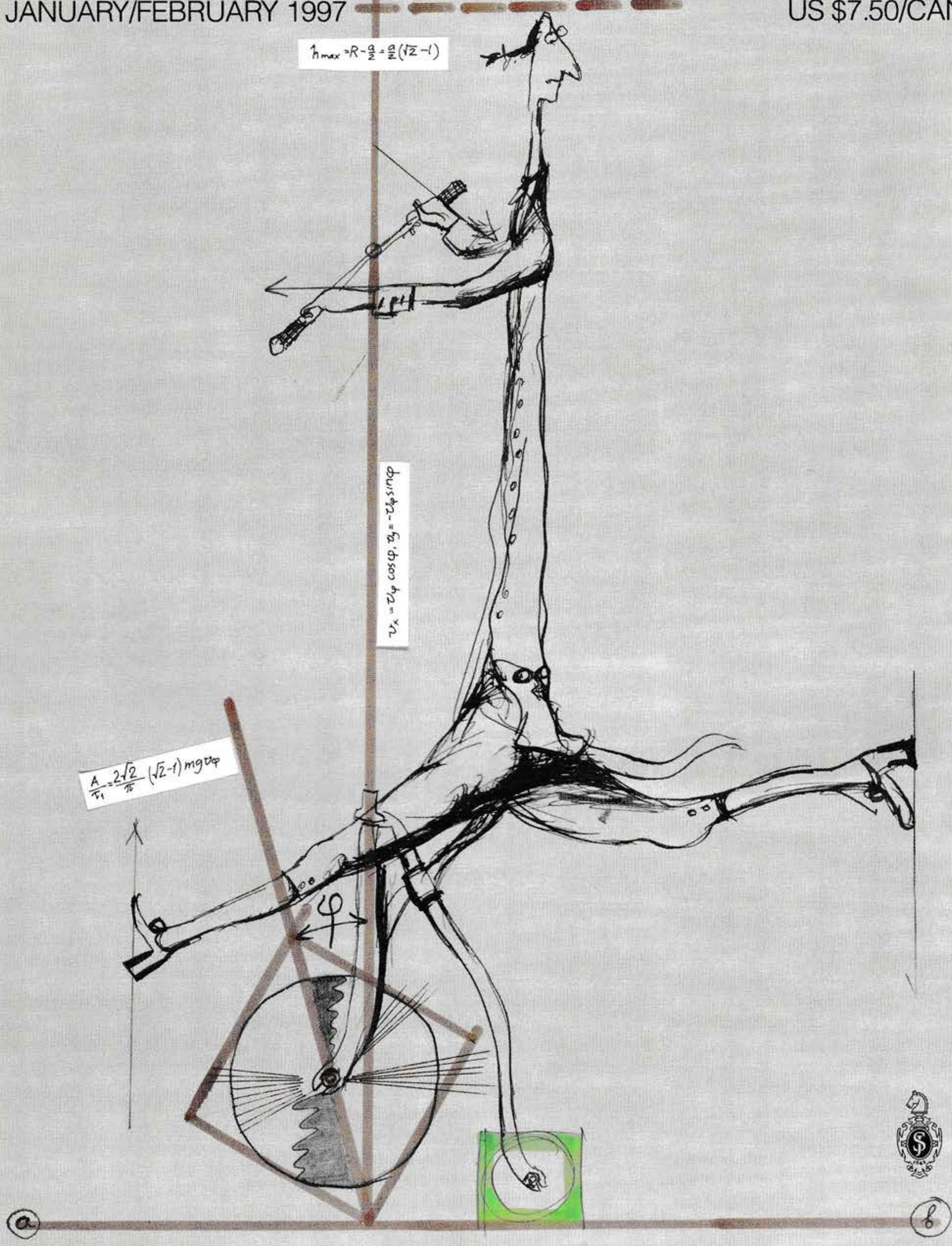
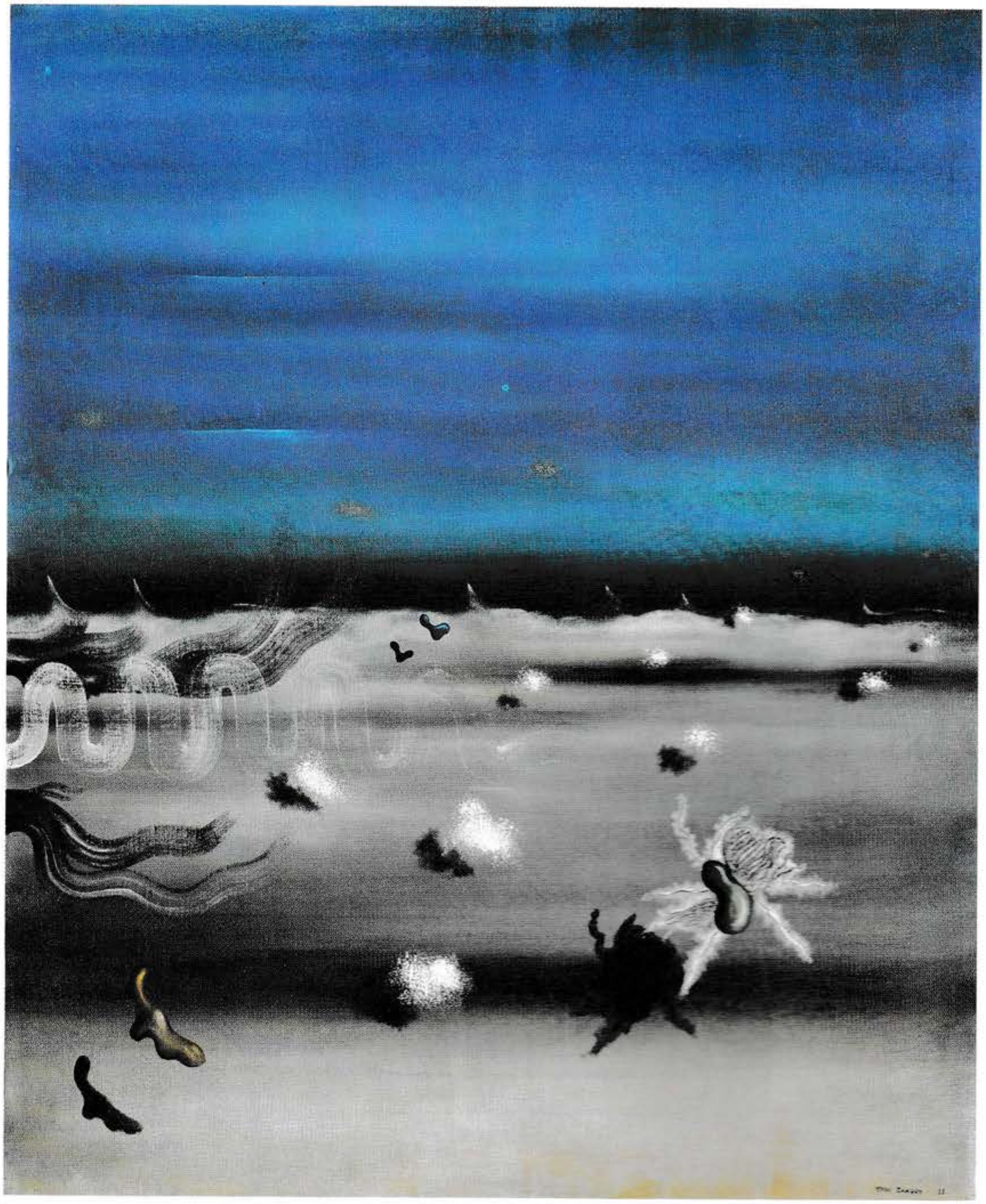
QUANTUM

JANUARY/FEBRUARY 1997

US $7.50/CAN $10.50



Springer



The Look of Amber (1929) by Yves Tanguy

THIS PAINTING ACTUALLY HAS LITTLE TO DO WITH outward appearance of the fossilized resin called "amber" in English. Even if Tanguy was unaware of its name in Greek (*elektron*), he clearly understood the role amber played in the history of electricity. The objects hovering in his dreamscape are charged with this mysterious force.

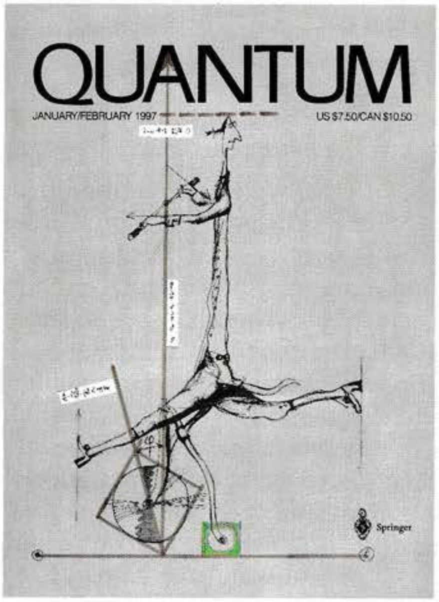
Around 600 B.C. the Greek philosopher Thales noted that amber gains the ability to attract feathers and other objects when it is rubbed with a piece of fur. Centuries later William Gilbert, who coined the word "electricity," found that glass also acquires this capacity when rubbed. In 1733 the French chemist Charles Francis de Cisternay Du Fay discovered that two charged pieces of amber repel each other, as do two charged pieces of glass, but that charged amber attracts charged glass; and if the pieces touch, they both lose their charge. This led him to speculate that there are two kinds of electrical charge, "resinous" and "vitreous."

The American scholar Benjamin Franklin thought otherwise. He suggested that a single "fluid" is involved. When glass is rubbed, it becomes "positively charged" (electricity flows into it); whereas when amber is rubbed, it becomes "negatively charged." When they touch, the fluid flows from the glass into the amber until a neutral balance is achieved. As Isaac Asimov notes in his guide to science, "If we substitute the word electrons for Franklin's 'fluid' and reverse the direction of flow (actually electrons flow from the amber to the glass), his guess was essentially correct."

Although the eighteenth century was full of electrical research (including Franklin's famous kite experiment), it wasn't until 1785 that Charles Augustin de Coulomb took the first step in quantifying electricity. In his honor the Kaleidoscope in this issue offers an assortment of shocking facts and electrifying problems.

QUANTUM

JANUARY/FEBRUARY 1997 VOLUME 7, NUMBER 3



Cover art by Yury Vashchenko

Sometimes an invention is slightly ahead of its time. At least, that's the tongue-in-cheek premise of "The First Bicycle" (page 44). This vehicle was designed and built, so the story goes, by the "great inventor Nga-Nga." The only problem was, the wheel hadn't been invented yet. The "wheels" of this first bicycle consisted of two sticks lashed together—square wheels, in effect. The ride was pretty rough, but the physics was rather interesting.

As always, there was plenty of room for improvement, and Nga-Nga's equally legendary descendants did indeed tinker with his original design. But if they eventually came up with a circular wheel, this achievement is lost in the mists of prehistory.

Indexed in *Magazine Article Summaries, Academic Abstracts, Academic Search, Vocational Search, MasterFILE,* and *General Science Source.* Available in microform, electronic, or paper format from University Microfilms International.

FEATURES

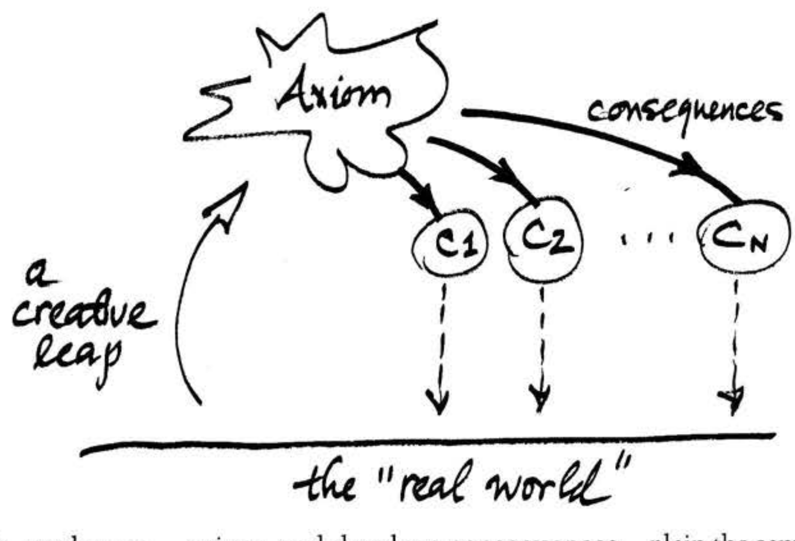
DEPARTMENTS

FRONT MATTER

The creative leap

Einstein's science

On November 25, a collection of 400 letters and papers from the Albert Einstein estate was offered at auction by Christie's in New York City. Before the event, media attention focused on Einstein's "love–hate" letters to his first wife, Mileva Maric. But it was a manuscript documenting his search for empirical proof of his theory of relativity that exceeded the expectations of the auction house (bringing in $398,500), while the Maric correspondence fell short. Somehow the media missed the point.



One interesting paper has never made it to the media's attention. A few years before his death in 1955 Einstein got a letter from a lifelong friend, Maurice Solovine. Solovine wasn't a scientist, but apparently enjoyed discussing science with his famous friend. In this exchange, Solovine wrote that he had trouble understanding a certain passage in one of Einstein's essays. The next week Einstein wrote back. First, he apologized to his friend for the confusion: it had to do with his friend's misunderstanding the structure of science and that it was his—Einstein's—fault for never having explained it carefully.

Einstein included a diagram with his explanation, which I've redrawn with labels in English.

In the text of his letter, Einstein explained that the lower, horizontal line represents the "real world." The curved line on the left signifies the creative leap somebody makes in attempting to explain a phenomenon. The leap is intuitive, and although it may be very insightful, it is not scientific. Many people take many leaps—not all are the beginning of a scientific insight. The scientific process begins, Einstein explains, when the scientist takes the idea, or axiom, and develops consequences based on it. These consequences are illustrated by a number of smaller circles connected to the axiom with lines. The next task in this process is to test these consequences against the material world. In Einstein's drawing these tests are vertical arrows returning to the world. If there is no match between the predicted consequences and the real world, the idea is (scientifically) worthless. If there is a match, the idea might have merit. With the publication of the idea, the community of scientists is brought into the process, and the original work is often modified.

There is beauty in science. If an idea is very general, having many consequences, it can replace many separate ideas and it's seen as more fundamental and thus more appealing. It's possible to construct a different explanation for each observation. For example, a scheme could be created to explain the disappearance of water from an open container, and another, unrelated idea could be employed to account for the fluidity of water, and so on. An idea about the structure of liquids (not just water) that could be used to explain these phenomena and many others would be a highly valued replacement for the collection of separate ideas. The simplicity of an idea also influences opinions about its worth. If more than one is proposed to explain the same phenomena and if they all predict the experimental results equally well, the most appealing idea is the simplest one. (In philosophy, this is known as Occam's razor.)

Einstein said it so much more succinctly in his diagram! And luckily, one doesn't need thousands of dollars to benefit from the wisdom this curious, fellow human being.

—Gerry Wheeler

Gerald F. Wheeler *is the Executive Director of the National Science Teachers Association and the Publisher of* Quantum.

Be a factor in the QUANTUM equation!

Have you written an article that you think belongs in *Quantum*? Do you have an unusual topic that students would find fun and challenging? Do you know of anyone who would make a great *Quantum* author? Write to us and we'll send you the editorial guidelines for prospective *Quantum* contributors. Scientists and teachers in any country are invited to submit material, but it must be written in colloquial English and at a level appropriate for *Quantum*'s predominantly student readership.

Send your inquiries to

Managing Editor
Quantum
1840 Wilson Boulevard
Arlington VA 22201-3000
E-mail: quantum@nsta.org

What's happening?

Summer study ... competitions ... new books ... ongoing activities ... clubs and associations ... free samples ... contests ... whatever it is, if you think it's of interest to *Quantum* readers, let us know about it! Help us fill Happenings and the Bulletin Board with short news items, firsthand reports, and announcements of upcoming events.

What's on your mind?

Write to us! We want to know what you think of *Quantum*. What do you like the most? What would you like to see more of? And, yes—what *don't* you like about *Quantum*? We want to make it even better, but we need your help.

What's our address?

Quantum
1840 Wilson Boulevard
Arlington VA 22201-3000

Be a factor in the QUANTUM equation!

QUANTUM

THE MAGAZINE OF MATH AND SCIENCE

A publication of the National Science Teachers Association (NSTA)
& Quantum Bureau of the Russian Academy of Sciences
in conjunction with
the American Association of Physics Teachers (AAPT)
& the National Council of Teachers of Mathematics (NCTM)

The National Science Teachers Association is an organization of science education professionals and has as its purpose the stimulation, improvement, and coordination of science teaching and learning.

Quantum (ISSN 1048-8820) is published bimonthly by the National Science Teachers Association in cooperation with Springer-Verlag New York, Inc. Volume 7 (6 issues) will be published in 1996–1997. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by Quantum Bureau (Moscow, Russia), as well as original material in English. **Editorial offices:** NSTA, 1840 Wilson Boulevard, Arlington VA 22201-3000; telephone: (703) 243-7100; electronic mail: quantum@nsta.org. **Production offices:** Springer-Verlag New York, Inc., 175 Fifth Avenue, New York NY 10010-7858.

Advertising:

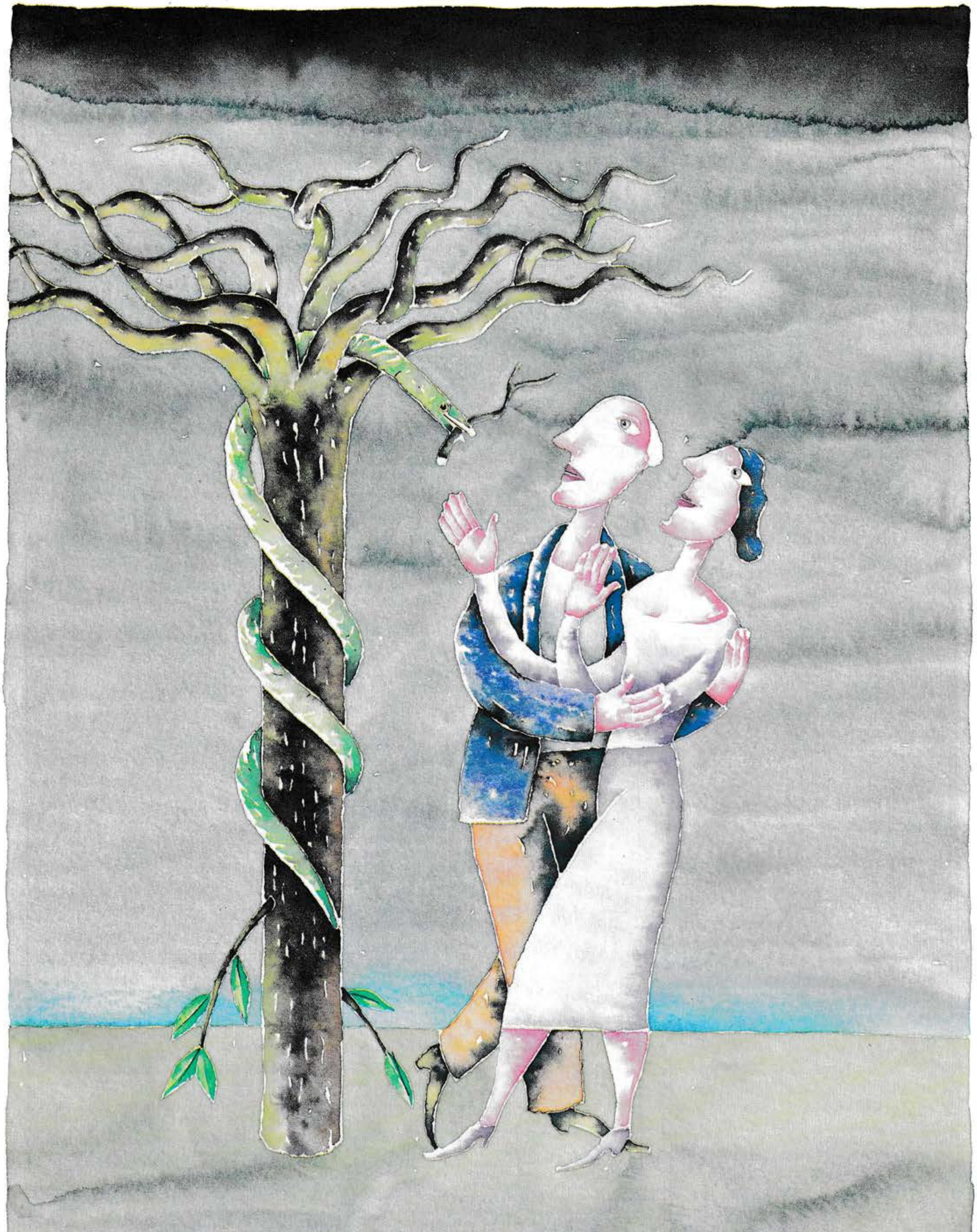
Advertising Representatives: (Washington) Paul Kuntzler (703) 243-7100; (New York) Brian Skepton (212) 460-1575; and G. Probst, Springer-Verlag GmbH & Co. KG, D-14191 Berlin, Germany, telephone 49 (030) 82 87-0, telex 185 411.

Periodicals postage paid at New York, NY, and additional mailing ofices. **Postmaster:** send address changes to: *Quantum*, Springer-Verlag New York, Inc., Journal Fulfillment Services Department, P. O. Box 2485, Secaucus NJ 07096-2485.

Subscription Information:

North America: Student rate: $17.70; Personal rate (non-student): $23.70; Institutional rate: $45.00; Single Issue Price: $7.50. Rates include postage and handling. (Canadian customers please add 7% GST to subscription price. Springer-Verlag GST registration number is 123394918.) Subscriptions begin with next published issue (backstarts may be requested). Bulk rates for students are available. Send orders to *Quantum*, Springer-Verlag New York, Inc., P.O. Box 2485, Secaucus NJ 07096-2485; or call 1-800-SPRINGER (777-4643) (in New York, call [201] 348-4033).

All Countries Outside North America: Personal rate is US$31.70 including surface delivery charges; institutional rate is US$45.00; airmail delivery is US$13.00 additional (all rates calculated in DM at the exchange rate current at the time of purchase). SAL (Surface Airmail Listed) is mandatory for Japan, India, Australia, and New Zealand. Customers should ask for the appropriate price list. Orders may be placed through your bookseller or directly through Springer-Verlag, Postfach 31 13 40, D-10643 Berlin, Germany.


S. Ivanov

Questioning answers

In every ending is a beginning . . .

by Barry Mazur

CAN YOU FIND *ALL* THE ROOTS OF A CUBIC polynomial in one variable, if someone gives you two of them? For example, can you find the third solution of the equation

$$x^3 - 9x^2 + 20x - 12 = 0$$

when you know that $x = 1$ and $x = 2$ are solutions? If so, you have all the background you need to read this article. (If you need a bit more background, see the chapter on "synthetic division" in any precalculus textbook.)

All of us who are fascinated by mathematics have the faith that math answers questions and solves problems like the one I have just asked. We work hard on some problem in mathematics because we are pursued by curiosity for the answer. Or, for the lesser reason that it was a problem that someone posed for us, perhaps as a challenge, perhaps as a test. We work hard, and then when we get "the answer" we might imagine that we can relax. I'd like to turn this picture upside-down and suggest that much of the art of mathematics only begins once we have "the answer." If we can manage, at that point, to ask the right *questions* of the answer that we have obtained, we may be led to even more interesting things.

Let's start with a frivolous-sounding question about numbers and think about new questions that its solution "invites" us to investigate. I won't try to prove things systematically, but I'll call upon you to make a few calculations at various times.

Question. The number 210 is both the product of two consecutive integers ($210 = 14 \cdot 15$) and the product of three consecutive integers ($210 = 5 \cdot 6 \cdot 7$). How many other numbers have this property of being expressible as both the product of three consecutive integers and the product of two consecutive integers?

Before I begin dealing directly with this question, let's think a bit about its nature. Why did I choose it? What do I have up my sleeve? Before reading on, try to find some other numbers that are the products of two and of three consecutive integers.

Our question, of course, can be rephrased as an algebraic problem. Think of our number N as the product of three consecutive integers. Let X be the middle one of those three integers, and we have

$$N = (X - 1) \cdot X \cdot (X + 1) = X^3 - X.$$

Now think of N as the product of two consecutive integers. Let Y be the smaller of those two integers, so we get

$$N = Y \cdot (Y + 1) = Y^2 + Y.$$

We have a solution "N" to our problem, then, every time we can find a pair of integers $[X, Y]$ having the property that

$$Y^2 + Y = X^3 - X.$$

We're faced, then, with an equation in two variables X and Y whose highest-degree term is a cube, and we're looking for solutions to this equation in whole numbers. We can visualize the real solutions to this equation as a curve C in the (X, Y)-plane (fig. 1). The graph of our

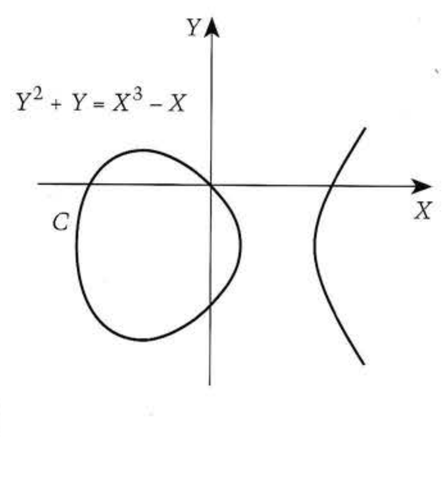


Figure 1

For Grace and Nick.—*Auth.*

Art by Sergey Ivanov

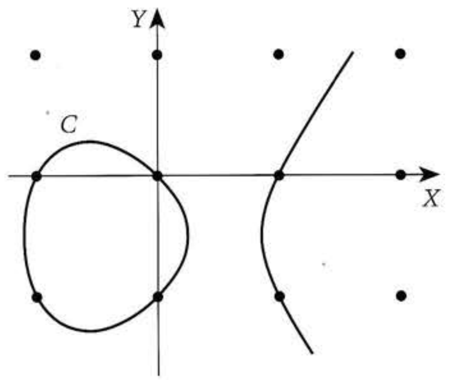


Figure 2

equation consists of two disjoint pieces. Nonetheless, we'll consider it a single "curve" (this is how mathematicians speak) and name the curve C (this is how mathematicians write).

We must remember that the graph we have just drawn, falling into two pieces as it does (the left-hand piece being oval-shaped and looking somewhat like a cartoon drawing of an eggplant, and the right-hand piece looking like an unstrung, infinite bow), traces out all *real* solutions to our equation. But we're looking for points on the curve C with *integral* (X, Y)-coordinates. So let's superimpose onto this graph the infinite array of points in the (X, Y)-plane that have *integral* (X, Y)-coordinates (these points are drawn as "big dots" in figure 2).

Having drawn this graph, we already "see" six solutions. But to "weigh" the problem confronting us, I want to pause to compare this equation with a much simpler equation that is the "standard fare" of high school algebra and is familiar to all of us.

The quadratic equation in one variable

Find the values of the variable X that "solve" the quadratic equation

$$a \cdot X^2 + b \cdot X + c = 0.$$

We all know the gambit here—an idea that has come down to us from Babylonian times: if we want to find numbers X that solve this equation (and we might be interested in integer solutions X, or rational numbers X, or real numbers, or complex numbers), we "complete the square" by rewriting this equation as

$$a \cdot \left(X + \frac{b}{2a}\right)^2 + \left(c - \frac{b^2}{4a}\right) = 0.$$

Problem 2. Check that this version of the equation is equivalent to the first.

This rewritten equation *visibly* has (at most) two beautiful solutions given by the quadratic formula. Of course, if we are specifically interested in integer solutions, or rational solutions, we must check whether the answers given by our quadratic formula are integers, or rational, and so on. I said "of course" in the last sentence, but I should remind you that this issue of whether or not the answers "given by our quadratic formula are integers, or rational" was historically, at least, not such a humdrum affair. For example, the fact that $X^2 - 2 = 0$ has *no* rational solution in X (that is, the fact that the square root of 2 is irrational) was viewed as devastating by the Pythagorean mathematicians who initially made this discovery. The irrationality of the square root of 2 was considered such a dark secret about the universe that when one of them revealed it to outsiders, the story goes, he was murdered as a betrayer.

This may be so, or may not be so, but one thing is certain: for the quadratic equation there are *at most* two solutions, and for any polynomial equation in one variable X of degree d there are *at most* d solutions.[1] One of the fascinations of the type of problem posed by the equation $Y^2 + Y = X^3 - X$ that we're considering is that we don't even have any idea *how many solutions to expect!*

How many solutions have you found?

Let's start with the six "easy" solutions to $Y^2 + Y = X^3 - X$ that are so modest, you might have overlooked them at first, if we hadn't discovered them by drawing our graph:

$$X = 0,\ Y = 0;$$
$$X = 0,\ Y = -1;$$
$$X = \pm 1,\ Y = 0;$$
$$X = \pm 1,\ Y = -1.$$

All of these solutions give $N = 0$:

$$\begin{aligned} 0 &= 0 \cdot 1 \cdot 2 = 0 \cdot 1 \\ &= 0 \cdot 1 \cdot 2 = (-1) \cdot 0. \end{aligned}$$

Did you get this relatively modest answer to our question 1? We'll come back later to the issue of exactly how modest or immodest these solutions are.

Did you also discover

$$X = 2,\ Y = -3;$$
$$X = 2,\ Y = 2?$$

These solutions give $N = 6$:

$$\begin{aligned} 6 &= 1 \cdot 2 \cdot 3 = (-3) \cdot (-2) \\ &= 1 \cdot 2 \cdot 3 = 2 \cdot 3. \end{aligned}$$

And then there are the solutions that were given in the statement of the problem itself:

$$X = 6,\ Y = -15;$$
$$X = 6,\ Y = 14.$$

These solutions give $N = 210$:

$$\begin{aligned} 210 &= 5 \cdot 6 \cdot 7 = (-15) \cdot (-14) \\ &= 5 \cdot 6 \cdot 7 = 14 \cdot 15. \end{aligned}$$

[1]This statement can be proved, for example, by induction on the degree of $f(x)$ and by using the "long division of polynomials."

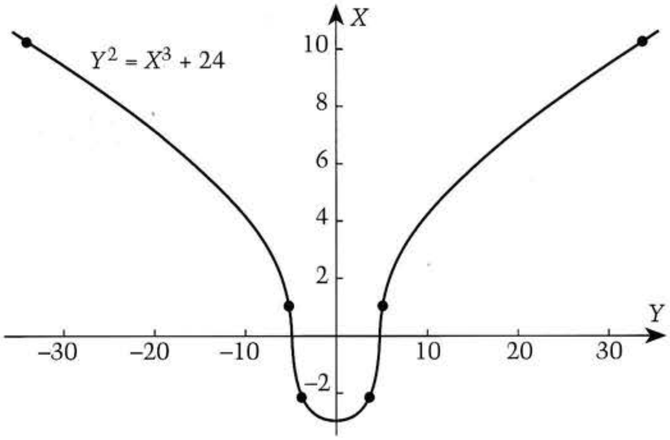


Figure 3

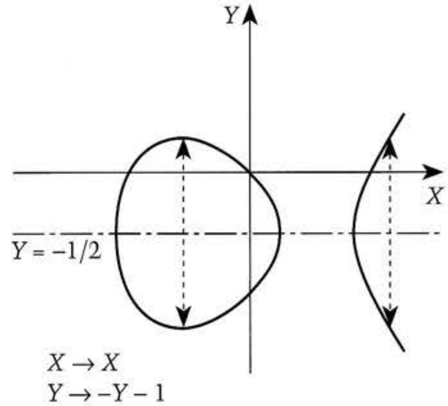


Figure 4
The "basic symmetry."

Did you find any others? Suppose you had tried out all numbers under a million and had found no further solutions. Would you then be confident that there were no others? Now, although "confidence" is a precious virtue that counts for a lot in mathematical work and in life, my recommendation in exactly this sort of calculation is that you should *not* be confident that you have found all the solutions. Let me illustrate why by bringing up a slightly different problem.

Problem. Find the integer solutions to the equation

$$Y^2 = X^3 + 24.$$

That is, find the perfect squares ("Y^2") that are 24 more than a perfect cube ("X^3"). Now, you'll surely guess a few of the solutions quite easily. For example, $X = -2$, $X = 1$, and $X = 10$ give solutions to this problem:

$$4^2 = (-2)^3 + 24;$$
$$5^2 = 1^3 + 24;$$
$$32^2 = 10^3 + 24.$$

But these are *not* all: there is one missing value of X that solves the equation, and if we wanted to plot the value of X on the graph in figure 3, we'd have to extend the "wingspan" of the graph[2]—from the three and a half inches it takes up on this page to a diameter of over half a mile!

A "basic symmetry"

The first thing that jumps to the eye, given the solutions of $Y^2 + Y = X^3 - X$ that we have already found, is that these solutions come in pairs, each pair giving the same value for N. In fact, the entire curve C is mapped onto itself by the "symmetry"

$$X \to X;\ Y \to -Y - 1$$

(see figure 4). In other words,

$$P = [X, Y] \leftrightarrow \overline{P} = [X, -Y-1],$$

and so we have

$$[1, 0] \leftrightarrow [1, -1];$$
$$[2, 2] \leftrightarrow [2, -3];$$
$$[6, 14] \leftrightarrow [6, -15];$$

and so on.

Using this symmetry, we can manufacture new solutions of our equation from old ones: given the solution $[1, 0]$ we can apply the symmetry to "discover" the solution $[1, -1]$, and so on. These discoveries are not too exciting, since the symmetry $[X, Y] \leftrightarrow [X, -Y - 1]$ is so elementary. But are there other geometric properties of the graph of our equation that we can use to force "old solutions" to somehow lead us to new ones?

Collinear points

I now want to use a geometric property of our curve C that is much subtler than the symmetry we have just discussed:

Any line L in the (X, Y)-plane intersects the curve C in at most three points.

The proof of this is easy: plug into our equation

$$Y^2 + Y = X^3 - X$$

the equation $Y = mX + b$ of the line L and solve for X. This gives us a cubic polynomial in the variable X that can have at most three solutions (see figure 5)!

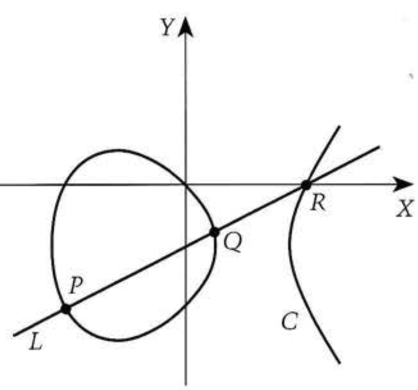


Figure 5

[2]Note that the axes are not the usual ones in figure 3. We have chosen the y-axis to be horizontal and the x-axis vertical.

Here now is a strategy that might potentially produce new solutions from old ones. It has some pitfalls, but first I'll try to formulate it. There will be time to criticize it later.

Let $P = [X_1, Y_1]$ and $Q = [X_2, Y_2]$ be two rational solutions of the equation $Y^2 + Y = X^3 - X$. Let $L = PQ$ be the straight line passing through P and Q. For obvious reasons we'll call L the *chord to the curve C passing through P and Q*. Consider the intersection of the line L and the curve C. There is at most one other intersection point R of L and C. Solving for R gives a "new" solution $R = [X_3, Y_3]$ of our equation.

Problem 3. If P and Q have integer coordinates, must R have integer coordinates?

This strategy gives us a new solution, but it seems to depend on having *two distinct solutions P and Q* so as to be able to produce a chord L passing through them.

Question. Without reading ahead, can you dream up a "natural" way of extending this strategy of "finding a third point of intersection" to make it work even when *"the two points P and Q are equal"*?

A bit of thought will suggest that yes, there is a natural extension of our strategy that allows us to work with one point "$P = Q$": take the line L to be simply the *tangent line* to the curve C at point P (fig. 6). This strategy of getting new solutions from old, by the way, is sometimes referred to as the *chord-and-tangent process*.

Let's look into this strategy in a little more detail. If we have the points P and Q in hand, we can very easily calculate R. Let's try an example. Take as our two points $P = [1, -1]$ and $Q = [2, 2]$. The line L passing through P and Q has the equation $Y + 1 = 3(X - 1)$. So, plugging this equation for Y into our equation $Y^2 + Y = X^3 - X$, we get a cubic equation in X:

$$Y^2 + Y = X^3 - X,$$
$$(3X - 4)^2 + 3X - 4 = X^3 - X,$$

or

$$X^3 - 9X^2 + 20X - 12 = 0.$$

Now, $X = 1$ and $X = 2$ are solutions of this equation

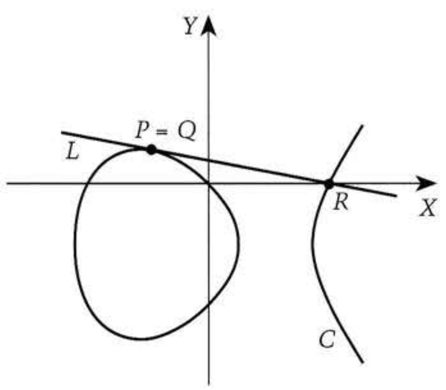


Figure 6
If P = Q, then L should be chosen to be tangent to C at P.

corresponding to the intersection points P and Q on the line L. If you answered the question at the beginning of this article, you have solved this cubic equation and know that its *third* solution is

$$X = 6.$$

Then, since $Y = 3X - 4$, we get $Y = 14$. That is, our third intersection point R of the curve C with the line L is

$$R = [6, 14].$$

The moral here is that, if we had *not* discovered this solution [6, 14] "on our own," we might perfectly well have been led to its existence by this strategy of finding new solutions from old . . . if we had previously obtained the solutions [1, –1] and [2, 2]. To put it another way, the chord-and-tangent process is a strategy that forces "old" solutions to "work for us" to possibly produce other solutions.

But there are a few tricky things about this strategy. The first tricky thing, which is minor, is that *for some choices of P and Q there is no third intersection point R*. This happens if and only if the line L is vertical. This is the same as saying that the X-coordinate of the points P and Q are equal. And this is the same as saying that P and Q are brought to each other by the symmetry of the curve C.

The second tricky thing, however, opens up a whole new issue: *sometimes the point R does not have integer coordinates, but only has rational coordinates.* That is, sometimes the X- and/or Y-coordinate of the new solution is a fraction rather than (what we want!) a whole number. You don't have to go far to run into this. For example, take $P = [1, 0]$ and $Q = [6, 14]$. Then the line passing through P and Q has the awkward equation $5Y = 14(X - 1)$, which gives us a denominator of 5 when we solve for Y:

$$Y = \frac{14}{5} \cdot (X - 1).$$

When we proceed as before and finally get the third intersection point of L and C, we find this point R to be

$$R = \left[\frac{21}{25}, -\frac{56}{125}\right].$$

In a word, our strategy does not preserve *integrality* of the solutions that are found, but does preserve *rationality*. Let's examine our arguments to see that in fact R has rational coordinates whenever P and Q have rational coordinates. Indeed, where can an irrationality creep in? It's easy to see that the slope and Y-intercept of the line through P and Q are rational (as long as the line isn't vertical, of course). Then we can express this line in the form $y = mX + b$ and plug this expression into the equation of our curve to get a cubic equation in X. We already had two (rational) solutions to this equation.

If you examine the division process that allows you to calculate the third root, you'll see that it involves only the operations of addition, subtraction, multiplication, and division. If we start with rational numbers, these operations can only give us rational answers. So the X-coordinate of the point R is rational. Since R lies on the line $Y = mX + b$, its Y-coordinate must then also be rational.

If we are going to make any systematic use of this strategy, we are curiously led to consider all *rational solutions* of the equation $Y^2 + Y = X^3 - X$ and then to think of the integral solutions (which we were originally after!) as a particular subcollection of rational solutions. You might think that this is a step backward in that there are, very likely, many more rational solutions than there are integral solutions, and therefore our chore is much, much harder. We'll follow this path to see where it leads. By a rational solution P to the equation $Y^2 + Y = X^3 - X$ we just mean a pair of rational numbers $P = [x, y]$ that "solve" that equation. If we want to think geometrically, we can also call P a *rational point* on C.

A theorem

Here is the surprising answer to the question: "What are all the rational solutions to $Y^2 + Y = X^3 - X$?" There are infinitely many rational solutions. Nevertheless, don't despair! The magic here is that you can get *all* rational solutions if you start with the single solution $P = [0, 0]$ and then go on to produce "new solutions from old" just by systematically applying the basic symmetry $P \rightarrow \overline{P}$ and the chord-and-tangent process to all pairs of points you get along the way.

"Nothing will come of nothing," according to King Lear, but as for *our* problem, the modest "double-zero" solution $[0, 0]$ to the equation $Y^2 + Y = X^3 - X$ *generates all the infinitely many rational solutions* by the chord-and-tangent process. I know of no better or more pleasant way to get a feeling for the mathematics behind the chord-and-tangent process, and to learn how effective it is in generating rational solutions to cubic equations in two variables, than to start reading the book *Rational Points on Elliptic Curves* by J. Silverman and J. Tate (Springer-Verlag, 1992).

You might imagine that constant application and reapplication of our chord-and-tangent process to the solutions of a cubic equation in two variables will lead to some unholy mess of solutions with disorganized heaps of them generated by our process, which works away with a sorcerer's apprentice–like zeal. But no: the end result of this process is amazingly "organized"—a minor miracle! Going back to our equation $Y^2 + Y = X^3 - X$, I'll restate it as a theorem (whose proof is, I must admit, *not* elementary!).

THEOREM. *There are infinitely many rational solutions to the equation*

$$Y^2 + Y = X^3 - X.$$

The "double-zero" solution $P_1 = [0, 0]$ generates all rational solutions via the chord-and-tangent process. There is a unique way of "listing" all these rational solutions by labeling them in a one-to-one correspondence with the set of all nonzero (positive and negative) integers

$$1 \leftrightarrow P_1 = [0, 0]$$

$$\vdots$$

$$n \leftrightarrow P_n = [x_n, y_n]$$

such that this one-to-one correspondence has these properties:

A. P_n is brought onto P_{-n} by the "basic symmetry. That is,

$$x_n = x_{-n} \text{ and } y_n = -y_{-n} - 1.$$

B. Three rational points P_n, P_m, and P_r (whose indices are distinct nonzero integers n, m, and r) lie on a straight line L in the (X, Y)-plane if and only if $n + m + r = 0$.

Here is the beginning of this listing for positive values of n (to get the listing for the corresponding negative values, just apply the "basic symmetry"—that is, replace the Y-coordinate with $-Y - 1$):

$P_1 = [0, 0]$
$P_2 = [1, 0]$
$P_3 = [-1, -1]$
$P_4 = [2, -3]$
$P_5 = [1/4, -5/8]$
$P_6 = [6, 14]$
$P_7 = [-5/9, 8/27]$
$P_8 = [21/25, -69/125]$
$P_9 = [-20/49, -435/343]$
$P_{10} = [161/16, -2065/64]$
$P_{11} = [116/529, -3612/12167]$
$P_{12} = [1357/841, 28888/24389]$
$P_{13} = [-3741/3481, -43355/205379]$
$P_{14} = [18526/16641, -2616119/2146689]$
$P_{15} = [8385/98596, -28076979/30959144]$
$P_{16} = [480106/4225, 332513754/274625]$
$P_{17} = [-239785/2337841, 331948240/3574558889]$
$P_{18} = [12551561/13608721, -8280062505/50202571769]$
$P_{19} = [-59997896/67387681, -641260644409/553185473329]$
$P_{20} = [683916417/264517696, -18784454671297/4302115807744]$

You can check my arithmetic,[3] because there are quite a number of miraculous constraints on this list of numbers: they are all solutions to $Y^2 + Y = X^3 - X$, but also if you take any pair of distinct integers n, m, by part B of the theorem we formulated we find that P_n, P_m, and $P_{-(n+m)}$ must lie on a straight line in the (X, Y)-plane. For example, since $3 + 5 + (-8) = 0$, the three points $P_3 = [-1, -1]$, $P_5 = [1/4, -5/8]$, and $P_{-8} = [21/25, -56/125]$ had better be collinear. Otherwise I made a mistake in compiling my list!

I'd like to pause a minute now to squint at the list of solutions P_n above. If we are to be as attentive as we

[3]I confess that this was *not* done by hand. I used the very convenient computer package PARI to compile this.

possibly can be to the answers we get to the questions we ask, there's something about that list—its general shape—that shouldn't escape our notice! Do you see a wiggly profile of a *parabola* hidden in it? To bring this out more clearly, let's compact our data a bit and consider this slightly more extensive list, where I give only the absolute values of the numerators of the X-coordinates of P_n for odd values of n, beginning with $n = 9$:

20
116
3741
8385
239785
59997896
1849037896
270896443865
16683000076735
2786836257692691
314892968128574031 6
34211575692760792742 0
2802511299225632914226 45
80428751803514156523619315 1
743043134297049053529252783151
3239336802390544740129153150480400
2613390252458014344369424012613679600
125187370946712398266830319435831525503 51
5969295654077588460781578504779882298363403 51
238585858632982963160807755393813926443135201015 5
5618605401843475352702275238228029188204880958285738 0
2389750519110914018630990937660635435269956452770356625916
65008789078766455275600750711306493793995920750429546912218291
863381503588680671392136126345657274078403806591767431591377541753 5
4327678343894888631258803040444144431340575553436625441643288092401906 5
59307604546964265894895676173979432448272923468711451231872777328587667138 9

Now take another look to check out the clean shadow of a parabola formed by the mere digits of our solutions. This shadow is a vivid indication that the rate of growth of the size of our solutions seems to be following a regular pattern. Can we prove this? Can we "question our answers" so rigorously as to get (heaven help us!) the equations of that parabola? Will doing this lead us to an even deeper understanding of the arithmetic behind our original problem? The answer to all these questions is *yes*, and following their lead would bring you into intimate contact with a good deal of the exciting work in number theory that has been taking place in this half-century!

Rational vs. integral solutions

Somehow we have wandered into the dazzling infinite array of *rational solutions* to our equation $Y^2 + Y = X^3 - X$, when at the outset of our investigation we had intended to study only the *integral solutions*. We did this because, infinite or not, the rational solutions to the problem have a certain orderly structure that was not in evidence when we focused only on integral solutions. Can we now go back and pick out the jewels—the integral solutions—in our infinite list? *What, in fact, is the actual answer to our original question?* That is, what are the numbers N that can be written both as a product of two consecutive and of three consecutive integers?

The answer to the first question is yes. It was apparent to Mordell half a century ago that we can indeed find all the integral solutions, and there is an elegant way to

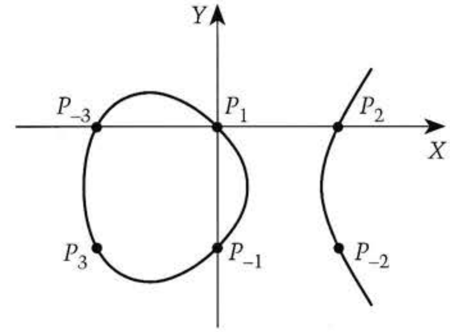


Figure 7

do this by "picking them out of" the infinitude of rational solutions. The answer to our original question is that $N = 210$ is the largest number that can be written as a product of two consecutive integers and also as a product of three consecutive integers.

Although the steps in the proof of this last statement are more involved than the mathematics of the rest of this article, I feel compelled to provide the bare bones of the argument. To clinch our problem, there are three things that we need to calculate.[4] Let's return to the graph of our equation in the (X, Y)-plane, which breaks into two pieces: the oval on the left and the piece "going off the page to infinity" on the right (fig. 7).

1. First check that the only integral points on the oval are the four integral points $P_1 = [0, 0]$, $P_{-1} = [0, -1]$, $P_3 = [-1, -1]$, and $P_{-3} = [-1, 0]$.

2. Check that the points P_n with *odd index n* lie in the oval on the left, while the points P_n with *even index n* lie in the piece going off to infinity on the right.

3. Check that any prime number dividing the denominators of the X- and Y-coordinates of P_n also divide the denominator of the X- and Y-coordinates of P_{2n}.

Suppose you have done these three chores, and suppose you know the theorem we formulated earlier and the list of P_n for small n. Equipped with all these facts, you are now in a position to prove the following theorem.

THEOREM. *The only numbers N that are both the product of two consecutive integers and three consecutive integers are N = 0, 6, and 210.*

Proof. We'll search among the infinite list P_m of rational solutions to $Y^2 + Y = X^3 - X$ to see which of these is integral—that is, which has the property that neither the X- nor the Y-coordinate has a denominator greater than 1. If a solution P_m is integral, then its image P_{-m} under the "basic symmetry" is also integral, so in searching for all integral solutions we can just try to determine all positive values of m for which P_m is integral. Write $m = 2^e \cdot m_0$, with m_0 odd and $e \geq 0$. Use fact 3 to see that if P_m has

CONTINUED ON PAGE 27

[4]The first of these you might try to do right now. The second and third would be something you'd naturally try your hand at while reading Silverman and Tate's *Rational Points on Elliptic Curves*.

Wobbling nuclear drops

"But if within your heart there should remain
The smallest drop of pity for my pain . . ."
—from Tatyana's letter to Onegin
(Eugene Onegin, *Alexander Pushkin)*

by Yuly Bruk, Maxim Zelnikov, and Albert Stasenko

WHEN WE THINK ABOUT drops, we usually think about rain, or fog, or the leaky faucet in the kitchen. However, the "liquid-drop model" is applied in physics to describe the life of atomic nuclei. You may find this surprising. What can droplets in a fog have in common with the nucleus of a uranium atom? Quite a lot, actually. The liquid-drop nuclear model developed by Niels Bohr, John Wheeler, and Y. I. Frenkel is very similar to the model of a drop of an ordinary liquid. On the basis of this model we can understand how nuclei oscillate, and we can also determine the masses of the nuclei. In this article we'll talk about the oscillations of drops, and we'll begin in a fog.

Drops in a fog

Assume that spherical drops are suspended motionless in the air. This experiment would be more accurate, of course, inside a spacecraft where the drops are weightless and don't fall to the ground. However, the small droplets dispersed in the Earth's atmosphere can be viewed as almost motionless because their weight is practically counterbalanced by the resistance (viscosity) of the air, and so the drops fall with a very small and constant velocity. In addition, the flows of air (for example, caused by convection) can slow the droplets and even cause them to stop falling (they can also force the droplets to move faster, and not only downward but upward as well).

We'll let our droplets "hover" in the air and have a spherical shape. Why must they be spherical and not, say, cubical, pyramidal, or flat as a pancake? Because of surface tension: any attempt to change the shape of the drop generates forces that try to restore its original shape. For our purposes water can be treated as an incompressible liquid. Since nature tries to do everything on a rational, economical basis, a droplet at equilibrium must have minimal energy associated with its surface tension. As a droplet is deformed (its volume remains constant due to incompressibility), its surface area varies. The minimal surface area for a body with constant volume is achieved when it is a sphere. Thus we come to the conclusion that a deformed droplet, when it tries to restore its equilibrium spherical shape, is striving to achieve the minimal surface (potential) energy.

Surface tension

Now let's recall how surface tension is defined in physics. Consider a light wire frame (fig. 1) with side CD that can move under the action of force $\mathbf{F}$. Assume that the area $ABCD$ is filled with a liquid film (a soap solution can be used to conduct such an experiment) and that the frame lies on the surface of a jar filled with liquid. Moving the wire CD in the direction of force $\mathbf{F}$, we perform work and increase the film's surface. The additional soap solution comes from the jar under the frame. The surface tension can be defined either by the ratio of the force $\mathbf{F}$ to the wire's length l, or by the ratio of the change in the film's energy in the rectangle $ABCD$ to the change in the area of this rectangle (caused by moving the side CD); or, in other words, the potential energy of a unit area of film. This is also confirmed by the dimensions for surface tension:

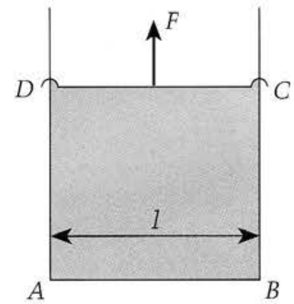
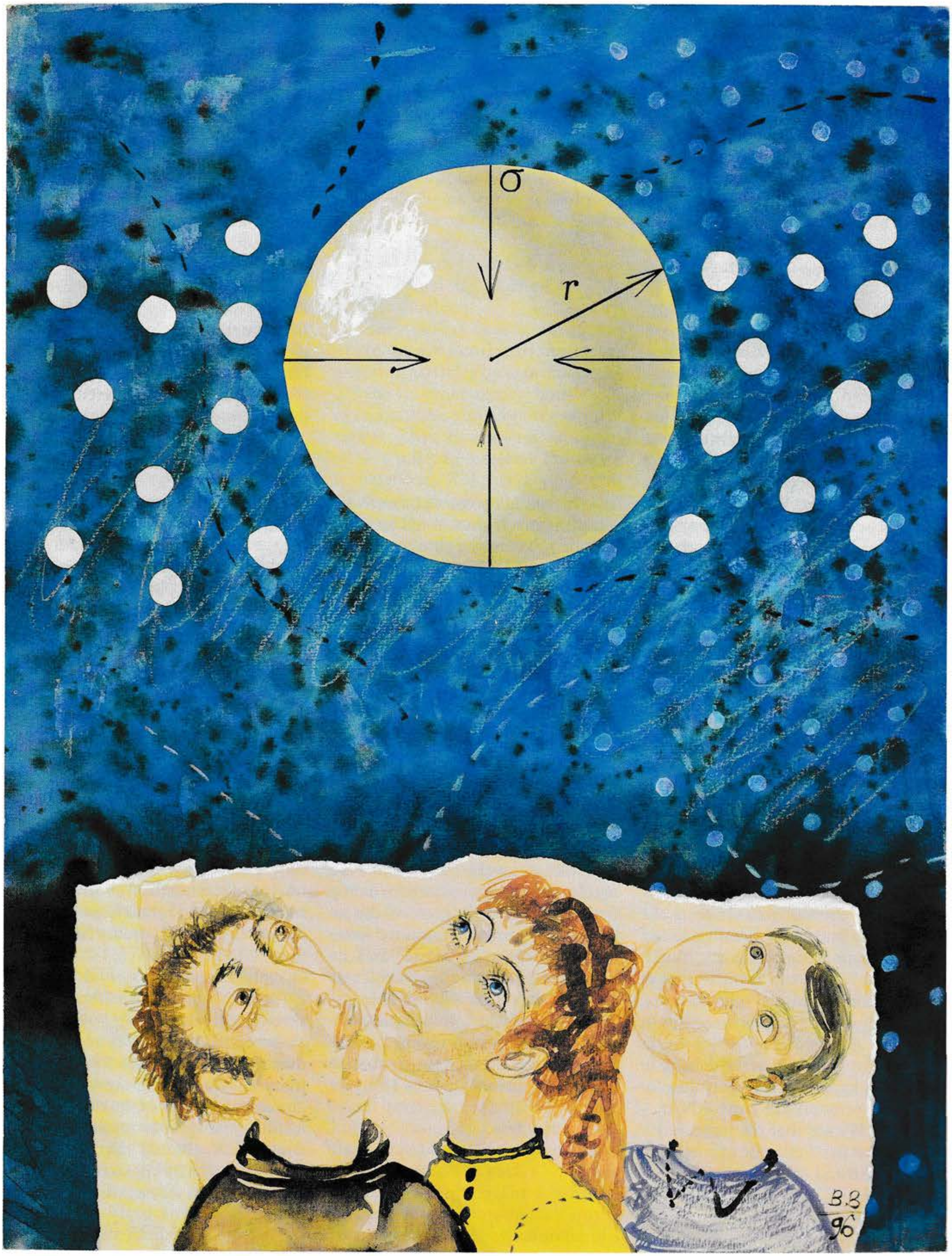


Figure 1

$$[\sigma] = \frac{\mathrm{N}}{\mathrm{m}} = \frac{\mathrm{N \cdot m}}{\mathrm{m^2}} = \frac{\mathrm{J}}{\mathrm{m^2}}.$$

Art by Vasily Vlasov


σ
r
B.B
96

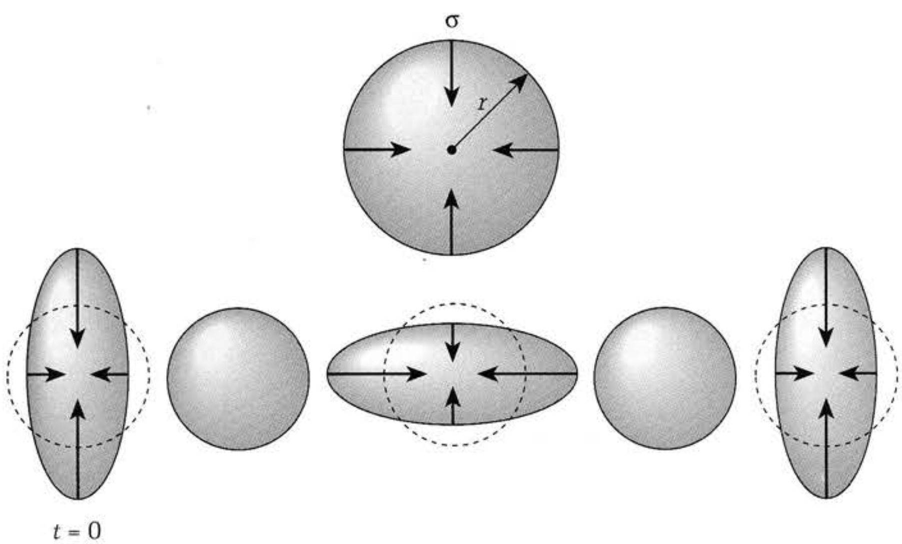


Figure 2

Now assume that initially the droplet was deformed due to, say, a clap of thunder or the shock wave produced by a supersonic jet. What will happen next? The drop begins to oscillate (fig. 2). But why does it "slide" through and past the equilibrium point—that is, the spherical shape shown by the dotted line? Evidently this is caused by the inertia of the moving particles in the droplet: after they acquire a velocity due to the action of surface tension, they just can't stop at the moment the drop assumes a spherical shape. They continue their motion and thus transform the drop from a stretched one to a squashed one. These oscillations persist until the water's viscosity and the resistance of the surrounding air transform the energy of the initial deformation into heat.

Frequency of a drop's oscillations

Now we're ready to estimate the frequency ν of a drop's oscillation and the corresponding period $T = 1/\nu$. To this end we use our favorite tool of dimensional analysis[1] with SI units. The radius r of a drop is measured in meters, and its mass $M = {}^4\!/_3\pi r^3\rho_0$ in kilograms (where ρ_0 is the density of water in kg/m^3). The dimensions for surface tension are J/m^2 = N/m = kg $\cdot$ s^{-2}. We see that the frequency ν measured in s^{-1} must be expressed in terms of σ and M as

$$\nu \propto \sqrt{\frac{\sigma}{M}}.$$

This is the only possible formula for our problem. Dimensional analysis cannot produce the dimensionless numerical coefficient in the frequency formula, so instead of the equal sign we use the proportionality symbol. In principle, we could write this formula in another way:

$$\nu \propto \sqrt{\frac{\sigma}{\rho_0 r^3}}.$$

Physicists usually take a risky step next—they assume that the numerical coefficient is equal to one. (In many "reasonable" cases it is indeed close to 1). Now we can obtain an estimate for the frequency of the oscillations. A typical size for a drop in a fog is $r \sim 0.1$ mm $= 10^{-4}$ m, the coefficient of surface tension for water $\sigma \cong 70 \cdot 10^{-3}$ N/m $= 0.07$ J/m^2, and the density of water $\rho_0 = 10^3$ kg/m^3. These values yield $\nu \sim 10^4$ s^{-1} = 10 kHz.

One can't help noting an interesting comparison. Let's rewrite our formula for the oscillation frequency using the relationship $T^2 = 1/\nu^2$:

$$T^2 \propto \frac{\rho_0 r^3}{\sigma},$$

or

$$\frac{T^2}{r^3} \propto \frac{\rho_0}{\sigma}.$$

Because ρ_0 and σ are constant for a particular liquid, we have the relationship $T^2 = \text{const} \cdot r^3$. Considering two droplets of the same liquid with different diameters, we can say that the ratio of the squares of the periods of their oscillations is equal to the ratio of the cubes of their radii. Sounds familiar, doesn't it? Of course—this is Kepler's third law, which describes the motion of planets around the Sun! True, this law deals with the orbits of the planets, not their radii. Use your gray cells, as the legendary Monsieur Hercule Poirot taught us, and perhaps you'll find something useful in this analogy!

Nuclear droplets

Now let's turn our attention to drops of nuclear liquid. The nuclei of atoms consist of nucleons—that is, protons and neutrons. Denoting the masses of protons and neutrons by m_p and m_n, respectively, the electric charge and mass number (that is, the total number of nucleons in the nucleus) by Z and A, the nuclear mass by $M(A, Z)$, and speed of light by $c = 2.998 \cdot 10^8$ m/s, we come up with the relationship

$$\Delta W = (Zm_p + (A - Z)m_n - M(A, Z))c^2,$$

known as the *nuclear binding energy* for the nucleons forming a nucleus. Nucleons are held tightly in the nucleus by the *strong interaction*. If we want to break apart a nucleus, we must expend energy equal to ΔW. Dividing ΔW by the total number of nucleons A, we get the binding energy per nucleon that keeps the nucleons in a nucleus: $\varepsilon = \Delta W/A$. For most stable and rather heavy nuclei the value of ε is approximately the same. More precisely, $\varepsilon(A)$ increases sharply from $\varepsilon = 0$ at $A = 1$ to $\varepsilon = 8$ MeV at $A = 16$, then passes through a maximum $\varepsilon_{max} = 8.8$ MeV at $A \sim 60$ (^{58}Fe and ^{62}Ni), and finally decreases slowly to $\varepsilon = 7.6$ MeV for uranium. We'll be dealing with heavy atoms only, so let's take the average binding energy per nucleon to be $\bar{\varepsilon} = 8$ MeV. (Recall that 1 MeV = 10^6 eV = $1.6 \cdot 10^{-13}$ J). So for our first approximation we can assume that

$$\Delta W \cong \bar{\varepsilon} A = 8A \text{ MeV}.$$

[1] See "The Power of Dimensional Thinking" in the May/June 1992 issue of *Quantum*.—*Ed.*

This accuracy is quite sufficient for the estimates we'll be making below.

The atomic nucleus is in some respects similar to an ordinary liquid. Just as with many liquids, the nuclear liquid is almost incompressible. This means that the density of the nuclear matter is practically constant in all kinds of nuclei. The fact that ΔW is proportional to A can be compared with the linear dependence of the energy needed to vaporize a drop of ordinary liquid on the drop's mass. The approximate constancy of the binding energy per nucleon in all nuclei also recalls a similar property of ordinary liquids.

The mass of any nucleus, and therefore its mass number A, are proportional to its volume. If R is the nuclear radius, $A \propto R^3$. Experimental data for different nuclei show just such a dependence $R \propto r_0 A^{1/3}$, where $r_0 = 1.5 \cdot 10^{-15}$ m. Qualitatively, such a dependence can be understood by assigning a "radius" r_0 to the nucleon; then for the nucleonic "balls" that are tightly packed within a sphere of radius R we get $R^3 = r_0^3 A$. The concentration of nucleons in a nucleus is $n = A/(^4/_3 \pi R^3) \sim 10^{44}$ nucleon/m^3; the mass of a nucleon is $m_p \sim m_n \sim 1.67 \cdot 10^{-27}$ kg; and the density of the nuclear matter is $\rho = nm_n \sim 2 \cdot 10^{17}$ kg/m^3.

At this point we should warn you that one can speak about the "radius" of elementary particles only with certain reservations, and that we mustn't treat these terms too literally. The "size" of nucleons, electrons, and other elementary particles cannot in principle be correctly determined within the framework of classical physics. This doesn't prevent us from obtaining estimates using "visual" notions from classical physics. Still, when we say certain words we mustn't forget that they must be justified by the more correct nonclassical considerations. For example, the sizes of atomic nuclei are determined by experiments in which elementary (and compound) particles are scattered in thin metal foils (recall Rutherford's experiments, which resulted in the planetary model of the atom).

When we think about how to determine the surface tension in a droplet of nuclear matter, we first should note that this droplet (the atomic nucleus) must contain many nucleons—otherwise the notion of surface energy becomes pointless. This is why our reasoning holds true for heavy nuclei ($A \gg 1$, $Z \gg 1$). Some nucleons (protons and neutrons) are located at the surface of the nuclear droplet, so their bonds with other particles are weaker than those of nucleons located deep inside the nucleus.

The nuclear forces acting on nucleons are similar to the forces affecting particles in an ordinary liquid in one other aspect. Both kinds of forces have the property of saturation. This means that each particle in a liquid interacts only with its neighbors. It is this property that causes the proportionality between ΔW and A. Were it not so—that is, if each nucleon interacted with all of its siblings in the nucleus—the total binding energy would be proportional to $A(A-1) \sim A^2$. The saturation of nuclear or molecular forces is related to their short ranges.

Along with nuclear forces, there is another kind of interaction within atomic nuclei. This is the familiar electric (Coulomb) repulsion between positively charged particles (protons). The forces of repulsion decrease the total nuclear binding energy. These electric forces act at large distances and do not become saturated. In other words, each proton interacts with all of its siblings, so the Coulomb contribution to the total binding energy is proportional to $Z(Z-1) \sim Z^2$. Keep in mind that the Coulomb interaction of protons in atomic nuclei differs from that of charged particles in plasma or a solid body, because the atomic nucleus contains only positively charged particles (protons), while plasma and metal have charged particles of both kinds (protons and electrons, positive and negative ions). The opposite charges interact in such a way as to "screen" each other. This is why the electric interaction operates over relatively short distances in metals or plasma. As for protons within an atomic nucleus, they have no screening countercharges, which is why the nuclear interaction operates at greater distances there.

Laplace pressure

We're still two steps short of obtaining the expression for the surface tension of a nuclear droplet. First let's look at the Laplace pressure. This additional pressure is produced by the forces of surface tension under the curved surface of a liquid. To understand where this pressure originates, imagine a small solid angle whose apex is at the center of a sphere of radius R, which "cuts" a tiny piece of the sphere's surface (fig. 3). The forces of surface tension are tangent to the sphere at each point of the circle cut into the surface. Summing all these forces (that is, integrating them) yields the total resultant force directed toward the center of the sphere. We can see now that the surface tension in a sphere results in an additional pressure ΔP within the liquid.

Now we recall that the surface tension σ can also be interpreted as the density of the surface energy. This means that the total surface energy of a spherical drop is $4\pi R^2 \sigma$. Let's assume that the Laplace pressure produced a very small symmetrical deformation of the drop and decreased its radius to $R - \delta R$. (How can this be done with an incompressible liquid? Here's where mathematics helps us: it works with *any* small value, so we just suppose that the deformation is much smaller than the radius of the atom.) The surface energy of the deformed

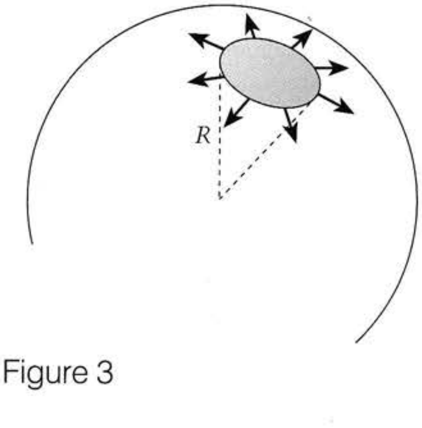


Figure 3

drop becomes $4\pi\sigma(R - \delta R)^2$. The corresponding energy change is $(4\pi R^2 - 4\pi(R - \delta R)^2)\sigma \cong 4\pi \cdot 2R \cdot \delta R \cdot \sigma$, because $\delta R \ll R$. On the other hand, the work performed by the pressure during the "shrinking" of the drop by δR is $\Delta P \cdot 4\pi R^2 \cdot \delta R$. Making the change in surface energy equal to the work of the pressure forces gives us the result for our first step—a formula describing the additional pressure under a curved surface:

$$\Delta P = \frac{2\sigma}{R}.$$

This surface pressure must be counterbalanced by an internal pressure, if we want the drop to be in equilibrium or near it (the latter case is called the *quasi-equilibrium* state).

Quantum pressure in atomic nuclei

Our second step is to obtain an estimate of the pressure that "resists" the Laplace pressure. At this point we must admit that nucleons in the atomic nucleus are somewhat different from the particles of ordinary liquids. As we know, pressure in an ordinary fluid is determined by the mean kinetic energy of its particles and by the energy of their mutual interaction. In macroscopic physical systems, the kinetic energy of the particles is determined by the temperature. The most obvious example is the classical ideal gas. Its pressure equals nkT, where n is the number density of the molecules (that is, their number per unit volume), T is the absolute temperature, and $k = 1.38 \cdot 10^{-23}$ J/K is Boltzmann's constant. Expressions like this are absolutely useless in describing the behavior of nucleonic liquids. Nucleons obey the laws of quantum mechanics, so their velocity, pressure, and energy are practically independent of temperature. So to evaluate the pressure in the nuclear system, we'll try a roundabout way and again use dimensional analysis.

When dealing with a quantum system, we naturally expect the pressure to depend on Planck's constant $\hbar = 1.05 \cdot 10^{-34}$ J · s. We'll assume that the pressure also depends on the concentration of particles n and the particle mass m. By the way, this dependence on n "hides" the influence of internucleonic distance (and mutual nucleonic interaction) on nuclear pressure. Here are the dimensions of the corresponding values: $[\hbar] = \text{kg} \cdot \text{m}^2/\text{s}^{-1}$, $[n] = \text{m}^{-3}$, $[m] = \text{kg}$, $[P] = \text{kg} \cdot \text{m}^{-1} \cdot \text{s}^{-2}$. Let's start with the formula

$$P \propto \hbar^\alpha m^\beta n^\gamma.$$

Comparing the dimensions on the left- and right-hand sides yields three equations used to find the numbers α, β, γ:

$$1 = \alpha + \beta,$$
$$-1 = 2\alpha - 3\gamma,$$
$$-2 = -\alpha.$$

It follows that

$$\alpha = 2,$$
$$\beta = -1,$$
$$\gamma = \frac{5}{3},$$

so

$$P \propto \frac{\hbar^2}{m} n^{5/3}.$$

This formula can be used to estimate the pressure of protons P_p and neutrons P_n, and the total quantum nucleonic pressure $P_N = P_p + P_n$. In calculating these values we assume that the number of protons in a nucleus is Z and the number of neutrons is $A - Z$, neglecting the difference between the masses of the protons and neutrons: $m_n \sim m_p \sim m$. Here are the resulting equations:

$$n_p = \frac{Z}{\frac{4}{3}\pi R^3} = \frac{Z}{\frac{4}{3}\pi r_0^3 A} = \frac{3}{4\pi}\frac{1}{r_0^3}\frac{Z}{A},$$
$$n_n = \frac{A-Z}{\frac{4}{3}\pi R^3} = \frac{A-Z}{\frac{4}{3}\pi r_0^3 A} = \frac{3}{4\pi}\frac{1}{r_0^3}\frac{A-Z}{A},$$
$$P_N = P_p + P_n$$
$$\propto \frac{\hbar^2}{m}\frac{1}{r_0^5}\left(\left(\frac{Z}{A}\right)^{5/3} + \left(\frac{A-Z}{A}\right)^{5/3}\right).$$

If we assume $\sigma/R \propto P_N$—that is, the Laplace pressure counterbalances the nucleonic pressure—we get (in order of magnitude, of course)

$$\sigma \propto \frac{\hbar^2}{m}\frac{A^{1/3}}{r_0^4}\left(\left(\frac{Z}{A}\right)^{5/3} + \left(\frac{A-Z}{A}\right)^{5/3}\right).$$

The expression in the large parentheses is of the order of 1. It turns out that more rigorous calculations don't improve the numerical coefficients in our formulas very much. In addition, more accurate (and far more complicated!) calculations confirm that our estimates are quite reasonable.

Electric pressure

Is the nucleonic pressure P_N the only counterforce that neutralizes the Laplace pressure ΔP? The creators of the liquid-drop nuclear model thought that the surface tension could also be offset by electrical pressure on the surface. This kind of pressure is produced by protons that have been pushed apart and equals the product of the electric field strength at the drop's surface and the surface density of the electric charge. We can estimate it as follows. At the surface of the nucleus (at a distance R from the center) the nuclear charge Ze generates an electric field with a strength of about $(Ze)/(4\pi\varepsilon_0 R^2)$, where $\varepsilon_0 = 8.85 \cdot 10^{-12}\ \text{C}^2/(\text{m}^2 \cdot \text{N})$ is the permittivity of free space. If all the protons were located on the nuclear surface, the surface charge density would be $(Ze)/(4\pi R^2)$. Thus the electrical pressure P_e is of the order of

$$\frac{Ze}{4\pi\varepsilon_0 R^2}\frac{Ze}{4\pi R^2} = \frac{(Ze)^2}{(4\pi)^2\varepsilon_0 R^4}.$$

If P_e were larger than P_N, we could estimate the surface tension as follows (again without numerical coefficients):

$$\frac{\sigma}{R} \propto \frac{(Ze)^2}{\varepsilon_0 R^4},$$

or

$$\sigma \propto \frac{(Ze)^2}{\varepsilon_0 R^3} = \frac{(Ze)^2}{r_0^3 A\varepsilon_0}.$$

Nuclear oscillation frequencies

In the general case both the quantum pressure P_N and electric pressure P_e contribute to the total pressure inside the nuclear droplet. Therefore, the condition of nuclear equilibrium looks like this:

$$\frac{\sigma}{R} = C_1 P_N + C_2 P_e,$$

where C_1 and C_2 are numbers that could be written explicitly if we took into account all the numerical factors in our estimate. The value of such a "precise" calculation isn't very great, however, because the model described is rather rough (for example, it neglects the possibility that the nucleons are unevenly distributed within the nucleus). Nevertheless, this simple model captures the qualitative dependencies and even produces reasonable numerical estimates.

Finally, we have arrived at our long-sought estimate for the surface tension:

$$\sigma = C_1 \frac{\hbar^2}{m} \frac{1}{r_0^4} A^{1/3} \left(\left(\frac{Z}{A} \right)^{5/3} + \left(\frac{A-Z}{A} \right)^{5/3} \right) + C_2 \frac{(Ze)^2}{\varepsilon_0 r_0^3} \frac{1}{A}.$$

Since in this article we're interested only in rough estimates, we assume that both C_1 and C_2 are of the order of 1.

There is another way to determine the nuclear surface tension if we suppose that an atomic nucleus is kept from disintegrating by the forces of surface tension alone. In this approach we have

$$4\pi R^2 \sigma_n \propto A\bar{\varepsilon} \cong \Delta W,$$

from which we get

$$\sigma_n \propto \frac{A\bar{\varepsilon}}{4\pi r_0^2 A^{2/3}} \propto \frac{\bar{\varepsilon} A^{1/3}}{4\pi r_0^2}.$$

For $A \sim 200$ (say, for isotopes of uranium with atomic mass 235 or 238), this formula yields $\sigma_n \sim 2 \cdot 10^{17}$ J/m^2. Compared to the surface tension of water, this is a prodigious force.

Let's assume that the value σ_n doesn't differ much from the value of σ obtained when we supposed that the main source of pressure was P_N:

$$\sigma \propto \frac{\hbar^2}{m} \frac{A^{1/3}}{r_0^4} \sim \frac{\bar{\varepsilon} A^{1/3}}{4\pi r_0^2} \propto \sigma_n.$$

From this we get

$$\bar{\varepsilon} \propto \frac{\hbar^2}{m} \frac{1}{r_0^2}.$$

As expected, the numerical value $\bar{\varepsilon}$ is about 10 MeV, which is close (for the accuracy accepted here) to the value of $\bar{\varepsilon}$ = 8 MeV mentioned above.

Now all that's left is to estimate the characteristic frequencies of nuclear droplet oscillation by means of the formula $\nu \propto (\sigma/M)^{1/2}$, where $M = Am$ is the mass of the nucleus. We'll do it for two limiting cases: (1) when $P_N > P_e$ and (2) when $P_e > P_N$. In the first case we have

$$\nu_1 \propto \frac{\hbar^2}{m} \frac{1}{r_0^2} A^{-1/3} \left(\left(\frac{Z}{A} \right)^{5/3} + \left(\frac{A-Z}{A} \right)^{5/3} \right)^{1/2},$$

while in the second we have

$$\nu_2 \propto \left(\frac{Z}{A} \right) \omega_0,$$

where

$$\omega_0^2 \propto \frac{e^2}{m r_0^3 \varepsilon_0}.$$

Since the expression in the large parentheses in the first formula for ν_1 is approximately equal to one, we can simplify the formula:

$$\nu_1 \propto \frac{\hbar}{m r_0^2} A^{-1/3}.$$

In the general case, the frequencies of nuclear oscillation are estimated to be about 10^{22} s^{-1} and are called *nuclear frequencies*. A more rigorous analysis shows that for the heavy nuclei we've considered, the pressure values are characterized by $P_N > P_e$. As in an ordinary liquid, the collective character of the motion of the nuclear particles in the incompressible nuclear liquid leads to oscillations of the droplet's shape (the volume of the nucleus doesn't change). In quantum mechanics the oscillating motion is quantized. The change of energy during the transition of a nucleus from one state to another $\Delta E = s\omega$, where $\omega = 2\pi\nu$ and s is an integer. So we can estimate the possible values of the energy corresponding to the modes of nuclear oscillation. (The simplest examples of such modes are quadrupole and octapole oscillations. In the first case the excited nuclear droplet assumes an ellipsoidal shape, while in the second case it looks like a pear.)

Summing up

A common feature of the oscillatory motion of drops in a mist and atomic nuclei is that the square of the frequency of oscillation is proportional to the ratio of two characteristic values: the surface tension and the mass of the droplet. We have seen that the oscillation of droplets of two completely different incompressible liquids can be described within the framework of the so-called liquid-drop model. We find it very significant that simple analogies and straightforward reasoning can lead us to the very depths of the atomic nucleus and even provide us with reasonable qualitative and quantitative estimates of its characteristic frequencies and energies.

Still we must remind our readers again that we were working with a model. In nuclear physics there are dozens of models, but still there is no united and completely consistent theory. So when you come up against other nuclear models, be attentive. Don't be in a hurry to find contradictions in the different approaches, but also don't avoid them. Of course, experimental results will have the final say. But alas, no one has yet incorporated into any physical model that tiny drop of love and sympathy that Pushkin's Tatyana sought.

A revolution absorbed

How non-Euclidean geometry entered the mathematical mainstream

by E. B. Vinberg

THE DISCOVERY THAT Euclid's geometry is not the only possible geometry, which was made in the beginning of the 19th century by Gauss, Lobachevsky, and Bólyai, had the same sort of effect on the world view of humanity as great discoveries in the natural sciences, such as the heliocentric system of Copernicus or Darwin's theory of evolution. However, there are very few non-experts who know that non-Euclidean geometry, along with Euclidean, has been one of the tools of modern mathematics since the end of the last century, despite the fact that "the space where we live" is, as far as we know it today, more Euclidean than non-Euclidean.

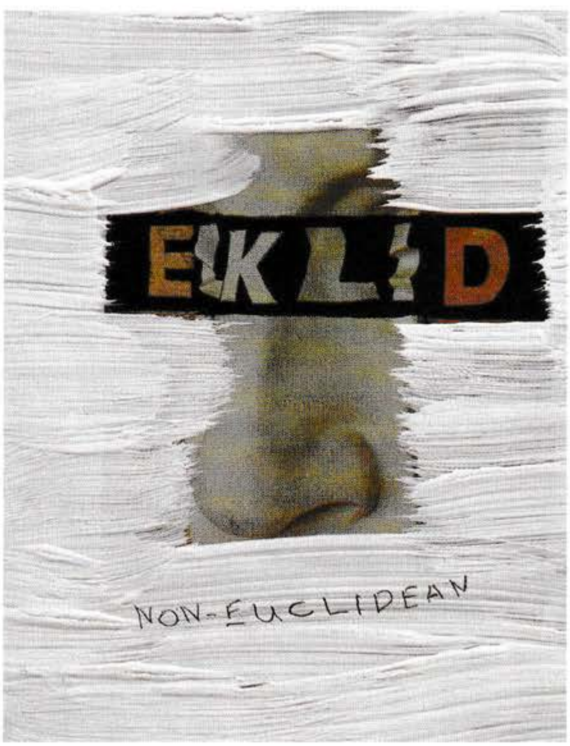
It is the nature of mathematical theories that their basic notions (in geometry, these are points, straight lines, transformations, and so on) can be interpreted in many different ways and applied to different kinds of objects. In particular, geometry may be applied not only to "the space where we live" but also to other spaces that appear in mathematical and physical theories. The geometries of these spaces may be of various types—in particular, they may be non-Euclidean.

If we let the expression "non-Euclidean geometry" denote any geometry whose axioms differ from those of Euclid, there would be infinitely many such geometries, and we could hardly be expected to say anything that would be true of all of them. In this article we'll use the term "non-Euclidean geometry" in a rather narrow sense, confining it to either Lobachevskian geometry or to geometry on a sphere (as we will see, these two geometries are in some sense equivalent). These two geometries, together with the Euclidean, occupy a special place among all the geometries that include the notion of distance between points. They could be described as geometries of maximum mobility, or as geometries of constant curvature.

The first applications of Lobachevskian geometry were made by Lobachevsky himself, who used it to calculate several integrals. They were rather particular results, and they were not developed further. However, some of Lobachevsky's integrals still appear in many tables.

At the end of last century Poincaré and Klein established that Lobachevskian geometry is closely connected to the theory of functions of a complex variable and to number theory (more precisely, to the arithmetic of indefinite quadratic forms). Since that time, Lobachevskian geometry and its methods have become indispensable parts of these branches of mathematics.

The importance of Lobachevskian geometry has increased even more during the last 15 years due to the work of the American mathematician William Thurston (recipient of the 1983 Fields Medal), who discovered its connection with the topology of 3-manifolds. Every year dozens of papers in this



Art by Yury Vashchenko

field are published. So it would be reasonable to say that the "romantic period" in the history of Lobachevskian geometry, when the efforts of researchers were largely aimed at its interpretation from the point of view of the foundations of geometry, is over. Modern investigations demand more and more practical knowledge of its methods.

In this article I'll give several examples of theorems in non-Euclidean geometry and formulate a principle that allows one to obtain theorems in Lobachevskian geometry from those in spherical geometry. After this we'll examine some problems in non-Euclidean geometry that play a central role in its applications.

The spherical equivalent of the Pythagorean theorem

Spherical geometry can be modeled within Euclidean geometry. That is, we can find objects that act like those of spherical geometry. In fact, this is easy to do: spherical geometry coincides with geometry on the surface of an ordinary Euclidean sphere (hence its name!). Choose any sphere and regard its great circles as straight lines and the lengths of their arcs (or the corresponding central angles multiplied by the radius of the sphere, which is the same) as distances. Thus, if one changes the radius of the sphere, all distances will be multiplied by a common factor, so there is no essential difference between geometries of spheres of different radii. Therefore, it's convenient to put the radius equal to 1. Unless stated otherwise, we'll assume that this condition holds.

Let's find an analogue of the Pythagorean theorem on the sphere—that is, a formula that expresses the hypotenuse of a right spherical triangle in terms of its legs.

Figure 1 depicts a right triangle ABC (C is the right angle[1]) on a sphere with center O. Put

$$|BC| = a,\ |CA| = b,\ |AB| = c. \qquad (1)$$

[1]Two great circles are said to form a right angle if the planes they determine are perpendicular.—*Ed.*

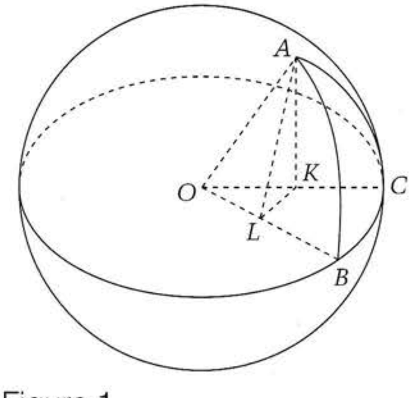


Figure 1

(Here we understand the length of a segment in the spherical sense—that is, as the length of the corresponding arc of a great circle.)

We make some constructions in the Euclidean space containing our sphere. First we draw line AK from point A perpendicular to radius OC. Since the planes AOC and BOC are perpendicular, AK is perpendicular to the plane BOC (the "equatorial plane" in figure 1). From point K we draw line KL perpendicular to radius OB. Since AK is perpendicular to plane BOC, plane ALK, which contains it, is also perpendicular to plane BOC. Now BO is a line perpendicular to plane ALK, so BO must be perpendicular to line AL as well.

Let

$$a = \angle BOC,\ b = \angle COA,\ c = \angle AOB. \qquad (2)$$

Then from right triangle AOK we have $OK = \cos b$. From right triangle AOL we have $OL = \cos c$, and from right triangle KOL we have $OL = OK \cos a = \cos a \cos b$. Comparing the two values of OL, we find that

$$\cos c = \cos a \cos b. \qquad (3)$$

This is the Pythagorean theorem in spherical geometry.

For a sphere of arbitrary radius R we have the following formula:

$$\cos\frac{c}{R} = \cos\frac{a}{R}\cos\frac{b}{R}. \qquad (4)$$

As $R \to \infty$ the sphere becomes flatter and flatter, and its geometry approaches the Euclidean. Let's assume that a and b are constant and use the approximate equality

$$\cos x = 1 - \frac{x^2}{2} + o\left(x^2\right), \qquad (5)$$

which is true if x is small.[2] (Here $o(x^2)$ stands for some polynomial in x whose terms are of degree 2 or higher.) Then from equation (4) we obtain

$$1 - \frac{c^2}{2R^2} + o\left(\frac{1}{R^2}\right) = \left[1 - \frac{a^2}{2R^2} + o\left(\frac{1}{R^2}\right)\right]\left[1 - \frac{b^2}{2R^2} + o\left(\frac{1}{R^2}\right)\right] = 1 - \frac{a^2+b^2}{2R^2} + o\left(\frac{1}{R^2}\right). \qquad (6)$$

Thus

$$c^2 = a^2 + b^2 + o(1/R^2), \qquad (7)$$

and as a limit as $R \to \infty$ we get the ordinary Pythagorean theorem:

$$c^2 = a^2 + b^2, \qquad (8)$$

as we might naturally expect.

The principle of equivalence

We now turn to Lobachevskian geometry. We can describe this geometry by saying that the Lobachevskian plane is just a "sphere of radius i" (where i is the imaginary unit). What might this outrageous statement mean? Essentially, all it means is that Lobachevskian geometry is the geometry derived from spherical geometry by taking all its formulas and dividing the linear values that appear in them by i. While there are other ways of describing the Lobachevskian plane, we will take this "principle of equivalence" as our starting point.

For all the functions we will be dealing with, it turns out that the algebraic formulas we already know for real numbers still hold true for complex numbers (this is guaranteed by the "principle of analytic continuation" for complex functions). It is also true that the sine and cosine of a complex number satisfy the following equations:

[2]This approximation comes from the Taylor series for cos x, which is studied in calculus.—*Ed.*

$$\cos ix = \cosh x,$$
$$\sin ix = \sinh x, \qquad (9)$$

where the *hyperbolic sine* ($\sinh x$) and *hyperbolic cosine* ($\cosh x$) are defined by

$$\cosh x = \frac{e^x + e^{-x}}{2},$$
$$\sinh x = \frac{e^x - e^{-x}}{2}. \qquad (10)$$

These formulas can be written as

$$\cos\frac{x}{i} = \cosh x,$$
$$\sin\frac{x}{i} = \frac{1}{i}\sinh x. \qquad (11)$$

Thus, when we turn a formula in spherical geometry into one in hyperbolic geometry, we need to replace all the trigonometric functions of linear values with their hyperbolic equivalents (imaginary units that appear from sines will automatically vanish). Readers unfamiliar with the theory of functions of a complex variable can regard this last statement (with a few additional details described below) as a formulation of the principle of equivalence.

In particular, formula (3) in spherical geometry thus yields the following hyperbolic equivalent of the Pythagorean theorem:

$$\cosh c = \cosh b \cosh a. \qquad (12)$$

On a "sphere" of radius iR, this formula becomes

$$\cosh\frac{c}{R} = \cosh\frac{a}{R}\cosh\frac{b}{R}. \qquad (13)$$

Letting $R \to \infty$, as we did above in the case of an ordinary sphere, we obtain the "usual" Pythagorean theorem at the limit.

The circumference of a circle

A set of points on a non-Euclidean plane, lying at a fixed distance r from a given point, is called the circle of radius r and center at the given point, just as in Euclidean geometry. It's readily apparent that a circle of radius r on the unit sphere is simply a Euclidean circle of radius $\sin r$. Therefore, its circumference C is given by the formula

$$C = 2\pi \sin r. \qquad (14)$$

Applying the principle of equivalence formulated above, we obtain a formula for the circumference of a circle of radius r on the Lobachevskian plane:

$$C = 2\pi \sinh r. \qquad (15)$$

A glance at equation (10) will show that the function $\sinh r$ grows very quickly as r approaches infinity. So in Lobachevskian geometry the circumference of a circle is not proportional to its radius, as it is in Euclidean geometry, but grows much faster. Accordingly, a circle on Lobachevskian plane is much more spacious than a Euclidean circle of the same radius.

The area of a triangle

All the non-Euclidean formulas that we've been discussing have Euclidean equivalents. However, there are many formulas in non-Euclidean geometry that have no equivalents in Euclidean geometry. The formula that describes the area of a triangle using only its angles is one such.

Let's find an expression for the area of a spherical triangle (fig. 2). Spherical triangle ABC can be thought of as the intersection of three hemispheres P, Q, and R whose boundary circles contain sides BC, CA, and AB, respectively. (In figure 2, P is the "upper" hemisphere, Q is the "front" hemisphere, and R is the "right" hemisphere.) The surface area of a sphere of radius r is given by $4\pi r^2$, so if the sphere has radius 1, a hemisphere has area 2π. It turns out that the "sliver" formed by the intersection of two hemispheres has an area twice the radian measure of the angle between their boundary circles. Thus the intersections of Q and R, R and P, P and Q have areas 2α, 2β, 2γ, respectively. (Here α, β, γ are the angles of our triangle—see figure 2).

The union of hemispheres P, Q, and R is the whole sphere minus the "polar" triangle $A'B'C'$. Let S be the area of triangle ABC. Then the area of $A'B'C'$ is also S, and we see that the area of the union of P, Q, and R is equal to $4\pi - S$.

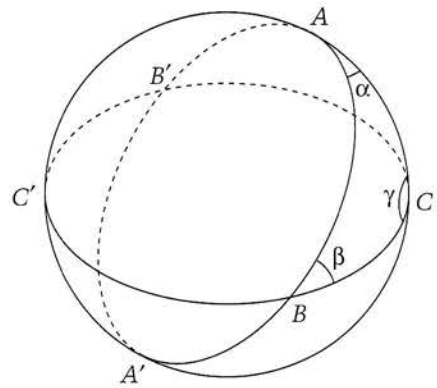


Figure 2

There is, however, another way to calculate this area: add up the areas of the three hemispheres, subtract the areas of their three mutual intersections (since each of them appeared in this sum twice), and add the area of triangle ABC, which was not taken into account so far (we added it three times while summing the areas of hemispheres, but later we subtracted it three times, when we subtracted the areas of the mutual intersections). Finally we have

$$4\pi - S = 2\pi + 2\pi + 2\pi - 2\alpha - 2\beta - 2\gamma + S, \qquad (16)$$

from which we get

$$S = \alpha + \beta + \gamma - \pi. \qquad (17)$$

We now see that the sum of the angles of a spherical triangle is always greater than π, and that the excess is equal to its area. If the triangle is very small, then the sum of its angles is close to π, because this triangle is almost Euclidean.

According to the principle of equivalence, if we want to obtain an expression for a triangle's area on the Lobachevskian plane, we must divide all linear values in equation (17) by i. The right-hand side of this expression contains no linear values and will not change (radian measure is dimensionless). The area will be divided by $i \cdot i = -1$. Multiplying both sides by -1, we obtain

$$S = \pi - (\alpha + \beta + \gamma). \qquad (18)$$

Thus the sum of the angles of a hyperbolic triangle is always less than α, and the shortfall is equal to its area. The sum of the angles of a very small hyperbolic triangle is almost equal to π.

Generally speaking, the geometry of a small domain in non-Euclidean space is in every respect very much like the Euclidean—the smaller the domain, the closer the approximation. That is why, if the "space where we live" is Euclidean, we can never prove this experimentally. Indeed we always deal with only a finite part of the space (although it may seem very large to us), and our measurements have only finite precision. Even if we find that the geometry in this part doesn't deviate from the Euclidean, we can still suppose that our space is non-Euclidean but that the domain we're exploring is too small (compared to the size of the universe) for us to discover non-Euclidean effects, taking into consideration the precision of our measurements.

(In reality the situation is even more complicated. According to the theory of relativity, space cannot be viewed as an entity apart from time—that is, one deals with something called space–time. So a question about "Euclidean space" is in need of some refinement.)

Parallel lines in Lobachevskian geometry

In order to understand what happens with parallel lines on the Lobachevskian plane, we first need to consider the matter on the Euclidean plane, although the reader might think it's trivial.

Two straight lines on the Euclidean plane are called parallel if they do not meet. It's well known that for any point A on the Euclidean plane that does not belong to a line l, there exists a unique line m containing A and parallel to l. This is known as Euclid's Fifth Postulate. The line m can be regarded as the limiting position of a straight line AB connecting A and a point B on l, while point B approaches infinity in a fixed direction. In fact, let's

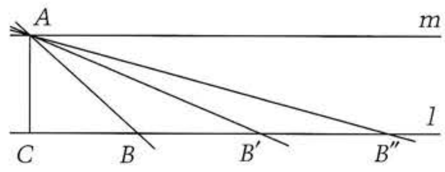


Figure 3

draw the perpendicular AC from A to l and watch how the angles ABC and BAC change. Let B' be the position of point B as it moves away from C (fig. 3). Angle ABC is external for triangle ABB' and so is equal to the sum of angles BAB' and $AB'B$. (This is equivalent to the statement that the sum of the angles of triangle ABB' is equal to π.) Therefore,

$$\angle AB'C < \angle ABC. \qquad (19)$$

Moreover, if $|BB'| = |AB|$—that is, if ABB' is an isosceles triangle—then $\angle BAB' = \angle AB'B$, and consequently

$$AB'C = \frac{1}{2}\angle ABC. \qquad (20)$$

From here it follows that as point B moves to infinity, angle ABC approaches zero.

Further, since the sum of the angles of triangle ABC is equal to π, then

$$\angle BAC = \frac{\pi}{2} - \angle ABC. \qquad (21)$$

Therefore, angle BAC approaches $\pi/2$. This means that the limiting position of AB is the line m perpendicular to AC. The same line appears when B approaches infinity in the opposite direction. The line m is the only line parallel to l that contains A.

Now let's make a similar construction on the Lobachevskian plane, where (as we noted above) the sum of the angles of a triangle is less than π. Inequality (19) only becomes stronger, and equality (20) turns into the inequality

$$\angle AB'C < \frac{1}{2}\angle ABC. \qquad (22)$$

Therefore, our final conclusion about the way angle ABC changes remains true: this angle monotonically approaches zero. Equality (21) becomes the inequality

$$\angle BAC < \frac{\pi}{2} - \angle ABC, \qquad (23)$$

and the difference between its right and left sides, which is equal to the area of triangle ABC, can only increase. Thus angle BAC tends to some acute angle δ. The limiting straight line m, which forms an angle δ with the perpendicular AC, does not intersect the line l. In Euclidean geometry, we call a line parallel to l if it doesn't intersect l. In Lobachevskian geometry, we agree that of those lines passing through point A, only the straight line m will be called *parallel* to l, although there are other lines through A that don't intersect l.

If point B moves in the opposite direction, the line AB approaches another limiting line m', which also forms an angle δ with the perpendicular, although this time the angle lies on the other side of AC. The line m' is also called parallel to l, but "in the opposite direction." This situation is schematically shown in figure 4 (we say "schematically" because it's not possible to depict a non-Euclidean figure in the Euclidean plane of a magazine page).

Thus, for any point A that does not lie on a line l, there are exactly two lines passing through A and parallel to l in the Lobachevskian plane. None of the lines that lie between them intersect l (but they are not called parallel to it).

The angle δ is called the angle of parallelism. It depends only on the distance $|AC| = d$. More precisely,

$$\delta = 2\arctan(e^{-d}). \qquad (24)$$

You may wish to derive this formula on your own. To do it you need to use

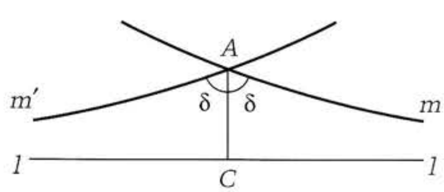


Figure 4

the methods we used to demonstrate the spherical equivalent of the Pythagorean theorem and prove certain relations between the sides and angles of a right spherical triangle. Then you can, with the help of the principle of equivalence, obtain the corresponding relations for a right hyperbolic triangle ABC (see figure 3) and pass to the limit.

Tilings of a plane with regular polygons

Graph paper and honeycombs are examples of tilings of the Euclidean plane with congruent regular polygons (squares in the one case, hexagons in the other).

Since the sum of the angles of a Euclidean p-gon is $(p - 2)\pi$, each angle of the regular p-gon is equal to $(1 - 2/p)\pi$. If q polygons meet at each vertex of a tiling, we must have

$$\left(1-\frac{2}{\pi}\right)\pi=\frac{2\pi}{q}, \qquad (25)$$

from which we obtain

$$\frac{1}{p}+\frac{1}{q}=\frac{1}{2}. \qquad (26)$$

This equation has only three solutions in positive integers:

$$(p, q) = (3, 6); (4, 4); (6, 3). \qquad (27)$$

The last two solutions correspond to the tilings with squares and regular hexagons mentioned above. The first solution corresponds to a tiling with equilateral triangles.

The formula for the sum of the angles of a p-gon in the Euclidean plane is derived from the corresponding formula for a triangle via the polygon's decomposition into $p - 2$ triangles by diagonals emerging from one of its vertices. It's possible to prove in the same way that the sum of the angles of a spherical (hyperbolic) p-gon is equal to $(p - 2)\pi$ plus (or minus) its area.

So it's clear that the angle of a regular spherical p-gon is greater than $(1 - 2/p)\pi$, and that, in contrast to the Euclidean case, it depends on the radius of the circle circumscribing the polygon. If the radius is small, the polygon is very much like a Euclidean polygon, and the difference between its angle and $(1 - 2/p)\pi$ is very small. As the radius approaches $\pi/2$ (the greatest possible value), the polygon itself approaches a hemisphere, and its angle approaches π. Thus the angle of a regular spherical p-gon can be equal to any number from $(1 - 2/p)\pi$ to π.

Therefore, a tiling of the sphere with congruent regular p-gons, such that q of them meet at each vertex, exists if and only if

$$\frac{1}{p}+\frac{1}{q}>\frac{1}{2}. \qquad (28)$$

This inequality has five solutions:

$$(p, q) = (3, 3); (3, 4); (3, 5); (4, 3); (5, 3). \qquad (29)$$

These tilings bijectively[3] correspond to the regular polyhedrons in Euclidean space. That is, if we project the surface of a regular polyhedron from its center onto a circumscribed sphere, we get a tiling of this sphere with regular polygons (images of the facets of the polyhedron). Conversely, each tiling of a sphere by congruent regular polygons defines a regular polyhedron, whose vertices coincide with those of the tiling.

So the result obtained above means that there exist only five regular polyhedrons. They are the tetrahedron, octahedron, icosahedron, cube, and dodecahedron (all of them known since antiquity).

Similarly, the angle of a regular hyperbolic p-gon is less than $(1 - 2/p)\pi$. It's very close to this value if the radius of the polygon is small, and it approaches zero if the radius tends to infinity. Thus the angle of a regular hyperbolic p-gon can be equal to any (positive) number less than $(1 - 2/p)\pi$.

And so we have the following inequality that describes tilings of the Lobachevskian plane with regular polygons:

$$\frac{1}{p}+\frac{1}{q}<\frac{1}{2}. \qquad (30)$$

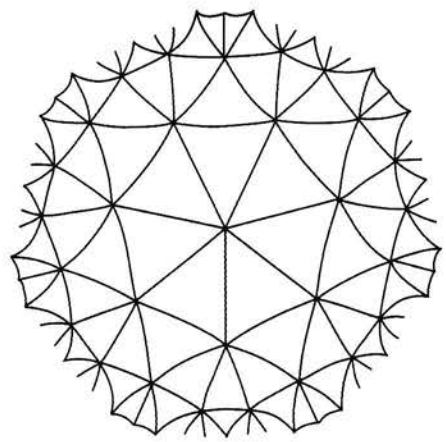


Figure 5

Solutions of this inequality are all pairs (p, q), except for those eight mentioned above that comply with equation (26) or inequality (28). Figure 5 gives a schematic representation of the tiling that corresponds to the solution (3, 7). We see now that at least as far as tilings are concerned, the Lobachevskian plane offers many more possibilities than the Euclidean plane or the sphere.

Even more possibilities present themselves if we eliminate the condition that the polygons in the tiling be regular (an artificial condition indeed), but retain the condition that they be congruent. Applications of Lobachevskian geometry to number theory and the theory of functions of a complex variable, mentioned in the beginning of this article, are connected with just this sort of tiling.

One can likewise investigate tilings of space with congruent polyhedrons. In the case of Euclidean space, such investigations are closely related to crystallography, and in the case of Lobachevskian space, to the topology of 3-manifolds. As to the latter, the theory of such tilings is still far from complete, although in 1954 Coxeter gave a description of all tilings of Lobachevskian space with congruent regular polyhedrons.

[3]A mapping is *bijective* if it is injective (that is, "one-to-one") and surjective (that is, "onto"). (For a primer on functions and mapping, see A. N. Kolmogorov's article "Home on the Range" in the September/October 1993 issue of *Quantum*.—*Ed.*)

Below absolute zero

They say you couldn't get there if you tried.
But maybe if you didn't try so hard,
or came at it from a different direction . . .

by Henry D. Schreiber

THE COLDEST TEMPERATURE is absolute zero—that is, zero on the Kelvin scale. Accordingly, if zero is the lowest possible value, absolute temperatures are always positive. Negative temperatures on the Kelvin scale would seem to be impossible. After all, the obvious avenue to get to negative temperatures would be to keep decreasing temperatures until they're below zero. But nature has another, more creative, way to get below absolute zero—and, once there, it exhibits even stranger properties!

Temperature

If asked to define temperature, most people equate it with the degree of hotness or coldness of substances. As such, temperature is the factor determining heat flow. When two objects are in contact, heat flows from the hotter object (the one at a higher temperature) to the colder object (the one at a lower temperature). The two objects eventually reach equilibrium when they achieve the same temperature.

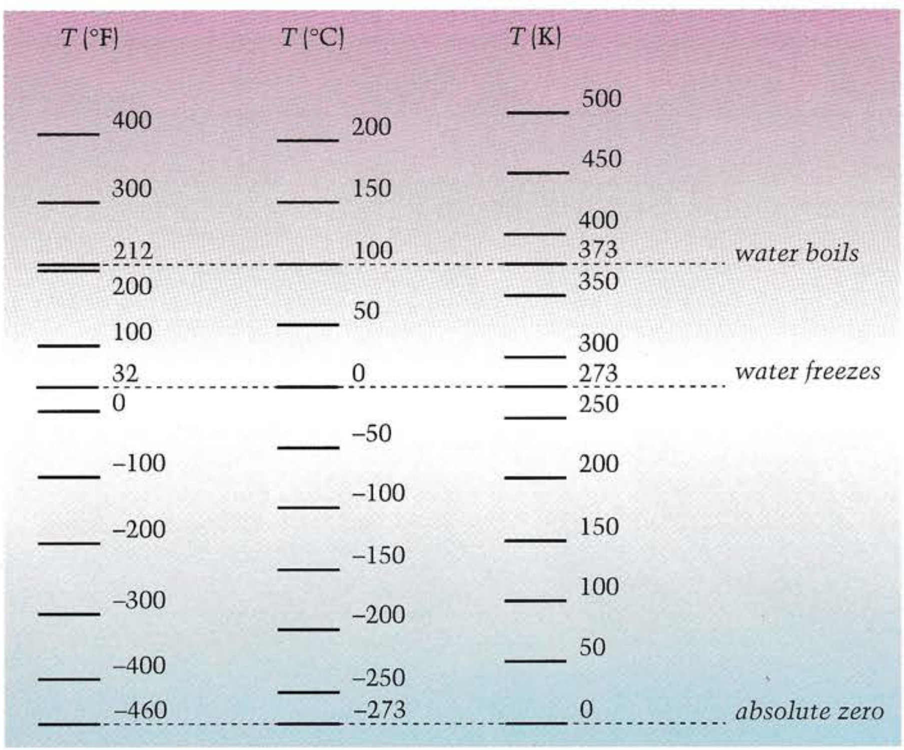


Figure 1

There are three scales that measure temperature: Fahrenheit, Celsius, and Kelvin, as shown in figure 1. Fahrenheit and Celsius are relative scales, defining temperatures with respect to certain reference points. In both cases, these values fix the freezing and boiling temperatures of water. Arbitrarily chosen by the developers of these scales are 32°F and 212°F versus 0°C and 100°C, respectively, for these reference points. Kelvin temperatures originated in order to explain Charles's law for gases, which states that the volume of a gas varies directly with its prevailing temperature. Upon plotting the volumes of various gas

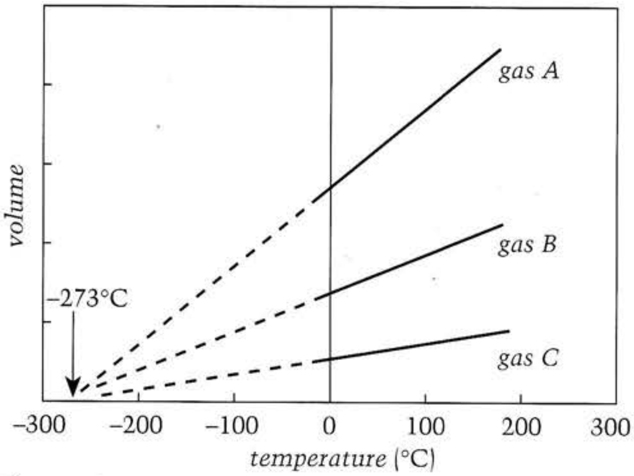


Figure 2

samples as a function of their temperatures in °C, as shown in figure 2, all lines extrapolated back to zero volume at –273°C. Absolute zero on the Kelvin scale then defined this temperature—temperatures lower than that were not possible, because volumes of substances are always positive. Thus, there is a 273-degree offset relating the Kelvin scale to the Celsius temperature. The Kelvin scale provides absolute temperatures, because its reference point of absolute zero is a fixed value for all substances. Absolute zero on the Kelvin scale is accordingly the coldest conceivable temperature.

One of the tenets of the kinetic molecular theory of gases (the principles describing gases as particles in constant motion) is that the kinetic energy of gas molecules depends only on their temperature when measured on the Kelvin scale. With decreases in temperature, molecules possess proportionally lower energies. Thus, an alternative way to envision absolute zero in this classical, albeit not quantum mechanical, approach is the temperature at which all molecular motion ceases. Temperatures can consequently approach absolute zero, but never quite get there. Temperatures as low as 0.00001 K have been recorded, but all motion or energy can never be withdrawn from a system.

With such perspectives on absolute zero and Kelvin temperatures, it seems reasonable never to imagine negative absolute temperatures. After all, you can't have less than zero volume for a substance, nor can you have less than zero motion. However, suppose you take an excursion into the molecular world—in particular, to see how molecules partition among themselves their available energy. You can then travel to the extremities of the absolute temperature scale—not just to the very cold but also the very hot. Surprisingly, you'll discover that negative absolute temperatures exist in certain systems. Not

only that, you'll find that there is a certain symmetry of the positive and negative temperatures about absolute zero, but not in the way you might have expected. Such is the life of molecules "below" absolute zero.

Energy distribution among molecules

Every molecule possesses a certain amount of energy. Suppose you consider a system in which the molecules have a choice of being in just one state of energy or another—a ground state or an excited state. An example is a magnetic system where some molecules align with the magnetic field, and so are in a low-energy or *ground state*; while other molecules align in opposition to the magnetic field, and so need more energy to enter the *excited state*. That is, at a given temperature some molecules have enough energy to go "against the grain," as molecular orientation defines the energy state. Figure 3 schematically illustrates such a system with only two energy states—a ground state with energy ε_0 and an excited state with energy ε_1. For convenience, consider the reference energy for the ground state to be zero, resulting in an energy of separation of ε.

The Boltzmann distribution law, which uses statistics to place molecules in energy levels, describes the number of molecules in each energy state at equilibrium. Applying this distribution law to the simple system with only two energy levels results in the equation

$$\frac{n_1}{n_0} = e^{-\varepsilon/kT},$$

where n_1 and n_0 are the populations of molecules in the respective energy levels, ε is the separation between the two energy levels, T is the temperature in kelvins, and k is Boltzmann's constant. Boltzmann's constant is a per-molecule version of the ideal gas constant and is just the ideal gas constant divided by Avogadro's number. Its numerical value is $1.380658 \cdot 10^{-23}$ J/K.

An examination of Boltzmann's distribution law shows that as the energy level separation ε increases, less of the molecular population exists in the excited state for the same temperature. Fewer molecules have sufficient energy to overcome the larger requirement of energy to get to the excited state. As also shown by this equation, the value of n_1 increases with respect to n_0 as the temperature increases. The energy available to the molecules increases with temperature, so that the number of the molecules in the excited state increases with respect to the ground state. Figure 4 plots n_1/n_0 as a function of the absolute temperature as described by the

Art by Sergey Ivanov

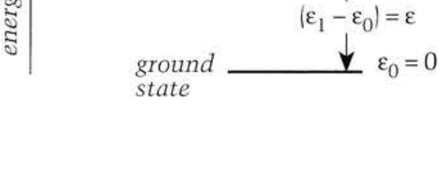


Figure 3

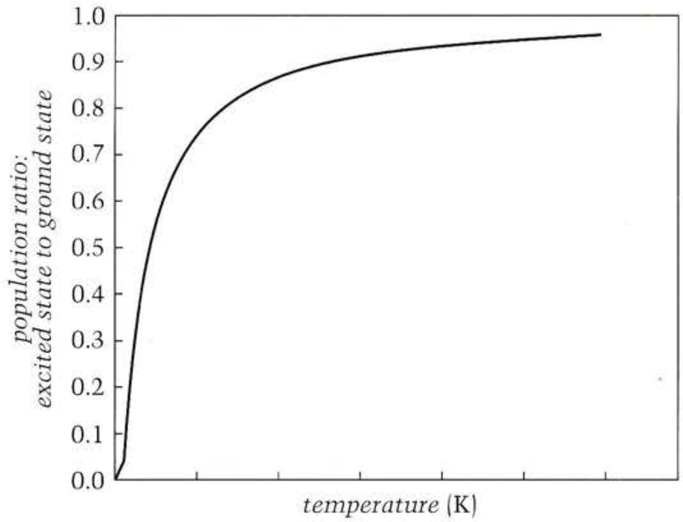


Figure 4

Boltzmann distribution law for a system with two energy levels at constant separation. This graph shows that as the temperature approaches absolute zero, n_1/n_0 also goes to zero. This seems reasonable, because at absolute zero there is no energy available for distribution to the molecules, so that the entire population drops to the ground state. At the other extreme, as the temperature gets very high, the mathematical expression for the distribution law states that n_1/n_0 approaches ε^{-0}, or unity. Consequently, at the limit of infinitely high temperatures, $n_1 = n_0$, and molecules populate the excited and the ground states equally. Figure 5 shows the temperature dependence for 10 molecules distributed between the two levels. Interestingly, there's a limit to the portion of molecules that exists in the excited state when at equilibrium.

Figure 5

Consequences of population inversions

It appears that no more than half of the molecules can exist in the excited state, even at infinitely high temperature. Is this limit real or artificial? A little of both! At equilibrium, there is indeed a limit, as quoted by the Boltzmann distribution law. On the other hand, can systems exist temporarily where enough energy has been absorbed by the molecules to put more in the excited state than in the ground state—that is, $n_1 > n_0$, or $n_1/n_0 > 1$? For example, suppose you pump energy into a system such that you excite the molecules to the

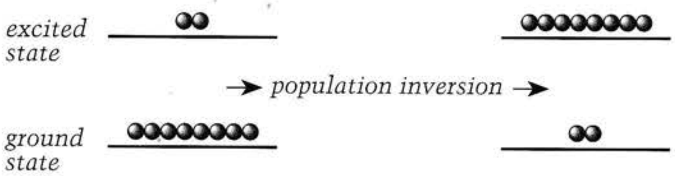


Figure 6

higher level faster than they can decay back to the ground state. In a sense, you trap the molecules in the higher energy state. Or alternatively, suppose you have an equilibrium distribution of molecules exposed to a magnetic field. Then you switch the direction of the magnetic field. What had been the excited energy state is now the ground state, and vice versa, at least momentarily. Figure 6 shows such occurrences, called *population inversions*.

What are the consequences of population inversions (which are admittedly not equilibrium situations)? First, taking the natural logarithm of both sides of the previously stated Boltzmann distribution law, we get

$$\ln\frac{n_1}{n_0} = -\frac{\varepsilon}{kT},$$

or, upon rearrangement,

$$\frac{1}{T} = -\frac{k}{\varepsilon}\ln\frac{n_1}{n_0}.$$

This equation describes the absolute temperature of the system by the relative populations of the two energy levels. But if n_1/n_0 is greater than unity, then $1/T$ is negative, as is the absolute temperature. There we have it: a negative absolute (Kelvin) temperature!

Going further, we can define fractions of the molecular population in each level—that is, $x_1 = n_1/N$ and $x_0 = n_0/N$, where N is $n_0 + n_1$, or the total number of molecules in the system. The previous equation then becomes

$$\frac{1}{T} = -\frac{k}{\varepsilon}\ln\frac{n_1/N}{n_0/N} = -\frac{k}{\varepsilon}\ln\frac{x_1}{x_0}$$

in terms of population fractions instead of actual populations. Because the population fractions of the two levels must add to one ($x_1 + x_0 = 1$), we can then rearrange this equation into

$$\frac{1}{T} = -\frac{k}{\varepsilon}\ln\frac{x_1}{1-x_1}.$$

Figure 7 is a plot of $1/T$ versus x_1, the fraction of molecules in the excited state. Note that the sign of $1/T$ is positive as long as x_1 is less than 0.5—that is, the excited state is less populated than the ground state as previously described. On the other hand, the sign of $1/T$, and

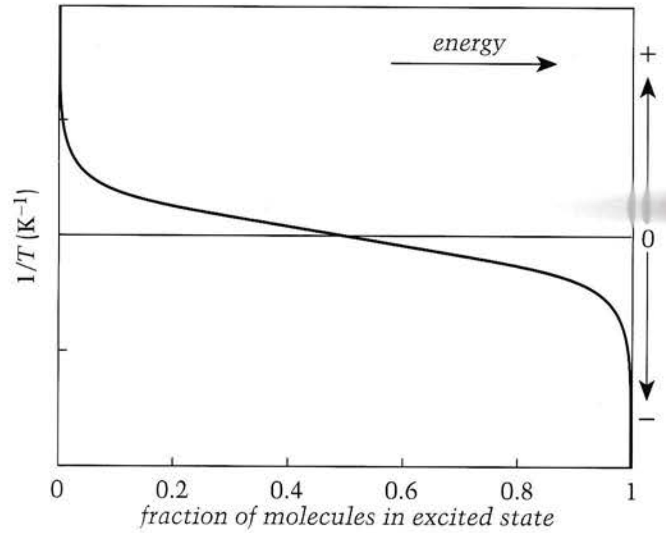


Figure 7

thus of the absolute temperature, becomes negative when x_1 is greater than 0.5—that is, when population inversions occur.

The total energy E for this system with two available energy levels is the sum of each level's population times its individual energy, or

$$E = n_1\varepsilon_1 + n_0\varepsilon_0 = n_1\varepsilon + n_0(0) = n_1\varepsilon.$$

Dividing by N (the total number of molecules) results in the expression

$$\frac{E}{N} = \frac{n_1}{N}\varepsilon = x_1\varepsilon,$$

or, if E' is the average energy per molecule (E/N),

$$\frac{E'}{\varepsilon} = x_1.$$

This equation shows that the energy of the system increases as x_1 increases. Thus, the x-axis in figure 7 also measures the energy of the system. This means that negative absolute temperatures represent much higher energies than do positive temperatures. This fact, of course, seems "reasonable," because more energy must be available to get a greater percentage of molecules in the excited state. In other words, curiously, negative absolute temperatures are hotter than positive temperatures.

Figure 8 (on the next page) plots T versus x_1—a rearrangement or modification of the previous figure. As expected, the left-hand side of the graph refers to positive temperatures; as more of the overall population goes into the excited state, the temperature systematically increases until defining an asymptote when x_1 approaches 0.5. Then the mathematical equation describes the mirror image of the function above $x_1 = 0.5$. Thus you get to negative absolute temperatures not by going

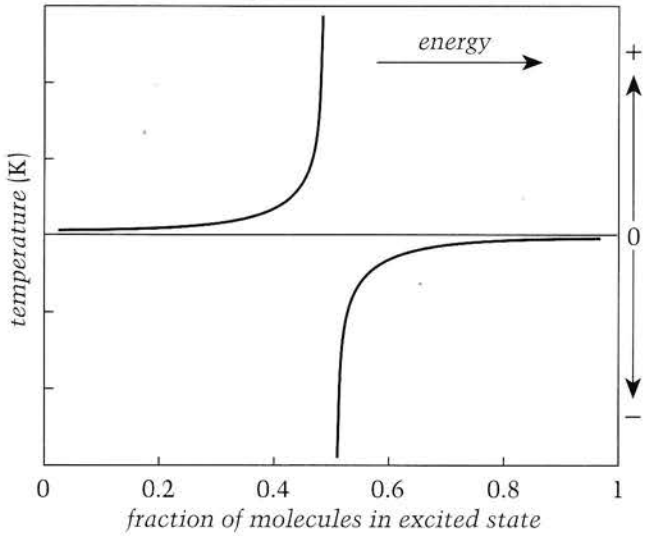


Figure 8

below absolute zero but by going past infinite positive temperatures. When $x_1 = 0.5$ (the two energy levels are equally populated), the temperature flips from positive to negative infinite temperatures as simply a mathematical quirk of the function relating the temperature to the population ratio. Once again, the x-axis also represents increasing energy to the right. Consequently, the coldest temperature is approaching absolute zero from the positive side (that is, as x_1 goes to zero, or when all molecules go to the ground state), and strangely, the hottest temperature is approaching absolute zero from the negative direction (that is, as x_1 goes to unity, or when all molecules go to the excited state).

Are negative absolute temperatures real?

Paradoxically, you don't attain negative absolute temperatures by the obvious route, by cooling from the positive temperatures and going through absolute zero. You can only get there by going through infinite temperatures. Furthermore, the hottest temperature is right below absolute zero, not very far numerically from the coldest temperature right above absolute zero.

Does this excursion into strangeness have any physical significance, or should you relegate it to the world of imagination, seeing it as a theoretical curiosity? Population inversions are real—they form the basis for the operation of lasers, in which more molecules populate the excited state than the ground state. One consequence is that the extremely high energies of the lasers impart properties describable by the world of negative absolute temperatures. Likewise, this concept of negative absolute temperatures has been verified experimentally in magnetic fields for nuclei whose energies are defined by their orientations.

Such phenomena, however, are not occurrences that are measured by ordinary thermometers. Thermometers are useful only for the positive absolute temperatures. Absolute temperatures are negative only under exceptional circumstances, when the molecular system wanders into the strange world of population inversions. Mathematical equations, such as the one relating the absolute temperature to the population ratio of the energy levels in the system—

$$T = -\frac{\varepsilon}{k\ln(x_1/x_0)}$$

—are the best way to assess the properties of a system with negative absolute temperatures. Mathematics, despite its abstract nature, is sometimes able to describe what is inherently harder to envision conceptually. Mathematical equations bring to science much the same understanding that poetic verses bring to literature. Both look at the real world from a different vantage point, lending beauty to what is being described and exposing hidden meanings.

Henry D. Schreiber *is a professor of chemistry at the Virginia Military Institute in Lexington, Virginia. His e-mail address is hs@vmi.edu.*

"QUESTIONING ANSWERS" CONTINUED FROM PAGE 10

integral (X, Y)-coordinates, then for each $j = e - 1, e - 2, \ldots, 0$, the point $P_{2^j \cdot m_0}$ also has integral (X, Y)-coordinates. In particular, P_{m_0} has integral (X, Y)-coordinates. Since m_0 is odd, fact 2 tells us that m_0 is either 1 or 3. Now let's consider the cases $m_0 = 1$ and $m_0 = 3$ separately.

(a) If $m_0 = 1$, then m is 2^e, and we saw above that P_{2^j} has integral (X, Y)-coordinates for all $j \le e$. But, quoting from our list,

$$P_{2^3} = P_8 = \left[\frac{21}{25}, -\frac{69}{125}\right]$$

does *not* have integral (X, Y)-coordinates. Therefore, $e \le 2$, and m is either 1, 2, or 4.

(b) If $m_0 = 3$, we use a similar argument, noting that

$$P_{2^2 \cdot 3} = P_{12} = \left[\frac{1357}{841}, \frac{28888}{24389}\right]$$

is *not* integral. This tells us that m is either 3 or 6.

Barry Mazur *is a professor of mathematics at Harvard University. This article is based on a lecture given in the spring of 1996 as part the series of Arnold Ross Lectures sponsored by the American Mathematical Society.*

Of combs and

"The mutual attraction of the electric fluid cal[...] called negative is inversely proportiona[...]
—Charles Augustin de C[...]

AS YOU MIGHT GUESS, COULOMB wasn't the first investigator who tried to find the law describing the interaction of charged bodies. His predecessors proposed interesting hypotheses, drew far-fetched analogies, and made fine experiments. Many outstanding scientific personalities took part in this scientific assault, among them Daniel Bernoulli, Joseph Priestley, Franz Aepinus, and Henry Cavendish. However, it was Coulomb who managed to complete the independent, careful, and convincing studies that laid the groundwork for quantitative electrostatics. The odd and predominantly qualitative mosaic of electric phenomena suddenly became united and harmonious. Now it was possible to speak of a unit of electric charge and explain the vast majority of accumulated data. Even more importantly, Coulomb's breakthrough helped introduce the well-developed ideas and methods of theoretical mechanics into the theory of electricity. Up until that time physics practically could not explain electrical demonstrations (it would be hard to decide whether to call them experiments or amusements). The law discovered by Coulomb paved the way for rapid and spectacular progress in the study of electrical phenomena.

Questions and problems

1. Given the (comparatively) vast spaces between elementary particles, and between atoms, why don't we simply fall through the floor?

2. What is the net force from two identical charges on a third charge located halfway between them?

3. A charged ball attracts a piece of paper (fig. 1). How does the attractive force change if a metal foil envelops (a) the ball, (b) the piece of paper?

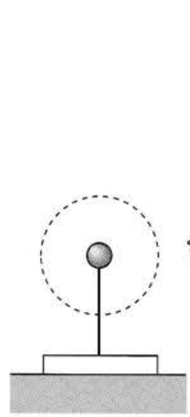


Figure 1

4. How does the attraction between the ball and paper in the previous problem change if the sphere around the ball is grounded?

5. All other conditions being equal, when will the force of electric interaction between two metal balls situated near one another be larger—when their charges are the same or the opposite?

6. A positively charged ball A is set near a metal ball B (fig. 2). Measurements detected no force of electric interaction between them. Is ball B charged or not?

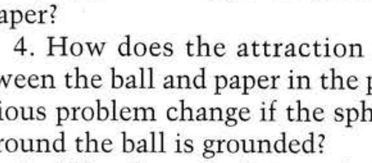


Figure 2

7. Can two conductors with like charges attract each other?

8. Two small weightless balls are suspended separately from a common point by thin nonconducting strings of the same length. What will happen if the balls are charged with like charges?

9. The charges of two identical small metallic balls located at some distance from each other differ by a factor of four. Will the force of their interaction change if the balls are briefly connected by a wire?

10. Two point charges q_1 and q_2 of equal strength but opposite polarity are fixed at some distance from each other (fig. 3). In which region (I, II, or III) can a third charge be in equilibrium with them? How will the answer change if the charges q_1 and q_2 have the same sign?

11. Two opposite point charges are attracted to each other with some force. Will the force affecting each charge change if a glass ball is placed between them?

12. An electron with a velocity v and located at infinity moves directly toward another electron that is free and at rest. What will happen to each electron?

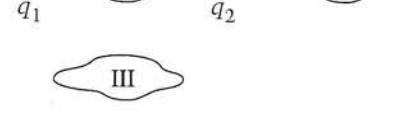


Figure 3

nd coulombs

*called positive and the electric fluid usually
nal to the square of the distance . . ."
e Coulomb (1736–1806)*

13. A thin wire ring carries a charge q. A like charge Q is placed at the center of the ring. What will result from the interaction of these charges?

14. A soap bubble connected with the atmosphere via a thin vertical tube shrinks and after a certain time turns into an almost flat film at the end of the tube. Will this time change if the bubble is (a) negatively charged, (b) positively charged?

15. Why can't the α-particles emitted by radioactive substances induce nuclear reactions in the heavy elements?

Microexperiment

Try to place a metal needle on the surface of water in a glass so as to make it float. Take a plastic comb and charge it with static electricity by running it through your hair. Put the charged comb near the needle. How does the needle behave? Why?

It's interesting that . . .

. . . an analogy with the Newtonian theory of gravity helped the German scientist Franz Maria Aepinus (who worked in St. Petersburg at the end of the 18th century) construct his own theory of electrical phenomena. Proceeding from the "harmony and economy" in Nature, Aepinus hypothesized that both electric and magnetic forces are inversely proportional to the square of distance.

. . . the law of interaction of electric charges was first established experimentally by Henry Cavendish.[1] Like many of his discoveries, however, this was something he did "for his own enjoyment"—he didn't publish his finding. It was James Clerk Maxwell who made this discovery public in the middle of the last century.

. . . according to the theoretical views held before Coulomb, electric interaction existed only in a special "atmosphere" immediately surrounding an electrified body.

. . . Coulomb called the device he built "to measure the minutest degree of force" a "torsion scale" and used it to study friction. The discoveries that immortalized his name were actually by-products of his main line of work—up to that time Coulomb hadn't been particularly interested in electricity and magnetism.

. . . repeating Cavendish's experiment, Coulomb found that electricity was distributed along the surface of conductors. Using the law of inverse squares he proved this property theoretically.

. . . Michael Faraday, convinced that all natural forces are interconnected, tried to find experimentally the interrelation between electricity and gravitation.

. . . although there is a formal resemblance between Coulomb's law and Newton's law of gravitation, a very deep gulf lies between them. All other conditions being equal, electric forces are much stronger than gravitational attraction, and gravitational repulsion has yet to be discovered. The existence of electric charges of two kinds and the strong interaction between them results in so precise a balance of these charges in any material body that it's not an easy thing to observe electric forces. The slightest disturbance of the bodies' neutrality induces the charges to restore it with all their indomitable might.

. . . scientists were able to explain (at least partially) the origin of elastic forces and friction only after they had come to understand the nature of the electric forces between neutral systems—that is, between molecules.

. . . in high school physics, and practically always in technology, electric and magnetic forces are handled separately. However, the question of which force—electric or magnetic—arises during the motion of free charges depends entirely on the system of reference.

. . . the phenomenon of superconductivity[2] can be explained by the coupling of free electrons in pairs that can move in metals without friction. In spite of Coulomb repulsion, the interaction of paired electrons with the crystal lattice reverses the sign of the force and results in their attraction.

. . . electrostatic experiments with conducting spheres have confirmed that the exponent in Coulomb's law is equal to 2 with an accuracy of 10^{-13}.

—Compiled by A. Leonovich

ANSWERS, HINTS & SOLUTIONS ON PAGE 53

[1]See "The Modest Experimentalist, Henry Cavendish" in the January/February 1991 issue of *Quantum*.—*Ed.*

[2]See "Meeting No Resistance" in the September/October 1991 issue of *Quantum*.—*Ed.*

Do you promise not to tell?

"Everything that happens is the message: you read an event and be one and wait, like breasting a wave, all the while knowing by living, though not knowing how to live."—William Stafford

by Arthur Eisenkraft and Larry D. Kirkpatrick

CAN YOU KEEP A SECRET? On the television show Seinfeld, George refused to tell his fiancé his code number to withdraw money from his bank with his ATM (automated teller machine) card. "They told me not to tell anyone." It's only when a person's life is in danger that George finally blurts out his secret code—Bosco. During World War II, the public was constantly reminded that "loose lips sink ships." It was through the work of Alan Turing (of computer fame) and others that the British were able to break the code of the Axis powers and win the war for the Allies. In the Pacific, the Japanese met defeat because they thought their code could not possibly be broken.

As we have more conversations electronically by e-mail or electronic bank transactions through the Internet, we assume somebody has successfully dealt with the question of security and secrecy. When the security is breached, we read about it in the newspapers and wonder how safe the whole system is. New interest has surfaced in the areas of random numbers and quantum properties of matter in hopes that more secrecy can be maintained.

How would you go about sending a message so that only one recipient can understand it? Although code books and number sequences have been somewhat successful in the past, we can look at the properties of waves for our secret transmissions. Is it possible to send a signal out so that one person will receive the signal but another will not? Let's rephrase the question as a physicist might: can we emit a wave that will be localized?

All waves emanating from a point source diverge from that source. A light bulb can be seen from all directions. We know that it is possible to send out a beam of light that is directional because we have used flashlights or observed the beam from a lighthouse. A collimated beam is produced by the careful placement of a lens or a mirror. If all the light emanates from the focal point of a lens, the transmitted beam will be parallel to the principal axis. Similarly, if the light is emitted from the focal point of a concave mirror, all of the light reflected from the mirror will emerge parallel to the principal axis. In both of these cases, the light that moves away from the lens or mirror without reflection will diverge, and spies could pick up our weaker signal from other positions.

What happens if we send out two signals? Two signals will produce places of constructive and destructive interference. If we know the location of our ally and the location of our enemy, we can, in theory, set up our pair of emitters such that our ally gets constructive interference of the two signals and our enemy gets destructive interference of the two signals.

As an example, we can look at a simple version of Young's double slit experiment. If the light emerges from two sources, we can calculate the positions of the constructive interference (antinodes) and the destructive interference (nodes). The conditions for the antinodes are that the distances from the two sources must differ by an integral number of wavelengths. The nodes must have distances that differ by an odd integral number of half-wavelengths ($^1/_2\lambda$, $^3/_2\lambda$, and so on).

Many people have experienced the destructive interference of radio waves when they are stopped at a red light near a large building and notice that the car radio now has lots of static. The direct signal from the transmitter and the reflected signal from the building produce the unpleasant static. Moving a short distance forward can bring your car to an antinode and away from the node so that you can hear your tunes.

In both of these examples, the two sources are in phase. When we send our secret messages, they needn't be. This gives us more flexibility in the placement of our transmitters.

This month's contest problem is adapted from one used at the XVI International Physics Olympiad held

Art by Tomas Bunk

LOOSE LIPS SINK SHIPS
Sound of Silence
TOM BUNK
TOP SECRET

in Portoroz, Yugoslavia, in 1985.

A young radio amateur maintains a link with two friends living in two towns. The two antennas are positioned such that when one friend, living in town A, receives a maximum signal, the other friend, living in town B, receives no signal, and vice versa. The two antennas transmit with equal intensities uniformly in all directions in the horizontal plane.

A. Find the distance between the antennas and the orientation of the antennas such that the electrical signals provide a maximum signal for one friend and no signal for the other. Assume that the two antennas transmit the signals in phase.

B. Find the parameters of the array (that is, the distance between the antennas, their orientation, and the phase shift between the signals supplied to the antennas) such that the distance between the antennas is a minimum.

C. Find the numerical solution if the radio station broadcasts at 27 MHz and the angles between north and the directions to town A and town B are 72° and 157°, respectively.

Please send your solutions to *Quantum*, 1840 Wilson Boulevard, Arlington VA 22201-3000 within a month of receipt of this issue. The best solutions will be noted in this space.

Boing, boing, boing . . .

The problem of the ball bouncing elastically down the inclined plane produced a very good response from our readers. Excellent solutions were submitted by Noah Bray-Ali from Venice High School in Los Angeles, California; Richard Burstein from the Commonwealth School in Boston; André Cury Maiali and Gualter José Biscuola, physics teachers in São Paulo, Brazil; Mary Mogge, a professor at Cal Poly Pomona; and Charles Thiel, a student at Montana State University–Bozeman.

Dr. Mogge points out that the problem is easiest to solve if we use the standard coordinate system for solving inclined plane problems, the x-axis pointing down the plane and the y-axis normal to the plane. She then writes that we should think of the problem as the projectile motion we all know and love, with the extra twist of an x-acceleration. We can then solve the problem in general and do not need to consider each bounce separately.

The motion in the y-direction is that of a ball bouncing with totally elastic collisions with the "floor" and a constant acceleration equal to the component of g perpendicular to the plane, $g/\sqrt{2}$. Therefore, the time for each bounce is always the same. Because of the symmetry of the motion in the y-direction, the time for each bounce is twice the time it took the ball to hit the plane the first time—that is, $t_1 = 2t_0$, where $t_0 = \sqrt{2g/h}$ as calculated in the article. Likewise, the magnitude of the y-component of velocity just before each bounce is always the same—it is the component of v_0 perpendicular to the plane, $v_y = v_0/\sqrt{2}$, where $v_0 = \sqrt{2gh}$ from conservation of energy. If we use n to denote the bounce number, counting the initial bounce as number zero, we can write this as

$$v_{yn} = \frac{v_0}{\sqrt{2}}.$$

The motion parallel to the plane has a constant acceleration equal to the component of g along the plane, $g/\sqrt{2}$. Therefore, the x-component of the velocity is given by

$$v_x = \frac{gt}{\sqrt{2}},$$

where t is measured from the time the ball is dropped. The zeroth bounce occurs at $t = t_0$, the first bounce at $t = 3t_0$, and the second bounce at $t = 5t_0$. Therefore, the x-component of the velocity just before (and after) each bounce is

$$v_{xn} = \frac{(2n+1)gt_0}{\sqrt{2}} = (2n+1)\sqrt{gh}$$
$$= \frac{(2n+1)v_0}{\sqrt{2}}.$$

This immediately tells us the speed of the ball at each bounce:

$$v_n^2 = \frac{\left[(2n+1)^2+1\right]v_0^2}{2}.$$

It's easiest in this coordinate system to calculate the tangent of the angle with respect to the plane. Just before each bounce, we have

$$\tan\phi_n = \frac{v_{yn}}{v_{xn}} = \frac{1}{2n+1}.$$

Note that both the tangent and the angle approach zero. As the x-component of the velocity increases, the ball hits closer and closer to parallel with the plane. You can obtain the tangent of the angle relative to the true vertical by rotating the coordinate system by 45° using the trigonometric identity

$$\tan(x-y) = \frac{\tan x - \tan y}{1+\tan x \tan y},$$

with $x = \phi$ and $y = 45°$, yielding

$$\tan\theta_n = \frac{n}{n+1}.$$

The coordinate of the ball along the plane is given by

$$x = \frac{gt^2}{2\sqrt{2}}$$

and

$$x_n = \frac{(2n+1)^2 h}{\sqrt{2}}.$$

Therefore, the distance traveled down the plane during each bounce is

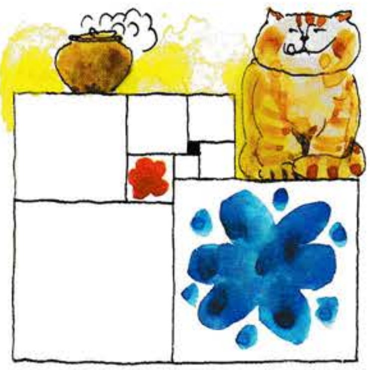
$$L_n = x_n - x_{n-1} = \frac{8nh}{\sqrt{2}} = nL_1.$$

Noah points out a nifty demonstration as an extension of this problem. Release a frictionless puck at $x = 0$ at time $t = 0$. Every time the ball hits the plane, the puck will be directly under the ball. If you cover the ball with ink, there will be no marks on the plane! Noah goes on to point out that the physics behind this demonstration is the same as that of the "monkey-shoot" described in most textbooks.

Just for the fun of it!

B191

Two friendly shadows. Two friends—one tall, the other not so tall—walk away from a lamppost in opposite directions at the same speed. Whose shadow moves faster?



B192

Spiral squares. Look at squares in the picture. Suppose the black one is a unit square. How long are the sides of the squares with the red and blue markings?

B193

Water and ice. One hundred grams of water at a temperature of 10°C is poured into a jar containing 50 g of ice at a temperature of –10°C. What will the equilibrium temperature of the water in the jar be? Neglect the heat capacity and heat conduction of the vessel.

B194

Big fish, little fish. The fish on sale at a certain market come in two sizes: big and little. Today you can buy three big fish and one little fish for the amount you would have spent yesterday on five big fish. On the other hand, two big fish and one little fish cost today as much as three big fish and one little fish cost yesterday. Which is more expensive: one big fish and two little fish today, or five little fish yesterday?

B195

Nail and magnet. A nail like the one pictured on the right lies on a table. How should you maneuver a horseshoe magnet so that the head and the point stick to the magnet simultaneously?

ANSWERS, HINTS & SOLUTIONS ON PAGE 53

Art by Pavel Chernusky

From a Roman myth to the isoperimetric problem

What to do when you have a chance to become "master of all you survey"

by I. F. Sharygin

IN ROMAN MYTHOLOGY there is a legend about Dido (or Elissa), daughter of the Phoenician king Belus of Tyre. She was married to Sychaeus (or Acerbas), the priest of Hercules, who was incredibly rich. Pygmalion, Dido's brother, killed Sychaeus in order to seize his wealth, and Dido had to flee from her motherland with part of her husband's fortune. At last she and her numerous fellow-travelers arrived in Africa, where she bought some land from the Berber king. The condition of this trade was that she could take as much land as a bull skin covered. So Dido cut that skin into several narrow strips and used them to fence in a large plot. There she founded Bursa, a stronghold of Carthage (the Greek word βυρσα means "(bull)skin").

That was the legend. Now here's a well-known puzzle.

Is it possible to make a hole in a sheet of $8^1/_2$-by-11-inch notebook paper such that an adult can easily walk through it?

The similarity between this puzzle and the problem solved by Dido is obvious. Despite the added restriction that you cannot cut the paper into pieces, you can still use Dido's method here. By the way, it's quite possible that Dido herself solved the very same problem, but later the story was misinterpreted or misunderstood. Mythologists, historians, and translators generally pay very little attention to the exact wording of problems. (If this puzzle is still too difficult for you, turn to page 54 and you'll find one possible solution.)

Although from the mathematician's point of view the puzzle is formulated more rigorously than Dido's challenge, it still lacks the precision it needs to be called "mathematical." In general, the business of "posing problems" is the primary task of mathematical modeling, the main question addressed by applied mathematics. In a certain sense, the ability to pose problems correctly is even more important than skill in solving them.

Before we proceed to the mathematical part of this article, let's recall another topic, one that can be found in folk tales and literature. The story "How Much Land Does a Man Need?" by Leo Tolstoy illustrates it very well. Briefly put, the hero of this story is offered as much land as he can walk around in one day.

We might invent a whole series of problems based on these mythological and literary plots. Here's one of the simplest.

Problem 1. Find a triangle with the greatest possible area among all triangles in which the length of one side and the sum of the lengths of the other two sides are fixed.

Solution. Let $2a$ be the given length and $2b$ the sum of the two other lengths. (Note that the Triangle Inequality assures us that $a < b$.) We'll denote one of these unknown lengths as $b + x$, and so the other will be equal to $b - x$ (fig. 1). Hero's formula gives us

$$S^2 = (a + b)(b - a)(a + x)(a - x)$$
$$= (b^2 - a^2)(a^2 - x^2)$$

(since the semiperimeter is $a + b$). Now it's clear that the maximum area is attained when $x = 0$—that is, when the triangle is isosceles.

Note that the correct choice of parameters has greatly helped us

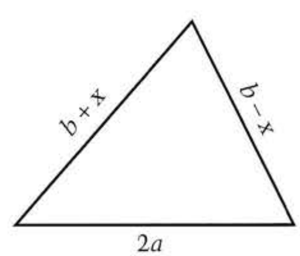


Figure 1

Art by Dmitry Krymov

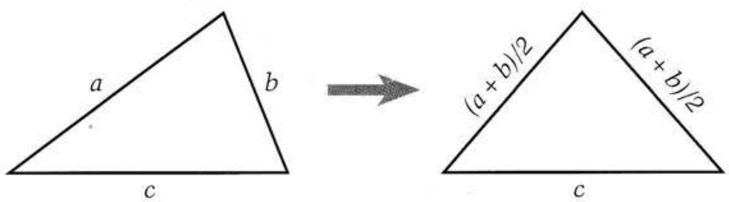


Figure 2

solve the problem. In particular, when the sum of two quantities (here $2b$) is given, we can introduce some symmetry into the situation by denoting them as $b + x$ and $b - x$.

Problem 2. Prove that of all triangles with a given perimeter, the equilateral triangle has the greatest possible area.

It's important to note that we'll use in our proof the fact that such a triangle *exists*.

Proof. Consider a triangle with the given perimeter and the greatest possible area. Suppose that it is *not* equilateral. This means that at least two of its sides do not equal each other (fig. 2). Let's denote their lengths as a and b ($a \neq b$) and let c be the length of the third side. It follows from the previous problem that the triangle with sides $(a + b)/2$, $(a + b)/2$, c has a greater area and the same perimeter $a + b + c$. But this contradicts our initial assumption.

Further, we can prove in the same way that the regular n-gon has the greatest area of all n-gons with the same perimeter. If we now let n increase to infinity (as one does to find the formula for the circumference of a circle), we'll reach a circle as the "limit" of this process. So we see that the circle bounds the greatest area among all closed curves of a given length. However, we won't take this long and cumbersome route, but rather proceed directly to the basic problem. First let's formulate it correctly.

Problem 3. Among all closed curves of a given length on the plane, find the curve that encloses a figure with the greatest possible area.

This is the famous *isoperimetric problem.* (This comes from the Greek ισοσ—"equal" and περιμετρεο—"measuring around." We know what the term "perimeter" means, when applied to polygons, but it can be extended to an arbitrary figure as well.) It's also called "Dido's Problem."

Once again we'll base our reasoning on the *existence* of such a figure, as we did above. In other words, we suggest that among all the lines of the given length, there exists one that encloses the greatest area. (This is easy to accept if we think of all the possible areas of figures with a given perimeter and note that the set values this area takes on are bounded.) Below we reproduce the solution found by the outstanding Swiss geometer Jacob Steiner (1796–1863).[1]

Solution. First of all, we note that the desired figure must be convex (that is, for any two points lying within it or on its border, the segment connecting them must also lie within this figure or on its border). Otherwise we would be able to construct a figure with the same perimeter and a greater area (see figure 3).

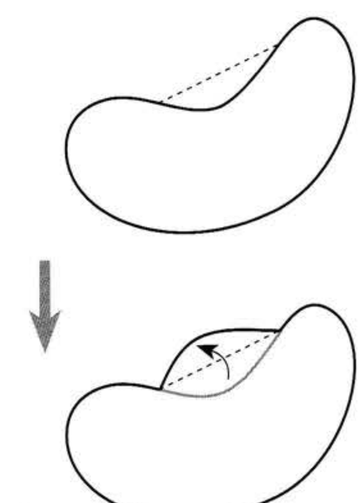


Figure 3

[1]To learn more about Steiner, see the *Quantum* article "From a Snowy Swiss Summit to the Apex of Geometry" (November/December 1993, p. 35).—*Ed.*

We note further that if a line divides its perimeter in half, then it must divide its area in half as well. Indeed, let AB be a line that divides the perimeter of our figure in half (A and B lie on the boundary—see figure 4). Suppose one of the parts created has a greater area than the other. If we now replace the part with the lesser area with the part having the larger area, reflected with respect to the line AB, we see that the area of the figure has increased, while its perimeter remains unchanged. So a line bisecting the perimeter of a figure with the largest possible area must bisect its area as well.

Let's look at figure 4 more closely. Suppose point M is taken on the boundary of our figure, different from A and B (fig. 5). We'll prove that $\angle AMB = 90°$. Suppose this isn't true. Draw the segments AM, MB, and AB. They cut our figure into four sectors, labeled I, II, III, and IV. Construct a new figure in the following way: draw the right triangle $A_1M_1B_1$, in which $A_1M_1 = AM$, $M_1B_1 = MB$, and $\angle A_1M_1B_1 = 90°$. Attach sectors equal to I and II on its legs. Finally, reflect all this with respect to the hypotenuse A_1B_1. We obtain a figure with equal perimeter and greater area, since the area of triangle $A_1M_1B_1$ is greater than that of triangle AMB. Thus we have shown that for any line AB dividing the circumference of a figure with the greatest possible area in half, having M as an arbitrary point on its border different from A and B, then $\angle AMB = 90°$. But this means that M lies on a circle with diameter AB. Therefore, the circle is the solution to the isoperimetric problem.

Lest you think this is an isolated puzzle with no real consequences, I

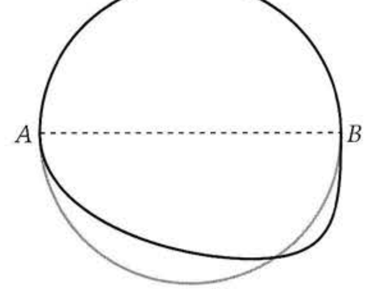


Figure 4

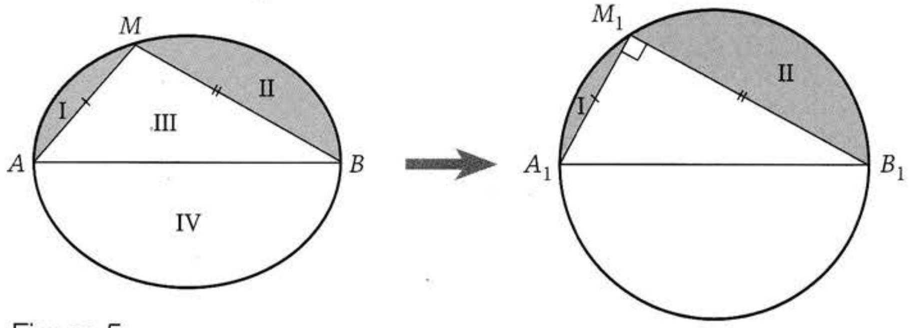


Figure 5

should point out that the isoperimetric problem essentially gave birth to one of the most important branches of modern mathematics—the calculus of variations.

At this point it would be reasonable to conclude our discussion, but we'll take just one more step (I don't even know whether it's a step forward or backward.) As I mentioned above, the route from polygons to the circle seemed most natural. Yet the method we used to solve the isoperimetric problem doesn't rely at all on the properties of polygons with maximal area. Quite the contrary, some properties of these polygons can be derived rather easily from this general result, while proofs of these qualities that do not rely on this general result are very unwieldy. Here's an example.

Problem 4. Consider all possible polygons with given sides in a given order. (One can imagine a polygon whose sides are connected by hinges. We consider any polygon that is the result of deforming such a polygon.) Prove that among them there exists one inscribed in a circle, and that it has the greatest area.

Solution. Consider a circle that is large enough that if we draw chords equal to the polygon's sides one after the other, in a fixed direction starting from an arbitrary point A on the circle, the sum of the corresponding arcs is less than the whole circle. Let B be the end of the last chord (fig. 6). Reduce the radius of this circle until points A and B coincide. Now we've got the desired polygon (fig. 7a). Let's prove that it has the greatest area. Consider an arbitrary polygon with the given sides. Construct on its sides sections equal to those on the corresponding sides of the inscribed polygon (fig 7b). The length of the border of these sections is equal to the length of the circumscribed circle. Thus, because of our last result, the area bounded by the arcs of these sections is less than that of the circle in figure 7a. Removing the sections, we see that the area of the inscribed polygon (fig. 7a) is greater than the area of any other polygon with the same sides (fig. 7b).

Now let's recall our basic isoperimetric problem (problem 3). The very fact that a solution *exists*—a fact evident enough for a mind unspoiled by mathematical truisms—allows one to find it. In connection with this, we'll tell you a story that belongs to the folklore of modern mathematics.

Once upon a time, a difficult scientific problem was assigned to a group of scientists, which included a mathematician and a physicist. When they met again some time later, the mathematician happily told the physicist that he had managed to prove the existence of a solution to this problem. The physicist remarked that if he had ever had the slightest doubt that a solution existed, he would never have bothered with such a problem!

Problems

1. In a triangle, one side is equal to a and the opposite angle is equal to α. Which such triangle has the largest area?

2. Consider an arbitrary triangle whose area is 1. What is the least possible perimeter of such a triangle?

3. Let AB be chosen on a fixed circle. Find points K and M on minor arc AB such that quadrilateral $AKMB$ has the largest possible area.

4. For any figure with perimeter l and area Δ, prove that $l^2 > 12.5\Delta$.

5. A set of segments is given. Consider all possible polygons whose sides are equal to these segments, taken in an arbitrary (but fixed) sequence. Prove that the greatest possible area of such polygons does not depend on the sequence chosen.

6. Consider all n-gons with $n - 1$ given sides. Prove that the greatest area is that of the n-gon that is inscribed in a circle whose "free" side coincides with the diameter of the circumscribed circle.

ANSWERS, HINTS & SOLUTIONS ON PAGE 54

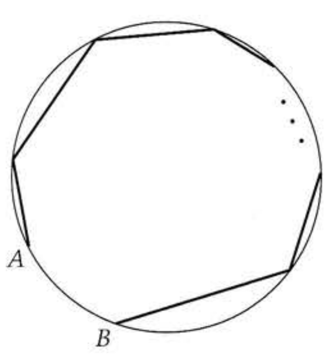


Figure 6

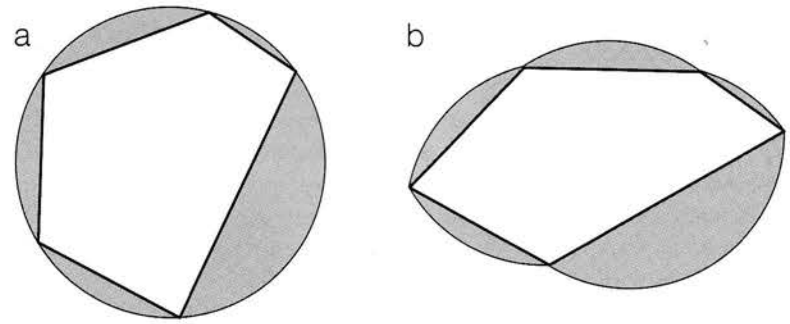


Figure 7

The green flash

An unusual spectacle at the close of day

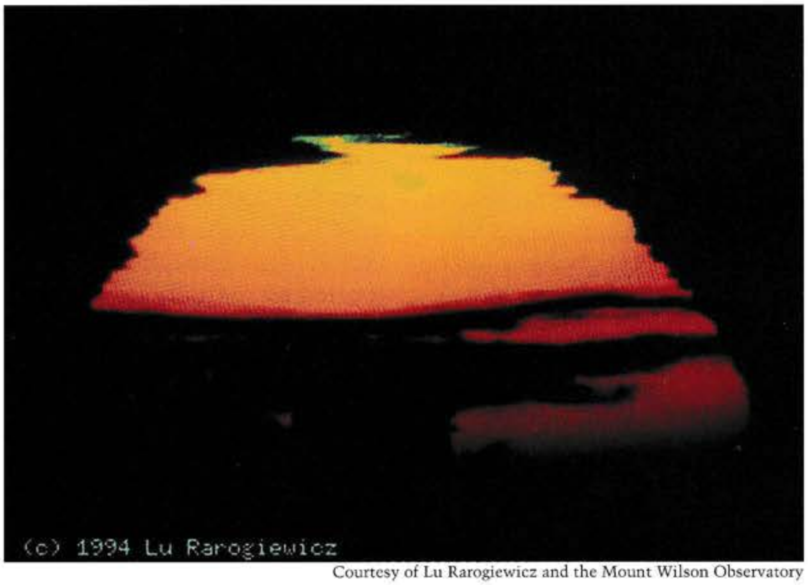
by Lev Tarasov

SOMETIMES AT SUNSET YOU can see a wonderful phenomenon—the so-called "green flash." When the disk of the Sun is almost entirely hidden beyond the horizon, suddenly a bright green light flashes out for a few seconds. The edge of the Sun becomes a startling bright green instead of the usual yellow-reddish color, and it emits green rays in every direction. One, two, three seconds . . . and the beautiful sight disappears.

This green light is a rare guest in the heavens. It's usually seen on evenings when the Sun shines brightly up to the very moment that it sets and doesn't change color, remaining yellow or yellow-orange. The Russian astronomer G. A. Tikhov studied this phenomenon for many years. He wrote: "If the Sun at sunset is red and it is easy to look at it, then one can say with certainty that there will be no green flash. On the other hand, if the white-yellow color of the Sun does not change markedly and the Sun sets in full brightness, one can expect the green flash to appear. It is important that the horizon be a clear-cut line without any variations in height (trees, buildings, and so on). These conditions are most readily encountered at sea, which is why the green flash is quite familiar to mariners."

The physics of the green flash is based on three phenomena: (1) the refraction of light in an optically nonhomogeneous medium (here, the Earth's atmosphere); (2) the dependence of light refraction on wavelength; and (3) light scattering in the atmosphere (or, to be more precise, the weakening of this effect in clear, calm air). Let's look at these factors one by one.

When a light beam passes through the atmosphere, it deviates in such a way that the curvature of its trajectory is always directed toward layers of air that are less dense. This is why the setting Sun seems a bit flattened in the vertical direction—its vertical diameter is seen at an angle of 26′, which is 6′ less than the angular measure of its horizontal diameter. What causes this discrepancy? The density of the atmosphere decreases with altitude, and this leads to a corresponding diversion of the light beams (that is, refraction). When we see, while admiring the sunset at the seashore, the lower edge of the Sun's disk touch the horizon, we usually don't realize that the observed



Courtesy of Lu Rarogiewicz and the Mount Wilson Observatory

Readers with Internet access can find photographs (like the one above) and additional information on the "green flash" and related phenomena at the Mount Wilson Observatory's Web site:

http://www.mtwilson.edu/Tour/Lot/Green_Flash

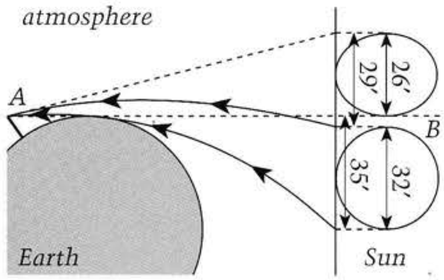


Figure 1

edge of the Sun is actually 2°35′ below the horizon at this point (which means that the entire disk is already well below the horizon). The number 2°35′ consists of two parts: the 2° portion is explained by the time it takes for sunlight to arrive at the Earth (a little more than 8 minutes), and the 35′ portion is due to the refraction of light in the atmosphere. Note that, because of refraction, the upper edge of the disk "rises" less than the lower one: not by 35′ but only by 29′ (refraction decreases with altitude). This is why the setting Sun looks like a flattened disk. Figure 1 illustrates our reasoning (*A* is an observer, *AB* is the horizon).

Now let's take into account point (2): the shorter the wavelength, the larger the refraction. This relationship explains how "white" light is split into a spectrum of colors when it passes through a prism. Green-blue rays are refracted more than red rays. For simplicity, imagine that sunlight consists of only two colors: green and red. In this simplification, the "white" disk of the Sun can be considered as a combination of two disks (green and red), one placed on top of the other. Atmospheric refraction "lifts" the green disk above the skyline higher than the red one.

As a result, an observer sees the setting Sun just as shown in figure 2. The upper edge of the disk is green, and the lower edge is red. The center is composed of both colors and looks "white."

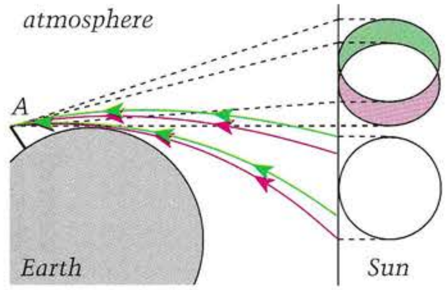


Figure 2

This scenario is correct as long as there is no scattering of light in the atmosphere. In reality, the atmosphere does scatter light. This means that rays with shorter wavelengths are more likely to be absent in the light coming from the Sun. (This law was formulated by Sir John Rayleigh: the intensity of scattered light is inversely proportional to the fourth power of its wavelength.) So we won't see the green edge on the upper part of the Sun, and the entire disk won't be "white" but reddish.

But imagine that almost the entire solar disk has disappeared below the horizon and only a tiny upper edge can still be seen. If the weather is clear and calm and the air clean, the scattering of light will be rather weak. These are the conditions that can produce a marvelous spectacle: a setting Sun with a bright green edge, casting a fan-shaped splash of green rays.

The "water worm"

"To appreciate how excellent it is I leave to those who have acquired a deep understanding of it."
—Galileo Galilei

by M. Golovey

THE EARLY DEVELOPMENT of physics is closely connected with the invention of various mechanisms to make difficult tasks easier. Using very primitive methods and materials, and applying the basic principles of geometry, mechanics, and hydraulics that they had developed in solving practical problems, the engineers of antiquity devised a variety of machines that impress us even today with their clever design and sound construction, and reflect a profound understanding of the problem at hand.

One of the most brilliant inventors in ancient times was Archimedes. Though nowadays his name is usually linked with the buoyant force acting on an object immersed in a fluid, Archimedes devised a vast number of remarkably clever inventions. The greatest of them is the device know through the centuries as "Archimedes' screw." In one of his papers Galileo wrote that this invention "was not merely beautiful, it was simply miraculous, because we see that the water rises in the screw while continuously flowing downward."

Indeed, at first glance it's hard to imagine how the water flowing down the screw's blade manages to ascend at the same time. Galileo described the working principle of this water pump in great detail and called it the "water worm" of Archimedes. It actually does look like a worm from a distance.

But how does the water worm work? Look at figure 1. It shows a cylinder *AKCB* whose side surface has a spiral channel *AEOPXTHC*. Water can flow down this channel without spilling out. A ball can likewise roll down the channel. If we incline the cylinder, immerse the end *A* of the helical channel into some water (fig. 2), and begin to rotate it around the supporting pivots *M* and *M′*, water will run up within the spiral channel and pour out of the upper end *C*. Paradoxically, during the entire time water indeed runs down relative to the surface of the spiral channel—but nevertheless it rises up to point *C* as a result!

Let's try to understand how this is possible. Let the segment *AE* of the spiral channel form an angle α with the cylinder's bottom *AB*. Incline the cylinder at an angle that is larger than α. Water now fills the segment *AE* and also part of *EO*.

Now begin to rotate the cylinder about its axis. All the points in section *AE* (except point *A*) start to rise, while all the points in section *EO* (except point *E*) drop, and water gradually flows from the "flooded" section farther on along the channel, dropping down relative to it but ascending relative to the ground. During rotation the different parts of the spiral channel, continually replacing one another, are positioned relative to the water in just the same way as

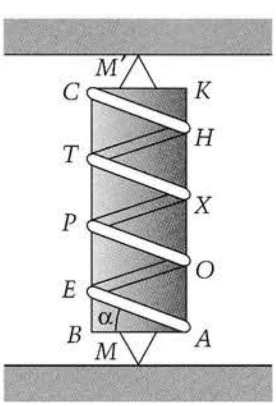


Figure 1

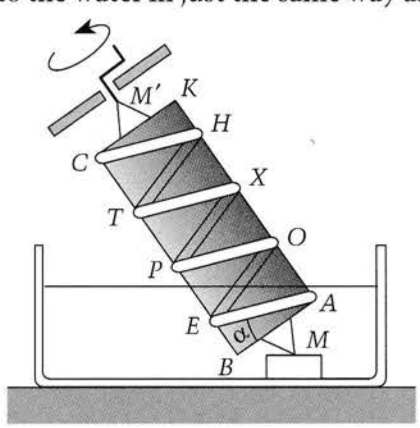
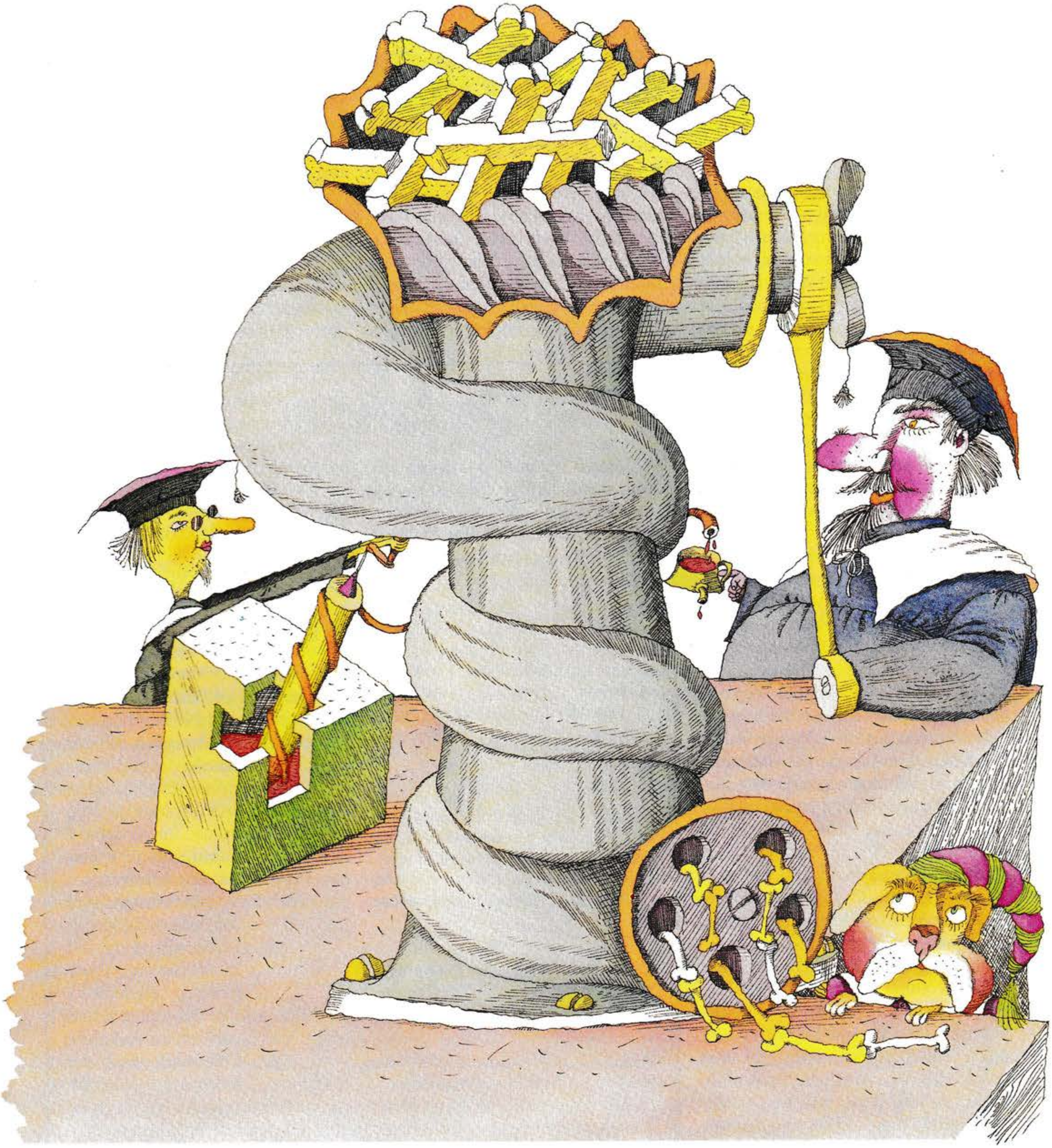


Figure 2

Art by V. Ivanyuk



section AE was, so the water simultaneously runs down the channel and rises up due to the channel's rotation. Finally, it rises from point A to point C.

Some readers are perhaps ready to pose a natural question: "Doesn't this device violate the law of conservation of energy, making it a kind of perpetual-motion machine?" It may seem that way at first. We rotate the cylinder, and due to its own weight the water runs down the conduit as if it were on an inclined plane, acquiring velocity and rising to the upper outlet of the worm (that is, its potential energy increases). We need only turn the extra energy from the water into the rotational energy of the worm to create a *perpetuum mobile*!

If you were asked in physics class to solve this problem, you might refer to the law of conservation of energy and point out that the reasoning above does indeed contradict it. You might even go further and add that the water rises in the conduit due to energy taken from some other source and expended on rotating the worm, such that the potential energy of the water is always less than the energy used to lift it.

This is all true enough, but it's too "theoretical" and doesn't explain the details. How does the conduit transfer energy to the water? Where does it occur? Where is the force acting on the water directed? Try to answer these questions yourself. To simplify the problem, you can replace the motion of water with the movement of a small ball in the spiral channel.

A simple model of the Archimedean worm can be made at home. Take a piece of tubing 3–5 cm in diameter and 20–30 cm long, or a round wooden rod of a similar size. The conduit can be made of any rubber or PVC (polyvinyl chloride) tube a few millimeters in diameter (the larger the diameter, the better). It would be best if you use a transparent flexible hose—then you can see the motion of the water or ball inside the Archimedean worm.

Wind the hose around the rod in a spiral and fasten its ends with wire or string near the upper and lower ends of the rod. That's it—your model is ready to use. Immerse one end into a jar of water and tilt the worm at an angle larger than that formed by the spiral and the cylinder's base. Now rotate the worm as if you were driving a screw into the water, holding it at a constant height. After a dozen turns, water will run out of the upper end of the worm. Instead of water you can put ball bearing inside the hose at its lower end. The ball will be lifted to the upper end of the worm and will drop out of this unusual conduit.

In much the same way you could make a real pump driven by wind power, say, or a gas engine. Compared to a piston pump, which only works with clean water, the Archimedean worm has an important advantage: it can pump water with particles (dirt, metal shavings, etc.) suspended in it. Compared to centrifugal pumps, however, the Archimedean worm is sadly outdated and cannot compete. It's very rarely used to pump water nowadays. But you can still come across a working specimen of the Archimedean worm in some developing countries.

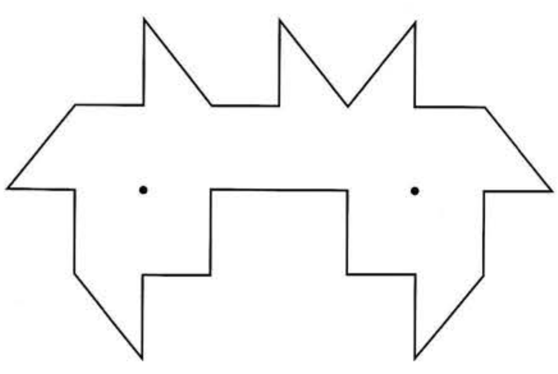
The type of device shown in the figures is also called a "spiral conveyer" in engineering parlance, and it's widely used in many machines to mix liquid, dry, and doughy substances. One example is the rotor of the common meat grinder.

Corrections

David Castro, a student at Macalester College in Minnesota, e-mailed us to correct a misstatement on page 29 of the November/December issue (in the article "Billiard Math"):

The question whether there exists a polygon with mirrored walls and internal points A, B such that a light at A will not illuminate B has been settled in the affirmative. The solution was published by George Tokarsky in the December 1995 issue of the *American Mathematical Monthly* ("Polygonal Rooms Not Illuminated from Every Point," pp. 867–79). In this article, Tokarsky shows how a class of such polygons can be constructed by taking mirror reflections of a certain set of triangles in the plane.

The problem in the *Quantum* article was pointed out to me by my professor, Stan Wagon, because I researched the question of illuminating polygons in his Geometry class (Spring 1996). In his article, Tokarsky shows a 26-sided polygon with the desired property; I saw that his example could be modified to produce a 24-sided polygon (see the figure below).



(Editor's note: the original Russian version of "Billiard Math" appeared in the November/December 1995 issue of our sister magazine *Kvant*.)

Professor Mario Velucchi e-mailed us from Pisa, Italy, to suggest that the figure for exercise 1 in the September/October Toy Store is problematic in a way not intended. The diagram as printed has "two solutions: 1. Kb3 and 1. Bd6, and this is bad for a chess problem." He suggests that the correct initial position for the white king is c3, "with the only right solution 1. Kb3."

We thank Mr. Castro and Prof. Velucchi for their feedback.

HOW DO YOU FIGURE?

Challenges in physics and math

Math

M191

Star qualities. An operation * is defined on a set S in such a way that (i) for any three elements a, b, c of S, $a * (b * c) = b * (c * a)$; (ii) if $a * b = a * c$, then $b = c$. Show that the operation * is (a) commutative—that is, for any two elements a, b of S, $a * b = b * a$; (b) associative—that is, for any three elements a, b, c of S, $a * (b * c) = (a * b) * c$.

M192

Erase and replace. A set of n positive numbers is written on a blackboard. A student is allowed to choose any two of them—say, a and b; erase them; and replace them with the number $(a + b)/4$. This operation is repeated $n - 1$ times, until a single number remains on the blackboard. Show that that if the n original numbers were all equal to 1, the single number left is no less than $1/n$. (B. Berlov)

M193

Onesy-twosy. Show that for any natural number n, there is a number whose decimal representation contains n digits, each of which is 1 or 2, and which is divisible by 2^n.

M194

Dividing in half. A quadrilateral $ABCD$ can be inscribed in a circle. Let straight lines AB and CD meet at M, and let BC and AD meet at K, so that B lies on the segment AM and D on the segment AK. Let P be the foot of the perpendicular from M to line AK, and let L be the foot of the perpendicular from K to line AM. Prove that LP bisects BD.

M195

Yes or no. A detective must question a witness about a crucial detail of a particular crime. The detective has concocted a series of at most 91 "yes/no" questions whose answers will guarantee him knowledge of the detail, provided the witness does not lie (the particular question asked may depend on an answer to one or more previous questions).

But suppose the witness lies no more than once. Show that the detective can revise his plan of questioning to include at most 105 "yes/no" questions and still ferret out the crucial detail.

Physics

P191

Sleigh on ice. A long sleigh glides on the surface of very smooth ice, drives onto a stretch of asphalt, and stops after traveling less than half its length. The sleigh is given a push so that it again acquires the same initial velocity, travels a distance, and stops for the second time. What is the ratio of the braking times and the braking distances in the first and second cases? (S. Krotov)

P192

Two metal bars. Two bars made of different metals with coefficients of linear expansion α_1 and α_2 have practically equal lengths (l_1, l_2) and cross-sectional areas (S_1, S_2) at 0°C. At what temperatures will the bars have equal (a) lengths, (b) cross-sectional areas, and (c) volumes? (B. Bukhovtsev)

P193

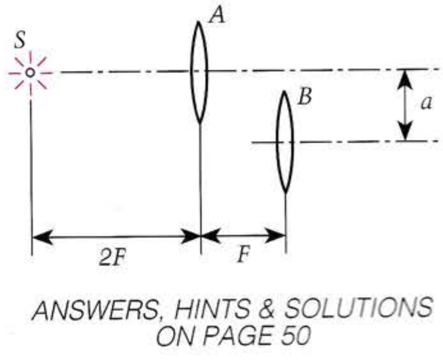
Three charged plates. A capacitor is formed from two large square plates of area S placed a small distance d apart. A third plate of the same material carrying a charge Q is located midway between them. The outer plates are connected via a resistor with a high resistance R. The central plate is quickly moved toward one of the outer plates, to a distance of $d/3$ from it. How much heat will be dissipated by the resistor after this shift? (A. Zilberman)

P194

Solenoid and cylinder. A constant current I flows through a long solenoid of radius r that has N turns per meter. It is known that outside such a coil the magnetic field is weak, while inside the coil it is practically uniform. On the same axis with the solenoid there is a long (though not as long as the coil itself) paper cylinder of radius R and length L whose surface is uniformly charged with a total charge Q. The current in the coil is decreased by a factor of three, which causes the cylinder to spin about its axis. What is the direction and angular velocity of this rotation? (A. Zilberman)

P195

Shifted lenses. Find the distance between a light source S and its image in the optical system shown in the figure below. Both lenses have the same focal length F. (V. Serbo).



ANSWERS, HINTS & SOLUTIONS ON PAGE 50

AT THE BLACKBOARD II

The first bicycle

It needed a little refinement . . .

by Albert Stasenko

ABOUT TWO MILLION YEARS ago (according to the Elders), the great inventor Nga-Nga constructed the first bicycle, without knowing that the wheels should be round . . . Well, how could he know? Since that time many questions have arisen in bicycle theory. For instance, how much power is expended in its motion? What is the most efficient way to ride a bike? How can one make it more stable? And so on. This article will address only a few of them.

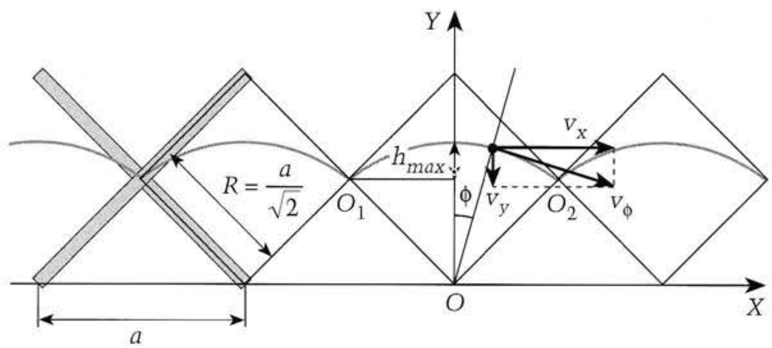


Figure 1

First, let's look at a "wheel" from the first bicycle—two sticks lashed together (fig. 1). Let's measure the angle of the wheel's rotation from the vertical axis OY to the right. Position 1 corresponds to the angle $\phi_1 = -\pi/4$, and position 2 corresponds to the angle $\phi_2 = +\pi/4$. In moving from position 1 to position 2, the "axle" describes the quarter-circle O_1O_2 of radius $R = a/\sqrt{2}$ with the center O (we assume that the wheel does not slip at all). Denoting the linear velocity along the arc by v_ϕ, we get the following formulas for the horizontal and vertical projections of the velocity of the wheel's center:

$$v_x = v_\phi \cos\phi, \qquad v_y = -v_\phi \sin\phi. \tag{1}$$

The dependence of these projections on the angle ϕ is shown in figure 2. We can see that after each quarter-turn the vertical velocity changes abruptly—it heads in the opposite direction (that is, changes sign) but preserves its magnitude.

If the bicycle travels with a constant linear velocity v_ϕ, the angle ϕ will be directly proportional to the time. In this case, figure 2 shows the oscillations of the projections of velocity versus time. Of course, these oscillations can't be called "harmonic" even though they're described by the trigonometric functions (1).

Starting from position 1, the wheel's center will reach the highest point, located a distance R from the ground, after one eighth of a turn. So, with respect to point O_1, it rises to the height

$$h_{\max} = R - \frac{a}{2} = \frac{a}{2}\left(\sqrt{2} - 1\right).$$

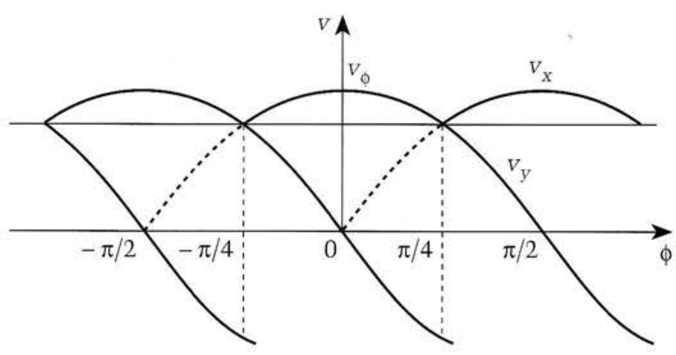


Figure 2

However, one wheel does not a bicycle make, as the people said in the time of Nga-Nga. Looking at figure 3, the great inventor guessed at once that at least two modes of operation are possible.

A. If both wheels are set initially in

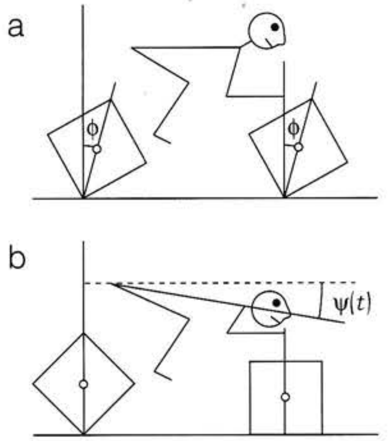


Figure 3

the same position and there is no slippage, the rotating "diameters" will always form the same angle ϕ with the vertical. Their centers will always be at the same height (which of course, depends on time). As the great Nga-Nga liked to say, the centers will oscillate *synchronously*, or *coherently*. In this case all the points of the bike and the cyclist will also oscillate vertically. The center of mass of such a bike will periodically rise to the height h_{max}, which we found earlier.

B. If the wheels are not set in the same position at the outset but are turned as in figure 3b, their centers will move with a phase shift of 45°. During this ride Nga-Nga's backbone will oscillate in a circle ($\psi(t)$), but at the same time the center of mass can move strictly horizontally.

Let's consider mode A. To lift the center of mass of the entire system to the height h_{max} we need to perform the work

$$W = mgh_{max} = mg\frac{a}{2}\left(\sqrt{2}-1\right).$$

If the velocity v_ϕ is kept constant, this work will be performed in the time required for a one-eighth turn of the wheel:

$$\tau_1 = \frac{1}{8}\frac{2\pi R}{v_\phi} = \frac{\pi}{4\sqrt{2}}\frac{a}{v_\phi}.$$

So the average power developed by the cyclist in that period is about

$$\frac{W}{\tau_1} = \frac{2\sqrt{2}}{\pi}\left(\sqrt{2}-1\right)mgv_\phi. \qquad (2)$$

What velocity v_ϕ can a cyclist generate on this bicycle out of her or his own energy resources? We know that a person used to physical labor (and Nga-Nga was no softy) needs about 5,000 kilocalories per day, and the efficiency of human labor is approximately 25%. So the power of a "human machine" is about

$$\begin{aligned} P_{av} &= 5000 \cdot 10^3 \text{ cal/day} \\ &\times \frac{4.2\,\text{J/cal}}{3600\,\text{s/h} \cdot 24\,\text{h/day}} \cdot 0.25 \\ &\cong 60\text{ W}. \end{aligned}$$

Estimating the total mass of the bike and bicyclist as $m \sim 100$ kg, we derive the velocity from equation (2):

$$v_{\phi_1} = \frac{P_{av}\pi}{mg2\sqrt{2}\left(\sqrt{2}-1\right)} \cong 15\text{ cm/s}.$$

However, this is true only for the first one eighth of the wheel's rotation. After reaching the height h_{max}, the biker can take a rest. Neglecting any energy loss, the rider drops from the height h_{max} in a time

$$\tau_2 \sim \sqrt{\frac{2h_{max}}{g}}$$

with a vertical velocity

$$v_{\phi_2} \sim \sqrt{2gh_{max}}.$$

If the shock from this velocity is too great, the cyclist can use reverse pressure on the pedals to descend more smoothly, but this mode requires extra work. Also, the period of oscillation will be increased, so the bicycle will go slower. This isn't a very comfortable mode of bike riding.

Mode B is certainly more attractive from the energy standpoint: one doesn't need to keep raising the center of mass and then slowing its fall. But all the same, this rumble of square wheels, this scrambling of the brain... This is why Ngo-Ngo, Nga-Nga's son, improved the design and made an octagonal "wheel." Nga-Nga's grandchild Ngi-Ngi went even further—her bicycle had 16-gon wheels. And so it went, until Archimedes himself proved that this succession of wheels tended to a circle. (But now we're getting into mathematics . . .)

It turned out that round wheels required roads, while the first bicycle of Nga-Nga could ride over stones and tree trunks. Could it be that the great Nga-Nga was a lot more clever than we give him credit for?

Revisiting the *N*-cluster problem

With thanks again to the creators and contributors of Math.Note

by George Berzsenyi

LET N BE A POSITIVE INTEGER, and let $C(N)$ be a set of N points in the plane with integer coordinates, so that no 3 of the points are on a line and no 4 of them lie on a circle. We will say that $C(N)$ is an N-cluster if the distance between each pair of its points is an integer. We will say that $C(N)$ is a primitive N-cluster if it cannot be obtained from another N-cluster by scaling—that is, by multiplying each of the coordinates of its points by a positive integer. Clearly, the property of being an N-cluster is invariant under translations (that is, adding the same integer X to the x-coordinate and the same integer Y to the y-coordinate of each of its points) under reflections in the lines $y = \pm x$ and under rotations of 90°. Furthermore, one can define the size of an N-cluster either by the longest distance between any two of its points or as the radius of the smallest circle centered at the origin that contains a translate of the N-cluster under consideration. Using the latter criterion, Stan Rabinowitz (who was the originator of the Math.Note file at DEC) found that the "smallest" 5-cluster (of size 56) is {(0, 0), (56, 0), (–16, 30), (16, 30), (0, –33)}, and the "smallest" 6-cluster (of size 1275) is {(0, 0), (1155, 540), (546, –272), (132, –720), (960, –720), (546, 1120)}. (For more information about Math.Note the reader is referred to page 38 of the September/October 1996 issue of *Quantum*.) Earlier, the originators of the concept of N-clusters (Bell and Chongo, whom I couldn't track down), found the 5-cluster {(0, 0), (0, –153), (136, 102), (–136, 102), (224, 207)}, which is definitely larger than Stan's.

The purpose of this column is to direct the attention of *Quantum*'s readers to interesting problems in the literature that deserve to be generalized and could lead to independent research and/or science projects in mathematics. Students who succeed in unraveling the phenomena presented are encouraged to communicate their results to the author either directly or through *Quantum*, which will distribute among them valuable book prizes and/or free subscriptions.

Clearly, 3-clusters can be constructed from Pythagorean and Heronean triples. A triple (a, b, c) is said to be Heronean if a, b, c are integers and if the area of the (Heronean) triangle with sides a, b, c is also an integer. My first challenge to my readers is to **prove that all Heronean triangles can be obtained by juxtaposing (putting together) two Pythagorean triangles**. My second challenge is to **prove or disprove that there are no other 3-clusters**.

When $N = 4$, the construction of 4-clusters is complicated by the fact that the points should not lie on a circle. Nevertheless, one can construct infinitely many different primitive 4-clusters by starting with Heronean triangles, which are not Pythagorean, and reflect them along their sides. My next challenge to my readers is to **find 4-clusters which cannot be split into two Heronean triangles in the manner indicated**.

There are many other open questions concerning N-clusters. Some of these were already considered by Stan Rabinowitz and his colleagues at DEC, but they have made no further advances on them. For example:

- **Are there N-clusters for $N > 6$?**
- **Are there infinitely many primitive 5-clusters and 6-clusters?**
- **Is there a way to generate an $(N + 1)$-cluster from a given N-cluster?**
- **What if we move our considerations to the 3-dimensional lattice?**
- **What is the situation in higher dimensions?**
- **What if we change to the triangular lattice structure in the plane, where the pre- and post-Pythagorean triangles[1] take the place of the Pythagorean triples?**

Please send your findings to me c/o *Quantum*, 1840 Wilson Boulevard, Arlington VA 22201-3000. Perhaps they will generate further discussions in a future column.

[1]Discussed in my March/April 1992 and September/October 1996 *Quantum* columns.

Bulletin Board

A summer PROMYS

The Program in Mathematics for Young Scientists (PROMYS) will soon be accepting applications for its 1997 program at Boston University, June 29 to August 9. PROMYS offers a lively mathematical environment in which ambitious high school students explore the creative world of mathematics. Through their intensive efforts to solve a large assortment of unusually challenging problems in number theory, the participants practice the art of mathematical discovery—numerical exploration, formulation and critique of conjectures, and techniques of proof and generalization. More experienced participants may also study group theory, dynamical systems, and combinatorics. Problem sets are accompanied by daily lectures given by research mathematicians with extensive experience in Prof. Arnold Ross's long-standing Summer Mathematics Program at Ohio State University.

PROMYS is a residential program for 60 students entering grades 10 through 12. Fifteen college-age counselors live in the dormitories and are available to discuss mathematics. Each participant belongs to a problem-solving group that meets with a professional mathematician three times a week. Special lectures by outside speakers offer a broad view of mathematics and its role in the sciences.

Admission to the program will be based on applicants' solutions to a set of problems, teacher recommendations, high school transcripts, and student essays. Tuition costs are still to be determined. Financial aid is available—PROMYS is dedicated to the principle that no student will be unable to attend because of financial need.

Application materials can be obtained by writing to PROMYS, Department of Mathematics, Boston University, 111 Cummington Street, Boston MA 02215, or by calling 617 353-2563. Applications will be accepted from March 1 until June 1, 1997.

The dirt on germs

How could a young woman catch tuberculosis just by breathing the air in her high school classroom? You will find the answer in *The Race Against Lethal Microbes*, the latest report published by the Howard Hughes Medical Institute (HHMI) and available free of charge to teachers. The beautifully illustrated report describes recent progress in our struggle against myriad bacteria, viruses, and parasites that cause infectious diseases. It tracks the ominous spread of drug-resistant microbes and shows how scientists are fighting these invaders with the aid of molecular biology.

A fold-out poster in the center of the publication highlights 29 bugs that cause disease, explaining how such microbes enter our bodies, how they injure us, how we fight them, and how microbes fight back (for example, by sharing drug-resistant genes).

HHMI is a scientific and philanthropic organization that conducts medical research. Previous titles in this series include *Finding the Critical Shapes*, *Blazing a Genetic Trail*, and *Blood: Bearer of Life and Death*.

Copies of the report can be requested by writing to Howard Hughes Medical Institute, Office of Communications, 4000 Jones Bridge Road, Chevy Chase MD 20815-9864.

A cozy CyberTeaser (B192)

To win a *Quantum* button in the CyberTeaser contest at our Web site, you not only have to be smart—you've got to be quick. These ten people were both:

Oleg Shpyrko (Cambridge, Massachusetts)
Pasquale Nardone (Brussels, Belgium)
Matthew Wong (Edmonton, Alberta)
Ken Wharton (Berkeley, California)
Leo Borovskiy (Brooklyn, New York)
Gary Sega (Oak Ridge, Tennessee)
Jim Grady (Branchburg, New Jersey)
Steve Hunter (Ascot, Berkshire, UK)
Graeme D. MacDonald (Wheaton, Illinois)
Joyce Bossom (Lincoln, Massachusetts)

In addition to making it into the top ten, veteran CyberUnTeaser Leo Borovskiy sent a nice ASCII graphic. And Jonathan Devor of Jerusalem, Israel, enhanced his entry with a handsome JPEG, which can be seen on the solution page at the CyberTeaser Web site.

Most of the answers, whether from Croatia or Connecticut, were pretty straightforward (and correct, we might add). But one correspondent had more than little squares on his mind:

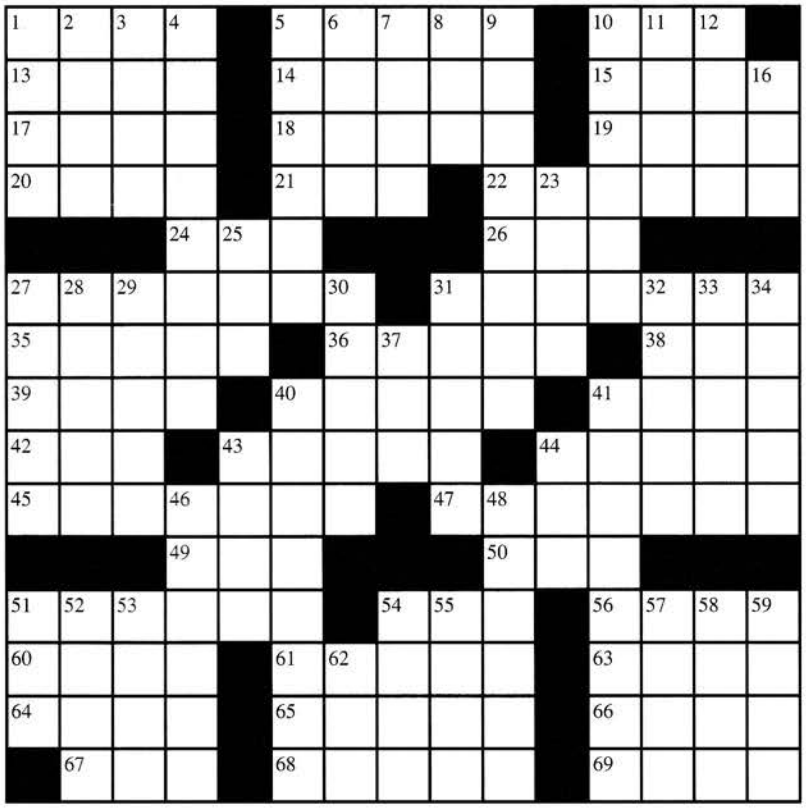
> After doing extensive research on anti-entropic, polymorphic, isotropic, splatter-colored squares, I have come to the conclusion that the red must have 4-unit sides and the blue, 16-unit long sides.
>
> If you say that this is not the correct answer, you are wrong and will never be accepted into our very prestigious Quasi-Scientific Society. You will walk the halls of ignorance forever and be forgotten by history.

Ouch!

A new CyberTeaser awaits your anti-entropic efforts at **www.nsta.org/quantum**. Good luck!

Criss × cross science

by David R. Martin



Across

1 Mild oath
5 699,325 (in base 16)
10 Soft-walled cavity
13 Venetian magistrate
14 Talus
15 Unit of pressure
17 Charles ___ (Blood researcher)
18 Robber
19 60,334 (in base 16)
20 Shout
21 Brain graph: abbr.
22 Micalike clay minerals
24 Row
26 After deductions
27 Conic section
31 Like uranium
35 Singer Frankie ___
36 Separates
38 Threading device
39 Doesn't exist
40 Receiver
41 Group of organisms
42 Earn with great effort
43 Taped movie
44 Oise tributary
45 ___ map (of genes)
47 ___ phase (of a liquid crystal)
49 Samuel's teacher
50 Dry single-seeded fruit
51 ___ lens system
54 Honest ___
56 Town on Stewart Island
60 Croat or Czech, e.g.
61 Repeat
63 Surgeons William James and Charles Horace ___
64 French director Jacques ___
65 Unpaid
66 Large wading bird
67 Bathroom sign
68 Donkey
69 Retina cell

Down

1 Type of current
2 Blood and guts
3 Gebang palm fiber
4 Condensation starts here
5 Herbaceous plants
6 Pain
7 Prison at sea
8 Flying insect
9 ___ integral
10 Makes metal
11 Both: pref.
12 Talk
16 Exclamation of surprise
23 Minus
25 Primate
27 Architect ___ Saarinen
28 British political scientist Harold ___
29 Type of closet
30 Lyric poem
31 Refrigerant
32 I. e., in full
33 Lake: comb. form
34 978,668 (in base 16)
37 Chemical suffix
40 LED predecessor
41 Like H_2 or HCl
43 Dale
44 Unit of mass: abbr.
46 Temperature unit
48 Force times istance
51 Aug. time
52 Stanislaw ___
53 Top of the head
54 Electron pair acceptor
55 Kind of retaining wall
57 Chemist Lambert ___ (1818–1899)
58 Hebrew letter
59 Facial feature
62 Female lamb

SOLUTION IN THE NEXT ISSUE

SOLUTION TO THE NOVEMBER/DECEMBER PUZZLE

B	A	B	O	■	A	D	A	P	T	■	A	I	D	S
E	X	O	N	■	S	I	N	A	N	■	B	R	I	E
T	O	N	E	■	S	N	O	U	T	■	B	I	O	T
A	N	N	I	E	■	A	P	L	■	T	E	S	L	A
■	■	■	R	M	S	■	L	I	C	E	■	■	■	■
S	E	C	■	F	E	A	■	■	E	N	E	R	G	Y
K	A	R	L	■	X	R	A	Y	S	■	A	A	A	E
O	G	E	E	■	T	E	R	A	T	■	S	I	T	A
A	L	E	S	■	A	S	T	R	O	■	Y	S	E	R
L	E	P	T	O	N	■	■	D	D	E	■	E	S	S
■	■	■	■	I	T	E	M	■	A	M	P	■	■	■
B	A	S	A	L	■	B	A	I	■	B	A	B	F	A
A	B	E	L	■	B	A	S	O	N	■	P	I	O	N
R	A	C	E	■	C	A	E	N	O	■	E	S	A	U
E	C	H	O	■	G	A	R	S	T	■	R	E	M	S

ANSWERS, HINTS & SOLUTIONS

Math

M191

First we will demonstrate commutativity. Take any elements a, c in S. If we let $a = b$ in condition (i), we find that $a * (a * c) = a * (c * a)$. From condition (ii) we can conclude that $a * c = c * a$. Associativity follows from condition (i) and commutativity. Indeed, for any a, b, c we have $a * (b * c) = b * (c * a) = c * (a * b) = (a * b) * c$ (this last from commutativity of the elements c and $a * b$).

We have seen that condition (i), together with commutativity, implies associativity of $*$. Certainly if $*$ is commutative and associative, then condition (i) must hold. However, it is not true that associativity and condition (i) imply commutativity, so that condition (ii) is necessary in this problem. For example, let $S = \{0, 1, 2, 3\}$, and the operation $*$ is defined so that $1 * 2 = 3$, while $a * b = 0$ for any other choices of a and b (and in particular $2 * 1 = 0$). The reader is invited to check that this operation is associative and obeys condition (i), but is certainly not commutative, and also to give an example to show that condition (i) alone does not imply associativity.

M192

We can show that the sum of the reciprocals of the numbers on the blackboard can only get smaller. Since the sum of the reciprocals of the original n numbers is n, the reciprocal of the single remaining number must be less than or equal to n. But this means that this number is at least $1/n$.

To show that the sum of the reciprocals of the numbers cannot increase, suppose the numbers that are erased are a and b. Since a square cannot be negative, $(a - b)^2 \geq 0$, or $a^2 + b^2 \geq 2ab$, or $a^2 + 2ab + b^2 = (a + b)^2 \geq 4ab$. Since both a and b are positive, so are their sum and product, and we can divide this inequality by $ab(a + b)$ to get

$$\frac{a+b}{ab} = \frac{1}{a} + \frac{1}{b} \geq \frac{4}{a+b}.$$

This completes the proof.

M193

Let's call a number whose decimal representation consists only of 1's and 2's an "appropriate" number. We will show that no two appropriate numbers with n digits can have the same remainder when divided by 2^n. Then, since there are 2^n such appropriate numbers, and also 2^n remainders when a number is divided by 2^n (including the remainder 0), one of these appropriate numbers must be divisible by 2^n.

Our proof that distinct appropriate numbers have distinct remainders proceeds by induction. For $n = 1$ we can check this directly: there are four numbers to check, and four remainders. Suppose now that if two $(n - 1)$-digit appropriate numbers have the same remainder when divided by 2^{n-1} that they must in fact be the same. We will show that the same holds for any two appropriate n-digit numbers. Call these number a_n and b_n. If they have the same remainder when divided by 2^n, then they are both even or both odd, so their last (units) digits must be the same. Thus we can write $a_n = 10a_{n-1} + i$, $b_n = 10b_{n-1} + i$, where i is either 1 or 2, and a_{n-1} and b_{n-1} are appropriate $(n - 1)$-digit numbers. Now $a_n - b_n$ is divisible by 2^n, so $10(a_{n-1} - b_{n-1})$ is also divisible by 2^n. It follows that $a_{n-1} - b_{n-1}$ is divisible by 2^{n-1}, or equivalently, that a_{n-1} and b_{n-1} have the same remainder when divided by 2^{n-1}. By the induction hypothesis, then, they must be equal. So all the digits of a_n and b_n are the same, and these two numbers are equal. This completes the proof.

M194

Without loss of generality, we may assume that $\angle KBA < 90°$, and therefore that L lies on AB. Then $\angle MDK = \angle KBA < 90°$, and P lies outside AD. Draw a line parallel to AM through D, and let it meet LP at E (fig. 1). We will prove that $DE = BL$. This will show that $DEBL$ is a parallelogram, and the conclusion follows from the fact that diagonals of a parallelogram bisect each other. First note that triangles PED, PLA are similar (because $DE \parallel AL$), so $DE/PD = AL/PA$. We can write this as $DE = (AL \cdot PD)/PA = (AL/PA) \cdot PD$. Now triangles AKL, AMP are both right-angled and have a common acute angle at A, so they are similar, and $AL/PA = KL/MP$. So we can write $DE = (KL/MP) \cdot PD = KL \cdot (PD/MP)$. Finally, we consider triangles ALB, MPD. They are also both right-angled. Since quadrilateral $ABCD$ is inscribed in a circle, $\angle CBA$ is supplementary to $\angle CDA$, which means that $\angle CBA = \angle CDP$. So triangles ALB, MPD are also similar, and $PD/MP = BL/KL$. Now we can write $DE = KL \cdot (BL/KL) = BL$, and the proof is complete.

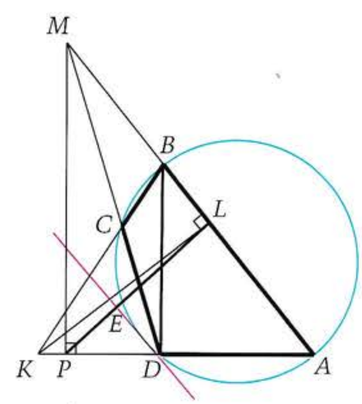


Figure 1

M195

The detective can ask, after any question or group of questions, "Did you lie in answering any of the previous group of questions?" Observe that if the witness answers "no" to this test question, then he must be telling the truth. For if he answers "no" and is lying, then he must have also told a lie also in answering the previous group of questions and thus would have exceeded his quota of lies. So this question is indeed a test of the veracity of the witness for any give set of questions.

With this in mind, the detective can proceed as follows. He follows his original line of questioning for the first 13 questions. For the 14th question, he gives his test question. If the answer is "no," he proceeds with the next 12 questions, asks his test question about these twelve, then the next 11, asking his test question about these eleven, and so on, until he asks his final original question (this makes 13 + 12 + 11 + ... + 1 = 91 original questions).

Note that the witness must give exactly one answer "yes" to a test question. If he lies in an answer to one of the original questions, he cannot also lie on a test question, so he must answer "yes" for the test question of the group that includes his lie, and for no other test question. If he saves his lie for a test question, then, as we have seen, he must answer "no" to all the other test questions, and he is not lying for in answering any of the original questions.

So the detective's plan can be to wait until a test question is answered "yes" and repeat the series of questions to which it refers. The witness must now answer all these questions truthfully (he has used up his lie, whether it was for the test question or one of the original questions). If the second set of answers contains one that differs from the first set of answers, the detective knows that the second set is correct. If the two sets of answers are the same, he knows that the witness lied on the test question. In either case, he now has the information he needed.

After a test question is answered with a "yes," the detective need not ask further test questions, since he knows that their answers would have to be "no."

How many questions has he used? If the kth test question is answered "yes," then there are $k + (14 - k) = 14$ "extra" questions that the detective must ask, so he has gotten the truth after 105 questions.

In the general case, if the detective's original plan involved N questions, this method will give the truth after $N + q$ questions, where q is the smallest natural number such that $N \le q(q-1)/2$. (A. Andjans, I. Solovyov, and V. Slitinsky)

Physics

P191

Let the sleigh have length l, mass m, and initial velocity v_0, and let the coefficient of sliding friction on asphalt be μ. From the moment the sleigh drives onto the asphalt, the equation of its motion (see figure 2) will be

$$mx'' = -\mu g \frac{m}{l} x. \qquad (1)$$

This is the equation for simple harmonic motion, which gives the dependence of distance on time as

$$x(t) = X_1 \sin \omega_0 t,$$

where $\omega_0 = \sqrt{\mu g / l}$ and X_1 is the coordinate of the sleigh at the moment it stops (this corresponds to the amplitude of "oscillation"). At that moment all the initial kinetic energy of the sleigh will have been dissipated in working against the frictional force. The value of X_1 can be obtained by the law of conservation of energy:

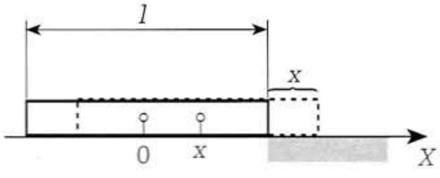


Figure 2

$$\frac{mv_0^2}{2} = \frac{1}{2}\mu g \frac{mX_1}{l} X_1 = \frac{\mu m g X_1^2}{2l}$$

($^1/_2 \mu g m X_1 / l$ is the average force of friction), from which we obtain

$$X_1 = v_0 \sqrt{\frac{l}{\mu g}} = \frac{v_0}{\omega_0}.$$

The period t_1 from the moment the sleigh slides onto the asphalt to the first time it stops is a quarter of the period of oscillation with frequency ω_0—that is,

$$t_1 = \frac{2\pi}{4\omega_0} = \frac{\pi}{2}\sqrt{\frac{l}{\mu g}}.$$

In the second case, when the sleigh is again given a velocity v_0, the equation of its motion will be the same (see equation (1)). The time-dependence of the x-coordinate $(x > X_1)$ will be

$$x(t) = \tilde{X}_1 \sin \omega_0 t,$$

where $\tilde{X}_1$ is the sleigh's coordinate at the moment it stops for the second time (the new amplitude of oscillation).

To find the braking time and the braking distance in the second case, we assume that the sleigh comes onto the asphalt with a velocity u_0, and after sliding a distance X_1, travels at a velocity v_0. So, according to the law of conservation of energy,

$$\frac{mu_0^2}{2} = \frac{1}{2}\mu g \frac{m\tilde{X}_1}{l} \tilde{X}_1 = \frac{\mu m g \tilde{X}_1^2}{2l},$$

from which we get

$$\tilde{X}_1 = u_0 \sqrt{\frac{l}{\mu g}} = \frac{u_0}{\omega_0}.$$

The velocity u_0 can be obtained from the equation

$$\frac{mu_0^2}{2} - \frac{mv_0^2}{2} = \frac{1}{2}\mu g \frac{mX_1^2}{l},$$

which reduces to

$$u_0 = v_0\sqrt{2}.$$

Thus

$$\tilde{X}_1 = \frac{v_0}{\omega_0}\sqrt{2}.$$

The distance the sleigh has slid between the first and second times it stopped is equal to

$$X_2 = \tilde{X}_1 - X_1 = \frac{v_0}{\omega_0}\left(\sqrt{2}-1\right),$$

and the ratio of the braking distances is

$$X_1 : X_2 = 1 : \left(\sqrt{2}-1\right).$$

Given an initial velocity of the sleigh $u_0 = v_0\sqrt{2}$, it travels a distance X_1 in a time $\tilde{t}_1$ defined by the condition $v_0 = v_0\sqrt{2}\,\cos\omega_0\tilde{t}_1$—that is,

$$\tilde{t}_1 = \frac{\pi}{4}\sqrt{\frac{l}{\mu g}}.$$

Before it stops for the last time, the sleigh moves for a period

$$t_1 = \frac{\pi}{2}\sqrt{\frac{l}{\mu g}}$$

(keep in mind that the period of oscillation does not depend on the initial velocity). Thus the time during which the sleigh moves after it is pushed until it stops again is

$$t_2 = t_1 - \tilde{t}_1 = \frac{\pi}{4}\sqrt{\frac{l}{\mu g}},$$

and the ratio of the braking times is

$$t_1 : t_2 = 2 : 1.$$

P192

Since the lengths, cross-sectional areas, and thus the volumes of the bars at 0°C are almost equal, the temperatures we seek obey the inequality $\alpha t \ll 1$. Under this condition the coefficient of volume expansion $\cong 3\alpha$, and the coefficient of surface expansion $\cong 2\alpha$.

The temperature at which the bars are the same length can be found from the equation

$$l_1(1 + \alpha_1 t_1) = l_2(1 + \alpha_2 t_1).$$

Consequently, to find the temperature t_2 at which the bars have the same area, we use the following equation:

$$S_1(1 + 2\alpha_1 t_2) = S_2(1 + 2\alpha_2 t_2),$$

while temperature t_3 at which the bars have identical volumes results from

$$V_1(1 + 3\alpha_1 t_3) = V_2(1 + 3\alpha_2 t_3).$$

Therefore,

$$t_1 = \frac{l_1 - l_2}{\alpha_2 l_2 - \alpha_1 l_1};$$

$$t_2 = \frac{S_1 - S_2}{2(\alpha_2 S_2 - \alpha_1 S_1)};$$

$$t_3 = \frac{V_1 - V_2}{3(\alpha_2 V_2 - \alpha_1 V_1)}.$$

P193

Because the outer plates are connected, the total force acting on the middle plate at the initial position is zero. By the statement of the problem, this plate is shifted very quickly, so the charges carried by the outer plates have no time to change. Therefore, the force remains zero, and no mechanical work is performed.

One thing should be noted: there is an electric field around the capacitor, because the total charge of all the plates is not zero. However, the intensity of the outer electric field does not vary as the middle plate is moved, so the energy of this field can be neglected in our reasoning. Thus the amount of dissipated heat is equal to the difference of the field energy within the capacitor at the final and initial positions of the middle plate.

The field strengths on both sides of this plate can be found from the condition of the absence of a potential difference between the outer plates. This yields the formula for the total energy:

$$U(x) = \frac{Q^2 x(1 - x/d)}{2\varepsilon_0 S},$$

where x is the distance between the middle plate and one of the outer plates. Calculating the difference between the energy values at $x = d/2$ and $x = d/3$, we get

$$\Delta U = \frac{Q^2 d}{72\varepsilon_0 S}.$$

P194

A change in the current flowing in a solenoid generates a vortex electric field that spins the cylinder. The rotating charged cylinder produces a magnetic field inside itself just like the solenoid with current running through it. The corresponding formulas for the magnetic fields are rather simple, but we can manage without them if we recall that a magnetic field is proportional to the value of the current produced—that is, to the product of the charge and the angular velocity of the cylinder.

By the statement of the problem, the cylinder is very light, so the change in the magnetic flux through it generated by the solenoid must be offset by the flux of its own magnetic field. From this we can derive the formulas describing the two different cases: (1) when cylinder is located completely inside the coil and (2) when its radius R is greater than that of the coil r:

$$\omega_1 = \frac{4\pi NIL}{3Q} \text{ for } R < r,$$

$$\omega_2 = \frac{4\pi NIL\left(r^2/R^2\right)}{3Q} \text{ for } R > r.$$

If the charge Q is positive, the cylinder will rotate in the same direction as the current in the coil.

P195

The image S_1 of the source S produced by lens A is located to the right of the lens at twice the focal length (fig. 3). Indeed, the lens formula yields

$$\frac{1}{2F} + \frac{1}{f_1} = \frac{1}{F},$$

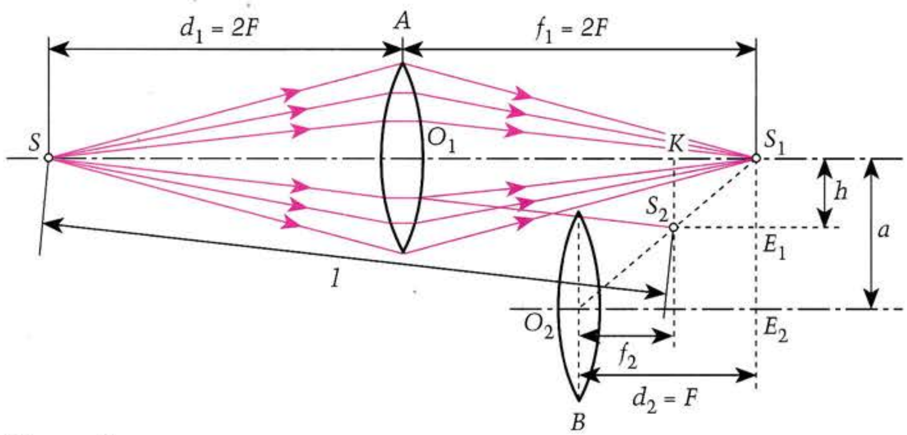


Figure 3

from which we get $f_1 = 2F$. This means that lens B receives a convergent light beam. This beam will be collected at point S_2, which is the final image of the source.

For lens B the point S_1 is a virtual source. So, according to figure 3,

$$-\frac{1}{d_2}+\frac{1}{f_2}=\frac{1}{F},$$

where $d_2 = F$. Therefore,

$$f_2=\frac{F}{2}.$$

The distance $l = SS_2$ between the source S and its image S_2 in this optical system is $l=\sqrt{SK^2+h^2}$, where $SK = 3.5F$. The distance h can be found by the similarity of triangles $O_2S_1E_2$ and $S_2S_1E_1$:

$$\frac{h}{a}=\frac{F-f_2}{F}, \quad h=0.5a.$$

Finally we have

$$l=SS_2=\sqrt{12.25F^2+0.25a^2}.$$

Brainteasers

B191

Let L be the height of the lamppost, h the height of the person, d the distance from the person from the lamppost, and s the length of the person's shadow. By similar triangles, we have

$$\frac{L}{d+s}=\frac{h}{s},$$

or

$$s=\frac{d}{1+L/h}.$$

Calculating the speed for a small movement of the person, we have

$$\frac{\Delta s}{\Delta t}=\frac{1}{1+L/h}\frac{\Delta d}{\Delta t}.$$

Therefore, the taller person's shadow moves faster.

B192

Let a be the side of the square that adjoins the black one on the left (fig. 4). Then, moving clockwise, we find the sides of the other squares touching the black one, one after the other. We find that the side of the red square is $a - (a - 4) = 4$. The side of the blue square can be found in a similar fashion. It's 15.

B193

In order to warm 50 grams of ice from −10°C to 0°C, one needs

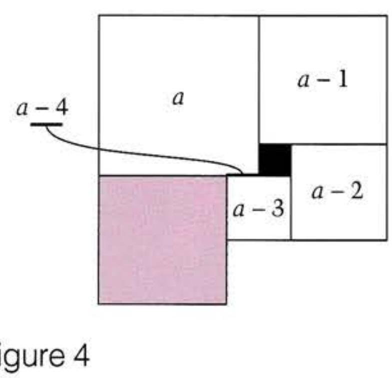


Figure 4

0.43 cal/(g · °C) · 50 g · 10°C = 215 cal, plus an additional 80 cal/g · 50 g = 4,000 cal to melt this ice. Cooling 100 g of water from +10°C to 0°C gives only 1 cal /(g · °C) · 100 g · 10°C = 1,000 cal. So not all of the ice will be melted, and the jar will contain a mixture of ice and water at a temperature of 0°C.

B194

Let x be yesterday's price for big fish and y the new price. Let a and b be the corresponding prices for little fish. The statement of the problem can be written as $3y + b = 5x$, $2y + b = 3x + a$. So we see that $b = 5x - 3y$, $a = 2y + b - 3x = 2y + 5x - 3y - 3x = 2x - y$. We have to compare $y + 2b$ and $5a$. If we express a and b in terms of x and y, we'll find that $y + 2b = y + 10x - 6y = 10x - 5y = 5a$. So one big fish and two little fish cost exactly as much today as five little fish did yesterday.

B195

See figure 5.

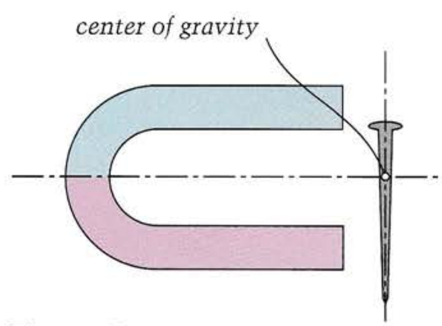


Figure 5

Kaleidoscope

1. The force keeping us from falling through the floor or ground is the electric repulsion between the atoms of the surfaces that come into contact.

2. The force is equal to zero.

3. (a) The force will not change; (b) the force on the paper will be zero, but the charged ball will attract the foil by induction.

4. The ball and paper will not interact.

5. The force of interaction will be larger for the oppositely charged

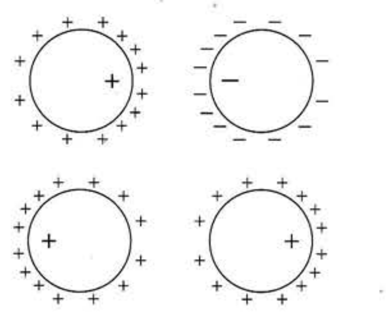


Figure 6

balls, because electrostatic induction results in the redistribution of charges on the balls, after which the like charges are located at greater distances from each other than are the opposite charges (fig. 6).

6. Ball B is charged and its charge is positive, because if the ball were neutral, it would be attracted to A.

7. Yes. For example, when the charge of one ball is far greater than that of the other.

8. The ball will jump away to a distance equal to twice the length of the string.

9. If the ball were oppositely charged, the force of attraction would be replaced by a smaller repulsion. If the balls had the same type of charge, the force of repulsion would increase after the brief electric connection.

10. In the first case, the third charge can't be at equilibrium in any region. In the second case, it can be at unstable equilibrium in region I (midway between q_1 and q_2).

11. The force will increase. Consider the polarization of the glass ball (fig. 7) and estimate the resultant force acting on, say, the charge q_2.

12. The electrostatic interaction of the two electrons slows one of them and accelerates the other. The electrons approach to a minimum distance, then the electron that initially was moving will stop, and the other one (initially at rest) will fly off with a velocity v.

13. The charged ring will be stretched, and the force affecting the charge Q will be equal to zero.

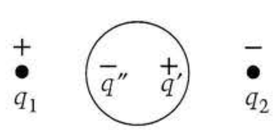


Figure 7

14. The time increases in both cases, because the forces of surface tension are opposed to the force of electric repulsion of like charges on the outside shell of the bubble.

15. The energy of α-particles isn't large enough to overcome the force of electric repulsion of a heavy atomic nucleus and penetrate it.

Microexperiment

The needle will move away from the comb, because the comb attracts both the needle and the water. A hump is formed under the needle, and the needle will start "surfing" downward.

Roman myth

Puzzle. One of the possible holes is shown in figure 8.

1. Suppose the required triangle is ABC, with $BC = a$, $\angle BAC = \alpha$. Then, if we fix points B and C, the possible positions for A lie on an arc of a circle with endpoints B and C. The area of $\triangle ABC$ is then half the product of BC and the perpendicular distance from A to line BC. This is largest when A is the midpoint of the arc, so the largest area is attained when the triangle is isosceles.

2. We can show that the least possible perimeter is attained if the triangle is equilateral. Suppose it is not, and that it has two unequal sides. Let AB be the third side and C the opposite vertex. Then, if C moves along line m *parallel* to AB, the area of $\triangle ABC$ remains unchanged. Let M be the point on line m such that $\triangle AMB$ is isosceles. We will show that the perimeter of AMB is smaller than ABC (if C does not coincide with M). To show this, take point B' to be the reflection of B in line m. It's not hard to see that A, M, and B' are collinear (for example, by showing that $\angle AMB' = 180°$). So $AB' = AM + MB < AC + CB' = AC + CB$ for any other position of point C.

This argument shows that if two sides of our triangle are unequal, the perimeter can be made smaller without changing the area. It follows that the minimal perimeter occurs when the triangle is equilateral.

3. We can show that points K and L must trisect arc AB. Suppose, for example, that K is closer to A than to L. Then $\triangle AKL$ is not isosceles. Let K' be the midpoint of arc AL. Then the

CONTINUED ON COVER 3

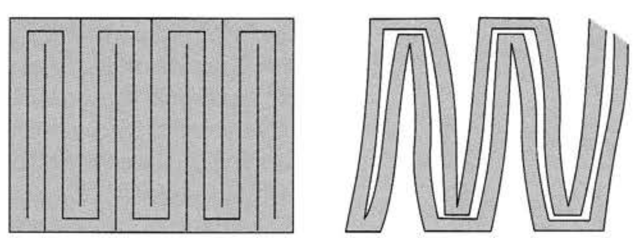


Figure 8

Superprime beef

Grading up the dairy herd

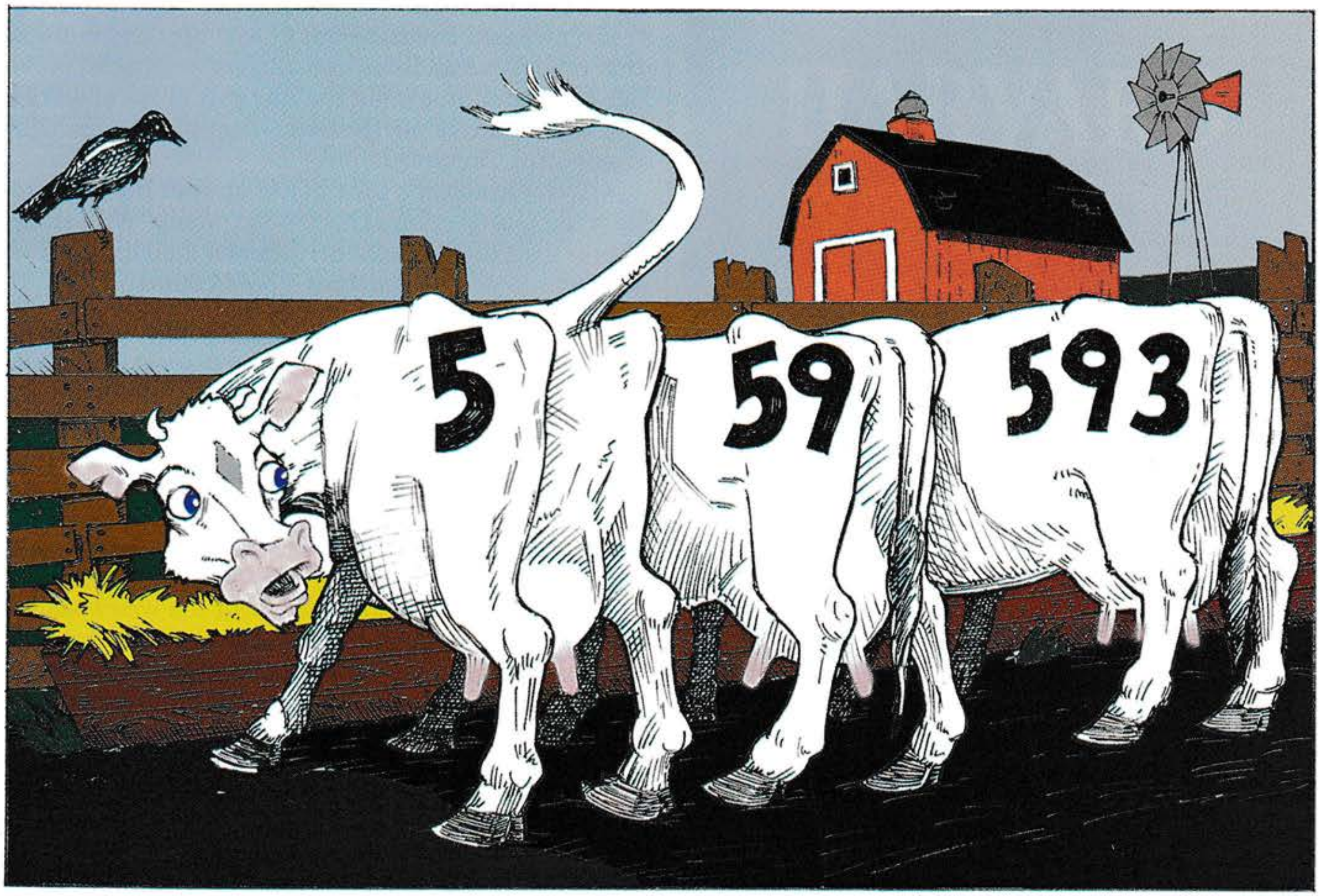
by Dr. Mu

WELCOME BACK TO COWCULATIONS, THE second column in a new series devoted to problems best solved with a computer algorithm. (I say "welcome back," but if this is your first visit, I should tell you that I live in Wisconsin, the Dairy State, and much of my ruminating revolves around the work of my colleagues—and the humans who attend to us.)

Any farmer trying to make a living milking cows has to pay attention to good breeding. Just as race horses are bred for speed, cows can be bred for milk production. By "grading up the herd," as we say in Wisconsin, the average yearly production of milk has climbed to the current level of 12 to 15 thousand pounds per cow. Betsy, a purebred Holstein and barnyard companion, produced a record 44,000 pounds of milk one year: now that's a lot of butter.

Superprimes

Farmer Paul, my boss, and a member of the Primeville Organization for Outstanding Purebreds, takes great pride in raising some of the best purebred Holstein cows in Wisconsin. He calls them affectionately his *superprime* beef,



Art by Mark Brenneman

and to distinguish them from the rest of the herd, he brands them with a very special prime number—a superprime. A superprime is any prime number that remains a prime when any number of digits are deleted from the right side of the number. For example 5939333 is a superprime because 5939333 and 593933 and 59393 and 5939 and 593 and 59 and 5 are all prime.

I tested the sample list with the Mathematica™ function **PrimeQ[p]**, which is True if and only if p is prime. Of course, all the numbers in the sample list are also by definition superprimes.

```
sample={5939333,593933,59393,5939,593,59,5};
PrimeQ[sample]
```

```
{True, True, True, True, True, True, True}
```

Farmer Paul has been in the business for some time now and has seen his herd of superprime beef grow considerably. Predictably, he is running into difficulty coming up with new superprimes and has asked me for some assistance.

This past month, between holiday shopping and all the barnyard parties, I've been ruminating about this problem. Of course, it has been known since the time of Euclid that there are an infinite number of primes. And the famous Prime Number Theorem asserts that the density of prime numbers, as measured by counting the number of prime numbers $\leq n$ (denoted in Mathematica by **PrimePi[n]**), grows like **n/Log[n]**.

For example, there are 78,498 prime numbers between 1 and 1,000,000, as shown in the Mathematica cowculation:

```
PrimePi[10^6]
```

```
78498
```

Let's plot the growth of the number of primes up to and including 500 (see figure 1):

```
Plot[PrimePi[n],{n,1,500},
  AxesLabel->{"N","Number of primes up to N"}]
```

So we are confident there are plenty of prime numbers to brand a herd of any size. But farmer Paul wants to use only superprimes. Will there always be enough numbers? What does your intuition tell you? OK, now go to your cowculator and find some, maybe all, superprimes. If you cowculate in Pascal or C/C++, you'll need to build a fast PrimeQ function, which is best constructed with the well-known sieve algorithm. (Look for it in almost any programming book.)

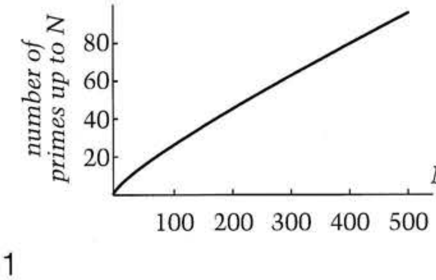


Figure 1

Repusprimes

Superprimes have a built-in right-handed bias. There are left-handed farmers, too, who prefer to knock off their digits from the left side. If they remain prime down to the last digit, they are called repusprimes. (Read repus backwards to understand why.) Let's check to see if 5939333 is a repusprime:

```
sample2={5939333,939333,39333,9333,333,33,3};
PrimeQ[sample2]
```

```
{True, False, False, False, False, False, True}
```

Nope, it fails. But don't despair here's one:

```
sample3={739397,39397,9397,397,97,7};
PrimeQ[sample3]
```

```
{True, True, True, True, True, True}
```

What can you say about the number of repusprimes? Will there always be enough repusprimes for the growing herds of purebred Holsteins belonging to our left-handed friends in the business?

COW 2a. Find an efficient algorithm to generate superprimes.

COW 2b. Find an efficient algorithm to generate repusprimes. Will both right-handed and left-handed farmers always have the option of "grading up the herd"?

Note: COW 2a was used as a problem at the training camp of the 1994 USA Computing Olympiad final round.

You can e-mail your cowculations to me, Dr. Mu, at drmu@cs.uwp.edu. Visit our home page at http://usaco.uwp.edu/cowculations to read the contributions that other cowhands have sent in concerning all previous problems. Stop off at the first part of the address if you want to check out the latest happenings at the USA Computing Olympiad.

OK cowhands, it's time to fire up your iron and produce the special primes I need—and pronto! With all the chores to be done around here, farmer Paul has no time to waste on poky algorithms. Speed counts!

In fields where we lay,
My bovine friends say,
"I can cowculate that!" And they do it.

So I say to you now,
Go after the COW,
Good luck and success—you can do it.

—Dr. Mu

Solution: COW 1

In the last issue, I told you about the weird feeding game farmer Paul played with the herd. The "Challenge Outta Wisconsin" or COW 1a, asked you to find an efficient algorithm that I could use to always win at chow time, assuming I have the advantage of going first. Here is the algorithm sent in by Catalin Drula, a 10th grader at the Bucharest High School of Computer Science in Bucharest, Romania:

My cowculation is simple. Initially the feed pails look like feed = {O, E, O, E, ..., O, E}, where O are the pails with odd indices (1, 3, 5, ..., $p - 1$) and E the ones with even indices (2, 4, 6, ..., p), where p is the number of pails.

The game starts, and Dr. Mu can take an even pail at p or the odd pail at 1. Let's say Dr. Mu takes the odd pail at the first position:

```
Dr. Mu={O}
feed={E,O,E,O,...,E}
```

Once Dr. Mu makes a selection, the opponent must take an even pail, since that is all that is showing at either end. Dr. Mu can force the opponent to stay with the even pails by always taking the odd pails. Thus if the sum of the odd pails is equal to or larger than the sum of the even pails, Dr. Mu will win. But if the situation is reversed and the sum of the even pails is larger than the sum of the odd pails, then switch Dr. Mu's strategy and tell the Doctor to take the even pails instead, which forces the opponent to get only the odd pails. Since Dr. Mu has the option of going first, have the Doctor make the initial calculation to compare the two sums and stay with the larger sum, either the evens or odds. That will guarantee a win for Dr. Mu.

Note by Dr. Mu: Catalin also submitted a dynamic programming cowculation that maximizes the amount that I can eat depending on the choices of the second cow. It appears on the web page at http://usaco.uwp.edu/cowculations. One can imagine how, using the same strategy explained above, I could maximize my total chow. Instead of making one computation of the sum of the evens and the sum of the odds at the beginning of the game, I should redo the computation at each stage of the game when it's my turn and switch from evens to odds, or vice-versa, if it's to my advantage to do so. That happens when the sum of the remaining evens is larger than the odds, or vise versa. The best the second cow can do is to play the Greedy algorithm, since I am in complete control.

A cowculation was also submitted by Tony Capra from White Station High School in Memphis, Tennessee. The complete program listings for all cowculations submitted can be found on the cowculations Web page.

In COW 1b I asked for a cowculation of my chances of winning over Betsy if we both follow the Greedy algorithm. What I expected was a simple simulation where I play Betsy a few times—say 1,000—and the proportion of wins for me are recorded. I carried out these cowculations for the number of pails from 4 to 100. As you might have expected, the chances of winning go down as the number of pails increases. Figure 2 is a graph of the results, done in Mathematica.

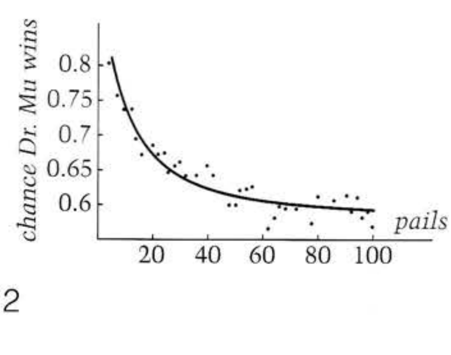


Figure 2

"ANSWERS, HINTS & SOLUTIONS"
CONTINUED FROM PAGE 54

perpendicular distance from K' to AL is greater than the perpendicular distance from K to AL, so the area of $\triangle K'AL$ is greater than that of $\triangle KAL$. Thus if K is not the midpoint of arc AL, the area of the quadrilateral can be increased. Similarly, L must be the midpoint of arc KB, so that K and L trisect arc AB.

4. For any given value of l, the figure with the largest area is the circle. For a circle, $l^2 = 4\pi^2r^2 = 4\pi(\pi r^2) = 4\pi\Delta > 12.5\Delta$. For any other shape, then, l^2 must be greater than this quantity.

5. Consider an inscribed polygon each of whose sides is equal to each segment in sequence (as we know from problem 4 in the article, it has the greatest area). Change the order of two of its adjoining sides. The polygon is still inscribed, its area is unchanged, but the sequence of the sides has altered. By changing the order of one pair after another, we'll obtain inscribed polygons with all possible sequences of sides. They all have the same area.

6. Take an arbitrary polygon of the given type and attach to it its image after a reflection with respect to the varying side. We obtain a polygon with fixed sides and twice the area of the original. Now we can apply the result of problem 4.