QUANTUM

MARCH/APRIL 1995

$5.95

SPRINGER INTERNATIONAL

La condition humaine (1933) by René Magritte

JUST AS IN MAGRITTE'S "THE BLANK SIGNATURE," which inaugurated Gallery Q in the May 1990 issue, there's no mystery about *what* is being depicted here. Once again, Magritte has fused the ordinary and the extraordinary, raising a flurry of questions with an "impossible" representation of everyday things.

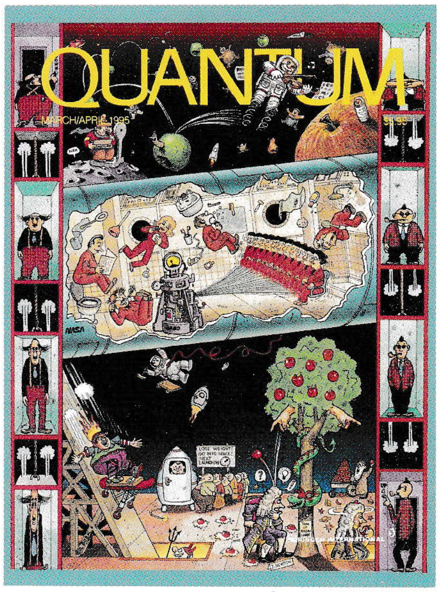
On one level, "La condition humaine" exposes the fallacy that art "captures a moment." One would have to be a fast worker indeed to produce the painting within the painting. Unlike a photographer, who truly works on the scale of fractions of seconds, a painter synthesizes impressions scattered in time. If we're deeply affected by this work, perhaps it's because it stirs a resonance in that part of us that cries out, in a moment of pure happiness: "Oh, moment, stay!" Such is the human condition: to be able to imagine timelessness and to be helplessly caught in the flow of time.

Our ability to imagine (or at least, to name) such absolutes as "infinity" and "perfection" sometimes impels us forward, sometimes leads us astray. On page 48, Gennady Myakishev wrestles with the concept of a "most inertial" frame of reference.

QUANTUM

MARCH/APRIL 1995

VOLUME 5, NUMBER 4



Cover art by Tomas Bunk

All we can say about our cover this time is: "No comment!"

Well, actually . . . we'd just like to point out that, although *Quantum* is a serious magazine full of important ideas, challenging problems, and sophisticated art, we are not above having a bit of fun. In fact, in every issue Tom Bunk has regaled us with his irreverent style of scientific illustration in the Physics Contest department. But this month, we offer our readers some April Tomfoolery right up front—on our hallowed cover.

Sir Isaac Newton, beaned by a legendary apple on our cover for the umpteenth time, would probably scowl at such horseplay. But we have a sneaking suspicion that Albert Einstein would have gotten a kick out of it. Then again, you don't have to be a genius to enjoy slapstick!

As for fooling around, we wonder if any of our readers will notice anything fishy in the Kaleidoscope . . .

FEATURES

4 Tools of the Trade
Nudging our way to a proof
by Galina Balk, Mark Balk, and Vladimir Boltyansky

10 The World in a Drop
The history of a fall
by Leonid Guryashkin and Albert Stasenko

16 Chance and Choice
Generalizing Monty's dilemma
by John P. Georges and Timothy V. Craine

22 Superconductivity Applied
Magnetic levitation comes of age
by Thomas D. Rossing and John R. Hull

DEPARTMENTS

2 Publisher's Page

9 Brainteasers

29 How Do You Figure?

30 In the Lab
Bobbing for knowledge

32 Kaleidoscope
The most mysterious shape of all

36 Physics Contest
Weighing an astronaut

39 Math Investigations
Distinct sums of twosomes

40 Mathematical Surprises
Lewis Carroll's sleepless nights

42 Innovators
The legacy of Norbert Wiener

48 In Your Head
The "most inertial" reference frame

49 At the Blackboard
Arithmetic on graph paper

53 Addendum
Hollow molecules

54 Happenings
Bulletin Board

55 Crisscross Science

56 Answers, Hints & Solutions

61 Toy Store
The last problem of the cube

PUBLISHER'S PAGE

Word and image

Hints on how to read Quantum

WHEN THE PREMIER ISSUE of *Quantum* appeared back in January 1990, many of our current readers were still in middle school. Chances are they didn't see that issue and so were denied the opportunity to read the very first Publisher's Page. There they would have learned where *Quantum* came from and what its aims are. They also missed the wonderful Letters from the Editors—words of welcome and advice from *Quantum*'s founding editors: Yuri Ossipyan, Sheldon Lee Glashow, and William P. Thurston.

For the sake of these readers, who are now finishing high school or completing their first year in college, I'd like to revisit these topics. The discussion may give them a deeper appreciation of *Quantum*, and it may calm any fears and frustrations they might have when reading it. As for the rest of our readers, they may find it interesting to hear what it is we think we're doing.

In 1989, the Russian-language *Kvant*, beloved at home and renowned abroad, had been around for almost 20 years, and there was nothing remotely like it in this country. It was a product of a Russian educational culture that encouraged top-notch scholars to teach and write for students at the pre-university level. The question was: would an English version work? I had been impressed by the seriousness of *Kvant*, and also by its whimsy and quirkiness. The illustrations were a perfect match, I thought, for the profound play that filled its pages. I was convinced that US students could handle such a magazine—that they *deserved* such high-quality material.

The founding editors knew that *Quantum* would make certain demands on its readers. Mathematician and Fields medalist William Thurston felt moved to "post a warning" at the outset:

> Science writing, and math writing in particular, tends to be dense and full of hazardous turns and treacherous sandpits. When I was a child I took pride in how many pages I read in an hour. In college I learned how foolish that was. When reading mathematics ten pages a day can be an extremely fast pace. Even one page a day can be quite fast. . . .
>
> *Quantum* articles aren't written like articles in scientific journals, but some of the same reading habits still apply. Don't be afraid to stop in midparagraph or midsentence when something surprises or puzzles you. Speed isn't the issue. Don't assume something is obvious just because an author treats it that way. What you work out on the side, even though it takes much more time, will have immensely more value than what you read straight through.

Thurston was saying: Slow down! But students—especially those who excel—are unlikely to take such advice. They like to go fast! One of the brilliant ideas embodied in *Kvant* is to intersperse the words and equations with illustrations that are attractive and unusual—sometimes even a little confusing! The goal is to do with art what you can't do by edict: slow the impatient reader down a bit. As a side benefit, you also enliven the numbing march of column after column of text, but that isn't the main purpose of the art.

The presence of high-quality art in *Quantum* is an outgrowth of our 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.

Thurston used a lovely image in his piece in that premier issue of *Quantum*. He wrote:

> With the modern emphases on test scores, on "basics," on mathematics as a competitive sport, on getting "ahead" in math, and so on, it often seems that the diversity, richness, liveliness, and depth of mathematics has been pruned away from the school experience. Mathematics isn't a palm tree, with a single long straight trunk covered with scratchy formulas. It's a banyan tree that has grown to the size of a forest, inviting us to climb and explore.

(I doubt that many of our readers have actually seen a banyan. You can find a picture of one in the 15th edition of the *Encyclopaedia Britannica*. It says there: "Aerial roots that develop from its branches descend and take root in the soil to become new trunks. The banyan reaches a height of 30 metres (100 feet) and spreads laterally indefinitely.")

As I reread the 1990 Letters from the Editors, I find it curious that both Thurston and Yuri Ossipyan, a physicist and former advisor to Soviet President Mikhail Gorbachev, were bored at school. (Palm trees rather than banyans in their classrooms?) "Whether you have already developed an interest in math and science," Ossipyan wrote, "or have gathered from school courses that these subjects are boring (as I did in my early teens), I hope *Quantum* will help you discover the excitement inherent in mathematics and the natural sciences." And Thurston says, "As a child, I often hated arithmetic and mathematics in school. Pages of exercises were tedious and dull. They weren't fun or challenging, they were just a chore." Without actively encouraging such behavior, I feel compelled to report that he would sometimes read books under the desk, or stare out the window, working through mental puzzles of his own devising!

I join Thurston in having "great hopes that *Quantum* will open up a road to some of the breadth, wonder, and excitement" of mathematics and the quantitative sciences. Physicist and Nobel laureate Sheldon Glashow wrote in January 1990: "I wish *Quantum* had been around when I was a student—it would have made it a lot easier for me to have found fulfillment as a physicist." But even if you have no intention of being a scientist or a mathematician, I believe *Quantum* can give readers of all ages great pleasure. You just have to earn it. It seems like a fair deal to me. As Ossipyan said, "Few experiences are as intensely exhilarating as the feeling 'I've got it!' that comes as a flash when you've solved an intricate problem or grasped a profound idea." It goes without saying that sometimes you *won't* get it. That's what makes it so much fun when you do.

—Bill G. Aldridge

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 5 (6 issues) will be published in 1994–1995. *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:

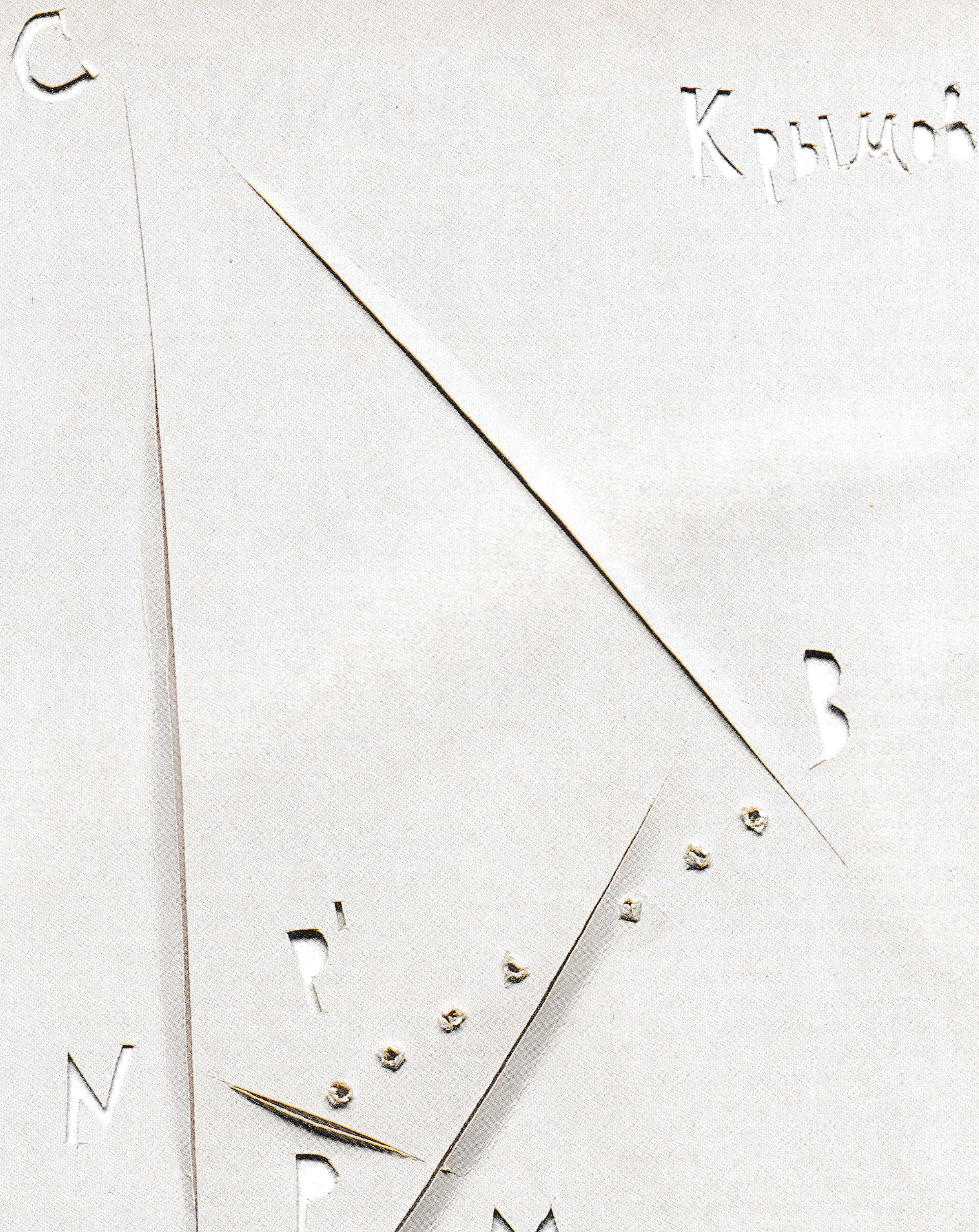
Advertising Representatives: (Washington) Paul Kuntzler (703) 243-7100; (New York) Brian Skepton (212) 460-1575; and G. Probst, Springer-Verlag GmbH & Co. KG, D-14191 Berlin, Germany, telephone (0) 30-82 07-1, telex 185 411.

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: $34; 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.


C
Крымов
B
P'
N
P
M
A

Nudging our way to a proof

Using the method of small perturbations

by Galina Balk, Mark Balk, and Vladimir Boltyansky

A MATHEMATICIAN TESTING geometrical statements for plausibility or trying to construct a counterexample often follows this line of reasoning: "I'll take the case where this statement is true and nudge one point (or segment, or some other figure). By doing this, can I obtain a case where the statement becomes invalid?"

As a first example, look at the following simple question.

Problem 1. A point P inside a convex polygon is orthogonally projected on its sides or their extensions. Let's say that such a projection is "pleasant" if it belongs to the corresponding side and "unpleasant" if it belongs to the side's extension. Is it true that at least two of these projections are always pleasant?

Solution. Consider first a point in a *triangle ABC*. If the triangle is acute, then all the projections of any of its interior points are pleasant. In an obtuse triangle ABC, it's easy to find a point with exactly two pleasant projections. (It can be chosen close to one of the acute angles—see figure 1.) Now we can construct a convex *quadrilateral* and a point on its perimeter that has only one pleasant projection. Take, for instance, the quadrilateral $MNCB$ in figure 2, where MN is an arbitrary line through P separating the projections of P on AB and AC from the side BC. Point P has only one pleasant projection with respect to this quadrilateral (point P itself). But it doesn't lie strictly inside the quadrilateral. So let's slightly "perturb" the diagram by moving point P slightly inside $MNCB$. If the shift is small, the projections of P will shift just a little, and we can make their displacements so small that the unpleasant projections will remain on the extensions of the corresponding sides—that is, they remain unpleasant. Thus, we come up with a point P' that has a single pleasant projection (on the side MN, to be exact).

Problem 2. It's well known that the altitudes of a triangle (or their extensions) meet at a point. Is it true that all four altitudes of an arbitrary tetrahedron (that is, the perpendiculars to its faces through their opposite vertices) also meet at a point?

Solution. Draw two altitudes AM and DN in an arbitrary tetrahedron $ABCD$ (fig. 3). Now let's perturb it by moving the vertex D into a new position D' in the same plane BCD. The altitude from A in the new tetrahedron is the same line AM (because the planes BCD and BCD' coincide). At the same time, the

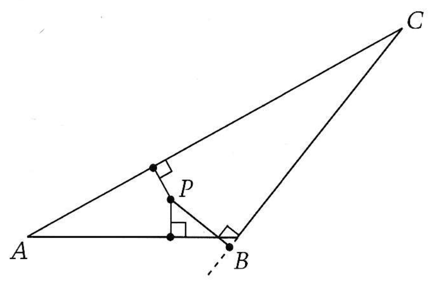


Figure 1

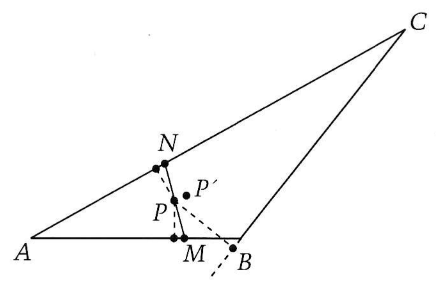


Figure 2

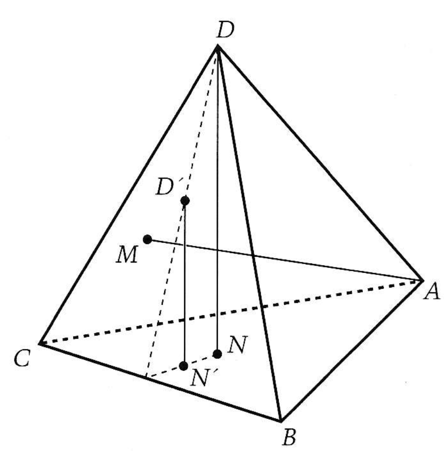


Figure 3

Art by Dmitry Krymov

altitude DN can be moved into any new position $D'N'$ parallel and close enough to the original one. Now it's clear that by properly moving point D we can make the altitudes AM and $D'N'$ of the new tetrahedron disjoint, even if they originally intersected. So the Altitude Theorem for triangles can't be extended to tetrahedrons (at least, not to all tetrahedrons).

Problem 3. (A. Kuzminykh) Does there exist a convex polyhedron whose orthogonal projection on any plane is a 1995-gon?

Solution. Instead of considering a fixed polyhedron and various planes on which it is projected, it would be more convenient to fix a plane α and vary the position of the polyhedron.

Suppose a polyhedron satisfying the requirement of the problem exists. Let's position it so that one of its edges, AB, is perpendicular to the plane α. Then the vertices A and B are projected onto the same vertex P_1 of the polyhedron's projection on α—a 1995-gon $P_1P_2...P_{1995}$. Now let's nudge the polyhedron so as to slant the edge AB and make its projection a new edge of the polygon in the plane α. If the perturbation is small enough, the different vertices $P_2, ..., P_{1995}$ will remain vertices (that is, they won't be absorbed into the body of the projection) and they'll remain different. This means that the new projection is at least a 1996-gon, which contradicts our assumption.

So a polyhedron with the requirements of the problem doesn't exist.

These simple examples suffice to show the essence of the "small perturbation" technique we applied. The properties of a geometric object fall into two classes: *stable* properties, which are preserved under any (sufficiently small) perturbations of the figure, and *unstable* properties, which can be violated under certain small perturbations. For instance, in problem 1, "to have its projection on the extension of a side" is a stable property of a point in a polygon, while "to lie on the boundary of a polygon" is an unstable property. In problem 3, the property "the projections of two vertices are distinct" is stable, while the property "the projections of two vertices coincide" is unstable. Classify the properties involved in problem 2 yourself.

Thus, in order to find a figure with a certain set of properties, we first try to achieve the desired *stable* properties and then perform a small perturbation that removes *unstable* properties that we don't need.

Now let's solve a more complicated problem.

Problem 4. Is it possible to make a hole in a regular tetrahedron that would let another congruent tetrahedron go all the way through it?

It will be better to reword it (we'll see below that both problems are equivalent):

Problem 4a. Is it possible to arrange two congruent regular tetrahedrons T and T' in space in such a way that the orthogonal projection of one of them on a given plane lies entirely inside the projection of the other tetrahedron on the same plane?

Solution. Attach the tetrahedron T' beneath the plane so that its face $\Delta = A'B'C'$ lies in the plane (fig. 4). Put the other tetrahedron above the plane with the projection of its face ABC exactly fitting Δ. Now we'll try to move T so as to squeeze its projection completely inside Δ. First, turn T about the edge AB until the edge CD becomes perpendicular to our plane (fig. 5). Then the projection of T becomes an isosceles triangle $A'B'C_1$ with C_1 inside Δ.

Next we make the projection of the edge AB shorter than $A'B' = AB$. This is done by slightly rotating T about the line through the midpoints M and K of AB and CD. The

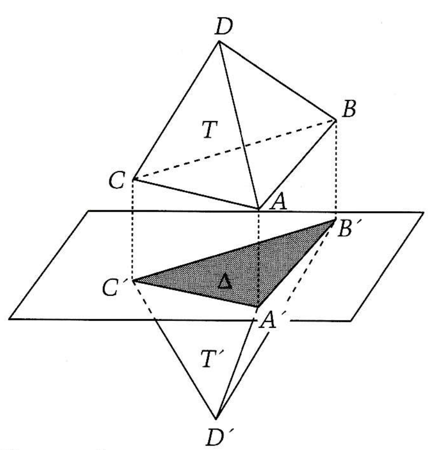


Figure 4

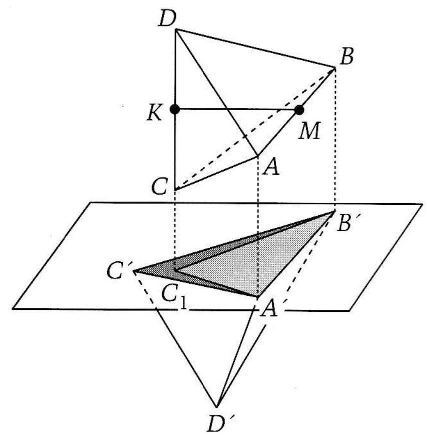


Figure 5

projection of T turns into an isosceles trapezoid $A_2B_2C_2D_2$ (fig. 6) whose base A_2B_2 lies on the segment $A'B'$. The other base, C_2D_2, will lie completely inside Δ for a sufficiently small angle of rotation.

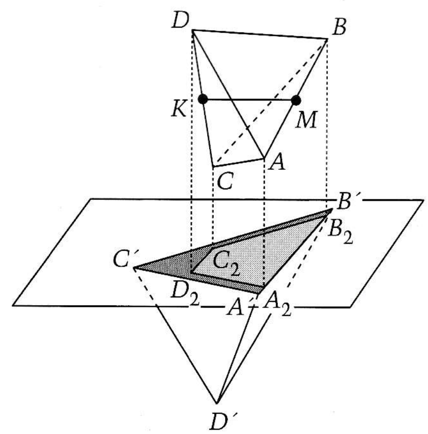


Figure 6

Finally, we can translate T by a sufficiently short distance along the line MK so as to bring the vertices A_2 and B_2 inside Δ without moving C_2 and D_2 out of Δ (fig. 7).

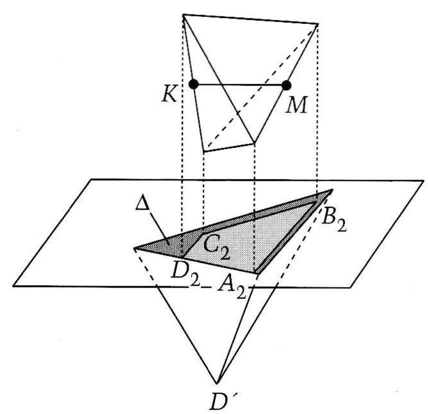


Figure 7

And this gives a positive answer to our auxiliary problem 4a.

Coming back to problem 4, it remains to note that if we remove the part of tetrahedron T' that consists of all the points directly beneath the projection of T on Δ, make the hole thus obtained just a little wider, and let tetrahedron T fall from the position we achieved (shown in figure 7), it will drop all way through this hole in T' without hindrance.

Another situation where small perturbations prove useful arises when you have to choose (or construct, or find) a figure from a given set of figures that is the "best" in one sense or another (has the smallest perimeter, the greatest area, and so on). The main difficulty in these problems is often to guess the right answer. And this is just where the method of small perturbations can help: we take any figure from the given set and try to improve it by a small perturbation. If we fail to do it, then it's quite plausible that the figure we've taken is the desired one. (Of course, the guess must be followed by a strict proof.)

Problem 5. Through a point P given inside an angle, draw a line that cuts the triangle with the smallest possible perimeter off the given angle.

Searching for a solution. The perimeter of a triangle ABC can be expressed in terms of the tangents drawn from, say, the vertex C to the circle touching side AB (on the outside) and the extensions of CA and CB (fig. 8). Using the fact that two tangents from the same point to the same circle are the same length, we derive that the perimeter $2s$ is equal to

$$\begin{aligned}2s &= CA + AB + BC\\ &= CA + AT + TB + BC\\ &= CA + AM + NB + BC\\ &= CM + CN\\ &= 2CM,\end{aligned}$$

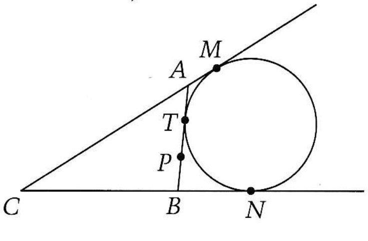


Figure 8

where T, M, and N are the points of tangency of the circle in question.[1] This circle is one of the three *escribed* circles—that is to say, one of the *excircles* of the triangle.

We can think of this triangle ABC as being cut off from the given angle (with vertex C) by a line AB through P. If we contract the circle toward C so that it remains inscribed in the given angle and correspondingly turn the line AB so that it remains a tangent to the circle, then the tangent length CM and thus the perimeter of $\triangle ABC$ will become smaller. This perturbation decreasing the perimeter is possible as long as point P remains outside the circle. As soon as point P gets on the circle, we're no longer able to continue the process. So it's quite plausible that the smallest perimeter is achieved for the triangle ABC whose excircle in $\angle ACB$ touches AB at P. Now an accurate solution can be derived without too much effort.

Draw any circle inscribed in the given angle. Let P' be the point of intersection with line CP that is closest to C. The dilation relative to center C by the factor CP/CP' takes this circle into a circle O through P inscribed in the given angle (with P on the arc facing vertex C—see figure 9). Let A and B be the points where the tangent to the circle O through P meets the sides of the angle. Then ABC is the required triangle.

Indeed, consider another triangle $A'B'C$ cut off from the angle by a line through P. Since line $A'B'$ is not tangent to the circle O, they have another point of intersection Q. So the excircle of $\triangle A'B'C$ is farther from the vertex C than the circle O; therefore, $\triangle A'B'C$ has a larger perimeter than $\triangle ABC$.

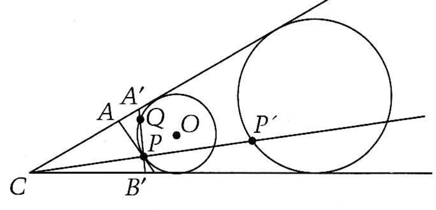


Figure 9

[1] This not-so-obvious relation is certainly the nicest point in the solution that follows and it can hardly be discovered by means of the small-perturbation method. However, once you become so lucky—or so smart—that the relation reveals itself to you, the method will help you use it properly.—*Ed.*

You can check through the solution once again to make sure that the triangle with the smallest perimeter is unique.

The method of small perturbations is useful in areas other than geometry, as the following example shows.

Problem 6. Three friends, Art, Billie, and Carmen, organized a chess tournament among themselves in which each played the same number of games with the others. After the tournament Art declared himself the winner because he had fewer defeats than the other two participants; Billie said she should be awarded first place because she had the greatest number of wins; and Carmen noted that her score was the highest (a player is given 1 point for a win, 1/2 point for a draw, zero for a defeat). Is it possible that all three friends were right, or did some of them miscalculate the results?

Solution. Let's try to construct a tournament satisfying all three conditions that the friends had formulated. It will be convenient to "assemble" it from separate *rounds*—that is, sets of three games between each of the three different pairs of players. A pair of rounds will be called a *type-A* (for Art) *double round* if (a) Art plays all four of his games to a draw and (b) Billie defeats Carmen in the first round and loses to Carmen in the second round. This double round, and any number of such rounds as well, satisfies Art's condition: he has the smallest number of defeats (no defeats at all). At the same time, in a type-A double round all three friends get the same score, Art has the smallest number of wins, and the other two players have the same number of wins and defeats (one win and one defeat). The same will be true for a tournament consisting of, say, 100 type-A double rounds.

Now let's perturb this big tournament by adding a comparatively small number of rounds (not necessarily type-A double rounds). Art's condition will remain true (because Art will have only a small number of defeats—no greater than the number of added rounds—while Billie and Carmen will have at least 100 defeats each). In other words, this condition is stable. Let's try to choose a perturbation that satisfies Billie's condition. Consider a type-C double round (in which Carmen achieves a draw in all games, while Art and Billie win once against each other). It retains the result of equal scores among all the players, but gives Billie an advantage over Carmen in the number of wins. So after 100 type-A double rounds and, say, 10 type-C double rounds, Art, Billie, and Carmen will have 10, 110, and 100 defeats, respectively, and 10, 110, and 100 wins, which means that both Art's and Billie's conditions are fulfilled. But the score of all the players is the same (220 points), so Carmen's statement isn't true yet.

So let's perturb this last tournament a bit more by adding another round (which certainly won't violate the first two conditions) chosen in such a way that Carmen gets the highest score. For instance, we can assume that in this additional round Carmen wins against Art and Billie, who play their game to a draw. This completes the construction of the required tournament.

Exercises

1. A quadrilateral *ABCD* has congruent and perpendicular diagonals and a pair of congruent opposite sides ($AB = CD$). Does it have to be a square?

2. In a convex pentagon all sides are congruent. Is this pentagon necessarily regular?

3. Does a convex pentagon with congruent diagonals have to be regular?

4. A convex hexagon has parallel opposite sides and congruent diagonals joining opposite vertices. Is it necessarily regular?

5. A quadrilateral circumscribed about a circle has congruent diagonals. Does it necessarily have parallel opposite sides? (Hint: start with a square!)

6. $2n$ points $A_1, \ldots, A_n, B_1, \ldots, B_n$ are given in the plane. Prove that these points can be moved an arbitrarily small distance so that no two of the segments $A_1B_1, \ldots, A_nB_n$ are parallel and no three of them have a common point.

7. Prove that any n points in space can be moved an arbitrarily small distance so that no four of them are in the same plane.

8. Prove that $n - 5$ vertices of a convex n-gon ($n > 5$) can be nudged so that no three of its diagonals intersect at one point.

9. Prove that the constant term a_n of the polynomial $x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n$ can be changed by an arbitrarily small amount so that there are no multiple roots. (A multiple root is a root corresponding to two or more identical factors of the polynomial.)

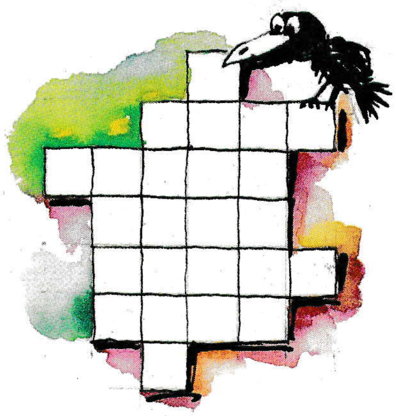
10. Draw a line through a point P in a given angle so as to cut off the triangle with the smallest possible area.

11. If a quadrilateral is convex, the sum of its diagonals is greater than its semiperimeter. Find a counterexample showing that the converse statement is wrong.

Just for the fun of it!

B136

Color logic. Four girls, Babs, Grace, Pam, and Winnie, are standing in a circle and chatting. The girl in the green dress (neither Babs nor Grace) is standing between the girl in the blue dress and Winnie. The girl in the white dress is standing between the girl in the pink dress and Grace. Which dress is each girl wearing?

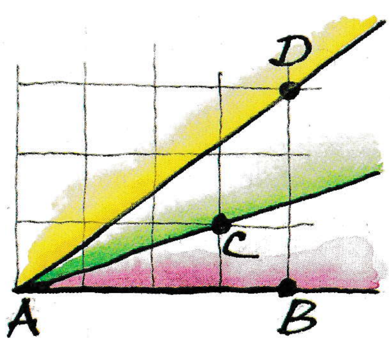


B137

Dissection on graph paper. The shape in the figure at left is to be divided into six congruent parts by cutting it along the grid lines. What shapes can the congruent parts have? (M. Koman)

B138

Run or walk. A group of hikers came upon a stream. A log stretched from one bank to the other. The first hiker started walking slowly across the log bridge and fell off, but managed to swim the rest of the way. Then the leader of the group ran across the log bridge. The rest of the group also ran across. Why is it better to run along the log rather than walk? (S. Krotov)



B139

Geometry on graph paper. Three rays AB, AC, and AD are drawn from a node A on graph paper as shown in the figure. Using the square grid, prove that the angles BAC and CAD are equal. (M. Koman)

B140

Redundant rook. Fifteen rooks are set on a chessboard such that there is at least one rook in each rank and file (that is, horizontal and vertical rows). Prove that it's possible to remove one rook so that the remaining rooks still satisfy the same property of no empty ranks or files. (V. Proizvolov)

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Art by Pavel Chernusky

The history of a fall

"You will look into our souls, and I beg that you remember that we are people too, that we love and we suffer like everyone else on the earth! I have given you the idea of the play. Now you be the judge of how well it is worked out. And now, let us begin!"
—Prologue to Leoncavallo's opera I Pagliacci (The Clowns)

by Leonid Guryashkin and Albert Stasenko

HERE BEGINS OUR LITTLE tragedy, in six acts with an epilogue.

Act 1: Not an easy birth

It would seem that there is nothing easier in the world than to make a drop of nitrogen. You just pour some liquid nitrogen into a tin can with a small hole in the bottom and drops will fall out one by one like water from a leaky faucet. Well, we tried it, and it didn't work—the nitrogen poured out in a thin stream, sliding along the layer of nitrogen vapor that had formed on the walls of the tin can. You see, these walls—which were pretty much at room temperature (T = 300 K)—were like a white-hot oven compared to the liquid nitrogen (T' = 77 K). We could wait until a fur coat of condensed water vapor and carbon dioxide from the air grew on the outside of the can, but in the meantime too much nitrogen will have poured out. To prevent that, we could have invented some kind of valve. Or we could have . . . Well, what we did was put a vertical tube, about 1 cm long, in the hole and filled it with a porous substance that quickly assumed the temperature of liquid nitrogen and let the liquid nitrogen flow through it with an average speed of two drops per second (fig. 1).

Falling freely through the air with an initial diameter of about 2 mm, these drops passed between a strobe light and some photographic film, leaving shadow pictures on the film. The duration of the light flashes was very short—just one-millionth of a second. This explains why the picture looks quite clear—during such a short period of time the drop itself and the picture of its streamlined flow changed hardly at all.

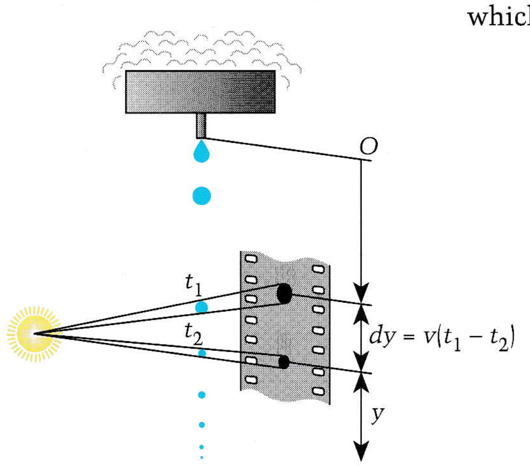


Figure 1

Figure 2 (on page 12) shows one of many pictures of nitrogen drops, and figure 3 shows a "colorized" version. We can clearly see the drop's "tail," which is due to the sharp difference in density between the surrounding air and the cold mass of gaseous nitrogen evaporating from the surface of the drop. (A similar tail can be seen with the proper lighting when you dissolve a lump of sugar in a glass of water.)

Experiments were also conducted with drops of water of the same size—absolutely no tails were left on the film by the drops. The reason is clear. The temperatures of air and water are almost the

Art by Vera Khlebnikova

Act 1. Not an easy birth

It seems to be nothing more simple than to make a nitrogen drop: just pour some fluid nitrogen into a tin with a small hole, and drops will fall one by one as water out of a holey tinware. There was an attempt but... nothing of that kind: nitrogen poured out by a thin stream, sliding along the layer of its own evaporation formed at the walls of the tin. You see, these walls with almost room temperature (T ≈ 300K), was «felt» by nitrogen [illegible] 77K) as a wh[illegible] oven. We co[illegible] till a fur-co[illegible] condensed [illegible] water and [illegible] acid gas from [illegible] will grow on the outer surface of a tin, but meanwhile too much nitrogen would pour out; for nitrogen not to pour out one could invent some kind of a value, or one could... But we fitted the hole up with a vertical tube, about 1 cm long (figure 1) and filled with some porous substance which assumed the temperature of fluid nitrogen and transmitted fluid nitrogen with the average speed of 2 drops per second.

Falling free in the air, these drops with initial diameter about 2 mm, flew past verti- the picture looked quite clear — during such a small period of time the drop itself and the picture of its stream-line flow hardly changed.

Figure 2,a shows one of many pictures of nitrogen drops, and figure 2,b — the result of its processing in the arbitrary colors by a computer, w[illegible] can «paint» the b[illegible] white photogra[illegible] cording to experim[illegible] talist's design. On these pictures drawing one can clearly see the trace after a drop; it is formed due to sharp [illegible] density [illegible] ding [illegible] lass [illegible] gaseous nitrogen, which evaporates from the surface of a drop (an analogous trace can be seen in the outcoming light beamed through a glass of water with a lump of dissolving sugar piece being sunk on an edge of a spoon).

There were [illegible] ments with th[illegible] drops of water of [illegible] same size — there [illegible] e no traces left by [illegible] p on the film. The [illegible] on is clear. The [illegible] ra-tures of air a[illegible] are almost the[illegible] and are far fro[illegible] boiling point, so [illegible] velocity of evaporat[illegible] of a drop of water a[illegible] room temperature is very small. While in damp air the drops don't evaporate at all, and more of that, they may even grow due to condensation (remember how long the drops of dew or rain hang on the trees and grass, and a drop of water on a saucer can be easily observed). It should be noted that air is an alien «gas» for a drop of water, while it is almost a bosomfriend for a nitrogen drop (because nitrogen is one of the basic component of the air). So, what ca[illegible] learned about [illegible] of a nitrogen d[illegible] its «portrait»? [illegible] «portrait» m[illegible] sible to m[illegible] size of a [illegible] radius [illegible] [illegible] ρ_0

(ρ_0 = 800 kg/m^3 is the density of liquid nitrogen) and the area of its surface $s = 4\pi r^2$. Second, as the period Δt between the flashes is known, we can easily find its velocity $v = \Delta y/\Delta t$ from the shift of successive pictures of the same drop. Then, changing the height of a dr[illegible] one can find [illegible] mentally r[illegible] velocity o[illegible] functions [illegible] the fall[illegible].

The r[illegible] ee urement[illegible] own on figure 3 b[illegible] es. One can see the scattering of experimental points, which result from the limited accuracy of measurements in any experiment (the errors [illegible] velocity v are shown [illegible] The [illegible] w, [illegible] op [illegible] o-no[illegible] e-locity [illegible] grows, then it decreases after passing through a maximum point.

Now we want to explain theoretically the results obtained, which will enable us to describe the characteristic periods in the life of a falling drop.

It's clear that the force of air resistance acting upon a moving body depends on its relative velocity. At least it is absent at zero velocity, and while the newborn drop still moves slowly, this force of air resistance can be neglected. Then its movement is influenced only by the force of gravity, so the drop will be uniformly accelerated, and every a schoolchild is an expert in such a motion:

$$v^{(0)} = gt = \sqrt{2gy^{(0)}},$$

$$y^{(0)} = \frac{gt^2}{2}$$

The upper zero here indicates that the estimate corresponds to zero force of resistance. In figure 3 this dependence of velocity from height is shown by dot-and-dash line [illegible] ibes the [illegible] rve of [illegible] section OA.

Act 2. Inevitable Braking

As the drop accelerates, the air resistance becomes m re and more appreciable. How this force can be described?

Many a time it was shown that the force of air resistance affecting the moving body (i.e. aerodynamic force) is proportional to air density, to square of its size (or the area of its cross section) and to square of velocity:

$$F_a = -C_a \rho \pi r^2 v^2$$

This formula can be obtained from the dimensional analysis. The sign minus accounts for the force being directed opposite to velocity, whereas the dimensionless coefficient C_a, broadly speaking, depends on velocity, but this dependence is

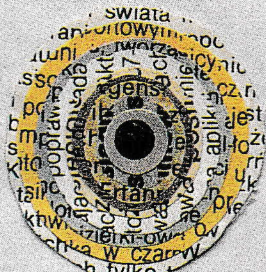


Figure 2

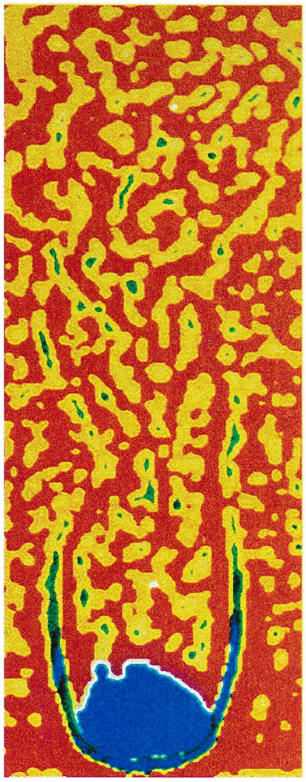


Figure 3

same and are far from the boiling point, so the evaporation rate for a drop of water at room temperature is very small. And in humid weather the drops don't evaporate at all—in fact, they may even grow due to condensation. (You've probably noticed how long the drops of dew or rain hang on the trees and grass, and you could take your own sweet time observing a drop of water on a microscope slide.) It should be noted that air is an alien "gas" for a drop of water, while it is almost a bosom buddy for a nitrogen drop (because nitrogen is one of the basic components of air).

So, what can we learn about the life of a nitrogen drop from its "portrait"? First, the portrait makes it possible to measure the size of the drop (its average radius r), and consequently its mass $m = (4/3)\pi r^3 \rho_0$ ($\rho_0 = 800\ \text{kg/m}^3$ is the density of liquid nitrogen) and its surface area $s = 4\pi r^2$. Second, since we know the period Δt between flashes of the strobe, we can easily find the drop's velocity $v = \Delta y/\Delta t$ from the distance between successive pictures of the same drop. And by changing the height of the dropper, we can determine the radius and the velocity of the drop experimentally as functions of y during the fall: $r(y)$ and $v(y)$.

The results of measurements are shown in figure 4 as little circles. We can see the scattering of experimental points, which results from the limited accuracy of measurements in any experiment (the errors in the velocity v are shown by horizontal bars). The measurements show that the size of a drop decreases monotonically, whereas its velocity increases at first, then decreases after passing through a maximum.

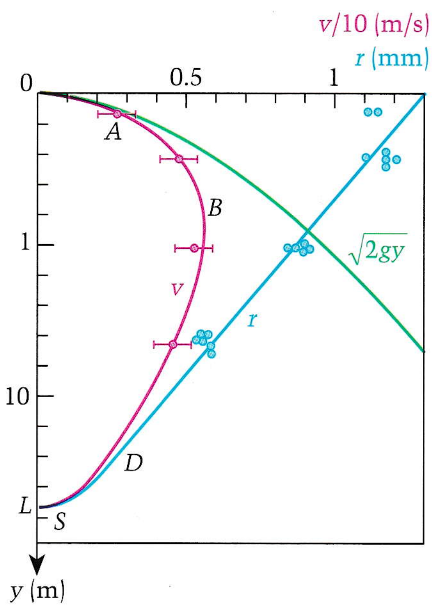


Figure 4

Now it's time to theoretically explain the results obtained. This will enable us to describe the characteristic periods in the life of a falling drop.

Act 2: Nonstop acceleration

It's clear that the air resistance acting on a moving body depends on its relative velocity. In any case, it's absent at zero velocity, and while the newborn drop is still moving slowly, the air resistance can be neglected. Therefore, its movement is influenced only by the force of gravity, so the drop will be uniformly accelerated, and you know all about that kind of motion (we hope!):

$$v^{(0)} = gt = \sqrt{2gy^{(0)}},$$

$$y^{(0)} = \frac{gt^2}{2}.$$

The superscript zero here indicates that this estimate is based on zero air resistance. In figure 4 this dependence of velocity on height is shown by the green line, which describes the experimental curve of the initial section OA.

Act 3: Inevitable braking

As the drop accelerates, the air resistance becomes more and more appreciable. How can this force be described?

Over and over you've heard it said that the air resistance affecting a moving body (that is, the aerodynamic force) is proportional to the density of air, the square of the object's size (or its cross-sectional area), and the square of its velocity:

$$F_a = -C_a \rho \pi r^2 v^2.$$

This formula can be obtained from dimensional analysis. The minus sign accounts for the force being directed opposite to the velocity; the dimensionless coefficient C_a, broadly speaking, depends on velocity, but this dependence is weak and can be obtained from experiments or other more complicated theoretical considerations.

We could also apply the following reasoning. The flow of an air mass moving inside a cylinder whose cross section is equal to that of the round drop (πr^2) is $\rho v \pi r^2$. But a unit of mass carries the momentum $mv/m = v$, so the total flow of momentum in this cylinder is $\rho v \pi r^2 v = \rho \pi r^2 v^2$. Of course, the flow lines of gas curve somewhat as they flow around the drop, so a catch-all factor C_a appears in the formula, accounting for every feature of the flow we can't consider here.

So, taking this aerodynamic force into account, the equation for the drop's movement (Newton's second law) can be written in the form

$$m\frac{dv}{dt} = C_a mg - \rho\pi r^2 v^2. \quad (1)$$

Dividing both sides of the equation by the mass of a drop $m = (4/3)\pi r^3 \rho_0$ yields

$$\frac{dv}{dt} = g - \frac{3}{4}C_a \frac{\rho}{\rho_0}\frac{v^2}{r} \quad (2)$$

—that is, the acceleration of a drop consists of two parts: the acceleration due to gravity g and the negative acceleration (braking) associated with the aerodynamic force.

We can write equation (2) in another way. Let's recall that the time interval dt can be obtained by dividing the distance traveled dy by the velocity v: $dt = dy/v$. Thus,

$$\frac{dv}{dt} = v\frac{dv}{dy} = \frac{d}{dy}\left(\frac{v^2}{2}\right).$$

But this is the change in kinetic energy per unit mass with distance. So equation (2) can be written as

$$d\left(\frac{v^2}{2}\right) = \left(g - \frac{3}{4}C_a \frac{\rho}{\rho_0}\frac{v^2}{r(y)}\right)dy, \quad (3)$$

and it seems that this would be a snap to integrate. But now is just the time to remember that the size of a drop also changes with distance. So as not to forget it, we've even written $r(y)$ on the right-hand side of the equation. The last equation can be read in this way: the change in the specific kinetic energy of a drop is due to the work performed by the force of gravity and the aerodynamic force over an elementary displacement dy.

Act 4: Continuous self-sacrifice and the pinnacle of glory

It's clear that the radius of a drop is nonnegative and doesn't increase with time. This means that by increasing the velocity and decreasing $r(y)$, the second term on the right side of equation (3) may become equal in absolute value to the first term, and at this point in the drop's trajectory the velocity will stop changing: $dv/dy = 0$. This point is marked with the letter B in figure 4. Now we can determine the value of the coefficient for the air resistance C_a, which was unknown when we obtained the aerodynamic force by dimensional analysis:

$$0 = g - \frac{3}{4}C_a \frac{\rho}{\rho_0}\frac{v_B^2}{r_B}$$

$$\Downarrow$$

$$C_a = \frac{4}{3}g\frac{\rho}{\rho_0}\frac{r_B}{v_B^2}.$$

Inserting the values $\rho_0 = 800$ kg/m^3, $\rho \cong 1$ kg/m^3, and $g \cong 10$ m/s^2, and taking the values for the velocity and radius from figure 3—$r_B \cong 0.9$ mm $= 9 \cdot 10^{-4}$ m, $v_B \cong 5$ m/s—we get $C_a \cong 0.2$.

What happens after point B? Well, the velocity of the drop must decrease, and maybe you've already guessed why. The force of gravity is proportional to the drop's mass—that is, to the cube of its radius (r^3), whereas the aerodynamic force is proportional to the drop's cross-sectional area—that is, to the square of its radius (r^2). So the force of gravity decreases faster than the aerodynamic force as the radius decreases. This means that after point B the drop slows down.

Now we should discuss the change in the drop's radius. First of all, the reason for its decrease is absolutely clear: the thermal energy of the air flowing around the drop evaporates its outer layers. Let's turn this thought into some formulas.

First let's determine the energy transferred to the drop. The air mass flowing past the drop is about $\rho v\pi r^2$. It is this mass of air that continuously transfers thermal energy to the drop. A unit mass of any substance carries energy consisting of the kinetic energy of chaotically moving molecules and the potential energy of their interaction. If we assume the surrounding air to be an ideal gas, its energy per unit mass will be c_pT (the energy of molecular interaction is zero for an ideal gas; the subscript p on the heat capacity indicates that the pressure is held constant). A unit mass of gas just after it evaporates from the drop, characterized by the temperature T', has energy c_pT', while in the condensed state (that is, in a liquid drop) the corresponding value is $c_pT' - L(T')$, where $L(T')$ is the heat of vaporization, which includes the potential energy of molecular interaction.

Figure 5 shows the qualitative dependence of the energy per unit mass on temperature for gaseous nitrogen (the straight line c_pT corresponds to an ideal gas) and also for nitrogen in the liquid state (the bottom curve). If necessary, these curves can be drawn from empirical data. But for now it's enough to have the values of c_p and L shown in the figure, which will prove useful for numerical estimates.

So, during the time dt thermal energy of about $\rho v\pi r^2 c_p(T - T')$ is transmitted to the drop. This energy is spent on evaporating a mass dm, which can be written as

$$-L\,dm = C_Q\rho v\pi r^2 c_p(T - T')dt. \quad (4)$$

This is the same approach we used earlier to estimate the flow of momentum (that is, the aerodynamic force F_a): we introduce a dimensionless coefficient C_Q, whose exact value we don't know, but surely it "conceals" fine details of the process that aren't essential for us at present).

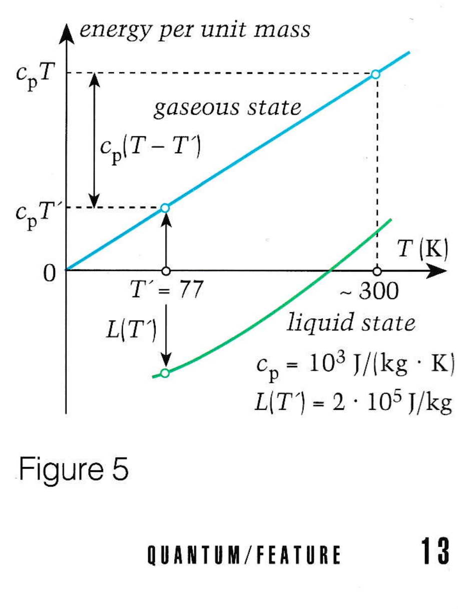


Figure 5

Substituting the formulas $dt = dy/v$ and $dm = 4\pi r^2 \rho_0 dr$ in equation (4) yields

$$\frac{dr}{dy} = -\frac{C_Q}{4}\frac{\rho}{\rho_0}\frac{c_p(T - T')}{L}. \quad (5)$$

Thus, the rate of change of the radius relative to distance is constant, which is confirmed by the straight line in figure 4. Measuring its slope, we find the dimensionless coefficient C_Q. As a matter of fact, that was the point of the experiment: we needed to make our theoretical considerations more exact—otherwise we need not have bothered! The solid curves $r(y)$ and $v(y)$ in figure 4 were drawn to create the best fit with the experimental data.

Act 5: A creeping life

The aforementioned equations describe the dynamics of a drop's heat exchange and mass exchange for a great part of its trajectory, but not all of it. Beginning from some point (denoted by D in figure 4), the laws governing the drop's life vary. Its further biography could be told in greater detail and with more exactitude than previously, but it would require that we introduce such complicated concepts as coefficients of viscosity (η) and thermal conductivity (k). We could also pass over this portion of the drop's life in discreet silence, citing the fact that every hero's life must have its mysterious periods. As a compromise between these two possibilities we'll give a mere outline of the main events of this period.

The most important fact is that after point D the drop becomes so light and small that the value $m(dv/dt)$ on the left-hand side of the equation stops playing any noticeable role, whereas the force of gravity mg is balanced by air resistance. But the latter also varies, and now it becomes proportional to the radius and velocity, and also to one more value—the viscosity coefficient. This new force $F_\eta = 6\pi\eta rv$ is known as the Stokes force, and the motion is referred to as "creeping," because it resembles the slow sinking of a pellet in honey.

So, after point D we get

$$0 = -6\pi\eta rv + mg,$$

from which we obtain

$$v = \frac{2}{9}\frac{\rho_0 g}{\eta} r^2. \quad (6)$$

Here is another point where we need to stop and think. You see, judging from figure 4, the drop's velocity continues to decrease, so $dv/dy \neq 0$ and $dv/dt = v(dv/dy) \neq 0$, and the mass also isn't zero yet—but we've written 0 instead of $m(dv/dt)$. What's up?

Let's recall that, if we really wind up, we can throw a stone pretty far, while the fluff from a dandelion wouldn't go very far despite our greatest efforts. The fluff immediately slows to zero velocity (provided there's no wind). Any physicist will tell you that the relaxation time for the fluff's velocity to equal the air velocity is very small due to its negligible mass (which in turn is the measure of inertia). A gust of wind would take this fluff away, while some period of time is needed to accelerate a sailboat. Likewise, when our drop becomes small enough, it will very quickly "get used to" the changing conditions of the flow and adapt to them almost without a time lag. (These conditions vary only because of the decrease in the drop's mass—otherwise the drop would fall with a constant velocity.) All this goes to show why the inertial properties of a drop become irrelevant at the end of its trajectory, and this allows us to drop the term $m(dv/dt)$ from the equation.

After point D the addition of heat to the drop will be determined basically by the process of heat transfer. In a time dt the drop will acquire thermal energy of about $4\pi r^2 k[(T - T')/r]$, which will be spent on evaporating a mass dm—that is,

$$-L\frac{dm}{dt} \sim 4\pi r^2 k\frac{T - T'}{r},$$

which leads to

$$r\frac{dr}{dt} = \frac{d}{dt}\left(\frac{r^2}{2}\right) = -\frac{k(T - T')}{L\rho_0} = \text{const},$$

$$r_D^2 - r^2 \sim t - t_D.$$

(Here k is the aforementioned coefficient of thermal conductivity.)

This law, which says that the surface area of the evaporating drop decreases proportionally with time, was formulated in 1883 by the Russian scientist B. I. Sreznevsky on the basis of dimensional analysis as well as many experiments with stationary drops of different substances. Comparing the last equation with equation (6) shows that the velocity of the drop decreases linearly with time, which corresponds to a quadratic law for the drop's coordinate (much like what takes place with uniformly accelerated motion): $v \sim \sqrt{y_L - y}$, where y_L is the point of complete evaporation (if further motion were described by this law, needless to say). So then $r \sim \sqrt[4]{y_L - y}$.

"Hey, wait a minute!" the thoughtful reader will say. "The drop's velocity has to be zero at the very beginning. So why didn't we pay attention to that in Act 2?" Well, it was okay, we reply, because as soon as the drop left the dropper, its velocity was already greater than zero, while it was so large that the mode of creeping motion in air wasn't observed from the very moment of its birth. Also, still left open is the delicate question of the influence of the close proximity of the dropper itself (in our reasoning we assumed all the bodies were "infinitely" distant from an isolated drop).

Act 6: The last microsecond

And now the pivotal moment in the drop's life arrives, when its size becomes equal to and then smaller than the mean free path l of the air molecules (that is, the mean distance between successive collisions of the molecules), which is approximately equal to 10^{-7} m. This means that now the drop is immersed not in a continuous medium but a rarefied one, which modifies the rules of the drop's life. Speaking of which, how much longer does the drop have to live?

At this point approximately $(1/6)n\langle u\rangle$ molecules strike a unit surface per unit time (where n is the density of the molecules and $\langle u\rangle = \sqrt{8RT/\pi M}$ is their average velocity). Each molecule carries energy of the order of $(3/2)kT$ (actually, $(5/2)kT$ if one recalls that the air consists of diatomic molecules—but this correction isn't crucial for our order-of-magnitude estimate). Thus, the entire spherical surface of the drop acquires per unit time an energy of

$$4\pi r^2 \frac{1}{6}\rho\langle u\rangle \frac{5RT}{2M}.$$

Making this energy equal to (in order of magnitude, as you might have guessed) to

$$-L\frac{dm}{dt} = -L4\pi r^2\rho_0\frac{dr}{dt},$$

we get the approximation

$$\frac{dr}{dt} \cong -\frac{5RT\rho\langle u\rangle}{12L\rho_0 M}.$$

It's true that some of the energy is carried away by the evaporating mass, and that the value for the latent heat of vaporization (given in reference books) should be reduced by the value of the work per unit mass performed against the pressure of the continuous medium (which is now absent)—but these two fine details wouldn't change the order of magnitude of the time we seek: τ_s. It's clear that the rate of evaporation is constant—therefore, r decreases proportionally with time, and

$$\frac{dr}{dt} = \frac{r_s}{\tau_s} = \frac{l}{\tau_s},$$

where r_s is the drop's initial radius at this stage of its life in the rarefied gas. So, the drop's disappearance time ($r \to 0$) is of the order of

$$\tau_s \sim \frac{r_s}{\frac{5}{12}\frac{RT}{LM}\frac{\rho}{\rho_0}\langle u\rangle}.$$

Inserting here $r_s = l \cong 10^{-7}$ m, $T \cong 300$ K, $M = 29 \cdot 10^{-3}$ kg/mol, and $\langle u\rangle \cong 500$ m/s, we get $\tau < 10^{-6}$ s = 1 microsecond. Alas, the last period of a drop's life turns out to be very short indeed. When time runs out, nothing is left of the almost stationary drop but one last molecule, and who can say which it belonged to—the drop or the air? (Perhaps the mystery could be solved by incorporating an atom of a radioactive isotope of nitrogen into the molecule.)

Well, the time has come to lower the curtain. And as the audience leaves, deeply affected by the tragedy, it has much to mull over. Was it an accurate portrayal? And if so, what was the point?

Epilogue

It goes without saying that our play, *The History of a Fall*, didn't cover everything. For example, as we mentioned above, one might examine such processes as the molecular transfer of mass (diffusion), momentum (viscosity), and energy (heat transfer). To do this, one needs the corresponding coefficients to describe the portion of the drop's trajectory from point D to point S, after which the mode of free molecular flow begins. Further, the influence of atmospheric oxygen wasn't discussed at all. One can expect that this "alien" gas (from the nitrogen drop's point of view), like any other substance, would try to diffuse into where it wasn't—that is, inside the drop. In any case, it does meddle in some way in the process of nitrogen evaporation. Another point wasn't taken into account: the drop isn't necessarily spherical—it oscillates during its fall, as you can see in figure 6. Also, we considered the drop's temperature at any period in its life to be equal to that of the saturated vapor at atmospheric pressure over a flat surface of liquid nitrogen—and this may well be incorrect.

Finally, who on earth would be interested in this drop and its tragic story? Well, sometimes it's necessary to know the behavior of

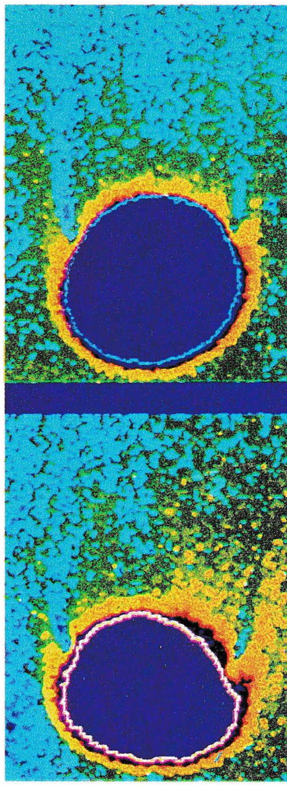


Figure 6

individual drops and entire clouds of drops. For example, if you need to visualize the flow around a model of a proposed aircraft, you can inject drops of liquid nitrogen into the flow and direct a beam of light on the area of interest—all the air vortices around the model become visible due to the scattering of light in the fog consisting of these drops and the particles of water and carbon dioxide that condense in their tracks. Liquid nitrogen is injected in cryogenic aerodynamic tunnels to cool the flow, and it's important to know the rate of evaporation of different-sized drops. In many industrial chemical processes, in gas "scrubbing" in factory smokestacks, and in many other instances, one needs to understand the behavior of an individual drop (of any substance) whose mass changes.

And for us, the audience, this little tragedy of a nitrogen drop exposed the workings of certain physical processes and showed how the experimental and theoretical approaches to a problem complement one another.


S. Kanov
JAGUAR

Generalizing Monty's dilemma

Strategies for dealing with a wily game-show host

by John P. Georges and Timothy V. Craine

SEVERAL YEARS AGO, MONTY'S DILEMMA—alternatively known as the "problem of the car and the goats"—generated controversy in the pages of the popular press and stimulated lively discussions in classrooms throughout the US. Widespread interest in this problem resulted from the fact that after *Parade* magazine columnist Marilyn vos Savant published her solution, she received thousands of letters, including many from mathematics professors, disputing her analysis. It turned out that in a sense, Marilyn was right, much to the surprise of many mathematicians with Ph.D.'s. In this article we consider generalizations of Monty's dilemma.

First, let's review the original problem. Monty Hall, a game show host,[1] asks a contestant to select one of three doors behind which are prizes. There is one goat behind each of two doors and a car behind the third. After the contestant selects a door, the host reveals a goat behind one of the unchosen doors. (In the event that there is a car behind the contestant's door, Monty may choose to open either of the other doors. We assume that each of these doors is chosen with probability 1/2.)

Monty then asks the contestant whether she wishes to switch her choice to the remaining door. The contestant, who we assume prefers to win the car, has two strategies: "stick" with the original door or "switch" to the remaining door.

What would *you* do? Do you figure your odds are fifty-fifty, so it doesn't matter what you do at this point?

Well, contrary to the intuition of many people, the switching strategy is superior to the sticking strategy. By switching, the contestant will win the car with probability 2/3, whereas by sticking, she will win the car with probability 1/3. Here is one explanation. Assuming that the contestant's initial choice is random, she will select the car with probability 1/3 and a goat with probability 2/3. By using the sticking strategy, which is essentially the same as ignoring the information given by Monty, the contestant should expect to win the car 1/3 of the time and to win a goat 2/3 of the time. If, on the other hand, the contestant uses the switching strategy, then for the 2/3 of the time when she initially selects a goat, she will switch to the car, and the 1/3 of the time when she initially selects the car, she will switch to a goat. In other words, under the switching strategy the events of initially selecting a goat and eventually winning the car are equivalent.

One car, many goats

Now let's suppose Monty decides to make the game more interesting by providing more doors. Consider the situation where there are n doors with $n \geq 3$, behind which appear $n - 1$ goats and one car. After the contestant has chosen a door, Monty opens one of the remaining doors behind which is a goat. (Again he chooses with equal probability from among those doors that have goats behind them.)

The contestant can either stick with her original choice or switch to one of the $n - 2$ remaining doors. The contestant will have chosen the correct door $1/n$ of the time and an incorrect door $(n - 1)/n$ of the time. As a result, the sticking strategy will be successful with probability $1/n$.

On the other hand, if the contestant adopts the switching strategy, the probability of success is computed by multiplying two probabilities, as we shall explain. The computation involves the concept of conditional probability. The notation described in the box below will be used throughout our discussion.

[1]His game show, "Let's Make a Deal," is no longer on the air (to the best of our knowledge). Monty Hall himself took a great interest in the dilemma named after him—see the front-page article in the *New York Times* of July 21, 1991.—*Ed.*

Art by Sergey Ivanov

$P(A)$ is the probability that event A occurs

$P(B \mid A)$ is the probability that event B occurs given that event A has occurred. Read this as "probability of B given A"

P_{stick} indicates the probability of winning the car if the contestant adopts the sticking strategy

P_{switch} indicates the probability of winning the car if the contestant adopts the switching strategy

The probability of two events occurring is equal to the probability that the first occurs multiplied by the conditional probability that the second occurs given that the first occurs—that is, $P(A \text{ and } B) = P(A)P(B \mid A)$. In this case the probability that the contestant chooses the car is found by multiplying the probability that the original choice is incorrect, $(n-1)/n$, and the probability that she chooses the car from the remaining $n-2$ doors, $1/(n-2)$.

Using our notation,

$$P_{switch} = P(\text{1st pick goat})P(\text{2nd pick car} \mid \text{1st pick goat})$$

$$= \frac{n-1}{n}\left(\frac{1}{n-2}\right) = \left(\frac{n-1}{n-2}\right)\frac{1}{n} > \frac{1}{n} = P_{stick}.$$

We conclude that the switching strategy is superior to the sticking strategy in this generalization.

Many cars, many goats

But what if there is more than one car? Does it still pay to switch? Suppose there are n doors behind which there are j cars and $n-j$ goats, for a suitable value of j. This time Monty may reveal either a car or a goat without giving away his secret. If he reveals a goat, then we must have $1 \le j \le n-2$. If he shows a car, then $2 \le j \le n-1$. In either case the probability of success is j/n for the sticking strategy. For each variation the switching strategy is successful whenever the contestant either first picks a goat and then switches to a car or first picks a car and then switches to another car—that is,

$$P_{switch} = P(\text{1st pick goat})P(\text{2nd pick car} \mid \text{1st pick goat}) + P(\text{1st pick car})P(\text{2nd pick car} \mid \text{1st pick car}).$$

In the case where Monty reveals a goat, the probability of a successful switch is given by

$$P_{switch} = \frac{n-j}{n}\left(\frac{j}{n-2}\right) + \frac{j}{n}\left(\frac{j-1}{n-2}\right)$$

$$= \frac{j}{n}\left[\frac{(n-j)+(j-1)}{n-2}\right] = \frac{j}{n}\left(\frac{n-1}{n-2}\right) > \frac{j}{n} = P_{stick}. \quad (1)$$

Problem 1. Show that in the second case, where Monty reveals a car,

$$P_{switch} = \frac{n-j}{n}\left(\frac{j-1}{n-2}\right) + \frac{j}{n}\left(\frac{j-2}{n-2}\right) \quad (2)$$

and that $P_{switch} \le P_{stick}$.

Consequently, when Monty reveals a goat, the probability of success for the switching strategy will be greater than j/n (the probability of success for the sticking strategy); however, when Monty reveals a car, the probability of success for the switching strategy will be less than j/n. Based on these results, we can conclude that the contestant should adopt a strategy based on what is behind the opened door: switch if it's a goat; stick if it's a car.

Many cars, many goats, uncertain revelations

In the previous example, Monty has decided in advance to reveal either a goat or a car. Suppose instead he tells the contestant, "I'll show you a goat with probability p and a car with probability $1-p$. But before I open a door you must decide whether to stick or switch."

Once again, when she sticks, the contestant is successful with probability j/n. When she switches, she is successful with probability

$$P_{switch} = p\left[\frac{n-j}{n}\left(\frac{j}{n-2}\right) + \frac{j}{n}\left(\frac{j-1}{n-2}\right)\right] + (1-p)\left[\frac{n-j}{n}\left(\frac{j-1}{n-2}\right) + \frac{j}{n}\left(\frac{j-2}{n-2}\right)\right]. \quad (3)$$

Note that expression (3) is derived from expressions (1) and (2) above.

Problem 2. Suppose there are eight doors with 5 goats and 3 cars. Monty tells you that he will roll a die to determine which door to open. If he rolls a multiple of 3 he will reveal a car. Otherwise he will reveal a goat. You must commit in advance to sticking or switching. What should you do?

Now suppose that $p = (n-j)/n$. Then $1-p = j/n$, and expression (3) becomes

$$P_{switch} = \frac{n-j}{n}\left[\frac{n-j}{n}\left(\frac{j}{n-2}\right) + \frac{j}{n}\left(\frac{j-1}{n-2}\right)\right] + \frac{j}{n}\left[\frac{n-j}{n}\left(\frac{j-1}{n-2}\right) + \frac{j}{n}\left(\frac{j-2}{n-2}\right)\right]$$

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

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

Consequently, whenever $p = (n - j)/n$, it doesn't matter whether the contestant decides to stick or switch.

Problem 3. Show that if $p > (n - j)/n$, then $P_{\text{switch}} > j/n$, and if $p < (n - j)/n$, then $P_{\text{switch}} < j/n$.

To appreciate the result demonstrated in problem 3, note that the extreme cases where $p = 1$ and $p = 0$ correspond to situations in which Monty's behavior is totally predictable, and the probability of success for switching is given by expressions (1) and (2), respectively, in the preceding section.

Many revelations

Here is yet another generalization of the original problem. Monty opens not just one door but rather some subset of unpicked doors. Assume, as before, that there are n doors with j cars. Monty will now reveal to the contestant k cars and $m - k$ goats with $1 \le m \le n - 2$ and $0 \le k \le m$. Once again, the probability of success for the sticking strategy is j/n. The probability of success for the switching strategy is

$$P_{\text{switch}} = P(\text{1st pick goat})P(\text{2nd pick car} \mid \text{1st pick goat})$$

$$+ P(\text{1st pick car})P(\text{2nd pick car} \mid \text{1st pick car})$$

$$= \frac{n-j}{n}\left(\frac{j-k}{n-m-1}\right) + \frac{j}{n}\left(\frac{j-k-1}{n-m-1}\right) \qquad (4)$$

$$= \frac{nj - nk - j}{n(n-m-1)}.$$

Problem 4. Show that expression (4) is equal to j/n when $k/m = j/n$, is greater than j/n when $k/m < j/n$, and is less than j/n when $k/m > j/n$.

Here is another way to view the result of problem 4. If the relative frequency of goats in the revealed set is greater than the relative frequency of goats in the original collection, the contestant should adopt the switching strategy. If this relative frequency equals that in the original collection, it doesn't matter which strategy is adopted, and if the relative frequency of goats in the revealed set is less than that in the original collection, the contestant should stick with her original choice.

Many different prizes

Suppose that the prizes available to the contestant aren't restricted to cars and goats but contain a wide variety. For instance, there may be six doors with two sports cars, each worth \$30,100; two motor boats, each worth \$15,100; and two goats, each worth \$100. In this case, by sticking with her original choice, the contestant can expect to win \$30,100 with probability 1/3, \$15,100 with probability 1/3, and \$100 with probability 1/3. The expected value of her prize is

$$\frac{1}{3}(30{,}100) + \frac{1}{3}(15{,}100) + \frac{1}{3}(100) = \frac{1}{3}(45{,}300) = \$15{,}100.$$

Again the question becomes, will she do better if she switches?

Suppose Monty decides to open a door with a sports car. Then the expected value of the prize the contestant wins by switching is determined by what lies behind the door she originally chose and the remaining four unopened doors.

If the prize behind the door she chose is a car, her expected value will be

$$\frac{2}{4}(15{,}100) + \frac{2}{4}(100) = \$7{,}600.$$

If it is a boat, her expected value will be

$$\frac{1}{4}(30{,}100) + \frac{1}{4}(15{,}100) + \frac{2}{4}(100) = \$11{,}350.$$

If it is a goat, her expected value will be

$$\frac{1}{4}(30{,}100) + \frac{2}{4}(15{,}100) + \frac{1}{4}(100) = \$15{,}100.$$

Each of these possibilities is equally likely, so overall her expected value for switching is

$$\frac{1}{3}(7{,}600) + \frac{1}{3}(11{,}350) + \frac{1}{3}(15{,}100) = \$11{,}350.$$

Since the expected value for the switching strategy is less than the expected value for the sticking strategy, if Monty reveals a car, she should stick with her original choice.

Problem 5. Again suppose there are six doors with two sports cars, each worth \$30,100; two motor boats, each worth \$15,100; and two goats, each worth \$100. This time, Monty decides to reveal a goat. Find the contestant's expected value for switching and determine what she should do.

Problem 6. In the same situation as in problem 5, suppose Monty reveals a boat. Compare the expected values for the sticking and switching strategies. What should the contestant do?

In order to generalize the problem with multiple prizes, we need some additional notation relating to expected value (see the box on the next page). In general, expected value E is found by multiplying the probability that each event occurs by the value of that event and finding the sum of all such products. If there are m events, this is often represented as

$$E = \sum_{i=1}^{m} p_i v_i,$$

where p_i is the probability that event i occurs and v_i is the value of the prize associated with event i.

Now suppose there are m types of prize, valued at v_1, v_2, ..., v_m. For $1 \le i \le m$, there are n_i prizes with value v_i.

$\sum$ means "take the sum of"

$$E = \sum_{i=1}^{m} p_i v_i$$

E_{stick} is the expected value of the prize with the sticking strategy

E_{switch} is the expected value of the prize with the switching strategy

$E_{\text{switch}}(1\text{st value} = v_i)$ is the expected value of the prize with the switching strategy given that the value of the first pick is v_i

We note that the total number of prizes is

$$n = \sum_{i=1}^{m} n_i.$$

The contestant strives to maximize the expected value of her prize.

The expected value of a prize for the sticking strategy is now given by

$$E_{\text{stick}} = \sum_{i=1}^{m} \frac{n_i}{n} v_i,$$

where n_i/n is the probability of picking a prize with value v_i. If we let

$$t = \sum_{i=1}^{m} n_i v_i$$

denote the total value of the prizes, then

$$E_{\text{stick}} = \frac{t}{n}.$$

After the contestant makes her original selection, Monty reveals a prize with value v_r. Presumably $n_r \geq 2$, so there remains the possibility that the prize chosen by the contestant also has value v_r. Now the expected value for the switching strategy is given by

$$E_{\text{switch}} = \sum_{i=1}^{m} P(1\text{st pick value} = v_i) E_{\text{switch}}(1\text{st pick value} = v_i).$$

Given that a prize with value v_i was behind the first door picked, the expected value of the prize obtained by switching is the average of the $n - 2$ prizes available. This average is found by taking t, subtracting $v_i + v_r$ (the sum of the values of the prizes behind the first door picked and the one opened by Monty), and then dividing the result by $n - 2$. Thus,

$$\begin{aligned} E_{\text{switch}} &= \sum_{i=1}^{m} \left(\frac{n_i}{n}\right)\left(\frac{t - v_i - v_r}{n-2}\right) \\ &= \frac{1}{n} \cdot \frac{1}{n-2}\left[nt - \sum_{i=1}^{m} n_i v_i - n v_r\right] \\ &= \frac{nt - t}{n(n-2)} - \frac{n v_r}{n(n-2)} \qquad (5) \\ &= \frac{t}{n}\left(\frac{n-1}{n-2}\right) - \frac{v_r}{n-2}. \end{aligned}$$

Problem 7. Show that $E_{\text{stick}} = E_{\text{switch}}$ whenever $v_r = t/n$. Also show that when $v_r > t/n$, the sticking strategy is better, and when $v_r < t/n$, the switching strategy is better.

Problem 7 demonstrates that the contestant should base her decision on how the revealed prize compares to the average value of the prizes.

What happens if Monty chooses to reveal more than one prize to the contestant in this multiple-prize situation? Suppose that s prizes are revealed with a total value of x. Equation (5) now becomes

$$\begin{aligned} E_{\text{switch}} &= \sum_{i=1}^{m} \left(\frac{n_i}{n}\right)\left(\frac{t - v_i - x}{n-s-1}\right) \\ &= \frac{1}{n} \cdot \frac{1}{n-s-1}\left[nt - \sum_{i=1}^{m} n_i v_i - nx\right] \\ &= \frac{t}{n}\left(\frac{n-1}{n-s-1}\right) - \frac{x}{n-s-1}. \qquad (6) \\ &= \frac{tn - t - xn}{n(n-s-1)}. \end{aligned}$$

Problem 8. Use equation (6) to show that the expected values for sticking and switching are equal whenever $x/s = t/n$.

The criterion for determining the better strategy is a comparison of the average value of the revealed prizes x/s with the average value of the entire set of prizes t/n. Whenever the former is greater than the latter, the contestant should stick. Whenever it is less, she should switch.

This last generalization encompasses all the previous cases that we have considered. For the situations in which the prizes consisted only of goats and cars, we can assign the value 0 to a goat and the value 1 to a car. Then the expected value associated with a strategy is equal to the probability of winning a car. Recall that in the previous section, in which the host opened m doors to reveal k cars, we determined that the criterion is to compare k/m and the original ratio of cars to doors. This is an instance of the more general criterion we just established.

Finally, the generalization in which Monty reveals a goat with probability p and a car with probability $1-p$ can be analyzed in terms of expected value by using the above approach. In this case there are $n_1 = j$ cars valued at $v_1 = 1$ and $n_2 = n - j$ goats valued at $v_2 = 0$. The total value of the prizes t is the number of cars j, and the expected value of the revealed prize v_r is the probability of winning a car $1-p$. Substituting into equation (5), we have

$$
\begin{aligned}
E_{\text{switch}} &= \sum_{i=1}^{m}\left(\frac{n_i}{n}\right)\left(\frac{t - v_i - v_r}{n-2}\right) \\
&= \frac{j}{n}\left[\frac{j-1-(1-p)}{n-2}\right] + \frac{n-j}{n}\left[\frac{j-0-(1-p)}{n-2}\right] \\
&= \frac{j(j+p) - 2j + n(j+p) - j(j+p) - n + j}{n(n-2)} \qquad (7) \\
&= \frac{nj + np - n - j}{n(n-2)}.
\end{aligned}
$$

Problem 9. Show that expressions (3) and (7) are equivalent.

An alternate approach

Equation (6) can be derived by using an alternative method. Instead of viewing the contestant's behavior as a conscious decision to stick or switch, let's consider it as entirely random. After the host has opened s doors to reveal prizes with a total value of x, the contestant is left with $n - s$ doors, including the one that she originally chose. If she chooses from these doors randomly, she will stick with probability $1/(n-s)$ and switch with probability $(n-s-1)/(n-s)$. The expected value of the prize she wins is the average of the prizes that remain unrevealed: $(t-x)/(n-s)$. Now, since we know that the expected value for the sticking strategy is t/n, we have

$$E_{\text{random}} = \frac{1}{n-s}E_{\text{stick}} + \frac{n-s-1}{n-s}E_{\text{switch}},$$

implying that

$$\frac{t-x}{n-s} = \frac{1}{n-s}\cdot\frac{t}{n} + \frac{n-s-1}{n-s}E_{\text{switch}}. \qquad (8)$$

Solving equation (8) for E_{switch}, we obtain equation (6).

Let's return to the original three-door problem for a moment. Immediately after Monty has revealed a goat, the contestant must decide whether to stick with original choice or switch to the remaining door. If at this point the contestant were to toss a fair coin to determine her behavior, she would expect to win the car 1/2 of the time—in other words, $P_{\text{random}} = 1/2$. If one accepts the fact that the contestant initially has a 1/3 chance of selecting the correct door (that is, $P_{\text{stick}} = 1/3$), then equation (8) shows that

$$\frac{1}{2} = P_{\text{random}} = \frac{1}{2}P_{\text{stick}} + \frac{1}{2}P_{\text{switch}} = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + \frac{1}{2}P_{\text{switch}},$$

which implies that $P_{\text{switch}} = 2/3$.

Many of the people who objected to Marilyn's solution of the original problem intuitively assigned 1/2 to the probability of selecting the correct door under the switching strategy. They had confused the random and switching strategies. In fact, the random strategy is a mixture of the pure strategies of sticking and switching, as shown in equation (8).

Conclusion

At first glance Monty's dilemma resembles problems that are best explained through conditional probability. As a result, most of the early discussions about Monty's dilemma were presented in this context. Indeed most of our generalizations were described in these terms—we compared the probability of success between the sticking strategy and the switching strategy.

In mathematical game theory, strategies are typically compared with respect to expected value—the bilinear combination of payoffs and probabilities. Our last generalization points out the true nature of the problem: should a person randomly select an alternative from a set of options A, or should she randomly select an alternative from the set of options $A - B$, where B is a given subset of A? The answer is quite simple: if the average value of the elements in A is strictly less than the average value of the elements in B, then the contestant should randomly select an element from A (sticking strategy); if the average value of an element of A is strictly greater than the average value of an element of B, then she should opt to randomly select an element of $A - B$ (switching strategy). Finally, if the average values of the elements in A and B are equal, then it does not matter which strategy is employed.

In conclusion, we note that other generalizations of Monty's dilemma can be given to include the notions of risk and utility. Risk involves the question of how willing a person is to take a chance. A "risk taker" will tend to prefer an alternative with high expected value, even if the probability of winning a large prize is quite small. A person who is "risk averse," on the other hand, may prefer a small prize that is highly probable to a much larger prize that is much less probable, even when the expected value of the second option is greater than that of the first.

Utility extends the concept of value to include personal preferences. Not all outcomes can or should be measured in dollar terms. After all, a goat may have greater value to a Tibetan monk than does a car! ◘

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Magnetic levitation comes of age

Reducing friction and increasing efficiency in vehicles and machines

by Thomas D. Rossing and John R. Hull

IMAGINE, IF YOU WILL, FLY-wheels that store megajoules of energy, maglev vehicles that fly on elevated guideways, motors that spin at more than 10^5 rpm. These are just some of the applications of magnetic levitation that are presently being developed in laboratories around the world. By providing us with a way to eliminate mechanical friction, magnetic levitation opens up new horizons in machinery design and transportation of people and goods at high speeds in an energy-efficient way.

Superconductors and levitation

In 1911 the Dutch physicist Heike Kammerlingh Onnes reported that the electrical resistance of mercury suddenly disappears below 4.2 K (–269°C)—the boiling point of helium. Even Kammerlingh Onnes, who was awarded the Nobel prize in 1913, could scarcely have dreamed of the many important technologies that would result from the use of superconductors. Magnets of superconducting wire cooled with liquid helium are now essential parts of high-energy particle accelerators and magnetic resonance imaging machines in hospitals, for example.

Another milestone occurred in 1986 when Alex Müller and Georg Bednorz (Switzerland) reported that superconductivity could occur in lanthanum–barium–copper oxide at 30 K. Their discovery led to a flurry of interest in high-temperature superconductors with ever increasing transition temperatures, some of them above the boiling point of nitrogen (77 K). Not only is liquid nitrogen much cheaper and easier to use than liquid helium, but these new materials belong to a class called Type II superconductors, which retain their superconductivity much better in the presence of magnetic fields than do Type I superconductors (such as mercury, lead, tin, and other elements).

The availability of materials that are superconducting at the boiling temperature of liquid nitrogen has made the levitation of a small magnet over a superconductor (or vice versa) a familiar demonstration in physics lectures.[1] This demonstration, which depends on Faraday's law or the Meissner effect (depending on how the experiment is done), was first done in 1945 by V. Arkadyev, who levitated a magnet over a concave lead plate in liquid helium.

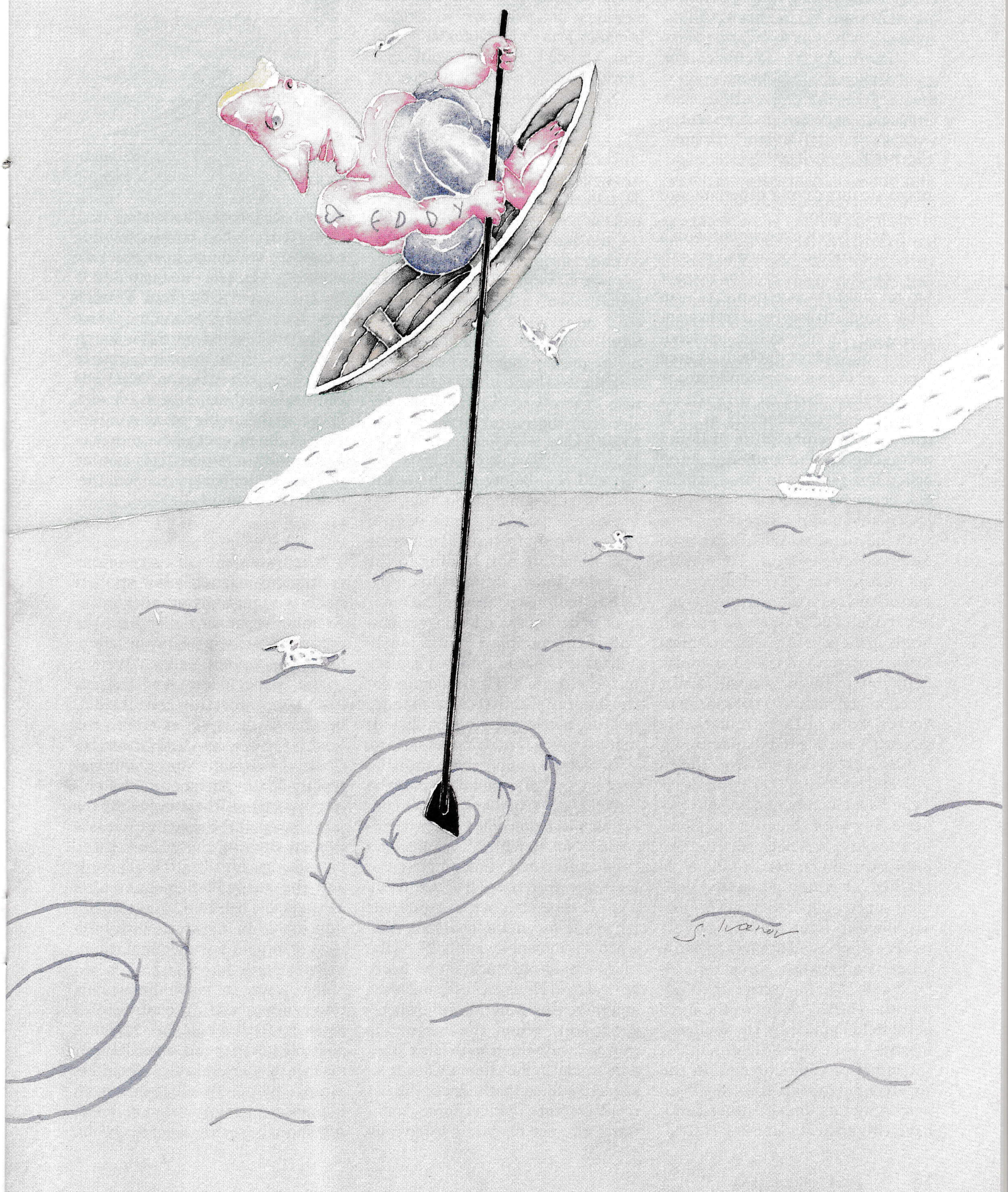
If a magnet is lowered onto a superconductor, shielding currents are induced on the surface of the superconductor (Faraday's law: a changing magnetic field induces a voltage in a conductor). These supercurrents create a magnetic field that repels the magnet and thus levitates it at a height such that the repulsive force equals the weight of the magnet. Supercurrents are possible because of the zero resistance of the superconductor.

If, on the other hand, a magnet already rests on the superconductor as it is cooled through its transition temperature and becomes superconducting, the magnet will mysteriously rise (but not quite as high as before) due to the expulsion of magnetic flux from the superconductor. This remarkable property of superconductors was discovered in 1933 by W. Meissner and R. Ochsenfeld and is called the Meissner effect.

In some discussions of magnetic levitation, both of these phenomena are attributed to the Meissner effect, but properly this term should apply only to the case in which the superconductor is cooled in a magnetic field (that is, field-cooled). Two important properties of a superconductor are its perfect conductivity (resistance $R = 0$) and its perfect diamagnetism (magnetic permeability $\mu = 0$, which means that

[1]The article "Meeting No Resistance" in the September/October 1991 issue of *Quantum* contains a photo of this phenomenon. See also the Physics Contest in the November/December 1994 issue.—*Ed.*

Art by Sergey Ivanov


EDDY
S. Ivanov

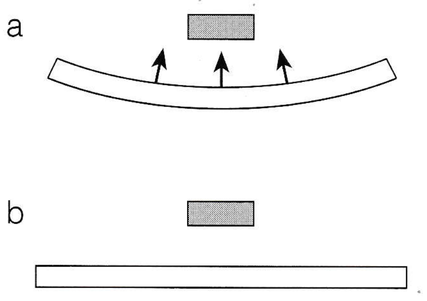


Figure 1

magnetic flux cannot penetrate into a superconductor).

In Arkadyev's experiment with a lead superconductor, a concave surface was required to give the magnet lateral stability (see figure 1a)—in other words, to prevent it from wandering off the edge of the superconductor. Nowadays the levitation of a magnet over a superconductor is usually demonstrated with oxides that are superconducting at 77 K (the boiling temperature of liquid nitrogen). Due to a phenomenon called *flux pinning* in these materials, a concave surface is not needed for lateral stability as was the case with the metallic lead samples used by Arkadyev (see figure 1b). A magnet floating over a Type II superconductor, such as yttrium–barium–copper oxide (YBCO), can be pushed nearly to the edge of the superconductor without loss of lateral stability. Furthermore, the magnet will float at different heights above the superconductor. Both of these remarkable properties are due to flux pinning.

Just what is flux pinning? Here's a brief description. Lead and other Type I superconductors expel magnetic flux (and thus the magnetic flux density, or B-field, goes to zero) until a certain critical H-field H_c is reached, at which point they lose their superconductivity and $B = \mu H$, as shown in figure 2a. (To avoid a rather complicated discussion of the difference between a magnetic B-field and H-field, suffice it to say that the H-field describes the magnetic field applied by the levitated magnet, while the B-field includes the effect of supercurrents in the superconductor as well.) Type II superconductors, on the other hand, have two critical values of H-field,

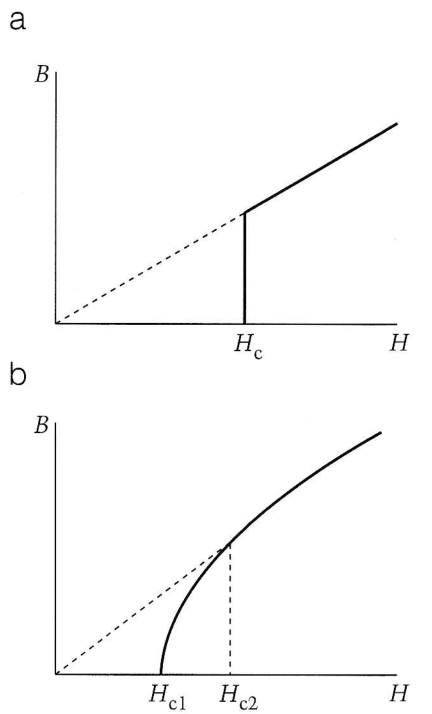


Figure 2

H_{c1} and H_{c2}. Below H_{c1} the entire specimen is superconducting, but between H_{c1} and H_{c2} some parts of the material are superconducting and others are not, leading to the situation shown in figure 2b. Magnetic flux lines penetrate the "normal" or nonsuperconducting regions, and if the sample is sufficiently "dirty" they become pinned in place in these regions. This accounts for the lateral stability of a magnet floating above a sample of Type II material even if it has a flat surface.

In a Type I superconductor the magnet will always levitate at the same height, whether the superconductor was field-cooled or zero-field-cooled. Not so with a Type II superconductor. Due to flux pinning, the levitation force on a magnet over a Type II superconductor is different when the magnet approaches than when it is moving away (fig. 3). As the magnet is brought nearer, the lower critical field H_{c1} is reached, and more and more flux penetrates the superconductor. When the magnet is moved away, the trapped flux lines tend to stay in the sample, causing an attractive force that reduces the net repulsive force. (Under some conditions, the net magnetic force may

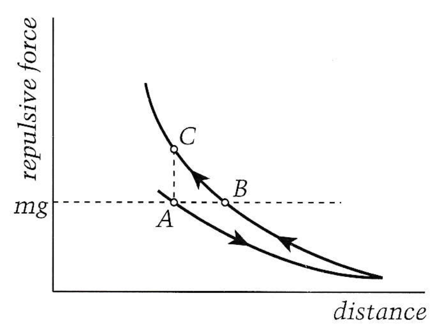


Figure 3

actually be one of attraction.)

In figure 3 the repulsive force equals the weight of the magnet mg at points A and B or at any point on the line connecting them. Point B represents the levitation height for a magnet lowered onto the superconductor. If the magnet is pushed down onto the superconductor and then released at point C, it will move to the stable point A. Likewise, if the magnet rests on the superconductor when it is cooled through its transition temperature, it will rise to point A.

Magnetic suspension

It is impossible to suspend one permanent magnet below another stably without applying other forces. At some separation distance, the attractive force equals the weight of the lower magnet, so the net force on the magnet is zero. However, this is a case of unstable equilibrium, because if the magnet moves the least bit above this equilibrium position, the attractive force increases rapidly and the magnet moves rapidly toward the other magnet. By the same token, if the spacing increases ever so slightly, the weight of the magnet exceeds the attractive force and the suspended magnet falls away. Stable equilibrium is actually impossible in a system with only inverse-square-law electrostatic or magnetostatic forces.

If, however, the current in the electromagnet is carefully controlled by means of feedback, it then becomes possible to suspend a permanent magnet or a ferromagnetic (ironlike) sample below an electromagnet. Likewise, an electromagnet can be suspended below a sheet of steel if the

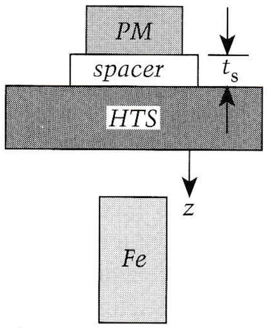


Figure 4

current in the electromagnet is carefully controlled. In fact, this is how the electromagnetic suspension (EMS) system in one type of maglev vehicle works. We'll take a closer look at it later in this article.

Another way to achieve stable suspension is to insert a superconductor between the magnet and the ferromagnetic material—a soft iron cylinder in the arrangement shown in figure 4. The magnet's field magnetizes the soft iron sample, which is then repelled by the superconductor. Stable suspension of the iron is possible due to a balance between the attractive force of the permanent magnet and the repulsive force of the superconductor. Lateral stability is supplied by flux pinning in the superconductor.

Such an arrangement was used by Loren Passmore, a summer student at Argonne National Laboratory, to demonstrate the levitation of a soft iron cylinder using a permanent magnet and a superconductor. Once levitated, such a cylinder could be rotated at high speed with negligible friction. Passmore found that adding a second magnet–superconductor pair below the cylinder improved the stability of the levitation.

It's also possible to suspend a magnet below a Type II superconductor (or vice versa) if the flux pinning is sufficiently strong. The magnet is generally placed next to the superconductor as it's cooled through its transition temperature. Much of the magnetic flux is then trapped within so-called flux vortices in the superconductor, causing them to act as small magnets, which together supply an attractive force for the permanent magnet. At the same time, the rest of the superconductor repels the magnet, so that a balance of attractive and repulsive forces is achieved over a range of separation distances. Lateral stability is again supplied by the flux pinning in the superconductor.

A new kind of bearing

One byproduct of superconductivity research is the magnetic bearing, which uses magnetic levitation to suspend rapidly rotating devices without mechanical friction. Although various types of electromagnetic bearings are now in use, the simplest type uses permanent magnets and superconductors. This bearing can be made stable without an active feedback system and is under development at several laboratories. If the rotor starts to drift away from its center position, flux pinning provides the force needed to restore its previous position. This ability to restore the rotor to its desired position is known as the magnetic stiffness of the bearing.

Superconducting magnetic bearings offer several advantages over other types of bearings, including long life, high efficiency, and low maintenance. Melt-textured YBCO, which promises high magnetic stiffness and large levitation pressure in magnetic bearings, is a particularly promising material. Melt-textured samples are prepared by heating YBCO to its melting temperature, cooling it slowly, and then annealing it in oxygen. This process helps eliminate nonsuperconducting phases and leads to large levitation forces.

Figure 5 shows a schematic of a simple motor constructed by

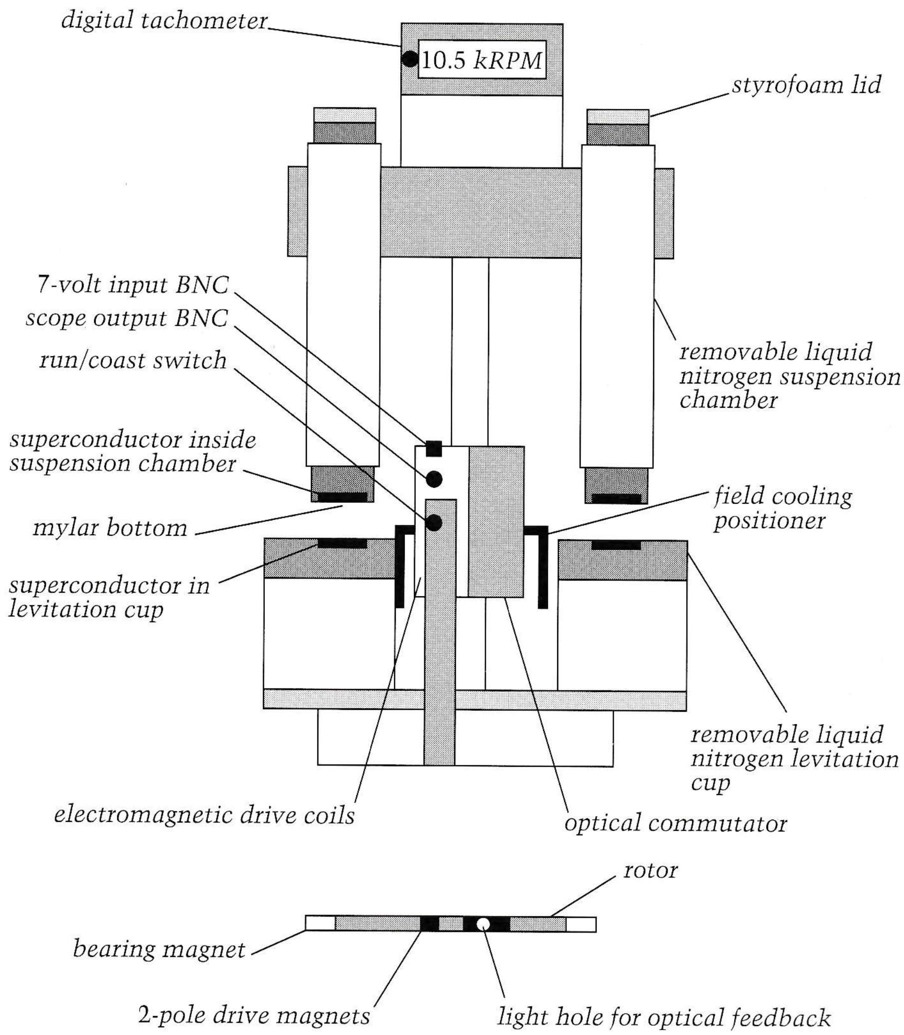


Figure 5

Argonne summer student Christopher Gabrys (University of Vermont) using magnetic bearings. The ends of the rotor have cylindrical NdFeB magnets that can either be supported by repulsive levitation over two high-temperature superconductors or by attractive levitation underneath the same superconductors. (To achieve attractive levitation, flux from the bearing magnets is allowed to penetrate the superconductors prior to cooling. Field-cooling allows an attractive force.) The rotor can be driven above 10,000 rpm by three magnetic coils controlled by optical feedback to turn the coils on and off at just the right time. This demonstration motor was a prize winner in the 1992 apparatus competition of the American Association of Physics Teachers.

Magnetic bearings promise to be one of the first practical applications of high-temperature superconductors, since they don't require that the superconducting material be fabricated in the form of wires, as many other applications do.

Low-loss flywheels

Another promising superconductivity application marks an improvement on a device that made its first appearance with the steam engine at the dawn of the industrial age. Flywheels store energy in the form of rotational kinetic energy in many devices, including buses and commuter trains. A large magnetically levitated flywheel, spinning in a vacuum chamber, could store large amounts of energy with little or no loss due to friction. Such flywheels are particularly attractive to electrical power companies for storing off-peak energy to meet peak load demands in order to make more efficient use of their power plants. Flywheels could also be used to store energy for making short trips in small automobiles. Their storage capacity is comparable to batteries of the same size, and they wouldn't have to be replaced periodically, as batteries must be. Like batteries, they wouldn't contribute to urban air pollution.

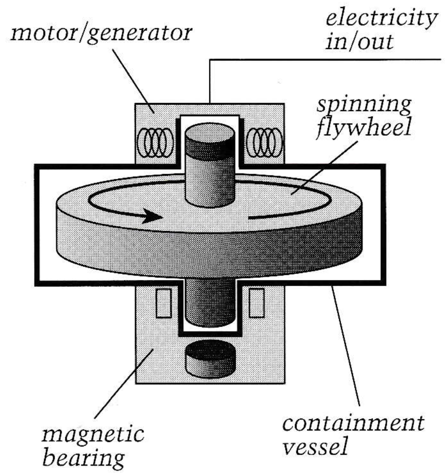


Figure 6

Figure 6 shows a small flywheel levitated by a magnetic bearing in a vacuum chamber. The bearing consists of a ring-shaped magnet levitated over superconducting disks of melt-textured YBCO. In experiments at Argonne National Laboratory, flywheels of this type have achieved coefficients of friction of less than 10^{-6}. In our tests, the flywheel is accelerated to high speeds with an electric motor or an air jet and then allowed to "coast" while we observe the rate of slowing down.

Levitation by induced eddy currents

According to Faraday's law, when a magnet moves over a conductor, the changing magnetic field induces a voltage in the conductor, which causes eddy currents to flow. These eddy currents in turn generate a magnetic field that opposes the change in field due to the motion of the magnet. The induced eddy currents flow in loops under the moving magnet. When the magnet moves at a moderate speed, eddy currents will flow in one direction under the leading edge of the magnet and in the opposite direction under the trailing edge, as shown in figure 7a. When the magnet moves at a very high speed, however, there will be only one set of loops, as shown in figure 7b. The renowned English physicist James Clerk Maxwell suggested that a model using images of the moving magnet could help us understand the eddy currents in a conductor under a moving magnet. According to Maxwell's model, when a magnet passes a point on the conducting plane, it induces first a "positive" image, then a "negative" image of the magnet. These images occur in the conductor much as an image occurs when one stands in front of a mirror.

The eddy currents in the conductor generate both a lift force and a drag force on the moving magnet. At low speed, the drag force predominates. You can easily demonstrate this by letting a disk magnet slide down a smooth aluminum plate (fig. 8). The magnet slides very slowly due to the eddy-current drag force. Another way to demonstrate eddy-current drag is to let a magnet drop through an aluminum or copper tube and note its slow speed of descent. As the speed of a moving magnet increases, however, the drag force decreases (approximately as

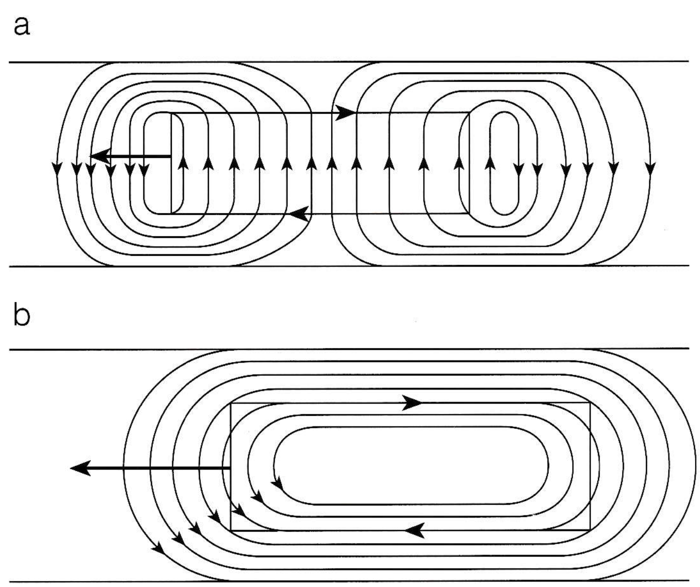


Figure 7

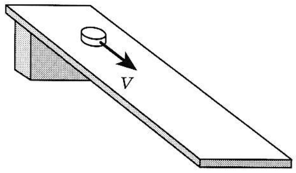


Figure 8

$\sqrt{v}$), and a lift force comes into prominence.

At sufficiently high speed, the lift force becomes essentially equal to the repulsive force between the moving magnet and an equal and opposite image magnet below the plane of the conductor. We mentioned above that bringing a magnet near a superconductor induces supercurrents whose magnetic field opposes the field of the magnet. The eddy currents induced in an ordinary conductor by a rapidly moving magnet show a striking similarity to the supercurrents induced in a superconductor. The major difference is that rapid motion isn't required with the superconductor.

Maglev vehicles

It now appears feasible to construct with present technology a high-speed ground transportation system using magnetically levitated vehicles operating at 500 km/h (300 mph). A national maglev system would do much to relieve congestion on highways and at airports. Such a system would be more energy efficient than most short-haul (100- to 600-mile) flights that operate in and out of major airports. Maglev vehicles could also be used to connect major cities with large international airports located away from congested urban areas.

Although maglev vehicles may resemble trains, in many ways their design is more closely related to airplanes. Consideration must be given to lift, drag, and guidance forces and to pitch, roll, and yaw, as in aircraft. One type of maglev system even uses "landing wheels" to support the vehicle at low speeds. Operation of high-speed magnetically levitated vehicles is sometimes referred to as "magnetic flight."

Magnetic flight can make use of either attractive or repulsive forces to levitate the vehicle. Electromagnetic suspension (EMS) depends on the attractive force between electromagnets and a steel guideway (fig. 9a), while electrodynamic suspension (EDS) depends on repulsive forces between moving magnets and eddy currents in the guideway (fig. 9b). EMS systems are inherently unstable, since the attractive magnetic force increases rapidly with decreasing spacing between the magnet and the guideway; EDS systems are more stable because the force of repulsion decreases with increasing spacing. In an electromagnetic (attractive) system, the magnet currents must be carefully controlled to maintain the desired suspension height; in an electrodynamic (repulsive) system, retractable wheels support the vehicle at low speeds and at rest.

Although research on maglev systems was pioneered in the United States during the late 1960s, support for this work was terminated in 1975 and did not resume again until about 1990. Research continued in Japan and Germany, however, and full-scale vehicles have been tested in both countries. A system based on the electromagnetic technology being planned for Orlando, Florida, will be the first public maglev system in the US.

The potential energy savings from a maglev transportation system are substantial. Flights of less than 500 miles tend to be very inefficient, because the aircraft spend a large fraction of time taxiing, climbing, and descending, and modern jet aircraft operate most efficiently at high altitudes. Maglev trains, on the other hand, could carry short-haul passengers to their destinations in less total time with far less energy. As our finite energy sources dwindle, we have every reason to hasten the development of maglev technology and intensify research in superconductivity.

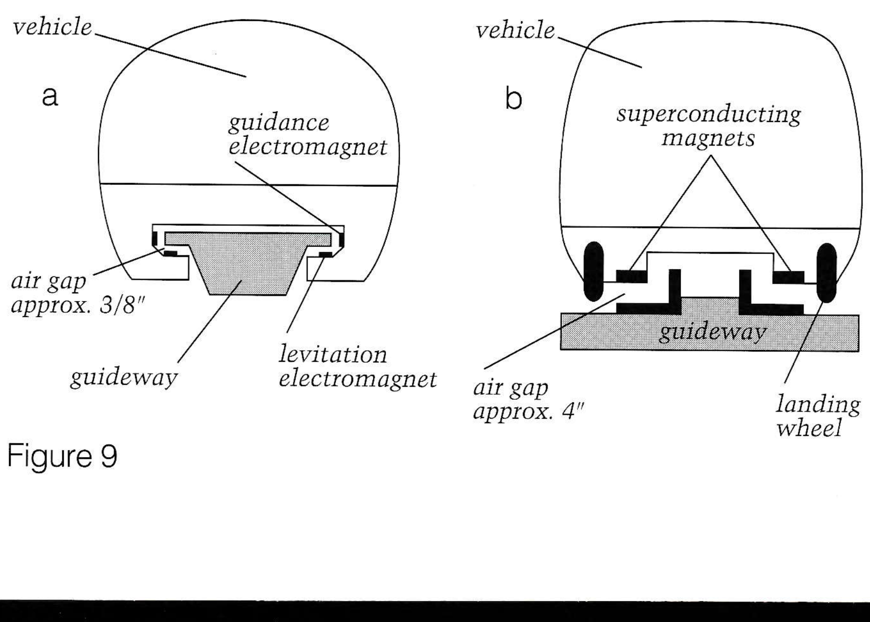


Figure 9

Mind-expanding books!

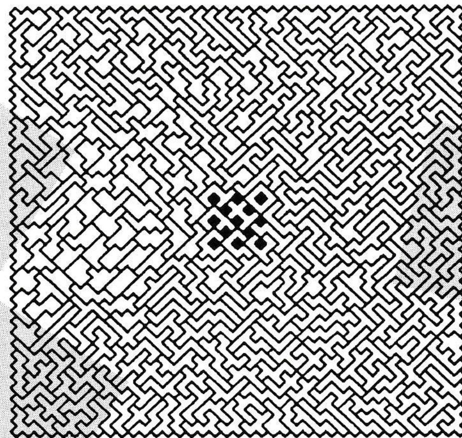
THE NATURE AND POWER OF MATHEMATICS

Donald M. Davis

In this engaging book, Donald Davis explains some of the most fascinating ideas in mathematics, highlighting their historical interest, their often surprising applicability, and their beauty.

The three main topics discussed are non-Euclidean geometry, with its application to the theory of relativity; number theory, with its application to cryptography; and fractals, which are an application of early mathematical work on iteration. Other topics include the influence of Greek mathematics on Kepler's laws of planetary motion, and the theoretical work that led to the development of computers.

Paper: $24.95 ISBN 0-691-02562-2



New edition

POLYOMINOES

Puzzles, Patterns, Problems, and Packings

Solomon W. Golomb

Inspiring popular video games, like *Tetris*, while contributing to the study of combinatorial geometry and tiling theory, polyominoes have continued to spark interest ever since their inventor, Solomon Golomb, introduced them to puzzle enthusiasts several decades ago. In this fully revised and expanded edition, the author takes a new generation of readers on a mathematical journey into the world of polyominoes, incorporating important recent developments.

"Those acquainted with polyominoes will, of course, snap up this volume. And if you haven't yet discovered the fascination of playing with these tantalizing tiles, this is the book to introduce you to them as well as to many significant aspects of combinatorial geometry." —Martin Gardner, author of *The Ambidexterous Universe*

Cloth: $24.95 ISBN 0-691-08573-0

DESIGNING THE MOLECULAR WORLD

Chemistry at the Frontier

Philip Ball

Some of the most exciting scientific developments in recent years have come from the chemistry lab. Chemists have created superconducting ceramics for brain scanners, designed liquid crystal flat screens for televisions and watches, fashioned metals from plastics, and produced drugs from crude oil. Philip Ball takes readers inside the exciting world of modern chemistry.

"This book offers an exhilarating, insightful tour of the vast wonderlands of modern molecular science. Philip Ball is a lucid and lively guide. He does not dally in museums but briskly roams the frontiers, pointing out splendid vistas." —Dudley Herschbach, Nobel Laureate in Chemistry

Cloth: $29.95 ISBN 0-691-00058-1

THE ENJOYMENT OF MATH

Hans Rademacher and Otto Toeplitz

What is so special about the number 30? How many colors are needed to color a map? Do the prime numbers go on forever? Are there more whole numbers than even numbers? These and other mathematical puzzles are explored in this delightful book by two eminent mathematicians. Requiring no more background than plane geometry and elementary algebra, this book leads the reader into some of the most fundamental ideas of mathematics, the ideas that make the subject exciting and interesting.

"A thoroughly enjoyable sampler of fascinating mathematical problems and their solutions."—*Science*

Paper: $12.95 ISBN 0-691-02351-4

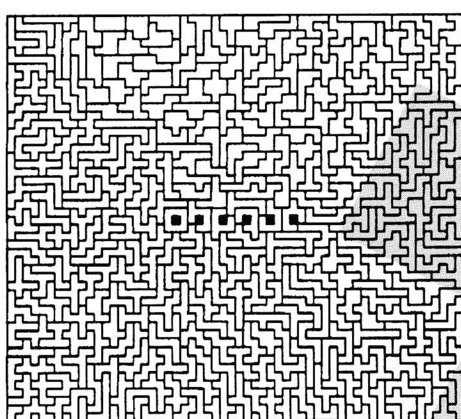
e: THE STORY OF A NUMBER

Eli Maor

The story of π has been told many times, both in scholarly works and in popular books. But the story of its close relative, the number e, has never before been written for a general audience. This fascinating book describes the story of e from a human as well as a mathematical perspective.

"Maor has succeeded in writing a short, readable mathematical story. He has interspersed a variety of anecdotes, excursions, and essays to lighten the flow. . . . It is like the voyages of Columbus as told by the first mate."—Peter Borwein, *Science*

Cloth: $24.95 ISBN 0-691-03390-0



PRINCETON UNIVERSITY PRESS

AVAILABLE AT FINE BOOKSTORES OR DIRECTLY FROM THE PUBLISHER: 800-777-4726

Circle No. 4 on Reader Service Card

Challenges in physics and math

Math

M136

Unknown natural powers. Solve each of the following equations in natural numbers x, y, z: (a) $x^y + 1 = (x + 1)^2$; (b) $2^y + 1 = 3^z$. (D. Fleishman)

M137

Distances to a trapezoid. Prove that the sum of the distances from an arbitrary point in the plane to three vertices of an isosceles trapezoid is always greater than the distance from this point to the fourth vertex. (S. Rukshin)

M138

Happy kangaroo reunion. Ten kangaroos set out jumping in turns along a straight road from a point A to point B in the following order: the first kangaroo jumps wherever it wants, the second jumps over the first twice the distance to the first (so that the first kangaroo finds itself exactly halfway between the takeoff and landing points of this jump), the third jumps over the second leaving it halfway behind, and so on, until the tenth jumps over the ninth, after which the first kangaroo again jumps at will, starting a new series of jumps. (a) Can all the kangaroos gather at B after ten series of jumps? (b) Can they gather at B earlier than that? (S. Eliseyev)

M139

Coprime-recycling cubic. Let $f(x) = x^3 - x + 1$. Prove that for any natural $m > 1$ the numbers m, $f(m)$, $f(f(m))$, $f(f(f(m)))$, ... are pairwise coprime. (A. Kolotov)

M140

Divisors of divisors. Let $d_1, \ldots, d_n$ be all the divisors of a positive integer N, and let δ_i, $i = 1, \ldots, n$ be the number of divisors of d_i. Prove that the numbers $\delta_1, \ldots, \delta_n$ satisfy this relation:

$$(\delta_1 + \delta_2 + \ldots + \delta_n)^2 = \delta_1^3 + \delta_2^3 + \ldots + \delta_n^3.$$

For instance, the number $N = 6$ has four divisors: 1, 2, 3, 6. The corresponding numbers δ_i are 1, 2, 2, 4, and we indeed have

$$(1+2+2+4)^2 = 81 = 1^3+2^3+2^3+4^3.$$

(V. Matizen)

Physics

P136

Horse runs in a circle. A horse runs with a constant speed v in a circle of radius R. A person stands at a distance r from the circle's center. What is the maximum speed at which the horse and the person approach one another? (A. Bytsko)

P137

Charged droplet. A charged droplet is suspended in the air by an electric field. At the moment $t_0 = 0$ the electric field begins to decrease and becomes zero at $t = t_1$. Figure 1 shows the droplet's acceleration versus time in relative units. Using this graph, find the droplet's maximum acceleration. Consider the air resistance to be proportional to the droplet's velocity. (A. Sheronov)

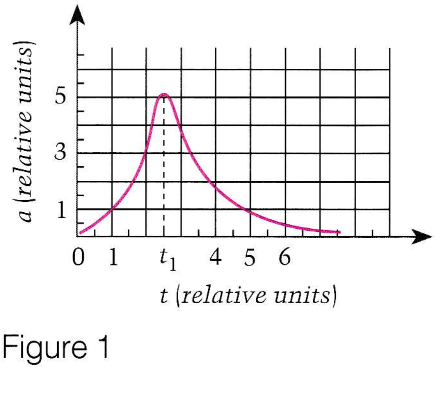


Figure 1

P138

Life of a soap bubble. Take a short tube of small diameter D and blow a soap bubble of radius $R_0 \gg D$. Now open the end of the tube and wait for the soap bubble to collapse. Evaluate the lifetime of such a bubble from the moment the tube is opened if $D = 2$ mm and $R_0 = 2$ cm. The surface tension of water is $\sigma = 0.07$ N/m. (V. Drozdov)

P139

Set of conductors. Several electrical conductors are located far from other physical bodies. The electrical potential of one of them is ϕ_1. This potential becomes zero when the charges on all the other bodies are changed to exactly the opposite ones. What will the potential of the first conductor be if its charge is now increased by a factor of four? (A. Zilberman)

P140

Sun screen. An image of the Sun is formed on a screen by means of concentric spherical mirrors (fig. 2). The

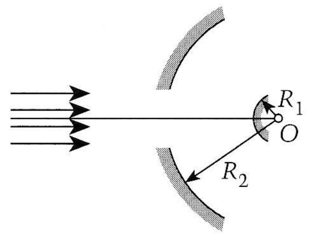


Figure 2

radii of the mirrors are $R_1 = 12$ cm and $R_2 = 30$ cm. What is the focal length of a thin lens that would give a solar image of the same size? (E. Kuznetsov)

ANSWERS, HINTS & SOLUTIONS ON PAGE 56

Bobbing for knowledge

Five experiments with a hollow plastic ball

by Pavel Kanayev

IN SPORTING GOODS STORES you can find a simple item that you'll need if you want to go fishing. It's a hollow plastic ball with two eyelets for attaching the fishing line and a little hole with a slightly protruding lip (fig. 1).[1] It's called a "bobber," and we can do some interesting physics experiments with it.

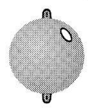


Figure 1

Before you start, you should learn how to fill the bobber quickly with water (or some other liquid). You can do this with a syringe (without the needle) or a medicine dropper, or you can rig up something with materials at hand (for instance, an empty ink tube from a ball-point pen and a small squeeze bottle).

Experiment 1. Dissolve several crystals of potassium permanganate in a small amount of hot water to obtain a dark-crimson solution. Fill the bobber with this solution, suspend it on a thread, and lower it into a glass jar filled with water at room temperature. (To make the bobber stable, tie a small "sinker"—which you can also get at the sporting goods store—or other weight to the bobber's lower eyelet.)

A colored stream will immediately begin to ascend from the opening, and its diameter is the same as the hole's (fig. 2). The stream eventually reaches the "ceiling" (that is, the surface layer), and then it trickles downward in several smaller streams. This goes on for a quite some time.

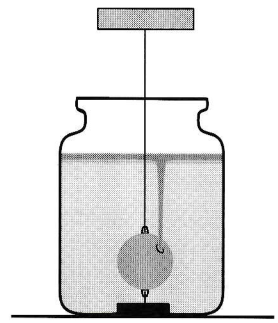


Figure 2

Now let's change the experiment slightly. Prepare the potassium permanganate solution at room temperature and drop the bobber into water at the same temperature—no colored stream will emerge from the opening.

As a third variant, add several drops of ethyl alcohol to the solution and repeat the experiment. The colored stream will rise up as before, and it will persist for a long time.

Why is the colored stream generated in the first and last cases, but not in the second?

In the first and third experiments, the densities of the hot water and of the water–alcohol mixture are less than that of the room-temperature water, so a buoyancy force arises. In the second experiment, the densities of the water inside and outside the bobber are the same, so there is no buoyancy.

Experiment 2. Fill the bobber with water up to the rim of the opening. Then, using a pipette, remove some water (about three pipettesful) and add the same amount of ethyl alcohol. Finally, seal the hole with modeling clay.

In about 20 minutes an air bubble (about the size of a pea) will form in the upper part of the bobber. Replacing the ethyl alcohol with a boric or acetic acid solution yields the same result. However, if the water is mixed with glycerin, no bubble will appear.

What is the explanation for this?

The kinetic theory of molecules says that there are gaps between molecules. When water is mixed with alcohol, the resulting volume of the mixture is less than the sum of the individual volumes of water and alcohol, so a bubble appears in the mixture. The same phenomenon can be observed when water is mixed with boric or acetic acid, but the total volume does not decrease when water is mixed with glycerin.

Experiment 3. Take the bobber and touch the edges of its opening to the surface of the water. The opening is covered with film, but in 30–40 seconds it breaks (with a little "pop") and then a drop of water appears near the lip of the opening.

Cover the hole with a film again, but this time insert a needle in the middle of the film and hold it there. Now the film remains intact for a very long time.

Why is the film so short-lived in the first case and so stable in the second?

The water film that forms over the opening is thicker near the lip

Art by Pavel Chernusky

[1] If you can't find a bobber with such a hole, simply drill a hole.—*Ed.*

than at the center, because the water "wants" to wet the edge. So if the film plane isn't strictly horizontal (which is quite probable), the water moves downhill and the film becomes thinner. At some point the film breaks and the water forms a drop.

The long life of the film in the second case is due to the fact that the needle is wetted, which causes the water to shift from the lip to the point where the needle makes contact with the water. The film becomes thicker and stronger.

Experiment 4. Fill two 0.5-liter glass jars up to the brim with water at room temperature. Using a pipette, add a few drops of shampoo diluted with water (in the ratio 1 : 4) to one of them.

Fill the bobber with pure water and carefully immerse it in the jar with the shampoo solution. The bobber immediately drops to the bottom.

Now wash off the bobber and immerse it carefully in the jar with clean water. Even though it's almost completely submerged in water, the bobber doesn't sink, despite the fact that the force of gravity is somewhat greater than the force trying to push it out of the water.

Why does the bobber filled with water sink in one jar but not in the other? Will the bobber sink in a weak sugar solution?

The result depends on the surface tension of the fluid and also on whether this fluid wets the surface of the bobber. The coefficient of surface tension of pure water is almost twice that of the soap solution. It's also important that plastic is weakly wetted by water. That's why the bobber doesn't sink in pure water.

Sugar increases the surface tension of water, so the bobber will not sink in a weak sugar solution either.

Experiment 5. Take a wire 2–3 mm in diameter and about one meter long. File down one end of the wire, bend it at a right angle, and stick it into the eyelet nearest the opening. Now submerge the bobber slowly in a tank or pail full of water and then (slowly!) lift it to the surface. You'll see air bubbles coming out of the hole when the bobber is ascending. The deeper you submerge the bobber, the more bubbles come out of it.

Why do the air bubbles form and break away as the bobber rises?

When the bobber descends, the air inside is compressed and water enters the bobber. As the bobber rises, the external pressure decreases, the air in the bobber expands, and bubbles form at the rim of the opening. The bubbles grow larger, but a collar forms around the opening, making the opening ever narrower. As this happens, the buoyant force acting on the bobber (directed upward) increases, while the surface tension restraining the bubble decreases. When these forces become equal, the bubble breaks away, assuming a spherical shape.

The most mysterious shape of all

A spiral primer

A SPIRAL IS A CURVE ON THE plane traced by a point that winds around a certain fixed point (the spiral's *pole*), approaching or receding from it depending on the direction of motion. The word "spiral" means a coil and sounds almost the same in Greek (σπειρα) and in Latin (*spira*). Spirality can be regarded as a type of symmetry[1] essentially different from the symmetry of a snowflake, atomic nucleus, or chessboard, which could be called spherical symmetry. While spherical symmetry is characterized by the constancy of the radius drawn from the center as the figure is turned to fit onto itself, spiral symmetry allows the radius to change. Spirality is thus a more fundamental property of matter, whereas spherical symmetry is a particular, exclusive case of spirality.

Mathematically, spirals are best described by means of polar coordinates. Let r be the distance from the pole O to a point M on a spiral, and let θ be the angle between OM and the fixed axis OA (the polar axis) (fig. 1). The interesting case of a spiral that reaches its pole only in the limit as $\theta \to \infty$—that is, after an infinite number of windings—can be given by the equation $\theta = kr^{-n}$, where $k > 0$, $n > 0$ are constant, $r > 0$, $0 < \theta < \infty$. Spirals with equations of this form are called algebraic. This form can be regarded as the first term in the expansion of a more general function $\theta = \theta(r)$ in powers of r; the remaining terms are small near the spiral's pole, but may play the main role far away from it. Depending on the exponent n, we recognize three types of algebraic spirals.

A spiral is called hyperbolic if $n = 1$. The equation then takes the form $\theta = k/r$. As $\theta \to 0$, the distance from a point on such a spiral to the polar axis stabilizes, since it's equal to $r \sin\theta \cong r\theta = k$. It follows that a hyperbolic spiral has an asymptote—the straight line it approaches as $\theta \to \infty$. The hyperbolic spiral was described by the French mathematician Pierre Varignon (1654–1722).

If $n > 1$, then $r \sin\theta \cong r\theta = kr^{1-n}$ for small θ—that is, in this case the asymptote coincides with the polar axis OA. For $n = 2$, this kind of spiral is called the *lituus*, which means "crook" (in the sense of a shepherd's staff). The term was used by Colin Maclaurin in 1722, but the curve was first described by Rodger Cotes in 1714.

For $n < 1$, an algebraic spiral has no asymptote; the distance from the spiral to the polar axis increases approximately as kr^{1-n} (as $r \to \infty$).

Another kind of spiral is the *pseudospiral*. In equations for pseudospirals θ isn't expressible as a power of r. An example of a pseudospiral is the *logarithmic spiral* defined by the equation $\theta = -k \ln r$. The angle θ here varies from $-\infty$ as $r \to \infty$ to ∞ as $r \to 0$ (fig. 2). A remarkable property of the logarithmic spiral is that it meets any ray from its pole at the same angle α (try to prove this yourself). That's why it's often called the *equiangular* spiral. This curve was described in 1638 by the great French philosopher and mathematician René Descartes (1596–1650). It has a lot of applications in engineering: rotating knives, milling cutters, and gears are often made in this shape.

If a logarithmic spiral is dilated A times and at the same time is rotated by an angle of $-k \ln A$, the spiral fits onto itself (because the equation remains valid). This is only one of the "reproductive" properties of this spiral. Many others were found by Jacob Bernoulli (1654–1705), who was so deeply impressed by his discovery that he asked that the curve be chiseled on his tombstone with the Latin inscription *Eadem mutate resurgo* ("Though changed, I shall arise the same" in E. T. Bell's translation).

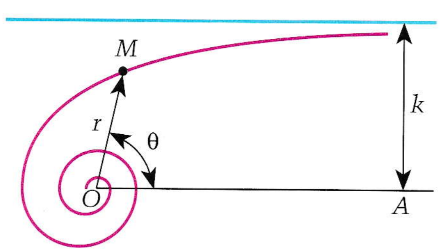


Figure 1

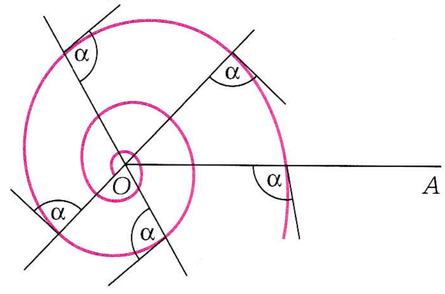


Figure 2

[1]That is, a way of making an object fit onto itself using a certain transformation.—*Ed.*

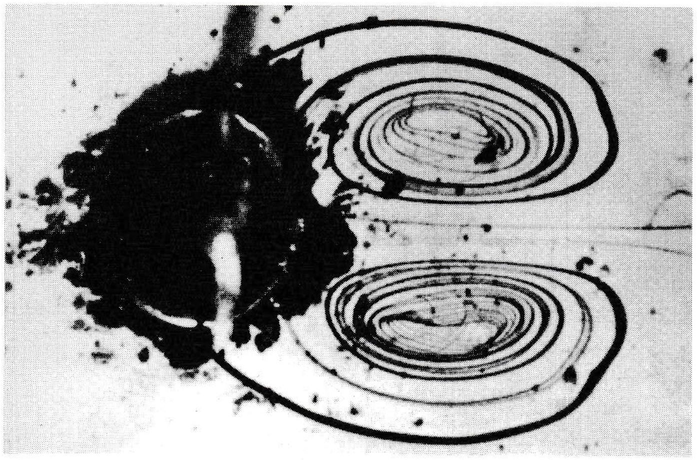


Figure 3

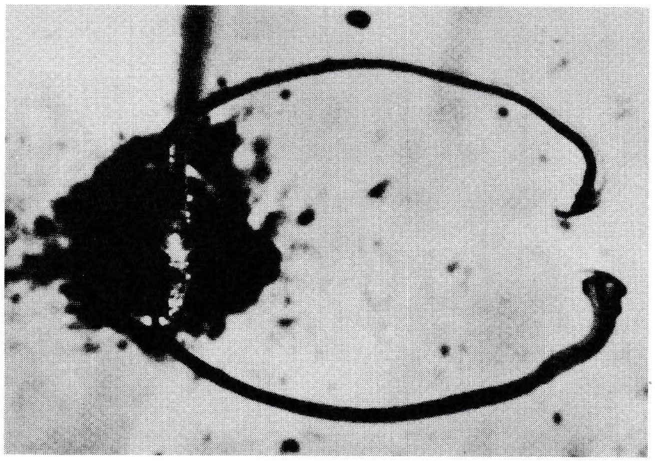


Figure 4

A spiral can have an infinite number of coils not only in the neighborhood of its pole, but also in the "neighborhood of infinity"—that is, as a point tracing it recedes from the pole to infinity. A few examples of this are the Archimedean spiral $\theta = kr$, the Galilean spiral $\theta^2 = k(r - r_0)$, and the parabolic spiral $\theta^{1/2} = k(r - r_0)$. It's also possible that a spiral winds infinitely many times about a certain curve approaching it in the limit as $\theta \to \infty$. For instance, the spiral $\theta(r - r_0) = k$ coils around the circle $r = r_0$.

On a molecular level, we find spiral (or helical—see next page) structures in DNA molecules, and on a galactic scale there are giant spiral galaxies. As to our own world, nestled between these two, we come across spiral structures at every turn. An acceleration eddy forms at the end of an oar thrust into the water; it gains strength during the stroke and moves behind the stern as the stroke is finished. Spiral eddies are intricately shaped. They can be both algebraic (fig. 3) or logarithmic near the pole. A logarithmic spiral with a small value of k is shown in figure 4: its coils are practically invisible, because the radius r decreases sharply (in mathematical terms, exponentially) with the growth of θ. The photographs in figures 3 and 4 were taken in a hydrodynamic tunnel as an accelerated stream of water flowed symmetrically around a plate, with pigment fed onto the plate's edges.

Spiral whirling occurs not only when a liquid flows around an obstacle, but also when it flows out of a slot. In figure 5 you see a photo of . . . an atomic explosion? No, it's only the leading edge of a jet issuing upward from a narrow slot. Initially, the liquid above the plane of the slot was colored with ink. The flow started from the steady state, where spiral structure is hardly seen. But it immediately reveals itself as soon as any pigment (for instance, our ink) is fed onto an edge of the slot. (See figure 6, where the slot has only one edge; the other one is replaced by a solid wall that can be considered the slot's symmetry, so it's a portrait of the flow out of a "half-slot.")

Figure 5

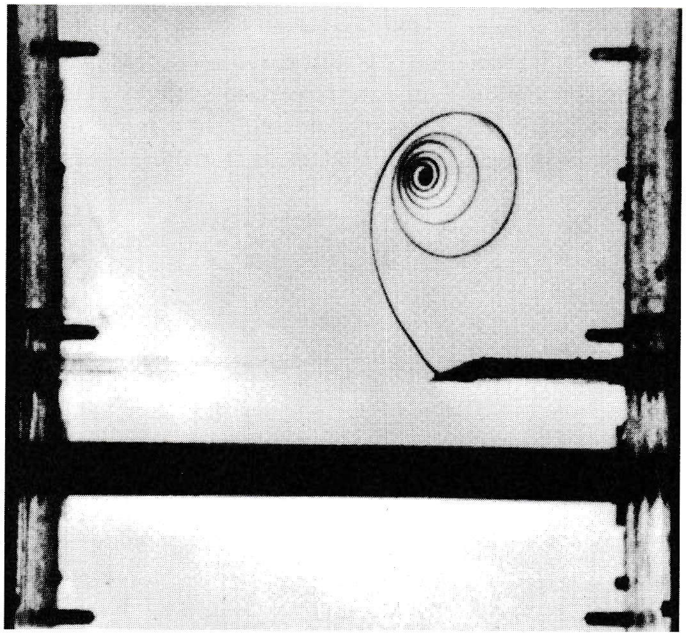


Figure 6

Nature, both organic and inorganic, is full of spiral shapes. We find them in the shells of most ordinary snails and in ancient fossils—the ammonite in figure 7 (on the next page) is about 180 million years old. Sunflower seeds

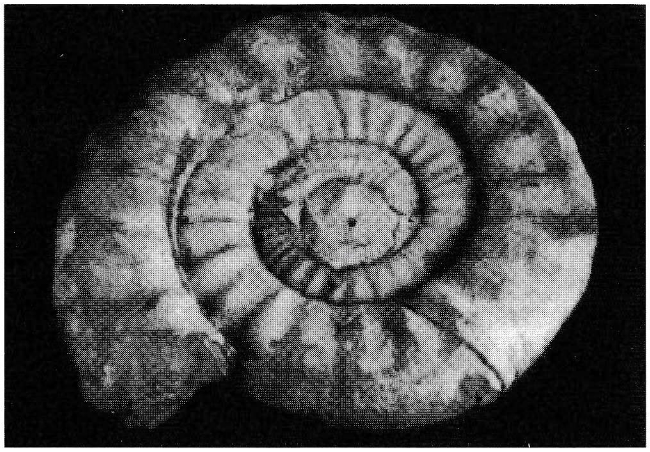


Figure 7

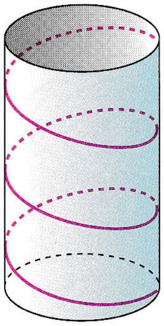


Figure 8

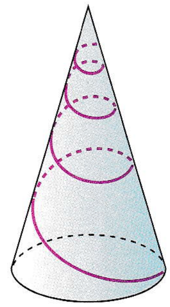


Figure 9

in their pod form two families of oppositely twisted spirals. In botany, the tendency to spirality is called phylotaxis. This phenomenon often manifests itself in *helical* arrangements, too. A *helix* is a kind of three-dimensional spiral—it's the curve traced by a point rotating about a certain axis and moving along this axis at the same time. More exactly, this is a cylindrical helix (fig. 8). Branches on a stem often grow along such a curve. If the point that traces a helix, in addition to moving around and along the axis, recedes from (or approaches) it, we get a curve on a circular cone—a conical helix (fig. 9). This curve is found in the arrangement of scales on a fir cone.

A helix can be twisted like the letter *S* (fig. 10) or *Z* (fig. 11), where the middle elements of the letters are thought of as lying on the visible side of the imaginary cylinder around which they wind. *S*- and *Z*-helices are mirror images of each other, as is seen quite clearly in the horns of antelopes and other horned creatures (fig. 12). The helical shape is taken by various lianas—flowering or fern plants unable to keep their stem erect on their own, without propping it up against a rock, building, or other plants. Hops, ivy, wild grapes, and blackberries are all lianas. They developed their knack for winding over the course of evolution as a part of their struggle for light. First a sprout, having emerged from the ground, stretches upward, then its tip begins to perform circular movements (in what direction?—make your own observations!) to find support. If a support isn't found, the plant leans back on the ground, grows up a little more, and resumes its "roundabout" exploration.

So far we've been considering curves. But there are surfaces winding about a certain, generally curved, axis (the dotted curve in figure 13) such that their transverse sections (with respect to the axis) are spirals. These are called spiral surfaces. Spatial spiral structure is found in certain atmospheric phenomena, such as cyclones and tornadoes. Figure 14 shows a cyclone over the Indian Ocean photographed by the Kosmos-144 satellite. A waterspout over a lake is seen in figure 15. This destructive column sweeps away everything it meets. In 1905 the oceanographer W. Eckmann discovered spatial spiral underwater currents characterized by a balance between Coriolis forces and frictional forces. The Eckmann current is often observed in the notorious Bermuda Triangle. Similar spiral motions in the upper layers of the atmosphere were investigated by the English scientist J. Taylor in 1915. These ocean and atmospheric spirals result from the Earth's rotation about its own axis.

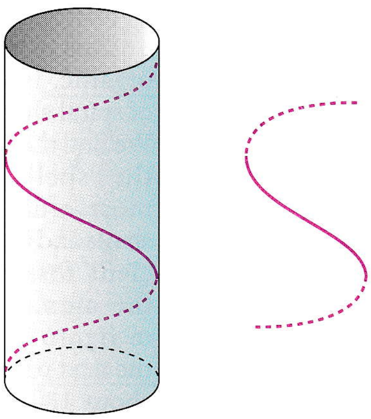


Figure 10

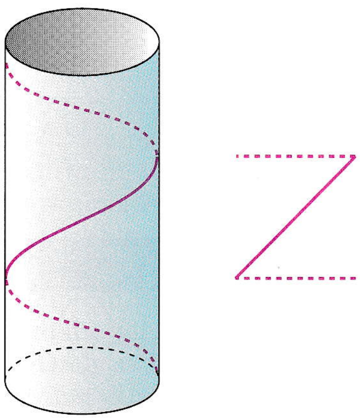


Figure 11

Figure 12

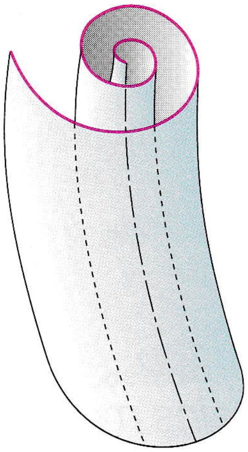


Figure 13

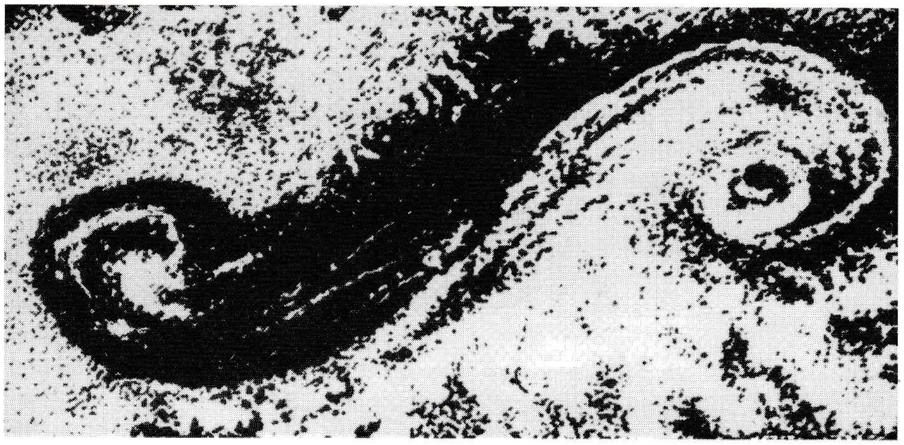


Figure 14

Spiral forms are encountered so frequently that it's impossible even to name the variety of their manifestations. A charged particle in a magnetic field (of the Earth, an accelerator, or a thermonuclear reactor) moves in a helix whose axis coincides with the direction of the field (the blue curves in figure 16). Spiral waves are generated by something called spin detonation. They are also observed in the fundamental Belousov–Zhabotinsky chemical reaction. A theory of this reaction has yet to be formulated, even though thousands of specialists are working on the problem. Many biologists think that spiral waves are responsible for heart arrhythmia and other biological phenomena.

Figure 15

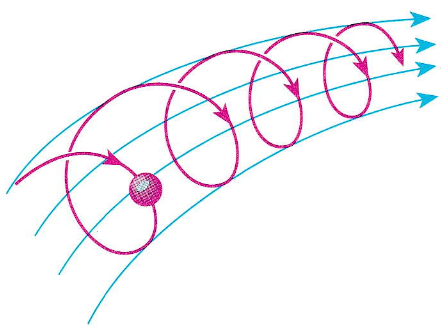


Figure 16

Perhaps this is where certain fundamental philosophical theories of spiral development of the spirit and of nature should be mentioned. What, indeed, is the world around us? Is it expanding? Infinite? Eleven-dimensional? Random? Protein? These questions have not yet been answered by scientists. As for spirality, the examples given here provide weighty proof that our world is spiral. However, we have a long way to go in explaining the mystery of how spirals emerge and persist. ◘

—Submitted by N. Bourbaki

Weighing an astronaut

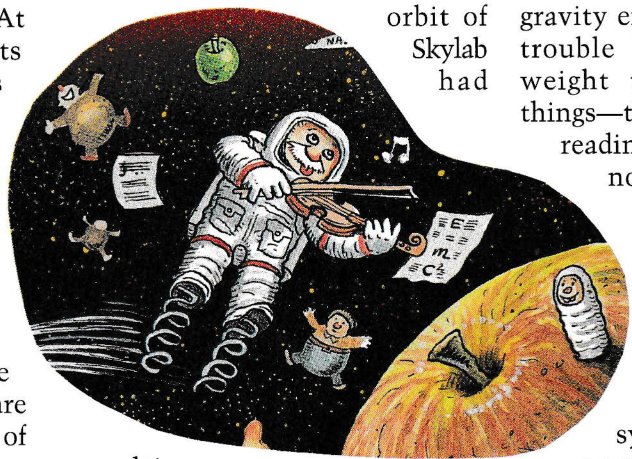
"The earth before us is a handful of soil, but it sustains mountains without feeling their weight and contains the rivers and seas without their leaking away."—Confucius

by Arthur Eisenkraft and Larry D. Kirkpatrick

ONE DAY SOME OF US WILL be living for extended periods of time in a space colony. What health problems might arise? Will we lose weight, or will our bones weaken in "zero gravity"? Medical questions were very important during the Skylab mission from May 1973 until February 1974. At the most basic level the scientists wanted to know if the astronauts would lose weight during prolonged stays in space. Let's begin by taking a closer look at the concept of weighing an astronaut.

When you want to know your weight in the morning, you simply step on a bathroom scale and read your weight. But how does the scale "know" your weight? You are actually measuring the amount of stretch or compression of a spring inside the scale. If we assume that the spring is ideal (that is, it obeys Hooke's Law), we know that $F = -kx$, where F is the force of the spring, k is the spring constant (a measure of the stiffness of the spring), and x is the extension of the spring. In this case, the applied force is just the force of gravity acting on you. Since the force of gravity is given by mg, if you have a mass of 60 kg and the local acceleration due to gravity is 9.80 m/s^2, the reading will be 588 N, or an equivalent value in pounds, stones, or kilograms. It seems that only physics teachers' scales are calibrated in newtons!

How much would you weigh if you had been a resident of Skylab in 1973? The orbit of Skylab had



Art by Tomas Bunk

an altitude of 386 km above the Earth's surface, or a radius of 6,378 km + 386 km = 6,764 km. Since gravity is an inverse-square force, you can calculate the force of gravity on you in Skylab:

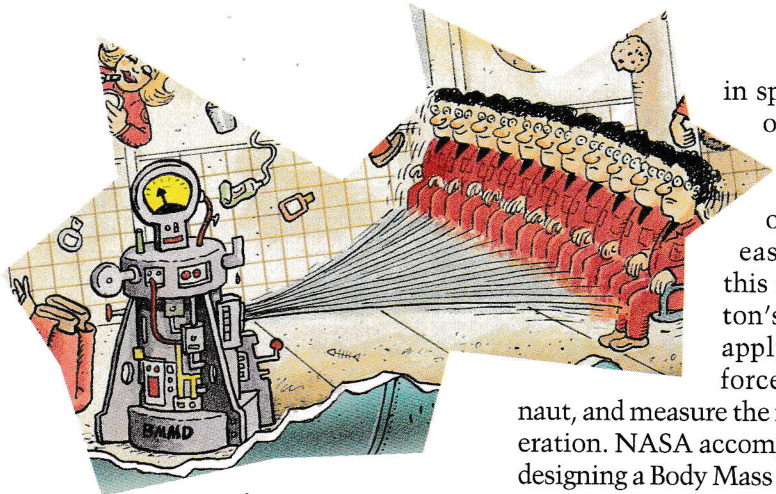
$$F = F_E\left(\frac{R}{R_E}\right)^2 = (588\ \text{N})\left(\frac{6{,}378\ \text{km}}{6{,}764\ \text{km}}\right)^2$$

$$= 523\ \text{N},$$

where the subscript E refers to Earth values. Therefore, you expect a scale reading of 523 N.

However, when you step on the bathroom scale in Skylab, it reads zero! In the vernacular of space, you are weightless. What does it even mean to "step on a scale" in a zero-gravity environment? We get into trouble because we have used weight to mean two different things—the force of gravity and the reading on the scale. This does not cause complications if you are in an inertial system—that is, a system that is not accelerating. We assume that your bathroom on Earth is an inertial reference system. (It is not really an inertial system because the Earth is rotating on its axis and revolving about the Sun, but it is approximately an inertial system.)[1] When you stand on the bathroom scale, you have no acceleration and Newton's Second Law tells us that there is no net force acting on you. Therefore, the downward force of gravity must be equal in size to the upward force of the scale acting on you, and it does not matter which

[1]See "The 'Most Inertial' Reference Frame" on page 48.—*Ed.*



force we measure.

Problems arise when your reference system is accelerating. Because Skylab is in orbit about Earth, it has a centripetal acceleration of v^2/R. And so do you—you are also in orbit. The force of gravity provides the centripetal acceleration needed to keep you in orbit and there is no need for the scales to hold you up (or anything else, for that matter).

We can create a similar situation on Earth. Some amusement park rides take you to the top of a structure and drop you over the edge. If you were to stand on a bathroom scale during this free fall, it would read zero. Both you and the scale are in free fall and the scale does not need to support you. For this reason, many physics teachers carefully distinguish between the force of gravity and the weight. Weight is the reading on the bathroom scale, or the support force needed to keep you at rest in the noninertial reference system. Other teachers use the term "apparent weight" to refer to the scale reading and use weight to refer to the force of gravity.

Let's return to the elevator and cause it to accelerate *upward* at 9.80 m/s^2. What will you weigh? The net upward force on you must be mg to give you an upward acceleration of g. Because the force of gravity is mg downward, the spring in the scale must push upward with a force of $2mg$. Therefore, the scale would read $2 \cdot 588$ N = 1,176 N.

We could have avoided our weight problem by realizing that the NASA scientists were really concerned with the mass of the astronaut. But how do we measure the mass of an astronaut in space? We obviously cannot ask the astronaut to stand on a scale. An easy way to do this is to use Newton's Second Law, apply a known force to the astronaut, and measure the resulting acceleration. NASA accomplished this by designing a Body Mass Measuring Device (BMMD for short). It is basically a chair mounted on a pair of leaf springs. Whenever a mass m is attached to a spring with a spring constant k and displaced a small distance from the equilibrium position, it executes simple harmonic motion with a period T given by

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

In practice, an astronaut sits in the BMMD and measures the period of oscillation to obtain the mass. This brings us to this month's contest problems.

A. The BMMD was calibrated by loading the chair with a known mass and measuring the corresponding period of oscillation. Graph the following data supplied by NASA to find the combined spring constant of the leaf springs and the mass of the empty chair for the BMMD:

Mass (kg)	Period (s)
0.00	0.90149
14.06	1.24979
23.93	1.44379
33.80	1.61464
45.02	1.78780
56.08	1.94442
67.05	2.08832

B. One of the crew members in the second group of three astronauts to live in Skylab was electrical engineer Dr. Owen K. Garriott. He measured periods of 2.012 s and 1.981 s while sitting in the BMMD at the beginning and end of a 58-day interval. How much weight (read mass) did he gain or lose during this time?

C. Assume that you ride a skateboard down an inclined plane that makes an angle θ with the horizontal. A scale has been mounted on the skateboard in a horizontal position as shown in figure 1. There is no friction between the skateboard and the inclined plane and you have a mass of 60 kg. What does the scale read? (A version of this question appeared on the preliminary examination used to select members of the 1995 US Physics Team that will compete in the International Physics Olympiad in Canberra, Australia, in July.)

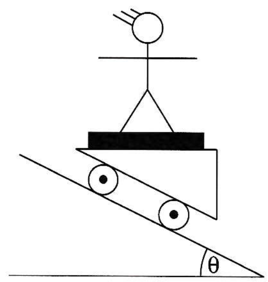


Figure 1

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*.

Rising star

In our problem in the September/October issue, the radio receiver records maxima and minima. This is our clue that some interference effect is occurring. The insight that solves the problem comes from the description of how the receiver is placed on an island near the shore. The interference is probably a result of the electromagnetic wave arriving directly from the star and a second electromagnetic wave arriving after reflecting from the surface of the water. Figure 2 shows these two rays.

The problem can be solved by recognizing that the reflected wave has traveled a longer path and has undergone a phase change at the surface

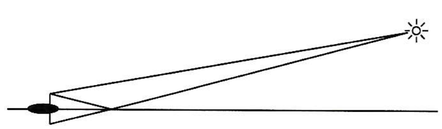


Figure 2

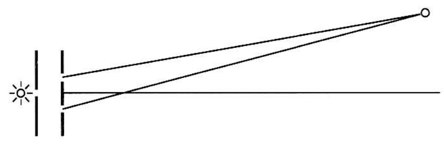


Figure 3

equal to $\lambda/2$. The total path difference must be equal to an integral number of wavelengths for constructive interference (producing maxima) and an odd half-integral number of wavelengths for destructive interference (producing minima.)

The geometry, though not difficult, is unfamiliar. By drawing reflections of the reflected ray and the radio receiver, we are reminded of Young's double slit experiment, in which light emerging from a pair of slits forms an interference pattern on a distant screen. Figure 3 shows Young's double slit geometry for comparison. Young's double slit geometry is analyzed in all physics texts covering optics. By drawing a line perpendicular to the line-of-sight to the star as shown in figure 4, we see that the path difference δ is given by

$$\delta \cong 2h \sin\theta,$$

where θ is the altitude of the star, h is the height of the radio tower, and we have assumed that for small θ the triangle is approximately a right triangle.

In our radio receiver problem, the conditions for maxima and minima (remembering the phase shift upon reflection from the water) are

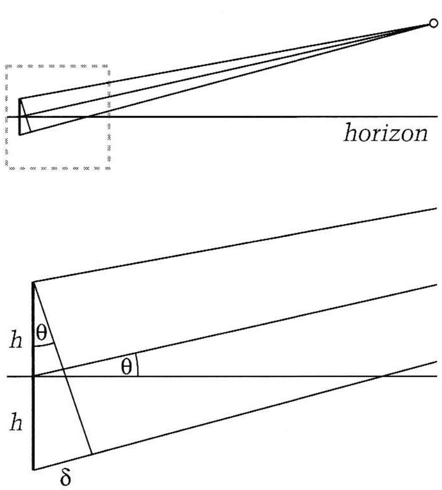


Figure 4

$$\delta_{max} = \left(k + \frac{1}{2}\right)\lambda$$

and

$$\delta_{min} = k\lambda,$$

where k is an integer 0, 1, 2, 3, … .

Part A of the problem asked for the altitude of the star when maxima and minima are observed:

$$\sin\theta_{max} = \frac{\delta_{max}}{2h} = \frac{\left(k + \frac{1}{2}\right)\lambda}{2h},$$

$$\sin\theta_{min} = \frac{\delta_{min}}{2h} = \frac{k\lambda}{2h}.$$

Part B asked whether the intensity increases or decreases as the star rises over the horizon. The star just peeks over the horizon when $\theta = 0$. Since this corresponds to a minimum, the intensity of the star will increase until it reaches its first maximum.

Part C asked for an investigation of the intensities of successive minima and maxima given that the ratio of the incident and reflected rays at the water's surface is

$$\frac{n - \sin\theta}{n + \sin\theta},$$

where $n = 9$.

The maximum amplitude can be found by adding the incident and reflected electric fields. Assuming that the incident electric field is E, the amplitude would be

$$amplitude_{max} = E + E\frac{n - \sin\theta}{n + \sin\theta}$$

$$= E + E\left(\frac{n - \frac{\lambda\left(k + \frac{1}{2}\right)}{2h}}{n + \frac{\lambda\left(k + \frac{1}{2}\right)}{2h}}\right)$$

$$= \left(\frac{4nh}{2nh + \lambda\left(k + \frac{1}{2}\right)}\right)E.$$

k	θ_{min}	θ_{max}	$intensity_{min}$	$intensity_{max}$
0	0	1.50°	0	3.9768
1	3.01°	4.52°	0.000135	3.9309
2	6.03°	7.54°	0.000532	3.8858
3	9.06°	10.59°	0.00118	3.8415

Figure 5

Since intensity is proportional to the square of the amplitude,

$$intensity_{max} \sim \left[\left(\frac{4nh}{2nh + \lambda\left(k + \frac{1}{2}\right)}\right)E\right]^2.$$

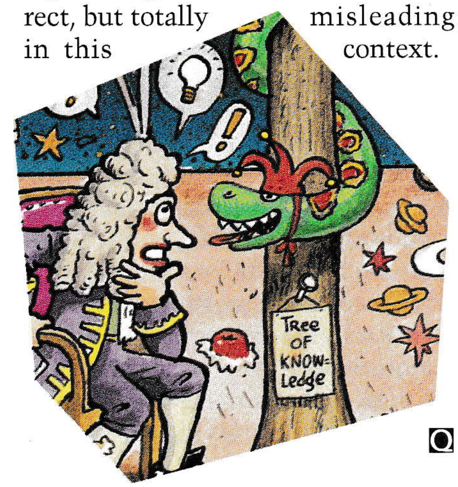
Similarly, the minimum amplitude can be found by subtracting the incident and reflected electric fields. The corresponding intensity is:

$$intensity_{min} \sim \left[\left(\frac{2k\lambda}{2nh + \lambda k}\right)E\right]^2.$$

Relative numerical values can now be computed. They are summarized in the spreadsheet chart in figure 5.

Correction

If you are having difficulties with Part D of the contest problem on the "superconducting magnet" that appeared in the November/December issue, it's probably due to our error. The last part of Part D should read: "However, we will destroy the switch if the current through the switch *in the normal state* exceeds 0.5 A. What steps can you use to shut the magnet down?" The parenthetical remark about large currents causing the magnet to switch states is correct, but totally misleading in this context.



Distinct sums of twosomes

Revisiting an old problem from the "Competition Corner"

by George Berzsenyi

IN THIS ISSUE I WANT TO REvive an old problem that was proposed (as Problem 547) in my "Competition Corner" in the December 1980 issue of the now defunct journal *Mathematics Student*. At that time I was not aware of the fact that a simpler version of the problem was already attempted in *The Fibonacci Quarterly*, or that a somewhat related problem was also posed in the 1979 British Mathematical Olympiad. Had I been aware of these, I would have been less ambitious in my original formulation thereof:

Let N be a set of twelve positive integers such that for distinct a, b, c, d in N, $a + b \neq c + d$. Prove that the largest element of N is greater than 67.

At that time 67 seemed a reasonable lower bound, since the best N that I could construct was

$$\{1, 2, 3, 5, 8, 21, 29, 37, 46, 60, 71, 83\}.$$

Unfortunately, there was a flaw in my solution, so I had to replace 67 with 35 in the problem to make it more reasonable. The choice of 35 was prompted by the observation that there are 66 distinct sums, the smallest of them being greater than or equal to $3 = 1 + 2$, hence they could—ideally—range from 3 to 68. Since 68 must be the sum of the two largest members of N, it is immediately clear that the largest member of N must be at least 35, and it takes only a bit of reasoning to show that it must be larger than 35. My first challenge to my readers is to provide the reasoning needed.

At that time I was fortunate to have seven of the eight winners of the 1980 USA Mathematical Olympiad and many other super-talented students among the participants of my program, and most of them did not settle for 35 as the best possible answer, especially since I first posed the problem with 67 as the intended solution. So a number of them came up with various improvements. I'll sketch some of these below in order to whet my readers' appetites for even more ambitious attempts, and to indicate the variety of possible approaches to the problem.

One student observed that if the elements of N are $a_1 < a_2 < a_3 < \ldots < a_{12}$, then at most two of the differences $a_{12} - a_{11}$, $a_{11} - a_{10}$, $a_{10} - a_9$, ..., $a_2 - a_1$ can be the same, and from this he could deduce that $a_{12} > 37$. Another student took a closer look at these differences, and (by considering whether 1 appears once, twice, or not at all as a difference) managed to prove that their pairwise equality is even more limited, and hence pushed the lower bound to 46. Several other participants went even further by observing that at least 56 of the differences $a - b$, with $a, b \in N$, must be distinct. To see this, observe that for $a, b, c \in N$ it is possible to have $c - a = b - c$ if c is not the smallest or largest element of N, and so there may only be 66 – 10 distinct differences. Their reasoning led to $a_{12} > 56$.

Other students managed to come up with better examples of N, like

$$\{1, 2, 3, 5, 8, 13, 21, 30, 39, 53, 67, 82\}$$

and the even better

$$\{1, 2, 3, 5, 12, 19, 36, 42, 48, 61, 69, 74\},$$

but clearly the gap is still too large. Therefore, I hereby challenge my readers to narrow the gap.

The answer is not even known for sets of $n < 12$ elements. Unfortunately, there doesn't seem to be any recursion present, hence no way to construct an $(n + 1)$-element N from an n-element N. It's true that if N is such a set of n elements, then so is $\{k + 1 - n \mid n \in N\}$, but this fact doesn't seem to be very helpful, except possibly to suggest that the smallest members of N need not be closer together than the largest ones.

At this point, some computer-generated data might also be helpful, and one might even consider selecting the members of N randomly at first and then making replacements. I fully believe that it is time to make further advances on this problem, and I encourage my readers to do so.

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

Lewis Carroll's sleepless nights

Two probability problems for insomniacs

by Martin Gardner

LEWIS CARROLL WAS THE PEN name of Charles Dodgson, who taught mathematics at Christ Church, one of the Oxford University colleges in England. He is best known, of course, as the author of two immortal fantasies about Alice and a long nonsense ballad called *The Hunting of the Snark*.

In 1893 Carroll published a little book of seventy-two original mathematical puzzles, many of them not easily solved. The book's title was *Pillow-Problems Thought Out During Sleepless Nights.* For the book's second edition he changed the last two words to "wakeful hours" so readers wouldn't think he suffered from chronic insomnia. A new preface was added to the fourth edition (1895). Carroll intended the book to be part II of what he called *Curiosa Mathematica.* Part I, *A New Theory of Parallels*, was too serious to be called recreational even though it was written with the usual Carrollian humor.

The most interesting puzzles in *Pillow-Problems* concern probability. The first one, problem 5, is simple to state but extremely confusing to analyze correctly:

> A bag contains one counter, known to be either white or black. A white counter is put in, the bag shaken, and a counter drawn out, which proves to be white. What is now the chance of drawing a white counter?

As Carroll writes, one is tempted to answer 1/2. Before the white counter is withdrawn, the bag is assumed to hold with equal probability either a black or white counter, or two white counters. If the counters in the bag are black and white, a black counter will remain after the white one is taken. If the counters are both white, a white counter will remain after a white one is drawn. Because the two states of the bag are equally probable, it seems that after a white counter is taken, the remaining counter will be black or white with equal probability.

Carroll claims correctly that the above argument, though intuitively plausible, is dead wrong. Let *A* stand for a white counter in the bag at the

Art by Sergey Ivanov

outset, B for a black counter, and C for the added white counter. After a white counter is taken, there are three, not two, equally possible states:

1. C has been taken, leaving A.
2. A has been taken, leaving C.
3. C has been taken, leaving B.

In the first two cases a white counter remains in the bag. In the third case, the remaining counter is black. The somewhat surprising answer, therefore, is 2/3.

The probability of first drawing a white counter is 3/4, and the probability that the remaining counter is white is also 3/4. Of course, as soon as you see that the counter taken is white, the probabilities alter. If black, the other counter is white with certainty. If white, the other counter is white with a probability of 2/3, and black with a probability of 1/3. All this can be made clear with an inverted tree diagram (see the figure below).

The fractions represent probabilities. The probability of each of the four outcomes (bottom row) is 1/2 times 1/2, or 1/4. The diagram shows that three times out of four a white counter will be drawn, and three times out of four a white counter remains in the bag. If drawing a black counter is not considered—assume that if this happens the black counter will be replaced and drawing continued until a white counter is taken—the remaining counter is white two times out of three.

The problem is easily modeled with playing cards. Shuffle a deck, spread it face down, and remove a card without looking at its face. Beside it place face down a card you know to be red. Turn your back while a friend mixes the positions of the two cards. Turn around and put a finger on one card. The chance that it's red is 3/4, and the chance the other card is red is also 3/4. Turn over the card you're touching. If it's black, the other card *must* be red. If it's red, the probability the other card is red goes down to 2/3.

The book's last problem, № 72, has been the subject of much controversy.

> A bag contains 2 counters, as to which nothing is known except that each is either black or white. Ascertain their colours without taking them out of the bag.

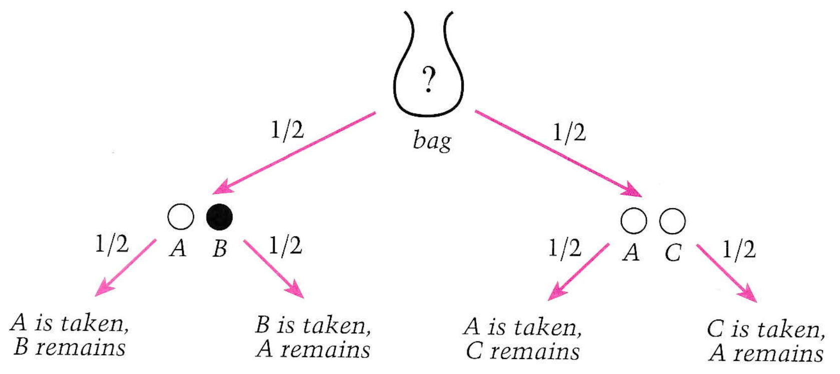
Here is Carroll's surprising answer:

> We know that, if a bag contained 3 counters, 2 being black and one white, the chance of drawing a black one would be 2/3; and that any *other* state of things would *not* give this chance.
>
> Now the chances, that the given bag contains (α) *BB*, (β) *BW*, (γ) *WW*, are respectively 1/4, 1/2, 1/4.
>
> Add a black counter.
>
> Then the chances, that it contains (α) *BBB*, (β) *BWB*, (γ) *WWB*, are, as before, 1/4, 1/2, 1/4.
>
> Hence the chance, of now drawing a black one,
>
> $= 1/4 \cdot 1 + 1/2 \cdot 2/3 + 1/4 \cdot 1/3 = 2/3.$
>
> Hence the bag now contains *BBW* (since any *other* state of things would *not* give this chance).
>
> Hence, before the black counter was added, it contained *BW*, i.e. one black counter and one white.

The proof is so obviously false that it's hard to comprehend how several top mathematicians could have taken it seriously and cited it as an example of how little Carroll understood probability theory! There is, however, not the slightest doubt that Carroll intended it as a joke. He answered all thirteen of the other probability questions in his book correctly. In the book's introduction he gives the hoax away:

> If any of my readers should feel inclined to reproach me with having worked too uniformly in the region of Common-place, and with never having ventured to wander out of the beaten tracks, I can proudly point to one Problem in 'Transcendental Probabilities'—a subject in which, I believe, *very* little has yet been done by even the most enterprising of mathematical explorers. To the casual reader it may seem abnormal, and even paradoxical; but I would have such a reader ask himself, candidly, the question "Is not Life itself a Paradox?"

It was characteristic of Carroll that he ended his book with a choice specimen of Carrollian nonsense. ◘



The legacy of Norbert Wiener

Part III: From feedback to cybernetics

IN 1933 WIENER BECAME ACquainted with Arturo Rosenblueth, a Mexican physiologist who was leading a series of interdisciplinary seminars at the Harvard Medical School. They hit it off well and began a long association during which Wiener's ideas on the relationship between mechanical and physiological systems—particularly in regard to the role of feedback—came to fruition. It appears that his interaction with Rosenblueth also set in motion the train of thought that would evolve into *cybernetics*. Thus, from an intellectual and scientific standpoint, their collaboration was an enormous success. In addition, judging from the warmth with which Wiener writes of him, Rosenblueth became the closest friend of his adult life.

The concept of a *feedback loop* was already familiar to James Watt in the 18th century, and today it is so deeply embedded in our thought processes that we hardly recognize it. An everyday example of a feedback loop is the one connecting a furnace to a thermostat. The furnace puts out heat, raising the temperature of the room. The thermostat senses the temperature, and if it gets too low, the thermostat completes a circuit and ignites the furnace. The furnace then continues to pump out heat until the temperature gets too high, at which point the thermostat breaks the circuit, and the furnace shuts down. In this way, the output of the furnace is fed back into the input.[1] What fascinated Wiener were unstable feedback mechanisms. Most of us know the difficulty of carrying a too-full bowl of soup to the dinner table: the soup begins to slosh and any attempt to settle it by tilting (negative feedback) only makes matters worse. Wiener and Rosenblueth proposed to model certain muscle spasms (intention tremors) using an unstable feedback loop. Later they used the same principles to study the heart muscle.

With the outbreak of World War II, Wiener had to defer these investigations. Confronted by what appeared to be the imminent collapse of European civilization, Wiener, like many scientists, searched for a way to contribute to the war effort. The problem he eventually chose was that of aiming anti-aircraft guns. This was a much more sophisticated problem than the ones he had worked on in World War I. Airplanes had become much faster and more dangerous, and so the human gunner had to be assisted by a machine. Moreover, it was no longer sensible to aim directly at the plane—by the time the shell got there, the plane would have moved on. The problem was one of *prediction*. That is, one had to determine the plane's position by radar signals and *predict* its future trajectory. Since it was clear that there was no hope of making a perfect prediction, Wiener decided to adopt a statistical approach. In other words, he devised a statistical model in which he could formulate precisely what it means to *maximize the probability of success*.

A central difficulty addressed by Wiener's statistical model was that if one tries to control the action of the gun too closely from the radar data, measurement errors can cause the gun to go into wild oscillations. Human gunners have no trouble adjusting to imperfect measurements, but a machine had to be designed specifically to prevent instabilities. Wiener compensated for the imperfections of the radar data by

Parts I and II appeared in the November/December 1994 and January/February 1995 issues, respectively. Reprinted from the program booklet for The Legacy of Norbert Wiener: A Centennial Symposium in Honor of the 100th Anniversary of Norbert Wiener's Birth, October 8–14, 1994, prepared by the MIT Department of Mathematics with the assistance of Tony Rothman.

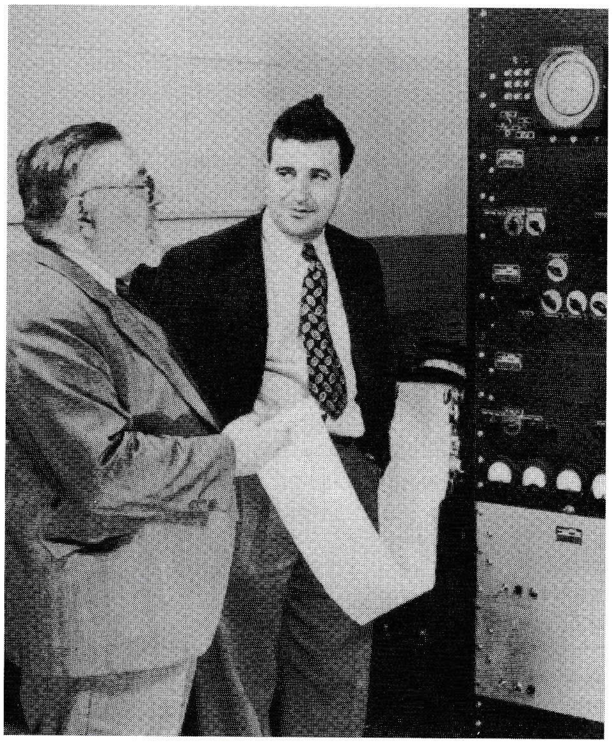
[1] This sort of feedback is called negative feedback because the thermostat reverses the action of the furnace.

In the lab with Norbert Wiener

The MIT Radiation Laboratory was founded in 1940 for the purpose of developing radar. In 1946, the Rad Lab was incorporated into the new Research Laboratory of Electronics (RLE). The following passage is taken from an 1992 essay by Jerome Wiesner, president emeritus of MIT, that appeared in *RLE Currents*.

At the start the RLE had two quite separate tracks: physics, in which researchers set out to exploit microwave tools in search of information about the physical universe, and the communications option, which primarily involved electrical engineers at the beginning, then quickly broadened to include speech and linguistics, neurophysiology, psychology and several other disciplines. . . .

Much of the communications work was inspired by Norbert Wiener and his exciting ideas about communications and feedback in man and machines. Wiener's theories, and those of Claude Shannon on information theory, spawned new visions of research for everyone interested in communications, including the neurophysiology, speech, and linguistics investigations. The work was both theoretical and experimental as well as basic and applied. For example, many early ideas about coding were developed in the RLE. So were broadband communications systems and the much earlier work about digital systems, as well as the interesting and exciting new ideas, such as the use of correlation functions to enhance weak signals, and the use of noise to measure system functions. This mix of the exploration of new ideas and their reduction to practice remains a hallmark of the present-day RLE.



Norbert Wiener and Jerome Wiesner with the "MIT Autocorrelator," built in the Research Laboratory of Electronics.

In the winter of 1947, Wiener began to speak about holding a seminar that would bring together the scientists and engineers who were doing work on what he called communications. He was launching his vision of cybernetics in which he regarded signals in any medium, living or artificial, as the same; dependent on their structure and obeying a set of universal laws set out by Shannon. In the spring of 1948, Wiener convened the first of the weekly meetings that [were] to continue for several years. Wiener believed that good food was an essential ingredient of good conversation, so the dinner meetings were held at Joyce Chen's original restaurant, now the site of an MIT dorm. The first meeting reminded me of the tower of Babel, as engineers, psychologists, philosophers, acousticians, doctors, mathematicians, neurophysiologists, philosophers, and other interested people tried to have their say. After the first meeting, one of us would take the lead each time, giving a brief summary of their research, usually accompanied by a running commentary by Wiener, to set the stage for the evening's discussion. As time went on, we came to understand each other's lingo and to understand, and even believe, in Wiener's view of the universal role of communications in the universe. For most of us, these dinners were a seminal experience which introduced us to both a world of new ideas and new friends, many of whom became collaborators in later years. Wiener's theme was that the organization of symbols, not their physical embodiment, bonded us together, whatever our disciplinary origins.

averaging them to remove *noise* (random measurement errors). When the data are averaged over time, the oscillations are dampened. His ideas were closely related to those he had about stabilizing unstable feedback loops. Of course, one has to be careful lest the averaging obliterate useful information. The whole point was to make a judicious choice of averaging procedure that retained as much information as possible.

In 1942 Wiener's collaborator Julian Bigelow[2] built a prototype to track an airplane for ten seconds and predict its location twenty seconds later. Sad to say, Wiener and Bigelow's efforts did not hasten the end of the war. It was only after the war that improvements in the speed and accuracy of airplane and radar equipment made systematic filtering and prediction devices very important. On the other hand, Wiener's ideas had ramifications far beyond their original motivation. On being confronted with a stream of data embedded in noise, the antiaircraft predictor is faced with the same problem as the communication engineer, who must send or extract a message from a noisy channel. In both cases it is possible to design a *filter* to exclude the noise, which is the engineer's term for what Wiener did. Filtering is any strategy to filter out the effects of random vibration or static from a mechanical or electrical system. Filters are needed in all sorts of devices, from stereo equipment to aircraft instrumentation.

Under the assumptions he made, Wiener's solution to the prediction and filtering problems was the best possible in a sense that is mathematically precise. Independently, at essentially the same time, A. N. Kolmogorov, the great Russian probabalist,[3] came up with a similar mathematical theory. Thus, Kolmogorov and Wiener developed the first systematic approach to the design of filters. However, their assumptions are not realistic in many applications. In technical jargon, their strategy is designed for random disturbances that are a linear function of white noise; it does not do a good job when the disturbances are nonlinear functions of white noise. Later on, Wiener addressed nonlinear problems with what he called the theory of *homogeneous chaos*, but neither Wiener nor Kolmogorov nor anyone else has achieved the kind of comprehensive success with nonlinear filtering that he did in the linear case.

Wiener wrote up his results in a 1942 report entitled *The Interpolation, Extrapolation of Linear Time Series and Communication Engineering*. The book was dubbed "the yellow peril" because of its yellow covers and its frightening mathematics. Wiener spend over a year working intensively on this report, only to have it be classified. Given Wiener's irrepressible urge to talk about his work and his desire to pursue it further, the classification was intolerable. From then on he frequently railed against military secrecy and proclaimed its incompatibility with free scientific inquiry.

In the mid-forties Wiener began to focus on neurophysiology. He advocated an interdisciplinary approach in which physicists, electrical engineers, and biologists could combine forces. He and John von Neumann started a series of conferences on "Circular Causal and Feedback Mechanisms in Biological and Social Systems."[4] Wiener also disseminated his ideas through a seminar associated with the Radiation Laboratory at MIT, which was founded during World War II to develop radar. To refine his ideas, Wiener coined "cybernetics" from the Greek word *kubernetes*, "helmsman." Webster's dictionary defines cybernetics as "the study of human control functions and of mechanical and electrical systems designed to replace them, involving the application of statistical mechanics to communication engineering."[5] Wiener says in *I Am a Mathematician* that the word was the best he could find "to express the art and science of control over the whole range of fields in which this notion is applicable."

The 1949 publication of *Cybernetics, or Control and Communication in the Animal and the Machine* turned Wiener overnight into something of a scientific superstar. At first glance this is puzzling. The book leaps between apparently unrelated meditations on time, entropy, and computers, interspersed with advanced mathematics. As Hans Freudenthal writes in *The Dictionary of Scientific Biography*: "Even measured by Wiener's standards, *Cybernetics* is a badly organized work—a collection of misprints, wrong mathematical statements, mistaken formulas, splendid but unrelated ideas, and logical absurdities."

Nevertheless, *Cybernetics* had a strong impact. In his informative article "Cybernetics" in *The Study of Information*, Murray Eden lists dozens of books and journals that took over the word "cybernetics" in the 1950s and 1960s: *Philosophy and Cybernetics*; *Cybernetic Principles of Learning and Educational Design*; *Cybernetic Modelling: The Science of Art*; *The Cybernetic ESP Breakthrough* . . . . Even this short list

[2]Bigelow was subsequently hired by John von Neumann to build the first programmable computer, the ENIAC.

[3]A. N. Kolmogorov was the cofounder (along with the physicist I. K. Kikoyin) of *Quantum*'s Russian-language predecessor and sister magazine *Kvant*.—*Ed.*

[4]In November 1946 von Neumann wrote Wiener a detailed letter suggesting that current laboratory technique was too primitive to yield a detailed picture of the brain and that it would be more worthwhile to develop electron microscopes and to perform X-ray crystallography on large organic molecules like proteins in viruses. As often happens in the history of science, this prophetic letter had no discernable influence on the brilliant course of molecular biology in subsequent years. Wiener's many ideas concerning computer design seem to have suffered a similar fate.

[5]Wiener's father, the philologist Leo Wiener, would have had a field day studying the way cybernetics has entered popular culture through words like *cyberspace* and *cyborgs*. But Arnold Schwarzenegger fans and detractors alike would be hard-pressed to say just what cybernetics is.

President Lyndon B. Johnson presents the 1963 National Medal of Science to Norbert Wiener and four other scientists, John R. Pierce, Vannevar Bush, Cornelius B. Van Niel, and Luis W. Alvarez. At the far left is Jerome Wiesner, an aide to President Johnson at the time.

leaves the impression that there was some confusion about what the word meant. Wiener himself never closely defined it, but his general train of thought is evident. As Eden notes, philosophers have always compared life to the dominant mechanical paradigm of the age. From Wiener's association with Rosenblueth and his work on communication theory and on anti-aircraft fire, he became convinced of the importance of feedback in diverse circumstances, physical and biological. From this point it is not a large leap to suppose that automata and living systems are governed by the same "laws."

But Wiener went further. Communication, he wrote, "is the cement of society." And since "sociology and anthropology are primarily sciences of communication[, they] therefore fall under the general head of cybernetics. That particular branch of sociology which is known as economics . . . is a branch of cybernetics." Here is where Wiener perhaps got himself into a bit of trouble. On the one hand, Jerome Wiesner recalls how Wiener's ideas fired up his MIT colleagues and resulted in research and courses whose descendants still exist. The viewpoint of cybernetics has been internalized in biology, where it has proved fruitful in many neurological and physiological studies, and the basic idea that information can be quantified has by now permeated our entire culture. On the other hand, in extending cybernetics to sociology, anthropology, and economics, Wiener exposed his idea to the kind of misuse evident in the list of titles above. As Dirk Struik says, Wiener was not pleased with some uses of cybernetics:

> There is a big idea behind it—control—which can be extended and overextended to society, which was not his idea. He was very worried at the time that there were people who saw in it some kind of universal panacea. He used to say to me, "I'm not a Wienerian." There was Deutsch for example at Harvard who made a whole social theory on that, and Wiener was very uneasy about these things. He had a feeling that the whole thing was flattened and became a little ridiculous. He may have had . . . himself some exaggerated ideas occasionally because he was always playing with ideas, but with all his fantasies he always had his heavy legs planted firmly on the ground.

The lighter side and the darker side

In 1926 Wiener married Margaret Engemann, who had emigrated at the age of fourteen from Germany and who had studied Russian literature with Leo Wiener. Finding a match for Norbert was no easy task. Her duties, as Norman Levinson recalls, were to "humor her husband when depressed, to allay his fears and anxiety and to tolerate him in his unbounded flights of fancy when cheerful." In short, she made her husband's life possible. Wiener dedicated *Ex-Prodigy* to his wife "under whose gentle tutelage I first knew freedom." After Norbert's death she remarked, "It was like caring for triplets."

There are, of course, the anecdotes. The most famous is the story of the day Wiener moved from a two-family house in Belmont to a quieter single-family house a few blocks away. When he left for work that morning his wife reminded him that he should return to the new house that evening. But by evening he had forgotten, and as he walked up to the old house he suddenly realized his mistake. He turned apprehensively to a child nearby and asked, "Little girl, do you happen to know where the Wiener family has moved?" And the girl replied, "Yes, Daddy, Mommy sent me to get you."

This story is not true, but many others are. For example, one day Wiener went to a seminar at Brown University in Providence. When he returned to Boston's South Station, he telephoned his wife to pick him up. "But Norbert," she said, "you drove to Providence." After an encounter with a friend outside Walker Hall on the MIT campus, Wiener asked, "By the way, which way was I going?" "Why, Norbert, you were heading to your office." "Thanks," replied Wiener, "that means I have finished lunch." David Cobb, a former MIT student, reports seeing him walking across campus in his tie and jacket "unaware of the snowstorm raging around him." Cobb also tells of the time Wiener walked into class, wrote a large "4" on the blackboard, and disappeared. Later the students discovered he was to be out of town for four weeks. There was the time Wiener gave a copy of *Cybernetics* to a famous colleague and, when this colleague had not looked at it by the following day, declared, "You are unworthy to read it." Several years later, Wiener asked his junior colleague Gian-Carlo Rota whether he had read Wiener's newly published novel, *The Tempter*. Rota replied that he had. Wiener paused and said, "Then tell me what happens in the section called '1908.'"

Everyone who remembers Wiener remembers his habit of walking up and down the long MIT corridors and buttonholing his colleagues with his latest theories. "Sometimes he would spout the most complete nonsense," says Struik. "At other times it would be almost prophetic." Fagi Levinson, Norman's widow, recalls how, on seeing Wiener, one colleague would literally hide under his desk. Her husband writes that another colleague found such Wiener-encounters so taxing that after one of them he would rush off to see his psychiatrist. Gordon Raisbeck, Wiener's son-in-law, remarks that this is insufficient information to identify the colleague uniquely. Nevertheless, Bose remembers these encounters as reconnaissance missions. "He had calibrated certain faculty in political science—or whatever areas he wanted to be up-to-date on—and he made his daily rounds. He'd talk to them for five or fifteen minutes, and he'd be up-to-date on everything."

Wiener did not mind interrupting. Donald Spencer, a student of Littlewood who came to MIT as an instructor in 1939, remembers that Wiener bounced into his office one day and announced, "Spencer, tell me what size animal can fall out of an airplane and survive. Is it a rat or a mouse? We should be able to do a Dedekind cut[6] on that." Spencer also recalls another day when he and Wiener were in the middle of a conversation in the hallway. Wiener needed to write, so he walked right into the nearest office and proceeded to use the blackboard, while the occupant, a physics professor, looked on incredulously. Wiener was concerned that eventually he would lose his sight completely, so he practiced being blind by burying his face in a book and walking the halls by following along with his finger. If he reached the open door of a classroom, he would simply forge ahead and circumnavigate the room while the entire class stared.

Wiener made forays into fiction. In 1952 Wiener pitched a movie script to Alfred Hitchcock. Joseph Kohn, the sole undergraduate in Wiener's graduate course on Fourier analysis the following year, said that Wiener would occasionally take time out from lecture to describe the plot of his latest pseudonymous detective novel.

Here is Hans Freudenthal's encapsulation of the public Wiener:

> In appearance and behavior, Norbert Wiener was a baroque figure, short, rotund and myopic, combining these and many qualities in extreme degree. His conversation was a curious mixture of pomposity and wantonness. He was a poor listener. His self-praise was playful, convincing and never offensive. He spoke many languages but was not easy to understand in any of them. He was a famously bad lecturer.

Wiener was extraordinarily solicitous of junior colleagues. He showered attention on new instructors in the mathematics department, inviting them to lunch and dinner and dropping by their offices frequently over the first few weeks. Norman Levinson writes:

> He would actually carry on his research at the blackboard. As soon as I had displayed the slightest comprehension, he handed me the manuscript of Paley–Wiener for revision. I found a gap in a proof and proved a lemma to set it right. Wiener thereupon sat down at his typewriter, typed my lemma, affixed my name and sent it off to a journal. A prominent professor does not often act as secretary for a young student.

Amar Bose recalls that when he arrived in India as an unknown postdoc, he was treated like royalty—given special editions of books, chauffeured to plays, invited to be a delegate to UN functions. It turned out that Wiener, who had spent the previous year in India, had paved the way for him by making weekly visits to the director of the Indian Statistical Institute. Wiener also exerted strenuous efforts on behalf of refugee mathematicians during World War II. For example, he persuaded the MIT administration to pay the trans-Atlantic fare for the well-known Polish Fourier analyst Antoni Zygmund, and he acted as a clearinghouse for Zygmund's job offers in the United States.

On the darker side, Levinson adds:

> If this picture of extreme kindness and generosity seems at odds with Wiener's behavior on other occasions, it is because Wiener was capable of childlike egocentric immaturity on the one hand and extreme idealism and generosity on the other. Similarly, his mood could shift from a state of euphoria to the depths of dark despair.

Wiener needed constant reassurance. He sought it from his colleagues, the loyal janitor down the hall, and the new crop of freshmen and graduate students who arrived each year. Everyone at MIT remembers his repeated lament, "Am I slipping?" and describes his hypersensitivity and dramatic mood swings. Paley said that whenever he needed a break from his intensive collaboration with Wiener, he would say that what they were doing was not working. This would plunge Wiener into a state of despair, and Paley was free to go off to his favorite nightclub, the Texguinan in New York. On his return, Paley would say that he had seen his way around the difficulties, and Wiener would spring back to life, confidence restored. Some family members suggest that his condition was a manifestation of mental illness, and that today it might be treated with proper medication, at potential risk to his creativity.

Cybernetics marked the high point of Wiener's fame, but it also marked the beginning of the end of his serious mathematical work. Much of his later career was occupied by the application of his earlier discoveries to diverse fields—for example, the application of the

[6] This is an insider's joke. A Dedekind cut is a term used in the formal mathematical construction of the real number system. Each cut splits the numbers into two halves, the ones below the cut and the ones above the cut.

autocorrelation function to electroencephelography. He became increasingly involved in literary efforts: his autobiographies, the semipopular treatments of cybernetic themes (*The Human Use of Human Beings* and *God and Golem, Inc.*), and a novel, *The Tempter*. In these works he comes across as a humane, even passionate man who saw perhaps more clearly than his contemporaries the impact of technology on society. He was a liberal in the best sense of the word with deeply held moral principles. He spoke out on issues that concerned him until the end of his life, and in this sense he was the antithesis of the cloistered academic.

As a prelude to cybernetics, Wiener had envisioned new kinds of prosthetic devices to replace the functions of vision and hearing by making use of the information channels of unimpaired senses. The best-known photograph of him on the "Infinite Corridor" at MIT shows him hooked up to what looks like a collection of pushbuttons, but is actually a device to receive messages through the sense of touch. In his last years, one of his favorite projects was the "Boston Arm," an artificial arm controlled by electrical signals from the user's upper arm muscles.

Norbert Wiener died of a heart attack on March 18, 1964, after giving a lecture in Stockholm. His scientific legacy is well documented and assures him a place in history. What distinguished him from other great contributors to twentieth-century mathematics was his ability to harness the power of abstract reasoning to practical matters. His colleagues and students have kept alive his memory as a teacher by recounting and embellishing the comical and eccentric aspects of his personality. But they also remember his inspiring enthusiasm for all rigorous intellectual activity. Amar Bose says:

> I could never have paid for the education Wiener gave me. More than anything, he gave me the belief in the incredible potential everyone has.

The "most inertial" reference frame

Is there such a thing?

by Gennady Myakishev

AN INERTIAL FRAME OF REFerence is a system where free bodies move with uniform velocities. Any frame that is usually assumed to be inertial is, strictly speaking, not inertial. One can speak only of systems that are *approximately* inertial.

The geocentric system is not inertial due to the circular movement of the Earth around the Sun and, even more important, the motion about its axis of rotation. The corresponding acceleration is maximal at its equator—only 3.4 cm/s^2. This value is much less than the acceleration due to gravity: $g = 980$ cm/s^2. The ratio of the centrifugal inertial force to the force of gravity is about 0.4%. Thus, in a great many cases the Earth can be assumed to be an inertial reference frame.

The Coriolis acceleration due to movement in the rotating reference frame does not exceed 0.1%. However, the Coriolis inertial force is of importance for movement over long periods of time.

The heliocentric reference frame is inertial to a very high degree. No conceivable experiment at the present time, or in the future, can detect the noninertial character of the heliocentric reference frame. But strictly speaking, it's not an inertial frame. Our sun is located at the outskirts of the galaxy and makes one revolution around its center every 200 million years. Perhaps, then, it's impossible to find a frame of reference "more inertial" than the one centered on the Sun (or the center of the galaxy).

Well, recently it became quite clear that this isn't so. Keep in mind that we're not talking about any practical need for such a "most inertial system of reference," but it *is* a matter of theoretical importance.

It's generally accepted that 15 billion years ago there was a "big bang," and from that moment on the universe has been expanding. At first the temperature of the universe was extremely high, but as it expanded the velocity of its component particles decreased and the universe began to cool down. At a temperature of 10^9 K neither atoms nor atomic nuclei can exist. The kinetic energy of the particles exceeded the binding energy of nucleons—when they formed, they were immediately destroyed by the next collision. Thus all the particles—protons, electrons, photons, and neutrons—were in dynamic equilibrium. The number of particles that emerged from collisions was equal on average to the number of particles that disappeared in collisions. Only when the temperature decreased further did atomic nuclei appear—helium nuclei first of all.

This went on for about a hundred thousand years, until the temperature fell to 3,000 K and the first hydrogen atoms were formed. An atom is an electrically neutral system, and so it interacts with the electromagnetic field—that is, photons—only slightly. Therefore, a "rift" occurred between radiation and matter at this time. The existing photons cooled gradually as the universe expanded, independent of other types of particles. The temperature of the equilibrium electromagnetic radiation dropped. This "relict" (or background) radiation is still present and can be detected everywhere in the universe.

The existence of the relict radiation had been predicted theoretically by the American scientists Alfer and Herman. In 1964 this radiation was discovered experimentally with a radiotelescope by Pensias and Wilson. The relict radiation comes to the Earth from every direction. It's an equilibrium thermal radiation with a maximum energy at a wavelength $\lambda = 1$ mm, which corresponds to a temperature of 3 K.

By means of the Doppler effect it's possible to detect the movement of the Earth relative to this background radiation. The wavelength of the radiation is shorter when the source and detector are approaching each other than when they are receding from each other. It's not possible to detect the movement of the Earth relative to the hypothetical ether or to a physical vacuum, but it's quite possible relative to this background radiation. It turns out that our solar system moves toward the constellation Cygnus at the unusually high speed of 200 km/s.

It is this relict radiation that forms the basis of a reference frame, and a "more inertial" frame is inconceivable. Such a system moves with a uniform velocity relative to the background radiation.

AT THE BLACKBOARD

Arithmetic on graph paper

Or graph tablets, or graph papyrus . . .

by Semyon Gindikin

IN SOME COUNTRIES, MATH notebooks are routinely made of "graph paper"—paper ruled with a square grid. Little first graders in those countries are always asking why this is. They find it rather inconvenient to write their numbers in the teeny little boxes. Later they discover that graph paper is very good for drawing geometric diagrams. I'll try to show you that we can also learn a lot of interesting things about arithmetic by drawing various figures on graph paper.

Representing numbers as certain figures on a square grid has its roots deep in antiquity—in the mathematics of ancient Babylon, Egypt, and Greece. Of course, at that time mathematicians didn't scratch lines in their clay tablets or draw squares on their papyrus—they made figures out of dots.

In ancient Greece the product of two natural numbers was called a *planar number*: it was associated with dots forming a rectangular grid (fig. 1). Here we'll depict a planar number as a rectangle on graph paper and count how many squares it contains.

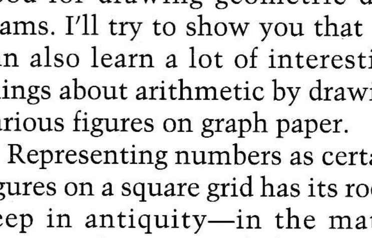
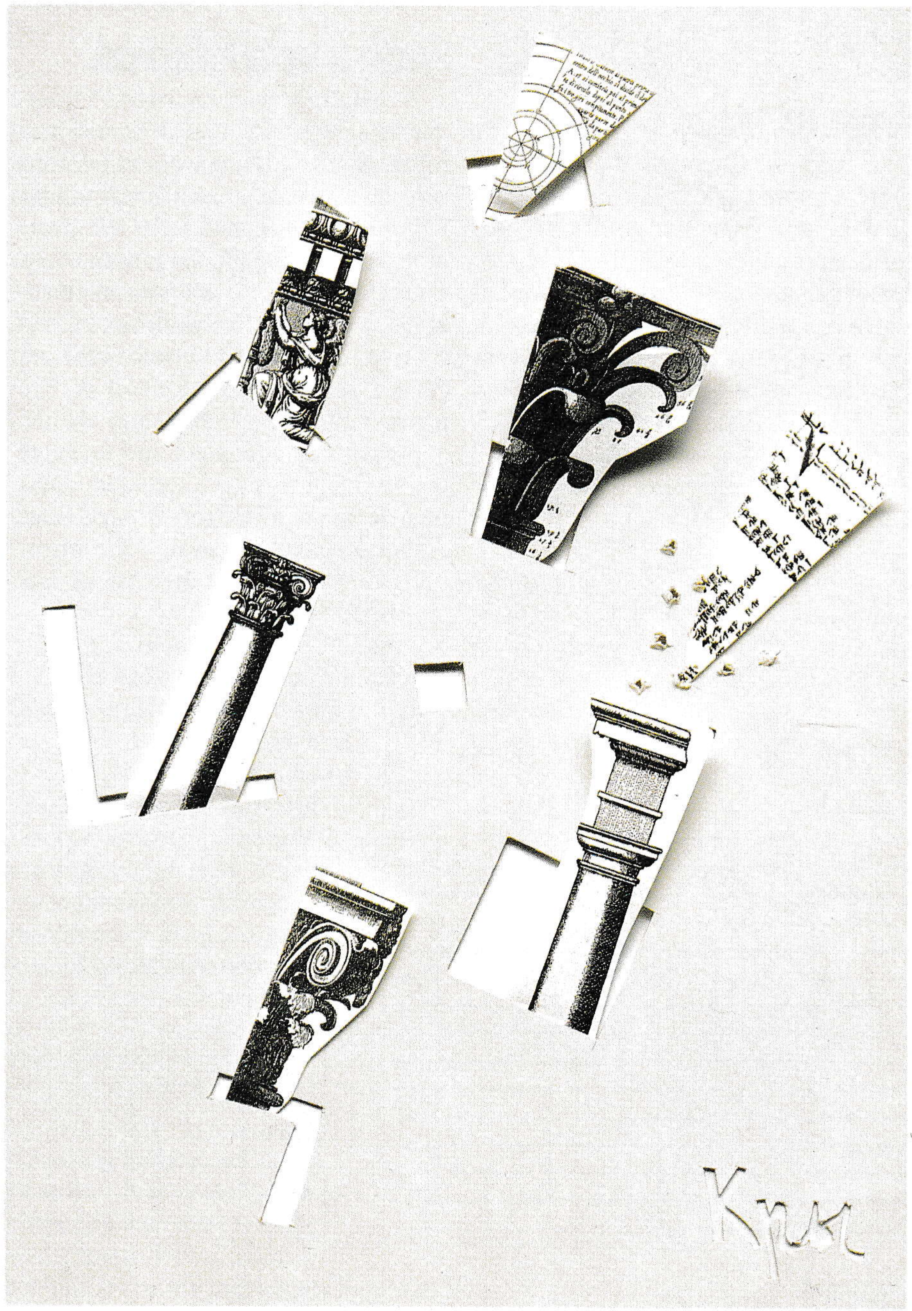


Figure 1



Art by Dmitry Krymov

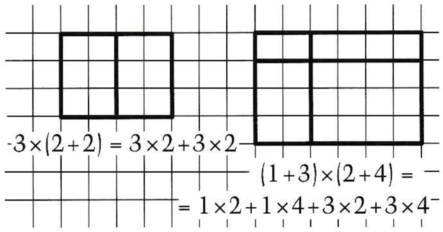


Figure 2

The properties of multiplication can now be neatly illustrated with diagrams. For instance, the distributive law (the rule for removing parentheses) corresponds to cutting a rectangle into smaller rectangles (fig. 2). The name "planar number" has now been forgotten, but the word "square" in the sense of a product of two equal factors still persists.

Squares and gnomons

In ancient Greece, odd numbers were depicted as dotted right angles with equal legs, called *gnomons*. On our graph paper these will be angular shapes made of unit grid squares, one square thick (fig. 3). The first argument I want to present involves

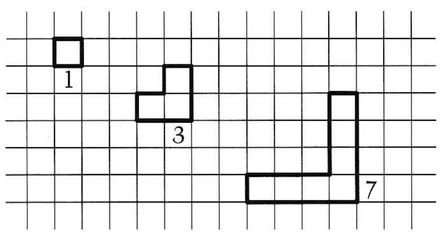


Figure 3

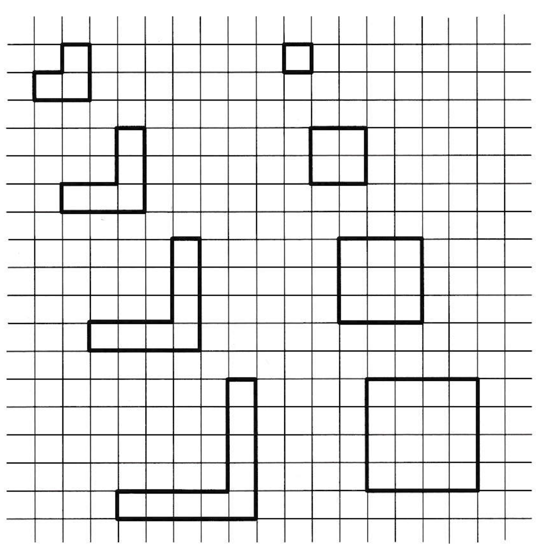


Figure 4

gnomons and squares. It is attributed to Nicomachus of Gerasa (ca. 100 A.D.).

Draw a number of gnomons and squares in succession (fig. 4). Doesn't it look as if each gnomon wants to be completed by the matching square? And this will result in the *next* square (fig. 5). This observation leads immediately to the following conclusion: *every odd number is the difference of two successive squares.*

Now let's fit a number of gnomons one inside the other, starting with the smallest. We get a square (fig. 6). So, *the sum of successive odd numbers starting with the first is a square of an integer: 1 + 3 = 4, 1 + 3 + 5 = 9, 1 + 3 + 5 + 7 = 16, and so on.*

Pythagoras and the Pythagoreans

The half-legendary sage Pythagoras (ca. 500 B.C.) liked to travel, and much of his teaching was culled from the wisdom of the East (that is to say, Egypt and the Near East). One characteristic of the Pythagoreans was their mystical attitude toward numbers. They collected all kinds of numerical curiosities, which were regarded as manifestations of divine powers. The Pythagoreans would express their thoughts and feelings through numerical images. Odd numbers were called *masculine*, even numbers *feminine*. The number 10 = 1 + 2 + 3 + 4 (fig. 7) was particularly valued by the Pythagoreans. They called such a tetrad *excellent*, and swore by "those who put the tetrad—the source and root of eternal nature—into our soul."

Numbers that are equal to the sum of their proper divisors (that is, excluding the numbers themselves), like 6 = 1 + 2 + 3, were called *perfect*. Nicomachus knew

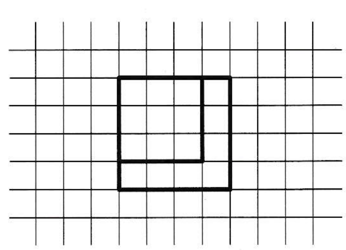


Figure 5

four perfect numbers: 6, 28, 496, and 8,128. Friendship was symbolized by pairs of *amicable* (or friendly) numbers—each number of such a pair equals the sum of the proper divisors of the other. For instance, the numbers 284 and 220 are amicable: 284 = 1 + 2 + 4 + 5 + 10 + 20 + 11 + 22 + 44 + 55 + 110 and 220 = 1 + 2 + 4 + 71 + 142.

But since there are "good" numbers, there must be "bad" numbers, too. A number that has no virtues is "bad," but a bad number surrounded by interesting numbers is even worse. Now, we all know that the number 13 brings bad luck.[1] But there used to be other numbers that struck fear into one's heart. Here's what Plutarch has to say: "The Pythagoreans have an aversion to the number 17, because 17 lies halfway between the number 16, which is a perfect square, and the number 18, which is a doubled square. These two numbers are the only planar numbers whose perimeter (of the corresponding rectangle) equals the area." In other words, it's asserted that if the product of two numbers (positive integers, of course) is equal to twice their sum, then these numbers are 3 and 6 or 4 and 4 (why?).

Pythagorean triples

Legend has it that Pythagoras celebrated one of his discoveries by sacrificing a bull. (Or was it one hundred bulls? It depends on which version

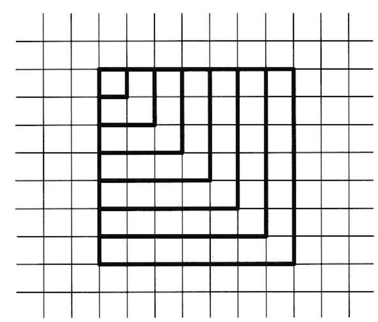


Figure 6

[1]The superstition involving the number 13 has managed to survive to the present day. A visitor from ancient times might be startled by a modern device like an elevator, but wouldn't be surprised to go from the 12th to the 14th floor, with no 13th floor in between.—*Ed.*

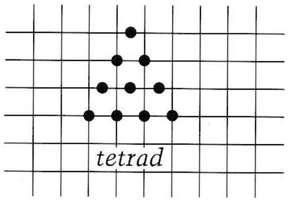


Figure 7

of the legend you hear.) Vitruvius asserts that it was the discovery of two squares whose sum is a third square that seemed so important to Pythagoras.[2] It has to do with the relation $3^2 + 4^2 = 5^2$. Nowadays the triples of natural numbers a, b, c such that $a^2 + b^2 = c^2$ are usually called Pythagorean triples. It turns out that they were known in ancient Babylon. Gradually the Greek mathematicians found them, too.

Let's try to understand how one can find Pythagorean triples. Recall that any odd number can be represented as the difference of two successive squares. Then an odd square together with the squares producing that odd square as a difference form a Pythagorean triple. For instance, $3^2 = 9 = 2 \cdot 4 + 1 = 5^2 - 4^2$ (fig. 8). Thus we get the triple 3, 4, 5. Similarly, $5^2 = 25 = 2 \cdot 12 + 1 = 13^2 - 12^2$, or $12^2 + 5^2 = 13^2$; $7^2 = 49 = 2 \cdot 24 + 1 = 25^2 - 24^2$ or $24^2 + 7^2 = 25^2$; and so on. In this way we can obtain *all* the Pythagorean triples a, b, c such that $c = a + 1$. Their general form is

$$a = \frac{m^2 - 1}{2}, \quad b = m, \quad c = \frac{m^2 + 1}{2}$$

(m is odd!). Prove this. And how can all Pythagorean triples be found?

The general problem

The experience we've acquired suggests that the difference of any (rather than successive) two squares $c^2 - a^2$ ($c > a$) needs to be examined. What we get is a "thick gnomon" $c - a$ squares thick (fig. 9). The problem has thus been reduced to a description of all possible transformations of a $b \times b$ square into a "thick gnomon" without changing the number of unit squares in it.

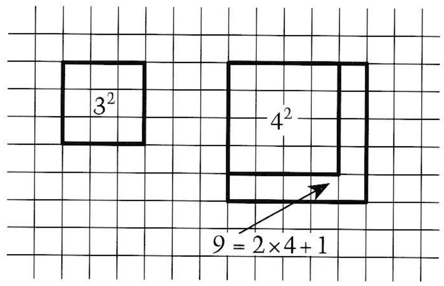


Figure 8

The first observation is that a thick gnomon can be reshaped into a rectangle (a planar number!) with side lengths $m = c - a$ and $n = c + a$ (fig. 10). This yields, by the way, a geometric proof of the formula $c^2 - a^2 = (c + a)(c - a)$. Clearly, the numbers $m = c - a$ and $n = c + a$ are different and are either both even or both odd, without any other limitations. So, a rectangle that is not a square, whose side lengths m and n are of the same parity, can be transformed into a thick gnomon that is the difference of the squares

$$c^2 = \left(\frac{m+n}{2}\right)^2$$

and

$$a^2 = \left(\frac{m-n}{2}\right)^2$$

(under our assumptions, the numbers $m + n$ and $m - n$ are even and nonzero).

Thus, the problem of finding Pythagorean triples has been reduced to transforming the square b^2 into a rectangle with side lengths m and n of the same parity ($m \neq n$). How? Let $r \neq b$ be a divisor of the number b^2 (but not necessarily of the number b itself) such that b^2/r has the same parity as r. Then r must have the same parity as b, although this condition alone is insufficient. (Can you explain why?) Then a rectangle measuring $m \times n$ with $m = r$, $n = b^2/r$ can be transformed into the thick gnomon that represents the difference of the squares

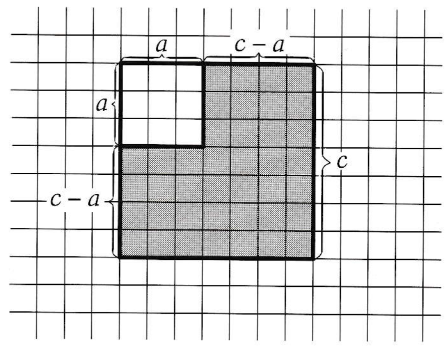


Figure 9

$$c^2 = \left(\frac{b^2/r + r}{2}\right)^2$$

and

$$a^2 = \left(\frac{b^2/r - r}{2}\right)^2.$$

Notice that r can be equal to 1 and that we can confine ourselves to the divisors r of b^2 less than b.

Finally, this is how an arbitrary Pythagorean triple can be written:

$$a = \frac{b^2 - r^2}{2r}, \quad b, \quad c = \frac{b^2 + r^2}{2r},$$

where r, $1 \le r < b$, is a divisor of b^2 such that r and b^2/r are the same parity (which is the parity of b). Using this rule we can automatically write out Pythagorean triples. Try to do it on your own. To make sure you're on

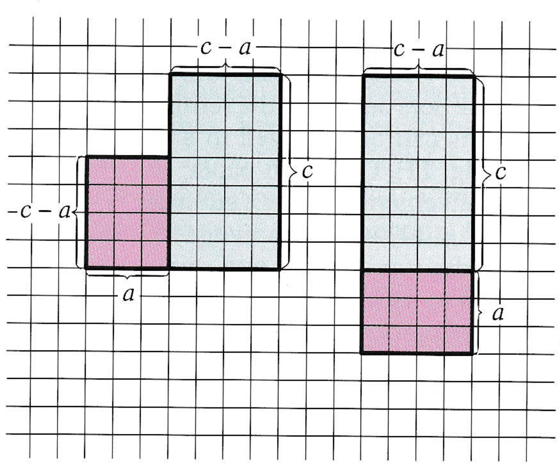


Figure 10

[2]Other sources, however, point to other reasons. And some are of the opinion that there wasn't any sacrifice at all—Pythagoreans didn't believe in them.

the right track, you can check your results against the table below, which lists the Pythagorean triples a, b, c based on the first ten values of b.

b	r	a	c
3	1	4	5
4	2	3	5
5	1	12	13
6	2	8	10
7	1	24	25
8	2	15	17
8	4	6	10
9	1	40	41
9	3	12	15
10	2	24	26

Since the numbers a and b in a Pythagorean triple a, b, c are interchangeable, pairs of triples differing only by a transposition of a and b occur in the table. We also notice that there is no triple for $b = 2$: in this case the number b has no suitable divisors (only $r = 1$ satisfies the condition $r < b = 2$, but this r is odd). For any other b there is at least one Pythagorean triple. If b is odd, we can take $r = 1$, which yields a and c differing by one (we looked at this case at the beginning of the article). As for an even $b \neq 2$, we can put $r = 2$: this yields triples with $c - a = 2$ (why?).

Triangular numbers

Babylonian cuneiform tablets contain a method for computing the sum of the first n natural numbers $1 + 2 + \ldots + n$. These sums received the name *triangular numbers* because the dots depicting the terms in the sums can be arranged in a triangle (fig. 11). On graph paper, it's more convenient to draw a staircase shape (fig. 12) in which every row except the top one has one square more than the row above it. To count the number of unit squares in such a staircase, which is equal to the sum in question, draw another, inverted staircase and move both staircases together (fig. 13). The steps of one staircase will fit those of the other, thus forming an $n \times (n + 1)$ rectangle. It consists of $n(n + 1)$ unit squares. The staircase has half that many—that is, $n(n + 1)/2$. So

$$1+2+3+\cdots+n=\frac{n(n+1)}{2}.$$

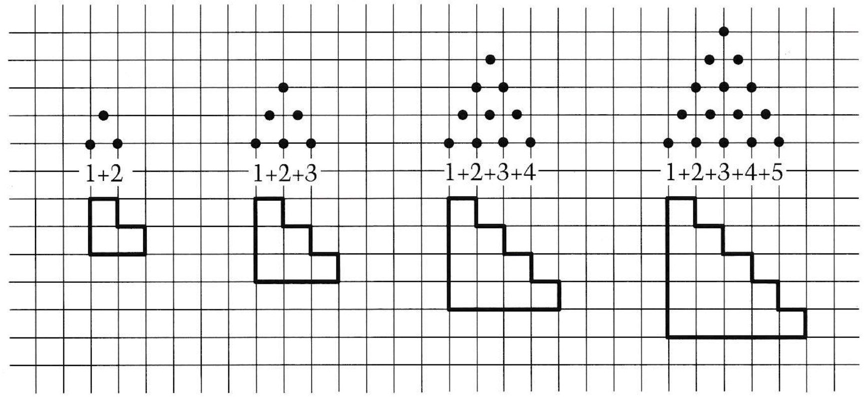


Figure 11

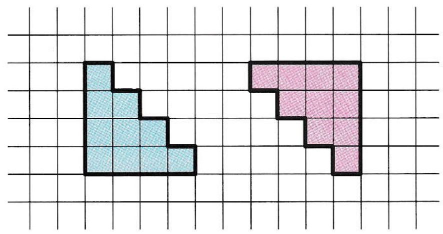


Figure 12

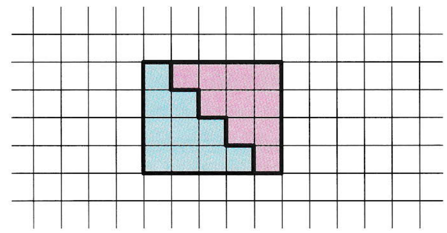


Figure 13

This becomes evident after we put together all the gnomons corresponding to adjacent triangular numbers.

Problem 1. Using staircases, find the sum $1 + 3 + 5 + \ldots + (2k + 1)$ of odd numbers, the sum $2 + 4 + 6 + \ldots + 2k$ of even numbers, and the sum $1 + 4 + 7 + \ldots + (3k + 1)$ of the numbers of the form $3i + 1$. (The first of these sums has already been found by using gnomons.)

The next problem is based on another interesting observation by Nicomachus.

Problem 2. Verify that the gnomon equal to the difference of the squares of two adjacent triangular numbers is a cube—more exactly,

$$\left(\frac{n(n+1)}{2}\right)^2-\left(\frac{n(n-1)}{2}\right)^2=n^3.$$

From this relation, Nicomachus derived a formula for the sum of successive cubes:

$$1^3+2^3+\cdots+n^3=\left(\frac{n(n+1)}{2}\right)^2.$$

Editor's note: Some of the topics covered in this article have been visited in previous issues of *Quantum*. For more on "amicable numbers" see "Kith and Kin" in the January 1990 issue. Another method for deriving Pythagorean triples can be found in "Genealogical Threes" (November/December 1990). And for some more interesting facts about "shape-numbers" see "An Old Fact and Some New Ones" (September/October 1990).

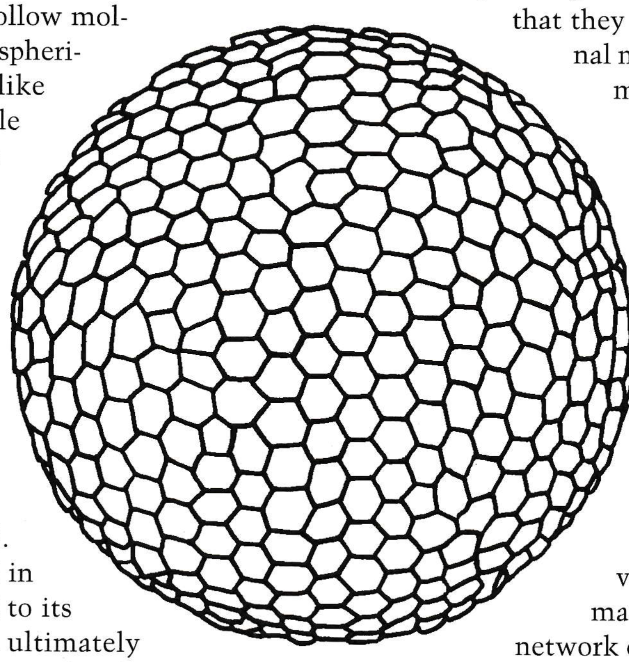
Hollow molecules

Editor's note: These speculations by David E. H. Jones (aka "Daedalus") were originally intended as an insert in "Follow the Bouncing Buckyball" (May/June 1994) but did not appear due to delays in obtaining the required permissions. Readers may wish to remove this page and insert it at page 10 of that issue.

THERE IS A CURIOUS DISCONTINUITY BETWEEN the density of gases (around 0.001 g cm^{-3}) and that of liquids and solids (from 0.5 to 25 g cm^{-3}). Daedalus has been contemplating ways of bridging this gap, and has conceived the hollow molecule. This would be a closed spherical shell of a sheet-polymer like graphite, whose basic molecule is a flat sheet of carbon atoms bonded hexagonally rather like chicken-wire. He proposes to modify the high-temperature synthesis of graphite by introducing suitable ill-fitting foreign atoms or molecular units into the sheets to warp them (rather like "doping" semiconductor crystals to introduce discontinuities). The curvature thus produced in the sheet will be transmitted to its growing edges so that it will ultimately close on itself. The radii of the molecules thus produced would be controlled by the level of impurities included in them. Daedalus calculates that a substance made of hollow molecules 0.05 micrometres across would have a bulk density of about 0.04 g cm^{-3}, about half-way between the densities of liquids and gases, and should constitute a vague fifth state of matter. These enormous molecules (molecular weights up to 100 million!) could hardly evaporate, but would interact so weakly at their few points of contact as not to be solid or even liquid. They should behave as tenuous fluids, retainable in open vessels but without any definite surface, and if heated would expand steadily, without boiling, into a gas-like state.

Such fascinating materials would find a host of uses, in novel barometers and shock-absorbers and fluidization-systems and so on; they might even be ideal as low-drag lubricants, where the rolling contact of the molecules would lower the friction even further in ball-race fashion. Daedalus was worried that they might deform under load until he realized that if synthesized in a normal atmosphere they would be full of gas and resilient like little footballs. So he is seeking ways of incorporating "windows" in their structure so that they can absorb or exchange internal molecules, thus acting as super molecular-sieves capable of entrapping hundreds of times their own weight of such small molecules as can enter the windows.

—*New Scientist*,
3 November 1966



Aulonia hexagona

From Daedalus's notebook

. . . Euler's Law states that for any polyhedron, (no. of corners) + (no. of faces) − (no. of edges) = 2. This prevents any polyhedron being made up entirely of hexagons, a network of which has $C + F - E = 0$. In that wonderful book *Growth and Form* (Cambridge University Press, pp. 708 and 738), W. D'Arcy Thompson discusses this problem in connection with radiolaria, those microscopic sea-creatures whose silica skeletons are frequently made up of hexagonal meshes. Even the beautifully symmetrical *Aulonia hexagona* (which is almost a perfect 100,000-fold scale enlargement of a 1,200-atom hollow graphite molecule) has some non-hexagonal faces.

From *The Inventions of Daedalus: A Compendium of Plausible Schemes* by David E. H. Jones (San Francisco: W. H. Freeman and Company, 1982)

Bulletin Board

New ARML site

The American Regions Math League announces the opening of a third site for its annual high school competition. The host for this new western site will be the University of Nevada at Las Vegas. The competition will be held on June 2 and 3, 1995.

The ARML competition is the largest on-site high school contest in the US. It currently draws students from 34 states and two Canadian provinces to its two other sites—Pennsylvania State University and the University of Iowa. The Las Vegas site will permit access to the event for teams from the western and Pacific states. All three sites will be linked by video communications to form a single national competition.

Among the unique features of this event is the teamwork involved. Teams of fifteen students typically represent an entire state, city, or geographical area, rather than a single school. Team members cooperate in solving a power question (a single long-answer question) and a team round (ten short-answer questions to be worked on by the whole team). Contest materials are typically taken back to home schools and form the basis of exploration and independent study.

For more information about forming an ARML team or joining an existing team, please contact

Mark Saul
711 Amsterdam Ave.
New York NY 10025
phone: 212 666-5188 (voice & fax)
e-mail:73047.3156@compuserve.com

or

Steve Adrian
55 Mill Plain Rd., Unit 29-5
Danbury CT 06811
phone: 203 792-6828.

Thinking computers?

The results from the fourth annual Loebner Prize Competition are in, and once again, the computers lost.

Pitting humans against computers, judges "conversed" with ten terminals on December 16, 1994, at the University of California–San Marcos. Some of the terminals were controlled by computer programs, some by hidden human "confederates." (Judges and confederates were all members of the media.) Conversation at each terminal was limited to a single topic. After the judges had conversed with all the computers, they rank-ordered the terminals from "least human" to "most human" and then tried to guess which terminal was which. While two humans were mistaken by some judges for computers, no computers were mistaken for humans.

A prize was awarded to the programmer whose terminal achieved the highest median rank. Prizes were also awarded to the "Most Human Computer" and the "Most Human Human." Spectators were able to view each conversation as it unfolded, try their hand at a "mystery terminal," and complete their own rating forms.

When a computer passes a test with unrestricted conversation, a grand prize of $100,000 will be awarded, and the contest will be discontinued.

Software is available that plays the conversations in real time exactly as they occurred in the last four competitions. To purchase this software, contact the Cambridge Center for Behavioral Studies, 675 Massachusetts Ave., Cambridge MA 02139; phone: 617 491-9020; fax: 617 491-1072; e-mail: 76557.1175@compuserve.com.

Programmers interested in communicating with other programmers planning submissions to the Loebner Prize Competition or wishing to get involved in team programming efforts should contact Ms. Kim Binsted (kimb@aisb.ed.ac.uk).

Stay at home—abroad

American–International Homestays, Inc. (formerly American–Soviet Homestays) offers travelers a unique opportunity to form friendships while experiencing history in the making. You stay in the private homes of English-speaking families and find out firsthand what the people are thinking and feeling during this time of great change in their nations. Among the locations offered are Moscow, St. Petersburg, Kiev, Siberia and Lake Baikal, Mongolia, Uzbekistan, Prague, Budapest, Krakow, the Baltics, and Beijing. This year, American–International Homestays is expanding its program to India and Australia.

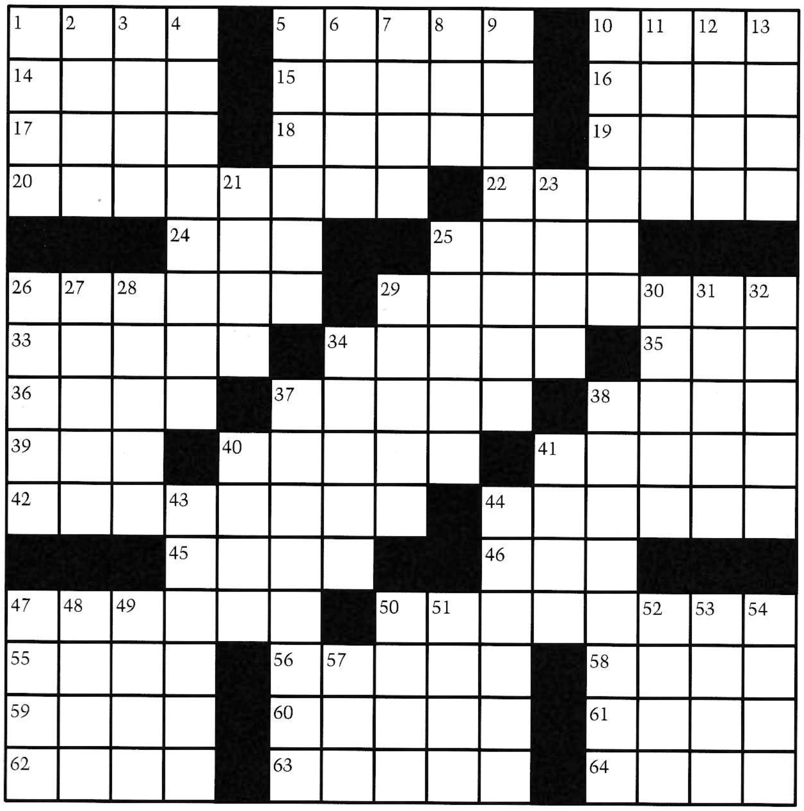
Travelers live one week each with two different families, in two cities, or five days each with three families in three cities, enjoying their warmth and hospitality, eating home-cooked meals, and entering their hosts' circle of friends and family. You have a private room, and the hosts will be off from work during the visit in order to spend time with you and introduce you to the daily life and sights of their culture.

Each scheduled group trip includes round-trip airfare from New York or San Francisco, internal transportation, all meals, and lodging. Other special packages are available. Most scheduled trips are 14 days long, with prices beginning at $2,090.

For a brochure call toll-free: 800 876-2048. From outside the continental US call 303 642-3088. Or write to American–International Homestays, Inc., PO Box 7178, Boulder CO 80306.

Criss×cross science

by David R. Martin



Across

1 Hypocritical act
5 Billiards shot
10 Closed
14 South American country
15 Hearing: comb. form
16 Infatuation (sl.)
17 ___ group (chem. group)
18 Provide with weapons again
19 Shortest distance
20 Pressure unit
22 mc^2
24 Japanese statesman Hirobumi ___ (1841–1909)
25 Snicker-___
26 Verbalized
29 A plane curve
33 Written research report
34 Tree dwelling primate
35 Perish
36 Ireland
37 Source of feelings
38 Prophetess (Scand. myth.)
39 Unit of mass: abbr.
40 Roman garments
41 More sensible
42 Austrian biologist Paul ___ (1880–1926)
44 Sex glands
45 Geologic time periods
46 Long inlet
47 Bird's respiratory feature
50 Element 81
55 Reverberation
56 Oaf
58 Pigment
59 Anthropologist Carleton ___ (1904–1981)
60 1000 kilograms
61 German river
62 Finishes
63 Keyboard word
64 ___ cone (missile section)

Down

1 Type of canned meat
2 Blood: comb. form
3 ___ group (chem. group)
4 1966 chem. Nobelist
5 Diamond element
6 Plant: suff.
7 Lion's sound
8 Belonging to us
9 Mass times velocity
10 Blood-forming organ
11 Dead keratinized cells
12 Military branch: abbr.
13 Those people
21 Roman road
23 ___-do-well
25 Greek island
26 Talk
27 Cleveland suburb
28 A narcotic
29 Immunologist ___ Milstein
30 Turkish city
31 Angered
32 Units of time
34 Theater boxes
37 Pyroelectric mineral
38 ___ radiation belt
40 10^{12}: pref.
41 *Podzol* or *chernozem*, e.g.
43 Fundamental particles
44 Kitchen tool
47 44,750 (in base 16)
48 Computer screen image
49 Red: comb. form
50 Hue
51 Sharpen
52 Villain in *Othello*
53 Hawaiian instruments
54 Simple
57 Wear

SOLUTION IN THE NEXT ISSUE

SOLUTION TO THE JANUARY/FEBRUARY PUZZLE

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

ANSWERS, HINTS & SOLUTIONS

Math

M136

(a) The answer is $x = 2$, $y = 3$. The equation is quickly reduced to the form $x^y = x^2 + 2x$ or $x^{y-1} - x = 2$ (since $x > 0$). It follows that x is a natural divisor of 2. The value $x = 1$ does not fit, so $x = 2$, and therefore $y = 3$.

(b) This equation has two solutions: $(y, z) = (1, 1)$ and $(y, z) = (3, 2)$. The second solution amounts to the same number relation as in (a): $2^3 + 1 = 3^2$. The case $y = 1$ is clear. Suppose $y > 1$; then the equation can be rewritten as

$$2^y = 3^z - 1 = (3 - 1)(3^{z-1} + 3^{z-2} + \dots + 1).$$

Since $y > 1$, the product on the right side must include more than one factor of 2. It follows that the second factor on the right side must be even, and so z must be even. Let $z = 2k$. Then we have $2^y = 3^{2k} - 1 = (3^k - 1)(3^k + 1)$. So $3^k - 1$ and $3^k + 1$ are powers of 2 differing by 2, which is possible only for $k = 1$ or for $z = 2$, $y = 3$.

Originally, the author of this problem proposed the much more difficult and general equation $x^y + 1 = (x + 1)^z$ in natural numbers x, y, z. Try to prove that it has only the following solutions: $(x, 1, 1)$ for any x, $(1, y, 1)$ for any y, and $(2, 3, 2)$. (Hint: use the binomial formula and considerations of divisibility.) Equations (a) and (b) are the particular cases for $z = 2$ and $x = 2$. (V. Dubrovsky)

M137

Let $ABCD$ be the given trapezoid (fig. 1). Using the observation that

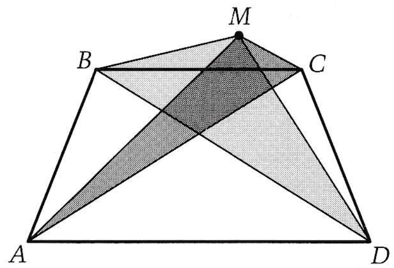


Figure 1

its diagonals AC and BD are the same length, and the Triangle Inequality for triangles MAC and MBD, where M is the given point, we get

$$\begin{aligned} MA &\le MC + AC \\ &= MC + BD \\ &\le MC + MB + MD. \end{aligned}$$

It remains to note that at least one of these two inequalities must be strict, because the first of them becomes an equality only if point M lies on the extension of AC beyond C, while the second only for M on the segment BD.

A nice, purely geometric solution is illustrated in figure 2: fitting the side BD of the triangle BDM to the side AC of triangle ACM, we obtain a quadrilateral whose side lengths are equal to the distances in question, which makes the statement of the problem self-evident. It is left to the reader to investigate what happens when point M is *not* on the plane of the trapezoid $ABCD$.

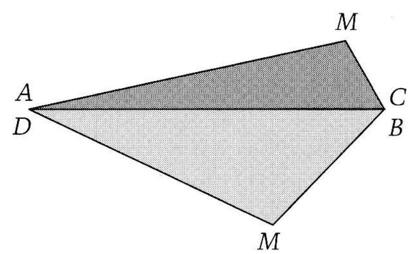


Figure 2

M138

This is a typical problem that can be solved "from back to front." We'll even begin the solution with the second question, the answer to which is no.

Denote by N_i the number of the ten-jump series in which the ith kangaroo k_i ($i = 1, 2, \dots, 10$) *finally arrives* at B—that is, jumps to point B from some other point to stay at B forever, perhaps hopping in place if the jumping is continued. Then $N_{10} < N_9 < \dots < N_1$, because to finally arrive at B, kangaroo k_i has to jump over k_{i-1}, which is not at B yet ($i = 2, 3, \dots, 10$). So the total number of jumping series is not less than $N_1 \ge N_2 + 1 \ge + N_3 + 2 \ge \dots \ge N_{10} + 9 \ge 10$.

It is almost obvious that at any time the next series of jumps can be organized so as to drive a chosen kangaroo k_i onto point B. If our kangaroos sit initially at points K_1, K_2, …, K_{10}, then, moving from the end, we find that k_{i-1} must be driven onto the midpoint M_{i-1} of the segment K_iB (to make k_i jump onto B; see figure 3 for $i = 4$), so k_{i-2} must be driven onto the midpoint M_{i-2} of $K_{i-1}M_{i-1}$, and so on up to k_1, which will have to jump onto the midpoint M_1 of the segment K_2M_2 determined by the above consideration.

Now, using this construction, we can send k_{10} to B in the first series of jumps, then send k_9 to B in the second series (k_{10} hops in place),

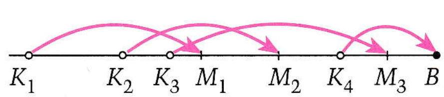


Figure 3

then send k_8 to B in the third series (k_9 and k_{10} hop in place), and so on. In the last (tenth) series, k_1 jumps onto B and all the other kangaroos hop in place. So it's possible to organize the required "migration" of the entire herd in ten series.

An algebraically minded reader may find it interesting to trace this "migration" in coordinates. This will involve Pascal's triangle and binomial coefficients. However, this isn't really necessary for our problem as it is. (V. Dubrovsky)

M139

Put $f_1(x) = f(x) = x^3 - x + 1$, and, for any $n = 2, 3, \ldots,$

$$f_n(x) = f(f_{n-1}(x)).$$

Then $f_1(0) = f_1(1) = 1$, and so (by induction) $f_n(0) = f_n(1) = 1$ for all n. This means that the constant term $f_n(0)$ of each of the polynomials $f_n(x)$ with integer coefficients is equal to 1. So, for any natural a the number $f_n(a)$ yields a remainder of 1 when divided by a. It follows that for any integers m, k, and l, $k > l > 0$, the number $f_k(m) = f_{k-l}(f_l(m))$ yields the remainder 1 when divided by $f_l(m)$—that is, $f_k(m)$ and $f_l(m)$ are coprime numbers, completing the proof.

Now, it's only natural to try to find all polynomials $f(x)$ with integer coefficients with the same property: *for any natural m the numbers m, f(m), f(f(m)), … are pairwise coprime.*

The solution above shows that any f satisfying $f(0) = f(1) = 1$—that is, having the form $f(x) = x(x-1)r(x) + 1$ (where $r(x)$ is a polynomial with integer coefficients)—will do. Notice that the construction in the problem generates infinitely many infinite sequences of natural numbers in which any two numbers are coprime. (For instance, for $m = 2$: $f_1(2) = 7$, $f_2(2) = f_1(7) = 337, \ldots, f_n(2), \ldots$.) The existence of such a sequence implies that the set of prime numbers is infinite (why?).

M140

First, consider a number $N = p^k$ for a prime p. The divisors of N are the $k + 1$ numbers $1, p, p^2, \ldots, p^k$, and the numbers of their divisors are equal, respectively, to $1, 2, 3, \ldots, k + 1$. So we have to prove that

$$[1 + 2 + \ldots + (k+1)]^2 = 1^3 + 2^3 + \ldots + (k+1)^3.$$

To do this, we denote the sum $(1 + 2 + 3 + \ldots + n)^2$ by s_n. Then, using the formulas for the difference of squares and for the sum of an arithmetic sequence we obtain

$$\begin{aligned} s_{k+1} - s_k &= (k+1)[2(1 + 2 + \ldots + k) + k + 1] \\ &= (k+1)[k(k+1) + k + 1] \\ &= (k+1)^3. \end{aligned}$$

So

$$\begin{aligned} s_{k+1} &= (k+1)^3 + s_k \\ &= (k+1)^3 + k^3 + s_{k-1} \\ &= \ldots \\ &= (k+1)^3 + k^3 + \ldots + 1^3, \end{aligned}$$

because $s_1 = 1$.

This establishes the result if $N = p^k$ for prime p. Now suppose that $N = AB$, where A and B are coprime numbers with the divisors $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$, respectively. Let α_i, $i = 1, \ldots, n$, and β_j, $j = 1, \ldots, m$, be the numbers of the divisors of a_i and b_j. Since A and B have no common divisors, the divisors of AB are all the products a_ib_j, $i = 1, \ldots, n$, $j = 1, \ldots, m$; notice that the number of these products is nm. For the same reason, the number of divisors of a_ib_j equals $\alpha_i\beta_j$. The sum of all these numbers can be written as

$$\alpha_1\beta_1 + \alpha_1\beta_2 + \ldots + \alpha_n\beta_m = (\alpha_1 + \ldots + \alpha_n)(\beta_1 + \ldots + \alpha_m),$$

so

$$(\alpha_1\beta_1 + \alpha_1\beta_2 + \ldots + \alpha_n\beta_m)^2 = (\alpha_1 + \ldots + \alpha_n)^2(\beta_1 + \ldots + \alpha_m)^2.$$

Similarly,

$$(\alpha_1\beta_1)^3 + (\alpha_1\beta_2)^3 + \ldots + (\alpha_n\beta_m)^3 = (\alpha_1^3 + \ldots + \alpha_n^3)(\beta_1^3 + \ldots + \beta_m^3).$$

It follows that if the statement of the problem holds for A and B, it holds for AB as well (because both sides of the relation in question for AB are equal to the products of the respective sides of the relations for A and B).

Now the proof can be completed by induction over N. Suppose our relation is true for any number less than N. We can always represent N as $N = p^kA$, where p is prime and A is not divisible by p. The case $A = 1$ was considered separately, and in the case $A > 1$, the coprime factors p^k and A satisfy our relation (they are less than N), so it's true for their product N as well.

It appears that the statement of this problem was first proved by the well-known French mathematician J. Liouville.

Physics

P136

Let the horse and the person be at points A and B, respectively (see figure 4). Denote $OA = R$, $OB = r$,

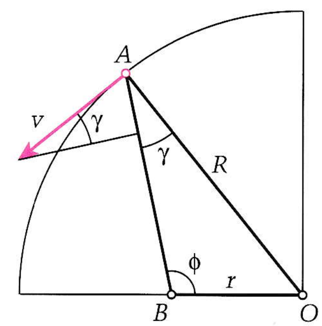


Figure 4

$\angle OBA = \phi$, and $\angle BAO = \gamma$. The speed at which the horse approaches the person is the component of the velocity **v** along the direction AB: $v_{appr} = v \sin\gamma$. We must find the position of point A where angle γ is the greatest:

$$\frac{\sin\phi}{R} = \frac{\sin\gamma}{r}, \quad \sin\gamma = \frac{r}{R}\sin\phi \le \frac{r}{R}.$$

So when $\phi = 90°$, the relative velocity yields a maximum of

$$v_{appr\,max} = v\frac{r}{R}.$$

P137

First of all we should carefully work with the curve for the droplet's acceleration and use it to graph its velocity versus time (again in relative units). To do this, let's recall that the increase in velocity per unit time is equal to the area under the curve for the acceleration in that period. We need the velocity curve to determine how much smaller the velocity v at the time t_1 of field shutoff is compared to the maximum velocity v_y. The author obtained the value of

$$v = 0.4v_y.$$

The air resistance F_r, which is proportional to the droplet's velocity, counterbalances the force of gravity mg. So the acceleration of the droplet at time t_1 (that is, the maximum acceleration) is

$$a = \frac{mg - F_r}{m} = g\left(1 - \frac{F_r}{mg}\right)$$

$$= g\left(1 - \frac{v}{v_y}\right) = 6\ \text{m/s}^2.$$

P138

Let's begin with qualitative considerations. Evidently, the time T that we seek is longer when the bubble's radius R_0 and the density of the gas inside are greater (for example, if the bubble is filled with hydrogen, it collapses more quickly than a bubble filled with air). On the other hand, this time is shorter when the surface tension of the soapy film σ is higher (a different type of soap can be used) and the tube's radius is smaller.

Now let's do some numerical calculations. According to the law of conservation of energy, the surface potential energy of the bubble changes during shrinkage into the kinetic energy of the escaping air (the kinetic energy of the bubble itself can be neglected, as we'll show later in our calculations). Thus, one can write

$$d(2\sigma S) = \frac{-v^2 dm}{2},$$

where $dm = -\rho_{air} dV$ is the mass of the air escaping the bubble in a short time dt; ρ_{air} is the density of the air, and dV is the decrease in the bubble's volume. The factor 2 in this formula is due to the fact that a bubble has two surfaces, and each surface has σS surface energy (of course, we neglect the thickness of the soap film and consider the radii of these surfaces to be equal). Since, on the one hand, $dV = 4\pi R^2 dR$ and, on the other, $dV = -\pi r^2 v dt$, we have

$$\frac{dR}{dt} = -\sqrt{\frac{\sigma}{2\rho_{air}}} \frac{r^2}{R^{5/2}}. \qquad (1)$$

Now let's compare the bubble's kinetic energy $E_k = M(dR/dt)^2/2$ with its surface energy $E_s = 2\sigma S = 8\pi\sigma R^2$ at the moment, say, when its radius is decreased by a half. As the mass of the bubble's shell is $M = \rho_w 4\pi R^2 h$, where ρ_w is the density of water and h is the thickness of the soap film, then

$$\frac{E_k}{E_s} = \frac{\rho_w}{8\rho_{air}} \frac{r^4 h}{R^5}.$$

If $r = 1$ mm, $R = R_0/2 = 10$ mm, $h = 0.01$ mm, $\rho_w = 10^3$ kg/m^3, and $\rho_{air} = 1.29$ kg/m^3, we get

$$\frac{E_k}{E_s} = 10^{-5} \ll 1$$

—that is, the assumption we made above is corroborated. To finish our solution, let's separate the variables R and t in formula (1) and integrate it, taking into account that $0 \le R \le R_0$:

$$dt = -\sqrt{\frac{2\rho_{air}}{\sigma}} \frac{1}{r^2} R^{5/2} dR,$$

$$T = \frac{2}{7}\sqrt{\frac{2\rho_{air}}{\sigma}} \frac{R_0^{7/2}}{r^2} = 4\ \text{s}.$$

The result obtained confirms the initial qualitative considerations and correlates rather well with experimental observations.

P139

To solve the problem, let's use the principle of superposition. At first,

$$\phi_1 = \frac{q_1}{C_1} + \sum_2^n \frac{q_i}{C_{ij}},$$

where C_{ij} are some factors (the so-called mutual electrical capacities of the conductors). When the electrical charges were replaced by the opposite charges, this equation became

$$0 = \frac{q_1}{C_1} - \sum_2^n \frac{q_i}{C_{ij}},$$

from which we get

$$\sum_2^n \frac{q_i}{C_{ij}} = \frac{q_1}{C_1}.$$

Thus, the desired potential is

$$\phi_4 = \frac{4q_1}{C_1} - \sum_2^n \frac{q_i}{C_{ij}} = \frac{3q_1}{C_1} = \frac{3}{2}\phi_1.$$

P140

We can solve the problem by two methods.

1. The image of the Sun that appears in the focal plane of the first mirror (of radius R_1) serves as an object for the second mirror (of radius R_2). Given an angular dimension of the Sun α, the diameter of this image is $l = \alpha F_1 = \alpha R_1/2$, and its distance to the second mirror equals $d_2 = R_2 - R_1/2$ The magnification of the second mirror is $\Gamma_2 = F_2/(d_2 - F_2) = R_2/(R_2 - R_1)$. The size of the image produced by the two mirrors is equal to

$$L_m = l\Gamma_2 = \frac{\alpha R_1}{2} \frac{R_2}{R_2 - R_1}$$

$$= \alpha \frac{R_1 R_2}{2(R_2 - R_1)}.$$

A thin lens forms in its focal plane a solar image of the size $L_l = \alpha F_l$. To get an image of the size $L_l = L_m$ one must choose a lens with a focal length

$$F_l = \frac{R_1 R_2}{2(R_2 - R_1)} = 10\ \text{cm}.$$

2. One of the rays parallel to the optic axis of the mirror system is shown in figure 5 (see line AB). Point M is the intersection of ray AB and ray DF, which leaves the system.

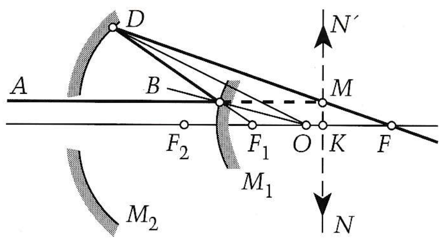


Figure 5

Let's place a thin lens of focal length $F_1 = |KF|$ in the plane NN', which passes through point M and is perpendicular to the optic axis of the system.

The ray AB that enters the lens (without mirrors) passes point F after refraction—that is, the path of this ray would be the same as when it leaves the mirror system. Therefore, the solar image formed with the lens will be the same size as that produced by the mirror system.

One can easily calculate the focal length of the lens to be $F_1 = 10$ cm.

Brainteasers

B136

Babs wears the white dress, Grace is in blue, Pam in green, and Winnie in pink. The first condition immediately implies that the girl in green is Pam. Then, from the second condition we derive that Grace is in blue (neither white nor pink—nor green, of course). Since Winnie is Pam's neighbor, she stands opposite Grace, so she wears a pink dress. (V. Dubrovsky)

B137

It's clear that each of the six pieces must consist of four squares. We see that they can have one of five different shapes (I-shape, L-shape, T-shape, Z-shape, or a 2×2 square). A bit of trial-and-error shows that only the L- and Z-shapes are possible (examine the filling of the top extreme right square). Two corresponding dissections are shown in figure 6 (other arrangements of these shapes are possible).

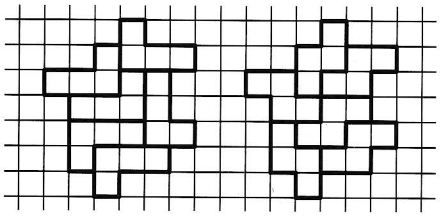


Figure 6

B138

A person running along the "bridge" and starting to fall ends up not in the water but on the opposite side of the stream, since it takes more time to fall than to run.

B139

We can see immediately from figure 7 that $AD = \sqrt{4^2 + 3^2} = 5$, where the unit length is the side length of a grid square). So $AD = AE$ (by the choice of point E), and $CD = CE = \sqrt{5}$. Therefore, $\triangle AEC \cong \triangle ADC$, which implies the required equality. (V. Dubrovsky)

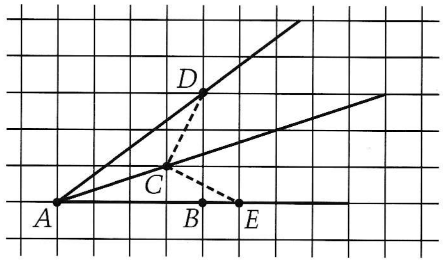


Figure 7

B140

A rook can be removed without violating the given property if there is another rook in its rank and another rook in its file. Such a rook necessarily exists. Otherwise, for each of the 15 rooks either the rank or the file in which it sits contains only this rook. This means that there are ranks or files each of which contains only one rook. Therefore, eight of them must be of the same kind—say, a rank (horizontal). Then each rank contains only one rook, and the total number of rooks is eight, which contradicts the statement of the problem. The arrangement of 14 rooks in figure 8 shows that the number 15 in the statement cannot be reduced.

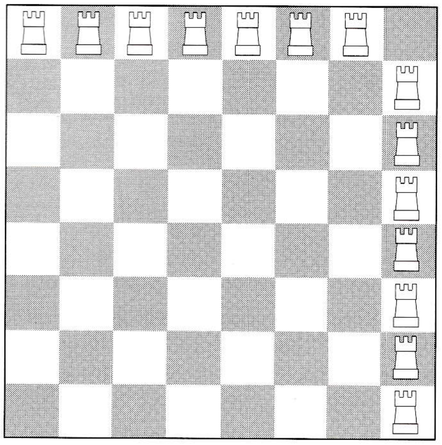


Figure 8

Toy store

1. All the n-move processes ($n \geq 3$) in question fall into two classes: (1) those in which the last rotated face is perpendicular to the previous one and (2) those in which the last two rotated faces are parallel. Any process of the first class can be specified by choosing its first $n - 1$ moves (S_{n-1} possibilities), then the last face, perpendicular to the $(n - 1)$st face (four ways), and then the angle of its rotation (three ways). This amounts to $3 \cdot 4 \cdot S_{n-1}$ processes in the first class. To specify a process of the second type, we choose an $(n - 2)$-move process (S_{n-2} ways), a pair of faces perpendicular to its last rotated face (two ways), and the angles by which these two faces are rotated ($3 \cdot 3 = 9$ ways). This amounts to $2 \cdot 9 \cdot S_{n-2}$ processes. The calculation for $n = 2$ is similar: $18 \cdot 4 \cdot 3 + 3 \cdot 9 = 243$.

2. The reasoning is similar to that for edge flips.

3. Verify and use the following fact: if two processes applied to a certain cube state produce different patterns, they will produce different patterns after being applied to any other state.

Monty's dilemma

1. See equation 1 in the box on the next page.

2. $p = 2/3$, $n = 8$, and $j = 3$. Substitute into equation (3) in the article to obtain $P_{\text{switch}} = 55/144 > 54/144 = 3/8 = P_{\text{stick}}$. You should switch.

1. $P_{\text{switch}} = P(\text{1st pick goat})P(\text{2nd pick car} \mid \text{1st pick goat})$

$$+P(\text{1st pick car})P(\text{2nd pick car} \mid \text{1st pick car})$$

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

$$< \left(\frac{j}{n}\right)\left(\frac{n-j}{n-2} + \frac{j-2}{n-2}\right) = \frac{j}{n} = P_{\text{stick}}.$$

2. $$p\left[\frac{n-j}{n}\left(\frac{j}{n-2}\right) + \frac{j}{n}\left(\frac{j-1}{n-2}\right)\right] + (1-p)\left[\frac{n-j}{n}\left(\frac{j-1}{n-2}\right) + \frac{j}{n}\left(\frac{j-2}{n-2}\right)\right]$$

$$= p\left[\frac{jn-j}{n(n-2)}\right] + (1-p)\left[\frac{jn-j-n}{n(n-2)}\right] = \frac{[p+(1-p)](jn-j)-(1-p)n}{n(n-2)}$$

$$= \frac{nj+np-n-j}{n(n-2)}.$$

3. Let $p = (n-j)/n + x$. Substituting for p in equation (3) we get $P_{\text{switch}} = j/n + x/(n-2)$. If $p > (n-j)/n$, then $x > 0$; thus, $P_{\text{switch}} > j/n = P_{\text{stick}}$. If $p < (n-j)/n$, then $x < 0$; thus, $P_{\text{switch}} < j/n = P_{\text{stick}}$.

4. Consider the equation $kn = jm + y$. If $k/m = j/n$, then $y = 0$; if $k/m < j/n$, then $y < 0$; and if $k/m > j/n$, then $y > 0$. Now, substituting for kn in equation (4) we have

$$P_{\text{switch}} = \frac{j}{n} - \frac{y}{n(n-m-1)}.$$

The result now follows.

5. $E_{\text{switch}} = \$18{,}850 > \$15{,}100 = E_{\text{stick}}$. The contestant should switch.

6. $E_{\text{switch}} = \$15{,}100 = E_{\text{stick}}$. The contestant can use either strategy.

7. Let $v_r = t/n + x$. Substitute for v_r in equation (5) to obtain $E_{\text{switch}} = t/n - x/(n-2)$. If $v_r = t/n$, then $x = 0$; if $v_r > t/n$, then $x > 0$; and if $v_r < t/n$, then $x < 0$. In all cases, the result follows. Note the similarity to the solution for problem 3.

8. If $x/s = t/n$, then $xn = st$. Substitute for xn in equation (6).

9. Combining terms in equation (3), we get equation 2 in the box above.

Index of Advertisers

American University in Moscow	47, Cover 4
Birkhäuser	8
Princeton University Press	28
U. S. Air Force Academy	Cover 3

Corrections

Brainteaser B131 in the last issue is (hopelessly, we fear) flawed. If you can find a valid solution, send it to us and we'll send you a handsome *Quantum* pin with our thanks.

Due to a copyediting error, the first sentence in the solution to brainteaser B134 says the opposite of what it should (as the rest of the solution makes clear). A tip of the foolscap to Andy Liu for bringing this to our attention. Andy also pointed out that this problem is closely related to problem 1 in the 1976 USAMO, which dealt with 4 × 7 and 4 × 6 rectangles.

"LAST PROBLEM OF THE CUBE" CONTINUED FROM PAGE 64

other if and only if their three invariants take the same values. In particular, accessible states are characterized by $F(S) = T(S) = 0°$, $P(S) = $ EVEN. This fact is far from obvious. Its proof virtually amounts to completing an algorithm for restoring the cube. Since there are $2 \cdot 3 \cdot 2 = 12$ triples of possible values of the invariants, all conceivable states of the cube fall into 12 classes such that by rotating faces we can reach any state within the class where we are, but we can't get into any other class.

Problem 3. Show that there are equally many states in all the classes.

This problem makes it clear why we had to divide the number N_0 by 12 to obtain the number of accessible states.

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Readers write . . .

M. Douglas McIlroy of Bernardsville, New Jersey, writes:

When I clapped my eyes on the cover of *Quantum* for November/December 1994, I recognized Norbert Wiener instantly. The likeness is uncanny: the pose, scrunched in a ship's wheel, is almost exactly as I first saw him forty years ago. There he sat in the faculty lounge, a goggle-eyed Humpty Dumpty enfolded low in an orange batwing chair, two short legs dangling over the edge. That ridiculous position only served to emphasize his authoritative presence. Thank you, Sergey Ivanov, for the memory.

The last problem of the cube

52, 44, 42, 35, ..., 21?

by Vladimir Dubrovsky

THE UNPRECEDENTED FAScination with Rubik's cube that seized the world some 15 years ago played a bad trick on the best puzzle of our century. Shortly after its appearance it fell prey to mathematicians, who jumped at a rare chance to demonstrate the power of their science to the general public. They eagerly subjected the toy to minute scrutiny and explained to one and all what could and couldn't be done with it. Puzzle inventors all over the world also took advantage of Rubik's idea and replicated it with numerous modifications. Soon the cube—too difficult to be solved by laypersons, completely chewed over by "cubologists," and stripped of its freshness and uniqueness by imitators—turned into a pretty knick-knack for some and a wonderful visual aid in group theory for others.

However, one important question about the cube remains unanswered, awaiting its . . . well, computer: what is the smallest number of moves (quarter- or half-turns of the cube's faces) sufficient to restore the original position of the cube from any scrambled state? This number is usually referred to as the "length of God's algorithm," an imaginary algorithm that always yields the shortest solutions.[1] In principle, such an algorithm can hardly be much different from a giant table of all patterns on the cube with their unscrambling processes obtained by a more or less economically organized but exhaustive search. Theoretically, it's no problem to write a computer program that would complete this task, though that wouldn't be very interesting. The problem is that the variety of patterns and processes turns out to be too great for such a program to be run in real time, even using the state-of-the-art hardware.

But human curiosity knows no bounds, and several cube addicts keep working on algorithms that are both realizable and close enough to their long-sought goal. Recent years have brought a considerable advance in this area. But to appreciate the latest achievements, we must understand how the closeness to God's algorithm can be estimated and review what's been done in this respect before.

The lower bound

It's very easy to estimate from below the number of states of the cube that can be produced by a sequence of n moves. The first move can be made in $6 \cdot 3 = 18$ different ways: we can turn any of the six faces by any of the three possible angles (90°, 180°, and 270°). So one move generates 18 different states. The second and each subsequent move can be chosen in $5 \cdot 3 = 15$ ways, because there's no sense in rotating the face used in the previous move again. So the number of two-move sequences is $18 \cdot 15$, the number of three-move sequences is $18 \cdot 15^2$, and so on. Adding the "no move" sequence, we get $M_n = 1 + 18 \cdot (1 + 15 + \ldots + 15^{n-1})$ sequences consisting of no more than n moves. These moves produce no more than M_n different states of the cube—fewer than M_n, in fact, because different sequences can generate the same state, even if they are very short. For instance, two successive rotations of parallel faces can be performed in any order with the same effect.[2]

Now suppose the total number of cube states is N. Then we can be sure that for any n such that $M_n < N$ there exists a state that can't be obtained from a given one (and can't be unscrambled) in n or fewer moves—n moves simply can't produce as many as N states.

To obtain a concrete lower bound, we have to calculate N. In accordance with cubist convention, we'll call the small cubes that form

[1]This term and most of the other "cubological" terms in this article were introduced and no doubt created by David Singmaster, the British mathematician who can truly be called the "leading force in cubism" (a kind of Picasso of the Rubik's cube!).

[2]Or a more interesting example: two successive half-turns of *adjacent* faces repeated six times bring the cube back to its initial state. So all thirty of these 12-move processes are in fact equivalent to the "no move" process and should not be counted at all.

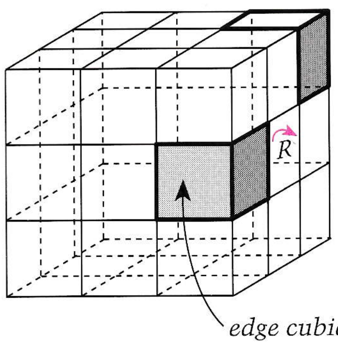


Figure 1

the large Rubik's cube "cubies" (see figure 1). There are $8! = 1 \cdot 2 \cdot \ldots \cdot 8$ rearrangements of the eight cubies in the corners of the big cube and, similarly, 12! rearrangements of the edge cubies. Now we must take into account the different orientations of cubies in their "nests": a corner cubie can be turned in its place in three ways, which makes a total of 3^8 possibilities, and an edge cubie can be "flipped" (rotated 180°), which gives 2^{12} possibilities for all 12 edges. So we come up with the product $N_0 = 8! \cdot 12! \cdot 3^8 \cdot 2^{12}$.

Actually, this is the number of all "conceivable" ways in which the entire cube can be reassembled after being taken apart. But only some of these possibilities can be realized in accordance with the rules of the game—that is, by rotating faces. Like many other transformational puzzles (for instance, the triads, discussed in the Toy Store in the last issue), Rubik's cube has a set of *invariants* that place certain restrictions on the possible patterns. We'll talk about them in the last section below. For now I'll just give the result: to obtain the number N of cube states produced by all possible processes starting from a certain fixed state, we must divide N_0 by 12: $N = N_0/12 \cong 4.3 \cdot 10^{19}$.

You can verify that $M_{16} < N < M_{17}$, so the length of God's algorithm is no less than 17 moves. This estimate can be slightly improved by noticing that the processes that contain three consecutive turns of parallel faces are redundant. Let R and L stand for clockwise quarter-turns of the right and left cube faces, respectively, and let the double move RR be denoted as R^2. The three-move sequence RLR, for example, can be replaced by R^2L or LR^2, so these three sequences should be counted as one. One can find the number S_n of n-move processes taking into account redundancies of this kind—that is, ignoring processes that include three successive turns of parallel faces and counting all processes that differ from one another only by the order of successive "parallel moves" as one process.

Problem 1. Show that $S_n = 12S_{n-1} + 18S_{n-2}$ for $n \geq 3$ with $S_0 = 1$, $S_1 = 18$, $S_2 = 243$. Check that, for $T_n = S_0 + S_1 + \ldots + S_n$, $T_{17} < N < T_{18}$.

It follows from this problem that some states of the cube can't be unscrambled in fewer than 18 moves.

On the way to God's algorithm

The "golden age" of Rubik's cube produced a plethora of restoring algorithms. However different they were, almost all of them could be characterized by two words: "geometric" and "manual." They were manual because they were created "by hand" and could be performed "by hand," and geometric because the order of restoration depended on the cube's geometry. In most of these algorithms you had to drive cubies to their destined positions one-by-one, except for the last stages, in which you worked on two or three cubies at a time using specially devised, tricky processes. Corners and edges, as well as locations and orientations, were usually treated separately, and as you moved along the lines prescribed by the algorithm, you could see each of the cube's faces taking on a single color.

The reported lengths of such algorithms ranged from several hundred to about 70 moves. (I don't think, though, that many of these figures ever were scrupulously verified by anybody except the authors of the algorithms, so I'll refrain from citing more exact numbers.) In his *Notes on Rubik's "Magic Cube"* (one of the classics of cubology), David Singmaster writes that another famous cubologist, Morwen B. Thistlethwaite, invented an algorithm of at most 63 moves that first set edges to rights, then corners. This is the shortest geometric, manual algorithm I've heard of.

Perhaps it shouldn't be called "manual": its author made extensive use of his computer in solving the cube, so I guess at least part of this algorithm should be attributed to the machine. And it was Morwen Thistlethwaite with his computer who made a real breakthrough in the pursuit of God's algorithm. In 1980 he (or should I say they?) created a 52-move algorithm that was neither geometric nor manual (in the sense specified above).

The algorithm consisted of four stages. As they were defined, the stages didn't give an explicit idea of how cubies should be rearranged. It was simply postulated that in the first stage you could use all face turns; in the second stage two parallel faces (say, front and back) were allowed to be turned only by 180° and the other four by any angle; in the third stage the same restriction was additionally imposed on another pair of parallel faces (say, right and left); and in the fourth stage the rotations of the third pair of faces were similarly restricted—only "square turns" were permitted. Adding for convenience the "fifth stage," where no turns at all are allowed, we can formulate the goal of stages 1 through 4 as bringing the cube into a position from which it can be unscrambled using only the processes of the next stage. So the fourth stage is intended to complete the solution.

One of the aspects of this approach that makes it well suited for computer treatment is that all achievements of each stage are automatically saved in subsequent stages (think why!). Figures 2–5 give geometric illustrations of what is done at each stage in this algorithm (see also the last section). However, the illustrations won't be much help to you in solving the cube by this method. You'll have to look up the current states of your cube in extensive tables with hundreds of entries

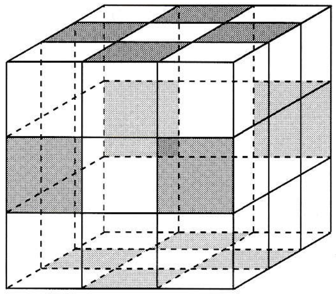


Figure 2

If this were the original coloration of the cube, then it will be restored after the first stage of Thistlethwaite's algorithm.

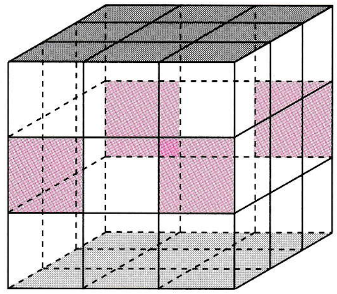


Figure 3

If this were the original coloration of the cube, the second stage will restore it.

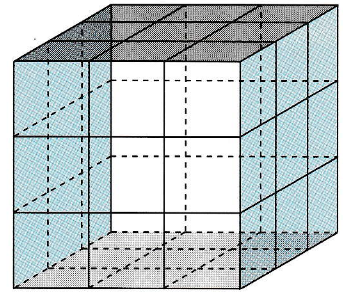


Figure 4

The third stage of Thistlethwaite's algorithm must (1) restore the original coloration in figure 4 and (2) put the corner cubies in an arrangement such that after joining originally opposite corners we get one of the four patterns of lines shown in figures 5a through 5d (the original pattern is 5a).

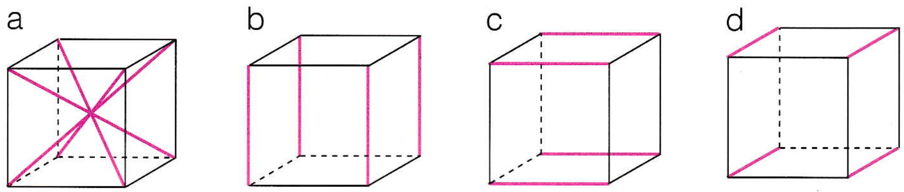


Figure 5

to find appropriate processes. I don't think there are many enthusiasts who would wish to do this, though perhaps it's still within human ("manual") capabilities.

But let's return to the history of record cubing. In 1982 a group of Donald Knuth's students confirmed Thistlethwaite's conjecture that the last stage of his algorithm can be done in 15 rather than 17 moves and proved that this number can't be reduced further.

The record of 50 moves held on for seven years as a new generation of computers entered the scene. In 1989 the Dutch cubist Hans Kloosterman, the "last of the Mohicans" (or at least, one of the precious few), published a description of his 44-move algorithm in *Cubism for Fun* (*CFF*), a newsletter of the Dutch Cubists Club. (These are perhaps the only cubist newsletter and club that survived after the cube craze died away.) It took him about a year to improve that by two moves. Basically, Kloosterman follows in Thistlethwaite's footsteps. He spends the same number of no more than seven moves for the first stage, but reduces the previous 13 moves in the second stage to 10 (as was predicted by the method's originator). More importantly, he eliminates the irreducible fourth stage, instead of which the third stage is "geometrically" subdivided (within the same set of allowed moves) into (1) putting the cubies of the up and down faces into their respective layers and (2) putting all the rest in order. This stage requires at most 25 moves altogether. Kloosterman asserts that these three numbers (7, 10, 25) are the best possible for the chosen scheme.

However, a short time later, from a letter from David Singmaster to the editors of *Quantum*, I found out that in 1991 someone in the Netherlands—presumably Kloosterman—had got down to 35 moves (he must have improved the scheme)! Since then, there has been no confirmation of this result (not in the *CFF* newsletter, at any rate). Nor have any other improvements been reported. So at this point we definitely know that the length of God's algorithm is a number between 18 and 42, perhaps between 18 and 35. We can only guess about more exact estimates. But there's something to base our guesses on: in the April 1992 issue of *CFF* Herbert Kociemba of Darmstadt, Germany, described an algorithm that had been solving all positions proposed to it in no more than 21 moves!

It should be made very clear that this does *not* mean that the length of Kociemba's algorithm is 21. There's no guarantee that some unknown pattern won't require more moves, even with this program. But so far it hasn't been found.

A few words about the program. Kociemba eliminated both the second and fourth stages of Thistlethwaite's algorithm. So the aim of his first stage is to restore the coloring shown in figure 3 (assuming that, before it was scrambled, the cube was colored this way), and the second stage restores the original coloring using any turns of the horizontal faces and only half-turns of the other four faces. The total number of variants to be examined according to this scheme still exceeds the capacity of modern computers. So instead of making an exhaustive search, Kociemba's program treats each given pattern individually, checking if the processes it generates allow the current stage of the algorithm (first or second) to be completed in a certain prescribed number of moves. So the solutions it finds are "short enough" (they really are!) rather than the shortest (it may find a shortest solution, of course, but it wouldn't tell us).

Although they're far from definitive, Kociemba's results make it very plausible that the length of

God's algorithm is somewhere in the low twenties, as group theorists conjectured back in the cube's heyday.

The invariants of Rubik's cube

Now, I suppose some math should be added to "*Quantum*ize" this article. I mentioned above that only some of all the conceivable states of the cube can be obtained from the original regular state by rotating faces. These states—let's call them *accessible*—can be described in terms of invariants. Invariants are certain values that depend on the arrangement of the pieces and are preserved under face rotations. The cube has three invariants—they restrict accessible orientations of edge and corner cubies and their permutations.

Let's begin with edge orientations. Consider the coloring of edges shown in figure 2—the *reference edge coloring.* Suppose that all the edge cubies were initially colored according to this pattern. Now scramble the cube (but memorize the reference coloring). Define the *flip* of each edge cubie as 0° or 180° depending on whether its colored facelet coincides with the reference colored facelet on its edge. The *total flip* $F(S)$ of a state S is the sum of all 12 individual flips counted modulo 360° (as we usually do with angles of rotation)—that is, $F(S) = 0°$ or 180° if the number of flipped cubies is even or odd, respectively. The total flip is our first invariant.

Indeed, consider a quarter-turn of any face. It can only affect the flips of the edge pieces in this face. But it clearly doesn't change these flips if, say, the up face is rotated, because this rotation takes the reference coloring of this face into itself. The same argument applies to the down, right, and left faces and to half-turns of the front and back faces as well. As to *quarter*-turns of the front (or back) face, we can argue as follows. Suppose we alter the reference coloring on the two vertical edges in this face (fig. 6). Then the flips of exactly two edge pieces (on the altered edges) will change by 180°, so the total flip with respect to the new coloring will stay the same as it was initially. But now the argument used above shows that front turns preserve the modified total flip and so preserve the original flip (which is equal to it) as well.

For accessible states, $F(S) = 0°$, because initially all edges are unflipped. There are 2^{12} conceivable ways of flipping the cube's edges, but this restriction leaves only half of this range of possibilities for accessible states. As follows from the proof above, all the processes allowed in Thistlethwaite's algorithm after the first stage leave invariant not only the total flip but the flips of all individual edge pieces. So the aim of the first stage can be formulated as "to unflip all edge cubies" (with respect to the reference coloring in figure 2), and there are $2^{11} = 2{,}048$ essentially different variants to be considered in this stage.

The corner-orientation invariant is defined in much the same way. Here the reference coloring is the one in figure 7. In a scrambled state each corner cubie must be turned 0°, 120°, or −120° about the corresponding diagonal of the cube to match the reference position. This angle of rotation is called the *twist* of this cubie. The sum of all eight twists mod 360° is the *total twist* $T(S)$ of the given state S; it takes three values: 0°, 120°, and −120° (or 240°, if you prefer).

Problem 2. Show that $T(S)$ is invariant under face rotations and all eight "individual" twists are invariant under the processes of the third and fourth stages of Thistlethwaite's algorithm.

For accessible states, $T(S) = 0°$, and the number of possible "twistings" of corners is 3^7.

The third invariant concerns only the locations of the 12 + 8 = 20 movable cubies rather than their orientations, or, to put it mathematically, the permutation of the cubies. Any permutation of any objects can be represented as a number of successive pair exchanges. The parity of this number is called the *parity of the given permutation*. It can be shown that the parity of a permutation is well defined (that is, does not depend on the representation in terms of pair exchanges) and that the parity $P(S)$ of the permutation of our 20 cubies in a given state is invariant under face rotations. Taking the first of these facts for granted, you can prove the second as an exercise (by representing a quarter-turn of any face in terms of pair exchanges). The first fact also has an elementary proof, but it's too long to be included here. (Details can be found in "Some Things Never Change" in the September/October 1993 issue of *Quantum.*)

The invariant $P(S)$ takes two values, EVEN and ODD, and the states of either kind constitute exactly half of the total number $12! \cdot 8!$ of the cubies' permutations (not 20!, because corners and edges don't mix). Accessible permutations are even, because the original identity permutation is representable as an even number (zero) of pair exchanges.

The three invariants $F(S)$, $T(S)$, and $P(S)$ form a complete system. That is, any two of the N_0 conceivable states (obtained by taking the cube apart and reassembling it at random) are convertible into each

CONTINUED ON PAGE 60

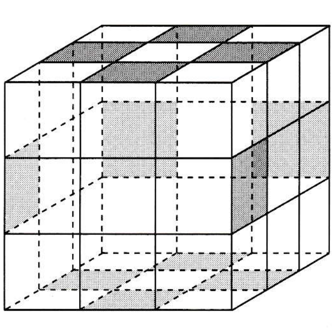


Figure 6

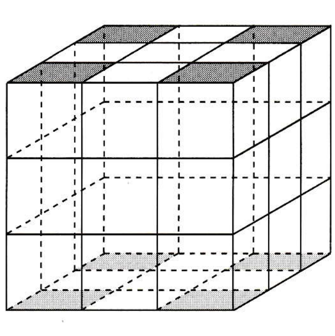


Figure 7