first-order term $y_{b1}(t)$ using all of this information simplifies nicely to

$$\frac{dy_{b1}(t)}{dt} = \frac{s\mathrm{f}(st)}{a} - \frac{sy_{b1}(t)}{a}$$

with initial condition $y_{b1}(0) = 0$. The solution can be written

$$y_{b1}(t) = \int_0^t \frac{s\mathrm{f}(su)}{a} e^{\frac{-s(t-u)}{a}}\, du.$$

Estimating this integral shows that $y_{b1}(t)$ is bounded by st/a, since $\mathrm{f}(st)$ is bounded by 1. We can get a better bound by doing some deeper analysis of the integral, which is a convolution.

Introduce the variable $w = \frac{s}{a}(t - u)$, and find that

$$y_{b1}(t) = \int_0^{\frac{s}{a}t} \mathrm{f}(st - aw)e^{-w}\, dw.$$

Then let $\mathrm{g}_t(w) = \mathrm{f}(st - aw)I_{[0,(s/a)t]}(w)$, where $I_{[0,S]}(w)$ is the indicator function on $[0, S]$, which is 1 for $w \in [0, S]$ and 0 otherwise. Note that $\mathrm{g}_t(w)$ is continuous, since as $w \to st/a$ then $\mathrm{f}(st - aw) \to \mathrm{f}(0) = 0$ by the assumption that $\mathrm{f}(\cdot)$ was C^1 and starts at the origin ($\mathrm{f}(0) = 0$). Also $\mathrm{g}_t(w)$ is bounded by the bound on $\mathrm{f}(\cdot)$, which we have assumed to be 1. With this notation

$$y_{b1}(t) = \int_0^{\infty} \mathrm{g}_t(w)e^{-w}\, dw,$$

and the nature of the function as a Laplace transform is clearly revealed. Then

$$\begin{aligned}
|y_{b1}(t)| &= \left| \int_0^{\infty} \mathrm{g}_t(w)e^{-w}\, dw \right| \\
&\leq \int_0^{\infty} |\mathrm{g}_t(w)|e^{-w}\, dw \\
&\leq \int_0^{\infty} e^{-w}\, dw = 1.
\end{aligned}$$

This means that the amplitude of the back-tire motion never exceeds the amplitude of the front-tire motion, a reasonable conclusion.

Another formulation and solution by iteration

Formulation and diagrams For a nonparametric front-tire path, given as $y_f = \mathrm{f}(x_f)$, an interesting alternative differential equation for the back-tire path results. Of course, any such front-tire path could be put in parametric form and studied using methods from the previous sections. However, looking at the bicycle equations in this new form provides a simple derivation of an unusual form of nonlinear delay-differential equation, interesting in its own right. The method of successive approximations is a typical way to solve nonlinear equations; the bicycle equation provides a nontrivial example in a situation where we have an answer to check against. A good principle of research in applied mathematics is to solve a problem in two different ways and compare the answers. Fortunately for us, both solution methods yield the same results!

As before, take a to be the wheel-base of the bicycle, and assume the front tire starts at the origin, so $\mathrm{f}(0) = 0$. Then the back tire starts at the coordinate $(-a, 0)$. Let the path of the back tire be given as a function $\mathrm{g}(x_b)$ of the x-coordinate of the back tire, x_b. Then we know for example that $\mathrm{g}(-a) = 0$.

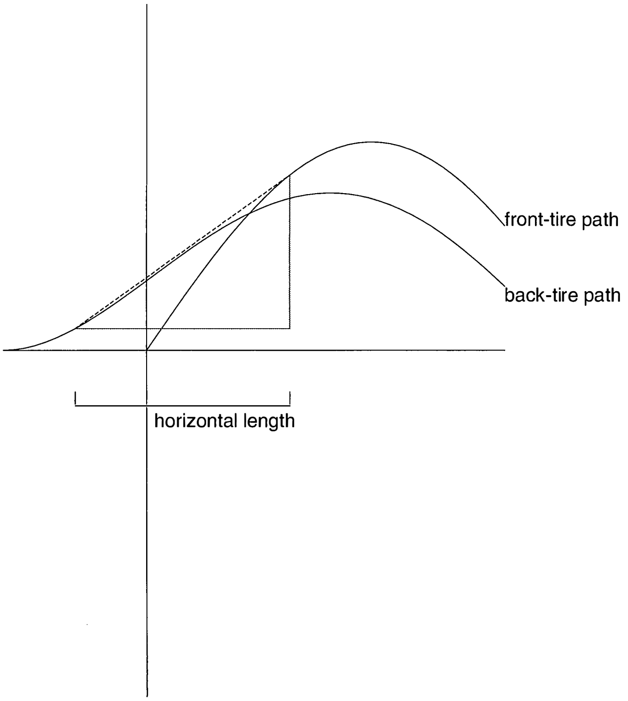


Figure 6 A schematic diagram of the front-tire path, the back-tire path, the bicycle spanning the paths (dashed line), and the projection of the bicycle along the axis giving the horizontal length

We use our notation to express the horizontal distance between the front and back tires, as seen in FIGURE 6:

$$x_f - x_b = \sqrt{a^2 - \left\{\mathrm{f}(x_f) - \mathrm{g}(x_b)\right\}^2}.$$

This gives an implicit relationship between x_f and x_b, assuming that we know the back-tire path $\mathrm{g}(x_b)$. Using the fundamental fact that the tangent vector to the back-tire path points in the direction of the bicycle, we can write

$$\frac{d\mathrm{g}(x_b)}{dx_b} = \frac{\mathrm{f}(x_f) - \mathrm{g}(x_b)}{\sqrt{a^2 - \left\{\mathrm{f}(x_f) - \mathrm{g}(x_b)\right\}^2}}.$$

Again, remember that x_f is given implicitly in terms of x_b, so that there is really only one independent variable x_b in the differential equation. We can rewrite this equation slightly more transparently if we set $x_f = x_b + \Delta$. Then the differential equation is

$$\frac{d\mathrm{g}(x_b)}{dx_b} = \frac{\mathrm{f}(x_b + \Delta) - \mathrm{g}(x_b)}{\sqrt{a^2 - \{\mathrm{f}(x_b + \Delta) - \mathrm{g}(x_b)\}^2}}. \tag{21}$$

Although the differential equation (21) now looks more familiar, it still hides quite a bit of difficulty. Since the right side contains not only the unknown function $\mathrm{g}(x_b)$, but also an advanced argument on the right side, $x_b + \Delta$, this might seem to be a simple delay-differential equation. But this hides an important fact: the delay itself depends on the unknown function $\mathrm{g}(x_b)$ through the implicit relationship $x_f = x_b + \sqrt{a^2 - \{\mathrm{f}(x_f) - \mathrm{g}(x_b)\}^2}$. Therefore, this differential equation has the unknown function appearing not only nonlinearly on the right side, but nonlinearly in the argument of the right side too!

The method of successive approximations is a typical way to solve such an equation. First, approximate the horizontal distance Δ between the front- and rear-tire contact points of the bicycle by the wheel-base a. Generally, this will be an overestimate for the horizontal distance, but it will be close if the amplitude of the paths is not large.

Equation (21) then becomes

$$\frac{d\mathrm{g}(x_b)}{dx_b} = \frac{\mathrm{f}(x_b + a) - \mathrm{g}(x_b)}{\sqrt{a^2 - \{\mathrm{f}(x_b + a) - \mathrm{g}(x_b)\}^2}}. \tag{22}$$

For a given front-tire path and a known wheel-base this equation is easy to solve numerically.

The next approximation replaces the horizontal distance Δ with

$$\sqrt{a^2 - \{\mathrm{f}(x_b + a) - \mathrm{g}(x_b)\}^2}.$$

This will also be larger than the true horizontal distance, but clearly a closer approximation. The equation becomes

$$\frac{d\mathrm{g}(x_b)}{dx_b} = \frac{\mathrm{f}\left(x_b + \sqrt{a^2 - \{\mathrm{f}(x_b + a) - \mathrm{g}(x_b)\}^2}\right) - \mathrm{g}(x_b)}{\sqrt{a^2 - \left\{\mathrm{f}(x_b + \sqrt{a^2 - \{\mathrm{f}(x_b + a) - \mathrm{g}(x_b)\}^2}) - \mathrm{g}(x_b)\right\}^2}}. \tag{23}$$

Now the differential equation is clearly more complicated since already the unknown function appears in the argument of the right-hand side as well as nonlinearly. Nevertheless, we can still solve the differential equation numerically. Conceptually, one could use a simple scheme such as the Euler method: Given an initial condition such as $\mathrm{g}(-a) = 0$, and knowing the given function $\mathrm{f}(x_f)$, the value of the right hand side can be calculated to give the slope of $\mathrm{g}(x_b)$ at $x_b = -a$. Then $\mathrm{g}(-a + h)$ can be estimated, and the process can be repeated. In practice, of course one uses a more sophisticated technique such as a Runge-Kutta method, or a multi-step predictor-corrector method.

Numerical computation of the back-tire path for comparison As an illustration, we use the iteration procedure to calculate the path of the back tire when the front-tire path is a sine curve as before. Then we can compare the numerical solution of the coupled Riccati equations for the parametric form with the solution calculated numerically by the iteration procedure described above.

We use $\mathrm{f}(x_f) = A_f \sin(x_b)$, with front-wheel oscillation amplitude $A_f = 0.3$ and wheel-base $a = 1$ (the same parameters as before). Substituting this information into (22), we find an equation that we can solve numerically.

Plotting the numerical solution along with the zeroth order iteration solution gives a figure that, in print, is virtually indistinguishable from FIGURE 2. For a color version that shows the slight difference, see http://www.maa.org/pubs/mathmag.html. Already, the zeroth order application of this method is almost good enough to fool the eye.

For better results, we can solve the differential equation numerically from the first-order approximation. We can even take one more step in the iteration process to obtain the complicated equation

$$\frac{g(x_b)}{dx_b} = \frac{0.3\sin\left[x_b + \sqrt{1^2 - \left[0.3\sin\left(x_b + \sqrt{1^2 - \{0.3\sin(x_b+1) - g(x_b)\}^2}\right) - g(x_b)\right]^2}\,\right] - g(x_b)}{\sqrt{1^2 - \left\{0.3\sin(x_b + \sqrt{1^2 - \left\{0.3\sin(x_b + \sqrt{1^2 - \{0.3\sin(x_b+1) - g(x_b)\}^2}) - g(x_b)\right\}^2}) - g(x_b)\right\}^2}}.$$

Despite its complicated appearence, the equation can still be solved numerically, with an expected improvement in results. For the purposes of comparison, we present in TABLE 1 all the different solutions for the case where the front tire follows a sine curve. This gives a comprehensive view of the quality of each of the approximate solutions.

TABLE 1: Comparison of the various solutions to the numerical solution. The rows are approximations at various x_b-values, the columns are the various solution methods. The methods labeled g_0, g_1, and g_2 are, respectively, the zeroth, first, and second order iteration methods. The values are rounded to 5 decimal places.

	Numerical	Perturbation	Iteration Sol.		
x_b	Sol.	Sol.	g_0	g_1	g_2
0	0.09998	0.10036	0.10097	0.09995	0.09997
1	0.21934	0.21912	0.21986	0.21932	0.21934
2	0.17756	0.17713	0.17720	0.17757	0.17756
3	−0.01191	−0.01273	−0.01520	−0.01174	−0.01192
4	−0.18601	−0.18538	−0.18780	−0.18593	−0.18601
5	−0.18603	−0.18557	−0.18642	−0.18600	−0.18603
6	−0.01541	−0.01440	−0.01268	−0.01555	−0.01540
7	0.17077	0.17028	0.17277	0.17068	0.17077
8	0.19890	0.19851	0.19947	0.19887	0.19890
9	0.04531	0.04426	0.04296	0.04544	0.04531
10	−0.15100	−0.15066	−0.15355	−0.15090	−0.15100

Conclusions We have shown that if the path of the front tire of a bicycle is specified, it is possible to derive the corresponding path of the back tire. In some geometrically simple cases, such as a large circular path for the front tire, it is possible to derive the corresponding back-tire path precisely. In some other reasonable geometric cases, such as a sinusoidal front-tire path, it is not possible to find the corresponding back-tire path precisely, but we can derive approximations to any desired degree of accuracy. In this paper, we have solved the approximation equations to first order, which seems sufficient for most purposes. In fact, the approximation techniques are easy to apply for any reasonably general front-tire path. The only limit to being able to express the solution analytically is the ability to evaluate a convolution, or equivalently to solve a first-order linear differential equation. Of course, in any case, the equations for the back-tire path can be solved numerically.

The coupled nonlinear differential equations for the back-tire path are easy to express and fairly easy to solve when the front-tire path is given parametrically. When the front-tire path is given directly as a function of the position down the road, the differential equations assume the more challenging form of an unusual delay-differential equation, where the delay even depends on the solution. Nevertheless, the problem can still be handled through successive approximations. The solutions found numerically, whether by regular perturbation or successive approximations all agree, and with about the same amount of work, so the choice of technique should be determined by the available information or purpose of the solution.

Our analysis sheds light on the relative distances traveled by the front and back tires in special cases. If the front-tire path is a large circle, the back tire follows a concentric circle, and should experience less wear than the front tire, because the ratio of the circumferences is $\sqrt{1-a^2/c^2}$. Likewise, if the front tire weaves back and forth along a sine curve, with an amplitude A_f and spatial frequency ξ, then the back tire also follows a sine curve with the smaller amplitude $A_f/\sqrt{1+a^2\xi^2}$. Although it is not possible to express the arc length of a sine curve with a simple analytic expression, the proportionality of the expressions for the functions show that the arc length traveled by the back tire is proportionately less than that traveled by the front tire. Of course, if the path is perfectly straight, both tires go the same distance.

Can we verify the folklore that on a long bike trip the back-tire wear is less than that of the front tire? Probably not, even though the analysis in this article supports the folklore. Too many other variables intervene in the reality to be modeled so simply. For example, if the back-tire inflation is less than the front tire's, it will wear more. The style of riding, including braking, sliding, and skidding, can affect the wear too.

However, we do offer two somewhat practical consequences from the solutions. First, presented with two intertwined sinusoidal functions, known to be the paths of the front and back tire of a bike, we can now confidently know that the path with the larger amplitude is the front tire, and the path with the proportionally smaller amplitude is the back-tire path. With additional inspection, knowing that the tangent vectors from the back-tire point with fixed distance to the front-tire track, we can find which way the bicycle went. Second, the solutions of the general front-tire path case show that the amplitude of the back-tire path never exceeds the amplitude of the front-tire path, that is, in this model the bike doesn't "fish-tail."

This bicycle problem shows that moderately complicated nonlinear differential equations can be found even in simple everyday experiences. Better yet, we were able to apply several different techniques, yielding solutions of various kinds, giving better understanding of both the everyday experiences and the techniques.

Acknowledgment. The work of S. Dunbar was partially supported by the National Science Foundation, DUE-9555404. The work of S. Nooij and R. Bosman was partially supported by a joint grant from the U.S. Department of Education FIPSE and the European Commission. The authors also thank the anonymous referees for numerous suggestions that improved the exposition of this paper.

REFERENCES

1. P. DiLavore, *The Bicycle; Teacher's Guide to the Bicycle*, American Association of Physics Teachers, College Park, MD, 1976.
2. G. Francke, W. Suhr, and F. Riess, An advanced model of bicycle dynamics, *European J. Phys.*, **11**:2 (1990), 116–121.
3. Thomas B. Greenslade, Jr., Exponential bicycle gearing, *The Physics Teacher*, **17** (1979), 455–456.
4. Robert G. Hunt, Bicycles in the physics lab, *The Physics Teacher*, **27** (1989), 160–165.
5. David E. H. Jones, The stability of the bicycle, *Physics Today*, **23**:04 (1970), 34–40.
6. Daniel Kirshner, Some nonexplanations of bicycle stability, *Amer. J. Phys.*, **48** (1980), 36–38.
7. J. Lowell and H.D. McKell, The stability of bicycles, *Amer. J. Phys.*, **50**:12 (1982), 1106–1112.
8. William T. Reid, *Riccati Differential Equations*, volume 86 of *Mathematics in Science and Engineering*, Academic Press, New York and London, 1972.
9. Dan Velleman, Joseph D. E. Konhauser, and Stan Wagon, *Which way did the bicycle go?*, MAA, 1996.
10. Y. Yavin, Navigation and control of the motion of a riderless bicycle by using a simplified dynamic model, *Math. Comput. Modelling*, **25**:11 (1997), 67–74.

NOTES

Volumes and Cross-Sectional Areas

WILLIAM T. ENGLAND
T. LEN MILLER
Mississippi State University
Miss. State, MS 39762

Calculating areas and volumes is the primary application of integration in a first calculus course. In numerous exercises, we slice planar regions into cross-sectional lengths perpendicular to one of the axes and integrate those lengths to obtain the area of the region. Similarly, we slice solids into planar cross-sections perpendicular to an axis and integrate the areas of those regions to calculate the volume of the solid. In the early history of area and volume calculations, reasoning in terms of the cross-sectional lengths or areas was known as the method of indivisibles and developed by Bonaventura Cavalieri, a student of Galileo. The basis for Cavalieri's computations is known today as Cavalieri's Principle:

> *If two plane figures have equal altitudes and if sections made by lines parallel to the bases and at equal distances from them are always in the same ratio, then the plane figures are also in this ratio.*

In elementary calculus courses—as with most early historical calculations—the slices of a planar or solid region are taken perpendicular to an axis, and consequently are all parallel. There is, however, a well-known theorem proved by the classical geometer Pappus (c. 400) that deals with nonparallel cross-sections:

> *If a region $\mathcal{R}$ is rotated about a line exterior to the region, the volume of the resulting solid $\mathcal{Q}$, denoted* $\operatorname{vol}(\mathcal{Q})$*, is the area of the region $\mathcal{R}$ multiplied by the length of the arc traced by the centroid of $\mathcal{R}$.*

In terms of line integrals, Pappus's theorem says:

$$\operatorname{vol}(\mathcal{Q}) = \int_C A(s)\,ds,$$

where $\mathcal{C}$ is the circular arc traversed by the centroid of $\mathcal{R}$, s denotes arc length along $\mathcal{C}$, and $A(s) = \operatorname{area}(\mathcal{R})$ is the (constant) area of the orthogonal cross section of $\mathcal{Q}$ at the point s units along $\mathcal{C}$.

This formula seems to be considerably less intuitive than the case of parallel cross sections—the volume swept out by the outer half of the region $\mathcal{R}$ is greater than that swept out by the inner half. This difference is apparent in FIGURE 1, which shows a cylindrical tube obtained by rotating a unit square $\mathcal{R}$ about the z-axis. Taking cross sections perpendicular to the z-axis, we compute the volumes swept out by the inner and outer halves of the square to be $5/4\,\pi$ and $7/4\,\pi$, respectively.

It is however a simple exercise to transform a cylinder with base $\mathcal{R}$ into a solid of revolution with cross section $\mathcal{R}$ and thus to obtain Pappus's theorem as a consequence of the change of variables theorem; indeed, this is just the familiar "cylindrical shells" method of calculating volumes.

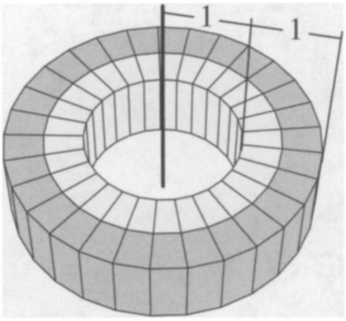


Figure 1 Solids of revolution of two halves of a square

The purpose of this note is to show that Pappus's theorem extends to more general curves and to solids with variable cross sections whose centroids do not lie on the given curve $\mathcal{C}$. All the tools needed are in a standard multivariate calculus course, specifically, the calculus of curves in space and the change of variables theorem for multiple integrals.

With the incorporation of systems such as *Mathematica* and *Maple* into calculus courses, the study of parameterized curves, surfaces, and solids has been facilitated. Now we can look at an object of interest from a variety of viewpoints, and computer algebra systems give us a way around the often formidable calculations necessary for its analysis. However, even for solids with explicit parameterizations, which can thus be plotted, naïvely attempting to compute the volume by calculating the Jacobian of the transformation directly can be impractical, and, of course, such an approach conveys none of the geometry that highlights Pappus's theorem and Cavalieri's principle.

A generalized method of Pappus In order to state our version of Pappus's theorem, we must introduce some notation. The basic facts that we use are available in any standard calculus text; for example, see Stewart [3]. Suppose that $\mathcal{C}$ is a curve in space parameterized by arc length as $(x, y, z) = \sigma(s), 0 \leq s \leq \ell$. In order to describe a solid in terms of orthogonal cross sections relative to $\mathcal{C}$, we use the so-called *moving frame*, or Frenet frame associated with $\mathcal{C}$, $\{\mathbf{T}(s), \mathbf{N}(s), \mathbf{B}(s)\}$, which can be constructed for any curve with a continuous unit normal. With this assumption, the vectors of the moving frame are defined by the equations

$$\begin{aligned} \mathbf{T}(s) &= \sigma'(s) \\ \sigma''(s) &= \kappa(s)\,\mathbf{N}(s) \\ \mathbf{B}(s) &= \mathbf{T}(s) \times \mathbf{N}(s), \end{aligned}$$

where $\kappa(s)$ denotes the curvature of $\mathcal{C}$ at $\sigma(s)$. Note that if $\mathcal{C}$ has zero curvature, then it is a line; then σ' is constant, and any unit vector $\mathbf{N} \perp \sigma'$ will do.

At each point $\sigma(s)$ on the curve, the vectors $\mathbf{T}(s)$, $\mathbf{N}(s)$ and $\mathbf{B}(s)$ form an orthonormal basis for Euclidean space. As illustrated in FIGURE 2, the plane $\mathcal{P}(s)$ orthogonal to $\mathcal{C}$ at $\sigma(s)$, known as the *normal plane*, can be described as the set of points of the

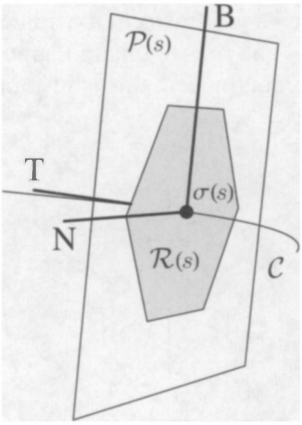


Figure 2 The Frenet frame for $\mathcal{C}$

form

$$(x, y, z) = \sigma(s) + r\,\mathbf{N}(s) + t\,\mathbf{B}(s), \quad \text{for real numbers } r \text{ and } t.$$

We propose to study solids $\mathcal{Q}$ that may be parameterized as the set of points (x, y, z) satisfying

$$(x, y, z) = G(s, r, t) = \sigma(s) + r\mathbf{N}(s) + t\mathbf{B}(s),$$

where $0 \le s \le \ell$, and the pair (r, t) lies in a region $D(s)$ of the r, t-plane. The cross section of $\mathcal{Q}$ at a point s units along $\mathcal{C}$ is then the region $\mathcal{R}(s)$ lying in $\mathcal{P}(s)$:

$$\mathcal{R}(s) = \{r\mathbf{N}(s) + t\mathbf{B}(s) : (r, t) \in D(s)\}.$$

We assume that these regions $D(s)$ are such that the cross sections $\mathcal{R}(s_1)$ and $\mathcal{R}(s_2)$ do not intersect whenever $s_1 \neq s_2$.

Finally, we can state our result.

Proposition. Let $A(s)$ denote the area of the orthogonal cross section $\mathcal{R}(s)$ of $\mathcal{Q}$ relative to $\mathcal{C}$, and suppose that $\mathcal{R}(s)$ has centroid $\sigma(s) + \bar{r}(s)\,\mathbf{N}(s) + \bar{t}(s)\,\mathbf{B}(s)$. Then the volume of $\mathcal{Q}$ is

$$\boxed{\operatorname{vol}(\mathcal{Q}) = \int_{\mathcal{C}} A(s)\,(1 - \kappa(s)\bar{r}(s))\,ds.}$$

Notice that when the curvature is 0, the formula above reduces to the familiar integration of cross sections along a line; also, when $\mathcal{Q}$ is a *tube*, generated by translating a planar region $\mathcal{R}$ through space, then we recover a version of Pappus' theorem: If $\mathcal{C}$ is the curve traced by the centroid of $\mathcal{R}$, then the volume of the tube $\mathcal{Q}$ is Area($\mathcal{R}$) length($\mathcal{C}$). Thus the volume formula in the proposition ties together the method of volumes by slicing and Pappus's theorem.

The key to the proof is the classic formulas of Frenet and Serret for $\mathbf{T}'(s)$, $\mathbf{N}'(s)$ and $\mathbf{B}'(s)$. These are themselves nice applications of the algebraic properties of the cross product and the product rule for vector-valued functions.

Proof of the Proposition. By the change of variables theorem [**2**] or [**3**], the volume of such a solid $\mathcal{Q}$ is

$$\operatorname{vol}(\mathcal{Q}) = \int_0^{\ell} \iint_{D(s)} J_G(s, r, t)\, dA(r, t)\, ds,$$

where J_G is the Jacobian of the transformation G and $dA(r, t)$ is area measure. As we mentioned before, trying to compute and then integrate J_G, even numerically, is hopeless except in the very simplest cases. We can, however, find a meaningful expression for the Jacobian by using its representation as

$$J_G(s, r, t) = |G_s \cdot (G_r \times G_t)|$$

where G_s, G_r and G_t are the (vector) partial derivatives of the transformation G. Clearly

$$\begin{aligned} G_s &= \frac{\partial G}{\partial s} = \sigma'(s) + r\mathbf{N}'(s) + t\mathbf{B}'(s) \\ G_r &= \frac{\partial G}{\partial r} = \mathbf{N}(s) \\ G_t &= \frac{\partial G}{\partial t} = \mathbf{B}(s). \end{aligned}$$

The first of these derivatives can be simplified with the Frenet–Serret formulas; see, for example, Grey [**1**, Theorem 7.3], or Stewart [**3**, Section 11.8, exercise 44]:

$$\begin{aligned} \mathbf{T}'(s) &= \kappa(s)\mathbf{N}(s) \\ \mathbf{N}'(s) &= -\kappa(s)\mathbf{T}(s) + \tau(s)\mathbf{B}(s) \\ \mathbf{B}'(s) &= -\tau(s)\mathbf{N}(s), \end{aligned}$$

where $\tau(s)$ is the torsion of $\mathcal{C}$ at $\sigma(s)$. These give

$$\begin{aligned} G_s &= \mathbf{T}(s) + r(-\kappa(s)\mathbf{T}(s) + \tau(s)\mathbf{B}(s)) + t(-\tau(s)\mathbf{N}(s)) \\ &= (1 - r\kappa(s))\mathbf{T}(s) - t\tau(s)\mathbf{N}(s) + r\tau(s)\mathbf{B}(s). \end{aligned}$$

Thus

$$J_G(s, r, t) = |G_s \cdot (\mathbf{N}(s) \times \mathbf{B}(s))| = |G_s \cdot \mathbf{T}(s)| = 1 - r\kappa(s). \qquad (*)$$

The volume can now be calculated

$$\begin{aligned} \operatorname{vol}(\mathcal{Q}) &= \int_0^{\ell} \iint_{D(s)} (1 - r\kappa(s))\, dA(r, t)\, ds \\ &= \int_0^{\ell} \left(\iint_{D(s)} dA(r, t) - \kappa(s) \iint_{D(s)} r\, dA(r, t) \right) ds \\ &= \int_0^{\ell} (A(s) - \kappa(s) M_B(s))\, ds, \end{aligned}$$

where $M_B(s)$ is the moment of the region $\mathcal{R}(s)$ relative to the $\mathbf{B}(s)$-axis. If $\bar{r}(s)$ is the $\mathbf{N}(s)$-coordinate of the centroid of the region $\mathcal{R}(s)$, then $M_B(s) = \bar{r}(s)A(s)$, and our result becomes

$$\operatorname{vol}(\mathcal{Q}) = \int_0^{\ell} A(s)[1 - \kappa(s)\bar{r}(s)]\, ds. \qquad \blacksquare$$

Applications Suppose $\mathcal{Q}$ is a section of pipe having an inside radius r, which is bent into an arc of a circle $x^2 + y^2 = R^2$ as in FIGURE 3.

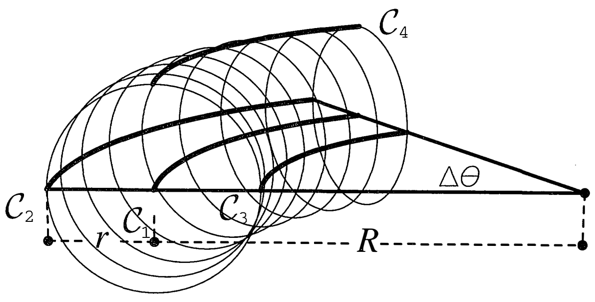


Figure 3 A curved pipe

We may represent the section of pipe as being generated by moving a disk of radius r centered on the curve $\mathcal{C}_1$ along the curve. Pappus's original result then calculates the volume of the pipe as

$$\operatorname{vol}(\mathcal{Q}) = (\pi r^2)(R\,\Delta\theta).$$

On the other hand, we may also consider the curve $\mathcal{C}_2$ as the generating curve. In this case, the parametrization of $\mathcal{C}_2$ by arc length is

$$\sigma_2(s) = ((R+r)\cos(s/(R+r)), (R+r)\sin(s/(R+r)), 0);$$

and the Frenet frame at $\sigma_2(s)$ is

$$\begin{aligned}\mathbf{T}(s) &= (-\sin(s/(R+r)), \cos(s/(R+r)), 0)\\ \mathbf{N}(s) &= (-\cos(s/(R+r)), -\sin(s/(R+r)), 0)\\ \mathbf{B}(s) &= (0, 0, 1).\end{aligned}$$

Since the $\mathbf{N}(s)$-coordinate of the centroid of the disk is $\bar{r}(s) = r$ and $\kappa(s) = \frac{1}{R+r}$, our result gives

$$\operatorname{vol}(\mathcal{Q}) = (\pi r^2)\left(1 - r\left(\frac{1}{R+r}\right)\right)[(R+r)\,\Delta\theta] = (\pi r^2)(R\,\Delta\theta).$$

Similarly, if we think of $\mathcal{C}_3$ as the generating curve then $\bar{r}(s) = -r$, the curvature is $\kappa(s) = \frac{1}{R-r}$, and so

$$\operatorname{vol}(\mathcal{Q}) = (\pi r^2)\left(1 - (-r)\left(\frac{1}{R-r}\right)\right)[(R-r)\,\Delta\theta] = (\pi r^2)(R\,\Delta\theta).$$

Finally, if we choose to regard $\mathcal{C}_4$ as the generating curve, the $\mathbf{N}(s)$-coordinate of the disk is $\bar{r}(s) = 0$; again our result reduces to the classical one.

Usually, curves are given in terms of a parameter other than arc length. If $\mathcal{C}$ has parameterization $\sigma(t)$, $a \le t \le b$, where $\|\sigma'(t)\| > 0$ for all t, let $\kappa(t)$ be the curvature of $\mathcal{C}$ at the point $\sigma(t)$, and so forth; in particular, $\bar{r}(t)$ denotes the $\mathbf{N}(t)$-coordinate of the centroid of the cross section of the solid $\mathcal{Q}$ at $\sigma(t)$. In this case our formula becomes

$$\operatorname{vol}(\mathcal{Q}) = \int_a^b A(t)\,(1 - \kappa(t)\bar{r}(t))\,\|\sigma'(t)\|\,dt.$$

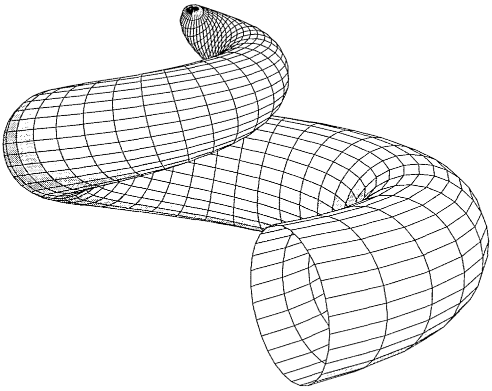


Figure 4 The horn $\mathcal{H}$

The *horn* $\mathcal{H}$ in FIGURE 4 has a spiral as its generating curve:

$$\sigma(t) = ((4-t)\cos(\pi t), (4-t)\sin(\pi t), t), \quad 0 \le t \le 4$$

The cross sections, $\mathcal{R}(t)$, are disks of radius $\sqrt{4-t}/2$ centered at $\sigma(t)$. We parameterize $\mathcal{H}$ as

$$(x, y, z) = G(r, s, t) = \sigma(t) + \frac{r}{2}\sqrt{4-t}(\cos s\,\mathbf{B}(t) + \sin s\,\mathbf{N}(t)),$$

for $0 \le r \le 1$, $0 \le s \le 2\pi$, and $0 \le t \le 4$, where $\mathbf{N}(t)$ and $\mathbf{B}(t)$ are the principal unit normal and binormal to the curve at $\sigma(t)$.

Because of the symmetry of the cross sections relative to the generating curve, the proposition lets us write the volume simply as

$$\begin{aligned}\operatorname{vol}(\mathcal{H}) &= \int_0^4 A(t)\,\|\sigma'(t)\|\,dt\\ &= \int_0^4 \pi\frac{4-t}{4}\sqrt{2+\pi^2(4-t)^2}\,dt = \frac{\pi(8\pi^2+1)^{3/2}-1}{3\sqrt{2}}.\end{aligned}$$

We invite the reader to verify this answer directly by calculating the Jacobian and integrating numerically. (Using *Mathematica*, our printout of the Jacobian was five pages long!)

As a final example, we spiral a square around a generating curve. Let $\mathcal{C}$ be the circle of radius 2 in the xy-plane, and let $\sigma(t) = (2\cos t, 2\sin t, 0)$ be the usual parameterization of $\mathcal{C}$. The normal and binormal for $\mathcal{C}$ at $\sigma(t)$ are $\mathbf{N}(t) = (-\cos t, -\sin t, 0)$ and $\mathbf{B}(t) = (0, 0, 1)$. Define $\mathbf{u}(t) = \cos(8t)\,\mathbf{N}(t) + \sin(8t)\,\mathbf{B}(t)$ and $\mathbf{v}(t) = -\sin(8t)\,\mathbf{N}(t) + \cos(8t)\,\mathbf{B}(t)$. The vectors $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are simply rotations of $\mathbf{N}(t)$ and $\mathbf{B}(t)$ by an angle of $8t$ radians around the point $\sigma(t)$ in the plane $\mathcal{P}(t)$. Take the square with sides $\mathbf{u}(t)$ and $\mathbf{v}(t)$ as the cross section, $\mathcal{R}(t)$. The solid $\mathcal{Q}$ is parameterized by

$$G(r, s, t) = \sigma(t) + r\,\mathbf{u}(t) + s\,\mathbf{v}(t), \quad 0 \le r \le 1, 0 \le s \le 1, \; 0 \le t \le \pi/16.$$

Four cross sections are represented in FIGURE 5.

Even though $\mathcal{Q}$ is not parameterized directly in terms of the Frenet frame, we can apply the proposition to compute the volume. The centroid of the cross section $\mathcal{R}(t)$ is

$$G(1/2, 1/2, t) = \frac{1}{2}(\cos(8t) - \sin(8t))\mathbf{N}(t) + \frac{1}{2}(\sin(8t) + \cos(8t))\mathbf{B}(t).$$

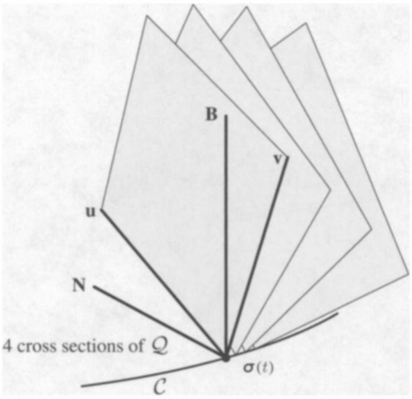


Figure 5 Cross sections of $\mathcal{Q}$

Since the curvature of the generating circle is $1/2$ and the area of each cross section is 1, we obtain

$$\operatorname{vol}(\mathcal{Q}) = \int_0^{\pi/16} \left(1 - \frac{1}{2}\frac{\cos(8t) - \sin(8t)}{2}\right) 2\,dt = \frac{\pi}{8}.$$

Using the change of variables formula and the formula for the Jacobian $(*)$, we can calculate other integrals associated with $\mathcal{Q}$. For example, let's compute the z-coordinate of the centroid of the $\mathcal{Q}$. Let F be the transformation from **N B**-coordinates to **u v**-coordinates in the plane $\mathcal{P}(t)$:

$$(r, s, t) = F(\mu, \eta, t) = \begin{pmatrix} \cos(8t) & \sin(8t) & 0 \\ -\sin(8t) & \cos(8t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \eta \\ t \end{pmatrix}.$$

Then the parameterization of $\mathcal{Q}$ in terms of the Frenet frame coordinates is

$$(x, y, z) = G(F(\mu, \eta, t)) = H(\mu, \eta, t),$$

and so, by the chain rule and the formula $(*)$ for J_H, we have that

$$1 - \mu\,\kappa(t) = J_H(\mu, \eta, t) = J_G(r, s, t) J_F(\mu, \eta, t).$$

But J_F is the constant function $J_F(\mu, \eta, t) = 1$; thus, solving for μ in terms of r, s and t, and using the fact that $\kappa(t) = 1/2$, we obtain

$$J_G(r, s, t) = 1 - (\cos(8t)r - \sin(8t)s)/2.$$

The z-coordinate of the centroid of $\mathcal{Q}$ is

$$\begin{aligned} \frac{1}{\operatorname{vol}(\mathcal{Q})} \iiint_{\mathcal{Q}} z\,dV &= \frac{8}{\pi} \int_0^{\pi/16} \int_0^1 \int_0^1 z(r, s, t)\, J_G(r, s, t)\, dr\, ds\, dt \\ &= \frac{8}{\pi} \int_0^{\pi/16} \int_0^1 \int_0^1 (\sin(8t)r + \cos(8t)s) \left(1 - \frac{\cos(8t)r - \sin(8t)s}{2}\right) dr\, ds\, dt \\ &= \frac{1}{\pi}. \end{aligned}$$

Use of computer algebra systems has transformed the core of the undergraduate mathematics curriculum for science and engineering students. The use of this technology makes abstract concepts more concrete, and applications previously out of reach now enliven traditional topics. Tubes generated by specifying a curve and cross section occur naturally in a variety of settings, and their construction is an excellent illustration of some of the basic ideas in vector calculus. The graphical capabilities of systems like *Mathematica* and *Maple* enable us to see the results of these constructions, complementing traditional analytic and geometric techniques. On the other hand, while technology can enhance the standard curriculum, it is not a replacement. The parameterization of tubes as above is based on local orthonormal coordinate systems; so it is natural to exploit this geometry in the analysis. Indeed, computations are often impossible otherwise, even given the power of a computer algebra system.

REFERENCES

1. A. Gray, *Modern Differential Geometry of Curves and Surfaces*, CRC Press, Boca Raton, FL, 1993.
2. J. R. Munkres, *Analysis on Manifolds*, Addison-Wesley Publishing Company, The Advanced Book Program, Redwood City, CA, 1991.
3. J. Stewart, *Calculus, 3rd Edition*, Brooks/Cole Publ. Co., Pacific Grove, CA, 1994.

Two Reflected Analyses of Lights Out

ÓSCAR MARTÍN-SÁNCHEZ
CRISTÓBAL PAREJA-FLORES
Departamento de Sistemas Informáticos y Programación
Universidad Complutense de Madrid
Madrid, Spain

The device was described to us as a beeping hand-held electromechanical puzzle, with buttons that turned lights on and off. But we subsequently found that it was nothing of the sort. After playing several games, we felt the urge to open the screws and peer inside. What we found packed in this device really had nothing to do with mechanics or electronics at all ...

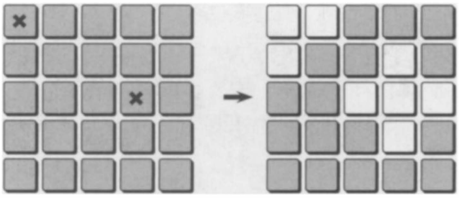
As someone with an *analytic mind* might point out, what is packed inside is strategy and empirical reasoning. To solve the puzzle, one needs nothing more than what is needed to solve a jigsaw puzzle: the human mind, some methodical work, and a little care.

As someone with *mathematical understanding* might point out, what is packed inside is a combination of matrices, vector spaces and scalar products. Indeed, this is just an instance of a well-known algebraic model—a system of linear equations—solvable with standard mathematical tools.

Two analyses follow: the left column addresses a person who is interested and methodical; the right one, a mathematician. These analyses are mirror images of one another: the concepts, examples, and figures on each side are designed to enrich their counterparts. Thus, for greater enjoyment and better understanding, we recommend a parallel reading of the two columns, passing fearlessly through the looking glass in the gap between the columns.

Our aim is to provide the reader with an understanding of the game, efficient algorithms to know when the game can be solved, and also how to find the solution. Our work is partially based on an article in the MAGAZINE by Anderson and Feil [**1**], but differs from that analysis in offering a way to solve the puzzle with the game in hand, without needing a computer or even pencil and paper. It also has elements of the theory of σ^+-automata [**9, 3**]. A collection of electronic resources for *Lights Out* can be found at http://www.maa.org/pubs/mathmag.html.

Rules of the game *Lights Out®* (Tiger Electronics) is a puzzle sold in toy stores. It consists of a 5 × 5 board of cells, where each cell is simultaneously a light and a button. Each light can either be on (white in the pictures) or off (gray). Pressing a button (marked with a cross in the figure), changes the on/off state of that light and also the lights of its horizontal and vertical neighbors. For example, if we start with all the lights off, then pressing the first button in the first row, and the fourth button in the third row, changes the lights according to the figure:



Such are the rules of the game.

The aim of the game is, starting with any given state of the board, to turn all the lights off (or *out*, if you prefer).

Although we are talking here about 5 × 5 boards, we do so in such a way that our methods can be applied to any $m \times n$ board.

First remarks To solve the puzzle we must press a number of buttons; during this process, some lights are switched several times, but we are only interested in the final result.

Pressing a button twice has no effect. Also, solving a given state is the same as reaching it from a fully unlit board: the same presses are needed.

Pressing one button and then another one has the same effect as pressing them in reverse order. This is because the final state of a cell depends only on (the initial state and) the number of buttons pressed that switch it; it has nothing to do with the order.

Statement of the problem *Lights Out* is a problem in matrix algebra. We work with vectors of $(\mathbb{Z}_2)^{25}$. We are given the following boxed matrix R, which encodes the rules of the puzzle:

$$R = \begin{pmatrix} A & I & O & O & O \\ I & A & I & O & O \\ O & I & A & I & O \\ O & O & I & A & I \\ O & O & O & I & A \end{pmatrix}, A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

where I and O are the 5 × 5 identity and zero matrices. For a vector $\vec{p}$, which we call a *press* vector, we calculate $R\vec{p}$, called the *effect* of $\vec{p}$. The effect of $\vec{p}$ is added (modulo 2) to vector $\vec{s}$, the initial *state*, to obtain a new state $R\vec{p} + \vec{s}$. For example, with the press vector that has 1s only in the 1^{st} and 14^{th} components, the null state is changed as follows:

$$R \times \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \vec{0} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(When needed, 25-component vectors are arranged in a 5 × 5 fashion.) The aim is, given a state $\vec{s}$, to find $\vec{p}$ such that $R\vec{p} + \vec{s} = \vec{0}$. Although we are talking here about 25-component vectors, our methods can be applied to vectors of any $m \times n$ size, and matrices resembling R.

First remarks First, note that the names we give to vectors and other elements (like press vectors, effects, states) are just a way of speaking: no physical meaning is needed.

An equivalent statement of the problem is, given $\vec{s}$, to find $\vec{p}$ such that $R\vec{p} = \vec{s}$. This is a well-known algebraic problem. (Observe that, modulo 2, $\vec{s} = -\vec{s}$.)

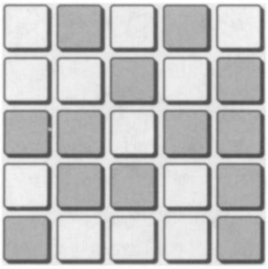
Observe that it makes sense to solve the problem, for a given $\vec{s}$, in more than one step: starting from the null state, $\vec{0}$, we might want to press $\vec{p}_1$, so that the new state, $\vec{s}_1$, is of a more convenient form, then press $\vec{p}_2$, etc.:

$$\vec{0} \xrightarrow{\vec{p}_1} \vec{s}_1 \xrightarrow{\vec{p}_2} \cdots \longrightarrow \vec{s}$$

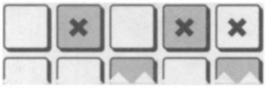
Therefore, a set of presses has the same effect if we remove all pairs of equal presses. Such a set of presses, where no cell is pressed more than once, we call a *procedure*.

The puzzle can be solved by trying out all the possible combinations. Since, for every button, we have to choose whether to press it or not, we have a huge total of 2^{25} procedures. We seek to simplify the solving method for someone with the game in hand.

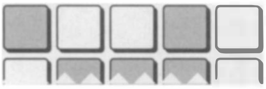
Reducing the search Given a state that we want to solve,



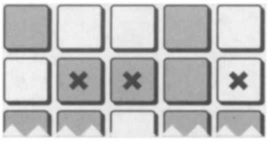
suppose we arbitrarily choose some buttons in row 1 (the top one), as below:



After pressing them, some lights in row 1 may be on



and the only way to turn them off (without further presses in row 1, and without switching on the lights that are off) is to press the buttons in row 2 that are exactly under any lights that are on in row 1:



The state of row 2 then determines what must be pressed in row 3, and so on through the rows. Therefore, given a particular state and chosing a set of buttons to press in the first row, the rest of the procedure is determined:

The solution would be the sum of the $\vec{p}_i$s:

$$\vec{0} \xrightarrow{\vec{p}_1+\vec{p}_2+\cdots} \vec{s}$$

As this is a known algebraic problem, we could use standard methods to solve it, as used in [1]. But these standard methods use matrices in general, whereas here we seek to exploit the particular features of R.

Finding linear dependences The problem can be written as follows:

$$R\cdot\begin{pmatrix}\vec{p}_{1\bullet}\\ \vec{p}_{2\bullet}\\ \vec{p}_{3\bullet}\\ \vec{p}_{4\bullet}\\ \vec{p}_{5\bullet}\end{pmatrix}=\begin{pmatrix}\vec{s}_{1\bullet}\\ \vec{s}_{2\bullet}\\ \vec{s}_{3\bullet}\\ \vec{s}_{4\bullet}\\ \vec{s}_{5\bullet}\end{pmatrix}$$

where, for the sake of clarity, $\vec{p}$ and $\vec{s}$ have been divided into subvectors $\vec{p}_{i\bullet}$ and $\vec{s}_{i\bullet}$, each consisting of 5 components.

Manipulating the previous equation leads to something more convenient: an expression for $\vec{p}_{1\bullet}$ depending only on $\vec{s}$, and others for $\vec{p}_{2\bullet}\ldots\vec{p}_{5\bullet}$ depending only on $\vec{p}_{1\bullet}$ and $\vec{s}$.

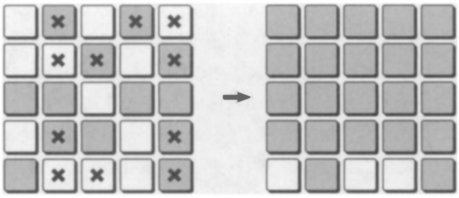
The equation $R\vec{p}=\vec{s}$ is equivalent to $J\vec{p}=(R+J)\vec{p}+\vec{s}$ for any matrix J. Using, in particular,

$$J=\begin{pmatrix}O & I & O & O & O\\ O & O & I & O & O\\ O & O & O & I & O\\ O & O & O & O & I\\ O & O & O & O & O\end{pmatrix}$$

we find:

$$\begin{pmatrix}\vec{p}_{2\bullet}\\ \vec{p}_{3\bullet}\\ \vec{p}_{4\bullet}\\ \vec{p}_{5\bullet}\\ \vec{0}\end{pmatrix}=\begin{pmatrix}A & O & O & O & O\\ I & A & O & O & O\\ O & I & A & O & O\\ O & O & I & A & O\\ O & O & O & I & A\end{pmatrix}\begin{pmatrix}\vec{p}_{1\bullet}\\ \vec{p}_{2\bullet}\\ \vec{p}_{3\bullet}\\ \vec{p}_{4\bullet}\\ \vec{p}_{5\bullet}\end{pmatrix}+\begin{pmatrix}\vec{s}_{1\bullet}\\ \vec{s}_{2\bullet}\\ \vec{s}_{3\bullet}\\ \vec{s}_{4\bullet}\\ \vec{s}_{5\bullet}\end{pmatrix}$$

Here, each $\vec{p}_{i\bullet}$ depends only on the subvectors of $\vec{p}$ and $\vec{s}$ with indices smaller than its own. Thus, we can use the equation from the first row to remove $\vec{p}_{2\bullet}$ from the second, then use the second row to remove $\vec{p}_{3\bullet}$ from the third, and so on.



But note that this *gathering* procedure only guarantees to turn off all the lights in the four upper rows: lights in the fifth row may be still on after this procedure.

We can thus deal with the solution to a given board in three steps:

- first, we apply the gathering procedure down to the last row, without doing any presses in row 1;
- then, we look for presses in the first row that, *when gathered*, would turn the remaining lights off;
- finally, we press those buttons in the first row, and gather the result.

But it could be that no set of presses in the first row can actually fulfill this task. In this case, there is no solution. We will return to this issue later.

The question is: which buttons must be pressed in the first row? In order to find out, for each button in the first row we start with an unlit 5 × 5 board, press that button, and then gather the resulting position down. These are the results in the *fifth* row, for each button in the *first* one:

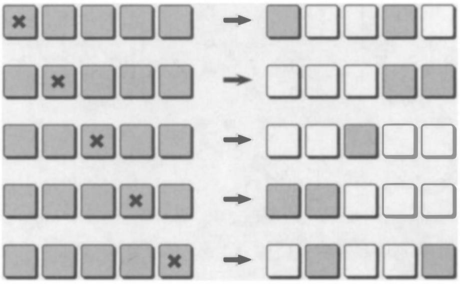


Figure 1

Therefore, $\vec{p}_{2\bullet} \dots \vec{p}_{5\bullet}$ are expressed depending only on $\vec{s}$ and $\vec{p}_{1\bullet}$:

$$\begin{pmatrix} \vec{p}_{2\bullet} \\ \vec{p}_{3\bullet} \\ \vec{p}_{4\bullet} \\ \vec{p}_{5\bullet} \\ \vec{0} \end{pmatrix} = \begin{pmatrix} B_1 & B_0 & O & O & O & O \\ B_2 & B_1 & B_0 & O & O & O \\ B_3 & B_2 & B_1 & B_0 & O & O \\ B_4 & B_3 & B_2 & B_1 & B_0 & O \\ B_5 & B_4 & B_3 & B_2 & B_1 & B_0 \end{pmatrix} \begin{pmatrix} \vec{p}_{1\bullet} \\ \vec{s}_{1\bullet} \\ \vec{s}_{2\bullet} \\ \vec{s}_{3\bullet} \\ \vec{s}_{4\bullet} \\ \vec{s}_{5\bullet} \end{pmatrix}$$

where $B_0 = I$, $B_1 = A$ and $B_{n+2} = A \times B_{n+1} + B_n$. Now, the equation

$$B_5 \vec{p}_{1\bullet} = B_4 \vec{s}_{1\bullet} + B_3 \vec{s}_{2\bullet} + B_2 \vec{s}_{3\bullet} + B_1 \vec{s}_{4\bullet} + B_0 \vec{s}_{5\bullet}$$

taken from the bottom row of the former matrix, involves only the $\vec{p}_{1\bullet}$ part of $\vec{p}$.

We can thus deal with this equation in three steps:

- first, we calculate the right-hand side, which only depends on $\vec{s}$; we will call this vector *gather*($\vec{s}$) for brevity;
- then, we try to find $\vec{p}_{1\bullet}$ such that $B_5 \vec{p}_{1\bullet}$ equals the now-known *gather*($\vec{s}$);
- finally, we calculate $\vec{p}_{2\bullet} \dots \vec{p}_{5\bullet}$ directly from $\vec{p}_{1\bullet}$ and $\vec{s}$, using the equations in the four upper rows of the previous matrix.

There exists some $\vec{s}$ for which no $\vec{p}_{1\bullet}$ satisfies the equation above. In such cases, the problem has no solution. We will study this issue later.

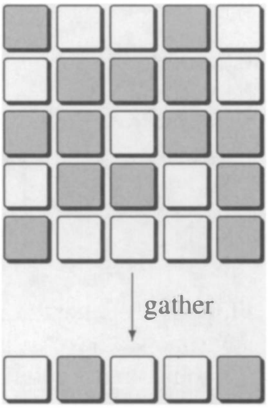
The only remaining question is how to obtain $\vec{p}_{1\bullet}$ such that it satisfies the equation

$$B_5 \, \vec{p}_{1\bullet} = gather(\vec{s}).$$

The matrix B_5, that will be necessary below, is the following:

$$B_5 = A^5 + A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

We have thus shifted from looking for which of 25 buttons need to be pressed in order to turn 25 possible lights off, to the simpler case of 5 buttons and 5 lights; this simplifies the search for solutions. In the following example, the gathering happens to produce the effect we calculated for button 5 above; thus, the solution is achieved by pressing button 5 in row 1 and gathering.



In a sense, we are now playing on a 1 × 5 board, with very different rules for how the lights change when a button is pressed, as can be seen in FIGURE 1. The new game inherits many properties from the original one: the order of presses is not relevant, and double presses have no effect.

Every solution to this game (1 × 5) gives a solution to the original game (5 × 5), and *vice versa.* Non-solvable states (if any) also correspond; the same can be said about states with more than one solution (if any). Procedures in the large board are called the *expansion* of the ones in the small one with respect to the state being solved.

In what follows, we are mainly playing with reduced boards, the results and conclusions of which correspond to those of the large boards.

Neutral but useful procedures Looking at the rules of the reduced game in FIGURE 1, some equivalences between procedures can be found, for instance:

The linear dependences found allow us to see the 25-dimensional problem

$$R\,\vec{p} = \vec{s},$$

as the 5-dimensional one

$$B_5\,\vec{p}_{1\bullet} = gather(\vec{s}).$$

As an example, the following state is reduced as shown below, leading to a much smaller problem:

$$\begin{tikzcd}
\begin{array}{ccccc} (0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0) \end{array} \arrow[d, "gather"] \\
(1\ \ 0\ \ 1\ \ 1\ \ 0)
\end{tikzcd}$$

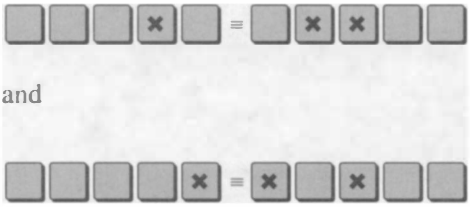
Each solution to this problem gives one and only one solution to the original problem, and *vice versa.* Thus, the following diagram is commutative:

$$\begin{tikzcd}
\vec{s} \arrow[r, "\substack{\text{solution}\\\text{(original rules)}}"] \arrow[d, "gather"'] & \vec{p} \\
gather(\vec{s}) \arrow[r, "\substack{\text{solution}\\\text{(reduced rules)}}"'] & \vec{p}_{1\bullet} \arrow[u, "expand"']
\end{tikzcd}$$

We call the 25-component vectors the *expansion* of the 5-component vectors with respect to the state being solved: $\vec{p} = expand(\vec{p}_{1\bullet}, \vec{s})$.

From the equivalence of these problems, we get that $null(R)$ and $null(B_5)$ have the same dimension. In the next section, we explore this issue further.

Image and null space of R We can perform Gauss-Jordan reduction on B_5 to get $XB_5 = E$, obtaining the matrices:



Several useful conclusions can be drawn from here. To begin with, these equivalences show that it is never necessary to push the two buttons farthest to the right: a solution using those buttons could be replaced by an equivalent one in which they are not used at all. For the purposes of our solution, we could pretend they were broken.

Also, we note that if two equivalent procedures are performed in turn, one undoes the other, and the overall result is null. The composition of these two procedures will be called a neutral procedure.

Combining each pair of equivalent procedures above gives two different neutral procedures. A third is obtained by combining these two. There is also a fourth, the trivial one, in which no buttons are pressed. A systematic search proves that there are no others.

When any procedure is followed by (or composed with) a neutral one, an equivalent procedure results. Therefore, when there is a solution for a given state, there are actually four solutions (as there are four neutral procedures).

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$$E = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We note that $null(E)$ equals $null(B_5)$, and its dimension is 2, and therefore so is $null(R)$.

Computation produces a basis of $null(E)$, the vectors:

$$\vec{i} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \vec{j} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

These allow us to generate additional solutions any time we have a single one: If $\vec{p}$ is a (5-component) solution for $\vec{s}$, there are three other solutions:

$$\vec{p} + \vec{i} \qquad \vec{p} + \vec{j} \qquad \vec{p} + \vec{i} + \vec{j}$$

Inspecting the form of $\vec{i}$ and $\vec{j}$, we see that one of these four solutions will have zeroes in the fourth and fifth components.

Is it even possible? We can now find out if a state is, or is not, solvable with just 8 tries. But we can achieve a more direct solution using more subtle reasoning.

First, note that the games, both the reduced and the original, have a certain symmetry: if button x toggles light y, then button y toggles light x. Any neutral procedure, because it is neutral, presses an even number of buttons in the neighborhood of *any* cell, leaving it unchanged. But then, symmetrically, pressing any button switches an even number of cells in N, the set of cells pressed in a neutral procedure. Therefore, the parity

Does a solution exist? We can reduce our search for solutions to 3-component vectors. Therefore, only 2^3 states are reachable out of the total posible 2^5 states.

In other words, the image of our matrix has the dimension of 3; we know this because the order of a square matrix (5 in our case) equals the dimension of its null space (2) plus the dimension of its image.

Therefore, not every state can be solved; it is thus worthwhile to look for criteria to identify the solvable states.

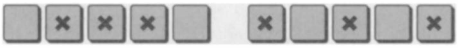
Algebra tells us that, given a symmetric matrix (like B_5), its image is orthogonal to its null space. So, for a given $\vec{s}$, there of the lit cells in N cannot be changed by any press.

So, from the fully unlit state, we can only get states with an even number of cells in common with any such neutral procedure N. Therefore, a state is solvable only if it has an *even* intersection with any neutral procedure.

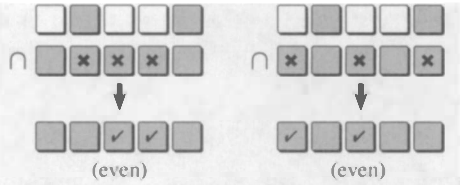
For a given neutral procedure, this test halves the number of possibly solvable states. Since any neutral procedure is the composition of the other two, it is enough to test it for two of them. Thus, we can halve the number of solvable states twice, and mark three quarters of the states as unsolvable.

On the other hand, each solvable state can be solved by means of exactly four different procedures. Therefore, one quarter of the states are solvable: those that pass the previous test.

Let's choose these two neutral procedures to perform the parity test:



We can use these results to show that the example we used before is solvable. The target state had lights 1, 3, and 4 on; comparing this with the first neutral state above, we find two lights in common: the third and fourth; comparing with the second again gives an even number: the first and third buttons. Thus the state is solvable:



All even ⇒ solvable

How to solve it? Now we have a puzzle with a state we want to solve. How are we going to do it? We know we need only focus on the three buttons on the left. Their effects are as follows:

exists a solution $\vec{p}$ for the equation

$$B_5\vec{p} = \vec{s},$$

if and only if $\vec{s}$ is orthogonal to the null space of the matrix. Because of the equivalence between the expanded and reduced problems, we know that this condition can be checked either with 5×5 matrices, or 25×25 ones.

Working with the convenient smaller system, for $\vec{s}$ to be solvable, the scalar product $\vec{s} \cdot \vec{n}$ should be 0 for every $\vec{n} \in null(B_5)$. And it is enough to test it for $\vec{n}$ in the basis of $null(B_5)$. This amounts to checking the following equations:

$$s_2 + s_3 + s_4 = 0$$

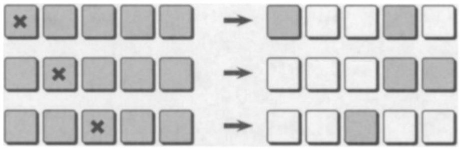
$$s_1 + s_3 + s_5 = 0.$$

Let's use this theory to show that the example we used before is solvable. In the figures below, in each box, we calculate the scalar product of the state under consideration and one of the neutral procedures. We find that both are null, that is, the two equations above are satisfied. Therefore, the state is solvable. And so is the corresponding expanded state whose *gather* yields this reduced state.

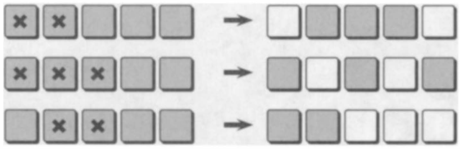
$$\begin{array}{cc}\boxed{\begin{array}{c}(1,0,1,1,0)\\ \cdot\,(0,1,1,1,0)\\ \shortparallel\\ 0+0+1+1+0\end{array}} & \boxed{\begin{array}{c}(1,0,1,1,0)\\ \cdot\,(1,0,1,0,1)\\ \shortparallel\\ 1+0+1+0+0\end{array}}\\ (s_2+s_3+s_4=0) & (s_1+s_3+s_5=0)\end{array}$$

All satisfied ⇒ solvable

Solution algorithm Consider again the matrix



Among the 8 possible ways of combining the presses of these three buttons, three combinations stand out:



Each is useful because it only switches one of the three lights to the left. This provides the first way to find a candidate for a solution, without using buttons 4 and 5. This candidate only depends on the state of the three first lights. We call it a candidate rather than a solution, because we have no assurance that the last buttons will turn out the right way. We will consider this below.

For a different way of finding the solution, look again at the effects of the presses of the first three buttons at the beginning of this section. Note that only button 1 can switch exactly one of the first two lights. This tells us that button 1 must be pressed if and only if exactly one of the first two lights is on. For button 2, we have to count how many of the first three lights are on: we press button 2 if and only if this is odd. Likewise, pressing button 3 depends on the parity of the set containing the second and third lights. This approach is also used in [**8**].

This gives us a second way to solve any puzzle, which we state for the nonreduced game: gather the state down to row 5; then, in row 1, press whichever buttons are required by the parity rules given above, based on which lights remain on in the last row; gather the new state down again. If the initial state was solvable, we have solved it in this way. A way to remember this technique is to note that the lights governing the press of a button are the ones in its *mini-neighborhood*, that is, the neighbors in the subset of the three cells on the left.

$$B_5 = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

of the reduced problem.

Following the approach of [**1**] we will look for a solution to our reduced problem such that $p_4 = p_5 = 0$.

Using the Gauss-Jordan reduction of B_5, we have

$$\vec{p} = E\vec{p} = XB_5\vec{p} = X\vec{s}.$$

This allows us to compute one of the solutions easily; the other three are obtained by adding the vectors in the null space of the matrix to $\vec{p}$.

Remembering the matrix

$$X = \left(\begin{array}{ccc|cc} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ \hline 0 & 1 & \multicolumn{1}{c}{1} & 1 & 0 \\ 1 & 0 & \multicolumn{1}{c}{1} & 0 & 1 \end{array}\right),$$

we can express the form of the solutions as follows:

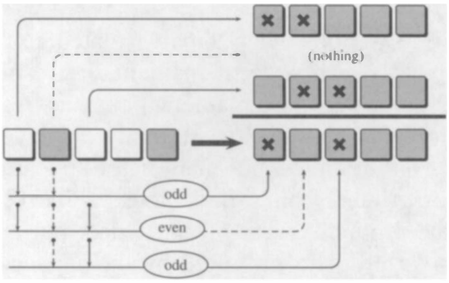
$$\begin{aligned} p_1 &= s_1 + s_2 \\ p_2 &= s_1 + s_2 + s_3 \\ p_3 &= \phantom{s_1 + {}} s_2 + s_3 \end{aligned}$$

In summary, to calculate a solution to any 25-component state $\vec{s}$ in the nonreduced problem, we calculate the first three components of *gather*$(\vec{s})$; then find p_1, p_2, and p_3; then, finally, the solution *expand*$(\vec{p}_{1\bullet}, \vec{s})$.

The matrix X has, embedded in it, a lot of information about the reduced problem. For the sake of clarity, we have divided it into three boxes above. Look first at the lower box. It determines, for a given state $\vec{s}$, the two last components of the presses $\vec{p}$ that solve it. But we have assumed above that $p_4 = p_5 = 0$. This gives another view of the condition for $\vec{s}$ to be solvable, because the two rows in this box are a basis of $null(B_5)$.

The relationship between these two solutions above is interesting: the set of cells that governs whether to press a button v constitutes the candidate solution for cell v. To prove this, let's reason about the symmetry of the game, and the parities of certain sets of cells. Suppose that v is the only button that changes the parity of the number of lights on in a set U. This means that v is the only cell whose neighborhood has an odd number of cells in U. And, therefore, U is *the* candidate solution for v.

This result can be applied to boards of any size. If we work out in advance the solutions (3, in our case) to the appropriate cells, possibly by trial and error, this is enough to solve, with a few further calculations, any other state of the board. The following figure shows the two ways previously described to find a solution: considering lights one by one (upper half), or considering buttons one by one (lower half):



Up to this point, we have not considered lights 4 and 5. We know that the first three lights uniquely determine a candidate solution (that is, something to try). When a solution does exist, all the lights are going to be left off, and our candidate will be an actual solution. If no solution exists, and thus some neutral procedure fails the parity criterion, then either one or both of the last two lights will remain on. The light that remains on denotes which of the neutral procedures fails the parity criterion: light 4 corresponds to the three central buttons, and light 5 to the three alternate ones.

Now, consider the upper right box of X. It is null, which means that the solution we are going to calculate does not depend on the last two components of $\vec{s}$. Thus, the last two components of a state are only useful to check whether the state is solvable or not.

So, in a sense, we have further reduced our original problem to one of dimension 3, whose solution is given by the upper left box of X. The rules of this new problem are given by the 3×3 box in the upper left corner of B_5. This new problem always has a solution; once we know it, we can attempt to solve the others. For our example:

$$\begin{aligned} p_1 &= s_1 + s_2 &&= 1 + 0 &&= 1 \\ p_2 &= s_1 + s_2 + s_3 &&= 1 + 0 + 1 &&= 0 \\ p_3 &= \phantom{s_1 +{}} s_2 + s_3 &&= \phantom{1 +{}} 0 + 1 &&= 1 \end{aligned}$$

But it is only an attempt. We calculate the solution just by looking at the first three components of $gather(\vec{s})$, so the other two can result 0 or 1. They are guaranteed to be null if there is a solution, that is, if the state is orthogonal to the null space of the matrix. When there is no solution, as can be deduced from X, the vector of the basis of the null space for which the orthogonality fails is denoted by the component, fourth or fifth, that remains nonnull.

Acknowledgment. We thank Gustavo Luengo, Pedro Palao, and Cristina Rodríguez for their helpful comments on a previous draft of this article; we are also indebted to the LaTeXmaster Pedro PALAO for his nice collection of macros.

Benjamin Lotto provided us with information about this game, as did Robert A. Moeser and Anya Hillery, who also helped to improve the English of the paper.

The authors of the left column also express their gratitude to these people for their contribution.

REFERENCES

1. M. Anderson and T. Feil, Turning lights out with linear algebra, this MAGAZINE **71** (1998), 300–303.
2. K. Barr, *Lights Out Fan Club*, at http://gbs.mit.edu/~kbarr/lo/
3. R. Barua and S. Ramakrishnan, σ-game, σ^+-game and two-dimensional additive cellular automata, *Theoretical Computer Science* **154**:2 (1996), 349–366.
4. C. Haese, paper in the sci.math newsgroup in 1998, available at http://www.math.niu.edu/~rusin/known-math/98/lights_out
5. B. Lotto, A computer puzzle as a paradigm for doing mathematics, *Vassar Quarterly*, Winter 1996, 18–23.
6. B. Lotto, personal letter, May 1999.
7. Ó. Martín-Sánchez and C. Pareja-Flores, *A Visual Basic Implementation of Lights Out*, available at http://dalila.sip.ucm.es/miembros/cpareja/lo
8. D.L. Stock, Merlin's magic square revisited, *Amer. Math. Monthly* **96** (1989), 608–610.
9. K. Sutner, Linear cellular automata and the garden-of-Eden, *The Mathematical Intelligencer* **11**:2 (1989), 49–53.

An Equilateral Triangle with Sides through the Vertices of an Isosceles Triangle

FUKUZO SUZUKI
Gunma College of Technology
580 Toriba Machi Maebashi
Gunma Japan 371-8530
suzuki@nat. gunma-ct.ac.jp

In the years 1603–1867, known as the Edo period, when Japan isolated itself from the western world, the country developed its own style of mathematics, especially geometry. Results and theorems of traditional Japanese mathematics, known as *Wasan*, were usually stated in the form of problems; these were originally displayed on wooden tablets (*Sangaku*) hung in shrines and temples, but many later appeared in books, either handwritten with a brush or printed from wood blocks. (See [**2**], [**3**], and [**4**] for more details.) Solutions to the problems were not provided, but answers were sometimes given. One of these is a problem proposed by the Japanese mathematician Tumugu Sakuma (1819–1896).

In this note, we take up a generalization of his problem: In FIGURE 1, triangle $\triangle ABC$ is an equilateral triangle and each of the sides CA, AB, BC (or their extensions) passes through three vertices L, M, N of an isosceles triangle $\triangle LMN$ with $ML = MN$. Find a relation among LA, MB, and NC. Our solution reveals an invariant property of this configuration.

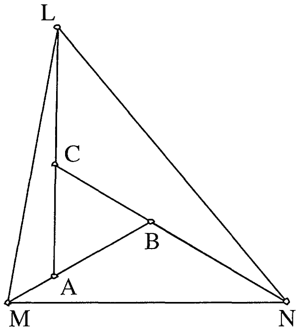


Figure 1

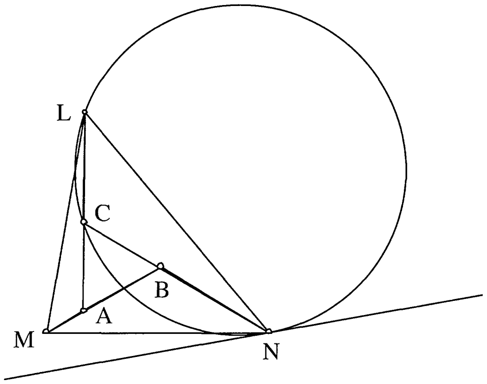


Figure 2

Sakuma's problem specifies L, M, and N as three vertices of a square, as can be found in Hirayama's book [**1**]. However, an incorrect relationship $NC = (\sqrt{3} - 1)(LA + MB + NC)$ appears there, along with some incorrect figures. We will solve the general problem, where $\angle NML$ need not be a right angle; we also take into account the possibility that the equilateral triangle may not fit nicely inside the isosceles one, as drawn. However, for simplicity the figures will be drawn with $\angle NML = \frac{4}{9}\pi$, and our first demonstration of the solution will assume that our triangle appears as in FIGURE 1.

Let α be the angle $\angle NML$ of the isosceles triangle $\triangle LMN$ and θ the angle $\angle LNC$. It may seem unfamiliar to put the summit angle of the isosceles triangle at the side of the diagram, but this matches the historical appearance, with the larger triangle forming the lower left half of a square.

From FIGURE 2, we see that the point C lies on the circumcircle of an equilateral triangle with side LN. (For now we will use the equilateral triangle that lies on the side of LN opposite M, leaving the other case for later.) This suggests how to construct the triangle $\triangle ABC$, given $\triangle LMN$ and a point C on this circle. Sliding the point C up the circle in FIGURE 2 creates the special case, $\theta = \frac{\alpha}{2} - \frac{\pi}{6}$, because then A coincides with M. On the other hand, when C slides down the circle (widening the angle θ), there is one position where A, B, and C all coincide. This is the Fermat point of the triangle $\triangle LMN$ and corresponds to $\theta = \frac{\pi}{6}$. In order to find ourselves in the situation of FIGURE 1, we will require that $\frac{\alpha}{2} - \frac{\pi}{6} < \theta < \frac{\pi}{6}$. Note that this is a nonempty range of values only if $\alpha < \frac{2\pi}{3}$; and to ensure for our first case that θ is a positive angle, we will require $\frac{\pi}{3} < \alpha < \frac{2\pi}{3}$.

In addition to the equilateral triangle, the large triangle also contains triangles $\triangle ALM$, $\triangle BMN$, and $\triangle CNL$. We will be applying the law of sines to these, and therefore will give names that remind us which angles are opposite which sides. We define lengths $a = LA$, $b = MB$, and $c = NC$, and corresponding angles θ_a, θ_b, θ_c opposite them in the triangles mentioned.

In FIGURE 2, observe that θ_c is the angle subtended by the arc NC in the circumcircle of $\triangle LNC$, a fact that will help later when we will need to orient all the angles and arcs to treat the general case.

PROPOSITION 1. *The invariant relationship in this case is*

$$a + b + c = \frac{1}{2}\left(3 + \sqrt{3}\cot\frac{\alpha}{2}\right)c. \tag{1}$$

Proof. Summing the angles in the three triangles mentioned above leads to the formulas

$$\theta_a - \theta_b = \alpha - \frac{\pi}{3}, \quad \theta_c - \theta_a = \frac{\pi}{2} - \frac{\alpha}{2} - \frac{\pi}{3}, \quad \text{and} \quad \theta_b - \theta_c = \frac{\pi}{2} - \frac{\alpha}{2} - \frac{\pi}{3}. \tag{2}$$

Therefore $\theta_c = \frac{1}{2}(\theta_a + \theta_b)$.

We denote the circumradius of the triangle $\triangle BMN$ by r. Then the circumradii of the triangles $\triangle ALM$ and $\triangle CNL$ are found to be r and $2r \sin\frac{\alpha}{2}$, respectively, from applying a well-known formula relating the circumradius of a triangle to an angle and opposite side.

Combining this with the law of sines, we have $c = 4r \sin\frac{\alpha}{2} \sin\theta_c$, $a = 2r \sin\theta_a$, and $b = 2r \sin\theta_b$.

Various trigonometric identities and the formulas in (2) above bring us to

$$a + b = 4r \sin\theta_c \cos\left(\frac{\alpha}{2} - \frac{\pi}{6}\right),$$

which quickly yields equation (1). ■

Triangles shaped differently from the one in FIGURE 1 require more care about the signs of quantities. We use the notation $\angle XYZ$ to denote the oriented angle from $\overrightarrow{YX}$ to $\overrightarrow{YZ}$, which is positive if the direction of rotation from $\overrightarrow{YX}$ to $\overrightarrow{YZ}$ is counterclockwise and negative otherwise. Let the orientation of triangle $\triangle ABC$ be counterclockwise, and assume that it is the same as the orientation of $\triangle LMN$. We generalize our definitions above as follows:

$$\theta_a = \begin{cases} \angle AML & \left(\theta \le \frac{\alpha}{2} - \frac{5\pi}{6}, \frac{\alpha}{2} - \frac{\pi}{6} < \theta\right) \\ \angle BML & \left(\frac{\alpha}{2} - \frac{5\pi}{6} < \theta \le \frac{\alpha}{2} - \frac{\pi}{6}\right) \end{cases},$$

$$\theta_b = \begin{cases} \angle CNM & \left(\frac{\alpha}{2} - \frac{5\pi}{6} < \theta\right) \\ \angle BNM & \left(\theta \le \frac{\alpha}{2} - \frac{5\pi}{6}\right) \end{cases},$$

$$\theta_c = \begin{cases} \angle CLN & (0 \le \theta) \\ \angle ALN & (\theta < 0) \end{cases}.$$

With these definitions θ_a, θ_b, and θ_c are always positive with the following exceptions: $\theta_b < 0$ and $\theta_c < 0$ if $\theta < \frac{\alpha}{2} - \frac{5\pi}{6}$, and $\theta_b < 0$ if $\theta > \frac{\pi}{2} - \frac{\alpha}{2}$.

FIGURE 3 shows the possible cases for $\frac{\pi}{3} < \alpha < \frac{2\pi}{3}$. The situations in (i), (ii), and (vii) were omitted in previous analyses [**1**] and [**5**].

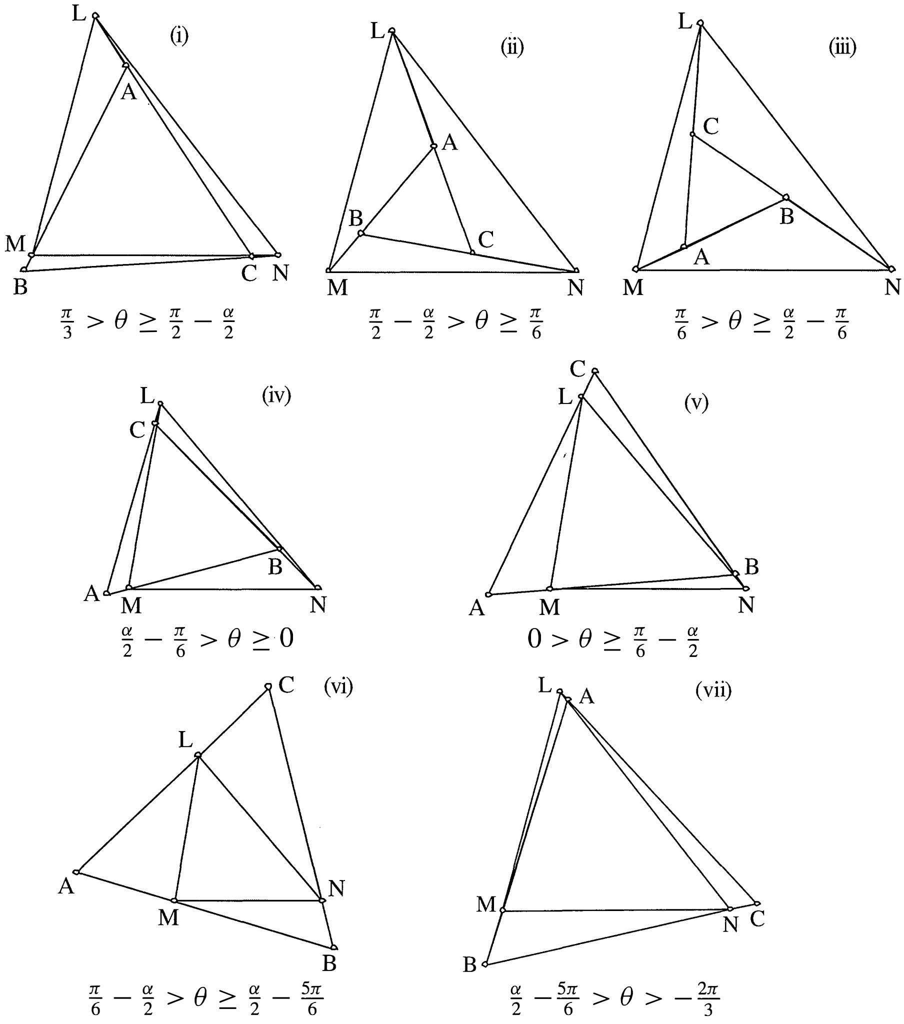
If $\theta = \frac{\pi}{6}$, the triangle $\triangle ABC$ degenerates to a point, but (1) still holds. If $\theta < \frac{\alpha}{2} - \frac{5\pi}{6}$, then $b = -MB$ and $c = -NC$, and if $\theta > \frac{\pi}{2} - \frac{\alpha}{2}$, then $b = -MB$.

These pictures can help the reader see that the definitions of the signs of the various quantities are what we need to make formula (1) hold for every value of θ. Rephrasing the result, we have

$$LA - MB + NC = \frac{1}{2}\left(3 + \sqrt{3}\cot\frac{\alpha}{2}\right) NC \qquad \left(\frac{\pi}{2} - \frac{\alpha}{2} < \theta < \frac{\pi}{3}\right)$$

$$LA + MB + NC = \frac{1}{2}\left(3 + \sqrt{3}\cot\frac{\alpha}{2}\right) NC \qquad \left(\frac{\alpha}{2} - \frac{5\pi}{6} \le \theta \le \frac{\pi}{2} - \frac{\alpha}{2}\right)$$

$$-LA+MB+NC=\frac{1}{2}\left(3+\sqrt{3}\cot\frac{\alpha}{2}\right)NC \qquad \left(-\frac{2\pi}{3}<\theta<\frac{\alpha}{2}-\frac{5\pi}{6}\right).$$



$\frac{\pi}{3} > \theta \geq \frac{\pi}{2} - \frac{\alpha}{2}$

$\frac{\pi}{2} - \frac{\alpha}{2} > \theta \geq \frac{\pi}{6}$

$\frac{\pi}{6} > \theta \geq \frac{\alpha}{2} - \frac{\pi}{6}$

$\frac{\alpha}{2} - \frac{\pi}{6} > \theta \geq 0$

$0 > \theta \geq \frac{\pi}{6} - \frac{\alpha}{2}$

$\frac{\pi}{6} - \frac{\alpha}{2} > \theta \geq \frac{\alpha}{2} - \frac{5\pi}{6}$

$\frac{\alpha}{2} - \frac{5\pi}{6} > \theta > -\frac{2\pi}{3}$

Figure 3 The possible cases

Reversing the inner triangle In the original problem the orientation of $\triangle ABC$ was counterclockwise, the same as the orientation of $\triangle LMN$. Let us consider the case where the orientation of $\triangle ABC$ is clockwise, that is, opposite the orientation of $\triangle LMN$. This will be the case if we constrain C to lie on the circumcircle of the *other* equilateral triangle with side LN (see FIGURE 4). As before, we name $\angle LNC = \theta$; the picture illustrates why $-\frac{\pi}{3} < \theta < \frac{2\pi}{3}$.

In this case we have $\angle A = \angle B = \angle C = -\frac{\pi}{3}$. Appropriate definitions for the angles θ_a, θ_b, and θ_c are as follows:

$$\theta_a = \begin{cases} \angle AML - \pi & \left(\frac{\pi}{6}+\frac{\alpha}{2}<\theta\right) \\ \angle AML & \left(\theta \leq \frac{\pi}{6}+\frac{\alpha}{2}\right) \end{cases},$$

$$\theta_b = \begin{cases} \angle BNM - \pi & \left(\frac{5\pi}{6} - \frac{\alpha}{2} < \theta\right) \\ \angle BNM & \left(\theta \le \frac{5\pi}{6} - \frac{\alpha}{2}\right) \end{cases},$$

$$\theta_c = \begin{cases} \angle CLN & (0 < \theta) \\ \pi - \angle NLC & (\theta \le 0). \end{cases}$$

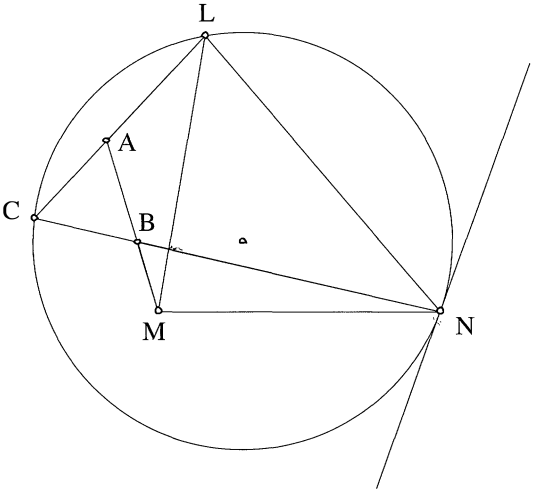


Figure 4

Also, let $a = sign(\theta_a)LA$, $b = sign(\theta_b)MB$, $c = sign(\theta_c)NC$. Then we have the following proposition. The proof is similar to that of Proposition 1, so we leave it to interested readers.

PROPOSITION 2. *With the quantities defined as above, the invariant relationship is*

$$-a + b + c = \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right)c. \tag{3}$$

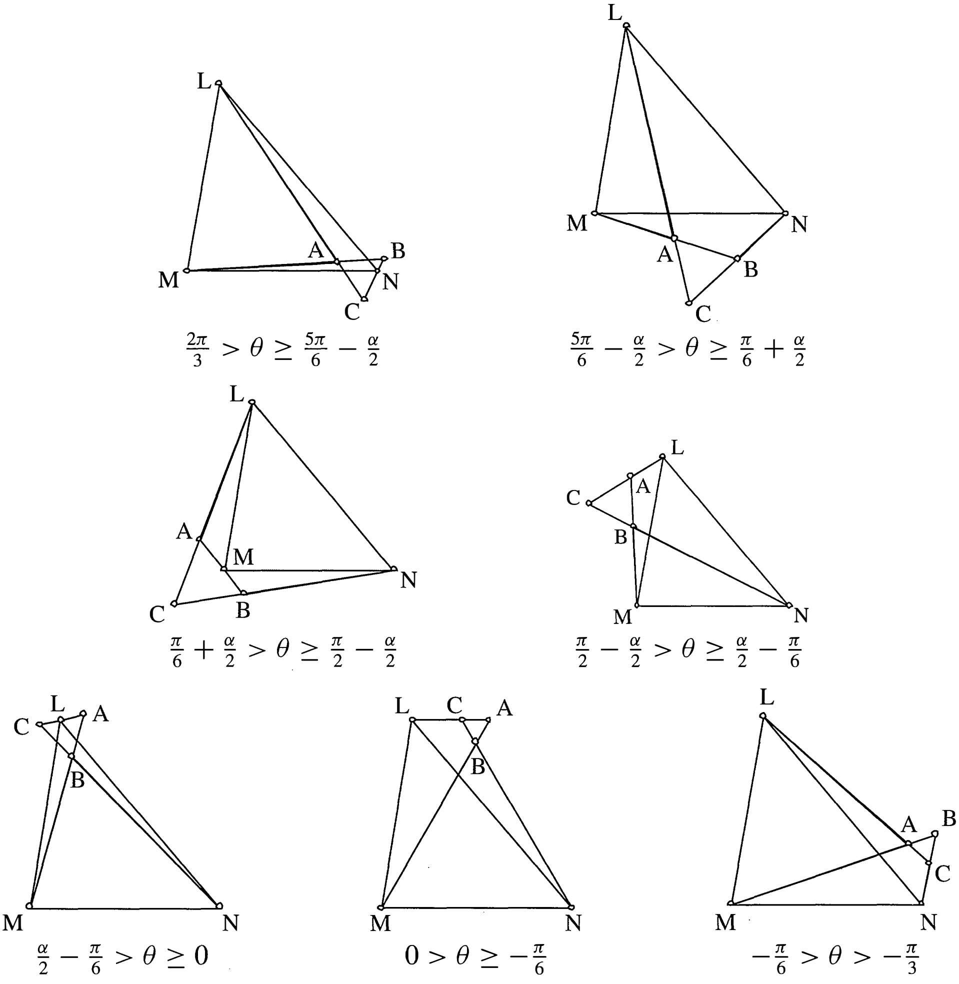
FIGURE 5 shows the various ranges of θ. They are drawn with $\alpha = \frac{4\pi}{9}$, but represent the general case when $\frac{\pi}{3} < \alpha < \frac{2\pi}{3}$. As in the previous proposition, the other possible ranges for α can be handled similarly. When $\theta = -\frac{\pi}{6}$, the triangle $\triangle ABC$ degenerates to a point, but (3) still holds. Furthermore $a = -LA$ if $\theta > \frac{\alpha}{2} - \frac{\pi}{6}$, and $b = -MB$ if $\theta > \frac{\pi}{2} - \frac{\alpha}{2}$. Also $-\frac{\pi}{6} + \frac{\alpha}{2} \le \frac{\pi}{2} - \frac{\alpha}{2}$ if $\alpha \le \frac{2\pi}{3}$ and $\frac{\pi}{2} - \frac{\alpha}{2} < -\frac{\pi}{6} + \frac{\alpha}{2}$ if $\frac{2\pi}{3} < \alpha < \pi$.

Thus if $\alpha \le \frac{2\pi}{3}$, we have

$$\begin{aligned} LA - MB + NC &= \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right)NC & &\left(\frac{\pi}{2} - \frac{\alpha}{2} < \theta < \frac{2\pi}{3}\right) \\ LA + MB + NC &= \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right)NC & &\left(-\frac{\pi}{6} + \frac{\alpha}{2} \le \theta \le \frac{\pi}{2} - \frac{\alpha}{2}\right) \\ -LA + MB + NC &= \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right)NC & &\left(-\frac{\pi}{3} < \theta < -\frac{\pi}{6} + \frac{\alpha}{2}\right), \end{aligned}$$

and if $\frac{2\pi}{3} < \alpha$, we have

$$\begin{aligned}
LA - MB + NC &= \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right) NC & &\left(-\frac{\pi}{6} + \frac{\alpha}{2} < \theta < \frac{2\pi}{3}\right)\\
-LA - MB + NC &= \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right) NC & &\left(\frac{\pi}{2} - \frac{\alpha}{2} \le \theta \le -\frac{\pi}{6} + \frac{\alpha}{2}\right)\\
-LA + MB + NC &= \frac{1}{2}\left(1 + \sqrt{3}\cot\frac{\alpha}{2}\right) NC & &\left(-\frac{\pi}{3} < \theta < \frac{\pi}{2} - \frac{\alpha}{2}\right).
\end{aligned}$$



$\frac{2\pi}{3} > \theta \ge \frac{5\pi}{6} - \frac{\alpha}{2}$

$\frac{5\pi}{6} - \frac{\alpha}{2} > \theta \ge \frac{\pi}{6} + \frac{\alpha}{2}$

$\frac{\pi}{6} + \frac{\alpha}{2} > \theta \ge \frac{\pi}{2} - \frac{\alpha}{2}$

$\frac{\pi}{2} - \frac{\alpha}{2} > \theta \ge \frac{\alpha}{2} - \frac{\pi}{6}$

$\frac{\alpha}{2} - \frac{\pi}{6} > \theta \ge 0$

$0 > \theta \ge -\frac{\pi}{6}$

$-\frac{\pi}{6} > \theta > -\frac{\pi}{3}$

Figure 5

As an exercise for the reader, we offer a similar problem. Suppose, as in FIGURE 6, that $LMNO$ is a kite-shaped quadrilateral with $OL = ON$, $ML = MN$, $\angle O = \frac{\pi}{3}$, and $\angle L = \angle N = \frac{\pi}{2}$. Let $ABCD$ be a parallelogram with $\angle A = \angle C = \frac{2\pi}{3}$ whose sides (possibly extended) pass through the vertices of the kite. Let a, b, c, and d be the (signed) lengths OA, LB, MC, and ND. In this situation, find a relationship among a, b, c, and d. You will find that a solution requires consideration of three cases for the angle $\angle NOA$, where the critical values are $\frac{2\pi}{3}$ and $\frac{5\pi}{6}$.

The answer appears on page 331.

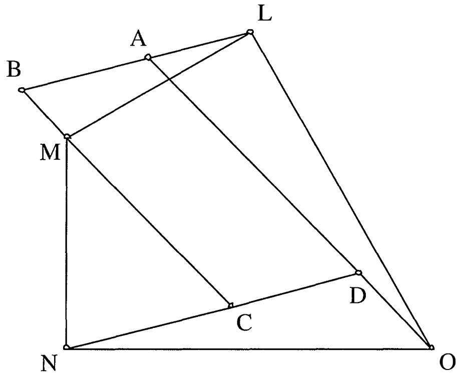


Figure 6 The kite problem

Acknowledgment. The author would like to thank the referees for their valuable suggestions in preparing this paper.

REFERENCES

1. A. Hirayama, *People of Wasan on record*, Fuji Junior College Press, Tokyo, 1965 (in Japanese).
2. H. Fukagawa and D. Pedoe, *Japanese Temple Geometry Problems*, The Charles Babbage Reseach Center, Winnipeg, 1989.
3. H. Fukagawa and J. F. Rigby, *Traditional Japanese Mathematics Problems of 18th–19th Century*, SCT Publishing, Singapore, 2000.
4. T. Rothman, Japanese temple geometry, *Scientific American*, **278** (1998), No. 5, 84–91.
5. T. Sakuma, Tōyō sampō, wood-block printed, 1853.

Cramer's Rule Is Due To Cramer

A. A. KOSINSKI
Rutgers University
New Brunswick, NJ 08903

Most freshmen who took the first calculus course know of Marquis de l'Hôpital as the man who did not invent l'Hôpital's rule. There is a certain core of truth to this assertion, but it is nevertheless somewhat unfair to the man, who was a productive mathematician in his time, highly respected by the Bernoullis and Leibniz, and an author of two excellent textbooks on calculus and analytic geometry.

It seems that the same fate now threatens Gabriel Cramer, who is in danger of becoming the man who did not invent Cramer's rule. Some authors[1] credit this invention to Colin Maclaurin on the basis of his *Treatise of Algebra*, edited from various manuscripts left at his death and published posthumously in 1748. (Recently B. Hedman [**4**] found among the unpublished papers of Maclaurin a manuscript of Part I of the *Treatise* dated 1729 and almost identical to the 1748 edition.) The *Treatise* was intended both as a textbook and as a sort of supplement to Newton's *Arithmetic*, pro-

[1] This claim has its origin in a note by C. B. Boyer [**1**]. M. Kline [**5**] asserts that "the solution of linear equations in two, three, and four unknowns by the method of determinants was created by Maclaurin."

viding proofs to various assertions that Newton did not bother to prove.[2] That it did, and much else besides. But one thing that it did not do was to provide Cramer's rule.

Let us review the evidence.

What is known as Cramer's rule is a formula expressing solutions of a system of n linear equations with n unknowns as a ratio of two quantities, each of which is a sum of products of certain coefficients provided with appropriate signs. The rule for forming the products is not difficult to state, especially for the denominator of the ratio, which is the same for all unknowns. The rule for signs, the heart of the matter, is almost impossible to state without an appropriate indexing of unknowns. Such indexing was already introduced by Leibniz in a 1693 letter to l'Hôpital and in 1700 in *Acta Eruditorum*, but it seems that nobody noticed.

Maclaurin certainly did not notice. In his *Algebra* (pp. 82–85) he solves, first, a system of two equations with two unknowns, then a system of three equations. In both cases coefficients are given by unindexed small letters a, b, c, ... Both solutions are arrived at by elimination. The solution of three equations is followed by a discussion of a rule for forming denominators and numerators. With a proper interpretation of his slightly confusing definition of "opposite" coefficients, the rule becomes correct—but for the lack of the convention concerning signs. This is followed by a breezy assurance that systems of four equations can be solved "much after the same manner by taking all the products that can be made of four opposite coefficients and always prefixing contrary signs to those that involve the products of two opposite coefficients." Since Maclaurin calls two coefficients opposite if they are attached to distinct unknowns in distinct equations, *every* product involves "the products of two opposite coefficients," and this "rule" makes no sense whatever. The most charitable explanation is that it is an attempt to describe what happens in the case of a system of two equations when two products are provided with "contrary signs." However, without stating which signs have to be affixed to which monomials, the "rule" is not adequate even in this case. In the general case, it indicates that Maclaurin did not know the correct rule for signs.

No more is said about linear equations in the *Treatise*; in particular, no notice is taken of the possibility that the denominator may vanish, rendering the formulas meaningless.

It would be incorrect to attach much blame to Maclaurin for this muddle. In the middle of the 18th century the solution by elimination of systems of linear equations did not present a problem to which a mathematician of Maclaurin's class would attach much attention. He was writing a textbook, and, in a hurry to get to some really interesting stuff, he certainly missed an opportunity to discover Cramer's rule.[3]

That much about Maclaurin. We now go to Gabriel Cramer, a Swiss mathematician known for his excellent editions of the works of James Bernoulli (2 vols., Geneva, 1744), and of John Bernoulli (4 vols., Geneva, 1742). However, his best known work is a hefty volume of 680 pages in-quarto, entitled *Introduction à l'Analyse des Lignes Courbes Algébriques* and published in Geneva in 1750. It is a well-organized and well-written book that contains most of what was known at the time about algebraic geometry, as well as Cramer's original contributions to the subject. In the appendix to this work, pp. 657–659 are devoted to a concise exposition of the theory of systems of

[2] Providing commentaries to the *Arithmetic* appears to have been a popular occupation in the eighteenth century. The first edition of the *Arithmetic* had a supplement by Halley. It was removed by Newton from the second (Latin) edition prepared by him in 1722, but reappeared, together with 7 other commentaries, in the edition published by s'Gravesande in 1732. The next edition commented by Castillione grew to two volumes in-quarto and the number of commentaries to 9.

[3] Already in 1901, M. Cantor [**2**] noted that, for lack of good notation, Maclaurin missed the general rule for solving linear equations. He also reproduced Cramer's solution [**2**, p. 607].

linear equations. The presentation is very clear. It may be summarized, with inessential modifications and retaining Cramer's notation, as follows.

We consider the system of n equations for n unknowns z, x, y, v, *&c.*:

$$\begin{aligned} A_1 &= Z_1 z + Y_1 y + X_1 x + V_1 v + \&c. \\ A_2 &= Z_2 z + Y_2 y + X_2 x + V_2 v + \&c. \\ A_3 &= Z_3 z + Y_3 y + X_3 x + V_3 v + \&c. \\ &\quad\&c. \end{aligned}$$

It is agreed that the letter Z always denotes the coefficient of the first unknown, the letter Y that of the second unknown, and so on.[4]

To find the solution, first form all expressions that can be obtained from the product $Z \quad Y \quad X \quad V \ldots$ (always in this order) by distributing as lower indices all permutations of numbers $1, 2, \ldots, n$. (For example, with $n = 2$ one obtains two terms: $Z_1 Y_2$ and $Z_2 Y_1$.) Now, count the number of transpositions (*dérangements*) in the permutation attached to a given term. If it is odd, then the term is provided with the minus sign, otherwise with the plus sign. The solution of the system is given by fractions which have as the denominator the sum of terms just obtained, and as a numerator the sum of terms formed, for the unknown z, by replacing the letter Z by A, for the unknown y, the letter Y by A, and so on.

This, of course, is Cramer's rule. Cramer did not provide a proof. However, he did consider what happens if the denominator vanishes (that is, in today's terminology, if the rank of the matrix of coefficients is less than n). He split this case into two according to whether the rank of the augmented matrix equals n or is less than n, and showed that in the first case the system will have no solution. The second case was called "indeterminate" and left at that.

Thus, Cramer's rule is due to Cramer. In fact, more is due to him. The procedure given above for attaching a number to a square array of numbers, of arbitrary size, is effectively the first definition of a determinant. Of course, any formula for a solution of a system of 2 equations with two unknowns and literal coefficients, whether arrived at by elimination or by any other method, will express this solution as a ratio of two determinants. Also, it is known that determinants of 2 × 2, 3 × 3, and 4 × 4 arrays were calculated earlier and in a different context by the Japanese mathematician Seki Takakazu, (see a note by Victor Katz in Fraleigh and Beauregard's *Linear Algebra* [**3**, p. 251]). But the first unambiguous, general definition of determinants of arbitrary size is due to Cramer. A formal recognition of this fact and the name of the object defined is missing in his book. This was provided later and is another story.

REFERENCES

1. C. B. Boyer, Colin Maclaurin and Cramer's rule, *Scripta Mathematica* **27** (1996), 377–399.
2. M. Cantor, *Geschichte der Mathematik*, vol. 3, Leipzig, 1901, p. 590.
3. Fraleigh and Beauregard, *Linear Algebra*, 3rd. ed., Addison-Wesley, Reading, MA, 1994.
4. B. Hedman, An earlier date for Cramer's rule, *Historia Mathematica* **26**, 365–368.
5. M. Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Press, New York, 1972, p. 606.

[4] This system of indexing is different from Leibniz's. However, as the editor of works of both Bernoullis, Cramer was certainly well acquainted with everything that had been published in *Acta Eruditorum.*

Dialog with Computer in the Proof of the Four-Color Theorem

J. D. MEMORY
North Carolina State University
Raleigh, NC

You've searched all cases, sage Machine,
List the largest set you've seen.
Yellow, red, blue, green.

O Computer, is it true?
Four Magic Markers will make do,
Green, yellow, red, blue?

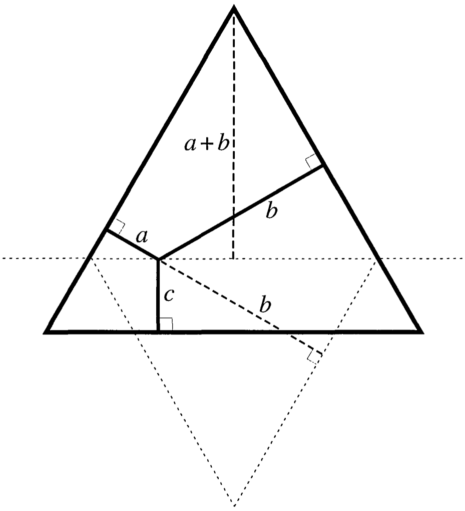
My terminal output sits there spread;
I'll say once more what I just said:
Blue, green, yellow, red.

So work your wetware till it's Jell-O,
Four's still enough; it's finished, Fellow:
Red, green, blue, yellow.

Proof Without Words: Equilateral Triangle

JAMES TANTON
The Math Circle
Boston, MA

For an equilateral triangle, the sum of the distances from any interior point to the three sides equals the height of the triangle.



Proof Without Words: How Did Archimedes Sum Squares in the Sand?

KATHERINE KANIM
New Mexico State University
Las Cruces, NM 88003

Archimedes deduced a formula for a sum of squares, used in his determination of the volume of a conoid (*Conoids and Spheroids*, Prop. 25) and areas of spirals (*Spirals*, Prop. 10 and Prop. 24). The modern transcriptions of his proof (Dijksterhuis, Heath, Heiberg) are completely algebraic and hard to follow. The geometry of Archimedes' proof is depicted visually below. Each step of his written proof is transparent in the geometry of the picture. One wonders if this is the picture Archimedes drew in the sand.

Spirals, Proposition 10 (Dijksterhuis, page 122)

> If a series of any number of lines be given, which exceed one another by an equal amount, and the difference be equal to the least, and if other lines be given equal in number to these and in quantity to the greatest, the squares on the lines equal to the greatest, plus the square on the greatest and the rectangle contained by the least and the sum of all those exceeding one another by an equal amount will be the triplicate of all the squares on the lines exceeding one another by an equal amount.

$$(n+1)n^2 + \sum_{i=1}^{n} i = 3\sum_{i=1}^{n} i^2$$

Archimedes also deduced corollary inequalities, which he used for his area and volume proofs by the method of exhaustion. Can you see them in the pictures?

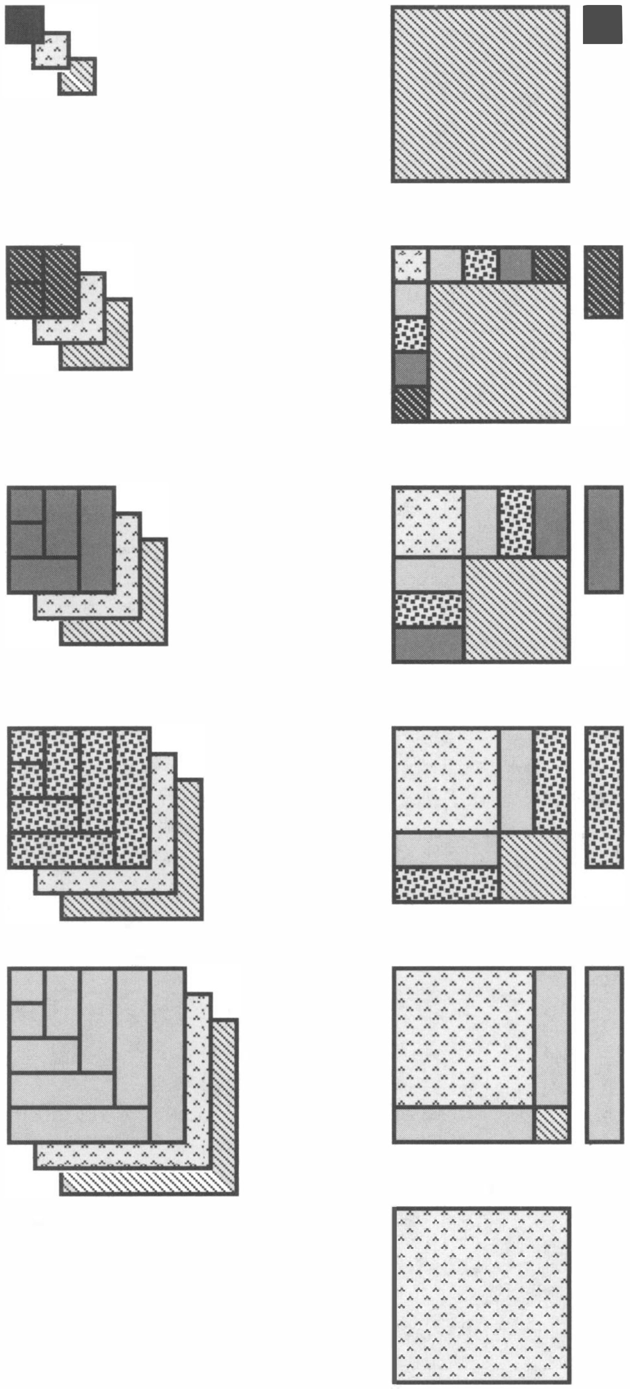
$$n \cdot n^2 < 3\sum_{i=1}^{n} i^2$$

$$n \cdot n^2 > 3\sum_{i=1}^{n-1} i^2$$

Acknowledgment. Thank you to David Pengelley for suggestions and support, to Steve Kanim for the electronic version of the figure, and to the referee for ideas about the layout of the figure.

REFERENCES

1. E. J. Dijksterhuis, *Archimedes*, Princeton University Press, Princeton, NJ, 1987.
2. Thomas L. Heath, *The Works of Archimedes*, in Great Books of the Western World, vol. 11, Encyclopaedia Britannica, Inc., Chicago, IL, 1952.
3. J. L. Heiberg, *Archimedes Opera*, vol. 2, Teubner, Leipzig, 1880–81.



Two Irrational Numbers From the Last Nonzero Digits of $n!$ and n^n

GREGORY P. DRESDEN
Washington & Lee University
Lexington, VA 24450

We begin by looking at the pattern formed from the last (that is, units) digit in the base 10 expansion of n^n. Since $1^1 = 1$, $2^2 = 4$, $3^3 = 27$, $4^4 = 256$, and so on, we can easily calculate the first few numbers in our pattern to be 1, 4, 7, 6, 5, 6, 3, 6 .... We construct a decimal number $N = 0.d_1d_2d_3\ldots d_n\ldots$ such that the n^{th} digit d_n of N is the last (i.e. unit) digit of n^n; that is, $N = 0.14765636\ldots$. In a recent paper [**1**], R. Euler and J. Sadek showed that this N is a rational number with a period of twenty digits:

$$N = 0.\overline{14765636901636567490}.$$

This is a nice result, and we might well wonder if it can be extended. Indeed, Euler and Sadek [**1**] recommend looking at the last *nonzero* digit of $n!$ (If we just looked at the last digit of $n!$, we would get a very dull pattern of all 0s, as $n!$ ends in 0 for every $n \geq 5$.)

With this is mind, let's define $\text{lnzd}(A)$ to be the last nonzero digit of the positive integer A; it is easy to see that $\text{lnzd}(A) \equiv A/10^i \bmod 10$, where 10^i is the largest power of 10 that divides A. We wish to investigate not only the pattern formed by $\text{lnzd}(n!)$, but also the pattern formed by $\text{lnzd}(n^n)$. In accordance with Euler and Sadek [**1**], we define the *factorial number*, $F = 0.d_1d_2d_3\ldots d_n\ldots$ to be the infinite decimal such that each digit $d_n = \text{lnzd}(n!)$; similarly, we define the *power number*, $P = 0.d_1d_2d_3\ldots d_n\ldots$ by $d_n = \text{lnzd}(n^n)$. We ask whether these numbers are rational or irrational.

Although the title of this article gives away the secret, we'd like to point out that at first glance, our factorial number F exhibits a suprisingly high degree of regularity, and a fascinating pattern occurs. The first few digits of F are easy to calculate:

$$\begin{array}{lll}
1! = \underline{1} & 5! = 1\underline{2}0 & 10! = 3628\underline{8}00 \\
2! = \underline{2} & 6! = 7\underline{2}0 & 11! = 39916\underline{8}00 \\
3! = \underline{6} & 7! = 50\underline{4}0 & 12! = 479001\underline{6}00 \\
4! = 2\underline{4} & 8! = 403\underline{2}0 & 13! = 6227020\underline{8}00 \\
 & 9! = 3628\underline{8}0\ldots & 14! = 87178291\underline{2}00
\end{array}$$

Reading the underlined digits, we have

$$F = 0.1264\ 22428\ 88682\ldots.$$

Continuing along this path, we have (to forty-nine decimal places)

$$F = 0.1264\ 22428\ 88682\ 88682\ 44846\ 44846\ 88682\ 22428\ 22428\ 66264\ldots.$$

It is not hard to show that (after the first four digits) F breaks up into five-digit blocks of the form $x\ x\ 2x\ x\ 4x$, where $x \in \{2, 4, 6, 8\}$, and the $2x$ and $4x$ are taken mod 10. Furthermore, if we represent these five-digit blocks by symbols ($\dot{2}$ for 22428, $\dot{4}$ for

44846, $\dot{6}$ for 66264, $\dot{8}$ for 88682, and $\dot{1}$ for the initial four-digit block of 1264), we have

$$F = 0.\dot{1} \quad \dot{2} \quad \dot{8} \quad \dot{8} \quad \dot{4} \quad \dot{4} \quad \dot{8} \quad \dot{2} \quad \dot{2} \quad \dot{6} \quad \ldots.$$

Grouping these symbols into blocks of five and then performing more calculations (with the aid of *Maple*) give us F to 249 decimal places:

$$F = 0.\dot{1}\dot{2}\dot{8}\dot{8}\dot{4}\ \dot{4}\dot{8}\dot{2}\dot{2}\dot{6}\ \dot{2}\dot{4}\dot{6}\dot{6}\dot{8}\ \dot{4}\dot{8}\dot{2}\dot{2}\dot{6}\ \dot{4}\dot{8}\dot{2}\dot{2}\dot{6}\ \dot{8}\dot{6}\dot{4}\dot{4}\dot{2}\ \dot{2}\dot{4}\dot{6}\dot{6}\dot{8}\ \dot{6}\dot{2}\dot{8}\dot{8}\dot{4}\ \dot{2}\dot{4}\dot{6}\dot{6}\dot{8}\ \dot{2}\dot{4}\dot{6}\dot{6}\dot{8}\ \ldots.$$

The reader will notice additional patterns in these blocks of five symbols (twenty-five digits). In fact, such patterns exist for any block of size 5^i. However, a pattern is different from a period, and doesn't imply that our decimal F is rational. Consider the classic example of 0.1 01 001 0001 00001 000001 ..., which has an obvious pattern but is obviously irrational. It turns out that our decimal F is also irrational, as the following theorem indicates:

THEOREM 1. *Let* $F = 0.d_1d_2d_3\ldots d_n\ldots$ *be the infinite decimal such that each digit* $d_n = \text{lnzd}(n!)$. *Then, F is irrational.*

We will prove this, but first note that our power number, P, might also seem to be rational at first glance. P is only slightly different from Euler and Sadek's rational number N, as seen here:

$$N = 0.14765\ 6369\underline{0}\ 16365\ 6749\underline{0}\ 14765\ 6369\underline{0}\ 16365\ 6749\underline{0}\ldots$$

$$\text{and}\quad P = 0.14765\ 6369\underline{1}\ 16365\ 6749\underline{6}\ 14765\ 6369\underline{9}\ 16365\ 6749\underline{6}\ldots$$

(Again, calculations were performed by *Maple*.) Despite this striking similarity between P and N, it turns out that P, like F, is irrational:

THEOREM 2. *Let* $P = 0.d_1d_2d_3\ldots d_n\ldots$ *be the infinite decimal such that each digit* $d_n = \text{lnzd}(n^n)$. *Then, P is irrational.*

Before we begin with the (slightly technical) proofs, let us pause to get a feel for why these two numbers must be irrational. There is no doubt that both F and P are highly regular, in that both exhibit a lot of repetition. The problem is that there are *too many* patterns in the digits, acting on different scales. Taking P, for example, we note that there is an obvious pattern (as shown by Euler and Sadek in [**1**]) repeating every 20 digits with $1^1, 2^2, 3^3, \ldots, 9^9$ and $11^{11}, 12^{12}, \ldots, 19^{19}$, but this is broken by a similar pattern for $10^{10}, 20^{20}, \ldots, 90^{90}$ and $110^{110} \ldots 190^{190}$, which repeats every 200 digits. This, in turn, is broken by another pattern repeating every 2000, and so on. A similar behaviour is found for F, but in blocks of 5, 25, 125, and so on, as mentioned above. So, in vague terms, there are always new patterns starting up in the digits of P and of F, and this is what makes them irrational.

Are there some simple observations that we can make about P and F to help us to prove our theorems? To start with, we might notice that every digit of F (except for the first one) is even. Can we prove this? Yes, and without much difficulty:

LEMMA 1. *For* $n \geq 2$, *then* lnzd(n!) *is in* {2, 4, 6, 8}.

Proof. The lemma is certainly true for $n = 2, 3, 4$. For $n \geq 5$, we note that the prime factorization of $n!$ contains more 2s than 5s, and thus even after taking out all the 10s in $n!$, the quotient will still be even. To be precise, the number of 5s in $n!$ (and thus the number of trailing zeros in its base-10 representation) is $e_5 = \sum_{i=1}^{\infty} \left[n/5^i\right]$, which is strictly less than the number of 2s, $e_2 = \sum_{i=1}^{\infty} \left[n/2^i\right]$ (here, $[\cdot]$ represents the greatest integer function). Hence, $n!/10^{e_5}$ is an even integer not divisible by 10, and so $\text{lnzd}(n!) \equiv n!/10^{e_5} \bmod 10$, which must be in $\{2, 4, 6, 8\}$. ■

Another helpful observation is that the lnzd function is at least sometimes multiplicative. For example,

$$\begin{aligned} &\text{lnzd}(12) \cdot \text{lnzd}(53) = 2 \cdot 3 = 6, \\ \text{and}\quad &\text{lnzd}(12 \cdot 53) = \text{lnzd}(636) = 6. \end{aligned}$$

However, we note that at times this would-be rule fails:

$$\begin{aligned} &\text{lnzd}(15) \cdot \text{lnzd}(22) = 5 \cdot 2 = 10, \\ \text{yet}\quad &\text{lnzd}(15 \cdot 22) = \text{lnzd}(330) = 3. \end{aligned}$$

So, we can only prove a limited form of multiplicativity, but it is useful none the less:

LEMMA 2. *Suppose a, b are integers with* $\text{lnzd}(a) \neq 5$, $\text{lnzd}(b) \neq 5$. *Then,* lnzd *is multiplicative; that is,* $\text{lnzd}(a \cdot b) \equiv \text{lnzd}(a) \cdot \text{lnzd}(b)$ *mod* 10.

Proof. Let x' denote the integer x without its trailing zeros; that is, $x' = x/10^i$, where 10^i is the largest power of 10 dividing x. (Note that $\text{lnzd}(x) \equiv x' \bmod 10$.) By hypothesis, a' and b' are both $\neq 0 \bmod 5$, and so $(a \cdot b)' \neq 0 \bmod 5$ and so $(a \cdot b)' = a' \cdot b'$. Thus,

$$\begin{aligned} \text{lnzd}(a \cdot b) &= \text{lnzd}((a \cdot b)') = \text{lnzd}(a' \cdot b') \equiv a' \cdot b' \bmod 10, \text{ while} \\ \text{lnzd}(a) \cdot \text{lnzd}(b) &= \text{lnzd}(a') \cdot \text{lnzd}(b') = (a' \bmod 10) \cdot (b' \bmod 10). \end{aligned}$$

The two are clearly congruent mod 10. ■

We are now ready to prove Theorem 1, to show that F is irrational. The proof is a little technical; it proceeds by assuming that F has a repeating decimal expansion with period λ_0, then choosing an appropriate multiple of λ_0 and an appropriate digit d, in order to arrive at a contradiction.

Proof of Theorem 1: Suppose F is rational, and thus eventually periodic. Let λ_0 be the period, so that for every n sufficiently large, then $d_n = d_{n+\lambda_0}$. Write $\lambda_0 = 5^i \cdot K$ such that $5 \nmid K$ (we acknowledge that K could be 1) and let $\lambda = 2^i \cdot \lambda_0 = 10^i \cdot K$. Then, $\text{lnzd}(\lambda) = \text{lnzd}(K)$, and since $5 \nmid K$, then $10 \nmid K$ and so $\text{lnzd}(K) \equiv K \bmod 10$. Note also that $\text{lnzd}(2\lambda) \equiv 2K \bmod 10$. Choose M sufficiently large so that both of the following are true: $\text{lnzd}(10^M + \lambda) = \text{lnzd}(\lambda)$ (this can easily be done by demanding that $10^M > \lambda$), and $d_n = d_{n+\lambda_0}$ for all $n \geq M$. Finally, let $d = \text{lnzd}((10^M - 1)!)$; since $10^M! = (10^M - 1)! \cdot 10^M$, then d also equals $\text{lnzd}(10^M!)$.

Since λ is a multiple of the period λ_0, if we let $A = 10^M - 1 + \lambda$ and $B = 10^M - 1 + 2\lambda$, then

$$\begin{aligned} &d = \text{lnzd}((10^M - 1)!) = \text{lnzd}(A!) = \text{lnzd}(B!) \\ \text{and}\quad &d = \text{lnzd}(10^M!) = \text{lnzd}((A + 1)!) = \text{lnzd}((B + 1)!). \end{aligned}$$

We will find our contradiction in the last two terms in the above equation. By Lemma 1, $d \in \{2, 4, 6, 8\}$, and so $\text{lnzd}(A!) \neq 5$. Also, since $\text{lnzd}(A + 1) = \text{lnzd}(10^M + \lambda) = \text{lnzd}(\lambda) \equiv K \bmod 10$, we know that $\text{lnzd}(A + 1) \neq 5$. Thus, we can apply Lemma 2 to $\text{lnzd}(A! \cdot (A + 1))$ to get

$$d = \text{lnzd}((A + 1)!) = \text{lnzd}(A!) \cdot \text{lnzd}(A + 1) \equiv d \cdot K \bmod 10.$$

Likewise, working with B, we find

$$d = \operatorname{lnzd}((B+1)!) = \operatorname{lnzd}(B!) \cdot \operatorname{lnzd}(B+1) \equiv d \cdot 2K \bmod 10.$$

Combining these two equations, we get $d(1-K) \equiv d(1-2K) \equiv 0 \bmod 10$. Since $5 \nmid d$, this implies that $5|(1-K)$ and $5|(1-2K)$, which is a contradiction. Thus, there can be no period λ_0 and so F is irrational. ■

We now turn our attention to the power number P derived from the last nonzero digits of n^n. This part was more difficult, but a major step was the discovery that the sequence $\operatorname{lnzd}(100^{100}), \operatorname{lnzd}(200^{200}), \operatorname{lnzd}(300^{300}) \ldots$ was the same as the sequence $\operatorname{lnzd}(100^4), \operatorname{lnzd}(200^4), \operatorname{lnzd}(300^4) \ldots$. This relies not only on the fact that $4|100$ but also on the easily proved fact that $a^b \equiv a^{b+4} \bmod 10$ for $b > 0$, used in the following lemma:

LEMMA 3. *Suppose* $100 \mid x$. *Then,* $\operatorname{lnzd}(x^x) \equiv (\operatorname{lnzd} x)^4 \bmod 10$.

Proof. As in Lemma 2, let x' denote the integer x without its trailing zeros; that is, $x' = x/10^i$, where 10^i is the largest power of 10 dividing x. Now,

$$\begin{aligned}\operatorname{lnzd}(x^x) &= \operatorname{lnzd}((10^i x')^{10^i x'}) \\ &= \operatorname{lnzd}((10^{i \cdot 10^i x'})(x')^{10^i \cdot x'}) \\ &= \operatorname{lnzd}((x')^{10^i \cdot x'}).\end{aligned}$$

Since $10 \nmid x'$, then $10 \nmid (x')^{10^i \cdot x'}$, and so

$$\operatorname{lnzd}(x^x) \equiv (x')^{10^i \cdot x'} \bmod 10.$$

Since $100 \mid x$, then $4 \mid 10^i \cdot x'$, and since $(x')^n \equiv (x')^{n+4} \bmod 10$ for every positive n, we can repeatedly reduce the exponent of x' by 4 until we have

$$\begin{aligned}\operatorname{lnzd}(x^x) &\equiv (x')^4 \bmod 10 \\ &\equiv (\operatorname{lnzd} x)^4 \bmod 10.\end{aligned}$$ ■

With Lemma 3 at our disposal, the proof of Theorem 2 is now fairly easy.

Proof of Theorem 2: Again, we argue by contradiction. Suppose P is rational. Let λ_0 be the eventual period, and choose j sufficiently large such that $10^j > 200 \cdot \lambda_0$ and such that

$$\operatorname{lnzd}((10^j + n\lambda_0)^{10^j + n\lambda_0}) = \operatorname{lnzd}((10^j)^{10^j})$$

for every positive n. Choosing $n = 200$, we get

$$\operatorname{lnzd}((10^j + 200\lambda_0)^{10^j + 200\lambda_0}) = \operatorname{lnzd}((10^j)^{10^j}).$$

We reduce the left side of the above equation by Lemma 3 (note that $\operatorname{lnzd}(10^j + 200\lambda_0) = \operatorname{lnzd}(2\lambda_0)$), and the right side is obviously 1, so we have

$$(\operatorname{lnzd} 2\lambda_0)^4 \equiv 1 \bmod 10$$

Note that $\operatorname{lnzd}(2\lambda_0)$ can only be 2, 4, 6, or 8, and raising these to the fourth power mod 10 gives us the contradiction $6 = 1$. Thus, P is irrational. ■

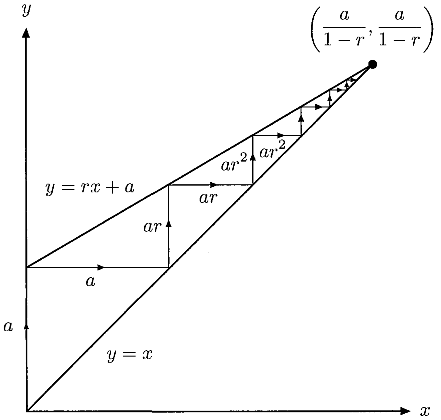
The obvious next question is far more difficult: Are F and P algebraic or transcendental? I suspect the latter, but it is only a hunch. Perhaps some curious reader will continue along this interesting line of study.

REFERENCES

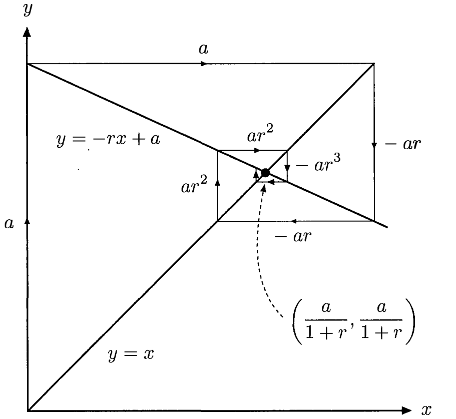
1. R. Euler and J. Sadek, A number that gives the unit digit of n^n, *Journal of Recreational Mathematics*, **29** (1998) No. 3, pp. 203–4.

Proof without Words: Geometric Series

THE VIEWPOINTS 2000 GROUP*



$a + ar + ar^2 + \cdots = \frac{a}{1-r}, \quad 0 < r < 1.$



$a - ar + ar^2 - \cdots = \frac{a}{1+r}, \quad 0 < r < 1.$

*__The VIEWPOINTS 2000 Group__ is a subset of the participants in the NSF-sponsored VIEWPOINTS Mathematics and Art workshop, held at Franklin & Marshall College in June, 2000:

Marion Cohen, Drexel University, Philadelphia, PA 19104
Douglas Ensley, Shippensburg University, Shippensburg, PA 17257
Marc Frantz, Indiana University, Bloomington, IN 47405
Patricia Hauss, Arapahoe Community College, Littleton, CO, 80160
Judy Kennedy, University of Delaware, Newark, DE 19716
Kerry Mitchell, University of Advancing Computer Technology, Tempe, AZ 85283
Patricia Oakley, Goshen College, Goshen, IN 46526

The Arithmetic–Geometric Mean Inequality and the Constant *e*

HANSHENG YANG
HENG YANG
Southwest China University of Science and Technology
Mianyang, Sichuan
China 621002
hsyang@swit.edu.cn

T. N. T. Goodman [1] and C.W. Barnes [2] gave two interesting proofs of the limit $\lim_{n\to\infty}(1+1/n)^n = e$ using the inequalities

$$\frac{e}{1+1/n} \le \left(1+\frac{1}{n}\right)^n \le e. \tag{1}$$

(See also [3, p. 354].)

In this note we present a very elementary proof that the inequalities

$$\left(1+\frac{1}{n}\right)^n < e \le \left(1+\frac{1}{m-1}\right)^m \tag{2}$$

hold for every integers $n > 0$ and $m > 1$. We use only the well-known *arithmetic–geometric mean inequality* (AGMI): For any n positive real numbers $x_1, x_2, \ldots, x_n$, we have

$$\sqrt[n]{x_1 x_2 \ldots x_n} \le \frac{x_1 + x_2 + \cdots + x_n}{n},$$

or, equivalently,

$$x_1 x_2 \ldots x_n \le \left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)^n; \tag{3}$$

equality holds if and only if $x_1 = x_2 = \cdots = x_n$. (See, e.g., [4] for more on the AGMI.)

Now set $x_n = (1+1/n)^n$. By the AGMI (3) with $n+1$ terms, we have

$$x_n = \left(1+\frac{1}{n}\right)^n \cdot 1 < \left[\frac{\overbrace{\left(1+\frac{1}{n}\right) + \cdots + \left(1+\frac{1}{n}\right)}^{n} + 1}{n+1}\right]^{n+1}$$
$$= \left(\frac{n+1+1}{n+1}\right)^{n+1} = x_{n+1},$$

which proves that the sequence $\{x_n\}$ is increasing.

For an arbitrary positive integer $q > 1$, let the equation $1/p + 1/q = 1$ determine the number $p > 1$; note that $1 + q/p = q$. Using (3) again, we see that

$$x_n \cdot \left(\frac{1}{p}\right)^q = \left(1+\frac{1}{n}\right)^n \left(\frac{1}{p}\right)^q \le \left[\frac{\overbrace{(1+\frac{1}{n})+\cdots+(1+\frac{1}{n})}^{n}+\overbrace{\frac{1}{p}+\cdots+\frac{1}{p}}^{q}}{n+q}\right]^{n+q}$$

$$= \left(\frac{n+1+\frac{q}{p}}{n+q}\right)^{n+q} = 1,$$

which implies that

$$x_n = \left(1+\frac{1}{n}\right)^n \le p^q. \tag{4}$$

Thus the sequence $\{x_n\}$ is increasing and bounded, and so converges to a limit we call e. By (4), e satisfies

$$\left(1+\frac{1}{n}\right)^n < e \le p^q.$$

If we set $q = m > 1$, then $p = m/(m-1)$, and we obtain the inequalities (2).

The special case $m = n+1$ in (2) gives the inequalities

$$\left(1+\frac{1}{n}\right)^n < e \le \left(1+\frac{1}{n}\right)^{n+1} := y_n.$$

Next, we show how to apply the AGMI again to prove the *strict* inequality

$$x_n = \left(1+\frac{1}{n}\right)^n < e < \left(1+\frac{1}{n}\right)^{n+1} = y_n, \tag{5}$$

and that $\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n$. Note that inequality (5) is stronger than (1). In fact,

$$\frac{y_{n+1}}{y_n} = \left(1+\frac{1}{n+1}\right)^{n+1}\left(\frac{1+\frac{1}{n+1}}{(1+\frac{1}{n})^{n+1}}\right) = \left(\frac{n(n+2)}{(n+1)^2}\right)^{n+1}\left(1+\frac{1}{n+1}\right)$$

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

Using the AGMI again, we have

$$\frac{y_{n+1}}{y_n} = \left(1-\frac{1}{(n+1)^2}\right)^{n+1}\left(1+\frac{1}{n+1}\right)$$

$$< \left[\frac{\overbrace{1-\frac{1}{(n+1)^2}+\cdots+1-\frac{1}{(n+1)^2}}^{n+1}+1+\frac{1}{n+1}}{n+2}\right]^{n+2} = 1.$$

Thus the sequence $\{y_n\}$ is decreasing and obviously positive, so it converges to a limit we call e, with $y_n > e$. On the other hand,

$$\frac{x_n}{x_{n+1}} = \frac{\left(1+\frac{1}{n}\right)^n}{\left(1+\frac{1}{n+1}\right)^{n+1}} = \left[\frac{(n+1)^2}{n(n+2)}\right]^n \cdot \frac{n+1}{n+2}$$

$$< \left[\frac{\overbrace{\frac{(n+1)^2}{n(n+2)} + \cdots + \frac{(n+1)^2}{n(n+2)}}^{n} + \frac{n+1}{n+2}}{n+1}\right]^{n+1}$$

$$= \left(\frac{\frac{(n+1)^2}{n+2} + \frac{n+1}{n+2}}{n+1}\right)^{n+1} = 1.$$

This shows that the sequence $\{x_n\}$ is increasing, and clearly $x_n < y_n < y_1 = 4$. Hence $\{x_n\}$ converges, and $x_n < \lim_{n\to\infty} x_n$ for all n. We can prove easily that $\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = e$, and (5) follows.

REFERENCES

→ T. N. T. Goodman, Maximum products and $\lim(1 + 1/n)^n = e$, *Amer. Math. Monthly* **93** (1986), 638–640.

→ C. W. Barnes, Euler's constant and e, *Amer. Math. Monthly* **91** (1984), 428–430.

→ Chung-Lie Wang, Simple inequalities and old limits, *Amer. Math. Monthly* **96** (1989), 354–355.

4. Tom M. Apostol, *Calculus*, Vol. 1, 2nd ed., John Wiley & Sons, Inc., New York, 1967, pp. 46-47.

50 Years Ago in the MAGAZINE

From John Lowe's article, "Automatic Computation as an Aid in Aeronautical Engineering," Vol. **25**, No. 1, (Sept.–Oct., 1951):

> A tremendous amount of numerical labor is involved in designing today's aircraft and missiles. . . . For example, in one phase of the fuselage stress analysis of a single configuration of the DC-6 airplane, 200,000 multiplications and additions were performed, and the flutter analysis required 1,000,000 multiplications and additions. In these facts we find our first reason for the use of automatic computers. . . .
>
> Machines must be programmed in the most minute detail. Problems which one does not consider in computing with a desk calculator become of paramount importance. For example, the determination of whether or not a given quantity is zero may require some planning. If data in graphical form are to be introduced into a computation, these data must be translated to numerical form, perhaps by some curve fitting method. . . .
>
> There exists a stringent shortage of people qualified for this work and this shortage shows every sign of becoming more acute. Capable people are being paid well. The field is so new that few people even know it exists. So little is known that ambitious people can be doing truly original work early in their careers. I hope that it will receive increasing recognition in school curricula and from student counselors. Today, few other fields offer technical or scientific college graduates the opportunity for advancement that is offered by computing.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN 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 March 1, 2002.

1628. *Proposed by Răzvan Gelca, Texas Tech University, Lubbock, TX.*

Find all functions $f:\mathbb{N} \longrightarrow \mathbb{N}$ satisfying

$$f\,(f\,(f(n))) + 6f(n) = 3f\,(f(n)) + 4n + 2001$$

for all $n \in \mathbb{N}$, where $\mathbb{N}$ is the set of positive integers.

1629. *Proposed by Peng Gao, Ann Arbor, MI.*

Define $f_1(x) = \dfrac{d}{dx}\sin^2 x$, and for integer $n \geq 2$, define

$$f_n(x) = \frac{d}{dx}\left((\sin^2 x) f_{n-1}(x)\right).$$

Find the value of $f_n(\pi/2)$.

1630. *Proposed by Geoffrey A. Kandall, Hamden, CT.*

Let P be in the interior of $\triangle ABC$, and let lines AP, BP, and CP intersect sides BC, CA, and AB in L, M, and N, respectively. Prove that if

$$\frac{AP}{PL} + \frac{BP}{PM} + \frac{CP}{PN} = 6,$$

then P is the centroid of $\triangle ABC$.

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.

1631. *Proposed by Hoe-Teck Wee, student, Massachusetts Institute of Technology, Cambridge, MA.*

For each integer $n > 3$ and not divisible by 3, how many ways are there to delete a square from an $n \times n$ chess board so that the remaining board can be tiled with 3×1 trominos?

1632. *Proposed by Erwin Just (Emeritus), Bronx Community College, New York, NY.*

Prove that there are an infinite number of integers n for which there exists a set of n distinct positive odd integers such that each member of the set divides the sum of all the members of the set.

Quickies

Answers to the Quickies are on page 330.

Q913. *Proposed by Murray S. Klamkin, The University of Alberta, Edmonton, AB, Canada.*

Let F be a function that has a continuous third derivative on $[0, 1]$. If $F(0) = F'(0) = F''(0) = F'(1) = F''(1) = 0$ and $F(1) = 1$, prove that $F'''(x) \geq 24$ for some x in $[0, 1]$.

Q914. *Proposed by Kelly Jahns, Spokane Community College, Spokane, WA.*

For $n, K \in \mathbb{N}$, define $\tau(n, K)$ to be the number of positive divisors d of n such that $d \leq K$ and $n/d \leq K$. Given fixed $K \in \mathbb{N}$, evaluate

$$\sum_{n=1}^{\infty} \tau(n, K).$$

Solutions

A Cubic Sum of Fifths — October 2000

1603. *Proposed by Ho-joo Lee, student, Kwangwoon University, Seoul, South Korea.*

Find all integer solutions to $x^5 + y^5 = (x + y)^3$.

Solution by Robert L. Doucette, McNeese State University, Lake Charles, LA.

The integer pair (x, y) is a solution of the given equation if and only if $x + y = 0$ or $(x, y) \in \{(0, \pm 1), (\pm 1, 0), \pm(2, 2)\}$. Clearly, if $x + y = 0$, then (x, y) is a solution. Assume now that (x, y) is a solution with $x + y \neq 0$.

We first show that $xy \geq 0$. Dividing both sides of $x^5 + y^5 = (x + y)^3$ by $x + y$ yields

$$x^4 - x^3y + x^2y^2 - xy^3 + y^4 = (x + y)^2.$$

This is equivalent to

$$(x^2 + y^2)^2 + x^2y^2 = (x + y)^2(xy + 1),$$

and it follows that $xy \geq 0$.

We next show that $|x + y| \leq 4$. The convexity of the function $t \to t^5$ on $0 \leq t < \infty$ implies that for nonnegative x and y,

$$\frac{x^5+y^5}{2} \geq \left(\frac{x+y}{2}\right)^5, \text{ or equivalently, } x^5+y^5 \geq \frac{1}{16}(x+y)^5.$$

If $x+y>4$, then $x^5+y^5>(x+y)^3$. Similarly, if x and y are both nonpositive with $x+y<-4$, then $x^5+y^5<(x+y)^3$.

Examining the cases where $xy \geq 0$ and $|x+y| \in \{1, 2, 3, 4\}$, we find the solutions $(x, y) = (0, \pm 1), (\pm 1, 0), \pm(2, 2)$.

Note. Jim Delany investigated nontrivial (i.e., $x+y \neq 0$) rational solutions to the equation, and showed that they correspond to rational points on the elliptic curve E defined by

$$r^2 = q(q^2-5q+5).$$

In particular a rational point (q, r) on E corresponds to a rational solution

$$x = 2(q^2-5q+5+r)/(q^2-5), \quad y = 2(q^2-5q+5-r)/(q^2-5)$$

to $x^5+y^5=(x+y)^3$. The group of rational points of E is the abelian group generated by the two solutions $(0, 0)$ and $(1, 1)$, the first of order two and the second of infinite order.

Also solved by The Assumption College Problems Group, Roy Barbara (Lebanon), Michel Bataille (France), Jean Bogaert (Belgium), Marc A. P. Bernstein (France), Owen Byer, John Christopher, Jeffery Clark, Con Amore Problem Group (Denmark), Knut Dale and Ivar Skau (Norway), Jim Delany, Daniele Donini (Italy), Petar D. Drianov (Canada), Arthur H. Foss, Joel D. Haywood, Tom Jager, Victor Y. Kutsenok, Stephen Maguire, Kevin McDougal, Can Ahn Minh, Kandasamy Muthuvel, Stephen Noltie, Robert L. Raymond, Kenneth Rogers, Shiva K. Saksena, Harry Sedinger, Heinz-Jürgen Seiffert (Germany), Ajaj A. Tarabay and Bassem B. Ghalayini (Lebanon), Joseph Wiener and Donald P. Skow, Rex H. Wu, Monte J. Zerger, Paul J. Zwier, and the proposer. Three incomplete solutions were also received.

An Intermediate Value October 2000

1604. *Proposed by Răzvan Tudoran, University of Timişora, Timişora, Romania.*

Let g be a differentiable function on the nonnegative reals such that $g(0) \in [0, 1]$ and $\lim_{x\to\infty} g(x) = \infty$. Let f be defined on the nonnegative reals and satisfy $f(0) > g(0)$ and, for some positive k and r and all nonnegative x and y,

$$|f(x)-f(y)| \leq k|g(x)-g(y)|^r.$$

Prove that there exists nonnegative c such that $f(c) = [g(c)]^{\lfloor r\rfloor+1}$.

Solution by Ivar Skau, Telemark College, Norway.

The continuity of g and the inequality in the problem statement imply that f is continuous on $[0, \infty)$. Let $h(x) = g(x)^{\lfloor r\rfloor+1}$, so h is also continuous on $[0, \infty)$. Note that there is an $\epsilon > 0$ such that $\lfloor r\rfloor + 1 = r + \epsilon > r$. Hence

$$f(0) > g(0) \geq g(0)^{\lfloor r\rfloor+1} = h(0)$$

and, for sufficiently large x,

$$\begin{aligned} f(x) &\leq f(0) + |f(x)-f(0)| \leq f(0) + k|g(x)-g(0)|^r < |g(x)|^{r+\epsilon} \\ &= g(x)^{\lfloor r\rfloor+1} = h(x). \end{aligned}$$

The desired conclusion follows from the Intermediate Value Theorem. Note that we only required that g be continuous, but not necessarily differentiable.

Also solved by Con Amore Problem Group (Denmark), Robert L. Doucette, Tom Jager, Kandasamy Muthuvel, Stephen Noltie, Marie Spong, and the proposer.

Maximal Area with Distance Constraint **October 2000**

1605. *Proposed by Chi Hin Lau, student, University of Hong Kong, Hong Kong, China.*

In $\triangle ABC$, $\angle A = 60°$ and P is a point in its plane such that $PA = 6$, $PB = 7$, and $PC = 10$. Find the maximum possible area of $\triangle ABC$.

(I) *Solution by Daniele Donini, Bertinoro, Italy.*

The maximum area is $36 + 22\sqrt{3}$. Let $\angle A = \alpha$, $PA = a$, $PB = b$, and $PC = c$, with $0 < \alpha \leq \pi/2$ and $0 < a < \min\{b, c\}$. Let $x = \angle BAP$. The conditions on a, b, c, and α imply that for any given x there is a unique configuration of points A, B, C, and P. For given x we have

$$AB = a\cos x + \sqrt{b^2 - a^2\sin^2 x} \qquad \text{and}$$
$$AC = a\cos(\alpha - x) + \sqrt{c^2 - a^2\sin^2(\alpha - x)},$$

so the area of $\triangle ABC$ is given by

$$f(x) = \frac{\sin\alpha}{2}\left(a\cos x + \sqrt{b^2 - a^2\sin^2 x}\right)\left(a\cos(\alpha - x) + \sqrt{c^2 - a^2\sin^2(\alpha - x)}\right).$$

The function f is strictly positive, differentiable, and has period 2π. Hence it takes on its extremes values at some critical points in the interval $[0, 2\pi)$. Direct calculation gives

$$f'(x) = af(x)\left(-\frac{\sin x}{\sqrt{b^2 - a^2\sin^2 x}} + \frac{\sin(\alpha - x)}{\sqrt{c^2 - a^2\sin^2(\alpha - x)}}\right).$$

The equation $f'(x) = 0$ leads to

$$\frac{\sin x}{b} = \frac{\sin(\alpha - x)}{c},$$

from which

$$\tan x = \frac{b\sin\alpha}{c + b\cos\alpha}.$$

There are two values, x_1, x_2 in $[0, 2\pi)$ that satisfy the equation. These values are characterized by

$$\sin x_1 = \frac{b\sin\alpha}{\sqrt{b^2 + c^2 + 2bc\cos\alpha}}, \qquad \cos x_1 = \frac{c + b\cos\alpha}{\sqrt{b^2 + c^2 + 2bc\cos\alpha}}$$

and

$$\sin x_2 = -\frac{b\sin\alpha}{\sqrt{b^2 + c^2 + 2bc\cos\alpha}}, \qquad \cos x_2 = -\frac{c + b\cos\alpha}{\sqrt{b^2 + c^2 + 2bc\cos\alpha}}.$$

Because $0 < \sin x_1 < \sin\alpha$, we have $0 < x_1 < \alpha$ and $x_2 = x_1 + \pi$. These values correspond to areas of

$$f(x_1) = \frac{\sin\alpha}{2}\left(bc + a^2\cos\alpha + a\sqrt{b^2 + c^2 - a^2 + 2bc\cos\alpha + a^2\cos^2\alpha}\right)$$

and

$$f(x_2) = \frac{\sin\alpha}{2}\left(bc + a^2\cos\alpha - a\sqrt{b^2 + c^2 - a^2 + 2bc\cos\alpha + a^2\cos^2\alpha}\right).$$

Because $f(x_1) > f(x_2)$, these values are the maximum and minimum values of f. With $\alpha = 60°$, $a = 6$, $b = 7$, and $c = 10$ we obtain a maximal area of $22\sqrt{3} + 36$ and a minimal area of $22\sqrt{3} - 36$.

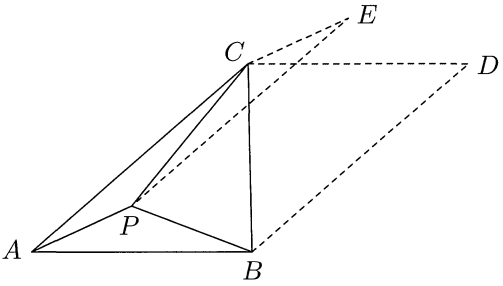
Note that for either extreme value,

$$\sin(\angle ABP) = a\frac{\sin x}{b} = a\frac{\sin(\alpha - x)}{c} = \sin(\angle ACP),$$

and it follows that $\angle ABP = \angle ACP$, because both angles are acute.

II. *Solution by the proposer.*

Let D and E be points such that $ABDC$ and $APEC$ are parallelograms. Then $PBDE$ is also a parallelogram.



Observe that

$$\begin{aligned}
&PD^2 - PC^2 + PA^2 - PB^2\\
&\quad= \|\overrightarrow{PD}\|^2 - \|\overrightarrow{PC}\|^2 + \|\overrightarrow{PA}\|^2 - \|\overrightarrow{PB}\|^2\\
&\quad= \|\overrightarrow{PA} + \overrightarrow{AD}\|^2 - \|\overrightarrow{PA} + \overrightarrow{AC}\|^2 + \|\overrightarrow{PA}\|^2 - \|\overrightarrow{PA} + \overrightarrow{AB}\|^2\\
&\quad= \|\overrightarrow{AD}\|^2 + 2\overrightarrow{PA}\cdot\overrightarrow{AD} - \|\overrightarrow{AC}\|^2 - 2\overrightarrow{PA}\cdot\overrightarrow{AC} - \|\overrightarrow{AB}\|^2 - 2\overrightarrow{PA}\cdot\overrightarrow{AB}\\
&\quad= \|\overrightarrow{AB} + \overrightarrow{AC}\|^2 - \|\overrightarrow{AC}\|^2 - \|\overrightarrow{AB}\|^2\\
&\quad= 2\overrightarrow{AB}\cdot\overrightarrow{AC} = 2AB\cdot AC\cos A = AB\cdot AC.
\end{aligned}$$

Thus, $PD^2 - 100 + 36 - 49 = CD \cdot PE$. By Ptolemy's theorem,

$$PD^2 - 113 \le CE\cdot PD + PC\cdot DE = PA\cdot PD + PC\cdot PB = 6PD + 70.$$

Hence $PD^2 - 6PD - 183 \le 0$, from which $PD \le 3 + 8\sqrt{3}$, and equality holds if and only if $PDEC$ is cyclic. Therefore

$$\begin{aligned}
[ABC] &= \frac{\sqrt{3}}{4}AB\cdot AC = \frac{\sqrt{3}}{4}CD\cdot PE\\
&\le \frac{\sqrt{3}}{4}(PC\cdot DE + PD\cdot CE) = \frac{\sqrt{3}}{4}(PC\cdot PB + PD\cdot PA)\\
&\le \frac{\sqrt{3}}{4}(70 + 18 + 48\sqrt{3}) = 36 + 22\sqrt{3},
\end{aligned}$$

where again equality holds if and only if $PDEC$ is cyclic. Note that this is the case if and only if $\angle CPE = \angle CDE$, that is, if and only if $\angle ACP = \angle ABP$.

Also solved by Herb Bailey, Jean Bogaert (Belgium), Con Amore Problem Group (Denmark), Knut Dale (Norway), Robert L. Doucette, Victor Y. Kutsenok, Jayavel Sounderpandian, Ajaj A. Tarabay and Bassem B. Ghalayini (Lebanon), Michael Vowe (Switzerland), and Li Zhou. There were two incorrect submissions.

A Series for $\frac{\sin x}{x}$ **October 2000**

1606. *Proposed by Anthony A. Ruffa, Naval Undersea Warfare Center Division, Newport, RI.*

For x real and nonzero, show that

$$\frac{\sin x}{x} = \cos^2(x/2) + \sum_{n=1}^{\infty} \sin^2\left(x/2^{n+1}\right) \prod_{m=1}^{n} \cos\left(x/2^m\right).$$

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

For positive integer n define

$$P_n(x) = \prod_{k=1}^{n} \cos\left(x/2^k\right) \quad \text{and} \quad S_n(x) = x\cos^2(x/2) + x\sum_{k=1}^{n} \sin^2\left(x/2^{k+1}\right) P_k(x).$$

Next observe that

$$\sin x = \cos(x/2)\left(x\cos(x/2) + x\sin^2(x/4) - x\cos^2(x/4) + 2\sin(x/2)\right), \tag{1}$$

so that

$$\sin x = S_1(x) - xP_2(x)\cos(x/4) + 2P_1(x)\sin(x/2). \tag{2}$$

It can be shown by induction that

$$\sin x = S_n(x) - xP_{n+1}(x)\cos\left(x/2^{n+1}\right) + 2^n P_n(x)\sin\left(x/2^n\right). \tag{3}$$

The case $n = 1$ is established in (1) and (2). For the induction step, replace x by $x/2^n$ in (1) and substitute the resulting expression for $\sin(x/2^n)$ into (3).

Because $|P_n(x)| \le 1$, we have

$$\begin{aligned}&\lim_{n\to\infty}\left(-xP_{n+1}(x)\cos\left(x/2^{n+1}\right) + 2^n P_n(x)\sin\left(x/2^n\right)\right)\\ &\quad = \lim_{n\to\infty}\left(P_n(x)\left(-x\cos^2\left(x/2^{n+1}\right) + 2^n\sin\left(x/2^n\right)\right)\right) = 0.\end{aligned}$$

It follows from (3) that

$$\sin x = \lim_{n\to\infty} S_n(x),$$

establishing the desired identity.

Note: Several partial solutions established the identity in the case when x/π is not an integer, but neglected special considerations needed when x/π is an integer.

Also solved by Michel Bataille (France), Jean Bogaert (Belgium), Con Amore Problem Group (Denmark), Daniele Donini (Italy), Robert L. Doucette, Mordechai Falkowitz (Canada), Tom Jager, Paul Zwier, and the proposer. Two incorrect results were submitted.

Ubiquitous Subsets **October 2000**

1607. *Proposed by Hassan A. Shah Ali, Tehran, Iran.*

Let n, k and m be positive integers satisfying

$$m > \binom{n}{k} - \binom{\lfloor n/2 \rfloor}{k} - \binom{\lceil n/2 \rceil}{k}.$$

Let A be a set with $|A| = n$ and let $A_1, \ldots, A_m$ be distinct k-subsets of A. Prove that if $a_1 \in A_1, \ldots, a_m \in A_m$, then there exists $i \in \{1, \ldots, m\}$ such that $A_i \subset \{a_1, \ldots, a_m\}$.

Solution by Reiner Martin, New York, NY.

Let $B = \{a_1, a_2, \ldots, a_m\}$ and $l = |B|$. The number of k-subsets of A that are contained in B or $A - B$ is

$$s(n, k, l) = \binom{l}{k} + \binom{n-l}{k}.$$

Because

$$s(n,k,l) + \binom{l}{k-1} - \binom{n-l-1}{k-1} = s(n,k,l+1),$$

it is clear that

$$s(n,k,l) \geq s(n,k,\lfloor n/2 \rfloor) = \binom{\lfloor n/2 \rfloor}{k} + \binom{\lceil n/2 \rceil}{k}$$

for all $0 \leq l \leq n$. Thus the number of k-subsets of A that are *not* contained in B or $A - B$ is less than m. Because B meets every A_j, we must have $A_i \subseteq B$ for some i.

Solution by Jean Bogaert (Belgium), Owen Byer, Robert L. Doucette, Tom Jager, Robert Pratt, John H. Smith, Li Zhou, and the proposer.

Answers

Solutions to the Quickies from page 324.

A913. Consider the Taylor series expansions about the points $x = 0$ and $x = 1$,

$$F(x) = F(0) + F'(0)x + \frac{F''(0)}{2}x^2 + \frac{F'''(c_1)}{6}x^3,$$

$$F(x) = F(1) + F'(1)(x-1) + \frac{F''(1)}{2}(x-1)^2 + \frac{F'''(c_2)}{6}(x-1)^3,$$

where $0 \leq c_1 \leq x$ and $x \leq c_2 \leq 1$. These reduce to

$$F(x) = \frac{F'''(c_1)}{6}x^3 \qquad \text{and} \qquad F(x) = 1 + \frac{F'''(c_2)}{6}(x-1)^3.$$

Setting $x = 1/2$, we find that there are c_1 and c_2 with $F'''(c_1) + F'''(c_2) = 48$. Thus at least one of $F'''(c_1)$ and $F'''(c_2)$ is greater than or equal to 24.

A914. Let $d \in \mathbb{N}$, and suppose that d is counted (exactly once) in $\tau(n, K)$ for some n. Then $d \leq K$, $d|n$, and $n/d \leq K$. It follows that $n \leq Kd$, and because n is a multiple of d,

$$n \in \{d, 2d, 3d, \ldots, Kd\}.$$

Now suppose that $d \in \mathbb{N}$, $d \leq K$, and $n \in \{d, 2d, 3d, \ldots, Kd\}$. Then $d|n$ and $n/d \leq K$, so d is counted in $\tau(n, K)$.

Thus, each $d \in \{1, 2, 3, \ldots, K\}$ is counted exactly once for each $n \in \{d, 2d, 3d, \ldots, Kd\}$. It follows that

$$\sum_{n=1}^{\infty} \tau(n, K) = K^2.$$

Solution to the problem from page 309.

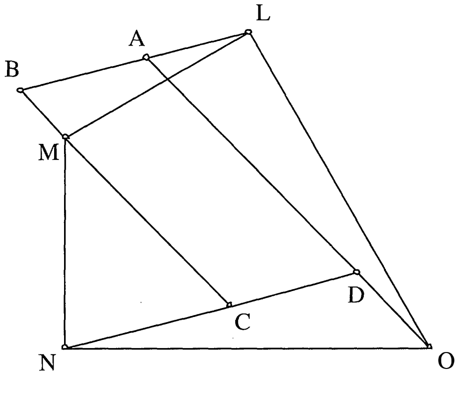
Let us define $\angle AON = \theta$. The figure is drawn with $\theta \approx \frac{\pi}{4}$. When θ is near this value the correct relationship is

$$OA + ND = 3MC \quad \text{and} \quad LB = MC.$$

However, this is only valid when point A is on the same side of line ON as L. The full solution is

$$\begin{aligned} OA + ND &= 3MC \left(0 \leq \theta \leq \frac{2\pi}{3}\right) \\ OA - ND &= 3MC \left(-\frac{\pi}{6} \leq \theta < 0\right) \\ -OA + ND &= 3MC \left(-\frac{\pi}{3} < \theta < -\frac{\pi}{6}\right). \\ LB &= MC \end{aligned}$$

This figure can be constructed by choosing A to be any point on the circumcircle of $\triangle NOL$, which explains the possible range of the angle θ.



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.

Engquist, Björn, and Wilfried Schmid, eds., *Mathematics Unlimited—2001 and Beyond*, Springer-Verlag, 2001; xv + 1237 pp, $44.95. ISBN 3–540–66913–2.

"What are the important developments in present-day mathematics? Where is mathematics headed?" Here are 62 essays on all kinds of mathematics, pure and applied—but mostly all kinds of applied mathematics, as specific as medical imaging, climate modeling, twistors, quantum computing, cryptography, and financial markets, and as general as computational aspects of number theory, experimental mathematics, astrophysics, molecular evolution, and mirror symmetry. [Dispensing with the little-used extra-wide margins would have made the book lighter than 3+ kg, and an index would have been useful.]

Albert, Jim, and Jay Bennett, *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*, Springer-Verlag, 2001; xviii + 350 pp, $29. ISBN 0–387–98816–5.

Despite its title, this book does not deal with the physics of baseball and the fact (long disputed) that curve balls curve but with baseball statistics. "[N]ext to stock market analysis, [baseball] may be the most widespread application of statistics in the United States." The statistician authors explore baseball simulations, data analysis of hitting data, probability models, situational effects, streakiness, alternative measures, state-space models, and prediction. They introduce some data analysis techniques (e.g., stem-and-leaf plots) and acquaint the reader informally with p-values but nevertheless keep the discussion at a popular level. Surprisingly absent are references or mention of Stephen Jay Gould's two notable articles on why there are no more .400 hitters and on Joe DiMaggio's remarkable streak of games with hits (see "Entropic homogeneity isn't why no one hits .400 any more," *Discover* (August 1986) 60–66, and "Streak of streaks," *New York Review of Books* **35** (18 August 1988) 8–12), both reprinted in Gould's books.) [This book, too, has no index! Despite indexing and concordance programs, and unless the book is produced by offset from pages prepared by the author, the author is not in a position to prepare the index. In any case, an index is better done by a professional indexer. Authors must insist on, and publishers should feel obliged to provide, an index.]

Peterson, Ivars, Circle game: Packing circles within a circle turns a mathematical surprise, *Science News* **159** (21 April 2001), 254–255, `http://www.sciencenews.org/20010421/bob18.asp` . Ivars Peterson's MathTrek: Temple circles `http://www.maa.org/mathland/mathtrek_4_23_01.html` .

Japanese *sangaku* problems usually involve multiple tangential circles within a large outer circle, examples of what are known in the West as Apollonian packings. The radii of such circles follow a formula found by Descartes in 1643, and the Japanese problems (which date from the late 18th century) are solved by a similar independently-discovered formula. Motivated by a German colleague's daughter's school assignment, Allan R. Wilks (AT&T Labs) conjectured a further relationship between the radii and the coordinates of the centers of the circles (with origin at the center of the large outer circle). He and other colleagues have proved and generalized the discovery to higher dimensions in euclidean, spherical, and hyperbolic geometry. Mysteries remain, such as whether some integers never appear as radii in any pattern.

Preuss, Paul, Are the digits of pi random? A Berkeley lab researcher may hold the key, `http://www.science.doe.gov/feature_articles_2001/July/Digits_of_Pi/Digits%20of%20Pi.htm` . Klarreich, Erica, Pi shared fairly: Mathematicians edge closer to proving that all numbers get an equal slice of pi, *Nature* (2 August 2001), `http://www.nature.com/nsu/010802/010802-9.html` . Bailey, David H., and Richard E. Crandall, On the random character of fundamental constant expansions, *Experimental Mathematics* **10** (2) (Summer 2001), 175–190, `http://www.expmath.org/restricted/10/10.2/baicran.ps` . Arndt, Jörg, and Christoph Haenel, *Pi—Unleashed*, Springer-Verlag, 2001; xii + 270 pp + CD-ROM, $29.95. ISBN 3–540–66572–2.

Are the digits of pi random? In particular, is pi *normal*—do all digits occur with equal frequency in its expansion? Arndt and Haenel's book came out just before the discovery by Bailey and Crandall that a hypothesis about dynamical maps implies base-2 normality of π, log 2, and $\zeta(3)$, and of $\zeta(5)$ if it is irrational. Their Hypothesis A concerns a rational function $r_n = p(n)/q(n)$ of polynomials with integer coefficients, with $\deg p \leq \deg q$ and $q(n)$ never 0. The Hypothesis is that for any $b \geq 2$ and starting point $x_0 = 0$, the iterates $x_n \equiv (bx_{n-1} + r_n) \bmod 1$ either have a finite attractor or are equidistributed in $[0, 1)$. Moreover, each famous constant corresponds to a specific function r_n. If the hypothesis is true, the digits of these constants "appear to be random because they are closely approximated by ... chaotic dynamics." Bailey and Crandall's work springs from an earlier discovery of Bailey and others of an algorithm to calculate an arbitrary digit in the binary expansion of π without calculating any of the preceding digits, based on

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k} \left[\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right].$$

The book by Jörg and Haenel, fascinating in its own right for its history and algorithms for π, explains and gives code for this algorithm, known as the BBP algorithm. [Hey—an index!]

Dewdney, A.K., The forest and the trees: Romancing the J-curve, *Mathematical Intelligencer* **23** (3) (2001), 27–34. A dynamical model of communities and a new species-abundance distribution, *Biological Bulletin* **198** (1) (2000), 153–163.

For virtually every ecosystem, a histogram of the number of species vs. their abundance has the shape of a backward J. Biologists have proposed various formulas for the curve, including the lognormal and log-series probability distributions. Motivated by his own biological collection and by computer models that produce J-curves, mathematician Dewdney set out to find a definitive model and formula. His search and discovery are illuminating; the result is nearly a hyperbolic curve. How will the ecologists react?

Peterson, Ivars, Surprisingly square: Mathematicians take a fresh look at expressing numbers as the sums of squares, *Science News* **159** (16 June 2001), 382–383.

Every positive integer is a sum of no more than 4 squares (Lagrange), and in 1829 Jacobi gave simple formulas for the number of representations of an integer as the sum of 2, 4, 6, or 8 squares. Subsequent results seemed to rule out further such simple formulas, but Stephen C. Milne (Ohio State University), using an elliptic-function approach of Jacobi's, has found new formulas for the number of representations by more than 8 squares.

Stubhaug, Arild, *Niels Henrik Abel and His Times: Called Too Soon by Flames Afar*, Springer-Verlag, 2001; x + 580 pp, $44.95. ISBN 3–540–66834–9.

The dust jacket rightly calls this "the definitive biography" of Abel. In addition to a wealth of material from unpublished sources, it has considerable ancillary material about Abel's time, places he knew, and friends; it even includes a plate of the wallpaper from Abel's death room! This, however, is not a "scientific biography"; Stubhaug includes a bibliography of Abel's published works, but for the mathematics readers will have to go elsewhere. [Another Springer book, this time with just an index of names.]

NEWS AND LETTERS

Carl B. Allendoerfer Awards — 2001

The Carl B. Allendoerfer Awards, established in 1976, are made to honor authors of outstanding expository articles published in MATHEMATICS MAGAZINE. Carl B. Allendoerfer, a distinguished mathematician at the University of Washington, served as President of the Mathematical Association of America, 1959–60. This year's awards were presented at the August 2001 MathFest, in Madison, Wisconsin. The citations follow, along with biographical information and the responses of the winners.

James N. Brawner, Dinner, Dancing, and Tennis, Anyone?, this MAGAZINE 73 (2000), 29–36.

In the 1996 men's draw for the U. S. Open tennis tournament, Andrei Medvedev was pitted against Jean-Philippe Fleurian. For reasons explained in the paper, a redraw was conducted. Again, Medvedev was pitted against Fleurian. A tennis official wondered about the probability of such an event and contacted the author of this paper. After rephrasing the question in two ways, the author answers both questions and connects them to two classical problems: Montmort's problem of coincidences and the *problème des ménages*. The reader is engaged throughout by substantial mathematics, as well as numeric examples. This well-written article concludes with some interesting problems, one of which has already inspired a follow-up article [see Barbara H. Margolius, Avoiding your spouse at a bridge party, this MAGAZINE, **74** (2001), 33–41].

Biographical Note Jim Brawner grew up in Atlanta, Georgia, and has spent his adult life bouncing up and down the Eastern United States. He was an undergraduate at Williams College in Massachusetts, where he majored in English and mathematical sciences, and he received his Ph.D. in algebraic geometry at the University of North Carolina at Chapel Hill. After spending several years at St. John's University, just down the road from the U.S. Tennis Center in Queens, New York, he is now an assistant professor at Armstrong Atlantic State University in Savannah, Georgia. In addition to algebraic geometry, his interests include combinatorics, number theory, balancing objects on his nose, and of course, dining, dancing, and playing tennis with his wife, Aubrey, and their two sons, Jimmy and Will.

Response from James N. Brawner I am thrilled and honored to receive the Carl B. Allendoerfer Prize, and I am very grateful to the prize committee for this honor. "Dinner, Dancing, and Tennis, Anyone?" grew out of a talk given for our wonderful weekly luncheon colloquium series at Armstrong Atlantic State University. I am grateful to Ed Wheeler for his encouragement in writing a paper based on the talk, and for suggesting MATHEMATICS MAGAZINE as an appropriate forum. I wanted to thank (former) editor Paul Zorn in the acknowledgments for his tremendous help, but he claimed he was just doing his job; I am pleased to be able to thank him now for his many helpful suggestions. I have been delighted by several excellent solutions to the problems I posed at the end of the paper, most notably by Barbara Margolius' article in this year's MAGAZINE. Lastly, I would like to thank the USTA for asking an intriguing question, and my family for their love and support in listening to the answer.

Carl B. Allendoerfer Awards — 2001 (continued)

Rafe Jones and Jan Pearce, A Postmodern View of Fractions and the Reciprocals of Fermat Primes, this MAGAZINE 73 (2000), 83–97.

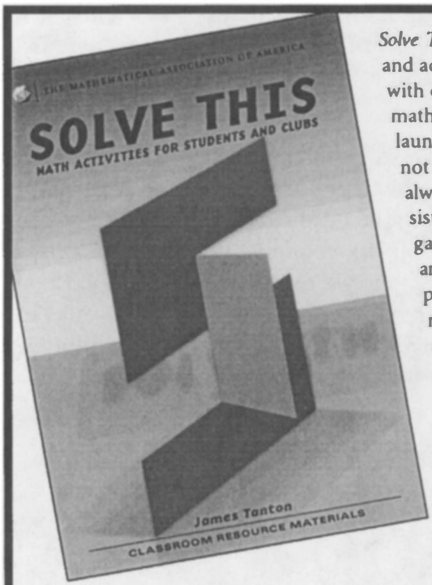
By means of a geometric approach, this article provides new ways of looking at an old subject: decimal expansions of fractions and expansions in bases other than 10. The approach is of great visual appeal and uses the geometry as the basis for weaving together a nice collection of facts from elementary number theory. The article is unusually well written, not only clear so far as its mathematics is concerned but also of excellent quality from a literary standpoint.

The spirit of the article is well conveyed by its final paragraph: "Thus ends our exploration of fractions and symmetry. Postmodernism has taught us that all ways of looking at a problem are not equivalent: different perspectives highlight different properties. Adopting our society's penchant for images led us to examine more closely the symmetries of certain fractions, and opened our eyes to unexpected visions."

Biographical Notes Rafe Jones received a bachelor's degree in mathematics and French from Amherst College in 1998. After a one-year interlude as a visiting student at the Ecole Normal Supérieure in Paris, he began his graduate studies at Brown University in 1999. His studies appear to be headed towards some brand of number theory. He is spending this summer working for Discover.com as the AAAS/AMS Mass Media Fellow.

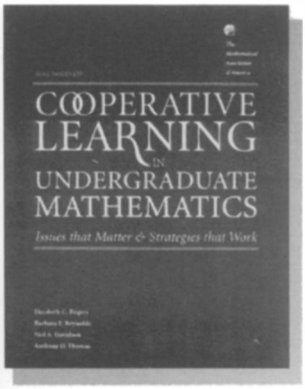
For the last nine years, Jan Pearce has been teaching mathematics and computer science at Berea College. She received her Ph.D. from the University of Rochester in algebraic topology in 1992 and her B.A. from Augustana College with majors in mathematics, computer science, and physics in 1987. She enjoys involving students in undergraduate research, and remembers vividly the day that Rafe Jones said that he was interested in doing research in "anything but number theory." Her current research interest is in Bayesian networks, an area of artificial intelligence, and she will be spending the next academic year on sabbatical at the University of Minnesota in pursuit of this interest.

Response from Rafe Jones and Jan Pearce We are both very pleased the Allendoerfer Prize Committee appreciated our article's mathematics and even more pleased they found the writing to be of good quality. In writing the article, we tried to take to heart one of postmodernism's lessons: language is not just a vehicle, but rather an essential part of a work's message. We'll admit, though, that we're not truly postmodernists; thus, we refrain from offering a deconstructive reading of the citation. We would like to thank the Prize Committee for this honor and also Paul Zorn for his fine editing.


THE MATHEMATICAL ASSOCIATION OF AMERICA
SOLVE THIS
MATH ACTIVITIES FOR STUDENTS AND CLUBS
James Tanton
CLASSROOM RESOURCE MATERIALS

The Mathematical Association of America

Cooperative Learning in Undergraduate Mathematics: Issues that Matter and Strategies that Work



Elizabeth C. Rogers, Barbara E. Reynolds,
Neil A. Davidson, and Anthony D. Thomas, Editors

Series: MAA Notes

This volume offers practical suggestions and strategies both for instructors who are already using cooperative learning in their classes, and for those who are thinking about implementing it. The authors are widely experienced with bringing cooperative learning into the undergraduate mathematics classroom. In addition they draw on the experiences of colleagues who responded to a survey about cooperative learning which was conducted in 1996-97 for Project CLUME (Cooperative Learning in Undergraduate Mathematics Education).

The volume discusses many of the practical implementation issues involved in creating a cooperative learning environment:

- how to develop a positive social climate, form groups and prevent or resolve difficulties within and among the groups.
- what are some of the cooperative strategies (with specific examples for a variety of courses) that can be used in courses ranging from lower-division, to calculus, to upper division mathematics courses.
- what are some of the critical and sensitive issues of assessing individual learning in the context of a cooperative learning environment.
- how do theories about the nature of mathematics content relate to the views of the instructor in helping students learn that content.

The authors present powerful applications of learning theory that illustrate how readers might construct cooperative learning activities to harmonize with their own beliefs about the nature of mathematics and how mathematics is learned.

In writing this volume the authors analyzed and compared the distinctive approaches they were using at their various institutions. Fundamental differences in their approaches to cooperative learning emerged. For example, choosing Davidson's guided-discovery model over a constructivist model based on Dubinsky's action-process-object-schema (APOS) theory affects one's choice of activities. These and related distinctions are explored.

A selected bibliography provides a number of the major references available in the field of cooperative learning in mathematics education. To make this bibliography easier to use, it has been arranged in two sections. The first section includes references cited in the text and some sources for further reading. The second section lists a selection (far from complete) of textbooks and course materials that work well in a cooperative classroom for undergraduate mathematics students.

Catalog Code: NTE-55/JR 250 pp., Paperbound, 2001 ISBN 088385-166-0 List: $ 31.95 MAA Member: $25.95

Name____________________ Credit Card No.____________________

Address____________________ Signature____________________Exp. Date____/____

City____________________ Qty________ Price $__________ Amount $__________

State__________ Zip __________ Shipping and Handling $__________

Phone ____________________ **Catalog Code: NTE-55/JR** Total $__________

Shipping and Handling: USA orders (shipped via UPS): $3.00 for the first book, and $1.00 for each additional book. Canadian orders: $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 2-3 weeks of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. Overseas Orders: $4.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $10.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. Credit card orders are accepted for all customers. All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.

Order Via:

Phone: 1 (800) 331.1622
Fax: (301) 206.9789
Mail: Mathematical Association of America
PO Box 91112
Washington, DC 20090-1112
Web: www.maa.org

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036