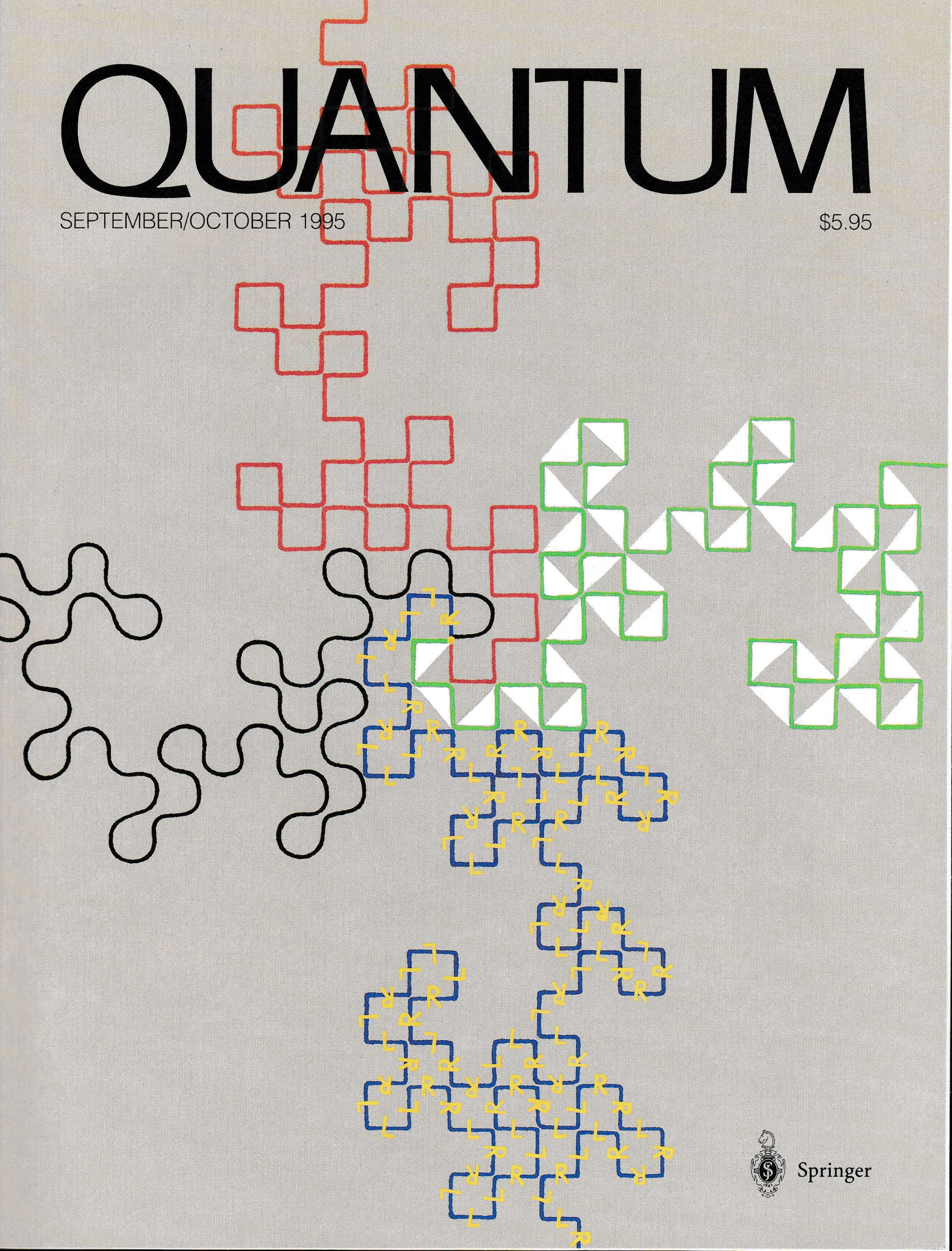
QUANTUM

SEPTEMBER/OCTOBER 1995

$5.95



Springer

GALLERY

Fanny/Fingerpainting (1985) by Chuck Close

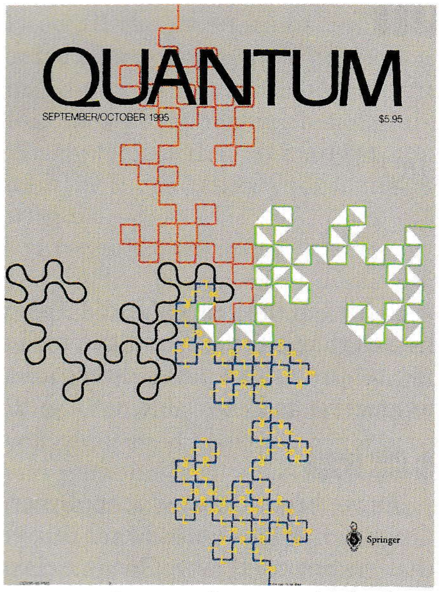
WHEN THIS WORK IS CONFINED TO THE PRINTED page, the illusion that it is a photograph is intensified. If you saw it in the gallery, however, it would be hard to mistake this large canvas—over two and a half meters tall—for a snapshot. And yet, even as it looms in the distance, it looks like a photo. Why? For one thing, the painting mimics the graininess of some black-and-white photographs. This can be attributed to Close's technique of applying paint with his fingertips and the remarkable subtlety of his touch. Also, parts of his subject are "out of focus"—something you rarely if ever see in a painting or drawing.

Clearly the artist used a photo of his subject as he painted, and he probably meant to create a photographic illusion. But he also chose to render this image with his bare hands, not entrusting it to an industrial process steeped in chemicals and burdened by hardware. Some see the portrait of Fanny, the artist's mother-in-law, as aggressively unflattering. Others think just the opposite.

You'll find another kind of digital image in the Happenings department. And Mark Biermann explores "depth of field"—one factor affecting the clarity of photographs—in the article beginning on page 26.

QUANTUM

SEPTEMBER/OCTOBER 1995 VOLUME 6, NUMBER 1



Cover art by Vera Khlebnikova

There's more than one way to skin a cat, as they say. The same goes for drawing "dragon designs," as our artist has demonstrated for us. Seeming to emerge from a common origin (though actually it's a matter of dispute "who begot whom"), the pure, abstract, "traditional" version heads "north," plowing right through our logo. A dandified rendering, pinwheels spinning and bow ties akimbo, saunters off to the right. A rather plodding dragon design walks off the lower edge, mumbling to itself "left, left, right, left . . .," as if trying to remember the way home. And exiting to the left is the elegant, self-assured *dragon curve*.

While many of us have squandered an afternoon finding shapes in clouds, mathematicians seem to see them just about everywhere. For an introduction to these curious designs and the mathematics behind them, turn to page 4.

Indexed in *Magazine Article Summaries, Academic Abstracts, Academic Search, Vocational Search, MasterFILE,* and *General Science Source*

FEATURES

4 *Kvant*'s 25th Anniversary
Dragon curves
by Nikolay Vasilyev and Victor Gutenmacher

12 Cosmic Charge
On the nature of space magnetism
by Alexander Ruzmaykin

20 Fermi Problems
Educated guesses
by John Adam

26 Fuzzy Physics
Clarity, reality, and the art of photography
by Mark L. Biermann

DEPARTMENTS

2 Publisher's Page

19 Brainteasers

32 Kaleidoscope
All bent out of shape

35 How Do You Figure?

36 Physics Contest
Split image

40 Math Investigations
Problems beget problems

42 In the Lab
A tell-tale trail and a chemical clock

45 At the Blackboard
The quadratic

49 Follow-up
Life on an accelerating skateboard

53 Happenings
Caught in the Web

55 Crisscross Science

56 Answers, Hints & Solutions

61 Toy Store
Winning strategies

The torch is passed

But I intend to stay close to its light!

SEVERAL TIMES OVER THE past year, an announcement has appeared in the pages of *Quantum* magazine, as well as the other journals published by the National Science Teachers Association (NSTA). Perhaps you noticed it. The announcement solicited applications for the position of executive director of NSTA—the position I have held for the past 15 years. I have been happy to serve science education in this capacity—the job is a wonderful combination of teaching (or "persuading"!), learning, traveling, and writing—all things I enjoy doing. But the time has come for me to cut back on my professional activities, and the announcements were part of an extensive search for a new executive director.

I'm pleased to announce that NSTA's board of directors has found a new executive director. He is Dr. Gerald F. Wheeler, a professor of physics at Montana State University, and he will have begun serving in his new capacity by the time this issue is printed. I have known Gerry for many years, and I am confident that he will be an energetic proponent of the association and of science education in the years ahead. In addition to being a close personal friend, Gerry was a site director in Montana for one of the projects I have directed—Scope, Sequence, and Coordination of Secondary School Science (SS&C).

Gerry's path in science has been interesting and somewhat unorthodox. In college he majored in science education and acquired a broad background in science. It wasn't until graduate school that he narrowed his focus, and I admire the fact that he earned a Ph.D. in experimental nuclear physics without majoring in physics as an undergraduate. He has taught science at every level, and he is the coauthor (with Larry Kirkpatrick, *Quantum*'s field editor for physics) of a college-level textbook (*Physics: A World View*).

In addition to his experience in the classroom, Gerry has served as president of the American Association of Physics Teachers (AAPT). This has undoubtedly helped prepare him for the administrative tasks he will face as executive director of the largest association of science teachers in the world.

One role that Gerry will *not* be taking on as the executive director of NSTA is that of publisher of *Quantum*. I will continue to serve in that capacity at least through December 1996. I will also continue to direct the SS&C project and serve as chair of the Global Summit on Science and Science Education, scheduled to take place in San Francisco in December 1996.

So I invite *Quantum* readers to join me in welcoming Gerry Wheeler as the new executive director of NSTA. I look forward to working with him in the years ahead. I'd also like to thank those who have made my tenure as executive director so productive and enjoyable.

In this issue . . .

As we begin the new school year, *Quantum* brings you several things that are not at all new. One is a classic, actually—the first feature article, "Dragon Curves." It first appeared in 1970, in Russian, in the second issue of *Kvant* (which we persist in calling our sister magazine, when it is in fact *Quantum*'s mother!). This year *Kvant* celebrated its 25th birthday, and we wish it 25 years more.

Another item that is certainly not novel but may prove useful as you start the school year is the article on quadratic equations, which begins on page 45. It's meant to refresh your recollection and solidify your understanding of this important topic.

I would direct the attention of teachers to the ad for the Duracell/NSTA Scholarship Competition on page 18 and to the call for manuscripts on page 39. (Of course, we hope all our readers patronize *all* the advertisers in *Quantum*!)

An award from Folio magazine

We all know that *Quantum* is a great magazine, but it's nice when "outsiders" recognize it. Recently we learned that *Quantum* was awarded a 1995 *Folio* Editorial Excellence Award (see the announcement on page 34). Congratulations to the writers, editors, translators, and artists in Moscow, Washington, and around the world who collaborate to produce this unique publication.

Survey results

I'd like to thank all those who responded to the reader survey in the May/June issue. We have tabulated all the evaluations and pondered every comment. While it was in no way a "scientific survey," it told us a lot about our readership. We appreciate that so many of you took the time to share your thoughts with us.

Even when we're not "surveying" you, we're interested in what you think. Whether you send your messages by letter, fax, or e-mail, or inscribe the guest book at our World Wide Web site, we read everything you send, even if we do not reply to you personally. Please be assured that your opinions matter, and that where possible we will try to implement any changes you suggest.

—Bill G. Aldridge

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
National Science Teachers Assoc.
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 6 (6 issues) will be published in 1995–1996. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by Quantum Bureau of the Russian Academy of Sciences, as well as original material in English. **Editorial offices:** NSTA, 1840 Wilson Boulevard, Arlington VA 22201-3000, telephone (703) 243-7100. **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 (0) 30-82 07-1, telex 185 411.

Second class 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: \$15; Personal rate (nonstudent): \$20; Institutional rate: \$39; Single Issue Price: \$5.95. 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: Subscription rates in U.S. currency as above (all rates calculated in DM at the exchange rate current at the time of purchase) plus postage and handling. SAL (Surface Airmail Listed) is mandatory for Japan, India, Australia, and New Zealand. Customers should ask for the appropriate price list. Air mail delivery to all other countries is available upon request. Orders may be placed through your bookseller or directly through Springer-Verlag, Postfach 31 13 40, D-10643 Berlin, Germany.

Dragon curves

Fear not! They're as tame as they are beautiful

by Nikolay Vasilyev and Victor Gutenmacher

EDITOR'S NOTE: THE PUBLIcation of this article marks the 25th anniversary of *Kvant*, the Russian-language sister magazine of *Quantum*, which was celebrated this year. We searched through a number of the early issues of *Kvant* and picked out this article (which originally appeared in the second issue of *Kvant* in February 1970) for several reasons. First, although dragon curves have been repeatedly described and discussed in the literature in the intervening years, they have lost none of their beauty and have acquired even greater mathematical significance with the development of fractal theory and other fields of mathematics. Also, this article is the very first example of the international collaboration in which *Kvant* has been involved since it was born and which eventually brought into existence the magazine you hold in your hands—it was based on the manuscript of the article "Number Representations and Dragon Curves" by the Canadian mathematicians Chandler Davis and Donald Knuth, which was published in *Journal of Recreational Mathematics* (vol. 3) in the same year. (See also Martin Gardner's articles in the March, April, and July 1967 issues of *Scientific American* and in *Mathematical Magic Show*, pp. 207–209, 215–220.)

But perhaps the most compelling reason for our choice was the manner in which the authors reworked the original material, a manner that seems to us the epitome of *Kvant*'s style and marks the difference between this presentation and many others. The authors drop some interesting proofs that are too sophisticated and complex for a popular magazine, but they still provide plenty of food for thought: ideas of proofs to be completed, hints to elaborate on, a lot of attractive problems (we can add one more: try to reproduce the dragon patterns using your computer, as we did in preparing the illustrations).

So, rediscover and enjoy the bewitching beauty of mathematical dragons!

What is a dragon curve?

Take a long paper strip, fold it in half and then in half again. Stand the folded strip on edge and open it to a right angle at each crease (fig. 1). Looking at it from the top, you'll see something like what's shown in figures 3a and 3b (which correspond to the foldings in figures 2a and 2b).

After three foldings we can get essentially different patterns (fig. 3c, 3d) depending on how we folded the

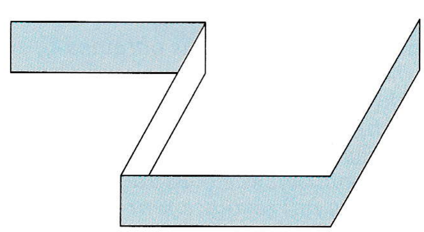


Figure 1

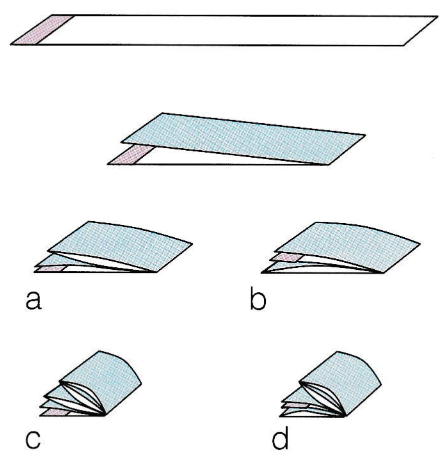


Figure 2

strip. If the strip is folded four times or more and opened to right angles, we can obtain many different

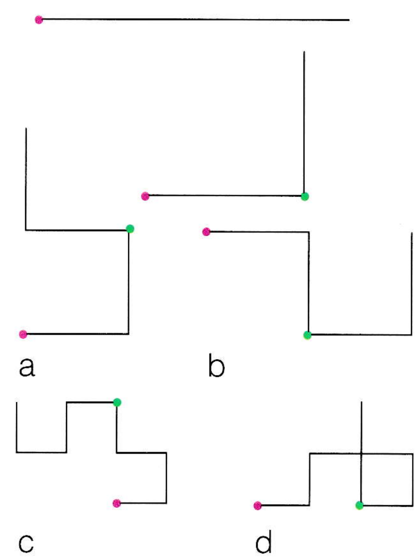


Figure 3

Art by Vera Khlebnikova

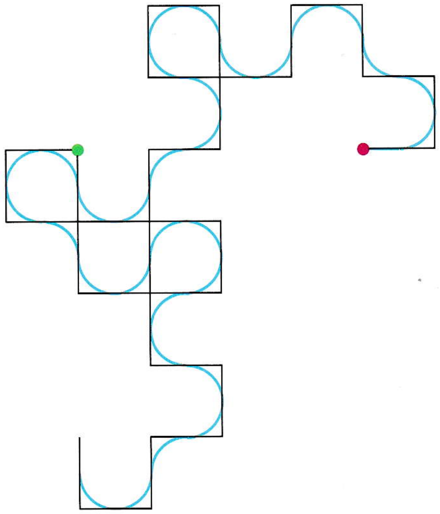


Figure 4

patterns. Figure 4 shows one of the patterns produced by folding in half five times.

It's practically impossible to fold the strip more than seven times—the eighth fold would already yield $2^8 = 256$ layers! But we'll soon learn how to draw patterns corresponding to multiple folds of the strip without the strip. In figure 5 you see one of the patterns that could emerge if we creased the strip 12 times. It consists of $2^{12} = 4{,}096$ segments.

If the strip is creased more than three times, then after unfolding it some of its corners will necessary touch one another (fig. 3d and fig. 4). Because of numerous contacts of this sort, large patterns have areas that look like a grid rather than one long rectangularily bending path. To make the path visible, we can round out its corners (as shown by the blue curve in figure 4). The illustration on page 4 contains curves generated in this way.

It was just such a pattern that suggested to John E. Heighway, an American physicist, the name "dragon curves." Anyone who has seen a dragon will readily confirm that this is exactly how that creature looks. The straight-segment patterns generating dragon curves are called "dragon designs."

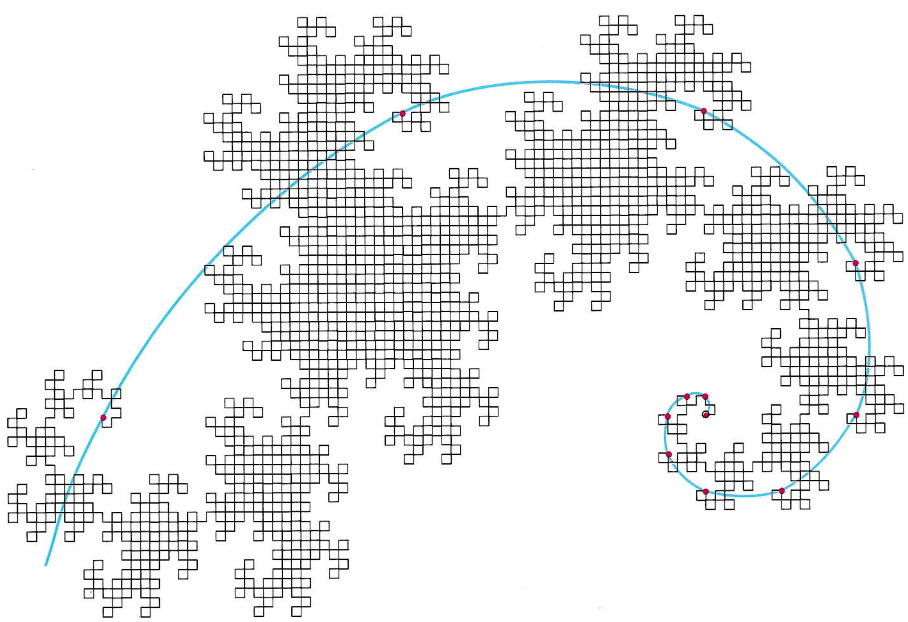


Figure 5

A polygonal path like this one (called the main dragon design*) is obtained if, starting with a segment, we always rotate the piece of the design already constructed in the same direction (here clockwise). This corresponds to the method of bending the strip in half shown in figures 2a and 2c: always "from right to left and upward." Figure 4 illustrates the beginning of the design (32 segments); here we have 4,096 segments. If we continue the construction, the path will slowly curve around its origin, making a full revolution in 8 "duplications." The red points lie on a logarithmic spiral (see the Kaleidoscope in the March/April 1995 issue of* Quantum*).*

Drawing long dragon designs

The dragon design obtained by folding a paper strip n times is said to be of the order n. Let's investigate the structure of dragon designs and learn how to draw them for large enough n.

The first method. A dragon design of order n consists of 2^n segments, and so it has $2^n - 1$ bends. Since this number is odd, the midpoint of a dragon design falls at one of its vertices (for $n > 0$). In figures 3 and 4 the midpoints are marked with green circles. You'll notice that they divide the corresponding patterns into two congruent halves obtained from each other by a 90° rotation. And this turns out to be a general rule.

THEOREM 1. *A dragon design of order n with the endpoint O extended by the same design rotated by 90° about O becomes a dragon design of order $n + 1$. Conversely, any dragon design of order $n + 1$ can be obtained in this way from a design of order n.*

This becomes clear when we analyze our construction. Suppose we want to fold the strip $n + 1$ times. Crease it once. Its two halves coincide and will be folded in exactly the same way further (fig. 6). Now open the last

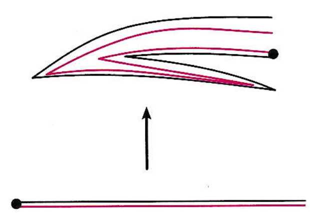


Figure 6

n bends to 90°. This yields two identical dragon designs of order n stuck together (fig. 7). It remains to pull them 90° apart—we get a dragon design of order $n + 1$. These considerations can be transformed into a more rigorous proof of both statements of the theorem, which is left to the reader.

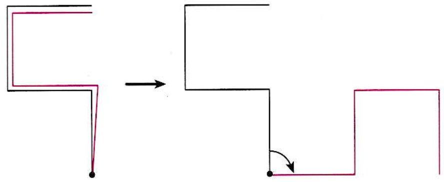


Figure 7

Now let's see how dragon designs can be drawn using this theorem. Since they run along the lines of a square grid, it's convenient to use graph paper.

Take any short dragon design—for instance, just one segment. One of its endpoints will be assumed to be its beginning, the other its end. Extend it by the same pattern rotated 90° about the end (the direction of rotation is chosen at will). Then the new design can be extended in the same way beyond its end. This process continues as long as desired and possible. The construction can be performed automatically and quickly if you have some tracing paper or, better still, slightly transparent graph paper.[1]

Of course, dragon *curves* can be drawn in quite the same way, only we must always round out the middle bend.

All dragon curves have one remarkable property: *they never intersect themselves*; or, equivalently, *dragon designs never traverse the same segment twice.* Thus, even though a dragon design may pass twice through the same point (a grid node), it never visits the same point *more* than twice. It's not clear from theorem 1 how to prove this property—on the contrary, the longer and more intricate designs or curves you draw, the more surprising it is how their bumps and pits fit each other. However, a proof of this property becomes less difficult if you use another theorem on the "duplication" of dragon designs—which, by the way, gives another method for drawing "long" designs.

The second method. In figure 8 alternate vertices of black paths are joined with red segments. You see that the red segments constitute dragon designs again. This turns to be a general law, too.

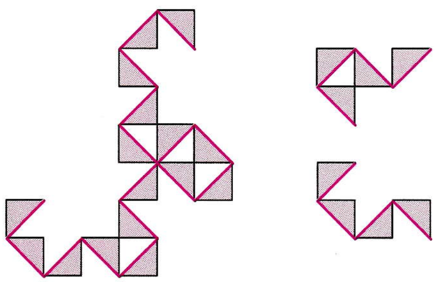


Figure 8

To give this an exact wording, notice that each red segment is the hypotenuse of an isosceles right triangle whose legs are segments of the initial design (these triangles are shaded in the figure). Each of these triangles is obtained from the neighboring triangles under a 90° rotation about their common vertex. In other words, moving along the red path, we'll meet these triangles alternately on the right and left sides of the path.

THEOREM 2. *Construct on each segment of a dragon design of order n as its hypotenuse an isosceles right triangle such that any two neighboring triangles are obtained from each other under a 90° rotation about their common vertex. Then the legs of all these triangles make a dragon design of order n + 1. Conversely, any dragon design of order n + 1 can be obtained in this way from an nth-order design.*

Indeed, let's examine the *last* folding of our strip. Look at figure 9. We want to fold the strip $n + 1$ times. Let's first make n creases and view the strip from the edge (the red line in figure 9). Then crease it once more and open the last fold to 90° (the black line in the figure[2]). Now the strip goes from bends A to bends B along the legs of the isosceles right triangle ABC rather than straight along its hypotenuse. Open the bends A and B to 90° as shown in figure 10 for a single bend. Then the legs of our right triangles will form a dragon design of order $n + 1$ and their hypotenuses a design of the nth order.

These considerations are readily turned into an accurate proof of both statements. You can try another approach and derive theorem 2 from theorem 1.

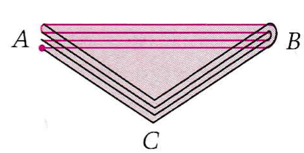


Figure 9

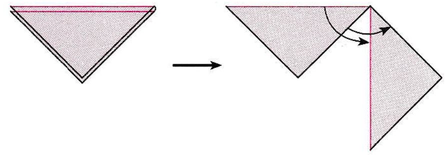


Figure 10

Using theorem 2, two dragon designs of order $n + 1$ can be obtained from each nth-order design, because the triangle on the first segment can be constructed on either side.

Notice that when we pass to a dragon design of the next order by theorem 1, the entire path becomes twice as long, while the length of each segment remains the same. When we apply theorem 2 for "duplication," the length of the design increases by a factor of $\sqrt{2}$, while each segment becomes shorter by a factor of $1/\sqrt{2}$.

Now practice drawing dragon designs using theorem 2.

Words

The property described in theorem 2 can be explained very simply if we look at dragon designs (or methods of paper folding, if you wish) from a somewhat different standpoint.

Imagine a turtle crawling along a dragon design (fig. 11). Every time it reaches a bend, it has to turn 90° to the left or right. From the turtle's point of view, its path is determined by the sequence of turns. For example, for the design in figure 3a (with the start at the red point) the sequence is left, left, right. Denoting the turns by their initials, we get the "word" LLR that defines this particular sequence. (A "word" in many branches of mathematics and logic

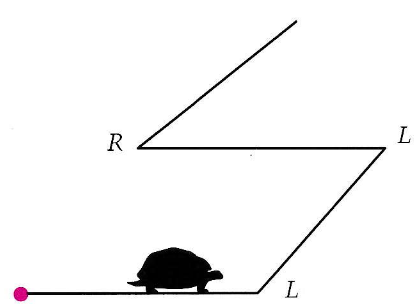


Figure 11

[1]A more state-of-the-art method would be to use a computer with graphing tools.—*Ed.*

[2]Our black line is $\sqrt{2}$ times longer than the red line, but this doesn't matter because we're interested in the line's shape rather than its size.

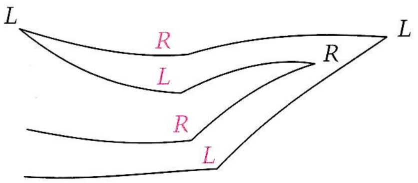


Figure 12

is an arbitrary sequence of letters.) So we can write out any dragon design symbolically using a sequence of the letters L and R. Note that a design of order n (that is, one in which the paper is folded n times) corresponds to a word with 2^{n-1} letters.

To obtain figure 3c from figure 3a, the strip must be folded once again from right to left (fig. 3a, 3c and fig. 12). But then a new fold appears on each segment of the strip. Not only that, figure 12 clearly shows that new bends have alternate directions. Thus we get

$$\begin{array}{l} \;\; L \;\; L \;\; R \\ \qquad\qquad\qquad \to LLRLLRR \\ L \;\; R \;\; L \;\; R \end{array}$$

that is, the word that codes figure 3c. One more fold to the left would give the design coded by the word

$$\begin{array}{l} \;\; L\,L\,R\,L\,L\,R\,R \\ \qquad\qquad\qquad \to LLRLLRRLLLRRLRR. \\ L\,R\,L\,R\,L\,R\,L\,R \end{array}$$

Draw this design. It's called the *main dragon design* of order 4. If you want, round out its corners as Heighway suggested.

If you repeat this procedure with the last word (starting the alternating sequence of letters with L), you'll get a 31-letter word—the code of the *main dragon design* of order 5 shown in figure 4.

Of course, we can start our alternating sequences of letters with R rather than L—this would simply yield other designs.

It's not hard to see that our method for producing a "dragon word" of order $n + 1$ from a word of order n corresponds exactly to the second duplication method described in theorem 2 (inserted letters correspond to added triangles). In general, the entire "theory of dragon designs" could have been developed algebraically rather than geometrically, by using operations on words composed of the two letters L and R—the codes of dragon designs—instead of rotation, triangle construction, and so on.[3]

You can continue on this path and get to know a number of interesting properties of "dragon words" and dragon designs by solving the following problems.

Exercises

Some of these exercises are simple, while others involve significant investigations. We supply full solutions to some, only hints to others.

1. Suppose that after 30 folds the distance between neighboring creases on our strip has become 1 cm. How long was the original strip? Was it longer or shorter than the distance from the Earth to the Moon?

2. How will the dragon design change if its generating paper strip is stood on the other edge? How will the corresponding word change?

3. Give an example of a "word" in L and R that does *not* define a dragon design.

4. Suppose a turtle has crawled along a dragon design and read a word consisting of the letters L and R. What word will it read if it crawls along the same design in the opposite direction? If the word of the original curve is w, we write $\overline{w}$ to denote this new word. For example, if $w = LLR$, then $\overline{w} = LRR$.

5. If u and w are dragon words, denote by wu the word obtained by writing u right after w. (For example, if $w = LLR$, $u = RL$, then $wu = LLRRL$.) Using the definition of $\overline{w}$ given in problem 4, prove that (a) for any w and u, $\overline{wu} = \overline{u}\,\overline{w}$. (b) If w_{n+1} is the word corresponding to a dragon design of order $n+1$, then $w_{n+1} = w_n x \overline{w_n}$, where w_n is the code word of a certain nth-order dragon design and x is a one-letter word. (c) If w is a code word for a dragon design, then $\overline{w}$ differs from w only in the middle letter. (d) The code word of a dragon design of order n is uniquely determined by the sequence of n letters in the 1st, 2nd, 4th, ..., 2^{n-1}th places of this word.

6. Theorems 1 and 2 each allow us to obtain two dragon designs of order $n + 1$ from a design of order n: with theorem 1 we can choose either endpoint for the center of 90° rotation, with theorem 2 we can choose either side of the first segment to construct the triangle. Are the two designs of order $n + 1$ in the first case the same as in the second or are they, generally speaking, different?

7. Let our turtle crawl from point A with a constant speed starting along the direction AB and making a 90° turn every 15 minutes. Prove that it can come back to A only in an integer number of hours and only perpendicular to AB.

The next few exercises are really full-length investigations and are left to the reader as open questions.

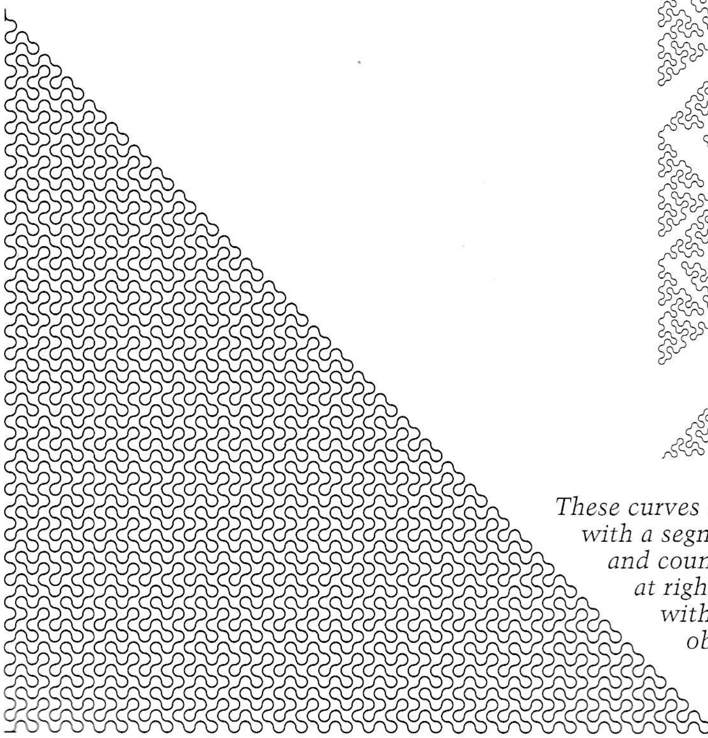
8. Prove that a dragon design traces no segment more than once.

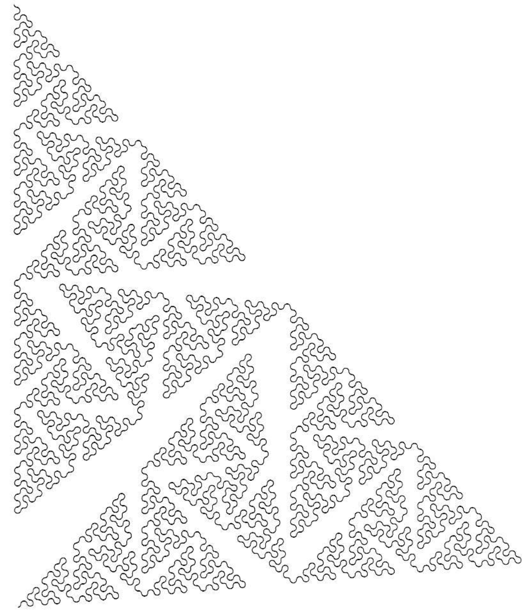
9. Consider a dragon design and rotate it about its endpoint O by 90°, 180°, and 270°. Prove that no two of the four designs thus obtained have a common segment.

10. Consider the set of all polygonal paths of 2^n equal segments that satisfy the following property: if we split any of them into 2^k pieces consisting of 2^{n-k} segments each, then any two adjacent pieces can be obtained from each other by a 90° rotation about their common point for any k, $2 \le k \le n$. Prove that this set coincides with the set of dragon designs of order n.

11. Let's draw the main dragon design on the coordinate plane so that its first three vertices are (0, 0), (0, 1), (1, 1). Using theorem 1, we can extend this design as many times as we wish, getting successively the designs of order 1, 2, 3, ... Imagine we've drawn all of them—that is, an *infinite dragon design* (the main dragon design of order ∞). Its first 2^n segments form the nth-order design. Prove that (a) The endpoint of the nth-order design has the coordinates $(2^{n/2} \cos (n\pi/4), 2^{n/2} \sin (n\pi/4))$ (the red points in figure 5). (b) The code word of this design can quickly be written by the following rule. First

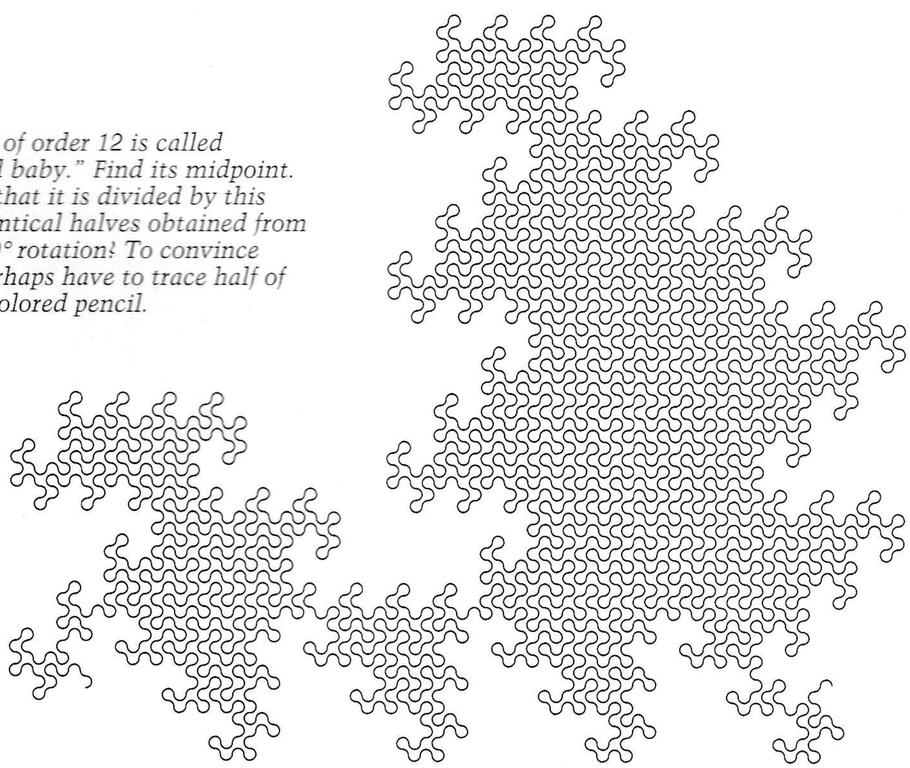
[3]The algebraic approach was used by Davis and Knuth (see the editor's note at the beginning of the article).





These curves are obtained after 12 "duplications" starting with a segment; the rotations are alternately clockwise and counterclockwise. The larger curve always bends at right angles; it fills up an isosceles right triangle with a uniform pattern. The other curve is obtained in exactly the same way except that right angles of rotation are replaced here with angles of 95°. This makes the structure of the curve dramatically visible.

This dragon curve of order 12 is called "Mama, papa, and baby." Find its midpoint. Can you imagine that it is divided by this point into two identical halves obtained from each other by a 90° rotation? To convince yourself, you'll perhaps have to trace half of the curve with a colored pencil.



we write out alternate letters L and R leaving blank spaces between them: $L\_R\_L\_R\_\ldots$. Point with a finger of your left hand at the first letter and enter this letter in the first space with your right hand. Then point at the second letter with your left hand and enter it in the second space with your right hand, point with the left hand to the third letter, and so on, through all the letters without skipping those written at the earlier steps:

$$L\underline{L}R\underline{L}L\underline{R}R\underline{L}L\underline{L}R\underline{R}L\underline{R}R\underline{L}L\underline{L}R\underline{L}\ldots$$

(c) If four main dragon designs are generated from the same point as in problem 10, then each segment of the integer square grid will be traced by exactly one of them (fig. 13). This is a difficult theorem first proved by Donald Knuth. The proof is based on a clever representation of complex numbers $n + mi$ with integer n and m in a special "number system" whose base consists of the numbers $1 + i$, $1 - i$, $-1 + i$, and $-1 - i$. This number system is similar in a certain sense to the "balanced ternary system" described in "Number Systems" (exercise 11) in the last issue of *Quantum*. ◘

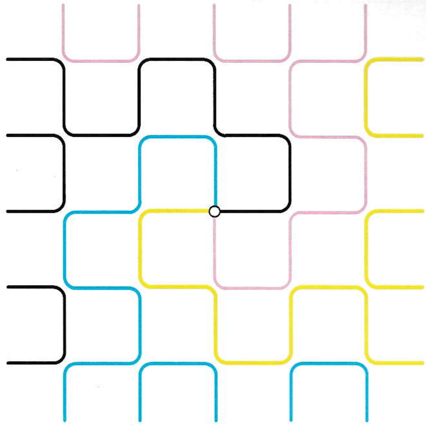
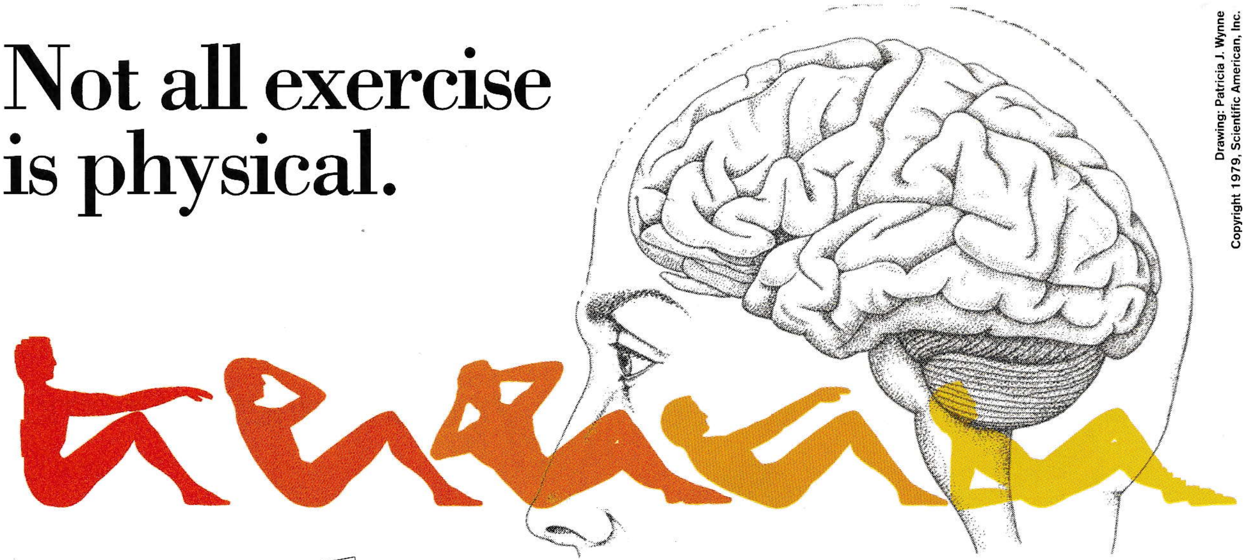


Figure 13

ANSWERS, HINTS & SOLUTIONS ON PAGE 60

Not all exercise is physical.



Drawing: Patricia J. Wynne
Copyright 1979, Scientific American, Inc.

In fractions of a second, your mind can have you slamming a ball off two walls, or sliding into second base.

It decides, directs and transmits the necessary brain waves needed to enjoy an hour of hard, body-toning physical exercise. It's what intelligent people do to keep themselves fit and healthy.

It's a well known fact that we use only a small portion of our minds' capabilities and like our bodies, our minds thrive on good healthy use. SCIENTIFIC AMERICAN provides an ideal workout for the mind, stretching the imagination and toning the intellect.

SCIENTIFIC AMERICAN probes the opposing mysteries of outer and inner space. Rocket technology, genetic manipulation, the socio-economic effect of earthquakes on populations or the origins of Indo-European languages–just a small sampling of the kinds of articles written by scientists and edited for your active mind.

Each month, SCIENTIFIC AMERICAN brings the individual talents of scientists, editor and artist together to keep you up-to-date on the new developments and important technologies that affect our everyday lives.

Each article is an opportunity to share in the act of science and observe first hand the arrival at discovery.

Join us at the frontiers of knowledge. Subscribe to

SCIENTIFIC AMERICAN

and exercise your mind.

12 issues per year $24.97

Subscriptions outside the U.S. are $47.00

(new subscribers only)

Save over $22.00 off the cover price.

Simply complete and mail the attached coupon to enter your subscription today!

SUBSCRIPTION ORDER FORM

Yes, I'll take a mental work-out. Enter my subscription for:

❑ **12 issues, one year $24.97**

Name ______________________________

Address ______________________ Apt ________

City ______________________________

State ______________________ Zip ________

❑ Payment enclosed ❑ Bill me

Please allow 4-6 weeks for delivery of first issue.
Rates outside the U.S.; 1 year $47 in U.S. funds drawn on a U.S. bank.
GST included. GST# R127387652. Quebec residents add QST. QST# 1015332537.

SEND THIS COUPON TO:
SCIENTIFIC AMERICAN · P.O. Box 3187 · Harlan, IA 51593 - 0378 **P5QM329**

On the nature of space magnetism

Untangling the turbulent source of the "hydromagnetic dynamo"

by Alexander Ruzmaykin

THE EARTH'S MAGNETIC field was discovered long ago. The magnetic fields of other planets, stars, and galaxies were discovered in our century, some of them quite recently. We know now that magnetic fields 10^{12} to 10^{14} times stronger than the Earth's are formed in pulsars. Cosmic magnetic fields are fundamentally different from those of permanent magnets made of steel or alloys. Like the fields in electromagnets, they are produced by currents flowing in an electrically conducting medium. However, cosmic electromagnets are not, as a rule, fed by an external emf, but operate as a self-excited dynamo (electromagnetic generator). The moving medium serves as the "armature" in this dynamo. This article is a brief introduction to cosmic dynamos as a source of large-scale magnetic fields in outer space.

Magnetic fields in nature

Your textbook tells you that magnetic fields are created by electric currents. Historically, though, people familiarized themselves with magnetic fields using permanent magnets. Recall how clearly the iron filings "drew" the magnetic lines of force. These lines are always closed. A simple observation shows that the needle of a compass orients itself tangent to the magnetic lines of force. In fact, this is how the Earth's magnetic field was discovered.

With the help of modern spacecraft equipped with devices more sensitive than a compass, the magnetic fields of other planets were discovered as well—those of Mercury, Mars, Jupiter, and Saturn. For example, the Martian magnetic field was first detected by magnetometers on the Soviet space station Mars-3. At the planet's equator the magnetic field is $6 \cdot 10^{-8}$ T. Jupiter, as it turned out, has the strongest magnetic field ($4 \cdot 10^{-4}$ T at the equator—approximately 10 times that of the Earth's field).

In the beginning of our century the Sun's magnetic field was discovered. The Sun is just a rank-and-file member of the stellar community. No wonder that the magnetic fields of other luminaries were also found. Certain stars have fields tens of thousands of times stronger than that of the Sun.

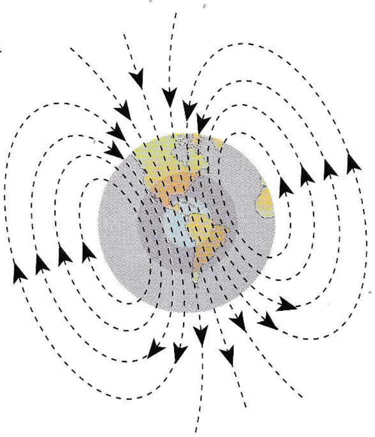
A magnetic field exists also in the interstellar medium. This is definitely not a field created by stars. The magnetic field of a star decreases with distance so radically (generally it's inversely proportional to the cube of the distance) that it usually doesn't reach even its nearest neighbor. At the same time, the magnetic field in our galaxy (the gigantic system of stars and gases seen at night as the great swath of the Milky Way) spreads over a huge distance exceeding by far the space between two neighboring stars. This field is rather weak—about 10^{-10} T.

The discovery of space magnetism raised the question of why it exists. There are few solid bodies in space, so the physical processes are occurring in liquid or gaseous media composed of electrons, positively charged ions, and neutral atoms and molecules. The Earth's magnetic field (figure 1a on page 14) arises from the motion of material (molten iron with impurities of carbon, sulfur, and silicon) in the liquid shell of the Earth's interior (it's a spherical layer between 0.19 and 0.55 of the Earth's radius). In gaseous stars (such as our Sun) the magnetic field is generated in the outer (partially ionized) shell, which is about 1/10 of a stellar radius thick.

Thus, cosmic magnetic fields are produced by currents flowing in an electrically conducting medium. In a similar way magnetic fields are generated in the coil of an electromagnet (figure 1b). However, to preserve the current in an electromagnet, one needs an external source of energy. Meanwhile, "battery" emfs

Art by Dmitry Krymov


Во
Вф



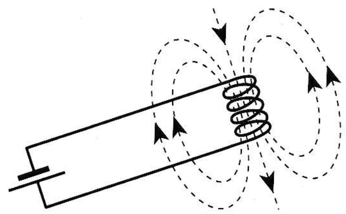


Figure 1

Comparison of magnetic fields. The magnetic field of the Earth (a) resembles that of a permanent magnet. However, the nature of these fields is different. The magnetism of a permanent magnet is explained by complicated quantum-mechanical properties of the particles in the solid body. The Earth's magnetic field, on the other hand, is generated by electric currents flowing in the liquid core. In this respect, the Earth's field is similar to that of an electromagnet (b).

produced from chemical or thermal effects are usually weak or absent in space. The main source of magnetic fields in space is the motion of the conducting medium itself.

Cosmic magnetic fields are not constant. For example, the large-scale solar field changes its polarity almost periodically once every 11 years. This process is connected with the cycling of solar activity, which manifests itself through dark sunspots, the solar corona, solar flares, and other phenomena. In recent years such cycles of activity were also found in many other stars. The familiar magnetic field of Earth also proved to be variable. Over long periods of time—about 200,000 years—it changes direction by 180°: the north magnetic pole becomes the south pole, and vice versa.

Why the magnetic field varies in the conducting medium

The electric field formed by a charge disappears when this charge is neutralized by another one of opposite sign. In plasma, which consists of equal amounts of moving positive and negative charges, any extra charge of one sign arbitrarily appearing in some area will quickly be neutralized. It's clear that the greater the electrical conductivity of the plasma, the shorter the period of neutralization of the electric field. The magnetic field in a plasma is generated by moving charges—electric currents. Clearly, this field will change with electric current.

Let's first examine how current changes in an ordinary metal wire that you can find in any room. Consider a winding of radius r made of wire with a cross section πl^2 and resistivity ρ. Let the winding be connected to a battery generating a current I, which produces a magnetic field B around it. The field is perpendicular to the plane of the winding and equals $B_{\max} = \mu_0 I/2r$, where $\mu_0 = 4\pi \cdot 10^{-7}$ T · m/A. The magnetic energy in this system can be estimated as

$$E \cong \frac{B_{\max}^2}{2\mu_0} \cdot \frac{4}{3}\pi r^3.$$

Then let's turn the current source off in such a way that the electric circuit remains closed. At first glance, the current will stop immediately due to the conductor's nonzero resistance R. In fact, however, the current won't stop all at once. Its rate of decrease is determined by the rate of the heat loss $P = I^2 R$. The current is fed by the magnetic energy stored around the winding. (Note that the magnetic energy is larger by far than the kinetic energy of the current carriers—that is, the electrons.) This energy will be consumed in the time

$$\tau \cong \frac{E}{P} \sim 10^{-6} \sigma l^2,$$

where $\sigma = 1/\rho$ is the conductivity of the material of the winding.

Let's estimate the damping time τ for copper wire of cross-sectional radius $l = 1$ cm. The conductivity of copper is about $6 \cdot 10^7\ (\Omega \cdot \mathrm{m})^{-1}$. This yields a time τ of less than one hundredth of a second.

Now let's turn to plasma in outer space. Usually its conductivity is the same as that of weak conducting metals, but the volumes occupied by the currents are huge. In this case the characteristic time for the field to change becomes very large. For example, the plasma conductivity in the upper shell of the Sun is about $10^6\ (\Omega \cdot \mathrm{m})^{-1}$ (mercury, nickel-chromium alloy, and bismuth have approximately the same σ at room temperature). The Sun's radius is about $7 \cdot 10^8$ m. So the damping time of the magnetic field produced by currents in a shell whose radius is one tenth of the Sun's radius is 10^{15} s, or 100 million years! Of course, this is just an approximation. Still, compare it with the fractions of a second characterizing the damping in copper wire. This means that ordinary electrical resistance has little effect on large-scale magnetic fields once they have been created. Then how do we explain the fact that the Sun's magnetic field, say, changes after a very short period of time—only 11 years?

The answer is that the medium is moving. The motion of a medium having free positive and negative charges is equivalent to the motion of a conductor. Textbooks on physics say that the motion of a conductor in a magnetic field brings about an electromotive force, and when the ends of the conductor are closed, electric current will flow in the circuit. This secondary current generates its own magnetic field, which adds to the original field. It's worth noting that in an ideally conducting plasma (that is, plasma whose resistance is zero) the total superimposed magnetic flux $\Phi = BS$ will be constant during these changes (S is the area of the circuit). So the magnetic field produced by a moving conductor seems to move along with it. Such a field is said to be

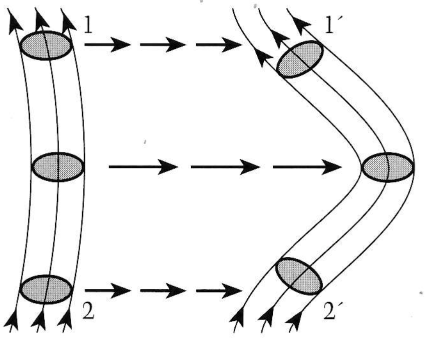


Figure 2

Change in the magnetic field when the particles of an ideally conducting plasma are shifted from position 1–2 to 1′–2′.

frozen in the plasma. One can picture the behavior of a magnetic field in a moving conducting medium as a thread thrown into a stream: not only is it carried along, it also stretches if the particles moving in the stream move away from one another.

When this "freezing" takes place, the movement of the plasma as a whole transports the magnetic lines of force. But the relative motions of different parts of the plasma deform the lines of force, curving and stretching them. A magnetic line connecting two previously adjacent plasma particles will connect them in the future as well—that is, it will follow the displacement of these particles (fig. 2). The number of magnetic lines of force permeating any area encircled by a closed circuit of moving liquid particles is constant (at $\rho = 0$). The magnetic field B can increase as a result of deformations that draw regions with the same field orientation closer together, and vice versa—it can decrease when regions with the opposite field orientation are brought together. It follows from the equality $B = \Phi/S$: if Φ is constant, then B changes in inverse proportion to S.

The simplest way to amplify the frozen magnetic field is to pinch (compress) a plasma layer (that is, decrease S). This idea is used as the simplest explanation for the huge magnetic fields of pulsars. A pulsar is a compact neutron star of radius $r \cong 10$ km. If we assume that the neutron star had been formed by compression of an ordinary star of radius $r_0 \cong 10^6$ km, we find that its radius decreased by a factor of 10^5 during compression. The conservation of magnetic field yields

$$\pi r_0^2 B_0 = \pi r^2 B,$$

where B_0 is the field before compression and B is the resulting field of the pulsar. Thus, the magnetic field is amplified by a factor of $(r_0/r)^2 = 10^{10}$. So the rather "modest" magnetic field of an ordinary star of, say, 0.01 T turns into a very strong field of 10^8 T.

On the other hand, plasma motion deforms and tangles the magnetic lines of force. For small-scale fields, as we saw, heat loss becomes significant.

It turns out that in certain kinds of plasma motion the effect of amplification of a large-scale field exceeds, or at least is not less than, its damping. In such a case the moving plasma works as an electromagnetic generator and so might naturally be termed a "dynamo" (at one time generators were called "dynamos"). The action of a dynamo is usually demonstrated by means of a wire frame with electric current rotating in a magnetic field.

Hydromagnetic dynamo

The dynamo in a conducting medium, in contrast to the usual generators, must work without wires, windings, and, most importantly, without an external magnetic field (which is a prerequisite for an ordinary dynamo[1]). As the energy source in the natural dynamo is only hydrodynamic plasma motion, it is called "hydrodynamic." To start the hydrodynamic generator one needs only a weak initial magnetic field. Such weak fields are usually due to various interactions among the charged plasma particles.

The idea of explaining the Earth's and Sun's magnetic fields on the basis of hydromagnetic motion was first proposed in 1919 by the well-known English physicist J. Larmor. But is such a self-inducing hydromagnetic dynamo possible—one that "works" without external sources of emf? The affirmative answer to this question is best illustrated by the following visual example.

Let's take a circular cord composed of individual rubber rings (fig. 3a). Let the rubber rings imitate the magnetic lines of force of the same orientation (marked with arrows). Now we curve the cord to form a figure eight (fig. 3b), place one ring of the eight on the other, and finally stretch the double-cord to restore its initial size (fig. 3c). Note that all the arrows point in the same direction. Thus, the number of magnetic lines of force passing through

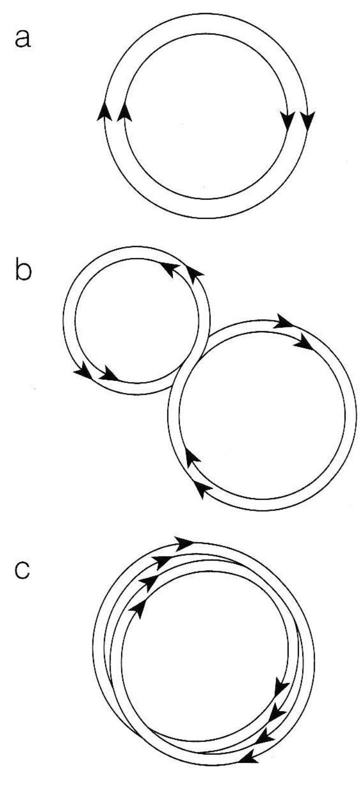


Figure 3

Doubling of the cord of magnetic lines of force.

[1]In a large-scale modern generator the current, which produces a magnetic field, is taken from the generator itself, and there is a close similarity between natural and commercial dynamos in this point. However, the placement of the windings and the overall design are matters of human ingenuity. Can nature achieve the same effect without insulating materials? That's the question!

any cross section of the cord is doubled. Repeating the procedure of folding and stretching, we can double in each step the number of magnetic lines in the cord until we reach the limit determined by the rubber's elasticity. Note that starting from any point in the initial cord, one could return to it making only one pass along the cord. However, it will take two passes to return to the starting point after the cord has been doubled. Thus, the new (doubled) cord isn't entirely equivalent to the initial one. To avoid this complication, in each step we can cut the cord at the joint of the figure eight, connect the free ends of the circles, and then put them together as one doubled cord that is equivalent to the initial one.

Under natural conditions the doubling operation can be performed by the movements of the medium in which the weak initial magnetic field is "frozen." It's not hard to imagine movements of this kind—say, repetitive turns about mutually perpendicular axes followed by stretching.

The elimination of nonequivalences of the doubled cord (in relation to the initial one) can be performed by electric resistance: the heat losses lead to the smoothing and removal of the magnetic lines where the curvature is greatest. Electric resistance also prevents the cord's cross section from becoming arbitrarily small after the repeated stretching that the doubling entails. The resistance will keep the cord's diameter almost constant.

Our example of the dynamo is simple and easy to picture. But is that how it's done in nature? If so, the motion of the conducting medium must be organized in a particular way. Let's see what kind of movements such a medium performs in space.

Nonuniform rotation

For all the planets, stars, and galaxies, the simplest motion is rotation. However, if all the parts of a body rotate with the same angular velocity (that is, rotate uniformly), the magnetic lines of force will rotate together with the body and remain practically constant. To bend, stretch, and fold the lines, one needs the relative motion of the neighboring plasma particles—that is, the rotation must be nonuniform. Now, solid bodies make only uniform circular motion—otherwise they would be destroyed. Liquid and gaseous bodies, on the other hand, can rotate nonuniformly. Observations of the Sun have shown that the equatorial parts of our star make a complete revolution several days earlier than the polar regions. Indirect data show that the unseen inner regions of the Sun rotate even faster. The gaseous disk of our galaxy also rotates nonuniformly: its inner regions perform more rotations in a given time than its outer regions.

To visualize more clearly this nonuniform rotation, think of the concentric circles of a carousel (fig. 4a). Let the circles that are closer to the center rotate more quickly than the outer ones.

Now let's seat several people on various circles such that they are arrayed on the same diameter of the largest circle. The passengers are given a rubber rope attached to the ends of the diameter, imitating a magnetic line of force. Now let's start the carousel. It's not hard to see that after several turns the rubber rope (magnetic line) will curve into a spiral (fig. 4b). At first the line lay along the diameter, and then it shifted closer to the circumference. Thus, the nonuniform rotation turned a radial magnetic field into an almost tangential (azimuthal) one.

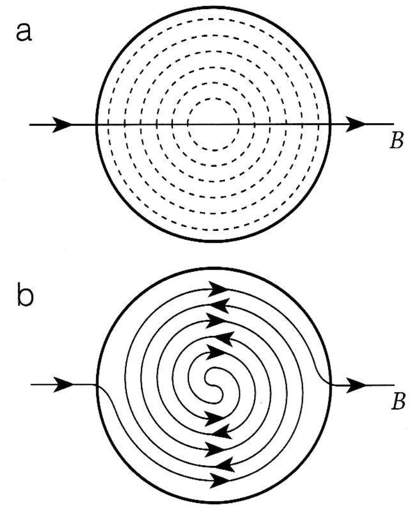


Figure 4

A horizontal magnetic line of force turning into a spiral after a number of turns of a nonuniformly rotating body.

Notice that the rotation forms pairs of circles with magnetic fields directed counter to each other. One circle will be located above the carousel, the other will be below it. However, as in the previous case, a problem arises: how do we keep the initial radial field constant?

Perhaps some other type of motion would help? Can one make something like a figure eight?

Helicity

The medium of outer space is characterized by chaotic, irregular, one might say turbulent, motion. We see such motion when observing water flowing in a mountain stream or near a whirlpool. It's also characteristic of the surrounding air. It is because of this chaotic motion that perfume or smoke quickly spreads.

Due to its chaotic, irregular nature, the turbulent movements of the conducting medium (plasma) usually entangle any magnetic field rather quickly and reduce to it to fragments. However, under space conditions the turbulence has a certain "regular" character.

Instead of real turbulent flow, let's imagine a set of individual spiral vortices that combine translational motion with rotation about their axes. Let some of them rotate clockwise, the rest of them counterclockwise. The spiral movement of the vortex is able to lift up the magnetic line so that it looks like the Greek letter omega (fig. 5) and then turn it. However, clockwise and counterclockwise vortices produce loops with opposing magnetic fields. The question is, what will be the effect of a large number of vortices whose direction of rotation varies—that is, real turbulent flow?

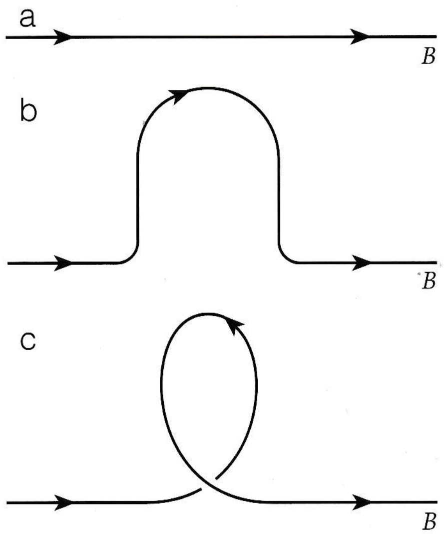


Figure 5

Spiral motion of a conducting medium stretching a magnetic line (a) into a loop (b) and rotating the loop (c). A current parallel to the field is generated.

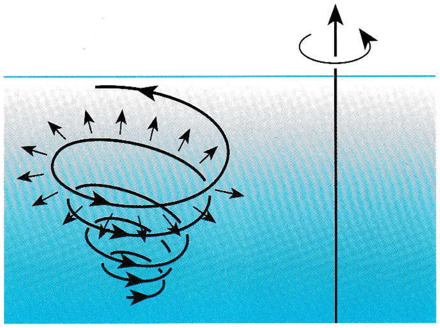


Figure 6

Rotating plasma layer whose density decreases upward. The overall rotation of the layer imparts additional angular velocity to an ascending and expanding vortex.

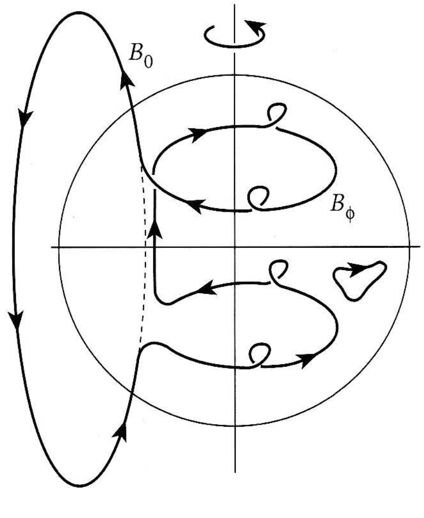


Figure 7

The combined effect of nonuniform rotation and spiral motion on the magnetic field. The rotation of the initial field (B_0) winds the azimuthal field (B_ϕ). The spiral motion lifts loops up from this field, then these loops tear off, turn, stretch, and merge to produce a new field directed parallel or counter to the initial field B_0.

In the homogeneous situation, the number of clockwise and counterclockwise vortices is the same. Such a set of vortices has the property of mirror symmetry. Indeed, let's install a mirror at any location and look at the reflection of the vortices: in the mirror the clockwise vortices become counterclockwise vortices, and vice versa. However, since the number of both kinds of vortices is the same, the picture remains basically unchanged after reflection.

Now let's recall that bodies in outer space rotate. Imagine this situation: in a layer of liquid the density is greater at the bottom than near the surface and the entire layer rotates about an axis running through it. There are vortices in this layer (both clockwise and counterclockwise). How will they behave?

Each vortex turns "by itself," and in addition it can rise or sink in the layer. Any ascending vortex will expand like a bubble that finds itself in a more rarefied medium. The lateral movement that arises during this process (fig. 6) will deviate from the purely radial due to the rotation of the layer as a whole (much as a person walking on a rotating disk deviates from a radial course). This means that the vortex is given an additional rotation caused by the overall rotation of the layer.

Clearly, the effect of the overall rotation is the same for all rising vortices—clockwise and counterclockwise. The vortices that are sinking are compressed, but the effect of the overall rotation will be the same as for rising vortices. In sum, the effect of the overall rotation on the aggregate of vortices in a nonuniform layer leads to an increase in the number of vortices rotating in the same direction as the plasma as a whole. However, a plasma with, say, counterclockwise vortices outnumbering clockwise can easily be seen to have lost the property of mirror symmetry.

The turbulence described is termed "spiral." Notwithstanding its chaotic nature, it has on average a certain added rotation. Helicity is the organizing feature that characterizes turbulence in rotating heavenly bodies with nonuniform densities such as planets, stars, and galaxies.

Spiral turbulence in heavenly bodies acts in concert with nonuniform rotation. As we saw previously (fig. 4), nonuniform rotation transforms a radial field into a field directed almost azimuthally—or to be more exact, along a spiral. Spiral turbulence lifts up many loops in the azimuthal field and then turns them in different directions (fig. 7). But as vortices of a certain orientation dominate, the summing of the loops will result in a common radial field with a particular direction. If the new radial field has the same direction as the initial one, the magnetic field becomes stronger. If the new field is directed counter to the original field, a reversal of the field is possible. This makes it possible to explain the repeated changes in field direction that occur in the Sun and other stars. Note that in the case of the figure eight we obtained a dynamo with continual amplification of the field—but with no reversals.

One shouldn't forget that the hydromagnetic dynamo works under marked turbulent tangling and destruction of magnetic lines of force. Thus, strengthening and weakening of the field take place simultaneously with the continual birth of new and the death of old magnetic fields.

The hydromagnetic dynamo is widely used to describe the magnetic fields of stars, planets, and galaxies. In particular, such a dynamo seems to be the main "motor" of the repetitive solar activity.

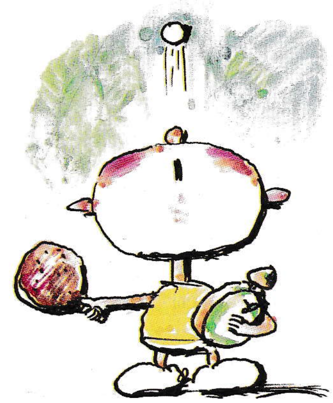
Just for the fun of it!

B151

Winnie wins? Winnie-the-Pooh and Rabbit took a bag of 101 candies to play a mathematical game. Each of them in turn takes a number of candies from 1 to 10 out of the bag. When the bag is empty, they count their candies: Winnie-the-Pooh wins if the two numbers are coprime, otherwise Rabbit wins. Which of them can force a win and how, if Winnie begins the game? (K. Kohas)

B152

Long printout. A computer printed out two numbers: 2^{1995} and 5^{1995}. How many digits in all were printed? (V. Pushnya, 10th grader)

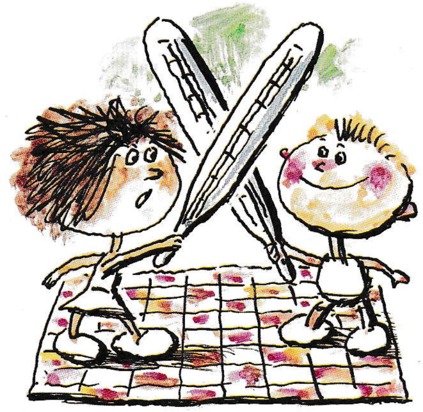


B153

Up or down? A Ping-Pong ball is tossed into the air. Will it take longer for it to go up or to come back down? (A. Savin)

B154

Square a cube. You are given a cardboard cubical box without a cover (its sides and bottom are squares of unit area). Cut it into three pieces that can be put together to form a square of area 5. (V. Proizvolov)



Art by Pavel Chernusky

B155

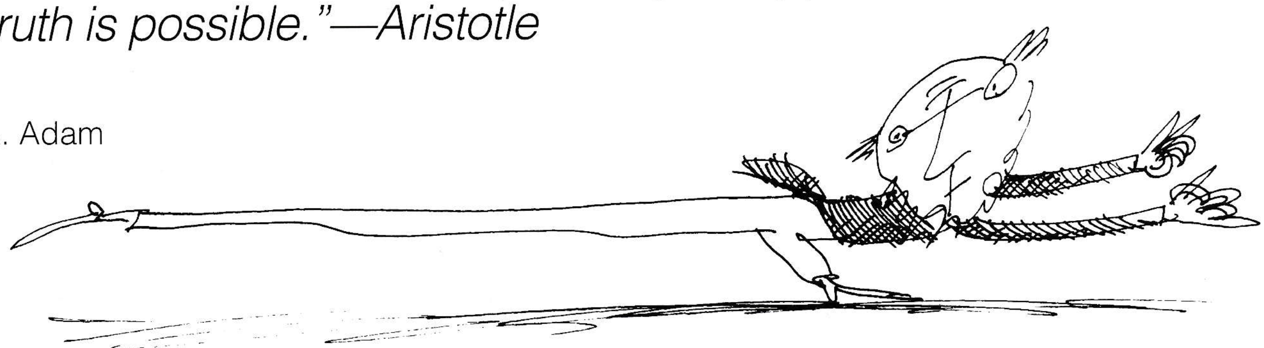
Sick of chess. Judith and Nigel played the same number of games in a chess tournament, fell ill, and quit. All the other participants played against one another, as was intended by the rules. The total number of games played is 23. Did Judith and Nigel play against each other during this tournament? (V. Bliznyekov, 10th grader)

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Educated guesses

"It is the mark of an instructed mind to rest satisfied with the degree of precision which the nature of the subject permits and not to seek an exactness where only an approximation of the truth is possible."—Aristotle

by John A. Adam



WE MAY NOT BE AS ERUdite as Aristotle, or as brilliant as Enrico Fermi, but we can learn to apply elementary reasoning to obtain "ballpark estimates" for problems (subsequently named "Fermi problems") in the manner attributed to that great physicist.

Several years ago a short article by David Halliday appeared in *Quantum* (May 1990). It was called "Ballpark Estimates," and in the context of a specific problem Halliday showed how to obtain order-of-magnitude answers to problems by breaking them down into their components and making appropriate common-sense estimates. The problem was to estimate how many "rubber atoms" are worn from an automobile tire for each revolution of the wheel. We shall consider a slight variant of this problem below, but what I find especially appealing in Halliday's article is the dialogue he provides en route with a typical reader's questions. While not necessarily a prerequisite to this article (having got you to read this far, I don't intend to let you go easily!), I urge you to read it nonetheless.

Of course, the ideas expressed and methods used in such Fermi problems go far beyond physics into the realm of everyday activities (though filling the Earth with sand may not qualify as an *everyday* activity). Two excellent resources I have enjoyed reading and using are *Innumeracy* by John Allen Paulos and *Consider a Spherical Cow* by John Harte. You'll recognize some of the problems cited here if you have already encountered these books. After a while you'll get comfortable with posing and estimating answers to your own Fermi problems. The book by Paulos will be an eye-opener for many: in particular, he shows the power of plausible assumptions coupled with simple calculations. The book by Harte is a good introduction to mathematical modeling (particularly environmental problem solving) with little or no use of calculus. While we're on the subject of interesting books, *The Universe Down to Earth* by Neil de Grasse Tyson has some chapters (1 and 3) relevant to the present article.

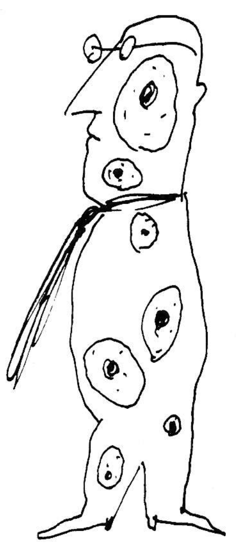
In much of what follows, letters are used to represent typical dimensions or other quantities. This will enable you to obtain your own estimates, though you should resist the temptation to just "plug in" your numbers in the formula without following the prior reasoning. Almost certainly we'll differ on typical sizes of objects (for instance, grains of sand). But almost as certainly we'll choose typical dimensions in the range (for this example) of 10^{-1} mm $\le d \le 2$ mm, so we probably won't differ significantly in our

Art by Sergey Ivanov

subsequent order-of-magnitude answers. Remember that it's to be understood that whenever ratios of dimensional quantities are to be sought, a conversion of units may be necessary in order to compare like quantities. For completeness, actual numerical estimates are given—some of their values may surprise you.

Needless to say, the question will be asked: so what if I know how to estimate the number of grains of sand that would fill Buckingham Palace? (Now there's a thought!) Apart from a spell in jail for attempting to verify such an estimate, it's a great encouragement to realize that a "back of the envelope" type of calculation can be carried out with a modicum of salient information for a "real world problem." Not only might this save a considerable amount of money and computer time on occasion, it might also give you a greater appreciation for the power of arithmetic. I've seen the "lights go on" when intelligent, educated people realize at last the distinction between 10^6 seconds (11½ days) and 10^9 seconds (32 years). Sometimes we need the right pegs to hang numbers (and concepts) on!

Among the simplest estimation problems are those arising from ratios of lengths, areas, and volumes. Thus, if D is a typical linear dimension of a given object (for example, a classroom), and $d < D$ is a typical linear dimension of a smaller object (for example, a piece of popcorn—we'll say popped!), then $N = D^3/d^3$ is the approximate number of smaller objects that would fill the latter. Thus, by using appropriate choices of D and d we can come up with estimates for the following questions.



1. *How many golf balls does it take to fill a suitcase?*

2. *How many pieces of popcorn does it take to fill a room?*

3. *How many soccer balls would fit in an average-size home?*

4. *How many cells are there in a human body?*

5. *How many grains of sand would it take to fill the Earth?*

Related problems involve volumetric measures of fluids.

6. *What is the volume of human blood in the world?*

7. *How many one-gallon buckets are needed to empty Loch Ness (and thus expose the monster)?*

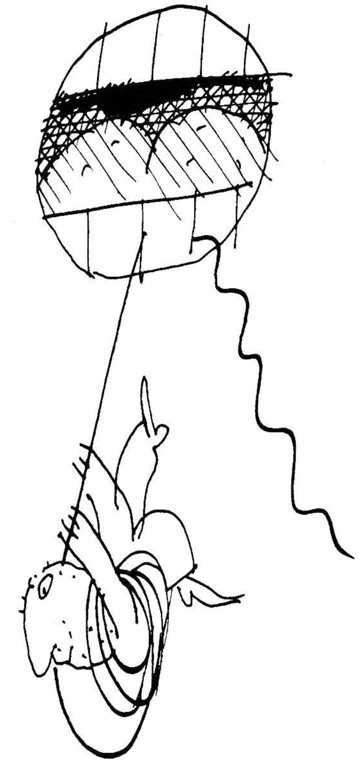
Sometimes everyday objects are obviously represented (or misrepresented) by cubes. Thus, if we are asking how many objects with a typical linear dimension d will fill a space with linear dimensions a, b, c, the formula $N = abc/d^3$ is appropriate. So for problem 1, we might suggest $a = 20$, $b = 24$, $c = 8$, and $d = 1.5$ inches, respectively, so $N \cong 10^3$. For problem 2, suppose $a = 10$ ft, $b = 20$ ft, $c = 15$ ft (classroom size), and $d = 1$ cm. Then, after conversion to metric units, $N \cong 3{,}000 \cdot 30^3 \cong 10^8$. For problem 3, consider $D = 30$ ft and $d = 1$ ft, which gives $N \cong 10^4$. Problem 4 yields 10^{14}, and the answer to problem 6 is less than $1/200$ mi^3 (both of these are discussed below). For problem 5, values of $D \cong 10^4$ km and $d = 1$ mm yield $N \cong (10^4 \cdot 10^3 \cdot 10^2 \cdot 10)^3 = 10^{30}$. A cubic Earth, you ask? Don't worry, you'll get over it without falling off (see the comment on problem 14 below). Using the fact that 1 ft^3 of liquid (water, soup, blood, and so on) is about 7.5 gallons, we arrive at $N \cong 10^{12}$ buckets to empty Loch Ness (problem 7). The loch has a volume of approximately 2 mi^3, so $2 \cdot 5{,}280^3 \cdot 7.5 \cong 10^{12}$. And while we're talking about gallons, here's problem 8.

8. *One gallon of paint is used to cover a building of area A. How thick is the coat?*

Clearly, if A is in square feet, then the thickness $d = 1/7.5A$ ft. For the "cubical house" of problem 3 (full of soccer balls by now, you'll recall), $A = 6 \cdot 30^2 \cong 5 \cdot 10^3$ ft^2, so $d \cong 10^{-5}$ ft $\cong 10^{-4}$ in.

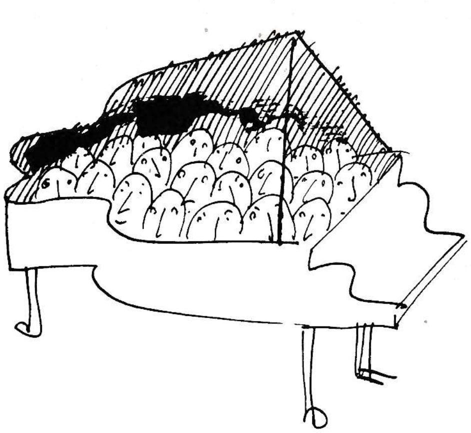
Questions of a more sophisticated nature require, not surprisingly, more terms in the estimation formulas. Thus we have the following problems.

9. *How much dental floss does a convict need?* A recent newspaper article featured the story of an inmate at a correctional center in West Virginia who escaped from the prison grounds by using a rope made from dental floss to pull himself



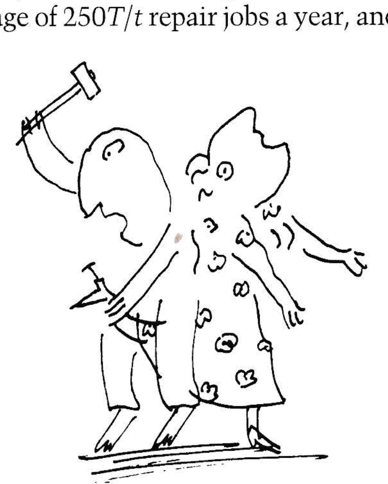
over the courtyard wall. The rope was estimated to be the thickness of a telephone cord, and the wall was 18 ft high. Taking 4 mm for the diameter of a telephone cord and 1/2 mm for the diameter of the floss, then the number of floss fibers in a cross section is $(4 \div 1/2)^2 \cong 60$, and if each packet of floss contains the standard length of 55 yards, the number of packets required is $N \cong (20 \cdot 60)/(55 \cdot 3) \cong 7$.

10. *Estimate the number P of piano tuners in a certain city or region.* Consider a population in the region totaling N, with an average of p pianos per family (generally $p < 1$).



Suppose that pianos are tuned b times a year on average (generally we expect $0 \le b < 2$), so the number tuned per year is approximately Npb/n_1, where n_1 is the average size of a household. If each tuner tunes n_2 pianos a day ($0 < n_2 < 4$ in general), this corresponds to $250n_2$ pianos per year (for a reasonable working year of $50 \cdot 5$ days). So the number of tuners in the region (city, town, country) is approximately $Npb/250n_1n_2$. Let's pop in some numbers. If, for New York City, say, $N \cong 10^7$, $n_1 = 5$, $b = 0.5$, $p = 0.2$, $n_2 = 2$, then $P \cong (10^7 \cdot 10^{-1})/(250 \cdot 10) \cong 4 \cdot 10^2$—that is, an order of magnitude of 10^2 to 10^3.

11. *Estimate the number C (for cobbler) of shoe repairers in a city or region.* If such a person spends on average t hours on a repair job in an average working day that's T hours long, T/t is the average number of repairs performed per day. Clearly, some shoes are worth repairing and some are not. Suppose the "average pair of shoes" is repaired on average every n years, leading to a repair rate of $1/n$ per year. For a 250-day working year, our cobbler can perform an average of $250T/t$ repair jobs a year, and

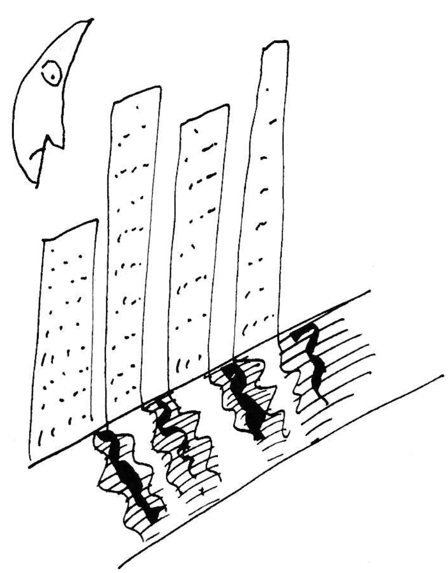


in a population of N, repairs N/n pairs of shoes each year. This leads to an estimate of $Nt/250nT$ cobblers in the region. Thus, if we take as our region this time the whole of the United States (we're being a little ambitious here, of course, but this is a question I'm constantly being asked), then $N \cong 2.5 \cdot 10^8$, $t \cong 1/2$, $T \cong 10$, $n \cong 2$; so $C = (2.5 \cdot 10^8 \cdot 1/2)/250 \cdot 2 \cdot 10) \cong 10^4$.

12. *Estimate how fast human hair grows (on average) in mph.* If the hair is cut every n months (usually $n \le 2$) and the average amount cut off is x inches, then x/n inches per month $\cong x/n \cdot 1/(5{,}280 \cdot 12) \cdot 1/(30 \cdot 24)$ mph $\cong 10^{-8}(x/n)$ mph. If $n = 2$ and $x = 1$, then the rate of hair growth is approximately 10^{-8} mph.

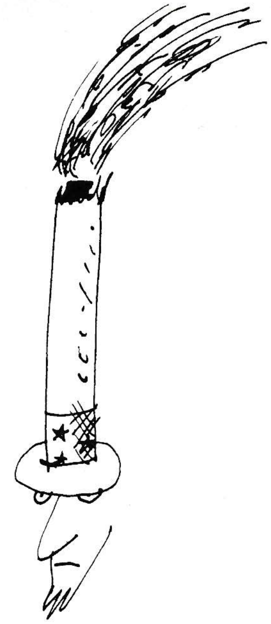
Now back to the blood problem (number 6).

6. (redux) *Estimate the total volume of human blood in the world.* For a population of $5 \cdot 10^9$ with an *average* of 1 gallon of blood per person,



$V \cong 5 \cdot 10^9/7.5 \cong 7 \cdot 10^8$ ft^3. This, as Paulos points out, could be contained in a cube of side length $(7 \cdot 10^8)^{1/3} \cong 900$ ft. Putting things a little more prosaically, since Central Park has an area of 1.3 mi^2, all this blood would cover Central Park to a depth of about $(7 \cdot 10^8)/[1.3 \cdot (5{,}280)^2] \cong 20$ ft. Hmm.

13. *Estimate the number of cigarettes smoked annually in the US.* Let f be the fraction of people in the population who smoke and n the average number of cigarettes smoked



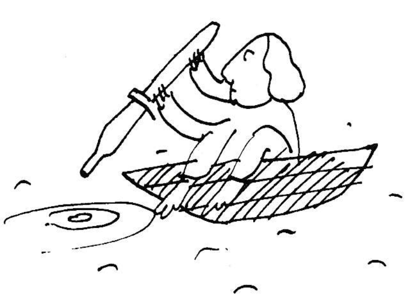
per day. Then $N = 2.5 \cdot 10^8 \cdot 365 \cdot fn \cong 10^{11}$, if $f \cong 10^{-1}$ and $n \cong 10$.

14. *The asteroid problem.* In the light of the impact(s) of ex-comet Shoemaker–Levy on Jupiter's outer atmosphere, the question has been raised: could it happen here on Earth? It may have happened already—one theory for dinosaur extinction (not Gary Larson's)[1] is that about 65 million years ago such an encounter occurred—this time with an asteroid. Eventually dust from the impact settled back on the surface of the Earth, having done a superb job of blocking sunlight and thus devastating plant and animal life. According to one hypothesis, about 20% of the asteroid's mass was uniformly deposited over the (now rather inhospitable) surface of the Earth—about 0.02 gm/cm^2. Question: how large was the asteroid? (You may feel that at this point, a more appropriate question would be: "What was the name of the bus driver?" But don't worry, we'll get to that later.) Okay—the mass is clearly about $4\pi R^2 \cdot 0.02 \cdot 5$ if R is the radius of the Earth in centimeters. This must be equated to density times volume for a cube of side length L (this is the simplest geometry to consider: the largest sphere that can be inscribed in a cube of side L differs in

[1]His memorable cartoon shows several tough-looking dinosaurs standing around, smoking cigarettes. The caption reads: "The real reason dinosaurs became extinct."—*Ed.*

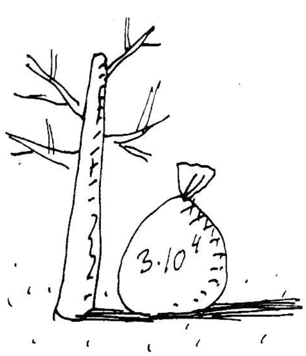
volume from that cube by a factor $\pi/6 \cong 1/2$, so this won't affect our order-of-magnitude estimate). Suppose we take a typical rock density of 2 gm/cm^3, so that $2L^3 \cong 0.4\pi R^2$, which gives us $L \cong (0.2\pi R^2)^{1/3}$. Since $R \cong 4{,}000 \cdot 1.6 \cdot 10^5$ cm (converting miles to centimeters) = $6.4 \cdot 10^8$ cm, then $L \cong 6 \cdot 10^5$ cm, or 6 km (10 km by order of magnitude). This is not unreasonable for an asteroid (even though the dinosaurs may disagree).

15. *Thickness of an oil layer.* Perhaps no one likes to take their medicine. Rumor has it that Benjamin Franklin noted that 0.1 cm^3 of oil (was it cod-liver oil?) dropped on a lake spread to a maximum area of 40 m^2. If d is the thickness of the



layer in meters, then $40d = 10^{-7}$, so $d = 25 \cdot 10^{-10}$ m, or 25 angstroms. Interestingly, this corresponds to a "monomolecular layer" of 10–12 atoms (with atom–space–atom–... for a molecule), which is about right for a molecule of "light" oil.

16. *The number of leaves on a tree.* If r is the typical radius of a tree's canopy, the surface area of the canopy is $4\pi r^2$; and if d is (in the same units as r) a typical leaf size, an estimate for the number of leaves is $4\pi r^2/d^2$. Clearly leaves don't cover the "surface" of the canopy continuously; this does, however, compensate for the fact that there are many leaves on branches inside the canopy.



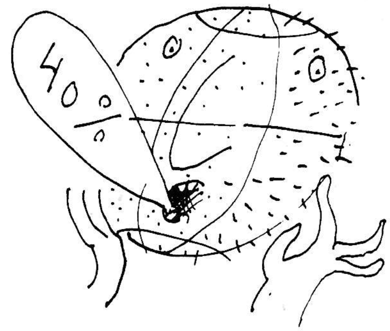
For a small tree (for example, a 15- to 20-year-old yew), the leaf canopy has a radius $r \cong 4$ ft and $d \cong 1$ in, so $N \cong 3 \cdot 10^4$—that is, an order of magnitude of 10^4–10^5 in general, if we include larger trees as well.

17. *Weekly supermarket revenue.* If there are n_1 checkout lines serving an average of n_2 customers per hour, the average customer receipt is x dollars, and the store stays open an average of n_3 hours a day, then in an average week $R \cong 7n_1n_2n_3x$ dollars. If, for example, $n_1 = 10$, $n_2 = 10$, and $n_3 = 14$, then we find that $R \cong 10^5$ dollars.

18. *Daily death rate in a city or region.* If in a city or region of population n_1 the average number of deaths per day (as listed, for example, in the obituary section of the local newspaper) is n_2, we can by a simple proportion get an estimate of the daily death rate d in the country (with a population N). Thus,

$d \cong Nn_2/n_1$. Clearly there are limits to the validity of this crude analysis. Death rates vary considerably from country to country. Nevertheless, one can get "lower bound" estimates for world death rates in a similar fashion. Thus, if $n_1 \cong 10^6$ and $n_2 \cong 30$, then $N \cong 2.5 \cdot 10^8$.

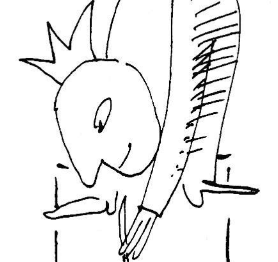
19. *The number of blades of grass on the Earth.* If 40% of the Earth's surface is covered by land, a fraction f_1 of this land is covered by grass. If the average number of blades of grass per square inch is n, then $N \cong (0.4)4\pi R^2 f_1 n$ for R measured in inches. Thus, for $R \cong 4{,}000 \cdot 5{,}280 \cdot 12$, $f_1 \cong 10^{-2}$ or 10^{-1} (this is difficult to estimate without a little research), and $n \cong 20$, then $N \cong 10^{16}$ or 10^{17}.



Now let's return to a variant of the car tire problem.

20. *What is the average depth of tread lost per revolution of a car tire?* This can be answered by a simple proportion: the distance d we require is to a typical tread t (for a new tire) as tire circumference $2\pi R$ is to length of useful mileage L. Thus, $d \cong 2\pi Rt/L$, which for $R = 1$ ft, $L = 5 \cdot 10^4$ mi, $t = 5$ mm corresponds to (after conversions!) to $d \cong 10^{-7}$ mm.

21. *Population square.* If each person on Earth were given enough space to stand comfortably on the ground without touching anyone else, estimate the length of the side of a square that would contain everybody in this way. If we give everyone a square 1/2 m on a side, then the side of the large square is $L \cong (5 \cdot 10^9)^{1/2} \cdot 1/2 \cdot 10^{-3}$ km $\cong 35$ km.

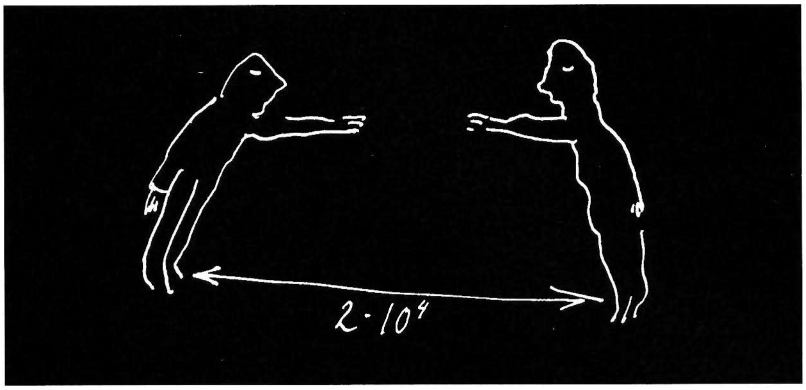


22. *Human surface area and volume.* To estimate these quantities crudely but quickly, consider a cylinder of radius r and height h: if $r \cong 1/2$ ft and $h \cong 6$ ft, then $V = \pi r^2 h \cong 5$ ft^3, and $S = 2\pi rh \cong 20$ ft^2. Since 1 ft $\cong 0.3$ m, $V \cong 0.1$ m^3. Now we're in a position to return to problem 4.

4. (redux) *Estimate the number of cells in the human body.* If we assume an average cell diameter of 10 microns, or 10^{-5} m, then since 1 ft $\cong 0.3$ m, V from problem 22 is

approximately 10^{-1} m^3, so $N \cong 10^{-1}/(10^{-5})^3 \cong 10^{14}$ cells.

23. *The average rate of growth of a child from birth to 18 years.* Over this time span the "speed" equals approximately $(h_{18} - h_0)/18 \cong 1/18$ m/yr $\cong 10^{-3}/(20 \cdot 400 \cdot 20)$ km/h $\cong 10^{-8}$ km/h—that is, about the same order of magnitude as the speed of hair growth! Perhaps we could label children as super- or subfollicular depending on whether or not they grow faster than their hair!



The remaining estimation problems concern SETI (the search for extraterrestrial intelligence) and interstellar launches. The astronomer Frank Drake has done the work for us in providing the famous Drake formula for the number N of extant technical civilizations in the galaxy. Here "technical" can be taken to mean at least as technologically capable as we are on planet Earth. Thus, if n_s equals the mean number of stars in the galaxy, f_p the fraction of these stars with planetary systems, n_p the mean number of planets suitable for life per planetary system, f_b the fraction of planets where life actually evolves, f_i the fraction of those $n_p f_b$ on which intelligent organisms have evolved, f_c the fraction of those intelligent species that have developed communicative civilization, and f_l the mean lifetime of those civilizations in terms of the age of the galaxy, then

$$N \cong n_s f_p n_p f_b f_i f_c f_l.$$

Of the seven quantities on the right of this expression, the first is astronomical in nature and well known to be about $4 \cdot 10^{11}$. The next two numbers are really educated astronomical guesses. The two following (f_b and f_i) are biological in nature, and here we're on pretty shaky ground, because we only have a sample space of one (ourselves!). The final two numbers are sociological in nature, and so in this context they're pure guesswork! Thus it happens that the numbers one puts in are indicative of one's philosophical stance: like it or not, we all have presuppositions about the universe we inhabit. Just for fun, let's see where this leads for the optimist and the pessimist. In both cases we might take $f_p \cong 0.2$ (remember that almost half the stars in our galaxy are thought to be binary systems at least) and $n_p \cong 0.1$. For the remaining four numbers, our optimist takes 1.0, 1.0, 0.5, and $10^6/10^{10} = 10^{-4}$, respectively, yielding $N \cong 10^5$–10^6. Our pessimist, on the other hand, takes the last four numbers to be 0.1, 0.1, 0.1, and $10^4/10^{10} = 10^{-6}$, respectively, yielding $N \cong 10$. Which are you?

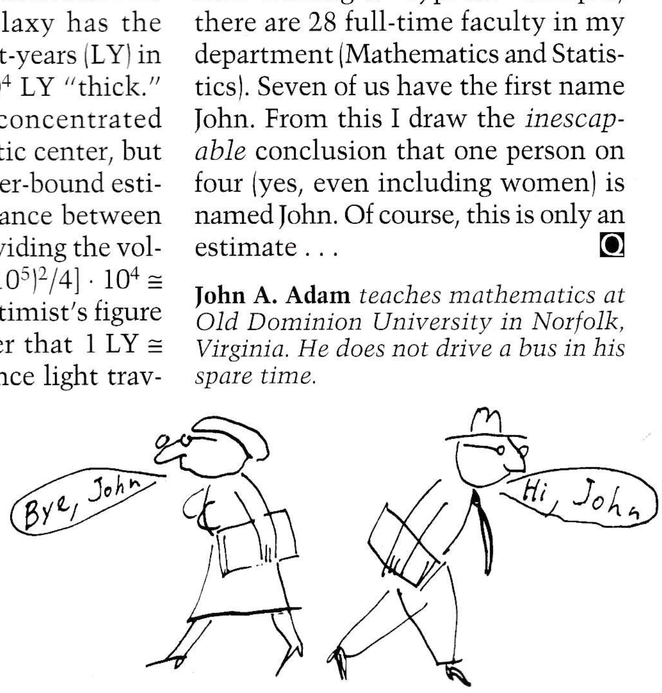
At this point, a timely reminder: whether for a debate or a mathematical model or merely an estimate, **the argument is only as good as the weakest assumption built into it.**

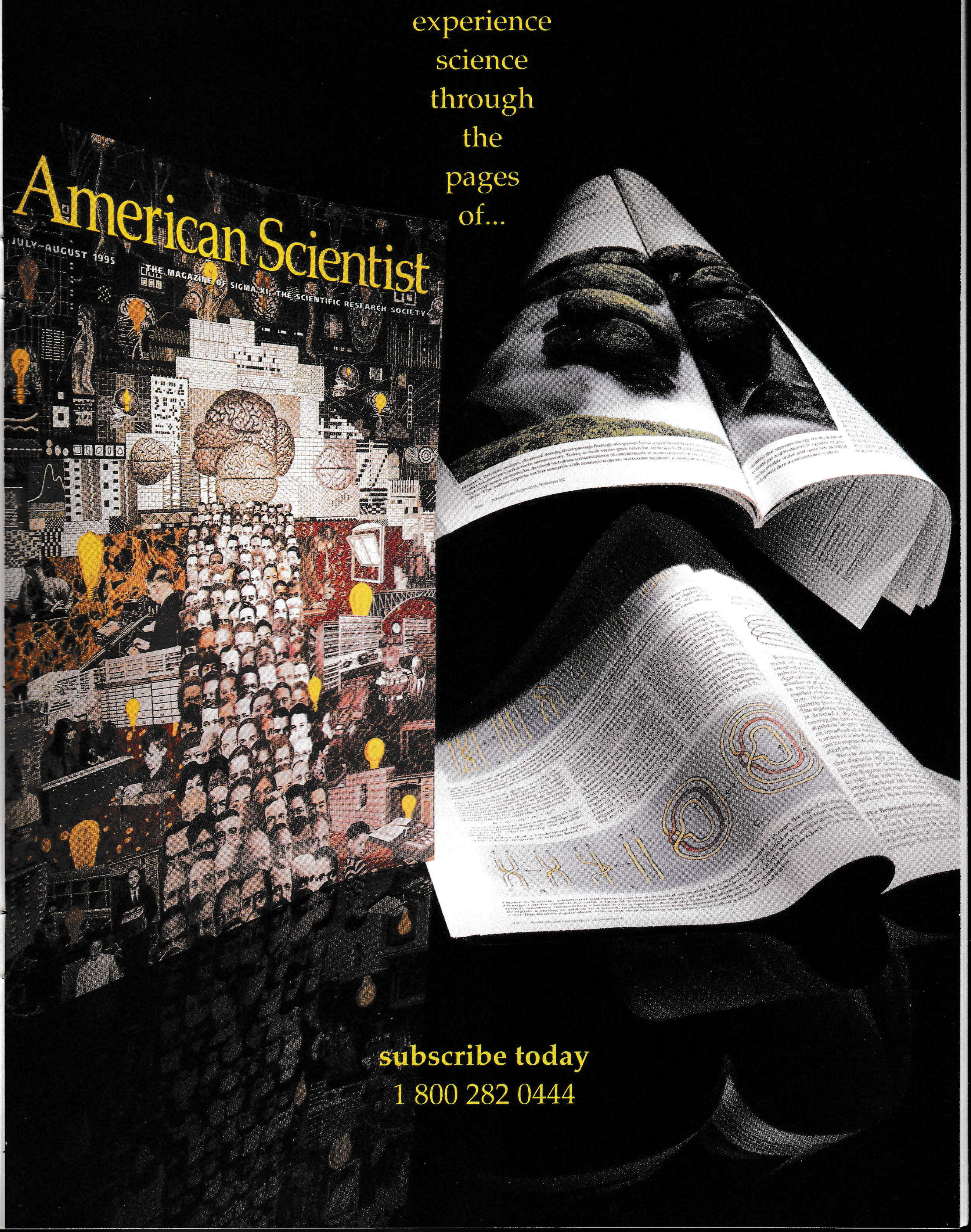
24. *Mean distance between two civilizations.* Our galaxy has the shape of a disk 10^5 light-years (LY) in diameter and about 10^4 LY "thick." Obviously stars are concentrated more toward the galactic center, but we can get a crude upper-bound estimate of the mean distance between two civilizations by dividing the volume of the galaxy $[\pi(10^5)^2/4] \cdot 10^4 \cong 10^{14}$ cubic LY by the optimist's figure of $N \cong 10^6$. (Remember that 1 LY $\cong 6 \cdot 10^{12}$ mi is the distance light travels in one year. Work it out for yourself.) Taking the cube root of 10^8 gives us approximately 500 LY. On the other hand, if $N \cong 10$ (the pessimist's estimate), the distance is $2 \cdot 10^4$ LY.

25. *How many launches of interstellar space vehicles might we expect per year?* Suppose that on average each civilization is able to launch s such vehicles per year. If $N \cong 10^6$—our most optimistic estimate—there will be (at steady state) some $10^6 s$ vehicles arriving per year somewhere or other within the galaxy. Suppose there are approximately 10^{11} interesting places to visit (each star!). Then we can expect $10^6 s/10^{11} = 10^{-5} s$ arrivals at a given "interesting place" per year. Suppose it is claimed that here on Earth we receive v such visits per year. The mean launch rate s should then be $10^5 v$ per year, or a total of $10^{11} v$ launches per year within the galaxy. This corresponds to 10^{11}–10^{14} if $v = 1$–10^3. All in all, it seems rather excessive, especially if you try to compute the quantity of material required to make such large numbers of spacecraft!

Oh, yes—one more thing. In problem 14 I asked (among other things) what was the name of the bus driver. There's a good chance it's John. Why? A simple estimate will suffice. Taking a "typical" sample, there are 28 full-time faculty in my department (Mathematics and Statistics). Seven of us have the first name John. From this I draw the *inescapable* conclusion that one person on four (yes, even including women) is named John. Of course, this is only an estimate . . .

John A. Adam *teaches mathematics at Old Dominion University in Norfolk, Virginia. He does not drive a bus in his spare time.*




experience
science
through
the
pages
of...
American Scientist
JULY–AUGUST 1995
THE MAGAZINE OF SIGMA XI, THE SCIENTIFIC RESEARCH SOCIETY
subscribe today
1 800 282 0444

Clarity, reality, and the art of photography

"Is there always an advantage in replacing a blurred image with a sharp-focused picture? Isn't the blurring frequently just what one needs?"—Ludwig Wittgenstein, Philosophical Investigations

by Mark L. Biermann

IS SEEING ALWAYS BELIEVING? Or for that matter, is seeing necessarily understanding? Recently, the veracity of images has been called into question because of digital image processing techniques. In movies like *Rising Sun*, based on a novel of the same name, the ability of computer experts to manipulate images was pivotal to the story line. *Time* magazine used image processing to alter a cover photograph and to change, or "morph," Bill Clinton into Jimmy Carter.

Yet the manipulation of images in order to convey specific information and impressions is not a recent development. For decades photojournalists and photographic artists have been manipulating cameras and film to produce "special effects" and ambiguous images. While digital image processing relies on image-capturing techniques and sophisticated computers, photographers can produce surprising images using basic characteristics of cameras and similar imaging systems. One such characteristic is the depth of field. In figure 1, I photographed everyday objects but controlled the depth of field to produce a rather strange photograph. By studying depth of field, we can understand how such a photograph is made.

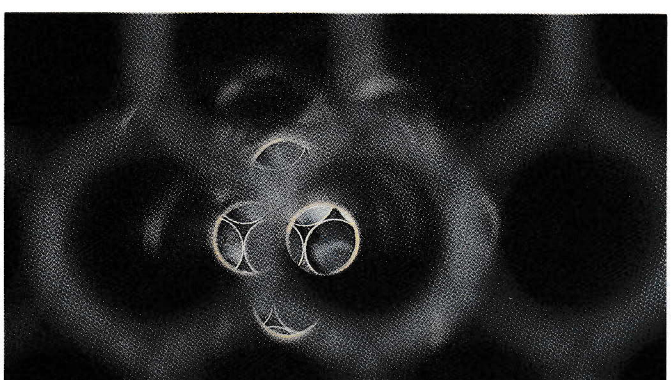


Figure 1

Throughout the following discussion, a still camera, such as a single-lens reflex camera, is used as an example of a typical imaging system. However, the results of this discussion can be applied equally well to other imaging systems, such as television cameras or camcorders.

Depth of field

In taking any photograph, the camera is focused for a single subject distance. However, a person viewing the photograph will see that objects appear to be in focus over a range of distances from the camera. For example, a camera focused on an object 4 m away may produce a photograph in which objects as close as 3 m from the camera and as far as 6 m from the camera all appear in focus. The range of distance from the camera over which subjects appear in focus is referred to as the depth of field. The depth of field for this situation is 3 m, running 3 m to 6 m from the camera.

Why does a range of distances appear in focus in a photograph even though the camera was focused at a single distance in front of the camera? The answer lies in the final imaging system for most photographs, the human eye. The eye, like all imaging systems, is limited in the fine detail it can detect. This effect is known as *resolution*—the finer the detail discernible by an imaging system, the greater its resolution. While the eye is impressive in the detail it can resolve, some things are just too small for the eye to image effectively. We use microscopes, magnifying glasses, and similar instruments to see detail finer than the resolution of our eyes permits.

But what does all of this have to do with depth of field? No imaging system, be it a camera or a projector, produces a perfect image. This is due to the fundamental wave nature of light and often is due to design limitations in the imaging system. As a result of this inability to produce a perfect image, a point on an object is never imaged as a point on the film in a camera. The image is a small blur, often referred to as the *circle of confusion* or *disk of confusion*. Hence, even objects that are exactly "in focus" are not imaged perfectly. A well-designed camera, however, forms a circle of confusion that is much smaller than the eye can perceive, so human beings perceive good photographs as being sharp and in focus. Using a magnifying glass or a microscope, one can in fact see these circles of confusion in photographs that are in focus. Further, if the camera or film is not functioning properly, the circles of confusion can be seen with the unaided eye. This degradation in photographic quality is one manifestation of what is commonly referred to as graininess in photographs.

Depth of field can be explained in terms of this circle of confusion. Cameras produce a photographic image in which the circle of confusion is much smaller than the human eye can perceive for the object located where the camera is focused. For subjects progressively closer to or farther from the camera, the images on the film become more and more out of focus—that is, the corresponding circles of confusion get larger and larger. Eventually, the circle of confusion becomes so large that it is visible to the unaided human eye and the image in the photograph becomes blurry. However, there is always a range of distances in front of and behind the plane of best focus that produces a circle of confusion too small for the eye to perceive. The resulting photograph therefore appears to be in focus over this range of distances, or depth of field. Optical engineers or designers, then, define the depth of field as the range of distances from the camera over which the circles of confusion in the resulting images are smaller than the human eye can detect.

The depth of field is closely related to another quantity, the depth of focus. Although the depth of focus is not discussed in detail here, defining the term is worthwhile in order to avoid confusion. The depth of field refers to the range of distances in focus in front of the camera lens, or in *object space*. Similarly, there exists a spatial range behind the lens, in *image space*, where one can place the film and obtain an image that is "in focus." The situation is illustrated in figure 2, which shows a cross section of an imaging system viewed from the side. The horizontal line across the center of the drawing represents the center of symmetry for the circular lens system and is called the *optical axis*. The distance from the subject in best focus to the lens is called the *object distance* S, while the distance from the lens to the image of this object is S'. Typically, the film in a camera is placed at S'. The circle of confusion that can be resolved by the eye has a diameter D. Subjects at X and Y in object space are within the depth of field. In order to see this we begin by following the rays of light that originate at those points. Light from X is brought to a focus at the point X'. The light rays from X then diverge from X' so that when they reach Y' they have spread over a circle of confusion of diameter D. However, the circle of confusion due to light from X has a diameter less than D at the film location—a distance S' behind the lens. The subject at X is therefore "in focus" on the photograph and within the depth of field. An analogous description holds for light originating at Y, except this light is focused behind the film plane and the limiting disk of confusion occurs while the light is converging.

The depth of focus is the distance range along the optical axis over which light imaged at S' forms a disk of confusion of diameter less than D. It is the range around the image plane that corresponds to the depth of field range in the object space. For example, if the film is too close to the lens, the light from S would not have converged sufficiently to form a circle of confusion too small for the eye to detect. Note that in figure 2, X' and Y' do not define the depth of focus but are located within it.

Before discussing depth of field in detail, two final quantities, the focal length and the aperture, must be defined. The *focal length* of a lens is the image distance S' for an object infinitely far away from the lens. The point at which the light is focused is called the *focal point*. A shorter focal length means that the lens brings light to a focus closer to the lens—that is, it is a more powerful lens. The requirement that S be infinitely large is met when incoming rays of light are parallel to the optical axis. In practice, optical

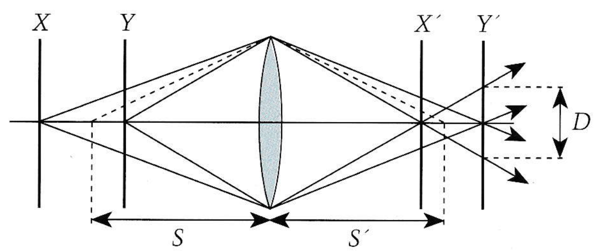


Figure 2

Cross section through the center of symmetry of an imaging system, viewed from the side. Objects at X, Y, and S are imaged to X′, Y′, and S′, respectively.

designers and engineers often assume "infinity" is 10 to 100 times the focal length. The *aperture* of a lens is determined by the area of the opening through that lens. A larger lens opening allows more light to pass through the lens. The aperture is usually described in terms of numerical aperture or a closely related quantity, the f-number. The aperture of a camera lens is usually given in terms of the f-number and photographers often refer to this quantity as the *f-stop*. For light coming from infinity, the f-number is given by the focal length divided by the diameter of the lens opening. Hence, the smaller the f-number, the larger the lens opening for a given lens. For light originating closer to the lens than infinity, the f-number is typically slightly larger for a given lens, allowing less light through the system. However, the effect is small enough that it is often ignored.

A photographer's perspective

Experienced photographers know that depth of field depends primarily on three aspects of an imaging situation: the focal length of the lens, the f-number being used, and the distance from the subject to the camera (denoted as *S* in figure 2). While many parts of an imaging system can play a role in determining the depth of field, these three are all that need be considered in most situations.

The depth of field depends inversely on the focal length of the camera lens. If one fixes the f-number and the subject distance, it's possible to view the effect of focal length alone on the depth of field. The two photographs in figure 3 illustrate the dependence of depth of field on focal length. In both cases, the subject distance—the distance to the card with the "M" on it—is fixed at about 5 m, and the f-number is 8 (this is usually written f/8). The "N" and "F" cards are 1 m in front of and behind the "M" card, respectively. The photograph in figure 3a was taken with a lens of focal length of 50 mm, while the focal length of the lens used for the photograph in figure 3b was 135 mm. It's obvious that the depth of field decreased when the focal length increased. The image made with the 50 mm focal length lens reflects a depth of field of about 10 m, while the 135-mm lens gives a depth of field of about 1.5 m. So, to maximize depth of field, minimize the focal length of the lens being used.

Now let's see what happens when we fix the focal length and the f-number and vary the subject distance. In figure 4a, a camera lens of focal length 55 mm is focused at a point about 0.5 m from the lens. The depth of field is limited to several centimeters. In figure 4b, the subject distance is now several meters. While the numbers on the meter stick are now too small to read, not only is the entire meter stick in focus, but the depth of field is now of the order of several meters instead of several centimeters. So, to maximize depth of field, maximize the distance to the intended subject.

Finally, we'll fix the focal length and the subject distance and vary the f-number. Both photographs in figure 5 were taken using a lens of focal length of 50 mm and a subject distance of about 2.5 m (again the distance to the card with the "M" on it). The photograph in figure 5a was taken using an f-number of f/2.8, while the photograph in figure 5b resulted from an f-number of f/22. Remembering that the larger the f-number, the smaller the aperture, it's obvious that the smaller lens opening leads to a greater depth of field. The f/22 photograph has a depth of field of about 4 m, while the f/2.8 photograph has a depth of field of less than 1 m. So, to maximize the depth of field, maximize the f-number, or, equivalently, minimize the aperture size.

An optical engineer's perspective

Having determined what we must manipulate in order to control the depth of field, it would be advantageous to understand why these aspects of the imaging system affect the depth of field as they do.

a

b

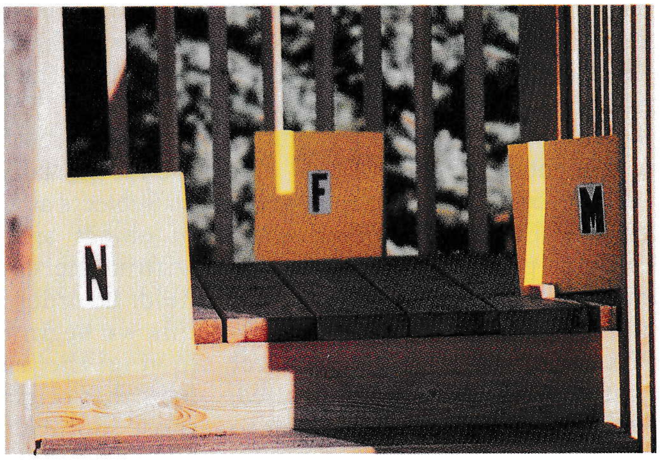


Figure 3

Depth of field as a function of lens focal length. The photographs were taken with lenses of focal length of (a) 50 mm and (b) 135 mm.

a b

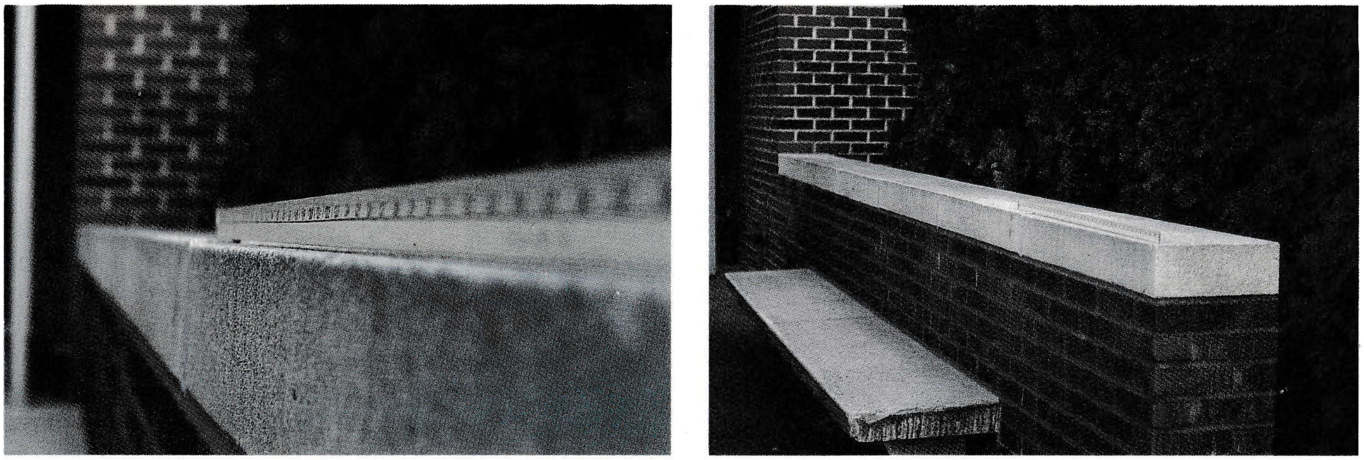


Figure 4

Depth of field as a function of subject distance. In image (a) the lens is focused at a point about 0.5 m from the lens. In image (b), the lens is focused at about 4 meters.

The optical engineer sees the effect of the lens' focal length on the depth of field as a consequence of a quantity known as the longitudinal magnification. An imaging system causes two types of magnification in the image it produces. The most familiar is the *lateral* (or *transverse*) *magnification*. In fact, it's so common that if someone refers to the magnification of an imaging system it's almost certain that the lateral magnification is being discussed. The lateral magnification is the ratio of the image height to the object height and is greater than 1 in magnitude if the image is larger than the object. This is illustrated in figure 6 (on the next page), where h is the object height and h' is the image height. The lateral magnification is then $M = h'/h$. Using similar triangles in figure 6, we can also see that the magnitude of M is S'/S.

The longitudinal magnification is closely related to the lateral magnification. *Longitudinal magnification* relates distances along the optical axis in object space to distances along the optical axis in image space. For example, light originating at a distance S in front of the lens in figure 6 is imaged at a distance S' behind the lens. Assume that the object is moved a small distance L closer to the lens. Then the longitudinal magnification M_L provides the relationship between L and the shift in the location of the image L'—that is, $M_L = L'/L$. Using a little calculus and algebra, it can be shown that $M_L = M^2$, or

$$M_L = \frac{S'^2}{S^2}. \tag{1}$$

The details of this derivation are straightforward but lengthy and can be found in introductory optics texts. It's important to remember that the magnitude of the longitudinal magnification is equal to the square of the lateral magnification.

a b

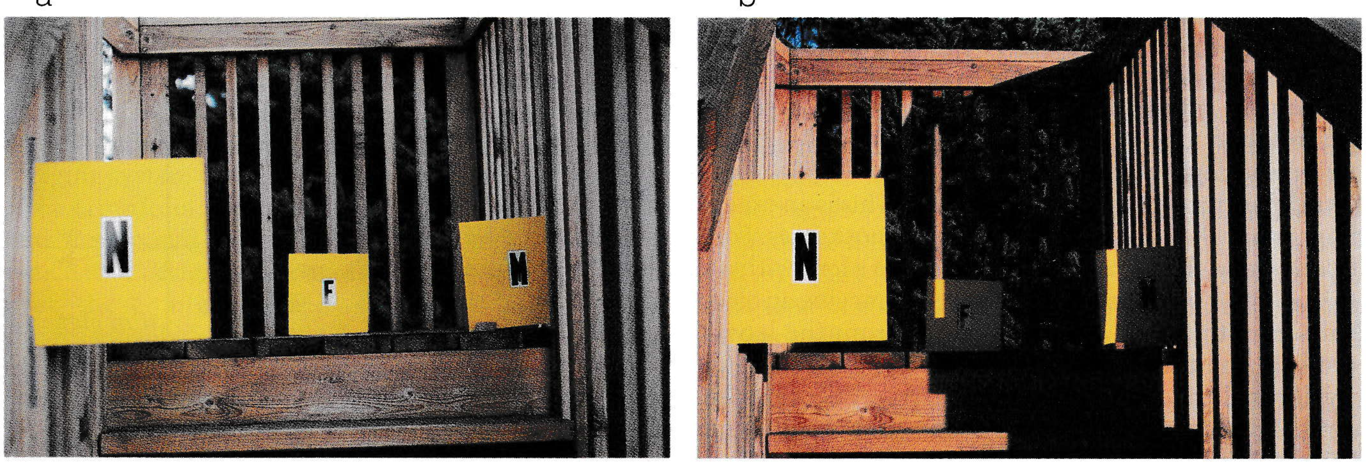


Figure 5

Depth of field as a function of aperture size (f-number). The f-numbers used in the two photographs were (a) f/2.8 and (b) f/22. (Remember, the larger the f-number, the smaller the aperture.)

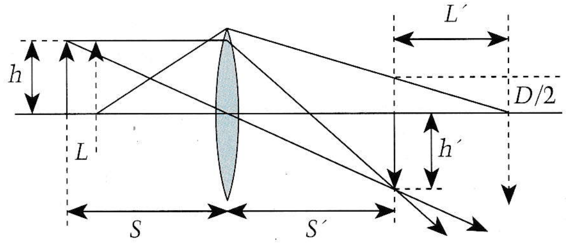


Figure 6

Cross section through the center of symmetry of an imaging system, viewed from the side. The lens forms an image of height h′ from an object of height h. Shifting the object a distance L along the optical axis causes the image to shift a distance L′ along the optical axis.

We can see the relationship between the depth of field and the longitudinal magnification in figure 6. As before, we assume that the film is at the plane of best focus, a distance S' behind the lens. Light originating a distance L closer to the lens than the original source will be imaged a distance L' behind the film. The light converging on the shifted image point will form a circle of confusion on the film plane. The greater the value of L', the larger the corresponding circle of confusion on the film and the more easily the human eye will be able to resolve this blur. Since a large value of the longitudinal magnification implies a large shift of the image location L' for a small shift in the object location L, it follows that a large value of M_L leads to a small depth of field. That is, even a small shift in the object location will cause a circle of confusion on the film that will be resolvable by the eye.

But where does the focal length of the lens fit into all of this? Having established the relationship between the depth of field and the longitudinal magnification, we must now relate the longitudinal magnification and the focal length. To do this, we use one of the simplest results from the study of imaging systems. In the early stages of the design or analysis of an optical system, optical designers and engineers often assume that an imaging system is perfect. This allows them to study some of the important characteristics of the optical system without being overwhelmed by the minute details that they will deal with later in their work. Two of the quantities that can be obtained in this simple analysis are the lateral and longitudinal magnifications. Within the "perfect imaging" assumption a simple relationship of the form

$$\frac{1}{S}+\frac{1}{S'}=\frac{1}{f} \tag{2}$$

holds true, where f is the focal length of the lens. The derivation of this equation, sometimes called the thin lens equation, can be found in most introductory physics and optics texts. Rearranging this equation we find

$$\frac{S'}{S}=\frac{f}{S-f}, \tag{3}$$

which gives the magnification in terms of the focal length and the object distance. Using equation (1), it follows that

$$M_L=\frac{S'^2}{S^2}=\frac{f^2}{(S-f)^2}. \tag{4}$$

This equation explicitly relates the longitudinal magnification to the focal length of the lens. We can make the relationship even simpler by recognizing that in most situations $S \gg f$ and ignoring f in the difference term to obtain

$$M_L=\frac{f^2}{S^2}. \tag{5}$$

This is an excellent approximation in most cases. For example, for the case of a lens with a focal length $f = 50$ mm imaging an object a distance $S = 2.5$ m from the lens, the error due to the approximation is only about 2%.

The relationship between the longitudinal magnification and the focal length is now clear. As one increases the focal length of the lens, the longitudinal magnification also increases. But we have seen that as the longitudinal magnification increases, the depth of field decreases. We have found the basis for what the photographer observes. In order to maximize the depth of field, one must minimize the longitudinal magnification, and this can be done by minimizing the focal length.

We can also understand the dependence of the depth of field on object distance S in terms of the longitudinal magnification. From equation (5), we see that as S increases, the longitudinal magnification decreases. This is consistent with the observations of a photographer. In order to maximize the depth of field, one must minimize the longitudinal magnification, and this can be done by maximizing the object distance.

In order to understand the effect of the f-number—that is, the size of the camera aperture—on the depth of field, we can't use equations (1) through (5), because equation (2) deals only with the locations of components in the imaging system along the optical axis. Characteristics of the imaging system that are off the optical axis, such as aperture size, are not described within the assumptions that lead to equation (2). However, the relationship between aperture size and depth of field is easily described in terms of the circle of

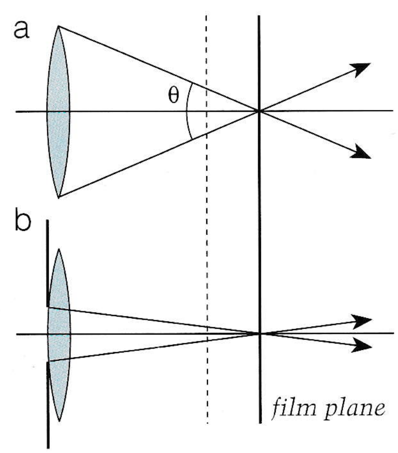


Figure 7

Depth of field as a function of aperture size (revisited). The angle subtended by the cone of light converging to form an image is controlled by the aperture size.

confusion. In figure 7a, light from across the entire lens is brought to a focus on the film. In three dimensions, this means that a cone of light that subtends a large angle θ converges on the film plane. This imaging situation provides abundant light to the film by collecting light over a large cone and focusing it to a single point on the film. However, this situation also leads to a cone of light that rapidly diverges as one moves in front of or behind the plane of best focus. The diameter of the circle of confusion increases rapidly along the optical axis. Light from subjects only slightly closer to, or farther from, the lens is focused at such a large angle that the circle of confusion is resolvable by the eye for subjects near the object that is in best focus. The depth of field can be increased by decreasing the size of the cone over which light is collected. This is done by decreasing the size of the aperture while leaving all other variables fixed. Figure 7b shows that for a small aperture the size of the circle of confusion increases very slowly along the optical axis. This leads to a large depth of field. In order to maximize the depth of field, one must minimize the rate at which the circle of confusion increases in size. This is done by minimizing the aperture size (maximizing the f-number), just as a photographer would suggest.

It should be noted that while focal length, object distance, and f-number are the most important quantities in determining the depth of field, other characteristics of an imaging system can also contribute. An example is the quality of the lens design and the accuracy with which that design is executed. However, these "higher order" effects can reasonably be ignored when discussing consumer cameras and similar systems.

Using depth of field

When one understands both what affects depth of field and why it does, it is possible to use depth of field to create intriguing photographs. The ability to control the depth of field leads to an ability to produce a variety of special effects. I have had some success in using depth of field to create interesting images. Figure 8 is an example of using a very shallow depth of field in order to emphasize an object that occupies little of the photograph. All of the grass and weeds in the background blend together and the clearly focused flower seems to jump out of the image. Figure 9 is a photograph of a stack of 3-inch plastic pipe taken from the end of the stack. During shipping, two of the pipes had shifted so that they were extending beyond the plane of most of the pipe ends. One pipe was about 20 cm beyond the other pieces and another extended about 60 cm. The two pipes appear increasingly blurry as their ends extend farther from the plane of best focus. Let's refer back to figure 1. It builds on the effects seen in figure 9, with a second stack of pipe visible. I took the photograph in figure 1 while standing about 0.5 m from the ends of a stack of pipe. The camera was focused on the pipe ends visible through the nearer pipe. The clearly focused pipes are about 5 m from the camera so that the nearer pipe ends are horribly out of focus.

Figure 8

Shallow depth of field. As a result, only the flower appears in focus.

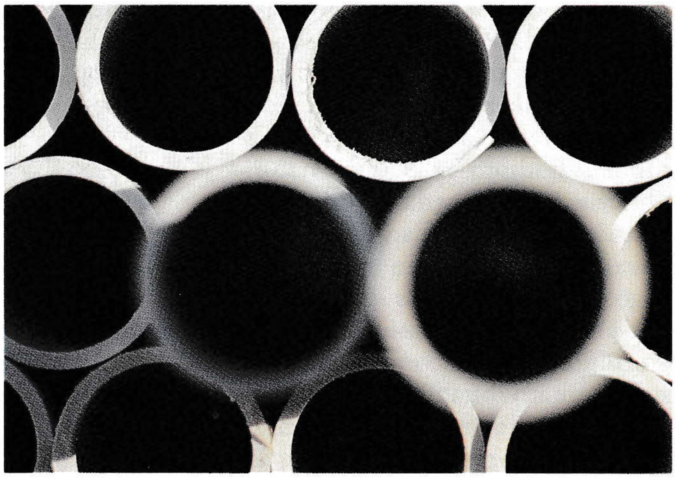


Figure 9

Objects lying off the plane of best focus. The two pipes that extend beyond the others are out of focus.

It's fascinating that the complex phenomenon of depth of field is described in relatively simple terms using basic theory and mathematics. With an understanding of depth of field, it's possible to view photographs and other images with a critical eye, aware that special effects aren't limited to the movies.

Mark L. Biermann *has recently taken a position as assistant professor of physics at Buena Vista University in Storm Lake, Iowa. At the time of writing this article, he was an assistant professor of physics at Whitworth College in Spokane, Washington.*

All bent out

"The force of any spring is proportional t

OF COURSE, ROBERT HOOKE was talking about one of the most obvious kinds of deformation—elastic deformation. But from our earliest school days we've heard about other kinds of deformation that play very important roles. Think back on those problems of elastic and inelastic collisions, bodies in equilibrium, changes in shape and volume, mechanical oscillations . . . The list goes on and on, and an understanding of deformation as it relates to the properties of physical bodies is so important for science and technology that we can say without exaggeration that the study of these properties is one of the most important tasks of modern physics.

When we study physics in high school we find that we need to understand deformation in many different areas. In mechanics and thermodynamics it's taught from a macroscopic viewpoint; when the discussion turns to explaining atomic and molecular interactions, it's presented on the basis of molecular kinetics; and when we poke into the nature of elastic forces, we find ourselves using electromagnetic theory. The occurrences of this notion in the world around us are indeed kaleidoscopic.

Questions and problems

1. When is a rope stressed more—when you pull it by its ends in different directions or if you pull it with both hands after attaching it to a wall? In both cases each hand develops the same amount of force.

2. A heavy cylinder falls freely. What forces act on each horizontal layer from the adjacent ones?

3. Two forces $\mathbf{F}_1$ and $\mathbf{F}_2$ are brought to bear on a board resting on supports (fig. 1). Will the board sag differently if the two forces are replaced by a single force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$?

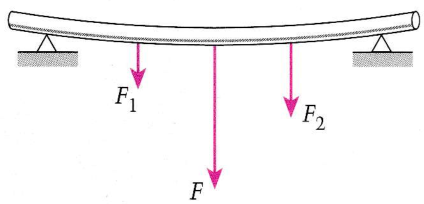


Figure 1

4. Iron and copper wires of the same size are suspended vertically, their bottom ends connected by a massless horizontal bar. Will the bar remain horizontal if a mass is attached to its midpoint?

5. How do the relative elongations of two wires with equal masses attached and made of the same material differ if the length and diameter of the first are twice those of the second? What about their absolute elongations?

6. In the manufacture of wire a metal rod is pulled through a set of openings that continually decrease in diameter. What kinds of deformation take place during this process?

7. Why do spring dynamometers have limiters that restrict the elongation of the springs?

8. Why do fishing rods have long, flexible ends?

9. Why does a bullet merely punch two small holes in a soft plastic cup of water, but smashes a glass cup to bits?

10. The ends of two weightless spiral springs of different lengths are connected as shown in figure 2. If one were to graph the stretching force F versus the displacement x of the point where the force is applied, what shape would it take?

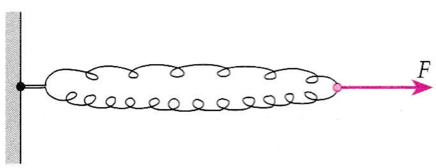


Figure 2

11. How does the period of vertical oscillation change for a load suspended from two identical springs if a connection in series is replaced by a connection in parallel?

12. A body A is suspended by a thread. Another body B is attached by a spring to A. Then the thread is burned. Will the accelerations of the falling bodies be the same?

13. Why can a quartz rod undergo drastic cooling without shattering?

14. It takes a lot of effort to tear off a piece of wire, but a red-hot wire can be broken much more easily. Why?

15. How does the stress in a rod change if you heat it without allowing it to expand?

16. How does a body's energy change during elastic deformation?

17. A compressed steel spring has a potential energy. What happens to this energy when the spring is dissolved in an acid?

18. Which of two massless springs with the same dimensions,

ut of shape

al to its elongation . . ."—Robert Hooke

one copper and one steel, acquires more potential energy under identical loads?

Microexperiment

Take a rubber tube and slide a metal ring onto it such that the ring stays put when you hold the tube vertically. Now stretch the rubber tube.. What happens to the ring, and why?

It's interesting that . . .

. . . the scientific interests of Robert Hooke were so wide-ranging that he often had no time to bring his research to a conclusion. This gave rise to acrimonious disputes with such authorities as Newton, Huygens, and others concerning the priority of their discoveries. However, "Hooke's law" was so convincingly validated by many experiments that Hooke's priority in this matter was never doubted.

. . . at the beginning of 18th century mining accidents due to broken elevator chains became more frequent. Many scholars, including the famous Gottfried Wilhelm Leibniz, tried to improve the iron chains, but without success. It was a senior mining adviser, W. Albert (a lawyer by training), who thought of replacing the chains with wire ropes or cables. This made it possible to exploit one of the most important properties of iron—its high tensile strength.

. . . the English physicist and engineer O. Reynolds was the first to explain why wet sand lightens in color when you walk on it. In 1885 he showed that the volume occupied by the grains of sand increased due to shear deformation, causing the upper layer of sand to rise above the water level temporarily.

. . . the individual crystals of a number of metals grown from a melt are so soft they can easily be bent by your fingers. But you can't unbend them! This is an example of the wonderful property of pliable deformable bodies to harden.

. . . an explanation for plastic deformation didn't come until the 20th century, when physicists discovered "dislocations"—that is, defects in a solid body's crystal lattice. From the modern point of view, this kind of deformation is the "movement of disorder" within a crystal.

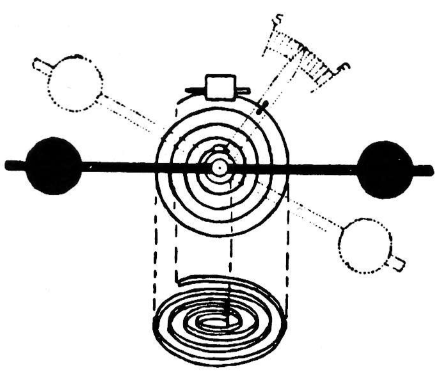
. . . nowadays superelastic alloys are available that behave like rubber and are able to endure huge elastic deformations—two orders of magnitude greater than ordinary metals. On the other hand, many kinds of alloys can be brought to a superelastic state, when they flow under very low pressure like heated glass.

. . . it's possible to combine opposite mechanical properties in "composites"—compound materials that include a light pliable base and a fiber filling made of a very strong material.

. . . one can measure deformations that are less than an atomic diameter—provided they're oscillatory and thus can easily be transformed into electric signals. By the way, the human ear can also "measure" similarly small deformation of the eardrum.

. . . the deformation of quartz and some other dielectrics results in the appearance of an electric charge on their surfaces. And the polarization of dielectrics in an electric field can produce a deformation. These phenomena are known as direct and inverse piezoelectric effects.

. . . when lead is bombarded by neutrons for a long time, it rearranges itself internally and becomes so elastic that a bell made of it might chime as resonantly as bells cast from the best bronze.



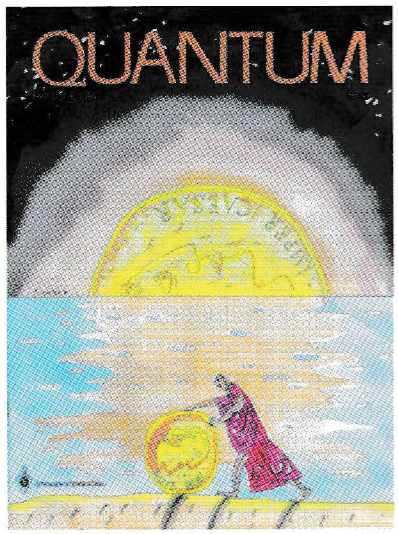
An early form of the balance spring of a watch. One historian of science called it the "most important single improvement ever applied to portable timekeepers," and it serves as an example of the disputes engendered by Hooke's working methods. Abbott Payson Usher writes: "The invention of the balance spring has been claimed by Huygens, Hautefeuille, and Hooke. But though Hooke's work may not have been known on the continent, owing to the secrecy with which he invested his experiments, there is little doubt of the priority of his work." (From A History of Mechanical Inventions, *New York: Dover, 1988)*

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Winner of

1995 Folio Editorial Excellence Award

For Outstanding Fulfillment of its Editorial Mission



Quantum received an award for editorial excellence from *Folio: The Magazine of Magazine Management* (published by Cowles Business Media). The award was based on *Quantum*'s fulfillment of its editorial mission.

Mission Statement

Quantum is a magazine of math and science for anyone who wants more than a textbook treatment of these subjects. *Quantum* articles are not written like articles in scientific journals; by *engaging* the readers (rather than dictating *to* them), they lead the reader to work out problems on the side. Some articles are elegant expositions of sophisticated concepts, and some give an unexpected twist to a well-known idea or phenomenon; others show that there is no such thing as a silly question. In addition to its feature articles, *Quantum* introduces "fun" to the sometimes mundane worlds of science and math, with departments like "Brainteasers" (fun problems requiring a minimum of math background), "Looking Back" (biographical and historical pieces), and "Gallery Q" (an exploration of links between art and science), among others.

A large part of the reader involvement comes from the beautiful illustrations that accompany the articles; the presence of high-quality art in *Quantum* is an outgrowth of the belief that a good science and math magazine should nourish the complete person; that good art will train the visual imagination, which is important in these disciplines; and that if *Quantum* art helped students become comfortable with (and even welcome) confusion and learn to "question their way out of it," such a habit of inquiry might carry over into their reading of scientific and mathematical texts.

May/June 1994 Issue

- *Exposition of sophisticated topic.* The article "Follow the Bouncing Buckyball" by Sergey Tikhodeyev (p. 8) describes and explains how the atomic structure of carbon lends itself to the formation of polyhedrons known as "buckyballs."

- *Engaging the reader.* The article "Six Challenging Dissection Tasks" by Martin Gardner (p. 26) gives readers the chance to formulate proofs using what they already know about the properties of geometry, combined with new twists from the man who for many years wrote the "Mathematical Diversions" column in *Scientific American.*

- *Fun.* The department Quantum Smiles offers "A Mathematical Handbook with No Figures," by Yuly Danilov (p. 42). After being introduced to a charming problem book filled with amusing mathematical abstractions, readers are given a sampling to try on their own.

- *High-quality art.* Throughout the issue, most articles are accompanied by sophisticated illustrations that serve to complement, represent, and transcend the text.

To order, call

1-800-SPRINGER

Challenges in physics and math

Math

M151

Successive felicity. A positive integer will be said to be felicitous if one can choose a number of digits from its decimal notation such that their sum equals the sum of all the remaining digits. (a) Find the smallest felicitous number a followed by a felicitous number. (b) Are three successive felicitous numbers possible? (N. Zilberberg)

M152

Concurrent perpendiculars. A point P is marked in a square $A_1A_2A_3A_4$ and joined to its vertices. Prove that the perpendiculars dropped from A_{i-1} on line PA_i, $i = 1, 2, 3, 4$ (of course, A_0 here should be read as A_4) all meet at the same point. (A. Vilenkin)

M153

Equispaced roots. Under what condition on the coefficients of a cubic equation $x^3 + ax^2 + bx + c = 0$ do its three roots form an arithmetic sequence? (M. Bezborodnikov)

M154

Down with twos! Written around a circle are n numbers, each equal to 0, 1, or 2. They are simultaneously put through the following transformation: all twos are replaced by zeros and then all the numbers next to them clockwise are increased by one. Initially there are $k \geq 1$ twos. (a) How many such transformations are sufficient to eliminate all twos? (b) Suppose in addition that there were no zeros initially. Prove that there will be k ones and $n - k$ zeros in the end. (N. Alexandru [Romania])

M155

Mathematics of floating timber. On a bank of a rectilinear river a number of logs lie, each of them making an angle less than 45° with the bank edge. (The logs are disjoint segments.) Prove that one of the logs can unobstructedly be rolled into the river in a direction perpendicular to itself (see figure 1) without changing this direction. (V. Ilyichev)

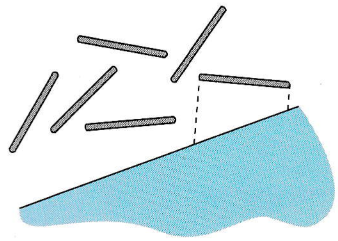


Figure 1

Physics

P151

Sheltered gun. A howitzer fires from under a deep shelter inclined at an angle α with the horizontal (fig. 2).

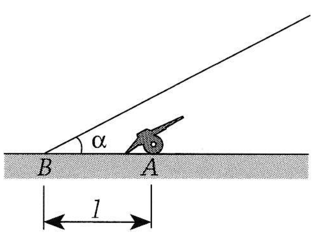


Figure 2

The gun is placed at a point A located a distance l from the base of the shelter (point B). The initial velocity of the shell is $\mathbf{v}_0$. Assuming that the shell's trajectory lies in the plane of the figure, find the maximum distance it will travel. (S. Krotov)

P152

Atomic battery. One element of an atomic battery is a capacitor with one plate covered with radioactive material that emits alpha rays with a speed of $v_0 = 2.2 \cdot 10^6$ m/s. Find the emf of this element. The ratio of the electric charge of an alpha particle to its mass is $k = 4.8 \cdot 10^7$ C/kg. (A. Grigorenko)

P153

Water under a piston. There is a small amount of water in a nipple at the bottom of a cylinder closed with a flat piston of diameter $D = 5$ cm (see figure 3). The diameter of the

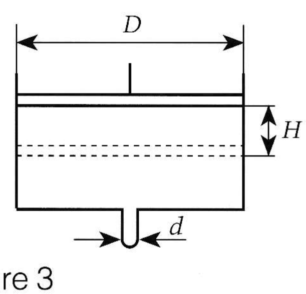


Figure 3

nipple is $d = 2$ mm. When the piston is lowered at a constant temperature by $H = 10$ cm, the water in the nipple rises by $h = 1$ mm. Find the saturated vapor pressure at a temperature of 20°C. (V. Belonuchkin)

CONTINUED ON PAGE 41

Split image

*"All these years, your left hand, modest
but sinister accompanist,
has seen itself in the mirror
grown stronger."*
—*Al Zolynas*, "Dream of the Split Man"

by Arthur Eisenkraft and Larry D. Kirkpatrick

AT SPECIAL MOMENTS WE'RE able to bring all of our attention to bear on a single item. We focus our minds, our gaze, and our efforts. Similarly, wouldn't it be interesting to take the light impinging on a surface and focus it to a point? Using the properties of transparent materials and our knowledge of Snell's law and geometry, we can construct an object for this purpose—a lens. A convex lens bends all rays of light parallel to the principal axis (the axis of symmetry of the lens) in such a way that they converge at a single point referred to as the focus (fig. 1).

The lens also takes the light emerging from one point and focuses that light to a point on the other side of the lens. This works whether the light source is on the principal axis (fig. 2) or off axis (fig. 3). This then provides the surprising and technologically vital property of image formation in lenses. All slide projectors, cameras, copy machines, microscopes, and binoculars are dependent on a lens being able to produce images.

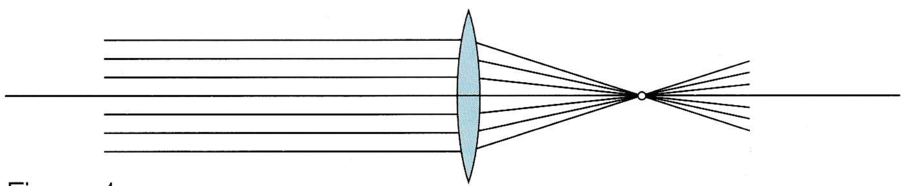


Figure 1

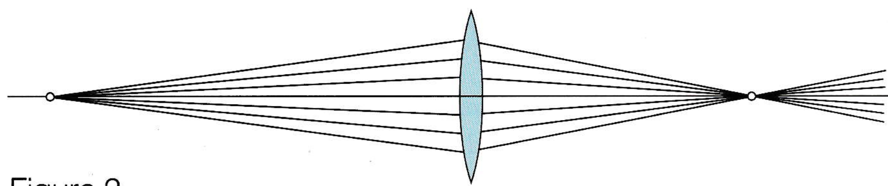


Figure 2

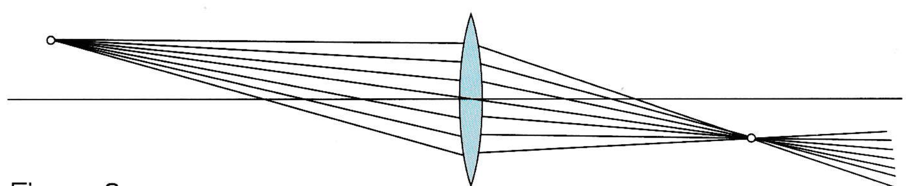


Figure 3

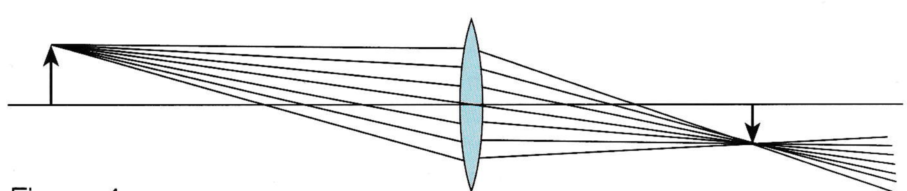


Figure 4

In forming a real image, all light leaving points of the object and passing through the lens come together on the far side of the lens (fig. 4). Snell's law ($n_1 \sin \theta_1 = n_2 \sin \theta_2$) can be applied to each ray to determine its direction upon leaving the lens. The intersection of any two of the rays determines the location of the image. However, it's much easier to use "special rays" that are easy to draw. In such a ray diagram, a ray of light leaving the "top" of the object parallel to the principal axis must pass through the focal point. A ray through the center of the lens is undeflected. This ray is, in essence, traveling through two parallel sides of a piece of glass. The intersection of these two rays gives the location

Art by Tomas Bunk

SPLIT IS A BEAUTIFUL OLD TOWN ON THE ADRIATIC COAST IN DALMATIA, WHERE THE ARTIST WAS BORN.

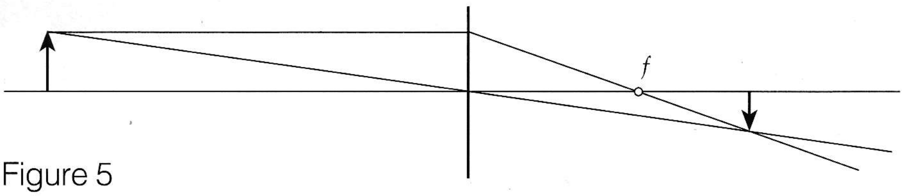


Figure 5

of the "top" of the image. Ray diagrams are quite an asset in determining the image's location and size (fig. 5).

Beware of the misconceptions that befall those students who begin to believe that this helpful ray diagram describes how the light actually behaves. The ray diagram is a map—it shouldn't be mistaken for the landscape. Those who place all their trust in a ray diagram may begin to believe that if part of the lens is removed, part of the image is also removed. Do our *Quantum* readers understand how the image changes if the top half of the lens is covered?

Those who focus only on the ray diagram may forget that the light diverges after passing through the image location, and may never discover other interesting optical phenomena. The tool of ray diagrams should not limit your understanding nor your curiosity.

This month's contest problem provides you with a broken lens and asks for a description of the resulting light pattern. It was first given at the 1972 International Physics Olympiad in Bucharest, Romania.

A. A lens of focal length f is cut into two parts perpendicular to its plane. The half-lenses are moved apart by a small distance δ (fig. 6—the gap is exaggerated due to typographic considerations). How many interference fringes appear on a screen at a distance L from the lens if a monochromatic light source (wavelength λ) is placed at a distance d ($d > f$) on the other side?

B. If f = 10 cm, d = 20 cm, δ = 0.1 cm, λ = 500 nm, and L = 50 cm, calculate the number of fringes.

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 and their authors will receive special certificates from *Quantum*.

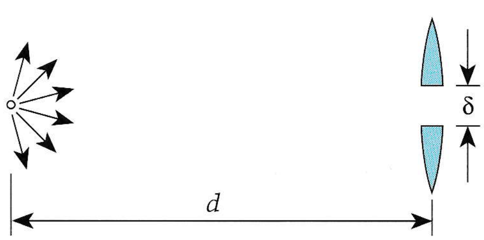


Figure 6

Weighing an astronaut

The problem in the March/April issue asked you to "weigh" an astronaut given calibration data from the Body Mass Measuring Device (BMMD). This problem was given as a class assignment by Art Hovey at Amity Regional High School in Woodbridge, Connecticut, and by your author (LDK) to sophomores at Montana State University. The best solutions at ARHS were submitted by Kurt Rohloff and Lori Sonderegger, and the best one at MSU was written by Dave Peters.

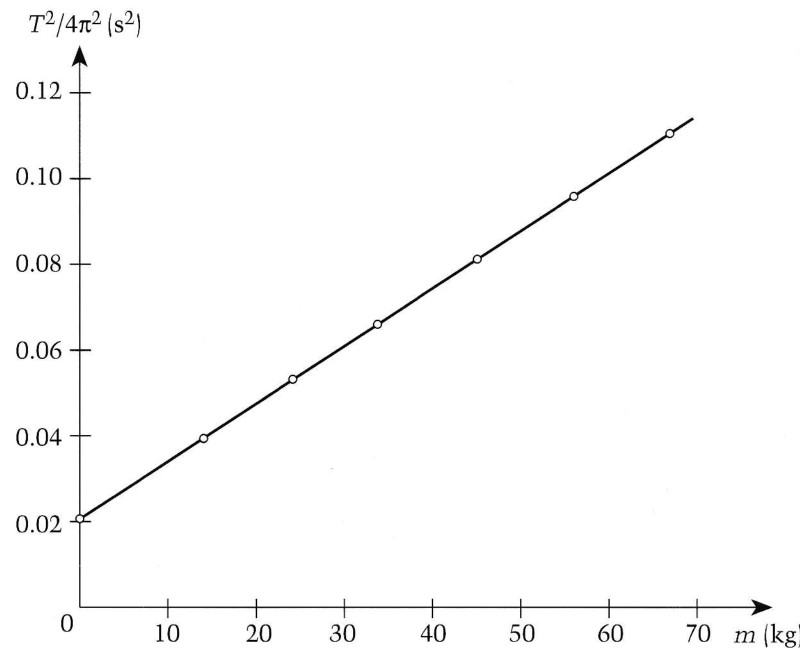


Figure 7

A. The equation for the period T of the simple harmonic motion of a mass m on a spring is given by

$$T = 2\pi\sqrt{\frac{m}{k}},$$

where k is the spring constant. In our case, the total mass m is the sum of the mass of the chair m_c and the mass of the astronaut m_a. Squaring both sides and solving for T, we obtain

$$\frac{T^2}{4\pi^2} = \frac{1}{k}m_a + \frac{m_c}{k}.$$

This is of the form $y = ax + b$. Therefore, if we plot a graph of $T^2/4\pi^2$ versus the mass used in the calibration of the BMMD, we should obtain a straight line with a slope of $1/k$ and a y-intercept of m_c/k (fig. 7). This allows us to obtain the numerical values k = 748 N/m and m_c = 15.4 kg.

B. We can now use these values and the numerical data for astronaut Garriott to discover that he lost 2.3 kg of mass during 58 days in space.

C. The third part of this problem asked you to calculate the reading on the scale as a person rides down

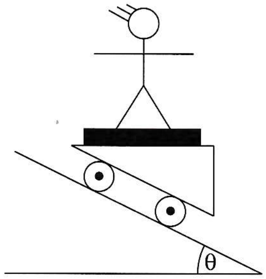


Figure 8

an incline as shown in figure 8 (which was figure 1 in the March/April contest problem). We begin by calculating the acceleration of the entire system (skateboard, scale, and person) down the incline. Because the component of the force of gravity down the incline is $m_s g \sin\theta$, the acceleration of the system down the incline is $g \sin\theta$. This is also the acceleration of the person, which allows us to use Newton's second law to find the forces acting on the person. Assuming that the scale reads the force applied normal to its surface, let's look at the vertical forces acting on the person. The force of gravity mg acts downward and the force of the scale F_s acts upward. Therefore, the difference in these two forces must yield the mass times the vertical component of the acceleration:

$$mg - F_s = (mg \sin\theta) \sin\theta,$$

or

$$\begin{aligned} F_s &= mg(1 - \sin^2\theta) \\ &= mg\cos^2\theta \\ &= (588\ \text{N})\cos^2\theta. \end{aligned}$$

This is the standard solution given by the top students on the preliminary exam to select the 1995 US Physics Team that competes in the International Physics Olympiad and the one expected in textbooks. However, there is a problem—the person will have a very difficult time standing vertically on the scale. This difficulty was pointed out by Dr. Albert A. Bartlett, professor emeritus at the University of Colorado and a past president of the American Association of Physics Teachers. Please refer to his follow-up article on page 49 for a more complete explanation.

Call for manuscripts

Nineteen ninety-six marks the 25th anniversary of The Club of Rome's study *The Limits to Growth.* To provide its young readers with both information and current perspectives on this study, *Quantum* invites the submission of papers for a special issue on *The Limits to Growth* and its 1992 sequel, *Beyond the Limits.*

Several authors have already expressed an interest in writing for such a special issue. Victor Gorshkov of the St. Petersburg Nuclear Physics Institute will prepare a paper that presents ideas from his recent book *Physical and Biological Bases of Life Stability* (Springer, 1995). Kurt Kreith will show how a spreadsheet investigation from "Look, Ma—No Calculus!" (*Quantum*, November/December 1994) illustrates "the four generic ways in which a population can approach its carrying capacity."

We seek additional papers that analyze this study and its implications from a variety of points of view. Such papers might address the changes (and growth!) that have occurred since the publication of *The Limits to Growth.* They might also address the advances in computer technology that make such models ("state of the art" in 1970) accessible via desktop computers available at most American high schools and in many secondary schools around the world. Or they might review both the study and its critics, shedding light on the ways in which science and public opinion interact in the search for solutions to the environmental challenges confronting the current generation of students.

Prospective authors are invited to send a query to

Managing Editor
Quantum
1840 Wilson Boulevard
Arlington VA 22201-3000

Fax: 703 522-6091

E-mail: quantum@nsta.org

MATH INVESTIGATIONS

Problems beget problems

"We shall have to evolve
Problem solvers galore—
Since each problem they solve
Creates ten problems more."
—Piet Hein

by George Berzsenyi

As I START PREPARING THESE columns, time and again I plan to comment on the reactions of my readers to previous columns, but time and again I fail to do so. I hereby apologize for these shortcomings, which are partially due to my desire to limit the columns to one page and thereby simplify my life and that of the managing editor. Clearly, such concerns are of little importance—hence the present departure from practice. In like manner, I will also start my future columns by first reporting on the

IN THE MARCH/APRIL 1995 ISSUE of *Quantum*, George Berzsenyi recalled a puzzle he first posed in 1980. He considered sets of n positive integers whose pairwise sums never match. That is, $p_i + p_j \neq p_k + p_l$ for distinct i, j, k, and l. Powers of 2 form such sets: $\{2^0, 2^1, \ldots, 2^{n-1}\}$. So do $\{f_2, \ldots, f_{n+1}\}$, where f_n are the Fibonacci numbers: $f_1 = f_2 = 1$ and $f_i = f_{i-1} + f_{i-2}$ for $i > 2$. A unique and "minimal" choice among all such n-element sets is obtained as follows. Choose those with the minimum value of p_n. From this selection, choose those with minimum p_{n-1}. Continue this process. For $n < 7$, but not otherwise, the Fibonacci sequences are minimal.

Let $P(n)$ be the value of p_n for the minimal n-element set. Berzsenyi examined the case of $n = 12$, for which he proved $57 \leq P(12) \leq 74$. His lower bound was obtained analytically; his upper bound from a student's example. Readers were challenged to "narrow the gap." For starters, our computer found the following minimal sets for $7 \leq n \leq 14$:

$n = 7$: $\{1, 2, 3, 5, 9, 14, 19\}$
$n = 8$: $\{1, 2, 3, 5, 9, 15, 20, 25\}$
$n = 9$: $\{1, 2, 3, 5, 9, 16, 25, 30, 35\}$
$n = 10$: $\{1, 2, 8, 11, 14, 22, 27, 42, 44, 46\}$
$n = 11$: $\{1, 2, 6, 10, 18, 32, 34, 45, 52, 55, 58\}$
$n = 12$: $\{1, 2, 3, 8, 13, 23, 38, 41, 55, 64, 68, 72\}$
$n = 13$: $\{1, 2, 12, 18, 22, 35, 43, 58, 61, 73, 80, 85, 87\}$
$n = 14$: $\{1, 2, 7, 15, 28, 45, 55, 67, 70, 86, 95, 102, 104, 106\}$

Thus, $P(12) = 72$, and we have met Berzsenyi's challenge!

Here's a related puzzle. As before, let $\{q_i\}$ be a set of n "matchless" positive integers. Suppose they satisfy the additional constraint that no q is the average of any two others. Denote the value of q_n for such a minimal n-element set by $Q(n)$. Clearly, $Q(n) \geq P(n)$. Some minimal sets of this kind are given below:

$n = 3$: $\{1, 2, 4\}$

comments received on previous columns.

Several readers found a couple of mistakes in one of the examples given in my "Distinct Sums of Twosomes" (March/April 1995). One of these was pointed out by an anonymous reader from Arizona on a beautiful postcard showing the Saguaro National Monument. Computer solutions for $n \le 12$ were provided by Curtis Cooper and his best student, Chris Campbell (Central Missouri State University in Warrensburg), James Cortese (Champaign, Illinois), Kang Su Gatlin and Ted Carlson (students at the University of San Diego, Department of Computer Science), David Reynolds (Beaverton , Oregon), Geary Younce (Fairfax, Virginia), and my most frequent correspondent, Brian Platt (Woods Cross, Utah). In particular, they all proved that for $n = 12$, the best possible value for a_{12} is 72, and that there are four such sequences. Most of them also sent me a copy of the computer programs that generated their results. Their responses were most appreciated.

We were also happy to receive a letter from Sheldon Glashow and Eric Carlson (Harvard University) addressing this problem and posing a related one, which I hereby offer as our new challenge. Their letter is reproduced below. Since Glashow (a Nobel laureate in physics) is one of the founding editors of *Quantum*, and since Carlson (also a physicist) was one of the most outstanding and enthusiastic participants in the competition I conducted through *The Mathematics Student*'s "Competition Corner" in 1978–1981, I was particularly pleased with their interest in this problem.

"HOW DO YOU FIGURE?" CONTINUED FROM PAGE 35

P154

Cylinders, charges, and magnetic fields. Two long, thin-walled, nonconducting cylinders can rotate freely about a common axis as shown in figure 4. The radius of the large cylinder is twice that of the small one. The cylinders have equal surface charge densities. The outer cylinder is set in motion with an angular velocity ω. In what direction and with what velocity will the other cylinder rotate? (V. Mikhaylov)

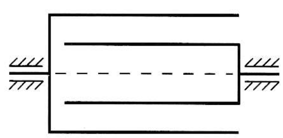


Figure 4

P155

Such a strange lens. A cylinder of radius $R = 5$ cm is composed of two identical semicylinders made of glass with a refractive index $n = 2$. The semicylinders touch along their plane surfaces. Without separating these surfaces from one another, one semicylinder is rotated such that the angle between the axes of the semicylinders is 90°. A pencil beam of light is directed onto the convex surface of one of the semicylinders so that it is perpendicular to the plane surfaces and passes through the intersection of the two axes. What does the emerging beam look like? By what factor will its cross-sectional area increase at a distance $L = 1$ m from the optical system? (A. Zilberman)

ANSWERS, HINTS & SOLUTIONS ON PAGE 56

$n = 4$: $\{1, 2, 5, 7\}$
$n = 5$: $\{1, 2, 5, 10, 12\}$
$n = 6$: $\{1, 2, 5, 11, 13, 18\}$

This problem, like its predecessor, can be attacked by computer, but that's not much fun. Let's turn to more general considerations.

The reader should be able to prove that $P(n + 1) > P(n) + 1$ for all $n > 2$. More interestingly, Berzsenyi's arguments lead to quadratic lower bounds on both $P(n)$ and $Q(n)$:

$$P(n) > \frac{n^2 - 3n + 4}{2}, \quad Q(n) > \frac{n^2 - n}{2}.$$

What about upper bounds on $P(n)$ and $Q(n)$?

Surprisingly, the upper bounds are quadratic in n as well. The proof is somewhat technical. Let p be the smallest prime number greater than n. Number theorists have shown that there is an integer x such that the smallest solution to $x^m \equiv 1 \pmod p$ is $m = p - 1$. (For example, if $p = 7$, we have $3^6 \equiv 1 \pmod 7$). A set of n positive integers may be defined in terms of x and p:

$$q_i \equiv [pi + (p - 1)x^i] \pmod{p(p - 1)}$$

for $i = 1, \ldots, n$. We have shown that the set $\{q_i\}$ satisfies the criterion of the second problem. That is, if $q_i + q_j = q_k + q_l$, then $i = k$ and $j = l$ (or $i = l$ and $j = k$). Furthermore, $1 < q_i < p(p - 1)$. In particular, we find that $Q(n) < n(n + 1)$ for n one less than a prime. There is always at least one prime between n and $2n$, and large primes are more closely spaced. Consequently, our result yields a quadratic upper bound on $Q(n)$ for all n. Roughly speaking, and for large n, $Q(n)$ and $P(n)$ must lie between $n^2/2$ and n^2. Refinements of these bounds would be welcome.

—Eric D. Carlson,
Sheldon L. Glashow
(Harvard University)

A tell-tale trail and a chemical clock

Two simple experiments with alternating current

by N. Paravyan

EVERYBODY KNOWS THAT alternating current (AC) periodically changes its direction and amplitude. Do you want to see with your own eyes that it indeed "changes direction"? It's not only possible, it's actually quite easy to do.

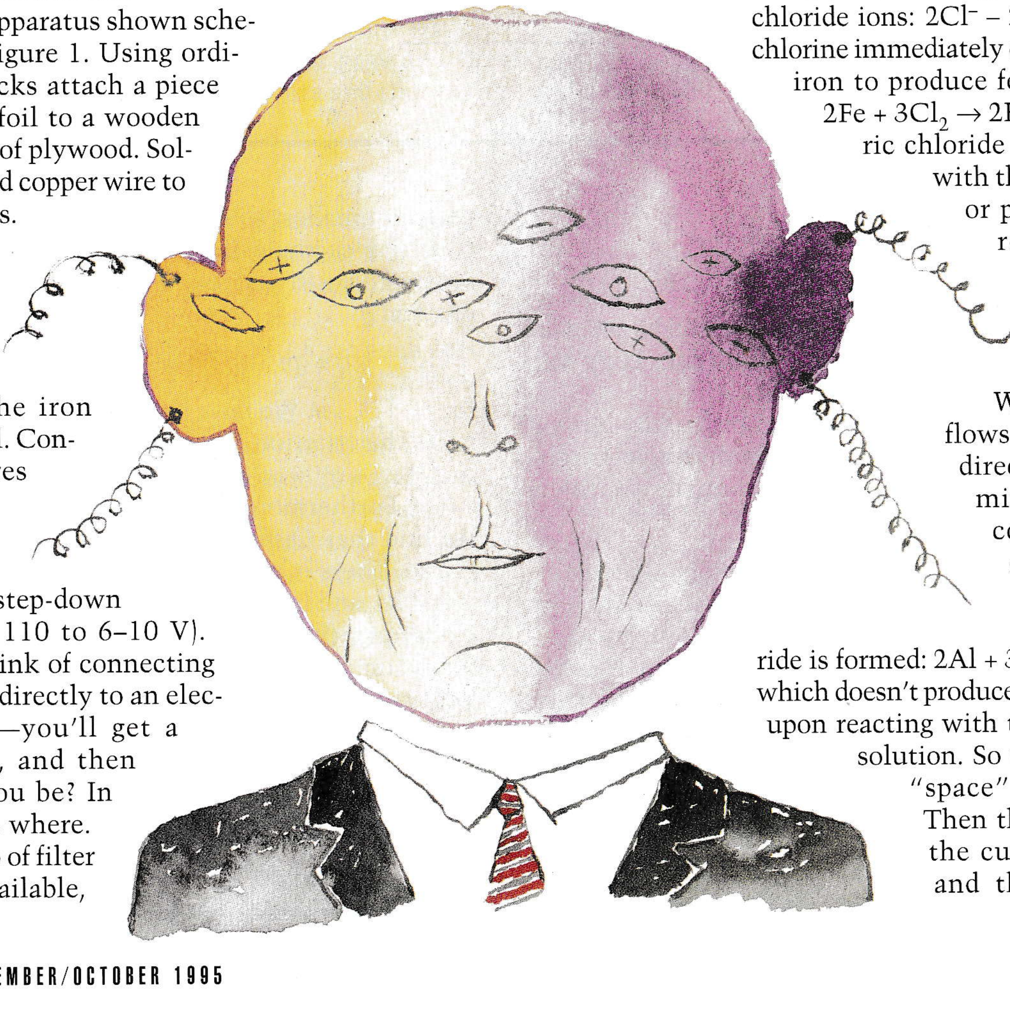
Seeing the current alternate

Set up the apparatus shown schematically in figure 1. Using ordinary thumbtacks attach a piece of aluminum foil to a wooden board or piece of plywood. Solder an insulated copper wire to one of the tacks. Using plastic or wooden clothespins attach another similar wire to the iron spike of an awl. Connect both wires to the terminals of the secondary winding of a step-down transformer (110 to 6–10 V). Don't even think of connecting the apparatus directly to an electrical outlet—you'll get a short circuit, and then where will you be? In trouble, that's where.

Take a strip of filter paper (if unavailable, use a strip of newspaper), moisten it with a 10% solution of ammonium chloride containing also 0.3 g of thiocyanate or potassium ferrocyanide, and place this paper on the foil. Now switch on the alternating current (through the transformer, please!) and quickly run the awl along the strip of paper (don't press too hard). A broken colored line will appear on the paper: magenta for thiocyanate or blue for potassium ferrocyanide. Why is it broken?

The solution used to moisten the paper contains ions—mostly NH_4^+ and Cl^-. When the current is turned on, the awl becomes an anode for one semiperiod, resulting in a discharge of chloride ions: $2Cl^- - 2e \rightarrow Cl_2$. The chlorine immediately combines with iron to produce ferric chloride: $2Fe + 3Cl_2 \rightarrow 2FeCl_3$. The ferric chloride in turn reacts with the thiocyanate or potassium ferrocyanide to produce the magenta or blue substance.

When current flows in the reverse direction, the aluminum foil becomes the anode, where colorless aluminum chloride is formed: $2Al + 3Cl_2 \rightarrow 2AlCl_3$, which doesn't produce colored agents upon reacting with the ions in the solution. So there will be a "space" in our line. Then the direction of the current changes and the iron spike



Art by Leonid Tishkov

becomes an anode again, resulting in a colored segment on the paper, and so on. These are the processes that "draw" the broken line with regular alternating dashes and spaces.

An electrolytic timepiece

There are many and various "watches" made by nature itself—geological, biological, and chemical. One can also make an electrolytic watch. Here's how.

Take a small jar—a rectilinear flat cuvette will be most suitable (fig. 2). Pour a saturated solution of table salt into the cuvette up to 2/3 its height. Cut two electrodes from a steel can (don't use an aluminum can!) so that they're almost the same height as the cuvette. Solder insulated copper wires to each electrode. Tape each electrode to a wooden rod and lower the electrodes into the solution. Connect one electrode to the terminal of a step-down transformer. Connect the other wire to the terminal of an AC ammeter rated at 5 A, connect its second terminal to the free terminal of the transformer.

Now our apparatus consists of a source of alternating current, a cuvette with sodium chloride electrolyte, two immersed electrodes, and an ammeter—all of them connected in series. Turn the power on and pay close attention to what's going on in the cuvette. Note the ammeter readings.

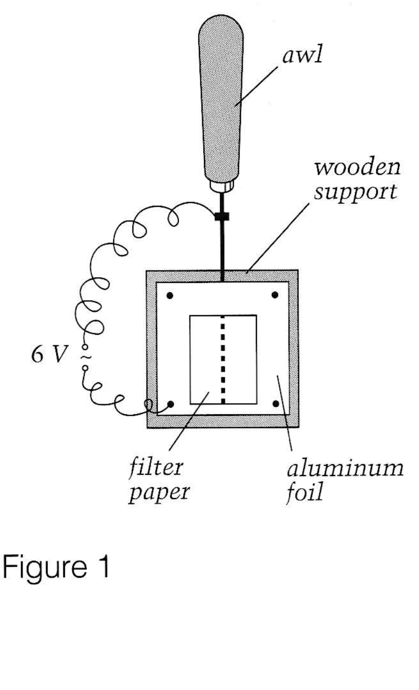


Figure 1

You'll see that the current gradually increases, and the solution begins to boil near the electrodes—in fact, quite a bit of steam is produced. After a while the boiling becomes very intense, the current reaches a maximum—and suddenly drops to almost zero. At the same moment the boiling stops. Then the current begins to increase again, the boiling becomes more and more intense and turbulent—and the process stops again. This will keep happening until the current is turned off. So, the apparatus is a true electrolytic timepiece. How can we explain this periodic boiling and simultaneous periodic change in the electric current?

The sodium chloride solution has a large resistance and so it becomes heated when the current passes through it. The temperature gets so high that the electrolyte boils. Because of the increasing number of gas bubbles, after a certain amount of time the electrodes are practically completely insulated from the solution (steam is a dielectric). Then the circuit is broken, the boiling stops—and the electrodes again make contact with the solution. Then everything repeats again . . . and again . . .

And why do the readings on the ammeter increase? The heating of the solution causes a drop in its resistance, and according to Ohm's law the electric current in the circuit increases.

Now continue with the experiment. Turn the transformer off, leave only one steel electrode in the cuvette, and take away the other. Connect a very thin bare copper or a steel wire 0.5 mm (or less) in diameter to the transformer's terminal in place of the electrode you removed. Wrap the wire with insulating tape and, holding it by the tape, turn the transformer on. Touch the surface of the electrolyte with the free tip of the wire and gradually immerse it to a depth of about one centimeter.

What occurs in the electrolyte isn't just boiling—the process is accompanied by the glowing of the red-hot tip of the wire, with spectacular bursts of yellow flame and a

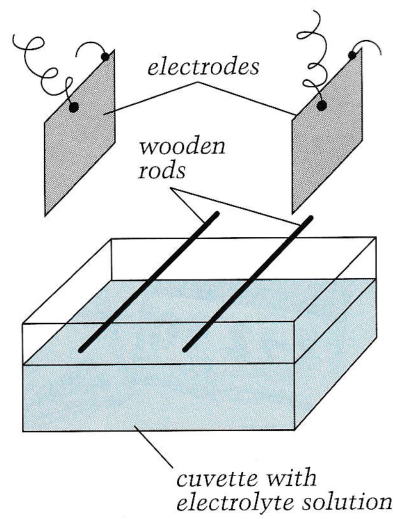


Figure 2

periodic sharp cracking sound. All of this can be easily explained.

First, the wire becomes red-hot due to its small diameter. However, the periodicity of the heating can't be seen by the naked eye because the interruptions in the current are too rapid. Second, in our system composed of one very large and one very small electrode, current rectification occurs (one-way conductivity). This leads to electrolysis of the table salt. In the process, hydrogen is produced at the small electrode (wire), which then mixes with very small drops of sodium chloride solution and explodes with a yellow flame and a cracking sound. In addition, at high temperatures water vapor decomposes at the surface of the red-hot wire: $2H_2O \rightarrow 2H_2 + O_2$. The resulting mixture of hydrogen and oxygen also flares at the tiniest spark—or rather, it explodes with a loud report.

Can any of these effects be observed if we used DC rather than AC? Using a rectifier for the 6–10 V output voltage, you can create all of these effects even more spectacularly, because an additional process occurs in parallel with the ones described—the electrolysis of the aqueous solution of sodium chloride.

One final note. If you don't have an AC ammeter, it doesn't matter. In its place you can use a light bulb rated for 6–10 V and 5–10 W. The effect you produce will be no less spectacular.

AT THE BLACKBOARD

The quadratic

Strengthen your grip on quadratic equations, quadratic functions, and their graph—the parabola

by Vladimir Boltyansky

LET'S BEGIN BY RECALLING how to solve the quadratic equation

$$x^2 + px + q = 0 \qquad (1)$$

with real coefficients p and q. This is one form of the general quadratic equation, although it may not be the form to which you've become accustomed. If our notation seems different and strange, you are welcome to recast our results using your own notation. You will find that anything you already know is expressed here in an equivalent form, although sometimes our forms are more convenient for our discussion. Many of our results will in fact be familiar to you, but some are bound to be new. The equation can be rewritten in the form

$$\left(x + \frac{p}{2}\right)^2 = \frac{1}{4}\left(p^2 - 4q\right) \qquad (2)$$

(which is easily verified by removing the parentheses). The number $D = p^2 - 4q$ is called the *discriminant* of equation (1). The equation now takes the form

$$\left(x + \frac{p}{2}\right)^2 = \frac{1}{4}D, \qquad (3)$$

and it becomes clear that if D is nonnegative, then equation (1) has two real roots

$$x_1 = \frac{1}{2}\left(-p + \sqrt{D}\right),\ x_2 = \left(-p - \sqrt{D}\right), \qquad (4)$$

which are different for $D > 0$ and coincide for $D = 0$.

In the case of negative D, equations (1) and (3) have no real roots, because their left sides cannot be negative for any real x. In this case we can rewrite equation (3) as

$$\left(x + \frac{p}{2}\right)^2 - \frac{1}{4}\left(\sqrt{-D}i\right)^2 = 0, \qquad (5)$$

where i is the *imaginary unit*—that is, $i^2 = -1$. Now we factor the left side and get

$$\left(x + \frac{p}{2} - \frac{\sqrt{-D}}{2}i\right)\left(x + \frac{p}{2} + \frac{\sqrt{-D}}{2}i\right) = 0,$$

from which it follows that the equation has two complex solutions

$$x_1 = \frac{1}{2}\left(-p + \sqrt{-D}i\right),$$
$$x_2 = \frac{1}{2}\left(-p - \sqrt{-D}i\right). \qquad (6)$$

This argument is summarized by the following theorem.

THEOREM 1. *The quadratic equation (1) with real coefficients p and q has two roots whose form depends on the value of the discriminant $D = p^2 - 4q$. If $D > 0$, the roots are real and distinct (see equation (4)); if $D = 0$, the roots are real and coincide; and if $D < 0$, the roots are complex numbers (with a nonzero imaginary part—see equation (6)).*

The next theorem was established by the famous French mathematician François Vieta (1540–1603), one of the pioneers in introducing letter notation and the modern system of algebraic symbols.

THEOREM 2. *The roots x_1 and x_2 of quadratic equation (1) satisfy the relations*

$$x_1 + x_2 = -p,\quad x_1x_2 = q. \qquad (7)$$

This statement can be proved by direct calculation. For instance, in the case of real roots (that is, for nonnegative D) we find from formulas (4)

$$x_1 + x_2 = \frac{1}{2}\left(-p + \sqrt{-D}\right) + \frac{1}{2}\left(-p - \sqrt{-D}\right)$$
$$= -p,$$
$$x_1x_2 = \frac{1}{4}\left(-p + \sqrt{-D}\right)\left(-p - \sqrt{-D}\right)$$
$$= \frac{1}{4}\left(p^2 - D\right) = q.$$

In the case of complex roots (for $D < 0$) Vieta's formulas (7) are similarly derived from equations (6).

THEOREM 3. *Any quadratic trinomial can be factored into linear factors*

$$x^2 + px + q = (x - x_1)(x - x_2),$$

where x_1 and x_2 are the roots of quadratic equation (1).

Indeed, by Vieta's formulas we have

$$x^2 + px + q = x^2 - (x_1 + x_2)x + x_1x_2 = (x - x_1)(x - x_2).$$

Sometimes the factorization is conveniently achieved by *completing the square* (compare with equation (2)) rather than calculating roots using formulas (4) or (6). For example,

$$x^2 + 8x - 33 = (x + 4)^2 - 49,$$

which leads to the factorization

$$x^2 + 8x - 33 = (x - 3)(x + 11).$$

And one more note. In the case of an even integer p it's often convenient to use the formulas

$$\frac{1}{4}D = \left(\frac{p}{2}\right)^2 - q$$

and

$$\frac{\sqrt{D}}{2} = \sqrt{\left(\frac{p}{2}\right)^2 - q}.$$

Now we're fully versed in all kinds of quadratics and can get to some problems.

Problems

1. Prove that the discriminant D of the quadratic $x^2 + px + q = 0$ is equal to $(x_1 - x_2)^2$, where x_1 and x_2 are the solutions to this equation.

2. Check by direct substitution that the numbers given by formulas (4) are solutions to equation (1) for $D \geq 0$. Do the same with the numbers from equations (6) for the case $D < 0$.

3. Solve the following quadratic equations:

(a) $x^2 - 5 = 0$;
(b) $x^2 + 7 = 0$;
(c) $x^2 + 3x = 0$;
(d) $x^2 - 7x + 12 = 0$;
(e) $x^2 - x - 30 = 0$;
(f) $x^2 + 4x + 5 = 0$;
(g) $x^2 + 10x + 25 = 0$;
(h) $x^2 + 2(a - 1)x - (6a + 3) = 0$;
(i) $x^2 + 2(a + 3)x + (a^2 + 2a + 9) = 0$;
(j) $x^2 - 2(a^2 - 1)x + (a^4 - a^2 + 1) = 0$;
(k) $2x^2 - 5x + 2 = 0$;
(l) $(1 + a)x^2 + 2x\sqrt{a^2 + 1} - (1 - a) = 0$.

4. Prove that for $D > 0$ the graph of a quadratic function $y = x^2 + px + q$ intersects the x-axis at two points $(x_1, 0)$, $(x_2, 0)$, where x_1 and x_2 are the roots of the equation $x^2 + px + q = 0$ (fig. 1). For $D = 0$ the graph touches the x-axis (fig. 2). Finally, for $D < 0$ the graph lies completely above the x-axis—they have no common points (fig. 3).

5. Write out the quadratic equation that has the following roots:

(a) $x_1 = 1$, $x_2 = -2$;
(b) $x_1 = x_2 = -4$;
(c) $x_1 = 2 - 3i$, $x_2 = 2 + 3i$;
(d) $x_1 = a + bi$, $x_2 = a - bi$;
(e) $x_1 = 3 - 4i$, $x_2 = 2 - 5i$.

6. Prove that for any p and q the system of equations

$$\begin{cases} y + z = -p, \\ yz = q \end{cases}$$

has two solutions: $y = x_1$, $z = x_2$; $y = x_2$, $z = x_1$ (real or complex, distinct or coincident), where x_1 and x_2 are the roots of the quadratic $x^2 + px + q = 0$.

7. Prove that the graph of the quadratic function $y = x^2 + px + q$ is symmetric about the line $x = -p/2$ (fig. 4).

8. Prove that the roots of the quadratic $x^2 + px + q = 0$ (with real p, q) are real and positive if and only if $D \geq 0$, $p < 0$, $q > 0$.

9. Find necessary and sufficient conditions for the roots of the quadratic $x^2 + px + q = 0$ (with real p, q) to be real, nonzero and (a) of the same sign, (b) of different signs.

10. Prove that if one root of the quadratic $x^2 + px + q = 0$ (with real coefficients p, q) is real, then the other is real, too.

11. Prove that if one root of the quadratic $x^2 + px + q = 0$ (with real p, q) is not a real number—that is, has the form $a + bi$ with $b \neq 0$, then the other root of this equation is equal to $a - bi$; in particular, it's not real either.

12. Find the set of all real x satisfying the inequality $x^2 + px + q < 0$ (with real p, q). (The answer depends on D!)

13. Solve the following strict quadratic inequalities (and make diagrams):

(a) $x^2 - 5x + 6 < 0$;
(b) $x^2 - 10x + 25 > 0$;
(c) $x^2 - x - 12 > 0$;
(d) $x^2 - 12x + 38 > 0$.

14. Solve the following weak quadratic inequalities:

(a) $x^2 - 3x - 18 \leq 0$;
(b) $x^2 - 8x + 16 \leq 0$;
(c) $x^2 + 6x + 5 \geq 0$;
(d) $x^2 - 14x + 50 \leq 0$.

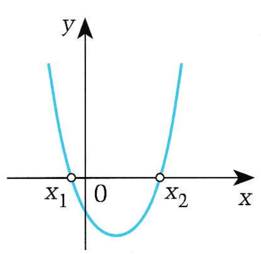


Figure 1

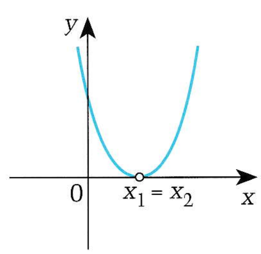


Figure 2

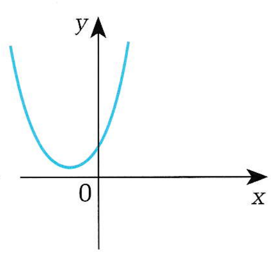


Figure 3

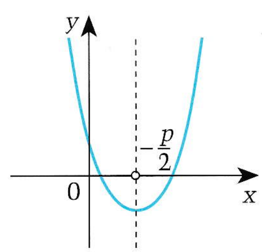


Figure 4

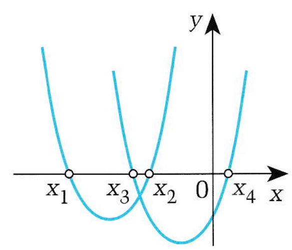


Figure 5

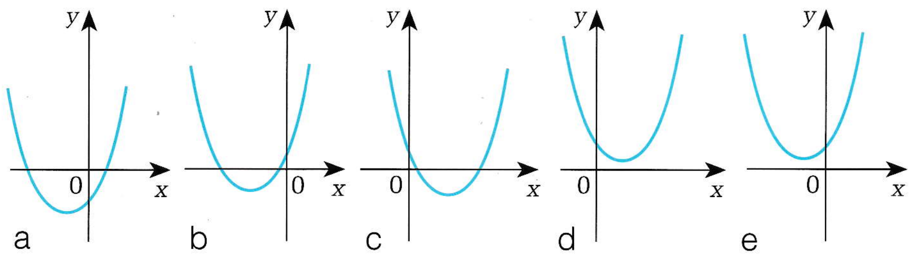


Figure 6

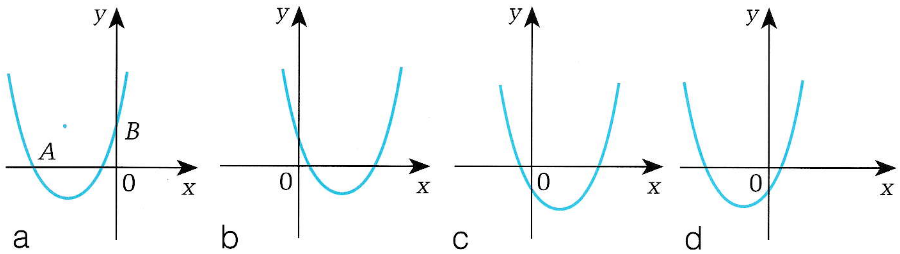


Figure 7

15. Prove that if the roots of both quadratic equations

$$x^2 + p_1x + q_1 = 0,$$
$$x^2 + p_2x + q_2 = 0$$

are real and belong to the segment $[a, b]$, then for any $k > 0$ the roots of the equation

$$x^2 + p_1x + q_1 + k(x^2 + p_2x + q_2) = 0, \quad (8)$$

if they are real, lie on the same segment.

16. Prove that if the roots x_1, x_2 of the equation $x^2 + p_1x + q_1 = 0$ and the roots x_3, x_4 of the equation $x^2 + p_2x + q_2 = 0$

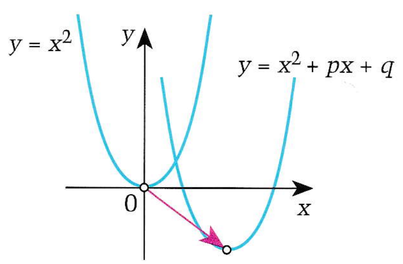


Figure 8

are all real and alternate—that is, $x_1 < x_3 < x_2 < x_4$ (fig. 5), then for any $k > 0$ the roots of equation (8) in problem 15 are real, one of them belonging to the segment $[x_1, x_3]$ and the other to $[x_2, x_4]$.

17. In the conditions of problem 16, prove that for any negative $k \neq -1$ the roots of equation (8) are real, one of them belonging to the segment $[x_3, x_2]$, the other lying outside $[x_1, x_4]$.

18. For what values of a does the equation $x^2 + ax + 6 = 0$ have integer roots?

19. For what values of a does the equation $(x - 10)(x - a) + 1 = 0$ have integer roots?

20. What are the signs of the numbers p and q if the graphs of $y = x^2 + px + q$ looks like the ones in figure 6?

21. Let x_1 and x_2 be the roots of the quadratic $x^2 + px + q = 0$. Find p and q knowing that $x_1 + 1$ and $x_2 + 1$ are the roots of the equation $x^2 - p^2x + pq = 0$.

22. Points A and B distinct from the origin are given on the x- and y-axes, respectively. Prove that there exists one and only one quadratic function $y = x^2 + px + q$ whose graph passes through A and B (fig. 7).

23. For what real values of a is the function $y = x^2 + 2ax + 1$ positive for all real x?

24. The equations $x^2 + p_1x + q_1 = 0$ and $x^2 + p_2x + q_2 = 0$ have real coefficients such that $p_1p_2 = 2(q_1 + q_2)$. Prove that at least one of the equations has real roots.

25. Consider a polynomial $f(x, y) = ax^2 + bxy + cy^2$ with real coefficients $a > 0$, b, c. Prove that one of the following statements is true:

(a) $f(x, y) = a(x - m_1y)(x - m_2y)$, where $m_1 \neq m_2$ are real numbers (here f is said to be a *hyperbolic polynomial);*

(b) $f(x, y) = a(x - my)^2$, where m is real (a *parabolic polynomial);*

(c) $f(x, y) > 0$ for all real x, y except $x = y = 0$ (an *elliptic polynomial*).

26. Prove that the graph of the trinomial $y = x^2 + px + q$ is obtained from the graph of $y = x^2$ under a translation (fig. 8) by the vector $\mathbf{a} = (-p/2, -D/4)$.

27. Prove that the system of equations

$$\begin{cases} y = x^2 + px + q, \\ y = ax + b \end{cases}$$

has no more than two solutions for any real a, b (fig. 9).

28. Prove that the interior of the parabola $y = x^2 + px + q$—that is, the set of all points (x, y) such that $y > x^2 + px + q$—is *convex* (which means that the segment joining any two points of this set lies entirely in this set—see figure 10).

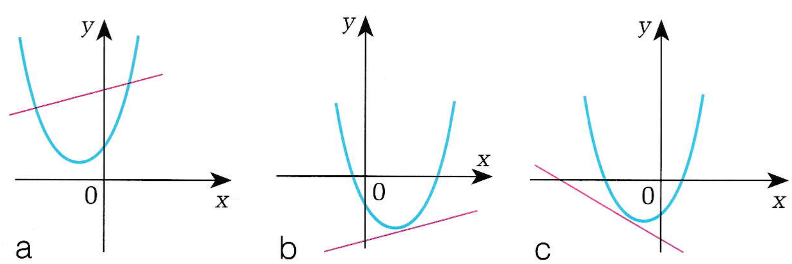


Figure 9

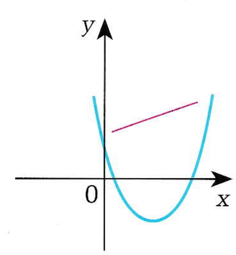


Figure 10

FOLLOW-UP

Life on an accelerating skateboard

What is "weight" anyway?

by Albert A. Bartlett

THE CONTEST PROBLEM "Weighing an Astronaut" in the March/April issue of *Quantum* contains a most interesting problem. A drawing on page 37 shows a person standing on the platform of a bathroom scale on a skateboard that is accelerating down an incline that makes an angle θ with the horizontal, as shown in figure 1. The question asked is, "What does the scale read?"

This problem contains two flaws:

1. A person could not stand as shown with his or her body in the orientation shown in figure 1.
2. The question of the reading of the bathroom scale cannot be answered without information about the mechanical construction of the scale.

A study of these flaws leads us to see serious inconsistencies that appear when one uses the conventional textbook definition of

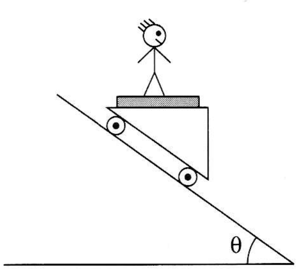


Figure 1

"weight." This leads us to a much improved definition of weight that is given by the International Standards Organization (ISO).

Analysis of the problem

Let's think about a frame of reference in which the measured free-fall acceleration is $\mathbf{g}_1$ (a vector). In this frame, the weight of a mass m is $\mathbf{W}_1 = m\mathbf{g}_1$. Now consider a second frame of reference that has an acceleration $\mathbf{a}_{21}$ (a vector) with respect to the first frame. The free-fall acceleration in the second frame is a vector $\mathbf{g}_2$, which is given by the vector equation

$$\mathbf{g}_2 = \mathbf{g}_1 - \mathbf{a}_{21}. \qquad (1)$$

The weight of the mass m in the second frame will be $\mathbf{W}_2 = m\mathbf{g}_2$. An observer at rest in the first frame will say that "up" is the direction of $-\mathbf{g}_1$, and an observer at rest in the second frame will say that "up" is the direction of $-\mathbf{g}_2$.

The utility of equation (1) can be verified by applying it to the common problem of a person standing on the platform of a bathroom scale in an elevator that is accelerating upward or downward as shown in figure 2. The beauty of equation (1) is that it is a vector equation that can be applied in cases where the accelerations are not collinear.

Let the laboratory be frame 1, in which we see the skateboard accelerating down an incline that makes

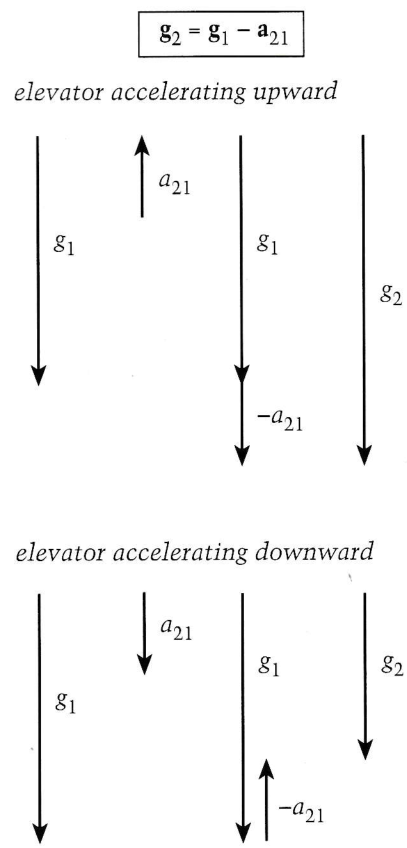
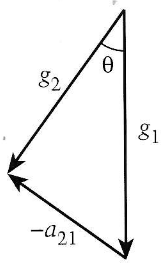


Figure 2

an angle θ with the horizontal. Let frame 2 be the frame in which the skateboard is at rest. For this case, $\mathbf{g}_1$ is the free-fall acceleration (approximately 9.8 m/s^2) measured in the laboratory. If we neglect friction and the rotational inertia of the skateboard's wheels, the skateboard will have a vector acceleration $\mathbf{a}_{21}$ of magnitude $g_1 \sin\theta$ down the incline.



$a_{21} = g_1 \sin\theta$

Figure 3

The vector $-\mathbf{a}_{21}$ will point up the incline.

Figure 3 shows the vector addition of equation (1) for this problem. The vector triangle of figure 3 is a right triangle because the magnitude of $\mathbf{a}_{21}$ has been shown to be $g_1 \sin\theta$. From figure 3 we can see that the magnitude of $\mathbf{g}_2$ is $g_1 \cos\theta$, and the direction of $\mathbf{g}_2$ is perpendicular to the surface of the incline down which the skateboard is rolling. If you were on the skateboard, you'd tell us that "down" is the direction of $\mathbf{g}_2$ and that "up" is the direction of $-\mathbf{g}_2$. In order to stand "up," you'd have to stand so that your body is perpendicular to the incline, which seems unnatural. This is why ski instructors tell you to "lean forward." In the frame of the ski slope, the direction of "up" is parallel to the trunks of the pine trees. In the frame of the skier (or skateboard rider) accelerating on a slope, "up" is not parallel to the pine trees but is perpendicular to the surface of the slope.

In figure 1, the person's body is shown as being parallel to the pine trees, which means that it is at an angle θ to "up" in that person's accelerating reference frame. If you tried to stand like that, you'd fall over to the left. This appears to be correctly represented in the lower-left corner of the cover of the March/April issue. The man shown riding a horizontal platform down the steep incline appears to be falling backward to the left.

In order to get to the question of the reading of the platform scale, let's avoid the difficulty of "which way is up" by replacing the person with a point mass m that rests on the platform on the scale. We must then assume that the coefficient of friction between m and the platform is of the order of unity so that m won't slide on the platform as the scale accelerates down the slope. Now we can ask, "What is the reading of the scale?"

To answer this, we need information about the mechanics of the operation of the scale. In particular, does the scale read the magnitude of the force acting on its platform independent of the direction of the force, or does it read only the component of the force that is perpendicular to the platform? If the scale reading R_t is the total force F_t acting on the platform, the scale will read

$$R_t = F_t = W_2 = mg_2 = mg_1 \cos\theta.$$

If the scale reads only the component of the total force that is perpendicular to its platform, its scale reading R_c will be

$$\begin{aligned} R_c &= F_t \cos\theta = mg_2 \cos\theta \\ &= mg_1 \cos^2\theta. \end{aligned}$$

Figure 4 shows how a person would stand in riding a skateboard down an incline. In this case, the scale reading is $R_t = mg_1 \cos\theta$.

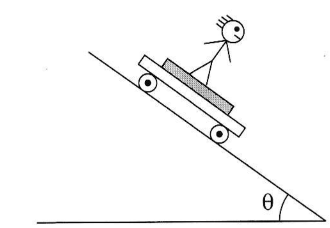


Figure 4

Problems in defining weight

There is much clarity and consistency to be gained by replacing the traditional definition of weight with the ISO definition that has been used above. Almost universally in textbooks, unnecessary confusion is caused by the use of the conventional definition of weight, which says that weight is the product of mass times the acceleration of gravity.

This can be illustrated by tracing seven steps in the development of weight and related concepts as they are encountered sequentially in typical introductory physics texts.

1. Texts start by describing measurements of the free-fall acceleration ($g = 9.8$ m/s^2) and this quantity is called the acceleration due to gravity.

2. Next, the texts define the weight of a mass m as the product of m and $\mathbf{g}$, which they call the force of gravity:

$$\mathbf{W} = m\mathbf{g}.$$

3. When a person of mass m stands on a bathroom scale in an elevator that has an acceleration upward or downward of magnitude a, the scale reading R is

$$R = m(g \pm a).$$

4. This quantity is a force that is much like a weight. But it does not match the definition of weight given in step 2 above, so textbook authors have to coin a new name for R. Texts generally call R the "apparent weight" of the mass m.

5. Now imagine cutting the cable so the elevator is in free fall. The apparent weight R goes to zero. Here is our first point of confusion. To be logical, we should say that when the apparent weight goes to zero, the resulting condition is "apparent weightlessness." Instead, texts call this condition "weightlessness."

6. Now the confusion becomes serious. An astronaut in orbit is observed to be weightless. But according to the textbook definition of weight (mass times the acceleration due to gravity), the astronaut still has weight. This has led to vigorous (and generally unproductive) debate as to whether or not an orbiting weightless astronaut has weight.

7. Now for the final confusion. Weight was originally defined to be due solely to gravity, and apparent weight was defined to be due to gravity plus the acceleration of a reference frame. But in the chapter where we learn that the laboratory frame is on the rotating Earth, the texts tell us that the reading of the bathroom scale in the laboratory is the result of gravity plus the effect of the centripetal acceleration that arises from the rotation of the Earth.

Therefore, the reading of the scale in the laboratory is really an apparent weight.

Here we have the ultimate confusion. In step 2 we defined the reading of the scale in the laboratory to be a weight. We then defined apparent weight to be quite different from weight. But now we find in step 7 that what we defined to be weight is really an apparent weight! The textbooks rarely point out this glaring internal inconsistency.

The definition of weight

The ISO gives us the definition of weight that we need to eliminate these inconsistencies: *The weight of a body in a specified reference system is the force that, when applied to the body, would give it an acceleration equal to the local acceleration of free fall in that reference system.* The main requirement of this definition is that we give up the use of the words "acceleration due to gravity" to describe the free-fall acceleration that is measured in the laboratory. In effect, we are asked to call the free-fall acceleration the "free-fall acceleration" in all frames of reference. This definition does not change the common relation that we use to calculate the weight of an object:

$$W = mg.$$

Let's retrace the textbook steps 1 to 7 above to see how this definition of weight eliminates the serious internal inconsistencies resulting from the use of the conventional definition. When the frame of reference is the surface of the Earth, the magnitude of the free-fall acceleration is approximately $g = 9.8$ m/s^2 and all weights have their conventional numerical magnitudes: $W = mg$.

In the accelerating elevator, the free-fall acceleration is $g \pm a$, and a mass m on a scale in the elevator causes the scale to read $m(g \pm a)$. This scale reading is called the weight of m in the frame of the accelerating elevator. The magnitude of the weight in the accelerating elevator has changed from its value in the laboratory frame, but this should cause no problem because we have always emphasized to students that weight is not an intrinsic or invariant property of the mass m. The confusing term "apparent weight" is no longer needed! When the elevator containing m is in free fall, the free-fall acceleration with respect to the elevator is zero, so the weight of m in the frame of the falling elevator is zero, and it is then consistent to say that m is weightless.

With reference to the frame of an orbiting spacecraft, an astronaut has zero free-fall acceleration, and so the astronaut is weightless. Finally, when we look at the rotating Earth, no confusion results from saying that part of the free-fall acceleration in the laboratory is due to gravitation and part is due to the Earth's rotation.

So, the ISO definition of weight gives internal consistency to our nomenclature and usage. It eliminates the confusion that is perpetuated in textbooks and has plagued generations of physics students.

Suggestions for further reading

Bartlett, A. A. "The Forces of Gravity in Some Simple Accelerated Systems: Which Way is Up?" *The Physics Teacher*, November 1972, pp. 429–37.

———. "Why the Ski Instructor Says, 'Lean Forward!'" *The Physics Teacher*, January 1987, pp. 28–31.

French, A. P. "On Weightlessness." *American Journal of Physics*, February 1995, p. 105.

Iona, M. "International Weight." *American Journal of Physics*, February 1995, p. 106.

Albert A. Bartlett *is an emeritus professor of physics at the University of Colorado.*

Caught in the Web

Mir, Discovery, and Atlantis

This has been a busy year for Russians and Americans working together in space. The pictures on this page are a small sampling of the images stored at various NASA sites on the World Wide Web. These images, taken with the IMAX large-format movie camera, were found at http://www.hq.nasa.gov/office/pao/NewsRoom/today.html.

Photo courtesy NASA and Russian Space Agency

The Russian Space Agency's Mir Space Station is seen over the Pacific Ocean during rendevous operations with the Space Shuttle Discovery in February.

If you are particularly interested in the shuttle missions, NASA has set up sites devoted to some of them. For instance, you can find video clips, still pictures, and tons of information at http://shuttle.nasa.gov/sts-71. If your interests lie elsewhere—the Hubble Space Telescope, the Ulysses solar mission, the Apollo 11 lunar landing, whatever—just make a bookmark for http://www.nasa.gov/hqpao/hqpao_home.html. You're sure to return time and again, because you'll find much more than you were looking for.

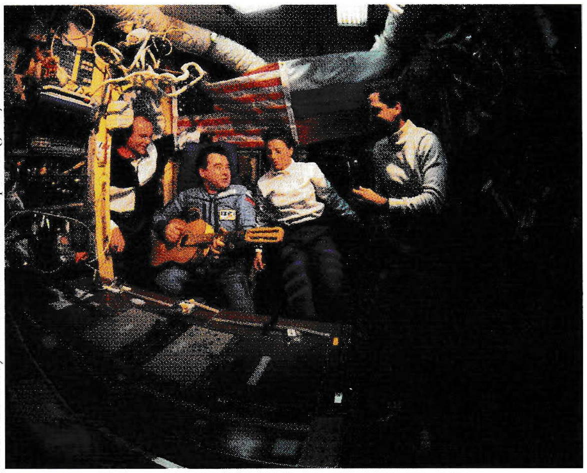


Photo courtesy NASA and Russian Space Agency

Gennady Strekalov plays guitar and sings with Atlantis astronauts Charlie Precourt, Bonnie Dunbar, and Greg Harbaugh aboard the Mir Space Station.

CyberTeaser winners

It seemed like a simple question, and several visitors to the *Quantum* home page handled it with aplomb. (See brainteaser B153 on page 19.) Others were a bit more tentative. And some, we're sorry to report, had it backwards. But we hope everyone enjoyed thinking about this little conundrum.

The following will receive a copy of this issue of *Quantum* and a button designed by staff artist Sergey Ivanov:

Jim Hanby (Lexington, South Carolina)
Louis Smadbeck (Martha's Vineyard, Massachusetts)
Jeff Dodson (Long Beach, California)
Hal Harris (St. Louis, Missouri)
Ben Davis (Wayne, Maine)
Paul Grayson (Urbana, Illinois)
Daniel Jordan (Roscoe, Illinois)
Cheng-Chih Chien (Columbus, Ohio)

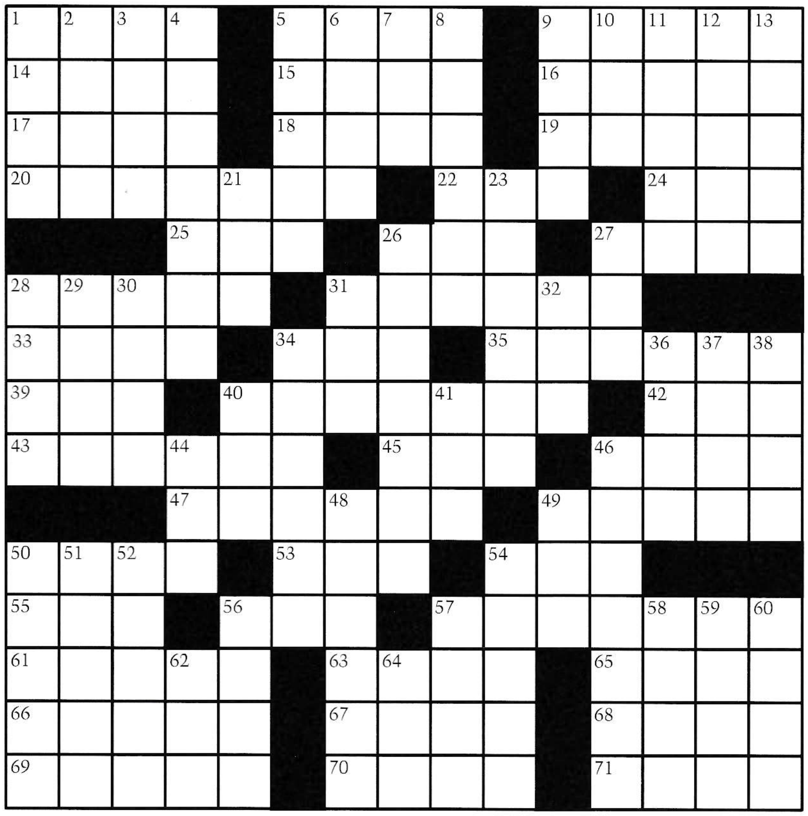
The next CyberTeaser has been posted and awaits your attempts to crack it. Go to http://www.nsta.org/quantum and follow the link.

Index of Advertisers

American Scientist	25
Dalog	51
Didax	10
Duracell/NSTA Scholarship Competition	18
HighText Publications	44
JETS	44
OnScreen Science	10
Scientific American	11
Suntrak	Cover 3
Telos	34
The Factory	44
US Air Force	Cover 4

Criss×cross science

by David R. Martin



Across

1 Phosphatase unit
5 Soviet linguist Nikolay ___ (1864–1934)
9 Type of spectroscopy
14 Type of parity
15 Pelvic bones
16 966,314 (in base 16)
17 Soil
18 53,181 (in base 16)
19 Musical group of nine
20 Math subject
22 Chemical suffix
24 Distilled coal
25 Earth: comb. form
26 Automobile
27 10^{-12} (pref.)
28 Sunny mountain side
31 Parabolas, e.g.
33 Mite: comb. form
34 State of matter
35 Gramophone inventor
39 10^{18}: pref.
40 Orbital path
42 Samuel's teacher
43 Like computer numbers
45 "Mr. Basketball" ___ Holman
46 Hawaiian wreaths
47 Biologist ___ Carson (1907–1964)
49 Cosmologist Sir Hermann ___
50 Element 30
53 Characterized by: suff.
54 For each
55 Zeta follower
56 Priest's robe
57 For hydrogen, it equals one
61 Domesticated
63 Anthropologist Franz ___ (1858–1942)
65 Lunch time
66 Parapet passage
67 Trails
68 About 4.9289 cm^3: abbr.
69 More hurt
70 Sea eagle
71 "Y" followers

Down

1 44,762 (in base 16)
2 Amoral
3 1980 Nobelist in Chemistry
4 FORTRAN data type
5 10^{-6}: pref.
6 ___ process (for butter manufacture)
7 Side bone
8 About 57
9 Microbiologist ___ Dubos (1901–1982)
10 Blood group system
11 Mind: comb. form
12 700,076 (in base 16)
13 Sodium: comb. form
21 Gamble wager
23 Least wet
26 Trig. functions
27 Pressure unit: abbr.
28 43,755 (in base 16)
29 611
30 Horse color
31 Unit of heat (abbr.)
32 Alphabet run
34 Antifreeze chemical
36 Spotted
37 Smelly
38 ___ *prius*
40 Eon subdivision
41 Friend
44 Circle part
46 ___ transformation (of special relativity)
48 Astronomer Edwin ___ (1889–1953)
49 Ten decibels
50 Epsilon followers
51 Novelist ___ Svevo (1861–1928)
52 City on the Sambre
54 Out-of-date
56 ___ wax (ozocerite)
57 Oceanographer ___ Ekman (1874–1954)
58 Facial feature
59 Manage
60 River near Schladming
62 Poet's before
64 Row

SOLUTION IN THE NEXT ISSUE

SOLUTION TO THE JULY/AUGUST PUZZLE

D	E	B	T		B	A	B	C	C		C	B	A	C
A	T	O	R		A	D	O	R	E		H	O	B	O
B	U	R	I		R	O	W	A	N		R	H	E	A
B	I	N	A	R	Y		L	I	T	T	O	R	A	L
			N	E	O	N		G	R	A	M			
D	I	A	G	O	N	A	L		O	C	E	A	N	S
A	R	I	L	S		D	I	G	I	T		L	O	T
D	A	M	E		M	I	N	E	D		C	I	N	E
A	T	E		R	A	R	E	R		S	O	B	E	R
B	E	D	S	I	T		S	M	I	T	H	I	T	E
			C	O	R	M		S	L	O	E			
S	O	L	U	T	I	O	N		L	A	R	M	O	R
O	R	Y	Z		C	R	O	L	Y		E	A	E	A
L	E	N	Z		E	A	D	E	E		N	Y	C	T
D	O	X	Y		S	L	E	D	S		T	A	O	S

ANSWERS, HINTS & SOLUTIONS

Math

M151

(a) The answer is $a = 549$. The sums of digits of felicitous numbers are always even. Therefore, if the numbers a and $a + 1$ are felicitous, then a ends in 9; otherwise the sums of their digits differ by one and are of different parity. The number a can't have only two digits, because $99 + 1 = 100$ isn't felicitous. If a has three digits, $a = \overline{xy9}$ (the bar here denotes decimal notation), then $y < 9$ and $a + 1 = \overline{x(y+1)0}$. So $x + y = 9$, $x = y + 1$. Thus, $x = 5$, $y = 4$. (This is the only 3-digit solution, not just the smallest.)

(b) The answer is no, because, as follows from the above discussion, of three successive felicitous numbers a, $a + 1$, $a + 2$, two—namely a and $a + 1$—must end in 9, which is impossible.

M152

Let's rotate the square (fig. 1) about its center by 90° so that A_1 goes into A_4, A_2 into A_1, and so on. Then the lines PA_1, PA_2, PA_3, PA_4 will be taken to the corresponding perpendiculars (PA_1 becomes the perpendicular to PA_1 through A_4 because it's rotated 90°, and so on). So the four perpendiculars meet at the image P' of P under this rotation.

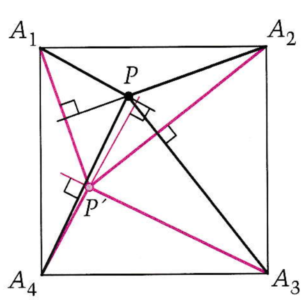


Figure 1

M153

The required relation is $c = ab/3 - 2a^3/27$; or, equivalently, the number $-a/3$ must be a root of the equation.

Indeed, if x_1, x_2, x_3 are the three roots, then by the Factor Theorem we can write

$$x^3 + ax^2 + cx + d = (x - x_1)(x - x_2)(x - x_3).$$

Multiplying out the product on the right and equating the corresponding coefficients of the two polynomials, we get

$$-(x_1 + x_2 + x_3) = a.$$

But since the roots form an arithmetic sequence, $x_1 + x_2 = 2x_3$, so $x_3 = -a/3$.

Try to derive similar conditions for quartic equations.

M154

The answer to part (a) is $n - k + 1$. We'll consider it together with (b).

Isolate the "series" of the given numbers that begin with a two and includes all the ones that immediately follow this two (clockwise). If a series includes m ones, $m \geq 0$ (see figure 2, where the number immediately after these ones, which is 2 or 0, is denoted by an asterisk), then

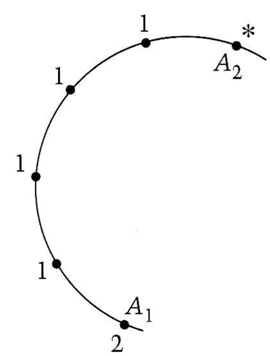


Figure 2

a

A_1	A_2
2	*
0	1

b

A_1		A_2
2	1	*
Ø	2	0
Ø	0	1

c

A_1					A_2
2	1	1	1	1	*
Ø	2	1	1	1	0
Ø	0	2	1	1	0
Ø	0	0	2	1	0
Ø	0	0	0	2	0
Ø	0	0	0	0	1

Figure 3

after $m + 1$ steps it will have zeros instead of all its ones. The subsequent asterisk number will turn into 1. Figures 3a–3c show the evolution of a series for $m = 0, 1, 4$, respectively (the sign Ø there denotes 0 or 1). Each series undergoes an evolution of this sort. Therefore, if there were k twos initially and the longest series contained $m \leq n - k$ ones, then after $m + 1$ transformations no twos will be left. If, in addition, the $n - k$ numbers distinct from 2 in the initial arrangement were all ones, they all will be replaced with zeros after at most $n - k + 1$ steps, while the k twos will turn into ones. (N. Vasilyev)

M155

Choose any log l_1 at random and try to roll it down into the river. Suppose it gets blocked by a log l_2. Then try to roll l_2 away. If it also gets blocked by some log, call this log l_3. Proceeding in the same way, we obtain a sequence l_1, l_2, ... of logs in which each log is blocked by the next one. Since the number of logs is finite, either this sequence will have to stop at a certain log l_n, in which case l_n can be rolled away unobstructedly, or it will close in a loop—that is, a certain log will appear in it for the second time. So it suffices to show that the second possibility never occurs.

To this end, for any log l we'll construct a certain domain $E(l)$ of the bank such that $E(l_1) \subset E(l_2)$ (and $E(l_1) \neq E(l_2)$) whenever log l_2 obstructs log l_1. So for the sequence of logs $l_1, l_2, \ldots$ considered above we'll get a "strictly expanding" sequence $E(l_1) \subset E(l_2) \subset \ldots$, which means that none of the logs can reappear in our sequence.

Think of the bank as the upper half-plane bounded by a horizontal line b and of the logs as line segments. Consider the *shadow* $S(l)$ of each log l—that is, the set of points that lie above and on the segment l and two rays at an angle of 45° to b: one drawn to the right (and upward) from the right endpoint of l, the other drawn to the left from the left endpoint (fig. 4). We can see that any log that has no common points with $S(l)$ will never get caught by l while it's rolled into the river.

Now define the *extended shadow* $E(l)$ of l. Start at any point on l and move to the right along l and along the right slope of $S(l)$ until we encounter the first log l_1 intersecting this slope. Then continue moving along the border of $S(l_1)$ up to the next log l_2 (intersecting the right slope of $S(l_1)$), and so on. The path thus traced is a part of the border of $E(l)$. The remaining part is traced similarly, starting from the same point and moving to the left. The set $E(l)$ consists of all points above and on the path we've traced (fig. 5). By

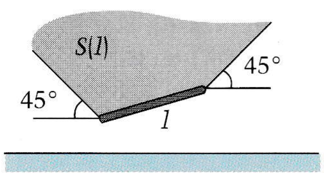


Figure 4

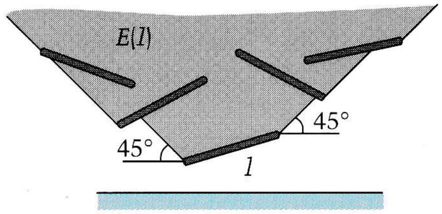


Figure 5

construction, any log l' either has no common points with $E(l)$ or lies entirely in $E(l)$. In the first case l' is disjoint with $S(l)$, too, and so l doesn't obstruct rolling l' into the river. In the second case the extended shadow $E(l')$ of l' lies in $E(l)$ (even if the borders of $E(l')$ and $E(l)$ have a common point A, the ascending parts of these borders must coincide starting from A).

So, if l obstructs l', then $E(l') \subset E(l)$, which completes the proof.

Physics

P151

Among all possible trajectories we take the one tangent to the shelter. Consider the flight of the projectile in the frame of reference with axes directed as shown in figure 6. In this system the "horizontal" projection (that is, directed along the OX-axis) of the shell's initial velocity is $v_{0x} = v_0 \cos(\phi - \alpha)$, and the "vertical" projection (along the OY-axis) is $v_{0y} = v_0 \sin(\phi - \alpha)$, where ϕ is the angle of the initial velocity vector relative to the horizontal plane.

The point of tangency C determines the maximum height of the shell's ascent over the "horizon," which according to figure 6 is $l \sin \alpha$. At this point the projection of the shell's velocity $\mathbf{v}$ upon the OY-axis is zero, and

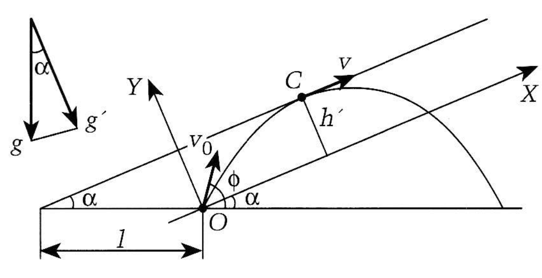


Figure 6

$$h' = \frac{v_{0y}^2}{2g'},$$

where $g' = g \cos \alpha$ is the "acceleration due to gravity" in the chosen reference system (the projection of vector $\mathbf{g}$ upon the "vertical" axis OY). Thus,

$$v_{0y}^2 = 2g'h',$$

and therefore

$$v_0^2 \sin^2(\phi - \alpha) = 2(g \cos \alpha)(l \sin \alpha).$$

In particular this means, that if an additional relationship holds—that is,

$$2gl \sin \alpha \cdot \cos \alpha = gl \sin 2\alpha > v_0^2$$

—there exists no tangent trajectory (a projectile will not "touch" the shelter). In this case the maximum travel distance corresponds to a shell fired at an angle $\phi = \pi/4$ with the true horizontal plane and is equal to $L_{max} = v_0^2/g$.

If the opposite inequality is true—that is,

$$v_0^2 \geq gl \sin 2\alpha$$

—then to touch the shelter the shell must be fired at an angle of

$$\phi_{tan} = \alpha + \arcsin \frac{\sqrt{gl \sin 2\alpha}}{v_0}.$$

If

$$\frac{v_0^2}{v_0^2 + 2gl} \leq \sin 2\alpha$$

(which means that $\phi_{tan} \geq \pi/4$—show this!), the initial slope corresponding to the longest flight again is $\phi_{init} = \pi/4$, and $L_{max} = v_0^2/g$. If the opposite takes place—that is,

$$\frac{v_0^2}{v_0^2 + 2gl} > \sin 2\alpha$$

(which in turn means that $\phi_{tan} < \pi/4$)—then

$$\phi_{init} = \phi_{tan}$$

$$= \alpha + \arcsin \frac{\sqrt{gl \sin 2\alpha}}{v_0},$$

and

$$L_{max} = \frac{v_0^2}{g}\sin 2\phi_{init}$$
$$= \frac{v_0^2}{g}\sin 2\left(\alpha + \arcsin\frac{\sqrt{gl\sin 2\alpha}}{v_0}\right).$$

P152

The electromotive force of an electrical source is equal to the work required for an external force to move a unit charge through the source. The emf is maximal when the alpha particle can't reach the opposite plate in the capacitor due to the electrostatic repulsion—in other words, when the kinetic energy of an emitted alpha particle is equal to its potential energy at the opposite plate of the capacitor. So,

$$\mathscr{E} = \frac{mv_0^2}{2q} = \frac{v_0^2}{2k} \cong 5\cdot 10^5 \text{ V} = 500 \text{ kV}.$$

P153

Since there is water in the nipple at the outset, the water vapor in the vessel is saturated. Its pressure doesn't change during the downstroke of the piston, and all the vapor in the "disappearing" volume

$$V = \frac{\pi D^2}{4H}$$

is converted to water. The mass of the condensed water is

$$m = \frac{\rho\pi d^2}{4h},$$

where $\rho = 10^3$ kg/m^3 is the density of water. The Clapeyron equation gives the pressure of the saturated vapor:

$$P = \frac{m}{\mu}\frac{RT}{V} = \rho\left(\frac{d}{D}\right)^2\frac{h}{H}\frac{RT}{\mu}$$
$$= 2.16 \text{ kPa}.$$

Note: In this formula the values D and d, H and h make dimensionless ratios. Hence, they can be measured in any units—but, of course, in the same ones. For example, D and d are expressed in millimeters, while H and h are expressed in centimeters. If all the other values (ρ, R, T, and μ) are expressed in SI units, the result will be in pascals.

P154

A charged nonconducting rotating cylinder is like a circular electric current that generates a magnetic field. The system looks like a long solenoid with a large number of tightly packed windings along its entire length l. The total current flowing in all the "windings" can be expressed in terms of the total charge on the cylinder:

$$I = \frac{Q}{T} = \frac{\sigma\cdot 2\pi Rl}{2\pi/\omega} = \sigma Rl\omega.$$

The field inside such a cylinder is homogeneous and proportional to the current:

$$B = \alpha I = \alpha\sigma Rl\omega,$$

where α is the proportionality constant.

As the outer cylinder accelerates to an angular velocity ω, the varying magnetic field generates a vortex electric field that acts on the electric charges on the inner cylinder and causes it to rotate. The angular velocity of this cylinder increases to a value ω_1 that corresponds to a zero magnetic field permeating the inner cylinder (which also generates a magnetic field). Recall that the inner cylinder is very light.

Thus we get

$$\alpha\sigma Rl\omega + \alpha\sigma rl\omega_1 = 0,$$

and therefore

$$\omega_1 = -\frac{R}{r}\omega = -2\omega.$$

So the inner cylinder rotates in the opposite direction with twice the angular velocity.

Try to solve a similar problem, when the inner cylinder is set in motion first. Remember that the outer magnetic field is very small.

P155

It will be convenient to consider the refraction in two mutually perpendicular planes. In doing so, we'll draw large angles to make the figures easier to interpret, but consider them small enough to replace the sines of angles by the angles themselves.

Thus, for the incident beam shown in figure 7, the angle of refraction is half of the angle of incidence ϕ, and the refracted ray crosses the principal axis just at the glass–air boundary. The beam emerges

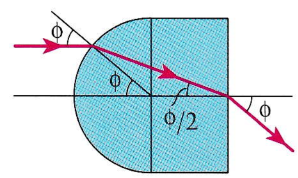


Figure 7

with the angle ϕ relative to the axis, and is deflected by ϕL from the axis at a distance L away from the system.

For the incident beam shown in figure 8, there is no refraction at the entrance of the system, but the beam emerges with the angle ϕ relative to the axis. The corresponding

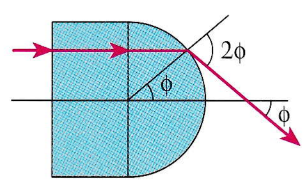


Figure 8

deflection from the axis at a distance L away from the system will also be ϕL (here we assume that $L \gg R$).

Thus, the narrow light beam is (almost) focused to a point shortly after it leaves the composite lens and then becomes a normal divergent conic beam. The ratio of the beam's area when it enters the system to its area at a distance L from the system is

$$S_1 : S_2 = R^2 : L^2 = 1 : 400.$$

Brainteasers

B151

Any two numbers whose sum is 101 are coprime, because if they had a common divisor, it would be a divisor of 101. But this number is a prime, so its only divisor less than itself is one. The rules of the game could as well be formulated as "do anything, then Winnie wins."

B152

If 2^{1995} consists of n digits and 5^{1995} of m digits, then $10^{n-1} < 2^{1995} < 10^n$, $10^{m-1} < 5^{1995} < 10^m$. Multiplying the corresponding terms of these inequalities, we get

$$10^{n+m-2} < 10^{1995} < 10^{n+m}.$$

It follows that $n + m - 2 < 1995 < n + m$, or $1995 = n + m - 1$. So the total number of digits in the two given numbers is 1996.

B153

The flying ball has to work against air resistance and thus continuously loses energy. So the total energy of the rising ball at a certain height is greater than that of the falling ball at the same height. Since the potential energy at these two moments is the same, the kinetic energy and, therefore, the speed of the ball at a certain height is greater when it rises than when it falls. So the time of descent is greater than the time of ascent.

B154

We can cut the box along four edges and unfold it to make the Z-shaped figure shown in figure 9. This shape is then transformed into the desired square by cutting along the dashed lines. There are other solutions as well.

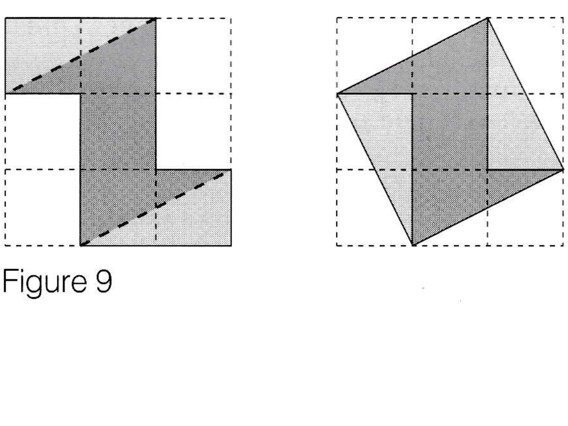


Figure 9

B155

The total number of games in a round-robin tournament with n participants is $T_n = n(n-1)/2$. If there were k participants apart from Judith and Nigel, we can write $T_k < 23 < T_{k+2}$. Since $T_6 = 15$, $T_7 = 21$, $T_8 = 28$, $T_9 = 36$, two values of k are possible: $k = 6$ and $k = 7$. In either case the number of unplayed games ($28 - 23$ or $36 - 23$) is odd. This means that Judith and Nigel haven't played each other—otherwise they'd both have an even number of unplayed games, each the same number.

Toy Store

1. The first player wins by putting the first coin at the center of the table and each subsequent coin such that it is symmetrical about this center to the coin placed immediately before that. This is always possible, because the position after each move of the first player is centrally symmetric.

2. The strategy described at the end of the article is equally applicable for any odd m with only this correction: the number four must be replaced by $m + 1$ and the intervals considered there must accordingly be replaced by $(m+1)k - m \le s \le (m+1)k$. It follows that the first player wins for all N with the exception of $N \equiv 0 \mod (2m+2)$,[1] when player A wins; $N \equiv 1 \mod (2m+2)$, when the second player wins; and $N \equiv (m+1) \mod (2m+2)$, when player B wins. The winner, whoever it is—player 1, 2, A, or B—plays the winning strategy from his or her first move, except for $N \equiv 2, 3, \ldots, m \mod (2m+2)$, when the first move of the winning (first) player is even.

The winning strategy for an even m is similar.

[1]The notation $x \equiv y \mod z$ means that $x - y$ is divisible by z.

All values of N are divided into intervals of the form $(m+2)k + 1 \le N \le (m+2)(k+1)$. If $N = (m+2)k + l$, where $2 \le l \le m + 1$, the first player starts with an *even* move that leaves a stock of $(m+2)k$ or $(m+2)k + 1$; if $N = (m+2)(k+1)$, the first player starts by taking *one* stone and in the next move takes an *odd* number of stones so as to leave a stock of the same size as above. Following this rule the first player will collect an even number of stones for any N except $N \equiv 1 \mod (m+2)$ and thus *forces a win*. In the exceptional case this strategy can be applied by the second player after any initial move of the first. So here *the second player wins*.

You can try to extend this analysis to the game in which player A wins by collecting an *odd* number of stones and player B by forcing an *even* score for A. The result will be similar but not the same.

Kaleidoscope

1. The rope will be stressed more in the second case.

2. None.

3. Yes, since the deformation of the board depends on where the forces are applied.

4. No, because the elongations of the wires will be different due to the different values of Young's modulus for iron and copper.

5. The relative elongation for the first wire is less by a factor of 4; the absolute elongation is less by a factor of 2.

6. In front of the opening, compression occurs, followed by stretching.

7. This is done so that the deformation of the springs does not exceed their elastic limits.

8. To decrease the force of the jerk when the fish is caught on the hook.

9. When a bullet passes through it, a plastic cup deforms, increasing in volume by an amount equal to the volume of the bullet. A glass cup is not capable of this, and under the force of pressure from the water it cracks.

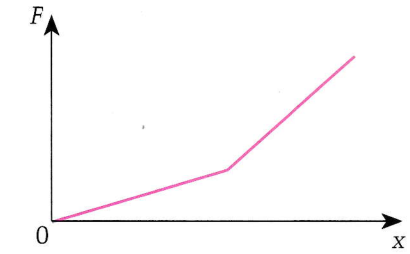


Figure 10

10. See figure 10.

11. The period is halved.

12. While the spring is stretched, body A falls with greater than free-fall acceleration; body B falls with less than free-fall acceleration.

13. Quartz has a low coefficient of linear expansion, so when the temperature changes the length of the rod hardly changes and no significant deformation occurs.

14. When the wire is heated, the mean distance between atoms increases and the attractive forces between them decrease.

15. The stress increases.

16. The internal energy of the body increases.

17. It turns into the internal energy of the solution.

18. The copper spring.

Microexperiment

Since the diameter of the tube's cross section decreases when the tube is stretched, the ring will fall.

Dragon curves

1. The length of the strip was smaller: 2^{30} cm = $1{,}024^3$ cm ≅ 10,700 km, whereas the distance to the Moon is 384,000 km.

2. The design is changed to its mirror reflection. In the code word the letters L are replaced with R's and vice versa So if we take a "dragon word" and exchange all the L's for R's and R's for L's, the new word will describe a similar dragon curve. That is, the two curves will have the same shape, regardless of their size or position.

3. An easy example is *LLLLL*, since a "dragon word" must have 2^{n-1} letters for some integer n. But even this is not sufficient. You can convince yourself that there is not a dragon curve of order 2 with the word *LLL*.

4. If w is the first word, then the second is $\overline{w}$ (see the statement of problem 5). This is true for any path; for dragon designs the transformation amounts simply to replacing the middle letter in w with the "opposite" (see problem 5c).

5. (a) Some examples will make this clear. (b) This is simply a restatement of theorem 1 in terms of words. (c) This is a restatement of theorem 2. We can write this result algebraically:

$$\overline{wL\overline{w}} = wR\overline{w}$$

and

$$\overline{wR\overline{w}} = wL\overline{w}.$$

Notice that a word must obey this condition in order to qualify as the word of some dragon curve. (d) An example will make this clear. Let's construct a word of order 4, given four arbitrary letters in positions 1, 2, 4, and 8:

RL_R_ _ _L_ _ _ _ _ _ _

Problem 5b shows that the letters in positions 7–15 must be the complements of those in positions 1–7. Hence we can fill in certain of them already:

RL_R_ _ _L_ _ _L_RL.

But problem 5b tells us that the letters in positions 1–7 also form a dragon word, so we can fill in more:

RL_R_ RLLRL_L_RL.

Going further with the result of problem 5b, we see that the first three letters must also form a dragon word. This allows us to fill in the remaining blanks:

RLLRRRLLRLLLRRL.

In fact, this construction is more easily done backwards, considering the first three letters, then the first seven, and so on. This effort is left to the reader.

6. In general, the two theorems yield different designs. (Consider, for instance, the design with the code word *RRLLRLLLRRLRRLL.*)

7. Assume that the line AB is horizontal. The number of segments in the turtle's path where it crawled to the right equals the number of segments where it crawled to the left, so their total number h is even. Similarly, the total number v of vertical segments in the path is even, too. But horizontal and vertical segments are alternating, and since their total number $h + v$ is even, we get $h = v$. So $h + v = 2h$ is divisible by four, the total time is an integer number of hours, and the last segment is vertical.

Lazy-day antidotes

(See the Kaleidoscope in the July/August issue)

Problems

1. Four coins. For example, the upper one, center one, and the two under the center one. Note that from every trio of coins that forms an angle, at least one coin must be removed, and one of them must be a vertex. But then it's necessary to remove the center coin as well.

2. There are no such numbers. Since the product is odd, it follows that all the numbers are odd, and the sum of four odd numbers is even.

3. 130°.

4. The middle finger.

5. Yes. For instance, a right triangle.

6. Three candles remained. The rest of them burned away.

7. Its area is 0, since 35 = 17 + 18.

8. The digit 5.

9. Four kilometers.

10. Six hundred kilometers.

11. The digit 0.

12. *Two Years Before the Mast* by Richard Henry Dana, *Twenty Years After* (a sequel to *The Three Musketeers*) by Alexandre Dumas, and *Twenty Thousand Leagues Under the Sea* by Jules Verne.

13. The triangle is an isosceles right triangle.

14. The angle remains equal to 1°.

Games

See the Toy Store article in this issue, "Winning Strategies," on page 61.

Winning strategies

A manual for the mathematical gambler

by Vladimir Dubrovsky

IN THE KALEIDOSCOPE ARTICLE in the previous issue of *Quantum*, five mathematical games were described. These sorts of games can just be played like any other games, but it's often more interesting to investigate them in order to find a winning strategy for one of the players. Although usually we find them in books on recreational mathematics, many of them are rather difficult to analyze and continually appear at math competitions. Some became the subject of serious research, and some remain yet unresolved. For all the diversity of mathematical games, there are only a few general methods of approaching them. We'll use our five Kaleidoscope games to illustrate some of these approaches.

All these games are played by two players who take turns making moves allowed by the rules and thus changing the current state (or position) of a certain, let's say, "object." Each player tries to bring the object into a state considered to be a winning one for this player. Our games have only a finite number of positions, necessarily come to an end after a finite number of moves,[1] and don't allow for draws. With this sort of game, we have only two logical possibilities: either the player who makes the first move (the *first player*) can win regardless of what the second player does, or there's a way to force a win for the second player. So the problem is to determine who wins and what the winning strategy is.

Game 1.[2] *The first player puts a white checker on any square of a chessboard, the second puts a black checker on any other square. Then they move their checkers horizontally or vertically, one square at a time. To win, a player must put a checker on the opponent's checker.*

If the players are in a peaceful mood, they can enjoy shifting their checkers as long as they wish. (Incidentally, this is impossible with the other four games.) Otherwise, the second player always wins. The strategy is very simple: initially, player 2 puts the black checker on any square diagonally adjacent to the initial square of player 1 (figure 1 shows two of the possible initial positions); then player 2 simply repeats the moves of player 1 (moves the checker in the same direction) until player 1 is forced to move a white checker onto the square adjacent to (alongside of) the black checker's square, when—bang!—the black checker jumps on the white one.

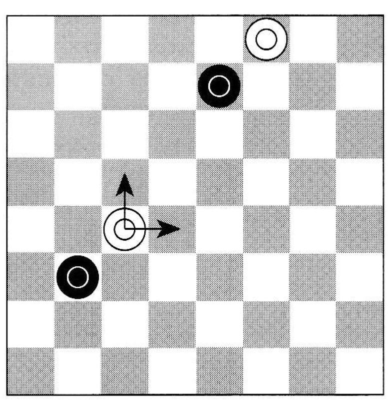


Figure 1

It's easy to see why this strategy is indeed a winning one. If the black checker was initially placed, say, "southwest" of the white one, they'll be in the same relative position after every exchange of moves. So player 1, to avoid an early defeat, will always have to go "north" or "east," which will inevitably drive that player's checker into a trap in the northeast corner of the chessboard (fig. 2).

This idea of repeating the opponent's moves can be regarded as a peculiar kind of *symmetric strategy*. These strategies are helpful in

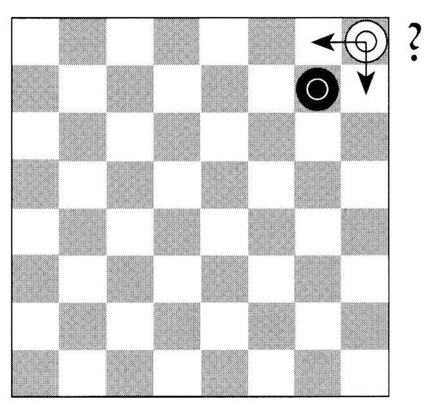


Figure 2

[1]For the first of the games below this is true only with a certain reservation.

[2]The order of the games here is different from that in the Kaleidoscope article.

various games. Here's a classic example of such a game.

Exercise 1. *Two players in turn place nickels on a rectangular table until there's no room for the next coin. The one who makes the last move wins. Which of the players can ensure a victory?*

Game 2. *Initially, there are two piles of nine candies each. Two players in turn move a candy from one of the piles to the other and eat two candies from either of the piles. The one who can't make a move loses.*

This is also a simple game to analyze. First, we notice that every single move decreases the number N of candies in the two piles by two, so N is always even. Then, it's clear that a move can always be made while $N \geq 4$. As for the case $N = 2$, there are two possibilities: (1) both "piles" consist of one candy; (2) one pile has two candies, the other is empty. The next move can be made only in case (1) and that would be the end of the game.

Now, the number N necessarily becomes equal to 2 after eight moves. But *each move changes the parities of the numbers of candies in both piles* (because one candy is shifted to the other pile). So after eight moves the parities are the same as they were initially—that is, both these numbers are *odd*. This means that only the first of the two cases for $N = 2$ is possible, and this allows for one more move—the ninth. So this game is, so to speak, counterfeit: it always takes exactly nine moves, and the last move is always made by the first player. So this player always wins, regardless of how the game was played!

Notice the idea of *invariant parity* used here. By the way, the winning strategy in the previous game implicitly used a similar idea: by choosing the initial square of the same color as the first player's, the second player ensures that the two checkers appear on squares of the same color after every exchange of moves. This makes it impossible for the first player to win even if the second player makes moves at random. In the "candy game" we can represent the numbers a and b of candies in the piles as the square (a, b) on the intersection of the ath column and bth row of a big chessboard. Then the ("chess") color of this square never changes during the game, so we can't get to squares (2, 0) and (0, 2) from square (9, 9), because it differs in color from the first two.

The next game is also, in a certain sense, "deceptive."

Game 3. *A chess knight is set on a corner square of a chessboard. Two players in turn mark a square such that any unmarked square can be reached by the knight according to the ordinary rules of chess without hitting the marked squares. The player who can't mark a square in this way loses.*

This is an example of a game that can be analyzed using the technique of *graphs*. Join the centers of any pair of squares that are one knight's move apart from each other with a line. This creates a set of 64 points (the centers) some of which are joined with lines (edges). Such objects are called graphs (a part of our graph is shown in figure 3). Initially, our graph is connected: each of its vertices can be reached along the edges from the knight's corner, and so any two vertices are connected by a path along edges. A move of the game is equivalent to erasing a vertex and all the edges issuing from it. The vertex must be chosen so as not to destroy connectedness of the graph. When can this be done? The answer is very short: always!

This is true for any connected graph. And here is a proof.

Define the distance between two vertices of a graph as the smallest number of edges in a path joining them (this number is well defined for any connected graph).

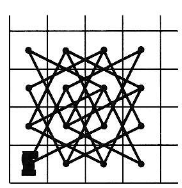


Figure 3

Fix a vertex K (in our case it's natural to choose the vertex that corresponds to the knight's location) and find the vertex A most distant from the fixed one (if there are several, take any of them). This vertex can be erased. Indeed, if after deleting vertex A and the edges issuing from it some vertex V becomes separated from K, then every path that joined K to V before we deleted A had to pass through A. But this means that V was farther from K than A had been, which contradicts the choice of A.

So this game, as well as the previous, always lasts a fixed number of moves—namely, 63—and always ends in a victory for the first player, regardless of what moves have been made, as long as they all were correct. However, as a game proper, it may be of some interest: if the time for a move is limited, it may simply become hard to find a correct move at a certain stage of the game.

The last two games are real interesting mathematical games. The first has some especially interesting theory behind it, which I'll only touch on. Both can be analyzed by means of another widely used method—reverse analysis, starting from the end of the game.

Game 4. *Two players in turn take stones from two piles. They are allowed to take either any number of stones from one pile or equally many stones from both piles at a time. The player who takes the last stone wins. (In the Kaleidoscope article the initial numbers of stones in the piles were 13 and 10.)*

This game was described by the Dutch mathematician W. A. Wythoff in 1907, and his name is often invoked when the game is discussed. He didn't know, however, that this is an ancient Chinese folk game, called *tsiangshitze* ("collecting stones"). Wythoff's paper was the first in a long series of works devoted to the theory of this game, which turned out to be surprisingly rich.

Each state of this game is described by a pair of nonnegative integers (n, m)—the numbers of stones

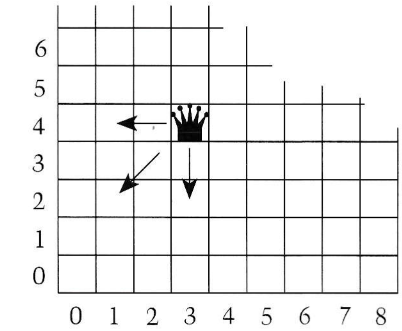


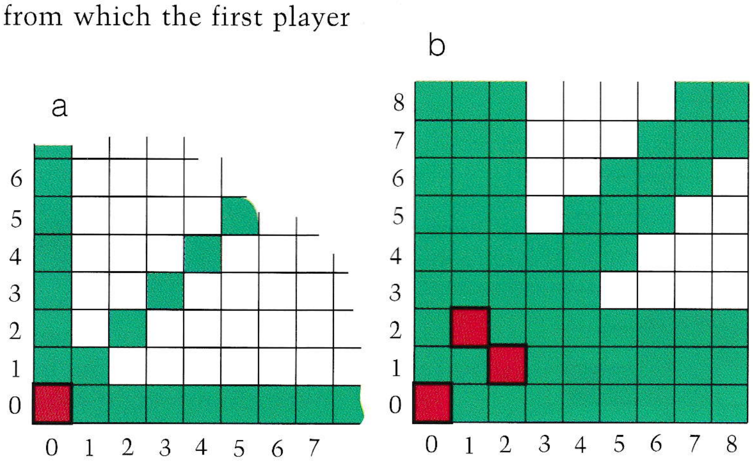
Figure 4

in the "first" and "second" pile. Of course, from a practical point of view there's no difference between the states (n, m) and (m, n), but for the graphic interpretation we'll use, it's more convenient not to unite them. This interpretation is quite natural: every pair (n, m) is in the usual way associated with a square in a "positive quadrant" of an infinite chessboard (fig. 4). Imagine that we put a chess queen on the square (n, m) corresponding to the current state of the game. If a number of stones are taken from the first pile, the queen goes the same number of squares to the left; taking stones from the second pile is equivalent to a downward move of the queen, and taking equal numbers from both piles moves the queen diagonally left and downward by the same number of squares. The player who first manages to put the queen on the (0, 0) corner square wins.

Since a queen, moving according to these rules, can't escape the corner, all squares (states) fall into two classes: *safe* squares, starting from which the first player can force a win, and *dangerous squares*, which bring victory to the second player.

Now, beginning from the end—that is, the final (0, 0) square—we'll scan the board and color the dangerous positions red and the safe positions green. The (0, 0) square is, of course, dangerous (actually, not just dangerous—it's disastrous: when your opponent reaches it, you lose). Then, any square that lies horizontally, vertically, or diagonally to the right of or above a dangerous square is safe, because starting at such a square you can put your opponent into a dangerous position in one move. Applied to (0, 0), this rule gives the coloring shown in figure 5a. Now we apply a second rule: color a square red if the file, rank, and diagonal through it are solid green below and to the left of it (any move from such a square creates a winning position for the opponent). So we color the new dangerous squares (2, 1) and (1, 2) red and color the squares from which these can be reached in one move green (by the first rule;—see figure 5b). Proceeding in the same way, sooner or later we'll get as far as we want; figure 5c illustrates the coloring of the 13 × 10 corner rectangle of the board that covers our particular initial state (13, 10). In fact, the color of this square becomes clear at the fourth step of the process, when the square (7, 4) = (13 – 6,10 – 6) is colored red. This shows that (13, 10) is a green, safe square—that is, the first player wins this game. The strategy is to make moves that lead to "red squares"—dangerous states. By our construction, this is always possible. For instance, in our particular case the first move can be made in three ways: by taking either 7 stones from the 13-stone pile, 2 stones from the other pile, or 6 stones from each of the piles.

The mathematically interesting part of this game is the sequence of "dangerous pairs" (1, 2), (3, 5), (4, 7), (6, 10), (8, 13), ... (by symmetry, it suffices to write out the pairs (n, m) with $n \le m$). The kth pair (a_k, b_k) in this sequence is given by the following unexpected relation: $a_k = [k\tau]$, $b_k = a_k + k$, where $\tau = (1 + \sqrt{5})/2$ is the famous "golden section." (See, for instance, "The Ancient Numbers π and τ" in the Kaleidoscope department of the January/February 1991 issue of *Quantum*.). Another description of this sequence can be given in terms of the "Fibonaccian number system," which was described in exercise 14 of I. M. Yaglom's "Number Systems" (in the last issue of *Quantum*): the numbers a_n are all those numbers whose Fibonaccian notation ends in an even number of zeros, and b_n is obtained from a_n by adding a zero at the end of the Fibonaccian notation of a_n. (See also exercise 15 in that article, where the numbers a_n, b_n are discussed from a different point of



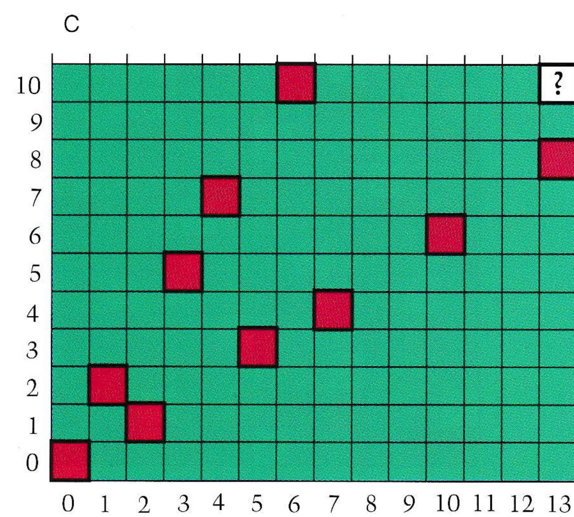


Figure 5

view.) All the pairs (a_n, b_n) can easily be written out using the following property: if (a, b) is a dangerous pair, then the pairs $(b - 1, a + b - 1)$ and $(b + 1, a + b + 2)$ are dangerous. I leave these remarkable properties as (not easy!) exercises.

Game 5. *There is one pile of stones, from which a player can remove one, two, or three stones in a move. The player who ends up with an even number of stones wins. The initial number of stones in the pile is $N = 25$.*

This game can also be modeled on a square grid and investigated with "reverse analysis." However, unlike the situation in Withoff's game, now we'll have to take into account how many stones are collected by each of the players. So it will be convenient to give names to the players—say, A and B—and denote by a and b the numbers of stones they've gathered by a given moment in time, respectively; these numbers will be called A's and B's (current) scores. Of course, the number s of stones left in the pile at the same moment is $s = N - (a + b)$, where N is the total number of stones; s will be called the (current) stock. Any current position is completely described by the corresponding pair of integers (a, b) and will be represented by the square (a, b) of the grid (fig. 6). The game starts at square (0, 0), proceeds in one-, two-, or three-square moves (A goes to the right, B upward), and stops at any of the terminal squares (a, b) with $a + b = N$. Terminal squares form the diagonal "line" joining squares $(0, N)$ and $(N, 0)$.

Notice that now we must provide for four rather than two (as in Wythoff's game) kinds of positions. Indeed, a position (a, b) with the scores of a and b of different parity is inequitable with respect to the players, because one of them must add an even and the other an odd number of stones to what they've already collected. Such a position can be favorable, say, to A regardless of whose turn it is to move. So we have to consider all of four logical possibilities: a position can be a winning one either for the first player (that is, whoever goes first, A or B, can force a win), for the second player, for player A (regardless of who goes first), or for player B. We'll color the corresponding squares green, red, white, or black, respectively.

So we start coloring from the end—that is, from the terminal row. The terminal squares (a, b) with even a ($(0, N)$, $(2, N - 2)$, ...) are left white; all the rest are colored black. Now A's goal can be described as "reaching a white terminal square," while player B is heading for a black terminal square.

It's interesting that this rule for defining a winner is equally applicable to the case of an even N, where the original rule is simply meaningless. In other words, if two players apply the rules of this game to an original stock of evenly many stones, then the final scores of both players will have the same parity. So in this case we can redefine the winner—say, by saying that A wins if each player has an even number of stones at the end and B wins if each has an odd number. In fact, both this rule and the original rule can be considered part of the same general rule, if we say that the winner of the game is A if A's final score is even, and it is B if A's final score is odd. Then it's not hard to see that our analysis applies just as well if N is even.

I leave it to the reader to extend the coloring—first to the "preterminal" diagonal row, then further down. The resulting pattern for $N = 17$ is shown in figure 6. Verify it and make sure that for $N = 25$ the square (0, 0) is also red—that is, *in our game the first player loses*. The strategy for A is: always go to a red or white square. Player B should step only on red and black squares.

Actually, the painting quickly becomes rather boring—you will notice that the pattern repeats itself (this is obvious in figure 6), which makes your work completely mechanical. So you may want to derive and prove a general rule for coloring any square directly, without building a painted road to it. For instance, any two squares in the same horizontal or vertical row eight squares apart ((a, b) and $(a + 8, b)$ or (a, b) and $(a, b + 8)$) or on a diagonal parallel to the terminal row two squares apart

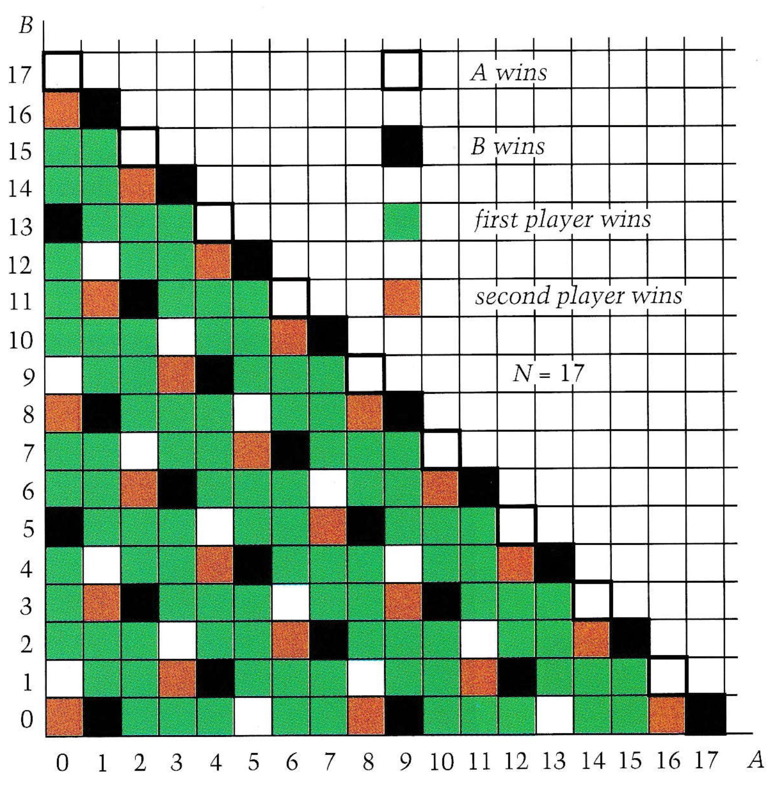


Figure 6

((a, b) and $(a + 2, b - 2)$) are always the same color.

All these regularities in the pattern boil down to the following observation.

If the game begins with an initial stock N in the interval $4k - 3 \le N \le 4k$, then the first player can score a number of the same parity as k. To achieve this, the first player should make only odd moves (take one or three stones) and always leave a stock divisible by four or greater than that by 1.

In other words, the first player must take one stone if $N = 4k - 3$ or $4k - 2$ and three stones for $N = 4k - 1$ or $4k$. This leaves a stock of $4(k - 1)$ or $4(k - 1) + 1$, so any reply of the second player brings the stock s into the interval $[4(k - 1) - 3, 4(k - 1)]$. Now the game starts anew, as it were, with $N = s$ and the first player making the first move again. Then another exchange of moves will bring s into the next interval of this form $[4(k - 2) - 3, 4(k - 2)]$, and so on. So there will be exactly k exchanges, in which the first player will score the sum of k odd numbers—a number of the same parity as k.

Our initial stock $N = 25$ can be represented as $4 \cdot 6 + 1$, so any first move leaves a stock s in the interval $4 \cdot 6 - 3 \le s \le 4 \cdot 6$. This means that the second player can apply the strategy described above to score an *even* number (because 6 is even) and force a win. The same conclusion is true for any N of the form $8l + 1 = 4 \cdot (2l) + 1$. The value $N = 8l$ is a winning one for player A (the rule for determining the winner in the case of an even N was described above), and $N = 8l + 4$ is a winning value for B. All other values of N are winning ones for the first player.

Exercise 2. *Classify all values of N according to the result of this game if the players are allowed to take as many as m stones, $1 \le m \le M$. Hint: consider separately the cases of odd and even m.*

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Readers write . . .

Stan Wagon, a professor of mathematics at Macalester College, writes:

On page 32 of the May/June issue of *Quantum* it is asked "how many angular minutes does the Earth rotate every minute?" The answer given is based on the assumption that the Earth spins 360 degrees on its axis in 24 hours.[1] But this is not true. In a "typical" day the Earth spins 361 degrees on its axis. The extra degree (approximately: this assumes 360 = 365) comes from the fact that the Earth has traveled a bit around the Sun, and so 360 degrees is not enough to bring us around to noon again. An extra degree is needed.

Things are a little more complicated on a daily basis, but since our 24-hour period, which we call a day, is based on a yearly average, it is in fact the case that the Earth spins (360 + (360/365)) degrees in one of our days. This means that, at New York's latitude, sunset is later each day beginning about December 7, not December 21.[2]

[1]The answer is correct to three significant figures.—*Ed.*

[2]Wagon, Stan. "Why December 21 is the Longest Day of the Year." *Mathematics Magazine*, 63 (1990), 307–311.