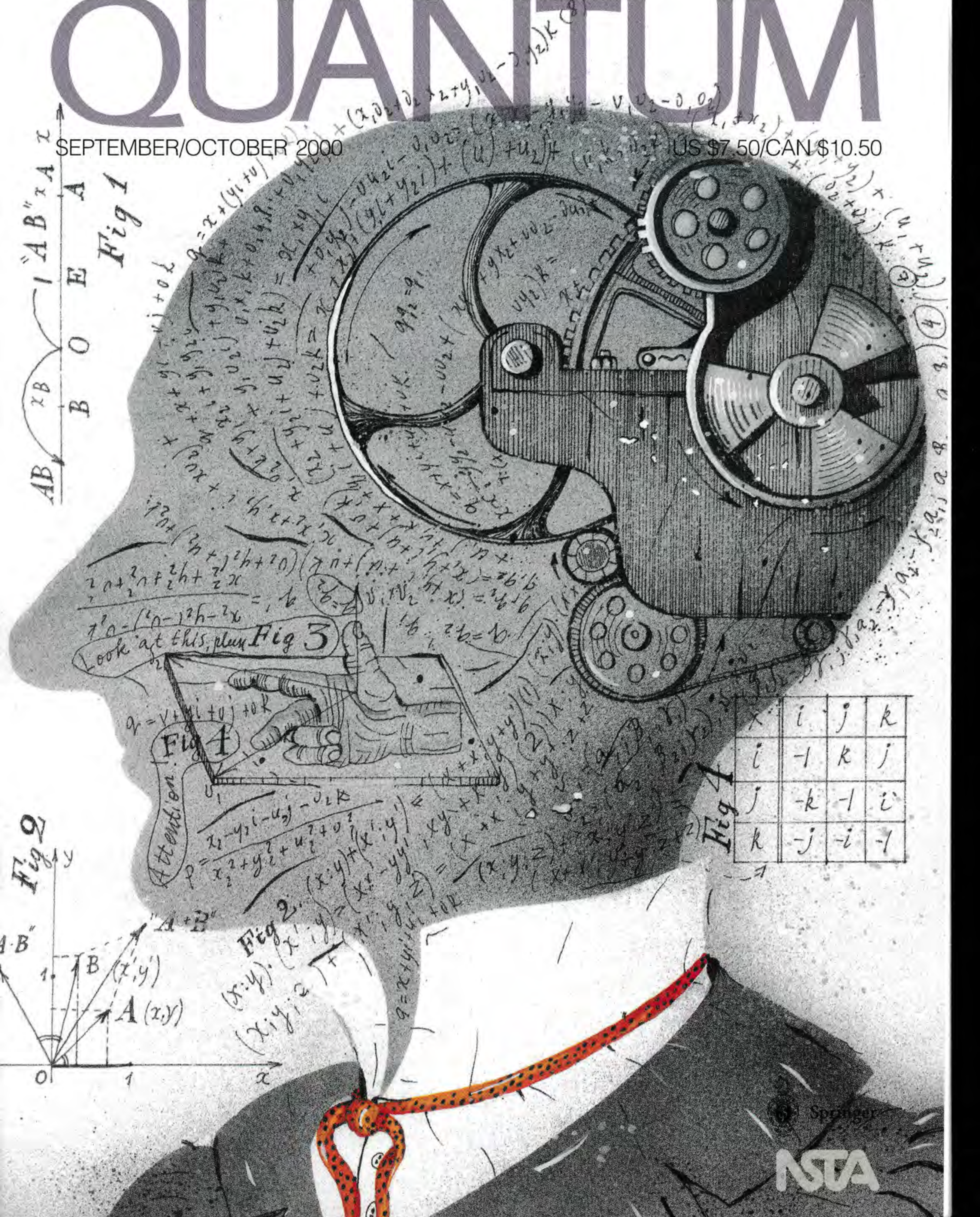
QUANTUM

SEPTEMBER/OCTOBER 2000

US $7.50/CAN $10.50



GALLERY Q

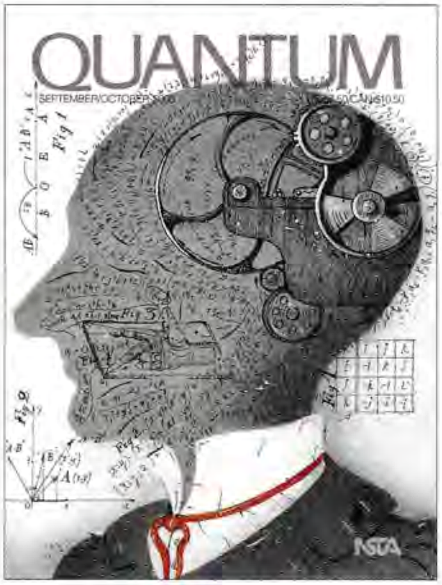
Watercolor conversion of print of painting by Montague Dawson, 25 × 18, Collection of The Salvador Dali Museum, St. Petersburg, Florida

The Ship (1942–43) by Salvador Dali

THIS ANIMATED FIGUREHEAD MAY ALSO HAVE a head for figures judging by the elaborate rigging she has woven between her limbs. Similar systems can be devised to synchronize the movement of the arms and legs to achieve a resonant frequency that will increase one's speed dramatically. Hopefully, the high winds and seas seen in the background will allow her to resume her more leisurely duties on the prow of the ship before long. If you'd like to learn more about her elaborate lines of locomotion, turn to "The physics of walking" on page 20.

QUANTUM

SEPTEMBER/OCTOBER 2000 VOLUME 11, NUMBER 1



Cover art by Ekaterina Silina

Explore yet another way to represent real numbers through quarternions, quadruples of real numbers that have a very interesting history. Learn how to obtain numbers from points in space and the secrets of adding, subtracting, and multiplying these peculiar points. Turn to page 4 for a closer look at these complex numbers.

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

FEATURES

DEPARTMENTS

Quantum (ISSN 1048-8820) is published bimonthly by the National Science Teachers Association in cooperation with Springer-Verlag New York, Inc. Volume 11 (6 issues) will be published in 2000–2001. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by Quantum Bureau (Moscow, Russia), as well as original material in English. **Editorial offices:** NSTA, 1840 Wilson Boulevard, Arlington VA 22201-3000; telephone: (703) 243-7100; electronic mail: quantum@nsta.org. **Production offices:** Springer-Verlag New York, Inc., 175 Fifth Avenue, New York NY 10010-7858.

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

Subscription Information:
North America: Student rate: $18; Personal rate (nonstudent): $25. This rate is available to individual subscribers for personal use only from Springer-Verlag New York, Inc., when paid by personal check or charge. Subscriptions are entered with prepayment only. Institutional rate: $56. Single Issue Price: $7.50. Rates include postage and handling. (Canadian customers please add 7% GST to subscription price. Springer-Verlag GST registration number is 123394918.) Subscriptions begin with next published issue (backstarts may be requested). Bulk rates for students are available. Mail order and payment to: Springer-Verlag New York, Inc., Journal Fulfillment Services Department, PO Box 2485, Secaucus, NJ 07094-2485, USA. Telephone: 1(800) SPRINGER; fax: (201) 348-4505; e-mail: custserv@springer-ny.com.

Outside North America: Personal rate: Please contact Springer-Verlag Berlin at subscriptions@springer.de. Institutional rate is US$57; airmail delivery is US$18 additional (all rates calculated in DM at the exchange rate current at the time of purchase). SAL (Surface Airmail Listed) is mandatory for Japan, India, Australia, and New Zealand. Customers should ask for the appropriate price list. Orders may be placed through your bookseller or directly through Springer-Verlag, Postfach 31 13 40, D-10643 Berlin, Germany.

Advertising:
Representatives: (Washington) Paul Kuntzler (703) 243-7100; (New York) Joe Antino (212) 460-1575; and G. Probst, Springer-Verlag GmbH & Co. KG, D-14191 Berlin, Germany, telephone 49 (0) 30-827 87-0, telex 185 411.

Printed on acid-free paper.

Visit us on the internet at www.nsta.org/quantum/.

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 mission of the National Science Teachers Association is to promote excellence and innovation in science teaching and learning for all.

Just for the fun of it!

B301

Be fruitful and multiply. One day all of Mrs. Brown's grandchildren came to visit her. There was a bowl of apples and pears on the kitchen table. Mrs. Brown gave each child the same number of pieces of fruit without keeping track of which kind. Billy got 1/8 of all the apples and 1/10 of all the pears. How many grandchildren did Mrs. Brown have?

B302

Overbooked dance card. A school dance was held where girls danced with boys. At the end of the evening, all the girls and boys were asked how many songs they danced to (with a partner, of course). Six said that they danced to three songs, one reported dancing to five songs, four danced to six songs, and one danced to nine. Prove that not all the answers were correct.

B303

Ratings game. Sixteen baseball teams were ranked 1 through 16 according to the results of a series of games. After the ranking, a playoff was held, in four rounds. In the first round, eight teams were eliminated; in the second round, four teams were eliminated; in the third round, two teams were eliminated; and the two remaining teams met in the championship game. It turned out that the team with the higher ranking won every game. A game was considered "interesting" if the ranking of the teams differed by not more than 4; otherwise, the game was considered "uninteresting." What is the minimum possible number of "uninteresting" games?

B304

Cross coverage. A unit cube is unfolded to form the figure held by the gentleman on the left. Use a number of copies of this figure to cover, without overlap, the surface of a cube with edge-length twice as big.

B305

Pirates with an image problem. On a roll of photographic film it says: "Develop in red light." Will such film reproduce the Jolly Roger in the picture on the left? How about the one on the right?

ANSWERS, HINTS & SOLUTIONS ON PAGE 50

Art by Pavel Chernusky

Quaternions

Simple operations with complex numbers

by A. Mishchenko and Y. Solovyov

WE HAVE RECENTLY PUBLISHED ARTICLES on the representation of real numbers as continued fractions (January/February 2000), and algebraic and transcendental numbers (July/August 2000). The present article completes this series (however, familiarity with the articles mentioned above is not required). It's about quaternions. These numbers include complex numbers and, therefore, all other kinds of numbers we've mentioned. It contains some information about applications of quaternions and of their peculiar history. The starting point was the desire to introduce an algebraic structure (addition and, most importantly, multiplication) for geometric objects (various sets of points).

How can we get numbers from points?

If we're talking about points on a line, it's easy. Choosing an origin ("zero") and a scale with a direction ("unit"), we can turn a line into a number line. Every point is assigned a real number—its coordinate (figure 1).

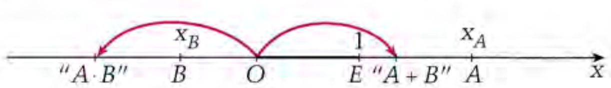


Figure 1

With points in the plane, the situation is more complicated. Choosing an origin ("zero") and a pair of perpendicular axes, we can assign a pair of coordinates (x, y) to every point in the plane. To make numbers out of these pairs, we must learn how to "add" and "multiply" them to retain the familiar properties of addition and multiplication: the commutative, associative, and distributive laws, and the existence of the inverse operations of division and subtraction.

Addition is easy: it's quite natural to add pairs coordinate by coordinate, just as we add vectors (figure 2):

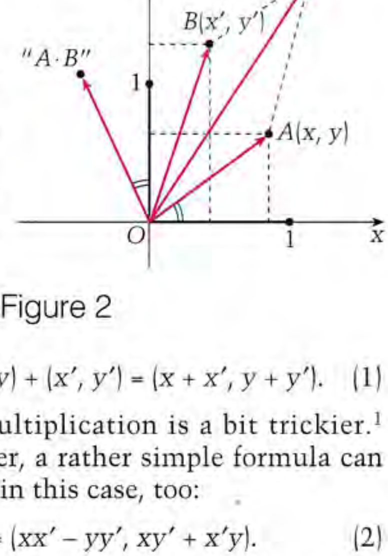


Figure 2

$$(x, y) + (x', y') = (x + x', y + y'). \quad (1)$$

Multiplication is a bit trickier.[1] However, a rather simple formula can be found in this case, too:

$$(x, y)(x', y') = (xx' - yy', xy' + x'y). \quad (2)$$

We can verify that this operation of multiplication, and that of addition (1), possess all the familiar properties listed above. Thus the set of pairs with operations (1) and (2) may be considered a number set.

[1]Multiplication by coordinates—that is, $(x, y)(x', y') = (xx', yy')$—is not satisfactory, because this operation has no inverse operation. (For example, division by the nonzero pair (0, 2) is impossible.)

Art by Ekaterina Silina

These pairs are, in fact, *complex numbers*. They are more often written as $x + yi$ rather than (x, y), where i is the imaginary unit (the pair $(0, 1)$) possessing the remarkable property $i^2 = ii = -1$. This property makes it possible to extract square roots from negative numbers (in the field of complex numbers).

Can we turn points in space into numbers? As before, we can introduce a coordinate system and describe a point as a set of three coordinates: (x, y, z). It's natural to add these triples coordinate by coordinate:

$$(x, y, z) + (x', y', z') = (x + x', y + y', z + z'). \qquad (3)$$

We will be able to consider triples as numbers if we find a method of multiplying them while retaining the familiar properties of addition and multiplication. In particular, multiplication must have an inverse operation (division by nonzero elements).

In 1833 the Irish mathematician William Rowan Hamilton (1805–1865) took an interest in this problem. The following is the story of this remarkable man's obsession with triples and quadruples.

W. R. Hamilton

Hamilton was a very capable, well-rounded person. At the age of 10 he knew many lines of Homer by heart, and at 14 he spoke nine languages. In 1824 he published a paper on geometric optics in the *Transactions of the Royal Irish Academy*, and in 1827 he became the royal astronomer of Ireland.

By 1833 Hamilton was the director of the Dunsink Observatory near Dublin and had published several papers on optics and analytical mechanics. Based on his work in geometric optics, Hamilton predicted the effect of double conical refraction in biaxial crystals, which was discovered soon after by his colleague Lloyd.

For ten long years Hamilton tried to devise a rule for multiplying triples, but without success. He later recalled, in a letter to his son: "Every morning, when I came down to breakfast, you and your brother would ask me: 'Daddy, have you learned to multiply triples?' And always I would have to reply, sadly: 'No, I can only add and subtract them.'"

Vector products

The problem that occupied Hamilton might not seem difficult at first glance. It's clear how vector addition can be defined (by means of formula (3)); all one needs is a formula for multiplication—something like formula (2). But none of the formulas that Hamilton tried worked—one of the properties was always violated.

The operation of vector multiplication was already well known at that time: the *vector product* $\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2$ of two nonzero vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is the vector perpendicular to both $\mathbf{v}_1$ and $\mathbf{v}_2$ and directed as specified by the right-hand rule (figure 3). The length of this vector is defined as $|\mathbf{v}_1| \cdot |\mathbf{v}_2| \cdot \sin(\mathbf{v}_1, \mathbf{v}_2)$. A basic result about vector products (sometimes called "cross products") is that if $\mathbf{v}_1$ and $\mathbf{v}_2$ are given by their coordinates in a Cartesian reference frame—

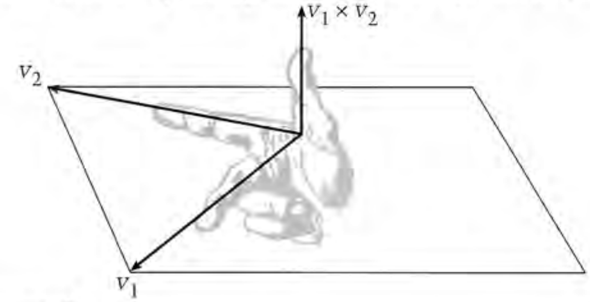


Figure 3

$$\mathbf{v}_1 = (\alpha_1, \beta_1, \gamma_1),$$
$$\mathbf{v}_2 = (\alpha_2, \beta_2, \gamma_2)$$

—then

$$\mathbf{v}_1 \times \mathbf{v}_2 = (\beta_1\gamma_2 - \beta_2\gamma_1, \gamma_1\alpha_2 - \gamma_2\alpha_1, \alpha_1\beta_2 - \alpha_2\beta_1). \qquad (4)$$

However, vector products weren't suitable for Hamilton because the inverse operation isn't valid. For example, if $\mathbf{v}_1 = \mathbf{v}_2 \neq \mathbf{0}$, then the angle $(\mathbf{v}_1, \mathbf{v}_2)$ between these vectors is zero. Therefore, the length of the vector product $\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2$ is zero (substituting into formula (4) will give the same result). If the operation of division by any nonzero vector existed, we would have $(\mathbf{v}_1 \times \mathbf{v}_2) : \mathbf{v}_2 = \mathbf{v}_1 \neq 0$. At the same time, $\mathbf{v}_1 \times \mathbf{v}_2 = 0$ and, therefore, $(\mathbf{v}_1 \times \mathbf{v}_2) : \mathbf{v}_2 = 0$. The contradiction obtained proves that division by $\mathbf{v}_2$ is impossible.

In the face of his disappointments and failures, Hamilton kept trying, with admirable persistence. Although his problem wasn't solved (and couldn't be solved, as we'll see below), his ten-year effort was rewarded. One day in 1843 Hamilton decided to try defining multiplication for quadruples rather than for triples. He called this new type of number quaternions. Here's how it happened.

An event on the Brougham bridge

In a letter to his son, Hamilton recalled that the idea of how the multiplication for quaternions could be defined occurred to him when he was walking along the Royal Canal to a meeting of the Irish Academy. Hamilton was so pleased with the idea, he couldn't resist taking out his penknife and scratching the fundamental formula in the soft stone of the Brougham bridge:

$$i^2 = j^2 = k^2 = ijk = -1.$$

Definition of quaternions

Quaternions are quadruples of real numbers (x, y, u, v) that can be conveniently written as

$$q = x + yi + uj + vk,$$

where i, j, and k are new numbers similar to the imaginary unit used in defining complex numbers. They must satisfy the relations

$$i^2 = j^2 = k^2 = -1, \qquad (5)$$

$$ij = -ji = k,\ jk = -kj = i,\ ki = -ik = j. \qquad (6)$$

These relations can be written in the form of a "multiplication" table:

	i	j	k
i	-1	k	$-j$
j	$-k$	-1	i
k	j	$-i$	-1

By definition, addition and multiplication of quaternions are performed according to the conventional rules of removing parentheses and collecting similar terms, taking into account the rules contained in formulas (5) and (6).

By this definition, if q_1 and q_2 are two quaternions, then

$$\begin{aligned} q_1 + q_2 &= (x_1 + y_1 i + u_1 j + v_1 k) + (x_2 + y_2 i + u_2 j + v_2 k) = \\ &= x_1 + y_1 i + u_1 j + v_1 k + x_2 + y_2 i + u_2 j + v_2 k \\ &= (x_1 + x_2) + (y_1 i + y_2 i) + (u_1 j + u_2 j) + (v_1 k + v_2 k) \\ &= (x_1 + x_2) + (y_1 + y_2)i + (u_1 + u_2)j + (v_1 + v_2)k. \qquad (7) \end{aligned}$$

This is conventional addition by coordinates, which we looked at earlier. The product of two quaternions q_1 and q_2 is calculated as

$$\begin{aligned} q_1 q_2 &= (x_1 + y_1 i + u_1 j + v_1 k)(x_2 + y_2 i + u_2 j + v_2 k) \\ &= x_1x_2 + x_1y_2 i + x_1u_2 j + x_1v_2 k + y_1x_2 i + y_1y_2 i^2 \\ &\quad + y_1u_2\, ij + y_1v_2\, ik + u_1x_2 j + u_1y_2\, ji + u_1u_2 j^2 \\ &\quad + u_1v_2\, jk + v_1x_2 k + v_1y_2\, ki + v_1u_2\, kj + v_1v_2 k^2 \\ &= x_1x_2 + x_1y_2 i + x_1u_2 j + x_1v_2 k + y_1x_2 i \\ &\quad - y_1y_2 + y_1u_2 k - y_1v_2 j + u_1x_2 j - u_1y_2 k \\ &\quad - u_1u_2 + u_1v_2 i + v_1x_2 k + v_1y_2 j - v_1u_2 i - v_1v_2 \\ &= (x_1x_2 - y_1y_2 - u_1u_2 - v_1v_2) + (x_1y_2 + y_1x_2 + u_1v_2 - v_1u_2)i \\ &\quad + (x_1u_2 + u_1x_2 - y_1v_2 + v_1y_2)j \\ &\quad + (x_1v_2 + y_1u_2 - u_1y_2 + v_1x_2)k. \qquad (8) \end{aligned}$$

That is, we act as if i, j, and k are variables, and expand using the distributive law. However, the order of multiplication of the "variables" i, j, and k must be preserved (although the order of the "coefficients" x_1, x_2, etc. may be changed).

A tedious but straightforward computation will show that multiplication of quaternions possesses the associative property:

$$(q_1q_2)q_3 = q_1(q_2q_3).$$

It is natural to consider real and complex numbers as particular cases of quaternions. The real number x is the quaternion

$$x = x + 0 \cdot i + 0 \cdot j + 0 \cdot k.$$

The complex number $z = x + iy$ can be written as the quaternion

$$z = x + iy + 0 \cdot j + 0 \cdot k. \qquad (9)$$

If you're not familiar with complex numbers, you can consider formula (9), together with (7) and (8) as their definition. Write the multiplication formula (8) for the case (9) and compare the result with formula (2).

It's clear that addition of quaternions has an inverse operation—subtraction. The difference of two quaternions q_1 and q_2 is defined by the formula

$$q_1 - q_2 = (x_1 - x_2) + (y_1 - y_2)i + (u_1 - u_2)j + (v_1 - v_2)k.$$

If $q_1 = q_2$, then the difference is the zero quaternion

$$q_1 - q_2 = 0 + 0 \cdot i + 0 \cdot j + 0 \cdot k = 0.$$

Division of quaternions

Consider the division of quaternions, which is the inverse of multiplication. In general, the quotient of division of a number a by a number $b \neq 0$ is the number c such that

$$bc = a. \qquad (10)$$

This is how the quotient for real and complex numbers is defined. Unfortunately, this definition doesn't apply to quaternions. In order for formula (10) to give a correct definition of the quotient, multiplication must be independent of the order of the factors. Otherwise, in addition to the quotient $c = b^{-1}a$ defined by formula (10), there exists the equally correct "left" quotient c' defined by the formula

$$c'b = a.$$

The left quotient can be different from c in formula (10). So besides the need to go beyond three-dimensional space, Hamilton had to make another sacrifice.

It turned out that quaternions lack another conventional property: the product of quaternions is *not* independent of the order of the factors. Indeed, we see from formulas (6) that the product changes sign when the order of the factors changes.

So we can speak only of "left division" and "right division." How can we find, say, the left quotient of

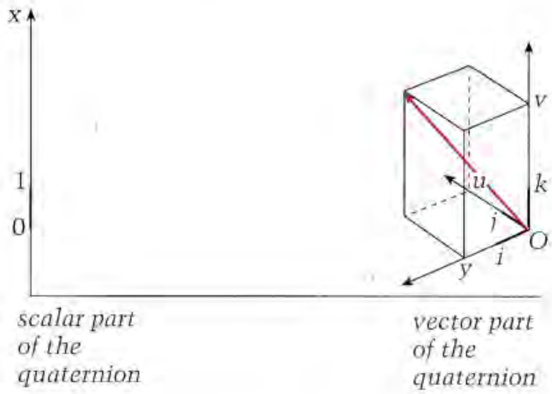


Figure 4

division of the quaternion q_1 by the quaternion $q_2 \neq 0$?

Denote the desired quotient by $q = x + yi + uj + vk$. Using the definition for multiplication of quaternions and the definition of left division, we obtain the following equality:

$$qq_2 = q_1$$

or

$$(xx_2 - yy_2 - uu_2 - vv_2) + (xy_2 + yx_2 + uv_2 - vu_2)i + (xu_2 + ux_2 - yv_2 + vy_2)j + (xv_2 + vx_2 + yu_2 - uy_2)k = x_1 + y_1 i + u_1 j + v_1 k.$$

This equality is equivalent to the following system of linear equations in the unknowns x, y, u, and v:

$$\begin{aligned} x_2x - y_2y - u_2u - v_2v &= x_1, \\ y_2x + x_2y + v_2u - u_2v &= y_1, \\ u_2x - v_2y + x_2u + y_2v &= u_1, \\ v_2x + u_2y - y_2u + x_2v &= v_1. \end{aligned}$$

We leave it to the reader to solve this system and find the left quotient of division of q_1 by q_2. Similarly, we can find the right quotient after division of q_1 by q_2.

If the dividend q_1 equals the real number 1, the quotient of left and right division of $q_1 = 1$ by the quaternion q_2 is the same quaternion

$$p = \frac{x_2 - y_2 i - u_2 j - v_2 k}{x_2^2 + y_2^2 + u_2^2 + v_2^2}$$

(we leave the proof to the reader). For this reason, the quaternion p is denoted by

$$q_2^{-1} = \frac{x_2 - y_2 i - u_2 j - v_2 k}{x_2^2 + y_2^2 + u_2^2 + v_2^2}.$$

The right quotient of division of q_1 by the nonzero quaternion q_2 is written as

$$q = q_2^{-1} \cdot q_1,$$

and the left quotient of division of q_1 by q_2 is written as

$$q = q_1 \cdot q_2^{-1}.$$

In practice, division of quaternions is performed by another method. In order to explain it, we have to make use of the scalar and vector parts of quaternions.

Scalar and vector quaternions

Like complex numbers, which can be broken down into their real and imaginary parts, the quaternion

$$q = x + yi + uj + vk$$

can be written as

$$q = x + (yi + uj + vk).$$

The first term in this expression is called the *scalar part* of the quaternion, and the second term (in parentheses) is called its *vector part*. The scalar part x is a real number, and the vector part can be considered a vector

$$\mathbf{r} = yi + uj + vk$$

in three-dimensional space, where i, j, and k are considered unit vectors in a Cartesian reference frame (figure 4).

Thus every quaternion q is written as the sum

$$q = x + \mathbf{r},$$

where x is the scalar part of q and $\mathbf{r}$ is its vector part. If $\mathbf{r} = 0$, then $q = x$, and q is called a *scalar quaternion*. If $x = 0$, then $q = \mathbf{r}$, and q is called a *vector quaternion*.

When quaternions are added, their scalar and vector parts are added independently.

Multiplication of quaternions is more complicated. If q_1 and q_2 are scalar quaternions, then their product q_1q_2 is a scalar quaternion as well. When $q_1 = x$ is scalar and $q_2 = \mathbf{r}$ is a vector quaternion, the product

$$q_1q_2 = x \cdot (yi + uj + vk) = (xy)i + (xu)j + (xv)k$$

is a vector quaternion and multiplication coincides with multiplication of the vector $\mathbf{r}$ by the real number x in three-dimensional space.

Finally, if both q_1 and q_2 are vector quaternions—that is,

$$\begin{aligned} q_1 &= \mathbf{r}_1 = y_1 i + u_1 j + v_1 k \\ q_2 &= \mathbf{r}_2 = y_2 i + u_2 j + v_2 k, \end{aligned}$$

—then

$$q_1q_2 = -(y_1y_2 + u_1u_2 + v_1v_2) + (u_1v_2 - v_1u_2)i + (v_1y_2 - y_1v_2)j + (y_1u_2 - u_1y_2)k.$$

As we can see from this formula, the scalar part of q_1q_2 equals the scalar product $(\mathbf{r}_1, \mathbf{r}_2)$ of the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ with the sign reversed. The vector part of q_1q_2 is the familiar vector product $\mathbf{r} \times \mathbf{r}_2$ written in the coordinate form (see formula (4)).

Combining all the cases, we obtain the general formula for multiplying quaternions. If $q_1 = x_1 + \mathbf{r}_1$ and $q_2 = x_2 + \mathbf{r}_2$, then

$$q_1q_2 = (x_1x_2 - \mathbf{r}_1 \cdot \mathbf{r}_2) + (x_1\mathbf{r}_2 + x_2\mathbf{r}_1 + \mathbf{r}_1 \times \mathbf{r}_2).$$

CONTINUED ON PAGE 18

Fertilizer with a bang

Investigating an explosive situation

by B. Novozhilov

IN JANUARY 1953 THE *TYRrhenia*—a steamer with a displacement of four thousand tons—exploded and sank in the Red Sea. There have been greater catastrophes in the history of sailing—explosions and fires in oil tankers and ships carrying ammunition. But the *Tyrrhenia* was carrying ammonium nitrate—not an explosive material (at least at first glance). If this were the only instance of such a catastrophe, one would look elsewhere for the cause of the explosion. However, Lloyd's Navigation Register, which records every major marine catastrophe, contained dozens of references to fires and explosions involving ammonium nitrate. They occurred more frequently in the forties and fifties, when production of this substance greatly increased. The worst year was 1947, when three powerful explosions occurred in port as the substance was being loaded, resulting in a large number of casualties.

Ammonium nitrate (NH_4NO_3) is used extensively in industry and agriculture. Tens of millions of tons of this chemical are produced every year. The formula itself gives a big hint about how it's used. Ammonium nitrate contains a large amount of nitrogen, and it is of particular importance that the nitrogen is present in forms that plants absorb most easily: ammonium and nitrate ions. Because of this, ammonium nitrate is one of the most potent mineral fertilizers. In addition, the presence of a large amount of oxygen in ammonium nitrate leads to another (not necessarily peaceful) use of this substance as an oxidant in the production of gunpowder and explosives. Without the addition of organic ingredients, however, ammonium nitrate is not explosive. In fact, it's often used in chemical demonstrations. On the other hand, mixtures of NH_4NO_3 with trinitrotoluene or aluminum are very powerful explosives.

Let's try to figure out how an absolutely safe mineral fertilizer, ammonium nitrate, which is stored for years in warehouses or in the open air, can turn into the cause of so many marine tragedies.

Art by Sergey Ivanov

An explosion is characterized by a rapid release of a large amount of energy, accompanied by a drastic change in the state of the explosive. The gases formed by the explosion can perform a huge amount of work as they expand.

Of course, terms like "rapid" and "large" are relative. One kilogram of coal burned in the air yields about thirty million joules. This energy is sufficient to heat the combustion products to many thousands of degrees. For a gas of normal density such temperatures correspond to a pressure of hundreds of atmospheres. The coal–oxygen system has enough chemical energy to produce an explosion, but in reality coal doesn't explode—it burns peacefully.

Something else is needed for an explosion to occur: the energy must be released rapidly. Indeed, coal begins to burn on its surface, so the deeper layers of coal start to burn only after they are heated by the energy released in the surface layers. Both the heating and the transfer of oxygen proceed so slowly that the combustion products have plenty of time to expand, so this reaction is not accompanied by a significant increase in pressure.

However, there are substances that need no external oxidant for their chemical transformation. Examples are solid rocket propellants and dynamite.

One such substance is trinitrotoluene $C_7H_5(NO_2)_3$, also known as tolit. If tolit is lit at the surface, it catches fire easily and burns quiescently. As it burns the inner part of the tolit slowly heats up. And despite the fact that the heat of reaction of this substance is far less than that of an ordinary fuel (it's only 4 MJ/kg), tolit can explode. To make it explode, you just need to increase the temperature of the entire volume of the tolit. You can do this in a number of ways—for example, by compression. In this case the reaction occurs throughout the entire volume and proceeds so quickly that the gases produced have no time to expand during the reaction. Occupying a volume approximately equal to that of the tolit before it was detonated, and heated to a high temperature, these gases create a pressure of tens or hundreds of thousands of atmospheres. The subsequent rapid expansion of the gases destroys whatever is nearby—a classic explosion.

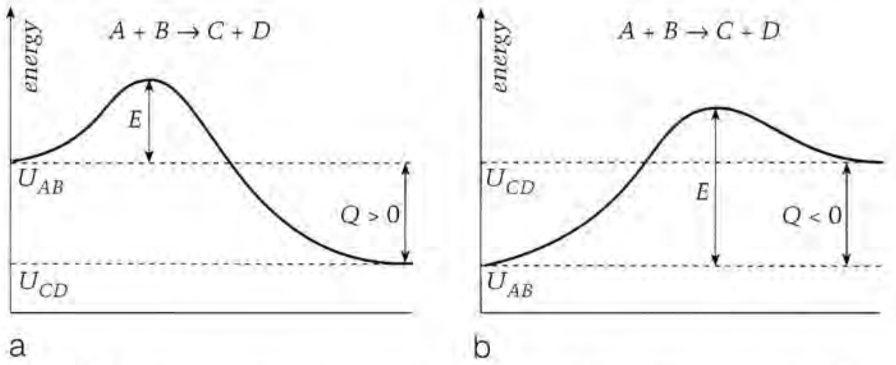


Figure 1. *Activation energy E and heat of reaction Q for the reaction $A + B \rightarrow C + D$ for (a) exothermic and (b) endothermic reactions.*

This article will examine the notion of a thermal explosion—a spontaneous explosive process linked to the release of chemical energy.

First, let's recall some basic facts about the rates of chemical reactions.

The first thing we notice about most chemical reactions is their slowness. For example, in a gas under normal conditions, every molecule participates in about 10^{10} collisions per second. If every collision resulted in a chemical conversion, the corresponding reaction would occur almost instantaneously. However, chemical experiments contradict this scenario. For example, a mixture of hydrogen and oxygen (oxyhydrogen gas, also called "detonating gas") can be stored at room temperature quite safely. This is because only a tiny fraction of all the collisions is efficient and results in a chemical reaction.

The basic chemical process is the conversion of one type of molecule into another. In this process some interatomic bonds are destroyed, while others are created. If the new bonds are stronger than the old ones, the reaction releases energy (it's an exothermic process). This doesn't mean, however, that the reaction must proceed rapidly. The fact is, before new (and stronger) bonds are formed, the old ones must be destroyed, which requires energy.

The durability of any composite system (atom, molecule, or nucleus) is characterized by its binding energy. This is equal to the work that must be performed to break the system down into its component parts. For example, the binding energy of the hydrogen molecule is $7.2 \cdot 10^{-19}$ J. This is the energy that must be expended to obtain two individual hydrogen atoms. Correspondingly, the same amount of energy will be released when a hydrogen molecule is formed from two hydrogen atoms.

The difference between the binding energies of the products and the initial substances in a chemical reaction is called the *heat of reaction* (because the internal molecular energy is usually converted to heat in the course of the chemical reaction). Recomputed per unit mass or volume of a substance, the heat of reaction is the *heat of combustion*.

If a reaction is accompanied by the release of energy, this doesn't necessarily mean it must proceed quickly. This is illustrated in figure 1. In the reaction $A + B \rightarrow C + D$ the bonds in molecules A and B must be destroyed or at least weakened. To meet this requirement, a certain amount of energy E (a definite amount for any particular reaction) must be supplied to the system. This is called the *activation energy* of the reaction. Clearly the activation energy must be of the order of magnitude of the sum of the binding energies of molecules A and B. Therefore, a chemical process runs

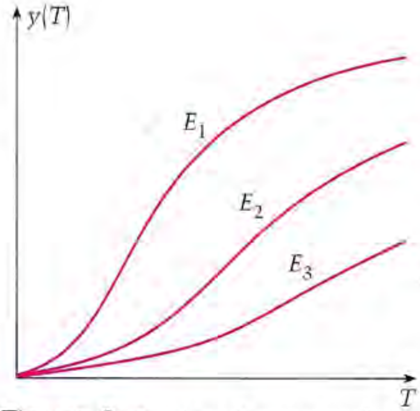


Figure 2. *Portion of active molecules at various energies of activation ($E_1 < E_2 < E_3$) (dependent on temperature).*

into an energy barrier at the very first stage. The molecules participating in the reaction must have sufficient thermal energy to overcome a hurdle of height E.

The chemical binds could be destroyed or weakened by the thermal (translational or oscillatory) motion of the atoms and molecules. Nature is so constructed that the energy of thermal motion at moderate temperatures is usually far less than the activation energy. The mean energy of thermal motion is of the order of kT—that is, $\sim 4 \cdot 10^{-21}$ J; while the activation energy is hundreds of times greater. This explains the negligible number of efficient collisions. The chemical reaction proceeds only with those molecules whose thermal energy is far greater than the mean energy. Such molecules are extremely rare.

The number of molecules that have thermal energy E at temperature T is given by a very simple formula:

$$y = e^{-E/kT}.$$

(The reader will have to take this formula on faith—a more or less strict deduction would be a large digression.) Figure 2 shows qualitatively the dependence of y on temperature T for various values for the activation energy E.

Thus the rate of a chemical reaction—that is, the amount of a substance (in kilograms, moles, or particles) that reacts in a unit volume per unit time—can be written as

$$w = ze^{-E/kT},$$

where z is determined either by the number of collisions (in gases) or by the number of molecular bonds that must be broken to make the chemical conversion possible in condensed bodies. The dimensions of w and z are the same: kg/(m$^3 \cdot$ s), mole/(m$^3 \cdot$ s) or 1/(m$^3 \cdot$ s). To get a feel for how much the rate of a chemical reaction depends on temperature, take a look at a specific numeric example—problem 1 at the end of the article.

Now we're ready to explain the phenomenon of thermal explosions. The explanation of thermal self-ignition is very simple. If a reaction is accompanied by heat release, the substance itself is heated. As we saw, heating increases the rate of a chemical reaction. This in turn leads to a more intense release of energy. The possibility of a progressively self-accelerating chemical reaction was pointed out at the end of the 19th century by the great Dutch chemist Jacobus Henricus van't Hoff. In 1928 Nikolay Semyonov provided a quantitative theory of thermal explosions that can justifiably be called classical. The theory is easy enough to explain, yet it remains the basis for investigating and solving new problems in physical chemistry—the science of the physical foundations of the chemical process—a century after it was formulated.

Now, if we jump to the conclusion that any mixture in which an exothermic reaction occurs will eventually explode, we'd be wrong. After all, the vessel in which the reaction occurs is limited, and the heat that is released can be absorbed by the surroundings. The competition between heat production and heat dissipation leads to very peculiar behavior in chemically self-intensifying systems.

Let's turn to the quantitative aspects of this phenomenon. Imagine a reaction that proceeds with a heat of combustion Q (J/kg) in some volume V. The amount of heat released in this volume per second is

$$P_+ = zQVe^{-E/kT} \text{ (J/sec)},$$

where T is the temperature of the reacting substance in volume V.

The temperature of the surrounding medium T_0 is assumed to be constant. The heat losses from the reacting volume are proportional to the difference $T - T_0$ and to the area S of the surface that defines the volume V. Therefore, every second the reacting substance transfers the following amount of heat to the surroundings:

$$P_- = \alpha S(T - T_0),$$

where α is the coefficient of heat transfer. This value naturally depends on the capacity of the reacting substance to transfer heat—the coefficient of heat transfer λ—and on the size of the object r (for a sphere, for example, this would be its radius). Thus

$$\alpha = K\lambda/r.$$

The coefficient K is a constant describing the effect of the body's shape on the process of heat transfer. It's clear why the parameter r is in the denominator: given the same temperature difference between two points of a body, the heat flow (that is, the amount of heat passing through a unit area per unit time) decreases as the distance between these points increases.

Figure 3 shows the heat production function $P_+(T)$ and a number of heat transfer lines $P_-(T)$ correspond-

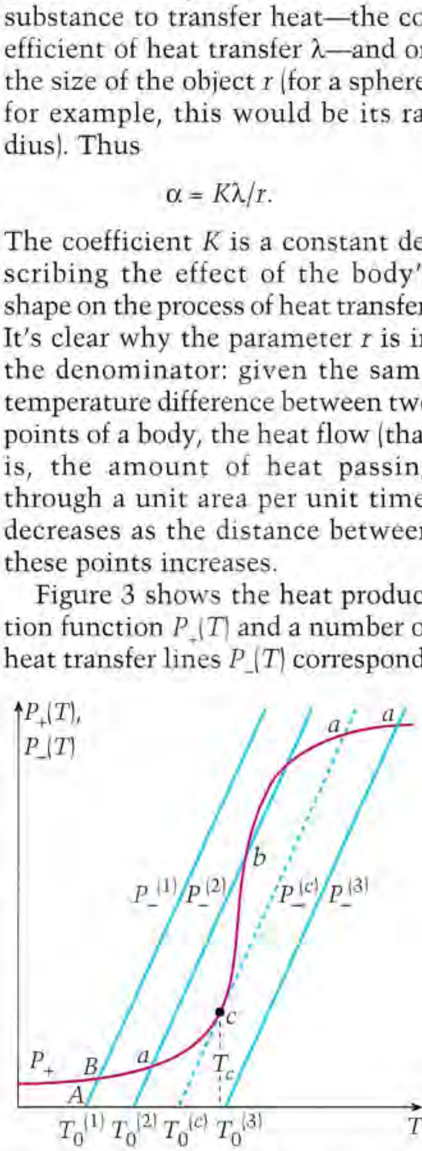


Figure 3. *Dependence of heat release* P_+ *and heat transfer* P_- *on temperature.*

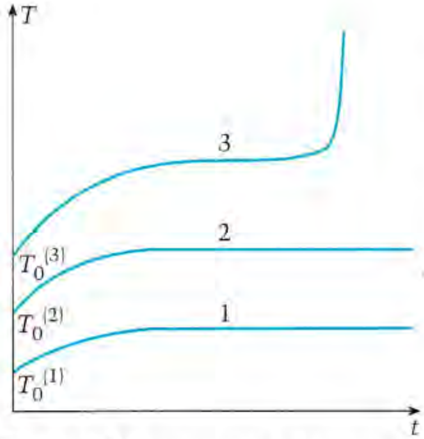


Figure 4. *Temperature changes in subcritical (curves 1 and 2) and supercritical (curve 3) modes of heating.*

ing to different temperatures of the surroundings. Pay particular attention to the number of points where the graphs of P_+ and P_- intersect—that is, the number of roots of the equation

$$P_+ = P_- . \qquad (*)$$

Because of the strong nonlinearity of the heat production function, the number of intersection points can be more than one. When the temperature T_0 is high enough or low enough ($T_0^{(1)}$ and $T_0^{(3)}$), equation (*) has only one root. However, at some intermediate temperature $T_0^{(2)}$ the graphs can intersect at three points. Equation (*) describes thermal equilibrium in which all heat generated is dissipated in the surroundings. It may easily be deduced that intersection points like a, where the heat removal graph is steeper than the heat release graph, are stable, while points like b are unstable.

Now let's consider the behavior of the reacting system over time. At the initial moment $t = 0$ the substance has a temperature $T_0^{(1)}$ equal to the temperature of the surroundings (point A in figure 3). Since at this point $P_+ > P_-$, the system will be heated. In figure 4 this process corresponds to a shift of the system along the curve 1. We see that the increase in temperature is slowing with time. It slows because the difference $P_+ - P_-$ decreases as the system heats up, and this difference determines the temperature increase. Heating will stop at point B (figure 3), which is stable.

A similar dependence of temperature on time (curve 2 in figure 4) will hold for a higher initial temperature $T_0^{(2)}$. In this case the system will be heated to some higher final temperature and will end up in a stable state.

By increasing temperature of the surrounding medium (and, correspondingly, the initial temperature of the reagents), we can arrange a qualitatively different mode of heating. Let's analyze how heating will proceed at an initial temperature $T_0^{(3)}$ (curve 3 in figure 4). At the beginning of the process, the dependence $T(t)$ is similar to that in both cases considered above: the difference $P_+ - P_-$ decreases over time, and therefore the increase in temperature is slowed. The system needs a lot of time to pass through the narrow gap between the heat production and heat dissipation curves—this corresponds to an almost horizontal portion of the temperature dependence $T(t)$. After the difference $P_+ - P_-$ has passed this minimum, the temperature increases sharply, and this increase is steeper than the exponential law. A self-accelerating chemical process of energy release has begun. And that is what a *thermal explosion* is: the thermal self-acceleration of an exothermic chemical reaction.

The most remarkable feature of this phenomenon is the existence of a critical condition. We see that a gradual increase in temperature T_0 produces an abrupt qualitative change in the system's behavior. The temperature dependence $T(t)$ can be only one of two types—either a weak heating or an abrupt heating after some "calm" (or latent) period—in other words, an explosion.

It's clear that there must be some critical temperature $T_0^{(c)}$ that separates these two radically different types of reactions. No doubt you've noticed the dashed line $P^{(c)}$ in figure 3, which corresponds to a temperature $T_0^{(c)}$. The value $T_0^{(c)}$ is the critical initial temperature we're looking for. The curves P_+ and $P_-^{(c)}$ have a common tangent point c. If the initial temperature is below $T_0^{(c)}$, slow heating occurs. A slight increase in T_0 above $T_0^{(c)}$ leads to an explosion.

Before we analyze the conditions for an explosion, note that changing the initial temperature isn't the only way to produce the critical state in a system. The functions P_+ and P_- depend on other parameters, and changing those parameters can move the system from the subcritical to the supercritical state (problem 3 explores this possibility).

Now let's obtain the quantitative criterion for an explosion. It should look like the inequality $T_0 > T_0^{(c)}$, which includes all the parameters of the problem. We want to obtain the condition for the tangency of the two curves representing the functions $P_+(T)$ and $P_-(T)$. This is a good problem for those who are familiar with differential calculus. If you don't know this mathematical technique, you should take the result for granted and try to understand its physical meaning.

For the graphs of functions to touch, two conditions must be met: first, the values of the functions must be equal at this point; and second, the values of the first derivatives at this point must be the same.

These two conditions determine the critical temperature—that is, the temperature of the substance corresponding to the tangent point of the graphs $P_+(T)$ and $P_-(T)$:

$$T_c = T_0 + \frac{kT_0^2}{E}$$

(the value of T_0 in this formula is equal to $T_0^{(c)}$).

This value of T_c gives the criterion for an explosion:

[1]Here's a hint for those who wish to deduce this formula and obtain the value of T_c and the criterion for thermal explosion. As we said, $kT/E \ll 1$, so in calculating T_c we can replace $\sqrt{1 - 4kT_0/E}$ with its approximation $1 - 2kT_0/E - 2(kT_0/E)^2$. Similarly, in deducing the criterion for explosion, we replace the expression $1/(1 + kT_0/E)$ with its approximation $1 - kT_0/E$.

$$\frac{zQV}{\alpha S}\frac{E}{kT_0^2}e^{-E/kT_0} \geq \frac{1}{e} \qquad (**)$$

Plugging T_c into (**) gives the condition for the critical mode; the sign ">" corresponds to an explosion.[1]

So, if we know the shape and the volume occupied by a substance with known physical and chemical properties, we can predict whether it will explode.

Now we can return to the exploding ships that were loaded with ammonium nitrate. As noted above, heat is produced when this substance breaks down. Since the rate of the chemical reaction increases with temperature, a thermal explosion of ammonium nitrate is possible. However, a possibility doesn't mean that an event will actually occur. Only a numerical estimate can tell us whether a thermal explosion took place or if we need to look for a reason elsewhere.

Let's write formula (**) another way. Clearly the ratio $V/\alpha S$ is proportional to the square of the linear size of the object—that is, r^2. Therefore, the criterion for a thermal explosion can be written as a condition determining the critical size r_c of an object:

$$r_c = \left(\frac{C\lambda kT_0^2}{zQE}e^{E/kT_0}\right)^{1/2}.$$

(The constant C characterizes the effect of the object's shape on the chemical reaction). If $r > r_c$, an explosion will occur.

The values of the parameters in the formula for r_c are known to different degrees of accuracy. The heat of combustion of ammonium nitrate is 1 MJ/kg, and its coefficient of heat transfer is $\lambda = 0.17$ J/(m · s · K). By contrast, the exact values of the kinetic constants z and E for ammonium nitrate are not known. Due to limitations in our current understanding of chemical conversions, it is not possible to find these constants theoretically. We can measure them, but the procedures are difficult and fraught with experimental error. So the tables of kinetic constants only give ranges for these parameters: $z = (3–5) \cdot 10^{16}$ kg/(m^3 · s) and $E = (2.4–2.7) \cdot 10^{-23}$ J.

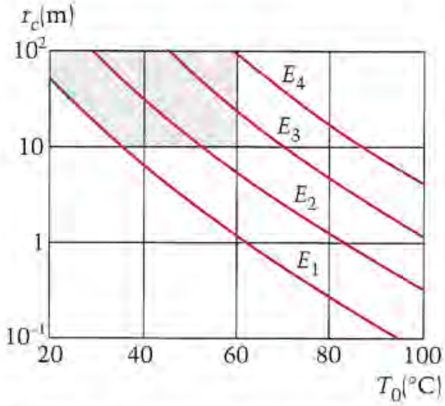


Figure 5. *Dependence of the critical size r_c on the temperature T_0 of ammonium nitrate for various values of the activation energy:*
$E_1 = 2.4 \cdot 10^{-23}$ J; $E_2 = 2.5 \cdot 10^{-23}$ J; $E_3 = 2.6 \cdot 10^{-23}$ J; $E_4 = 2.7 \cdot 10^{-23}$ J.

Thus the problem isn't calculating the critical size r_c of an object, it's determining the range of values of r_c that correspond to the ranges of z and E given above.

The critical size is most strongly affected by the activation energy and the temperature of the surrounding medium, because they are located in the index of the exponential function. The contribution of the other parameters is relatively insignificant. Let's plot the dependence for various values of the activation energy (figure 5). The region above this curve (large r and high temperature) corresponds to an explosion. The critical size drops as the activation energy decreases.

This plot shows that the characteristic sizes of the compartments in the ships (of the order of dozens of meters) corresponds to the region of r_c for reasonable estimates of activation energy and initial temperature. As noted above, the series of explosions accompanied the beginning of large-scale production of ammonium nitrate. The process was accelerated, and it included cooling the ammonium nitrate obtained from neutralizing nitric acid with ammonia. It is quite possible that the product was still warm when it was loaded on the ships. In addition, to prevent caking, the granules are covered with wax and packed in paper bags. Experimental studies show that adding organic substances lowers the activation energy. All these factors make an explosion more likely.

And here we'll stop with the hope that our readers have felt the power and practical value of theory, for there's nothing more useful than a good theory.

To conclude, we note that the methods developed in the theory of thermal explosions are widely applied in studying similar phenomena in other fields of science. Here are a few processes where nonlinear (with respect to temperature) heat sources play a key role: thermonuclear reactions; thermal breakdown in dielectrics; and critical phenomena accompanying the motion of viscous liquids (hydrodynamic thermal explosion). All of which shows the power and vigor of the theory of thermal explosions.

Problems

1. The rate of the gas-phase reaction of hydrogen with iodine is given by the equation

$$w = \kappa n_{H_2} n_{I_2} e^{-E/kT},$$

so

$$z = \kappa n_{H_2} n_{I_2},$$

where $\kappa = 10^{-10}$ cm^3/s is some characteristic of the reaction and $E = 2.69 \cdot 10^{-19}$ J. Assuming the concentrations of the initial substances to be

$$n_{H_2} = n_{I_2} = n = 1.35 \cdot 10^{19} / \text{cm}^3$$

(under normal conditions), find how long it will take 1% of the mixture to react at the following temperatures: (a) 273 K; (b) 600 K; (c) 800 K.

2. Prove that the intersection points of the graphs of heat production and heat transfer correspond to stable states if the curve for heat transfer is steeper than the curve for heat production, and vice versa.

3. Plot the graphs of the transition from subcritical to supercritical conditions as dependent on (a) the heat of combustion and (b) the coefficient of heat transfer α.

ANSWERS, HINTS & SOLUTIONS ON PAGE 51

HOW DO YOU FIGURE?

Challenges

Physics

P301

Suspended cord. A heavy cord is suspended at points A and B (figure 1). The cord's tension at point C is 20 N. Find the mass of the cord. (I. Slobodetsky)

P302

Isothermal and adiabatic curves. Transform an ideal gas from state 1 with temperature T_1 to state 2 with temperature $T_2 > T_1$ in such a way that, during the entire reversible process $1 \rightarrow 2$, the temperature doesn't fall and heat isn't lost from the gas. The minimum heat transferred to the gas in such a process is Q_1. What is the maximum heat that can be added to the gas under these conditions? (O. Shvedov)

P303

Welding in water. Sometimes a mixture of oxygen and hydrogen, obtained by electrolysis of water, is used for small welding jobs. Find the efficiency of electrolysis cell if the voltage drop between the electrodes in the cell is U = 2V. Take into account the fact that heat Q = 0.29 MJ is released by burning m = 2 g of hydrogen in oxygen. (V. Pogozhev)

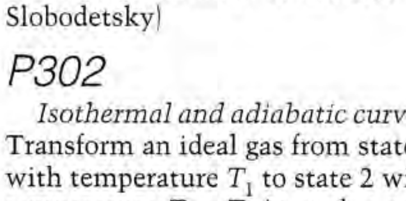


Figure 1

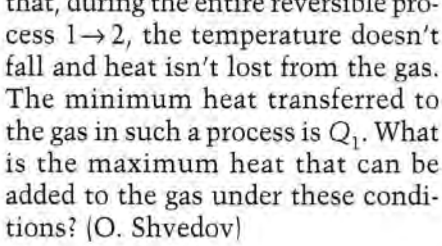


Figure 2

P304

Electricity on the rails. Parallel rails of length $2L$ and resistance per unit length ρ are fixed on the horizontal plane at a distance l from each other. Their ends are connected to identical batteries with emf = $\mathscr{E}$ (figure 2). A jumper of mass m and resistance R lies on the rails and can slide along them. The entire system is placed in a vertical homogeneous magnetic field B. Find the period of the small oscillations that result from moving the jumper from the equilibrium position. Neglect the damping of the oscillations, internal resistance of the batteries, resistance of the contacts, and inductance of the circuit. (A. Yakuta)

P305

A striking problem. A light beam traveling in the plane normal to the axis of a glass semicylinder strikes its flat surface at an angle of 45°. From what part of the curved surface of the semicylinder can the beam emerge? The refractive index of glass is n.

Math

M301

Making progress. The number N is divided into the number 201. The quotient, the remainder, and the divisor (i.e. N itself) form a geometric progression, taken in some order. Find all possible values of N.

M302

Tread lightly. It's known that the rear tires on a car wear out after 42,000 km, while the front tires wear out after 58,000 km. What is the maximum distance a car can travel if it is given four new tires plus a new spare? (All the tires are identical.)

M303

Check your calculations. A two-round chess tournament was held. In each round, every participant played a game with every other participant. The players earned points for each game: one point for a victory, 0.5 for a draw, and 0 for a defeat. Is it possible that, when the tournament was over, the participants' rankings were the opposite of those they held after the first round (that is, the participant who was last after the first round was the champion; the participant who was next to last after the first round ended up in second place, and so on). Solve this problem for (a) 9 participants; (b) 10 participants.

M304

Circumscribed circle. In triangle ABC the altitude CM is drawn. The line that is symmetric to the altitude drawn from vertex A about line CM intersects line BC at point K. Find angle OMK, where O is the center of the circle circumscribed about triangle ABC (the points O, M, and K are all distinct).

CONTINUED ON PAGE 27

Shape numbers

Exploring a Fermat hypothesis

by A. Savin

THE PROMINENT FRENCH mathematician Pierre Fermat is famous not only for the theorems he proved but for the hypotheses he proposed. Some of them turned to be false—for example, the hypothesis that all numbers

$$2^{2^n}+1$$

are primes. Another hypothesis, called Fermat's last theorem, was proved at the very end of the 20th century.

In this article we'll look at another Fermat hypothesis:

Any natural number can be represented as a sum of not more than three triangular numbers, not more than four square numbers, not more than five pentagonal numbers, and so on.

This hypothesis was proved by Augustin-Louis Cauchy some 300 years after it had been proposed. Before this, the hypothesis was proved for triangular numbers by Carl Friedrich Gauss; for square numbers it was proved by Joseph-Louis Lagrange. It seems that square numbers are of particular interest to mathematicians. In 1738 Leonard Euler established all cases when a number can be represented as a sum of two squares:

A natural number can be represented as the sum of two squares if and only if each of its prime factors of the form $4k + 3$ *occurs as an even power.*

In 1798 Adrien-Marie Legendre proved that numbers of the form $4^k(8n - 1)$ cannot be represented as the sum of less than four squares.

It might seem that everything in this field is clear now, and that the hypothesis is of historical interest only. But all we need to do is add the word "different" to Fermat's hypothesis to obtain another difficult problem:

Which natural numbers can be represented as the sum of not more than three different triangular numbers, not more than four different square numbers, and so on?

It's clear that not every number can be represented in this form—for example, the number 2.

Before considering this problem, we should recall the notion of "shape numbers" (or "number shapes"). If you have a number of similar things, such as coins, stones, or buttons, you sometimes feel like arranging them in a certain order—as a triangle, rectangle, or some other geometric figure.

In ancient Greece, mathematicians did their calculations on an abacus, repositioning stones on a set of rods. Nowadays, programmers play computer games when they want to take a break, but in the old days the only "professional" entertainment was to build number shapes. This pastime led to some serious results. For example, building rectangles from stones suggested the idea of divisors as well as prime and composite numbers.

Arranging stones in the form of triangles (figure 1), we obtain the triangular numbers 1, 3, 6, 10, ... The general formula for the nth triangular number is written as the sum of the arithmetic series

Figure 1

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

For square numbers (figure 2), we

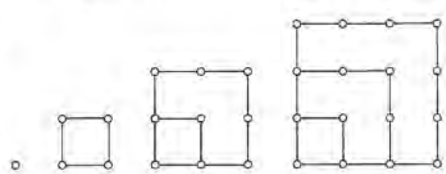
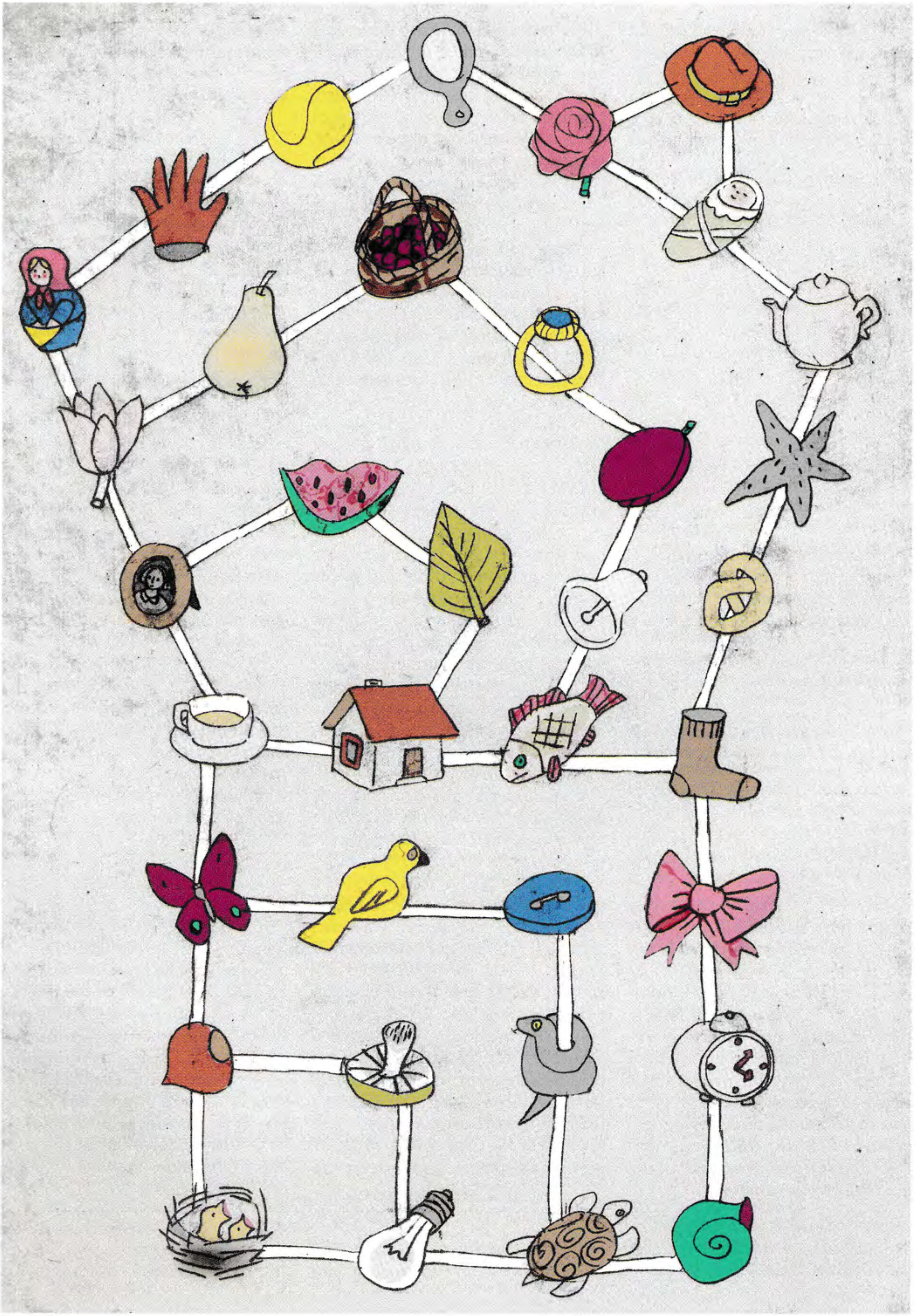


Figure 2

Art by Vera Khlebnikova



have 1, 4, 9, 16, … , and the general formula is very simple:

$$s_n^4 = n^2.$$

For the pentagonal numbers 1, 5, 12, 22, … (figure 3), the formula can

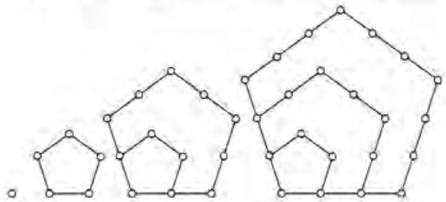


Figure 3

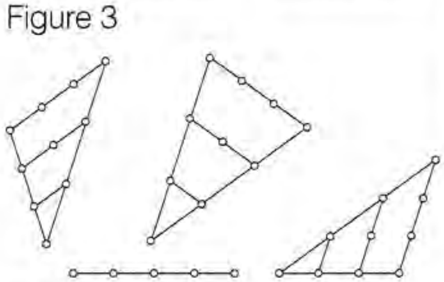


Figure 4

be obtained by examining figure 4. It's sufficient to add three $(n-1)$th triangular numbers and one "copy" of n. Thus we get

$$s_n^5 = n + 3\cdot\frac{(n-1)n}{2} = \frac{n(3n-1)}{2}.$$

Similarly, we can obtain a formula for the nth k-gonal number:

$$s_n^k = n + (k-2)\frac{(n-1)n}{2} = \frac{n(nk-2n-k+4)}{2}.$$

Now let's get to work on our problem. We'll start with the part about breaking the number down into the sum of squares. We can check directly that the small numbers 2, 3, 6, 7, and 8 cannot be represented as a sum of squares of different natural numbers. It's not hard to write a computer program for finding such numbers. The program shows that among numbers not exceeding 1,000, the following numbers cannot be represented as the sum of different squares:

2, 3, 6, 7, 8, 11, 12, 15, 18, 19, 22, 23, 24, 27, 28, 31, 32, 33, 43, 44, 47, 48, 60, 67, 72, 76, 92, 96, 108, 112, 128.

All other numbers not exceeding 1,000 can be broken down into the sum of different squares. What about larger numbers? Checking numbers up to 10,000 doesn't produce any more negative results. Maybe there aren't any more? This turns out to be the case. It's sufficient to prove the following theorem.

Theorem 1. *If, for a given number k and a particular function $f(k)$, all natural numbers between k and $f(k)$ can be represented as the sum of squares of different natural numbers, then k and all natural numbers after it can be represented in this form.*

We'll determine the function $f(k)$ in the course of proving this theorem.

Proof. Let k be the given number, and let p be an integer such that $p^2 > k$. Let's suppose that the numbers from k to $(p+1)^2$ can be represented as the sum of different squares. Then the numbers $p^2 + k$, $p^2 + k + 1$, ..., $2p^2 - 1$ can also be represented in this form. Indeed the representation of the numbers k, $k + 1$, ..., $p^2 - 1$ cannot include p^2. In the same way, we can show that the numbers $(p+1)^2 + k$, $(p+1)^2 + k + 1$, ..., $2(p+1)^2 - 1$ can be represented as the sum of different squares as well.

Consider the position of these numbers on the number line (figure 5). We know that the numbers in the intervals from k to $(p+1)^2$, from $(p^2 + k)$ to $(2p^2 - 1)$, and from $(p+1)^2 + k$ to $2(p+1)^2 - 1$ can be represented as a sum of different squares. Does this mean that all the numbers from k to $2(p+1)^2 - 1$ can be represented as a sum of different squares? That depends on how the numbers that bound our intervals are arranged on a number line. Figure 5 shows one possible arrangement. If the three intervals we are studying overlap, then the answer is yes. But suppose there are intervals containing integers that are skipped, as figure 5 shows? We can assure ourselves that the skipped intervals do not contain integers if we make sure that the following inequalities hold:

$$(p+1)^2 \geq p^2 + k - 1, \qquad (1)$$

$$2p^2 - 1 \geq (p+1)^2 + k - 1. \qquad (2)$$

The first inequality is equivalent to the inequality

$$p \geq k/2 - 1, \qquad (1')$$

and the second is equivalent (for $p > 0$) to the inequality

$$p + 1 \geq \sqrt{k+3}. \qquad (2')$$

For $k \geq 13$, inequality (1') implies (2'), since in this case $k/2 - 1 \geq 1 + \sqrt{k+3}$.

We formulate this result as the following lemma.

Lemma. *Let an integer number $k > 0$ and an integer p satisfying inequality (1') be given. If all integers from k through $(p+1)^2$ can be represented as the sum of different squares, then all integers from k through $2(p+1)^2 - 1$ can be represented in this form as well.*

Now we can complete the proof of theorem 1 by explicitly specifying the function $f(k)$. Since we assumed that all integers from k to $(p+1)^2$ can be represented as the sum of different squares, and that inequality (1'), which can be written as $p + 1 \geq k/2$, was satisfied, we get

$$(p+1)^2 \geq k^2/4.$$

Therefore, we can take $k^2/4$ as the function $f(k)$.

We complete the proof of our theorem with a peculiar form of induction, in which we keep extending the upper bound of the interval on which our proposition is true.

Let k be a positive even integer, and suppose all the numbers from k to $k^2/4$ can be represented as the sum of squares of different natural numbers. Consider the number $p = k/2 - 1$. By the lemma, all the numbers in the range k to $k^2/4 = 2(p+1)^2$

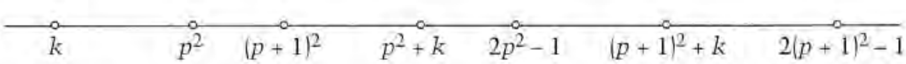


Figure 5

-1 can be represented in this form as well. Let's increase p by 1. A little algebra will show that

$$(p+2)^2 < 2(p+1)^2 - 1$$

for all $p > \sqrt{3}$. The values of p we are interested in are certainly in this range, so our lemma tells us that all the integers from k to $(p+2)^2$ can be represented as the sum of squares of different natural numbers. In addition, some more algebra will show that the inequalities resulting by substituting $p_1 = p + 1$ for p in the inequalities (1') and (2') are satisfied. Therefore, according to the lemma, all numbers from k to

$$\begin{aligned} 2(p_1+1)^2 - 1 &= 2(p+2)^2 - 1 \\ &= 2(p+1)^2 - 1 + 4(p+1) \end{aligned}$$

are representable as the sum of squares of different natural numbers. Thus the interval containing integers that can be represented as the sum of different squares is increased by $4(p+1) = 2k$. The same reasoning makes it possible to pass from p_1 to $p_2 = p_1 + 1 = p + 2$ and increase again the interval where all integers can be represented as the sum of different squares. This process increases the length of the interval in question by 4. Indeed,

$$4(p_1+1) = 4(p+2) = 4(p+1) + 4.$$

Now it's clear that taking $p_3 = p_2 + 1$, then $p_4 = p_3 + 1$, and so on, we prove that any integer greater than k can be represented as the sum of different squares.

It remains to show that such a number k exists. This job was done by a computer that calculated that all integers in the range from 130 to $130^2/4 = 4{,}225$ can be represented as the sum of different squares.

Readers who don't trust computers (and there may be a few!) can do the computations by hand.

We have proved that all integers beginning with 129 can be represented as the sum of different squares. However, we don't yet know the minimum number of different squares in the representation of each of these numbers. To analyze as many numbers as possible and get a sense of the answer to this question, we wrote the following program in BASIC.

```
N=0
INPUT "INPUT N" N
R=INT(SQR(N))
DIM S(N),T(N)
FOR A=1 TO N
S(A)=100
NEXT A
K=1
B0:
 P=K*K
 Q=P
B1:
 M=Q-P
 T(Q)=S(M)+1
 Q=Q+1
 IF Q<=N GOTO B1
 M=P
B2:
 IF T(M)>S(M) GOTO B3
 S(M)=T(M)
B3:
 M=M+1
 IF M<=N GOTO B2
 K=K+1
 IF K<=R GOTO B0
  FOR L=100 TO 6 STEP-1
  FOR A=1 TO N
   IF S(A)=L THEN PRINT "L=" L; "M=" A;
  NEXT A
 NEXT L
END
```

To use this program, you should input a number N, which serves as the upper limit for the computations. For every number $L \le 100$, the program prints all integers less than or equal to N for which the minimum number of squares in the representation is L. Numbers that cannot be represented as the sum of squares are printed on the row $L = 100$. These numbers have already been listed above.

Only the numbers 124 and 188 are represented as the sum of no fewer than 6 different squares. The following numbers are represented as the sum of 5 different squares:

55, 88, 103, 132, 172, 176, 192, 240, 268, 288, 304, 368, 384, 432, 448, 496, 512, 752.

All numbers obtained by multiplying the above numbers by a power of 4 possess the same property. These are

220, 352, 412, 528, 688, 704, 768, 880, and so on.

No other numbers less than 100,000 have been discovered that are represented as the sum of no fewer than 5 different squares.

Now we prove the following theorem.

Theorem 2. *If an even number cannot be represented as the sum of fewer than 5 different squares, this number multiplied by 4 possesses the same property.*

Proof. Notice that the remainder upon division of the square of an even number by 4 is zero, while for odd numbers the remainder is 1. Let a be an even number that cannot be represented as the sum of fewer than 5 different squares. Assume that $4a$ can be represented as the sum of 2, 3, or 4 different squares. Note that $4a$ cannot be a perfect square, since in this case a would be a perfect square as well.

Assume that $4a = b^2 + c^2$. Consider three cases. If both b and c are even—that is, $b = 2p$ and $c = 2q$—we have

$$4a = 4p^2 + 4q^2,$$

and, therefore, $a = p^2 + q^2$, contrary to our assumption. If both a and b are odd, the number on the righthand side of the equality $4a = a^2 + b^2$ gives a remainder 2 upon division by 4, and the lefthand side is divisible by 4, which is a contradiction. If one of the numbers a and b is even and the other is odd, the righthand side of this equality is odd, and its lefthand side is even, which is a contradiction as well.

The case $4a = b^2 + c^2 + d^2$ can be considered in a similar way. If all the numbers a, b, and c are even, then a can be represented as the sum of three different squares. In the other cases, the remainder upon division of the righthand side by 4 is either 1, 2, or 3, whereas the lefthand side is divisible by 4.

The case $4a = b^2 + c^2 + d^2 + e^2$ is quite similar except for the case when all the numbers a, b, c, and d

are odd. In this case $b = 2p + 1$, $c = 2q + 1$, $d = 2r + 1$, and $e = 2s + 1$. Substitute these expressions in the formula $4a = a^2 + b^2 + c^2 + d^2$ to obtain

$$4a = 4p^2 + 4p + 1 + 4q^2 + 4q + 1 + 4r^2 + 4r + 1 + 4s^2 + 4s + 1.$$

Dividing both sides by 4, we have

$$a = p(p + 1) + q(q + 1) + r(r + 1) + s(s + 1) + 1.$$

Every term among the first 4 terms on the righthand side is even; therefore, the righthand side is odd. However, the lefthand side is even, according to our assumption. This completes the proof.

Among the 18 numbers (listed above) that cannot be represented as the sum of fewer than 5 different squares, 16 numbers are even. Therefore, the numbers obtained from them by multiplying by a power of 4 possess the same property by theorem 2.

To prove this fact for the remaining two odd numbers 55 and 103, it's sufficient to show that the even numbers $220 = 4 \cdot 55$ and $412 = 4 \cdot 103$ cannot be represented as the sum of fewer than 5 different squares. This fact can be verified by direct calculations.

Examination of the set of numbers not exceeding 100,000 produced some interesting results. It turned out that the minimum number of different squares in the representation of numbers of the form $8n - 2$ (for $n > 13$) is three; for numbers of the form $8n - 5$ (for $n > 80$), it is also three; and for numbers of the form $4^m(8n - 1)$ (for $m \geq 0$ and $n \geq 14$), it is four.

It would be interesting to prove these theorems for numbers greater than 100,000 and to prove that there are no other numbers apart from the sequence specified above that cannot be represented as the sum of less than 5 different squares.

Consider other shape numbers. Some thought will show that theorem 1 can be proved for any shape number by choosing an appropriate function $f(k)$. It's also not difficult to modify our BASIC program to enable it to perform calculations for arbitrary k-gonal numbers. However, the results obtained are not as interesting as for square numbers.

For triangular numbers, only six numbers 2, 5, 8, 12, 23, and 33 cannot be represented as the sum of several different triangular numbers. There exists a single number

$$20 = 1 + 3 + 6 + 10$$

that can be represented as the sum of not fewer than 4 different triangular numbers.

We found 61 numbers that cannot be represented as the sum of several different pentagonal numbers. These are

2, 3, 4, 7, 8, 9, 10, 11, 14, 15, 16, 19, 20, 21, 24, 25, 26, 29, 30, 31, 32, 33, 37, 38, 42, 43, 44, 45, 46, 49, 50, 54, 55, 59, 60, 61, 65, 66, 67, 72, 77, 80, 81, 84, 89, 94, 95, 96, 100, 101, 102, 107, 112, 116, 124, 136, 137, 141, 142, 147, and 159.

Other numbers not exceeding 100,000 can be represented in this form. Only two of them, 241 and 206, can be represented as the sum of 6 different pentagonal numbers and cannot be represented as the sum of fewer of them.

Ambitious readers can extend our investigation to 6-, 7-, and more-gonal numbers.

I am grateful to Maria Kurkina and Archil Maisuradze, students at the Moscow Institute of Physics and Technology, for their help in writing the BASIC program and performing calculations.

CONTINUED FROM PAGE 7

What about triples?

Why wasn't Hamilton able to find a reasonable method for multiplying triples? It was not for lack of inventiveness or effort. We mentioned earlier that this problem cannot be solved for "three-dimensional" numbers. Indeed it has been proved that no multiplication operation for points in three-dimensional space exists that possesses the associative, commutative, and distributive (with respect to coordinate addition) properties and allows division by any nonzero element. Moreover, we now know all the cases where multiplication of this sort exists. The German mathematician F. G. Frobenius (1849–1917) proved that there exist only three such cases: for dimensions one (real numbers), two (complex numbers), and four (quaternions).

Further developments

Hamilton and others placed great hopes on quaternions. Quaternions were expected to bring rich results, even deeper than complex numbers. Indeed quaternions were used to obtain elegant formulas that describe a number of important physical phenomena. However, visions of further development of the algebraic and functional calculus of quaternions failed to materialize.

For quaternions, the fundamental theorem of algebra regarding the existence of roots of any polynomial with quaternion coefficients doesn't hold. On the other hand, for any quaternion, there exists a nonzero polynomial with quaternion coefficients, such that the given quaternion is a root.

Optimism gave way to skepticism. At the beginning of the 20th century, mathematicians lost interest in quaternions. However, some time later the need arose in theoretical physics to find a mathematical formalism for describing certain effects related to the so-called *spin* of elementary particles. Quaternions received recognition again when their role in constructing various geometric transformations of space important for quantum physics was understood. The geometric properties of quaternions is a separate topic, and a big one, to which we hope to devote a special article.

NSTA press™
NATIONAL SCIENCE TEACHERS ASSOCIATION

Building the Best Learning Environment...

with the Science by Design series

A hands-on approach to the physics of product design. Students apply concepts such as heat transfer, buoyancy, elasticity, or insulation to the development of everyday items. All materials necessary are inexpensive and easy to find.

PK152X
Members: $63.85 Non-Members: $71.82

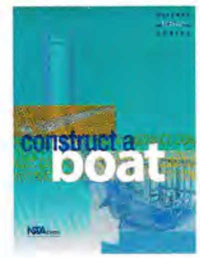
Buy any 3 in the series and get the 4th volume FREE!

Construct-a-Glove

(grades 9–12, 2000, 96 pp.)
Physics and technology go hand-in-hand in this practical demonstration of basic concepts in thermodynamics. By testing a simple prototype of an insulated glove, students learn about the process of homeothermic regulation and the variables that influence heat transfer. The challenge to improve upon their initial model introduces them to the design process and the relations between form and function.

#PB152X1
Members: $17.96 Non-Members: $19.95

Construct-a-Boat



(grades 9–12, 2000, 104 pp.)
By working with a simple model powered by a battery-driven fan, students get a feel for the forces involved in moving a boat through water. Students use their understanding of mass, buoyancy, friction, and acceleration to analyze parameters, improve performance, and design a faster boat. This immersion in learning-by-doing translates abstract concepts into tangible objectives and teaches students seaworthy lessons in modeling and design.

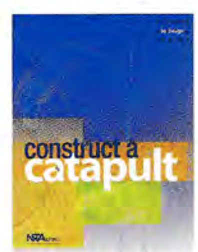
#PB152X2
Members: $17.96 Non-Members: $19.95

Construct-a-Greenhouse

(grades 9–12, 2000, 104 pp.)
Students design structures that will convert light to heat during germination and then reconfigure their greenhouses to promote photosynthesis. By comparing the effects of design characteristics on plant growth, students observe the connections between plant biology and basic principles in thermodynamics and energy transfer.

#PB152X3
Members: $17.96 Non-Members: $19.95

Construct-a-Catapult



(grades 9–12, 2000, 104 pp.)
Launch your students into a hands-on application of concepts such as torsion and elasticity as they learn the physics behind overcoming gravity and hurling objects through the air—SAFELY. By relating the effects of different force settings to distance projections, you can also introduce your students to the analysis of frequency distributions.

#PB152X4
Members: $17.96 Non-Members: $19.95

To order, call: 1-800-722-NSTA or visit www.nsta.org

The physics of walking

Why you're bound to run faster

by I. Urusovsky

CAN A PERSON BOUND HAND and foot run faster than an "untied" person? Yes! You don't believe it? Think about it...

From the viewpoint of oscillation theory, a person running in a reference system attached to the runner is just legs swinging. Therefore, when one runs or walks the legs play the role of pendulums attached to the body by hinges and oscillating in the gravitational field near the equilibrium position (the vertical position). However, a leg is not "a mass suspended by a weightless inelastic cord," and it's not the mathematical pendulum you've come across in your studies—the leg's mass is not concentrated at one point. The leg is a physical pendulum.

The most important characteristic of any pendulum is the frequency of its free oscillations. For a mathematical pendulum the angular frequency of free oscillation $\nu_0 = \sqrt{g/l}$ (the natural frequency) depends only on the length l of the pendulum. By contrast, the natural frequency of a physical pendulum depends not only on the length but also on the distribution of mass along the pendulum: the nearer the mass is to the axis of oscillation, the higher the frequency. However, even for a physical pendulum we can write the frequency as $\nu_0 = \sqrt{g/L}$, where L is a very important parameter of the pendulum: the "reduced length." This is the length of a mathematical pendulum that has the same frequency as the physical pendulum.

For proportionally built people the reduced length of their pendulum-leg is proportional to their height. Therefore, in different people, the natural oscillation frequency of their legs varies inversely as the square root of their height.

This feature yields a simple way to find ν_0 for a person of any height. You just need to measure the natural frequency of your own leg. It's not hard to do: stand on one leg and swing your other leg, relaxing your muscles in that leg. While the leg is swinging, count the mean number of oscillations per second. For example, for a person 175 cm tall, the frequency turns out to be 0.8 Hz. Therefore, for a person of height h (cm) we have

$$\nu_0 = 0.8\sqrt{175/h}\ \text{(Hz)}.$$

For a bent leg (as is the case when you run), the reduced length is smaller, while the natural frequency is higher by about one quarter.

When you walk, your legs perform forced oscillations under the action of your muscles. The characteristic feature of forced oscillations is that their amplitude depends not only on the size of the force but also on the frequency with which it varies. The frequency producing the largest amplitude of forced oscillations is called the resonant frequency.

The resonant frequency almost coincides with the natural frequency:

$$\nu_r = \nu_0\sqrt{1-\alpha^2},$$

where

$$\alpha = \frac{1}{2\pi}\ln n \approx 0.37\log n,$$

and n is a number showing by how much the amplitude of free oscillations decreases during one period. The amplitude of free oscillations of a human leg decreases about twofold during one period, so in the present case $n \cong 2$ and $\alpha \cong 0.11$. Thus the difference between natural and resonance frequencies is only 0.6%.

From the viewpoint of energy savings, the best way of walking is to move the legs at the resonant fre-

Art by Pavel Chernusky

quency. At this rate of walking, the potential energy accumulated while a leg is being raised converts most efficiently into kinetic energy at the moment it passes through the equilibrium position. After a quarter-period (that is, after a half-step) the kinetic energy of oscillation converts most efficiently into potential energy of the raised leg. In this case the maximal values of the potential and kinetic energies are almost identical, while the extra load to the muscles needed to "swing the legs" during walking is minimal.

When people walk, they actually do move their legs at the resonant frequency. The deviations of the "operational" frequency from the resonant frequency are quite small. The reason is clear: the greater this deviation, the greater the extra load on the muscles to overcome the inertia of the legs when one is walking too fast or, in the opposite case, to cope with the gravitational force affecting the legs when one is walking too slow.

You can encounter a similar phenomenon on the playground. It's comparatively easy to swing on a swing at its resonant frequency. But if you try to swing at, say, twice the resonant frequency, you'll get tired very quickly, and the swing will hardly move.

Now it's clear that when you need to go faster, you should increase the length of your stride and not the rate at which you step. As a result, your speed increases, but the resonant frequency changes only slightly: physics tells us that the resonant frequency is practically independent of the amplitude of the oscillations.

In addition to increasing the length of your stride, there's another way to walk faster: bend your leg at the knee as soon as you lift it off the ground. This trick won't affect your stride length, but it increases the resonant frequency. By "tuning" the resonant frequency of your legs to a higher value, you can walk faster.

A person's arms are also involved in the process of walking. Since the legs are moving in two parallel planes located some distance apart, the motion of the legs produces a torque that "tries" to turn the body about the vertical axis. To stay upright while walking, people swing their arms counter to their legs. This motion of the arms produces a torque that counterbalances the torque produced by the legs and compensates for it. Also, when you walk this way the vertical displacement of your center of gravity decreases, and along with it the mechanical work involved.

In fact, moving your legs apart while you walk lowers your center of gravity, while spreading your arms raises it. Of course, you can maintain a strictly vertical posture without swinging your arms, but only at the expense of an extra load on your muscles, and this makes your walking inefficient.

Swinging your arms while walking is easiest if the resonance frequencies of your legs and arms are identical. And indeed these frequencies do coincide. Moreover, when you "tune" the resonant frequency by bending your legs when you walk fast, you also bend your arms to keep the resonance frequencies of your legs and arms equal.

The faster you walk, the more you need to bend your extremities (legs and arms) to tune (that is, to increase) the resonance frequencies to the pace you choose. When you are running fast, your legs are bent at a very acute angle. Nevertheless, however we may try, the construction of our legs won't permit us to bend them more than double, so the most we can do is halve the reduced length. Therefore, we cannot increase the resonant frequency of the legs by more than a factor of $\sqrt{2}$. (Taking into account that the pushing leg should be straight, we would guess that the actual increase in frequency is significantly less than $\sqrt{2}$.)

So we see that "tuning" the resonant frequency to the walking speed is possible only to a rather limited extent. If one needs to run very fast, the potential energy accumulated by a raised leg isn't enough to give the leg the necessary kinetic energy of oscillation. The muscles must work to compensate for this deficiency. The kinetic energy required quickly increases with speed: it's proportional to the square of the runner's speed. Bearing in mind that for constant stride length the speed is proportional to the number of strides per second, we see that the energy "wasted" in swinging the legs at a rapid rate is proportional to the cube of the speed. This means that to increase the running speed by a mere 10%, one needs a one-third increase in energy expenditure; doubling the speed requires an eight-fold increase in energy. As for the forces the muscles must exert to move the legs when you run fast, they are proportional to the square of the speed.

This steep nonlinear dependence of energy expenditure on speed explains the relatively small difference in the speeds attained by record-holding sprinters and average runners. Now it's clear why it's so hard to shave another hundredth of a second from your time in the 100-meter dash. By the way, the steep energy–speed dependence also explains why long-distance runners conserve energy by taking long strides and moving their legs at nearly the resonant frequency. They do it despite the fact that the longer stride requires extra energy.

To find the quantitative dependence of the forces developed by muscles on the frequency of the strides, we need to know not only the natural frequency of the legs but also the decrease in the amplitude of these oscillations. For humans the amplitude is approximately halved during a single period.

The force of the muscles pulling the legs forward and backward acts periodically. We may assume that this periodicity is described by a sine function. Given these assumptions, we can calculate the force needed for any rate of walking using formulas developed from the theory of oscillations.

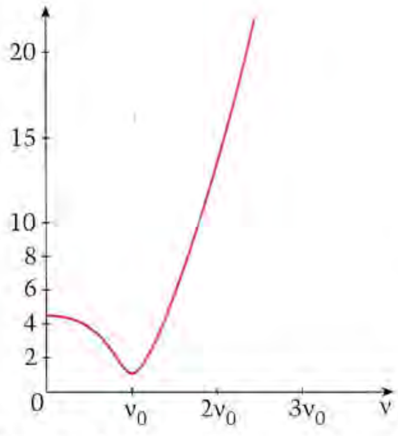


Figure 1. *Dependence of the forces setting our legs in motion on the walking speed, where ν is the frequency of oscillations of the legs—that is, the number of steps per second; and ν_0 is the resonant frequency of a leg. The vertical axis gives the factor by which the muscular force must increase in comparison with walking at the resonant frequency.*

The results of such calculations are shown in figure 1. It shows that the force required is minimal when the walking frequency is equal to the resonant frequency, and it increases drastically both for higher and lower frequencies. This graph makes it particularly clear how much more efficient it is to walk by moving the legs at the resonant frequency, and how the forces needed for walking increase catastrophically with speed. For example, doubling the stride rate (compared to the resonant frequency)—that is, increasing the number of strides per second by a factor of two—without changing the stride length increases the load on the muscles by a factor of 14! Even if the stride rate deviates from the resonant frequency by only 10%, the muscular forces required increase significantly. Now we see why people walk at a pace right at their individual resonant frequency: that pace is easiest and most efficient for them. (The graph in figure 1 qualitatively describes the process of running as well. Here we should remember that the resonant frequency for running is slightly higher than that for walking because your legs bend a bit more when you run than when you walk.)

The lower the reduced length of the legs (practically speaking, the more the legs taper from top to bottom), the higher the resonant frequency and the greater the speed that can be attained with the same effort. On the other hand, lighter legs require less muscular effort to move at the same speed, since the force is proportional to the mass of the leg. Now it's clear why swift-footed animals (for instance, antelopes, cheetahs, racing horses, and greyhounds) have strongly tapered and rather light legs.

True, the second condition is less important than the first, because the decisive role in running speed is played by the resonant oscillations of the legs. For example, elephants with their massive yet well-proportioned legs are almost frisky when they run—in India it's a compliment to say a person "moves like an elephant."

You needn't travel to Africa or Asia for examples: just try chasing a piglet sometime! The pig's legs (its "hams") have a cone shape, tapering sharply downward. The pig needs its hams not to grace our plates but to increase the resonant frequency of its legs. The short legs of a piglet have such a high resonant frequency that it can hold its own in a race with a human being, whose legs are much longer.

Thus the resonant frequency of the legs plays a key role in the speed at which creatures run and walk. A question naturally arises: can it be increased? It's tempting: increase the resonant frequency and increase the speed of runners and walkers.

In principle there are at least two ways to approach the problem. The first is to increase the force of gravity. The resonant frequency of a pendulum is proportional to the square root of the acceleration due to gravity. Therefore, doubling the gravitational force means increasing the leg's natural frequency by a factor of $\sqrt{2}$. True, it's not that easy to increase the Earth's gravitation (to put it mildly), and maybe it wouldn't be worth the effort. But for space travelers who will run and walk on other planets, the dependence of the resonant frequency of their legs on gravity isn't a theoretical issue—every planet has its own value for the acceleration due to gravity.

The second approach is much simpler. If the potential energy stored every time a leg is lifted is not sufficient to give the leg the necessary speed of oscillatory motion, we can tie the legs together with an elastic cord or attach some other elastic devices that help restore the equilibrium position of the legs. In this way we can increase the potential energy accumulated by the raised leg (at the expense of the kinetic energy). This will decrease the useless extra load applied to the muscles that is needed to overcome the inertia of the moving legs.

The elastic devices attached to the legs play the role of potential energy "batteries" that are "charged" and "discharged" twice during each period of oscillation. As in any oscillatory system, these "batteries" will increase the resonant frequency. The greater the elasticity of the device (say, the thicker the rubber cord between the legs), the higher the resonant frequency. So we've arrived at a paradoxical conclusion: in order to run fast, you should bind your legs—with elastic cord, of course.

An old German fairy tale describes a man who could run amazingly fast. To walk at a normal speed, he attached weights to his feet. Well, this fellow knew what he was doing! By adding the weights, he increased the reduced length of his legs and thereby decreased the resonant frequency and, along with it, his walking speed. If he attached a rubber cord instead of weights, he would run even faster. However, one might suspect that the secret of this fast runner consisted in the fact that his legs had a much higher resonant frequency than other people's legs had.

Any healthy person can become a fast runner by increasing the resonant frequency of the legs with some elastic device. Imagine doing it yourself: just before the start of a track-and-field event, you tie your feet together. When the judges ask you to explain your odd behavior, you modestly reply that, because you're such a fast runner, you tie your legs together so as not to run too far ahead of the other competitors.

As you take your starting position, you pull your leg back with some effort, thereby "charging" it with potential energy. When the gun goes off, you lift this leg from the ground and it flies ahead as if fired from a slingshot. This also decreases the acceleration period. You'll arrive at the finish line first—you, with your legs tied together!

Actually, the hard part is stopping yourself at the end of the race—your elastically bound legs "run by themselves."

If we wanted to take this experiment to its logical conclusion, it would make sense to attach the elastic devices to your arms as well, because you also use your arms when you run. You would simply attach elastic cords to your arms (above your wrists) and tie them to your waist. Our overall conclusion: if you want people to run faster, bind them hand and foot.

When I mentioned this to some of my colleagues, they laughed, but they agreed with me in principal. One of them, however, wasn't content to just stretch his mind—he wanted to put my idea to the test. So we got some thick rubber shock-absorbing cord, which is made of many thin parallel elastic rubber filaments. Then we cut 57-cm segments of these filaments and used them to make several rubber rings about 1 cm thick. It was a simple device: a ring was made of each rubber filament by tying the ends, and seven or eight rings were used to make a composite ring. At regular intervals this ring was tied with string (just as you tie up a bundle of twigs).

If you put this ring on your feet and stand with your legs together, the ring is neither stretched nor slack—it lightly grips your legs. When you walk or run, it gets stretched periodically. The elastic forces generated tend to restore the equilibrium position of your legs and thus increase the resonant frequency. In our experiments, the ring went around the legs at calf level and increased the resonant frequency by a factor of 1.5.

To keep the ring from constricting the blood vessels, we attached it to the legs with stiff bandages made of cardboard or a similar material. To prevent the ring from sliding upward, we attached it to the bandages, and also ran a string from the ring to our shoes. The strings didn't impede the motion of the ring in the horizontal plane, because the angle between the strings and the leg was always small—the strings were kept practically parallel to the legs.

The rubber ring increased the resonant frequency regardless of where it was attached—above the knee or below. Since the oscillatory kinetic energy of the lower part of the leg (below the knee) is 5–6 times greater than the kinetic energy of the upper part of the leg (above the knee), it makes sense to attach the ring below the knee.

I tried walking in such a device and found that my walking speed increased. I felt like something was helping me move my legs. When I wanted to stop, I found myself taking a few extra steps.

The rubber ring pulls the legs to the equilibrium position, so it alternately presses the front and rear of each leg. After a while this continual "massage" may be harmful. For long distances it's better to attach the device in such a way that its parts don't compress areas of muscle that have many blood vessels. Perhaps the easiest way to attach it is to tie the ends of the rubber cord to straps wrapped around your feet.

In order to prevent the cord from hanging loosely between your feet when it's not stretched, attach a rubber band to the middle of it and tie the other end of this slightly stretched rubber band to your belt. In this design the middle part of the elastic cord is somewhat elevated, but it doesn't impede motion. We should point out, however, that under great tension the elastic cord "tries" to rotate your feet about the vertical axis. The muscles must compensate for this torque; otherwise you'll look like the police dogs in A. N. Tolstoy's *The Golden Key*, which "ran with a peculiar gallop, their hind legs swinging out to the side." You can eliminate this defect by replacing the straps around your feet with a rubber ring that goes around the feet and rolls around them when you move.

You could use this device with horses as well, attaching the elastic straps to their hooves (one strap for the front hooves, one for the hind hooves).

If technology ever brings us "walking devices" and robots, the future designers should bear in mind that they can increase their speed or economize the energy needed by tying the "legs" together with an elastic cord.

Of course, we don't necessarily have to bind our legs to increase our speed. There are many possible designs for such elastic devices and methods of attaching them. However, all such systems must share this common property: the elastic devices must develop forces that act to restore equilibrium. In other words, they must enhance the resonant frequency of that wonderful set of biological pendulums—our legs. ◘

Quantum on oscillations and parametric resonance:

L. Aslamasov and I. Kikoyin, "Wave Watching," January/February 1991, pp. 12–16.

A. Chernoutsan, "Swinging Techniques," May/June 1993, pp. 64–65.

P. Mikheyev, "A Magical Musical Formula," January/February 1995, pp. 30–31.

A. Stasenko, "Hurling at the Abyss," November/December 1998, pp. 43–44.

Tycho, Lord of Uraniborg

by J. D. Haines

FOR AS LONG AS TYCHO could remember, he had loved to gaze at the stars. As he watched the stars night after night, he noticed patterns in their movement. There must be a way to understand the movements of the stars, he thought. And a way to measure how far away they were. More than anything else, Tycho wanted to study the heavens.

When Tycho Brahe was born in Sweden in 1546, very little was known about astronomy. People still thought that the ancient Greek philosopher Aristotle's laws ruled the universe. Aristotle had believed that the Earth was the center of the universe. He also thought that the heavens were unchangeable. Tycho was destined to change the way humans understood the universe.

Tycho's father was a nobleman, and the family lived in a castle. His mother was in charge of the Queen's court. He was his parent's first child. But a very strange thing happened when Tycho's little brother was born. Tycho's uncle was an important Danish Admiral. He and his wife had no children, but they badly wanted a child. They wanted to adopt Tycho, but instead of seeking his parents' approval, they kidnapped Tycho shortly after his brother was born.

The family worked out an agreement to let the uncle and his wife raise Tycho as their own son.

Tycho's parents went on to have ten more children. But only one of the eleven was destined for greatness. Tycho never forgot his fascination with the stars. At fifteen, Tycho left his uncle's home to attend the university. He announced to his uncle that he intended to study astronomy.

"What is this astronomy?" his uncle asked. "Foolish gazing at stars? Please be serious, Tycho. You are a nobleman. You need a real profession, like the law and statesmanship. It has already been decided. You will study law at Leipzig."

Tycho was very disappointed. He didn't want to spend his time poring over dry, dusty law books. His passion was the stars. But he knew that it was useless to argue with his uncle. Soon after arriving at the university, he secretly purchased books on astronomy and studied them instead of the law. Just before he left Denmark, Tycho was lucky enough to witness a partial eclipse of the Sun. When Tycho was nineteen, his uncle died. He was now free to study astronomy openly.

Tycho had a fiery temper as a young man. He once got into a disagreement with another Danish nobleman that resulted in a sword duel. The two fought with broadswords. Fortunately, they both survived. But Tycho had part of his nose cut off. He made a prosthetic nose out of gold and silver and wore it the rest of his life.

In astronomy, Tycho's first goal was to improve upon the poor instruments available to measure the stars. His greatest invention was called an armillary sphere. It had a ring nine feet across and was used to accurately locate stars.

He spent every night charting the heavens with his instruments. Then, on November 11, 1572, Tycho made an incredible discovery. He ob-

CONTINUED ON PAGE 33

A partial history of fractions

by N. Vilenkin

THE FIRST FRACTION PEOPLE used was a *half*. While the names of all other fractions are related to the name of the number in the denominator (one-third, one-fourth, etc.), this is not the case for half. In many languages, its name has nothing in common with the word for *two*. The next fraction was a *third*. These and some other fractions are encountered in the most ancient mathematical texts that have been handed down, which were written more than 5000 years ago on Egyptian papyri and Babylonian clay tablets. Both the Egyptians and Babylonians used a special notation for the fractions 1/3 and 2/3, different from the notation used for other fractions.

The Egyptians tried to write all their fractions as sums of *unit fractions*, that is, fractions of the form $1/n$. The only exception was 2/3. For example, 8/15 was written as 1/3 + 1/5. This was sometimes convenient. For example, the papyrus copied by the Egyptian scribe Ahmes includes the following problem: *divide seven loaves among eight people.* If we cut each loaf into 8 parts, 49 cuts are needed. This is more cuts than is necessary. The Egyptians solved this problem as follows. They wrote 7/8 as 1/2 + 1/4 + 1/8. Now it is clear that 4 loaves should be cut in half, 2 loaves should be cut into quarters, and just one loaf should be cut into 8 parts, which makes 17 cuts in total.

However, it was inconvenient to add fractions written as sums of unit fractions. If the same unit fraction appeared in both terms, the sum would contain the disallowed fraction $2/n$. For this reason, the Ahmes papyrus begins with a table in which all fractions of the form $2/n$ from 2/5 to 2/99 are written as a sum of unit fractions. The same table was used to divide integers. For example, this is how 5 was divided by 21:

$$\begin{aligned}\frac{5}{21}&=\frac{1}{21}+\frac{2}{21}+\frac{2}{21}\\&=\frac{1}{21}+\left(\frac{1}{14}+\frac{1}{42}\right)+\left(\frac{1}{14}+\frac{1}{42}\right)\\&=\frac{1}{21}+\frac{2}{14}+\frac{2}{42}=\frac{1}{7}+\frac{1}{21}+\frac{1}{21}\\&=\frac{1}{7}+\frac{2}{21}=\frac{1}{7}+\frac{1}{14}+\frac{1}{42}\end{aligned}$$

The Egyptians also knew how to multiply and divide fractions. They multiplied by expressing the fractions as sums of unit fractions, then multiplying the unit fractions. Again, two identical unit fractions may appear, so they needed the table of $2/n$ once more.

The Babylonians had a different approach. They used the *sexagesimal* (base-60) notation. A unit in each position was sixty times greater than the unit in the preceding position. For example, the notation 14″42′38 meant $14 \cdot 60^2 + 42 \cdot 60 + 38$, which is 52958 in our ordinary notation (of course, the symbols used by the Babylonians for the digits were different from ours: they were based on cuneiform characters). The Babylonian fractions were also sexagesimal rather than decimal. Actually, we still use such fractions in the notation of time and angles. For example, the time 3h 17 min 28 s can be written as 3.17′28″, which reads as *three whole plus 17 sixtieths plus 28 thirty-six hundredths hours.* Instead of the words *sixtieth* and *thirty-six hundredth,* the shorter terms *first small part* and *second small part* were used. This is where our words *minute* (from Latin *minutus*—small) and *second* (from Latin *secundus*) come from.

Not every fraction can be written as a finite sexagesimal fraction, just as not every fraction can be written as a finite decimal fraction. For example, 1/7, 1/11, and 1/13 cannot be written as finite sexagesimal fractions. However, they can be approximated by finite sexagesimal fractions with an arbitrary small error. That is how the Babylonians wrote them.

Art by Sergey Ivanov

Sexagesimal fractions were also used by Greek and Arab mathematicians and astronomers. However, it was inconvenient to use the decimal notation for integers and the sexagesimal notation for fractions. Computations with ordinary fractions were even more difficult: for example, try to add or multiply 785/2213 and 8917/3411.

For this reason, in 1585 the Dutch mathematician and engineer Simon Stevin suggested using decimal fractions. At first, notation was rather unwieldy; then, gradually, modern notation became common. As early as 150 years before Stevin, the astronomer Jamshid al-Kashi, who worked at Ulugh Beg's observatory in Samarkand, used decimal fractions, but his work remained unknown to European mathematicians.

In the present day, binary fractions are used in computers. In the binary notation, the unit in each next position is twice as much as the unit in the preceding position. For example, the notation 100101 stands for the number $1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2 + 1 = 37$. Although this notation is rather long, it uses only two digits. This kind of notation is easy to implement on a computer, using electrical currents. For example, the digit 1 could be represented by a current flowing through some object, and the digit 0 by the lack of such a current. An example of a binary *fraction* is 0.101101. (Try to write this fraction in decimal notation.) It is interesting that binary fractions were common in old Russia, where such fractions as *a half, a quarter, a half-quarter, a half-half-quarter,* and so on, were used.

An interesting system of fractions was used in ancient Rome. It was based on dividing the weight unit *as* into 12 parts. A twelfth part of the *as* was called an *uncia* (ounce). Distance and time were treated by analogy with the familiar weight system. For example, a Roman could say that he traveled 7 ounces of the way or read 5 ounces of a book. This clearly meant that he had traveled 7/12 of the way or read 5/12 of the book. Special names were used for fractions resulting from some number of 12ths to lowest terms, and also for those resulting from dividing 12ths into smaller units. The 1/288 part of the *as* was called *scrupulus* (from which we get our word *scruple*). Other names used were *semis* (half of the *as*), *sextans* (1/6 of the *as*), *semiuncia* (half of an *uncia*, or 1/24 of the *as*), and so on. Eighteen different names were used. Computations with fractions were performed with the aid of addition and multiplication tables. Thus a Roman merchant would remember that the addition of a *triens* (1/3 of the *as*) and a *sextans* gave a *semis,* while multiplication of a *bes* (2/3 of the *as*) by a *sesquiuncia* (3/2 of an *uncia*, or 1/8 of the *as*) gave one *uncia.* Of course, they understood completely that they were multiplying fractions represented as certain weights rather than the weights themselves (it wouldn't make sense to multiply weights). Special tables were used to facilitate computations, and some of them have come down to us.

Thus in ancient Rome the number 12 played the same role that was played by the number 60 in ancient Babylon and by 2 in old Russia. The Roman system of weights and measures was duodecimal (but the notation was decimal, though different from that of the present day). Since the numbers of the form $1/10^n$ cannot be represented as finite duodecimal fractions, the Romans could not represent the result of dividing a number by 10, 100, etc. in the form of a fraction. For example, this is how a Roman mathematician would divide 1001 by 100. First, he obtained the integer part, 10 *ases*; then he divided the 1 *as* remainder into *unciae*, etc. But he was always left with a remainder.

In the mathematical works of ancient Greece, fractions were not used. Greek scholars thought that mathematics must deal only with whole numbers. Fractions were left to merchants, craftsmen, astronomers, and mechanics. Thus fractions found their way into the scholarly works of the Greeks in a roundabout way. Besides arithmetic and geometry, Greek mathematics included *music.* That was the name they gave to the branch of present-day arithmetic that deals with ratios and proportions.

Why this strange nomenclature? The Greeks had developed a scientific theory of music. They knew that the longer a string, the lower the sound, and the shorter a string, the higher the sound. Musical instruments often have several strings. For them to be in tune, their lengths must have a certain relation to each other. For example, for the pitch of the sounds produced by two strings to differ by an octave, their lengths must be in the ratio of 1:2. In a similar way, the ratio of 2:3 corresponds to the fifth, and the ratio of 3:4 to the fourth. Thus the Greeks considered the part of arithmetic dealing with ratios and fractions to be a part of music.

The modern system of notation of fractions with a numerator and denominator was developed in India. However, they wrote the denominator at the top and the numerator at the bottom, and they didn't use the fraction bar. The Arabs were the first to write fractions exactly as we do today. ◘

CONTINUED FROM PAGE 13

M305

Positively unknown (twice). Find positive solutions $x_1, \ldots, x_n$ to the equation

$$\cfrac{1}{x_n + \cfrac{1}{x_n + \cfrac{1}{x_{n-1} + \cfrac{1}{x_{n-1} + \ldots \cfrac{1}{x_1 + \cfrac{1}{x_1}}}}}} = \frac{n}{n+1}.$$

Every unknown occurs twice in the equation.

ANSWERS, HINTS & SOLUTIONS ON PAGE 46

Matter and

THIS INSTALLMENT OF OUR Kaleidoscope inaugurates a series of articles describing the interaction of two aspects of the material world: fields and matter. We begin with the behavior of objects in a gravitation field. In this article we'll discuss not only *material points*, whose motion in the gravitational field is described in most physics textbooks, but also *extended objects*. The motion of such objects reveals odd features of the "interrelations" between the gravitational field and matter, which have occupied scientists for centuries.

The motion of objects near the Earth, weightlessness, tidal phenomena, the synchronous rotation of planets and their satellites, the evolution of stars, the past and future of the Universe—all these problems are linked by the phenomenon of gravity. Although great scientists of the past have predicted many phenomena involving gravitation (for example, Laplace predicted the existence of "black holes" 200 years before they were discovered), the stream of unexpected and downright amazing discoveries hasn't dried up yet. And the increasing power of Earth-based and orbiting telescopes forces us to prepare our minds for new and amazing discoveries "out there."

So that these future discoveries will not catch us napping, let's try to get a handle on at least some of the questions raised by gravity. After that, we'll move on to electricity and magnetism in future Kaleidoscope articles.

"Nothing composed of terrestrial matter and raised to a height can avoid the strong ties of gravity." —Johannes Kepler

Questions and problems

1. Two objects start to fall simultaneously from two points located on the same vertical line near the Earth's surface. How will the distance between them change during their free fall?

2. Where is it easier for a person to swim—on Earth or on the Moon?

3. Are the gravitational forces on two identical balls—one of them floating in water and the other lying on a table—identical?

4. A rocket leaves Earth along a radial path with constant acceleration equal to the acceleration due to gravity at the Earth's surface. How does the weight of a passenger change with distance from the Earth?

5. A vessel with a floating object starts to fall with acceleration $a < g$. Will the object rise up out of the water?

6. Why do astronauts sleep without pillows for some time after they return to Earth?

7. Can a conventional medical thermometer be used in an orbiting space station?

8. Will the air density in a spacecraft change when it enters the state of weightlessness?

9. Why does an object weigh less at the equator than at the poles?

10. Why are the planets spherical, while comets and asteroids have very irregular shapes?

11. Two travelers started a round-the-world voyage from the same point on the equator and traveled at the same speed. One of them traveled along the equator, while the other traveled along a meridian. Will they arrive at the starting point at the same time at the end of their travels?

12. How rapidly must a planet rotate on its axis to be destroyed?

13. What causes the tides?

14. Why are the tides the highest when the Moon is new and full and lowest at the first and last quarters?

15. How would solar eclipses differ if the radius of the Moon's orbit were half the present value?

16. Imagine that the Earth stopped rotating on its axis and kept moving around the Sun. How would

nd gravity

the acceleration due to gravity at the Earth's surface vary from place to place?

Microexperiment

Next time you go swimming, carry a weight to help you submerge. Try to stay motionless in a state of neutral equilibrium, even for just a moment. Can we say that this is a state of weightlessness?

It's interesting that ...

...long before Newton, Johannes Kepler thought about the "gravity" acting between celestial bodies and used the notion to explain the tides, asserting that the Moon attracts the waters of the oceans.

...there are still doubts about whether Galileo ever really dropped anything from the Leaning Tower of Pisa. However, it's known that the tower was used in Galileo's time by an Aristotelian demonstrating that different objects fall at different rates. Be that as it may, it was certainly Galileo who managed to establish that the acceleration of freely falling objects in a vacuum does not depend on the nature of the objects (such as their size, shape, or mass).

...according to Aristotle, the force of gravity doesn't change with distance as the object approaches the Earth's center, but suddenly changes direction when it passes the center. This was the scientific description used by Dante Alighieri (1265–1321) in his description of the deepest region of Hell (the center of the Earth). By contrast, Newtonian theory says that the force of gravity becomes zero at the center, as Newton had proved that an object placed inside a thin spherical shell does not experience any forces.

...in 1801—almost two hundred years ago—the German astronomer I. Zoldner calculated the deflection of light rays in the Sun's gravitational field on the basis of Newtonian theory. More than a century later the English astronomer O. Lodge introduced the term "gravitational lens" and predicted that extensive celestial objects such as galaxies may produce several images of distant stars because of the deflection of light passing from these stars through such "lenses."

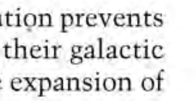
...the low-orbiting satellites of Jupiter have fewer craters on their surfaces than the high-orbiting ones. This is due to tidal forces and the magnetic field of the giant planet, which promote tectonic activity in the nearer satellites. Their surfaces are constantly being "reworked", and the old craters disappear.

...although about ten thousand tons of space debris falls to Earth every day in the form of meteorites and space dust, the resulting increase in the Earth's mass during its lifetime has had virtually no effect on the rate of the planet's rotation about its axis.

...the Earth's daily rotation is slowing due to tidal friction. Billions of years from now the same hemisphere of Earth will continually face the Moon, as Pluto faces its satellite Charon. By that time the Moon will have moved away from the Earth to a distance of 553,000 kilometers, and a month will last 47.2 days.

...according to the theory of relativity, the accelerated motion of objects should generate gravitational waves. But because of the weak nature of gravitation, these waves have not yet been detected from even the largest celestial objects. (By the way, the operating principle of the gravitational wave "receivers" is based on the tidal effect.)

...the Newtonian theory of gravity does not apply to "black holes," whose monstrous gravitational force prevents even light from escaping, even though the very possibility of their existence follows from this theory.



...the destructive effect of tidal forces was "demonstrated" by Jupiter, which in July 1992 destroyed a comet that dared to approach it too closely.

...for many years astronomers looked unsuccessfully for the so-called dark matter in the Universe, whose gravitation prevents galaxies from leaving their galactic clusters and slows the expansion of the Universe. However, recent astronomical data indicate that gravitation is not sufficient to bring the rapidly dispersing galaxies back together. The Universe expands at an increasing rate, and this process will go on forever. ◘

ANSWERS, HINTS & SOLUTIONS ON PAGE 51

Curved reality

by Arthur Eisenkraft and Larry D. Kirkpatrick

There are always two choices, two paths to take. One is easy. And its only reward is that it's easy.
—Unknown

ALL CURVES ARE NOT EQUIvalent. As young babies, perhaps as young as two or three months, we can distinguish between straight lines and curves. As we enter school, we learn geometry based on straight lines and the shapes that they create—triangles, squares, hexagons, and the like. During our fascination of discovering the mundane and esoteric properties of these polygons, we forget to ask about their relevance to the natural world. Certainly, we require squares to map out our rooms and gardens. We need triangles to determine the height of a flagpole from its shadow. We need straight lines to find the shortest distance between two points.

Does nature share our enthusiasm for straight lines? How often do we find straight lines in nature? The horizon appears to be a line, but we know that it must curve if the Earth is a sphere. The rays of light piercing through the clouds provide one example of the straight line in nature. The quarter moon illuminated by the Sun at just the correct angle provides us with another natural straight line. Is the edge of a crystal a straight line? Are there other examples? We would enjoy hearing from our readers as they expand this short list of natural straight lines.

Perhaps nature's straight lines are more subtle. Perhaps the lines are there not for our visual eyes but for our mind's eye. As a ball falls, its path is a vertical line. If we could only see the ball at all places at once, how beautiful it would be. When we measure the stretch of a spring with varying weights, we discover Hooke's law, which states that the stretch is directly proportional to the force. If we write that as an algebraic equation, we succinctly state $F = -kx$. When we graph this, we find nature's straight line. Every direct relation, from $x = vt$ to $F = ma$ to $V = IR$, is a discovery of a straight line in nature.

Visually, if not straight lines, then does nature favor curves? All curves are not equivalent. Does nature favor the circle (the Greek's symbol of perfection) over the parabola? Does the hyperbola appear more often than the ellipse? Does the cycloid make more appearances than the catenary? Let's embark on a brief tour of some simple physics with an eye toward the curves we may discover along the way. Following the earlier notion of the path of a falling ball, what would be the paths of other moving objects?

The thrown ball, without air resistance, travels along a parabola. This is simply proven with the equations of motion for horizontal and vertical motion.

$$x = v_0 t\cos\theta,$$
$$y = -\frac{1}{2}gt^2 + v_0 t\sin\theta.$$

Eliminating the time between the two equations—allowing us to see the thrown object at all times—the equation of motion becomes:

$$y = \frac{-gx^2}{2v_0^2\cos^2\theta} + \frac{x\sin\theta}{\cos\theta}.$$

This path is identical to the general equation of a parabola: $y = ax^2 + bx$. The curve of the parabola is determined by the initial angle and velocity of the throw. But all throws result in the parabolic shape.

A charged particle shot into a region of space containing a magnetic field experiences a force that is perpendicular to both its instantaneous velocity and the magnetic field. This perpendicular force serves as the centripetal force as the particle moves in a circle.

$$F = qv \times B = \frac{mv^2}{R},$$
$$mv = qBR.$$

Knowing the charge and magnetic field, we can measure the radius of the curved path and deter-


Art by Tomas Bunk


TOM BUNK

mine the momentum of the charged particle.

An orbiting satellite also follows a curved path. In this situation, the gravitational attraction between the Earth and the satellite is the centripetal force. Since the speed is dependent on the distance, we can determine the mass of the Earth by the behavior of the satellite.

$$F = \frac{GM_E m}{R^2} = \frac{mv^2}{R},$$

$$v = \frac{2\pi R}{T},$$

$$M_E = \frac{4\pi^2 R^3}{T^2}.$$

A moon orbiting a planet provides us with the mass of the planet if we are able to measure the radius of the moon's orbit and its period.

The moon about the planet and the planets about the Sun do not travel in perfect circles. It is a testament to the experimental accuracy of Tycho Brahe's measurements that Kepler could not "settle" for the approximate circular path of the planets, but determined that they travel in ellipses. The eccentricity of Halley's comet with the Sun at one focus provides strong evidence of Kepler's mathematical insight.

We find hyperbolic paths when we observe the scattering of an alpha particle from the nucleus of a gold atom. This Rutherford scattering experiment demonstrated the existence of a nucleus, which contains (almost) all of the mass and all of the positive charge of the atom packed into an extremely small volume. Knowing the deflection angle of the alpha particle, we can determine the impact parameter (how close the particle comes to the nucleus) or vice versa. The derived equation is:

$$\tan\frac{\Theta}{2} = \frac{q_1 q_2}{msv_0^2},$$

where Θ is the scattering angle and s is the impact parameter.

A particularly interesting curve is traced out by a point on a rolling wheel (figure 1). You sometimes observe this shape at night from the reflectors on bicycle tires.

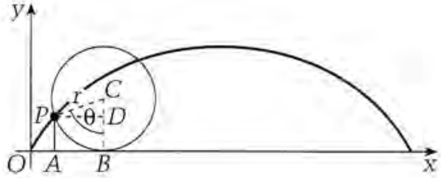


Figure 1

$$x = OA = OB - AB = OB - PD = r\theta - r\sin\theta = r(\theta - \sin\theta),$$

$$y = AP = BD = BC - CD = r - r\cos\theta = r(1 - \cos\theta).$$

This month's contest problem challenges you to traverse the terrain of some of these shapes.

1. A trajectory with air resistance can be assumed to have a frictional force proportional to the velocity.

$$F_t = -bv.$$

Derive an equation for the trajectory and sketch the path for different values of b.

2. A charged particle is projected into a region of space containing crossed (perpendicular) electric and magnetic fields. It enters the region perpendicular to both fields.

a) If the particle traverses this region without any deflection, show that the speed of the particle equals E/B. This is a way we can create a velocity selector.

b) If the particle traverses the crossed fields from the opposite direction with the speed E/B, does it remain undeflected?

c) If the particle traverses the crossed fields in the original direction, but with speed other than E/B, determine the curved path.

Magnetic vee

In the March/April issue of *Quantum*, we asked you to calculate the direction and magnitude of the magnetic fields at points P and P^* produced by the current in the magnetic vee shown in figure 2.

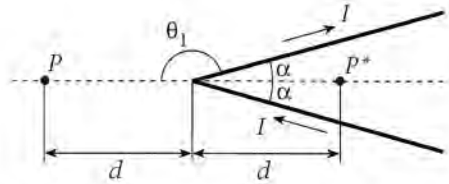


Figure 2

Part A: Using the right-hand rule, we see that the direction of the magnetic field produced by each segment of both wires is out of the page at point P and into the page at point P^*.

Part B: Symmetry tells us that the total field at point P is twice that generated by each half of the vee. To calculate the magnetic field produced by the upper wire, we use the formula for a current segment derived in the March/April issue. The first angle θ_1 is defined for the left-hand end of the segment. It is the angle between the current and the position vector of the point P. Therefore,

$$\cos\theta_1 = \cos(\pi - \alpha) = -\cos\alpha.$$

The second angle θ_2 is defined in the same way for the right-hand end of the segment. Because the wire is infinitely long in this direction, the angle is effectively 180° and

$$\cos\theta_2 = \cos\pi = -1.$$

Finally, the perpendicular distance between the (extended) wire and the point P is equal to $d\sin\alpha$. Remembering to multiply by 2 for the two current segments, we now have

$$B = 2\frac{\mu_0 I}{4\pi a}(\cos\theta_1 - \cos\theta_2) = \frac{\mu_0 I}{2\pi}\frac{1-\cos\alpha}{d\sin\alpha} = \frac{\mu_0 I}{2\pi d}\tan\frac{\alpha}{2}.$$

For point P^*

$$\cos\theta_1 = \cos(-\alpha) = \cos\alpha$$

and

$$\cos\theta_2 = \cos(-\pi) = -1.$$

Therefore,

$$B = \frac{\mu_0 I}{2\pi}\frac{1+\cos\alpha}{d\sin\alpha} = \frac{\mu_0 I}{2\pi d}\cot\frac{\alpha}{2}.$$

You can also solve part B by treating point P^* as if it were outside a vee with half-angle $\pi - \alpha$ carrying current I in the opposite direction. Then

$$B = \frac{\mu_0 I}{2\pi d}\tan\left(\frac{\pi - \alpha}{2}\right) = \frac{\mu_0 I}{2\pi d}\cot\frac{\alpha}{2}.$$

As a third method, you can use the superposition principle. The problem is equivalent to two crossed infinite wires plus a vee on the left carrying current I in the clockwise direction as shown in figure 3. The currents in the two vees on the left cancel each other, leaving the original situation. The superposition of the fields due to the vee on the left and the crossed infinite wires gives us the answer.

Each of the infinite wires produces a magnetic field given by

$$B = \frac{\mu_0 I}{2\pi d \sin\alpha}$$

into the page. The contribution to the magnetic field due to the vee on the left is out of the page and given

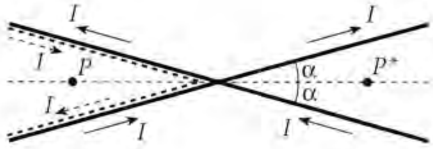


Figure 3

by our answer for point P. Therefore, the total field is

$$B = \frac{\mu_0 I}{2\pi d}\left(\frac{2}{\sin\alpha} - \tan\frac{\alpha}{2}\right) = \frac{\mu_0 I}{2\pi d}\cot\frac{\alpha}{2}.$$

CONTINUED FROM PAGE 25

served a "new" star that was as bright as Venus. Today we know that he saw a supernova, a variable star that suddenly increases in brightness before it dims. The discovery allowed Tycho to make two breakthroughs. First, he proved that the heavens were indeed changeable, contrary to what Aristotle had predicted. Secondly, he measured the new star to be as far away as the other stars.

At twenty-seven, Tycho was the most famous astronomer in all of Europe. The King of Denmark gave him the island of Hven and built him a castle that became the greatest observatory in the world. Tycho named the observatory Uraniborg, which means castle of the heavens. His next big discovery was the observation of the comet of 1577, which further challenged Aristotle's view of a static universe. For over twenty years Tycho taught astronomy and carefully mapped the sky. He engraved the stars' positions on brass plates that covered a celestial globe five feet in diameter.

Tycho also pioneered a method for measuring the distance of stars called stellar parallax. He observed that a nearby star that is sighted from opposite sides of the Earth's orbit about the Sun appears to shift in position (see figure 1). By calculating the parallax angle using a mathematical formula, a star's distance from the Earth could be determined.

Tycho was forced to leave Uraniborg when the king supporting him died. He moved to Prague to become the Imperial Mathematician for Emperor Rudolph II. Tycho was an old man now, and he wanted a younger man to continue his life's work. He met Johannes Kepler, who became his assistant and heir. After Tycho died in 1601, Kepler was able to use Tycho's observations of Mars to help formulate the laws of planetary motion. Kepler's discoveries then led to Isaac Newton's explanations of the workings of the solar system due to gravity. Newton's work in turn served as the basis for Albert Einstein's great leap forward in the twentieth century with the theory of relativity.

But it was Tycho's love of the stars and his careful observations that laid the foundation for many historic discoveries made by later astronomers. It has been said that each generation stands on the shoulders of the generation before them in order to see a little further into the unknown. There are many scientists today standing on the shoulders of Tycho Brahe.

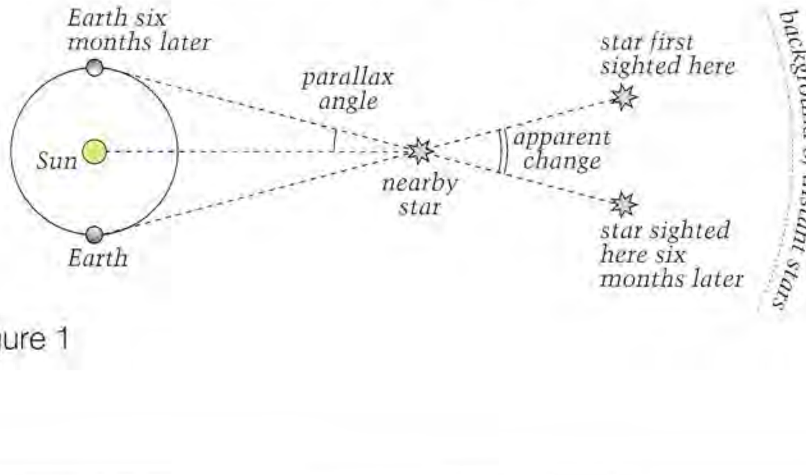
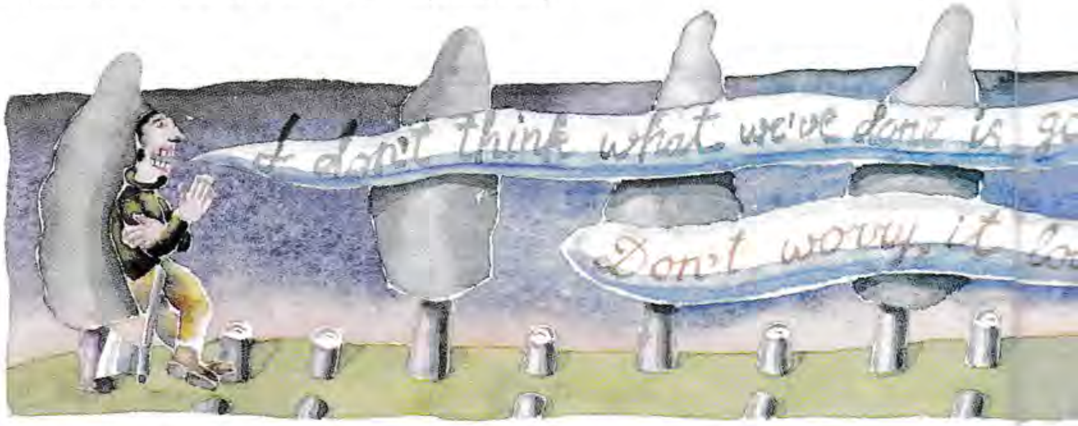


Figure 1

AT THE BLACKBOARD I

On quasiperiodic sequences

by A. Levitov, A. Sidorov, and A. Stoyanovsky



Art by Sergey Ivanov

LET'S PLAY A BIT WITH SEquences consisting of ones and zeros. Our sequences can be finite or infinite (or even infinite in both directions)[1]. Here's one example of an infinite sequence (this one is periodic):

...011001011001011001011O... .

A more interesting example of such a sequence can be constructed as follows. Define a transformation Π that assigns a new sequence $S' = \Pi(S)$ to every sequence S. This transformation *simultaneously* replaces every 1 in S by the three digits 100, and it replaces every zero in S by the two digits 10. For example, the sequence $S = 11010$ is transformed into $\Pi(S) = 1001001010010$. It's clear that $\Pi(S)$ is longer than S.

Now take $S_0 = 1$, and set $S_1 = \Pi(S_0)$, $S_2 = \Pi(S_1)$, $S_3 = \Pi(S_2)$, $S_4 = \Pi(S_3)$, and so on (table 1). We see that every sequence is a continuation of the preceding one.

Exercise 1.

(a) Prove this fact.

(b) Formulate a similar proposition for $S_0 = 0$ rather than 1.

Thus all sequences S_n form the beginning segments of an infinite sequence S_∞. This is our first interesting example of a sequence.

Exercise 2. Prove that $\Pi(S_\infty) = S_\infty$.

Consider the sequence S_∞ in more detail. First of all, it is not periodic.

Exercise 3. Prove this fact (or wait for the proof below).

Second, let's find the average density of ones in this sequence. The average density c can be calculated (to a high accuracy) as $c \approx$ (the number of ones in a 1 km of the sequence)/(the number of digits in a 1 km of the sequence).

In mathematical language, the definition can be formulated as follows: this density is the *limit* $\lim_{n\to\infty} c_n$, where c_n = (the number of ones in S_n)/(the number of digits in S_n). The first several values of c_n are given in the table. Notice that c_n rapidly approaches a number $c = 0.41421...$. You may have guessed by now that this number is $c = \sqrt{2} - 1$.

Exercise 4. Prove this fact.

$$\left(\textit{Hint: } c_{n+1} = \frac{1}{2 + c_n}.\right)$$

Thus we have constructed a sequence with a remarkable property: it has an irrational density of ones. We note that periodic infinite sequences do not possess this property.

n	S_n	c_n
0	1	1
1	100	1/3 = 0.33333...
2	1001010	3/7 = 0.42857...
3	10010101001010010	7/17 = 0.41176...
4	10010101001010010100101010010100101010010	17/41 = 0.41463...
5	...	41/99 = 0.414141...
6	...	99/239 = 0.41422...
7	...	239/577 = 0.41421...
8	...	577/1393 = 0.41421...

Table 1

[1]An infinite sequence is a function that assigns to every positive integer a number 0 or 1. A doubly infinite sequence is such a function defined for all integers (positive, negative, and zero).

Exercise 5. Prove this fact (and derive the assertion of exercise 3 from it).

The following problems involving further properties of the sequence S_∞ may be skipped on first reading.

Problem 1. Prove that the sequence S_∞ is quasiperiodic—that is, every finite segment of S_∞ occurs an infinite number of times at different places in the sequence.

Problem 2. Number the digits in S_∞ using the natural numbers. Let m_i be the number of the ith digit in the sequence. Prove that the following formula gives m_i explicitly:

$$m_i = \left[i\left(\sqrt{2}+1\right) - \frac{\sqrt{2}}{2} \right],$$

where $[a]$ denotes the integer part of a.

We now have made our acquaintance with the first inhabitant of the fantastic world of quasiperiodic sequences. Next, we'll deal with doubly infinite sequences.

Statement of the main problem

Let S be a doubly infinite sequence of ones and zeros. Does a sequence S' exist such that $\Pi(S') = S$? It's clear that S' exists if and only if the sequence S can be broken down into segments of the form 100 and 10. We call such a sequence *felicitous* and denote it as $S' = \Pi^{-1}(S)$. Suppose the sequence $\Pi^{-1}(S)$ itself is felicitous. Then there exists a sequence S'' such that $\Pi(S'') = S'$, and, therefore, $\Pi(\Pi(S'')) = S$. In this case, we call S *very felicitous.* As you might guess, a sequence S for which there exists a sequence S''' such that $\Pi(\Pi(\Pi(S'''))) = S$ is called *very very felicitous*, and so on. Finally, the most important point:

Definition. The sequence S is called *good* if the transformation that is the inverse of Π can be applied to it an infinite number of times. In other words, S is good if there exist sequences S', S'', S''', ... such that $S = \Pi(S')$, $S' = \Pi(S'')$, $S'' = \Pi(S''')$, and so on.

The main problem that we'll focus on is to classify all good sequences. Naturally we won't differentiate between sequences that can be obtained one from another by a shift.

The first question is whether at least one good sequence exists. The answer is given in the following exercise.

Exercise 6. Prove that if we adjoin the sequence from exercise 1(b) (which is infinite to the left) to the sequence S_∞ (which is infinite to the right), we obtain the good sequence S, and that $\Pi(S) = S$.

The next question is: do other good sequences exist, and if so, how many are there? Let's try to answer this on the fly. Let S and T be two good sequences.

Exercise 7. Prove that any finite segment of S occurs in T and vice versa.

The result of exercise 7 has the following interpretation. Let's imagine that the sequences S and T are written on strips of paper that are placed on a doubly infinite table. We'll also imagine that an infinitely large committee is sitting at the table, and the committee's aim is to determine whether the sequences are identical or not. If every member of the committee can examine only a finite segment of the sequences, it's impossible for this committee to find any differences between S and T.

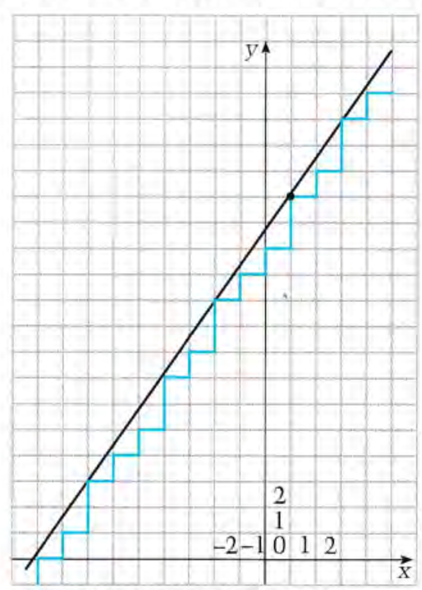
Can the committee draw the conclusion that $S = T$? (We, on the other hand, can take the entire sequence T in view and make certain that it cannot be made identical to S by shifts.)

So, do different good sequences exist? To answer this question and others, we'll use an unexpected yet powerful tool—graph paper.

A geometric interpretation

Let's take a big sheet of graph paper. Introduce a Cartesian reference frame with a unit along the coordinate axes equal to one square of the graph paper. We'll depict each sequence as a polygonal path consisting of line segments parallel to the coordinate axes. If the next digit in the sequence is 1, the polygonal path moves one square to the right; if the next digit is 0, the polygonal path moves one square up (see figure 1). The polygonal path corresponding to a good (felicitous) sequence will also be called good (felicitous) as well.

One thing we can notice when we look at a good polygonal path (such as that in figure 1) is that it looks very much like a straight line. More precisely, there are no nodes of the graph paper between the polygonal path and some straight line.



...101001010010101001010010100101010010...

Figure 1

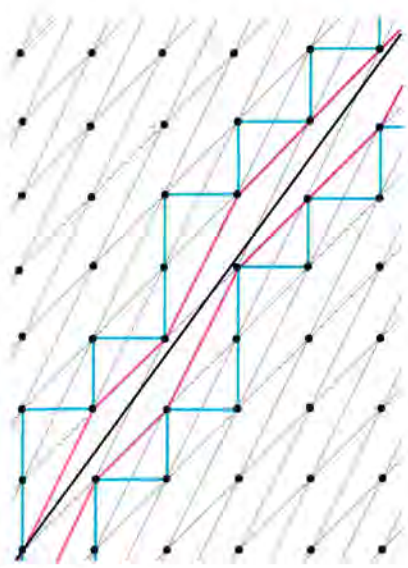


Figure 2

The polygonal path presses up against the straight line. To describe this straight line, we must, first of all, find its slope k. It's clear that the slope is approximately equal to the ratio $k \approx$ *(the number of vertical sections in* 1 *km of the polygonal path)/(the number of horizontal sections in* 1 *km of the polygonal path)* $\approx$

$$\frac{1-c}{c},$$

where c is the average density of ones. We know that $c = \sqrt{2} - 1$ (see exercises 4, 6, 7). Therefore,

$$k = \frac{2-\sqrt{2}}{\sqrt{2}-1} = \sqrt{2}.$$

This sort of inexact reasoning (which mathematicians call *heuristic* reasoning) suggests a method for constructing many examples of good sequences.

Theorem 1. *Let l be a straight line described by the equation $y = \sqrt{2}x + b$ (where b is an arbitrary number), and let $\Lambda_A = \Lambda_A(l)$ and $\Lambda_B = \Lambda_B(l)$ be polygonal paths that press up against l from above and from below, respectively. Then Λ_A and Λ_B are good.*

The idea of the proof. First of all, Λ_A and Λ_B are felicitous.

Exercise 8. Prove this fact.

We will prove that the polygonal paths $\Pi^{-1}(\Lambda_A)$ and $\Pi^{-1}(\Lambda_B)$ also press up against a certain straight line l' from below and above, where the equation of l' is $y = \sqrt{2}x + b'$ (for a certain b'). It follows from this fact that $\Pi^{-1}(\Lambda_A)$ and $\Pi^{-1}(\Lambda_B)$ are felicitous, and so on.

Proof. Connect the endpoints of the sections corresponding to 100 and 10 of the polygonal paths by segments (shown in red in figure 2). We obtain two new polygonal paths $\tilde{\Lambda}_A$ and $\tilde{\Lambda}_B$. Their sections lie on a new grid consisting of lines with slopes of 1 or 2 (the nodes of the new grid coincide with the nodes of the initial grid). Now imagine that the drawing in figure 2 is made on a sheet of transparent rubber. Stretch the sheet to make the new grid coincide with the initial (rectangular) grid. Then turn the sheet upside down as shown in figure 3. After these transformations, the polygonal paths Λ_A and Λ_B go into $\Pi^{-1}(\Lambda_A)$ and $\Pi^{-1}(\Lambda_B)$, respectively. Indeed, the sections corresponding to 100 on the old polygonal path Λ_B are assigned the horizontal section of the new line, while the section corresponding to 10 is assigned the vertical section. The same applies to Λ_A.

The transformation F of the coordinate plane that we performed (stretching and turning) takes the point (x, y) to the point

$$F(x, y) = (-x + y, 2x - y). \qquad (*)$$

In particular, $F(1, 2) = (1, 0)$ and $F(1, 1) = (0, 1)$. The transformation F has a fixed point at the origin. It

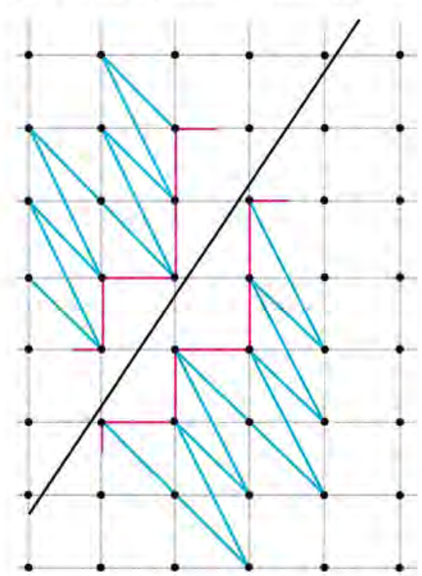


Figure 3

takes every straight line into a straight line and retains the lengths of the segments on the line.

Exercise 9. Prove these properties of F on the basis of formula (*).

Transformations of this kind are called *linear.*

Exercise 10. Prove that F takes the line l described by the equation $y = \sqrt{2}x + b$ into the line l' described by the equation $y = \sqrt{2}x + b'$, and find the coefficient b'.

It remains to note that there are no grid nodes between the polygonal paths $\Pi^{-1}(\Lambda_A)$, $\Pi^{-1}(\Lambda_B)$, and the straight line l' (see figure 3) (the same applies to $\tilde{\Lambda}_A$ and $\tilde{\Lambda}_B$ and the line l in figure 2). Thus, $\Pi^{-1}(\Lambda_A)$ and $\Pi^{-1}(\Lambda_B)$ press up against the line l'. The theorem is thus proved.

Solution to the main problem

It turns out that the converse theorem also holds.

Theorem 2. Let Λ be a good polygonal path. Then there exists a line l described by the equation $y = \sqrt{2}x + b$ against which Λ presses up from above or below.

The proof of this theorem will be given at the end of this section. In the meantime we'll explain how this theorem helps classify good sequences (or polygonal paths). It remains to find which polygonal paths of the form $\Lambda_A(l)$ and $\Lambda_B(l)$ give identical sequences (that is, sequences that can be brought into coincidence by a translation).

Exercise 11. Prove the following propositions.

(a) The line l contains a node of the grid if and only if $b = m + n\sqrt{2}$ for certain integer m and n.

(b) If the line l does not contain any nodes of the grid, then the polygonal paths $\Lambda_A(l)$ and $\Lambda_B(l)$ are equivalent. That is, the upper line can be obtained from the lower line by translation by the vector (–1, 1).

(c) Let l_1 and l_2 be two lines described by the equations $y = \sqrt{2}x + b_1$ and $y = \sqrt{2}x + b_2$, respectively. Then, the polygonal paths $\Lambda_A(l_1)$ and $\Lambda_A(l_2)$ are equivalent if and only if $b_1 - b_2 = m + n\sqrt{2}$, where m and n are integers.

(d) The polygonal paths $\Lambda_A(l_0)$ and $\Lambda_B(l_0)$ for the line l_0 described by the equation $y = \sqrt{2}\,x$ are not equivalent.

Thus every number b is assigned a good sequence (except for the numbers b of the form $m + n\sqrt{2}$, which are assigned two different good sequences), and two sequences are equivalent if and only if the corresponding numbers b_1 and b_2 differ by $m + n\sqrt{2}$. This is the complete solution to the classification problem.

Exercise 12. (a) Prove that the set of good sequences is infinite.

(b) Find an explicit formula for the position of the ith one in the sequence corresponding to the number b.

(c) Find the number b for the sequence S in exercise 6.

Now let's prove theorem 2.

Lemma. Any good polygonal path Λ can be placed in a strip between two straight lines with slope $\sqrt{2}$, and the upper line is obtained from the lower line by translation by the vector $(-1, 1)$.

Proof. Let's try to place Λ in a strip as specified in the lemma (it will be called the *strip of standard width*). Take a line l_1 with the slope $k = \sqrt{2}$ that passes very high (above Λ) and start moving it downward until it comes up against Λ. (As we do this, a strange situation can occur where the polygonal path doesn't touch the straight line. At the same time, it's impossible to move the straight line downward because different "teeth" of the polygonal path approach arbitrarily close to the straight line.) We perform the same operation with the line below: take a line l_2 with the slope $k = \sqrt{2}$ and move it right up to Λ. We assert that the width of the strip containing Λ is not greater than the standard width. Indeed, suppose that the strip containing Λ is wider. Move the lines l_1 and l_2 a bit closer to each other. Then the polygonal path will intersect both boundaries of the strip, which is still wider than the standard one. Consider a finite part of the polygonal path that contains the intersection points with both lines l_1 and l_2. But the finite parts of Λ are identical to those of some polygonal path $\Lambda_B(l)$ (see exercise 7), which can be placed in a strip of the standard width (the proof is left to the reader). Thus we have arrived at a contradiction, which proves the lemma.

Proof of theorem 2. Place the polygonal path Λ between the lines l_1 and l_2, which can be done by virtue of the lemma. Then either $\Lambda = \Lambda_B(l_1)$ or $\Lambda = \Lambda_A(l_2)$, or both.

Exercise 13. Prove this fact.

Thus theorem 2 is proved.

Conclusion

Let us summarize. We have become acquainted with quasiperiodic sequences of a special type (related to the transformation Π). In addition, we studied the structure of polygonal paths that press up against a straight line with slope $\sqrt{2}$.

We can continue our study in two directions. The first is to consider transformations different from Π. For example, consider the transformation $\tilde{\Pi}$ that takes 1 into

$$1\underbrace{00\ldots0}_{n\text{ zeros}}$$

and 0 into

$$1\,\underbrace{00\ldots0}_{n-1\text{ zeros}}.$$

Problem 3. Classify the sequences that are good relative to $\tilde{\Pi}$. Everything is similar to the case of Π, except that the role of $\sqrt{2}$ is played by the number

$$\frac{n + \sqrt{n^2 + 4}}{2} - 1.$$

Problem 4 (for study—which means that the authors don't know the solution). Consider more general transformations instead of Π (for example,

$$1 \to \underbrace{11\ldots1}_{p\text{ ones}}\underbrace{00\ldots0}_{q\text{ zeros}}$$

and

$$0 \to \underbrace{11\ldots1}_{r\text{ ones}}\underbrace{00\ldots0}_{s\text{ zeros}}$$

and similar transformations) and classify the corresponding good sequences.

The other direction is to consider lines with other irrational slopes k and study polygonal paths that press up against these straight lines. It turns out that the structure of these polygonal paths depends on how the number $k + 1$ is represented as a *continued fraction*. Here is an example of a continued fraction:

$$\sqrt{2} + 1 = 2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ldots}}}.$$

Similarly, $k + 1$ can be written as

$$n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \cfrac{1}{\ldots}}},$$

where n_i are natural numbers. The study of continued fractions is a separate and interesting topic. We can't discuss it here due to space limitations. For a detailed discussion of this subject, see the article by Y. Nesterenko and E. Nikishin in the January/February 2000 issue of *Quantum*.

Here's the last problem.

Problem 5. Examine the structure of polygonal paths that press up against straight lines with an irrational slope k. For example, it's clear that such polygonal paths can be broken down into sections of the form

$$1\underbrace{00\ldots0}_{n_1\text{ zeros}}$$

and

$$1\,\underbrace{00\ldots0}_{n_1-1\text{ zeros}},$$

where $n_1 = [k + 1]$. What happens if every section

$$1\underbrace{00\ldots0}_{n_1\text{ zeros}}$$

is replaced by 1 and every section

$$1\,\underbrace{00\ldots0}_{n_1-1\text{ zeros}}$$

is replaced by 0?

The problem discussed in this article was solved by Alexander Sidorov, then a student at Moscow School No. 57. He presented a report on this problem at the 19th Student Conference in Batumi (November 1988). On July 1, 1990, his life came to a tragic end. We dedicate this article to his memory.

The price of resistance

by S. Betyayev

THE GREAT FRENCH SCIENtist Henri Poincaré (1854–1912) once said: "The experiment is the only source of truth: nothing else can teach us something new, and only it can equip us with trustworthiness." And we can't help agreeing—experimentation is truly the cornerstone of every science.

Experiments great and small

On the basis of the material expenditures on experimental research, one often sees science divided into "big science" and "small science." Research in small science is carried out by small research groups, or even by an individual, with little money. Big science involves the realization of grand projects—building huge particle accelerators, or space vehicles, or nuclear power plants.

Without big science it would be impossible, for instance, to build airplanes, ocean tankers, and rockets. The design of each arose out of industrial research conducted in modern aerodynamic test facilities. According to the principle of relativity, there are two equivalent methods of creating flow: moving a model in a medium that is at rest, or pushing a medium past a model that is at rest. The first method is used in ballistics facilities, hydrochannels, and experimental tanks, while the second is used in hydro- and aerodynamic tunnels.

Figure 1. *A MiG-29 jet fighter in a wind tunnel.*

In a hydrochannel, a special cart moves along rails set above the channel and pulls a model of a hydro-airplane; the flow around it is photographed, and the forces and moments acting on the model are measured. Aerodynamic tunnels, like the one shown in figure 1, have made it possible to move from primitive airplanes to modern airliners in a single century.

However, many physical laws can be discovered in small-scale experiments. So-called "kitchen experiments" belong to the category of scientific (as opposed to technological) research. Most of them can be done at home or in a school physics lab, as the American physicist Robert Wood said, with "a stick, a rope, sealing wax, and mica." The German physicist Hermann Helmholtz once said of Michael Faraday that "old pieces of wire, wood, and iron are seemingly all he needs to make great discoveries."

Great "ball tossers"

The easiest way to determine the resistance acting on a body moving relative to a medium is to watch how balls fall in the medium.

Leonardo da Vinci (1452–1519) was one of the first researchers who carried out such experiments. However, he experimented not only with falling objects but also with objects moving in water, and even with flat surfaces moving in air at an angle of attack.

Galileo Galilei (1564–1642) continued Leonardo's work. Throwing heavy and light balls from the leaning tower at Pisa, he discovered that the speed of a falling object is independent of its weight and formulated one of the great physical principles—the principle of inertia: *if no net force acts on a body, it moves uniformly in a straight line*. Galileo attached great significance to the logical explanation of experimental results, to an understanding of the physical essence of a phenomenon.

"Nature gave us eyes to see her works," he told his students. "But she also gave us a brain capable of understanding them."

Another great "ball tosser" was Sir Isaac Newton (1643–1727), the founder of classical physics and, together with Gottfried Leibniz (1646–1716), the founder of calculus. He threw balls in St. Paul's Cathedral in London. For Newton the question of moving objects and resistance wasn't a matter of idle speculation. He wanted to prove that (contrary to the assertions of the Aristotelians) interstellar space is not filled with matter. If it were, the matter would resist the motion of the heavenly bodies, and Newton's elegant mechanical system of the Universe would collapse like a house of cards.

According to Newton, the resistance of an object moving in water consists of three parts: the first part is constant, while the second and third are proportional to the speed and the square of the speed, respectively. The constant part of resistance is negligible; the resistance proportional to velocity is due to friction; and the resistance proportional to the square of velocity is due to the forces of inertia.

Today it's known that it's impossible in general to subdivide resistance into separate parts, because the effects related to flow around a body are woven together too tightly. Nevertheless, it was Newton who discovered that the force of resistance of a ball moving with speed v is directly proportional to its cross-sectional area S and fluid density ρ:

$$F = C_D \frac{\rho v^2}{2} S,$$

where C_D is the coefficient of proportionality (often called the drag coefficient).

What happens if the flow conditions are different? For example, we can spin an object about its vertical axis. Which ball will hit the ground first—one that is spinning or one that isn't? It turns out that rotating balls fall more quickly—you can demonstrate this on your own by throwing spinning and nonspinning tops from a tall building.

Figure 2. *Tsiolkovsky's wind tunnel.*

The law saying that resistance is less the faster an object spins is qualitative. Quantitative measurements of this phenomenon are fraught with difficulties. Because of the resistance force, there is a loss of momentum in the *trail* behind the object. In a way, the flow "remembers" how the trail was formed.

If the speed of a falling object doesn't vary along some portion of its trajectory, the resistive force remains the same. We know that speed is constant if the sum of all the forces acting on the object is zero; thus the force of resistance equals the force of gravity. For the same object this equality is reached at greater heights in air than in water. It seems that Newton took this feature into account when he chose water as a resistive environment in his ball-tossing experiments (he used a wooden barrel 4.5 meters high filled with water).

Tsiolkovsky's blower

In the 19th century the idea arose that the forces of air resistance (aerodynamic drag) could be measured in wind tunnels.

In Russia the first such tunnel was built by Konstantin E. Tsiolkovsky in 1897. He used a hand-driven winnowing machine to construct his blower. Unfortunately, the wind speeds generated in Tsiolkovsky's experiments were too low for use in aviation research. However, this kind of blower is very useful in carrying out many laboratory experiments on air resistance.

You can make a wind tunnel out of an ordinary vacuum cleaner by connecting a tube to its exhaust vent. You then place the object being tested in the stream of air coming out of the tube.

It's interesting that the laws of air resistance are manifest in various ways. For example, an airplane should have low drag but large lift. On the other hand, a parachute should have high air resistance to slow its fall.

Experiments with objects having a simple shape can be quite instructive. For instance, by blowing air or water on dihedrals with different opening angles β but the same face width l, we find that the highest drag is reached not for the flat plate ($\beta = 0$) but for a dihedral with some negative opening angle β (figure 3).

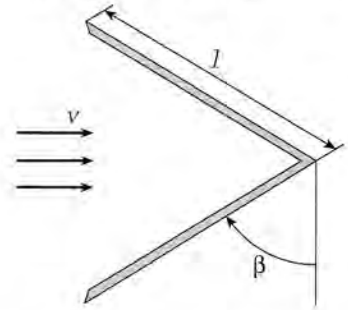


Figure 3. *Air flow around a dihedral.*

Along the same lines, we might think about airplane wings with a swept-forward shape (figure 4)—in the plane perpendicular to the air movement, the flow is similar to that around a dihedral with $\beta < 0$, and lift will correspond to the dihedral's drag. But now we've moved into the realm of "big science"...

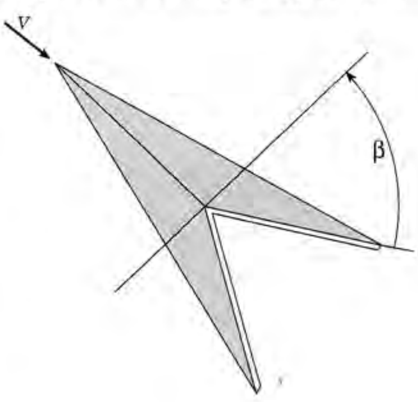


Figure 4. *Aircraft wing with forward-swept V shape.*

Field pressure

by A. Chernoutsan

GLANCING AT THE TITLE OF this article, you might think it's about the pressure of light (which is an electromagnetic field). You may have come across this phenomenon in your physics textbooks. You can grasp the nature of light pressure if you keep in mind the dual nature of light, which can be seen both as a wave and as a stream of particles (photons) that have momentum. When the photons are reflected or absorbed by an object, their momentum changes, which means that some force is exerted on the object.

This approach treats light pressure like the pressure of an ideal gas, which is explained within the framework of the kinetic theory as molecules striking the walls of a vessel. But if we try to explain light pressure without leaving the confines of electromagnetic theory, the creation of forces must be linked to the action of the wave's magnetic field on the uniformly moving charges of the substance, which is induced by another component of the wave—its electric field.

However, we won't get bogged down in a discussion of the pressure of electromagnetic waves, because this article is about something else—the pressure of a *static* field (either electric or magnetic). Clearly in this case there can be no talk of changes in momentum, so the very term "pressure" must be taken with a grain of salt. Nevertheless, this concept crops up in popular science magazines and books. For example, reading about the generation of extremely strong magnetic fields, we learn that one of the basic problems is the pressure this field exerts on the walls of the solenoid. This problem is closely related to controlled thermonuclear fusion, where the hot plasma must be held by a strong magnetic field (the "magnetic bottle"). However, let's begin not with magnetic field but with our old friend, the electric field.

Electric field pressure

We start with the case where pressure is exerted by an electric field on a charged surface. It arises if the field strength is different on each side of the surface. As always, the simplest case is a charged parallel-plate capacitor. The electric field inside such a capacitor is

$$E = \frac{\sigma}{\varepsilon_0},$$

where $\sigma = q/S$ is the surface charge density. When we calculate the force affecting a unit area of one of the plates, we need to take into account only the field generated by the other plate ($E/2$), because a plate doesn't act on itself:

$$P = \frac{F}{S} = \frac{E}{2}\sigma = \frac{\varepsilon_0 E^2}{2}.$$

Let's examine this result.

First, pressure is expressed as the intensity of the field located on one side of a plate (the field outside the capacitor is negligible). Second, the force affecting the plate is directed into the capacitor—that is, the plates are mutually attracted. This means that if we want to ascribe pressure to the electric field, we should treat this pressure as negative. In other words, the electric field doesn't "push" an object, it "pulls" it. Third, the field pressure is equal to the volume density of the electric field energy U stored by the capacitor:

$$P = -u = -\frac{U}{V} = -\frac{\varepsilon_0 E^2}{2}. \qquad (1)$$

These features of an electric field can be illustrated by means of energy conservation. Consider a parallel-plate capacitor isolated from any electrical source. Using an external force, let's slowly increase the distance between its plates by some value x. Since the strength of the electric field between the plates is constant (it's determined entirely by σ), the energy of the field will increase by $uS\Delta x$. So the external force must perform positive work $F\Delta x$, while the force of the field pressure must perform negative work $-PS\Delta x$. Therefore, the pressure of the electric field is negative, and its magnitude is equal to the volume density of the electrical energy.

Formula (1) is valid also in the case of a charged surface of any shape, provided the field strength is zero on one of its sides. Here's an important example: a fragment of

Art by Leonid Tishkov


The magnetic yoke: press and pull"
Тишков 2011

the surface of a conductor with area ΔS, at which the electric field is E, is acted on from the outside by a force $\Delta F = (\varepsilon_0 E^2/2)\Delta S$. Without proving this formula, let's work out the general result: If the electric fields on two sides of a charged surface are E_1 and E_2, then a force acts in the direction from the first to the second region, which corresponds to the pressure

$$P = \frac{\varepsilon_0 E_2^2}{2} - \frac{\varepsilon_0 E_1^2}{2}. \qquad (2)$$

This formula can be proved in three ways. The most simple and natural way uses energy. We mentally shift the surface a distance Δx and equate the work of the external force to the change in the energy of the electric field. (The work due to the pressure generated by the electric field is equal to the work performed by the external force but with the opposite sign).

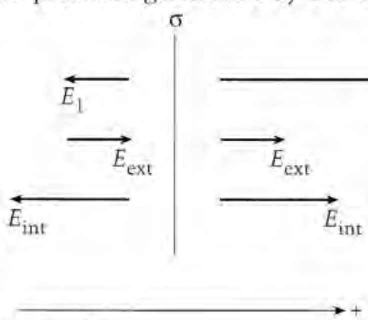


Figure 1

It's also possible, as in the case of the parallel-plate capacitor, to separate the external and internal fields (figure 1). Assume that both fields are perpendicular to the charged surface. In fact, the tangential component of the electric field (if it exists) must be identical on both sides of the charged surface (this follows from the conservative nature of the electrostatic field—prove it on your own), so it cancels out from the pressure formula. For the internal field E_{int} and external field E_{ext} we get

$$\begin{cases} E_1 = E_{\text{ext}} - E_{\text{int}}, \\ E_2 = E_{\text{ext}} + E_{\text{int}}. \end{cases}$$

The electric field near the surface is indistinguishable from the field generated by the plate—that is, $E_{\text{int}} = \sigma/2\varepsilon_0$. Both equations yield E_{ext} and σ:

$$E_{\text{ext}} = \frac{E_1 + E_2}{2}, \ \sigma = \varepsilon_0(E_2 - E_1), \quad (3)$$

from which we get the pressure

$$P = \sigma E_{\text{ext}}$$

and formula (2).

To make it clearer that it is the total field that determines the pressure, while the separation of the field into external and internal components is only a mathematical trick, let's consider the force affecting a thin layer of a volume charge (figure 2).

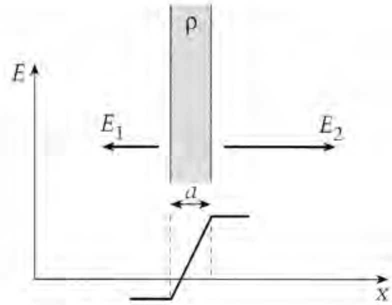


Figure 2

Inside the layer the field strength gradually changes from E_1 at one surface to E_2 at the other. If the volume charge density ρ is constant, the field strength varies linearly and the force affecting a segment of area S can be expressed by the mean field intensity:

$$F = \sigma S \frac{E_1 + E_2}{2}$$
$$= \varepsilon_0(E_2 - E_1)S\frac{E_1 + E_2}{2}$$
$$= \left(\frac{\varepsilon_0 E_2^2}{2} - \frac{\varepsilon_0 E_1^2}{2}\right)S,$$

where $\sigma = \rho a$ is the charge of a unit surface layer. The relationship between σ, E_1, and E_2 can be deduced by means of Gauss's law (see the list of references below) or by considering the external and internal fields, as we did in deducing formula (3).

If the function $\rho(x)$ is arbitrary, we divide the layer into many very thin layers of thickness dx and add the forces affecting these layers (figure 3):

$$F = \int_0^a E(x)\rho(x)S\,dx.$$

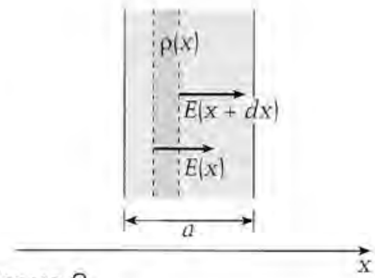


Figure 3

The change in the field strength at the next layer is

$$dE = \frac{\rho\,dx}{\varepsilon_0},$$

which is derived from Gauss's law and formula (3). For the pressure we get

$$P = \frac{F}{S} = \int_1^2 E\varepsilon_0\,dE = \frac{\varepsilon_0 E_2^2}{2} - \frac{\varepsilon_0 E_1^2}{2}.$$

Note an important feature here: In the case of a volume charge we need not separate out the internal field. The reason is that by decreasing the thickness of the layer, its internal field tends to zero.

Magnetic field pressure

In the case of a magnetic field we encounter two difficulties. One is purely pedagogical. Textbooks usually don't give the formulas for a magnetic field generated by an element of a current-carrying conductor (the Bio–Savart relation) or electric current flowing in a straight wire or a coil (solenoid). They also don't give the formula for the density of the magnetic field energy.

Therefore, we'll restrict our discussion to the case of a long solenoid (all our generalizations will proceed as in the case of electric field pressure). The solenoid's magnetic field is uniform almost everywhere (except at the ends) and is equal to

$$B = \mu_0 I \frac{N}{l} = \mu_0 i, \qquad (4)$$

where μ_0 is the permeability of free space, l is the length of the solenoid, N is the number of turns, and i is the surface current density (current per unit length), which is similar to the surface charge density in electrostatics. The direction of the magnetic field can be found from the right-hand rule. The magnetic flux in the solenoid is $\Phi = NBS$, from which we can find both the inductance $L = \Phi/I$ and the magnetic field energy stored by the solenoid $U = LI^2/2$. Dividing the energy by the solenoid's volume gives us the formula for the density of the magnetic field energy:

$$u = \frac{U}{V} = \frac{B^2}{2\mu_0}.$$

The second difficulty arises from the careless application of energy relationships in problems involving a magnetic field. If you're not careful, you may run into apparent paradoxes and contradictions. And that's exactly what threatens to occur here.

Our calculation of the force affecting a small rectangular area ΔS of a solenoid's surface will be based on

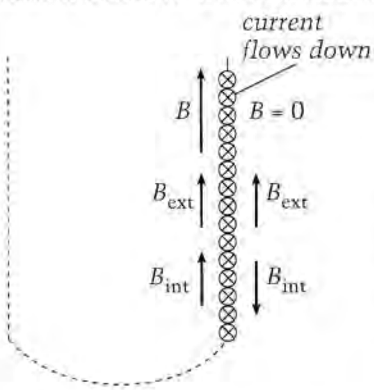


Figure 4

reasoning that is similar to that used in electrostatics (figure 4). We subdivide the magnetic field near the surface into an internal field B_{int} (very near the surface it should be equal to the field generated by an infinite current-carrying plane) and an external field B_{ext} (the field generated by the other parts of the solenoid). Thus we have

$$\begin{cases} B = B_{ext} + B_{int}, \\ 0 = B_{ext} - B_{int}, \end{cases}$$

from which we get $B_{ext} = B_{int} = B/2$. We can obtain the magnitude of the force from Ampere's law (taking formula (4) into account):

$$\Delta F = B_{ext}(i\Delta l)\Delta d = \frac{B^2}{2\mu_0}\Delta S,$$

where Δl is the width of a segment along the solenoid's axis and Δd is its length in the direction of the current.

At first glance, everything is okay and is similar to what we saw in electrostatics: the field pressure is numerically equal to the density of the magnetic field energy. However, if we determine the direction of the force according to the right-hand rule, we shall see a significant difference: the force is directed outward, so that in contrast to the electric field, the pressure of the magnetic field must be considered positive.

It would seem there's nothing wrong with this result—it agrees better with conventional views of pressure. However, it's not hard to see that this result is at odds with energy conservation. Indeed, if we mentally move the current-carrying surface, say, in the direction of the field (that is, if we decrease the radius of the solenoid), the external forces perform positive work against the magnetic forces, while the volume of the solenoid containing the field decreases—which means the energy of the field decreases as well! How can we resolve this contradiction?

The contradiction arose because we didn't take into consideration the work performed by the source of electric energy needed to maintain direct current in the solenoid—without this condition the value of the magnetic field in the solenoid cannot be held constant. The extra work of the source must compensate for the work of the induced emf arising due to the decrease in the magnetic flux in the solenoid. Moving the segment a distance Δx corresponds to a change in magnetic flux

$$\Delta\Phi = -B\Delta x\Delta d,$$

which generates in this segment an induced emf equal to

$$\mathscr{E}_{s\text{-}i} = -\frac{\Delta\Phi}{\Delta t}.$$

Taking formula (4) into account, the work performed by the source against the induced emf is

$$W_s = -\mathscr{E}_{s\text{-}i}\Delta q = -\frac{B\Delta x\Delta d}{\Delta t}i\Delta l\Delta t$$
$$= -\frac{B^2}{\mu_0}\Delta V,$$

where Δq is the electric charge that passed through the segment during time Δt. Now we see that the work performed by the external force and the source is exactly equal to the change in the energy of the magnetic field!

It's interesting to note that a similar problem arises in formulating energy conservation (the first law of thermodynamics) for an isobaric change in an ideal gas, where the pressure is also positive. It's convenient to consider the isobaric process because in this case the density of the internal energy remains constant:

$$u = \frac{U}{V} = \frac{nC_VT}{V} = \frac{C_V}{R}p,$$

where C_V is the molar heat capacity of the gas at constant volume. For example, compression of the gas is performed by the positive work of an external force. (Correspondingly, the work performed by the gas is negative). However, the internal energy of the gas decreases during this process. This "paradox" can easily be explained: the amount of heat transferred from the gas is exactly equal to that needed to satisfy energy conservation. In this case, the heat reservoir plays the same role as the electrical source in the solenoid problem.

Quantum on electromagnetic phenomena:

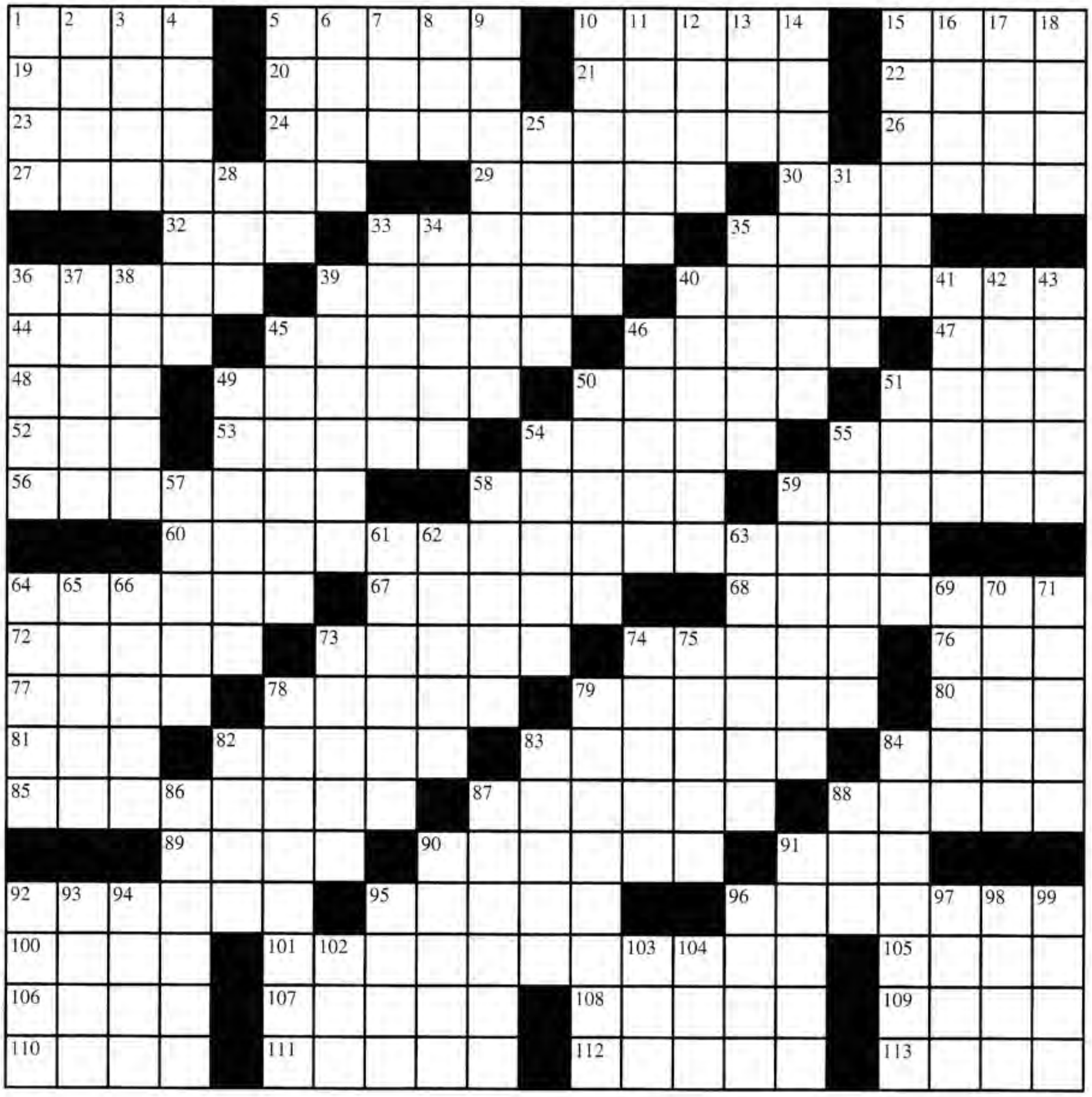
T. D. Rossing and J. R. Hull, "Magnetic Levitation Comes of Age," March/April 1995, pp. 22–27.

D. Tselykh, "Magnetic Fieldwork," September/October 1998, pp. 46–47.

A. Stasenko, "A Rotating Capacitor," May/June 1999, pp. 34–36.

Criss × cross science

by David R. Martin



ACROSS

1 Large fish hook
5 Assert
10 Pivoted
15 Complex silicate
19 More stable isomer: pref.
20 Archaeologist ___ Breuil (1877–1961)
21 Actress ___ Wasserstein
22 Type of exam
23 Brew
24 **F** = *m*a discoverer
26 Medicine amount
27 A solvent
29 Cowboy show
30 Representatives
32 Nucleotide chain: abbr.
33 Homes
35 Actor ___ Sharif
36 Aquatic plants
39 Annapolis freshmen
40 Three-sided figure
44 Beginning of the day
45 Town in Oman
46 Spoiled children
47 Phase difference
48 Dutch town
49 Sea duck
50 Hackneyed
51 Tennis score
52 Sheep cry
53 Actresses Mulgrew and Winslet
54 Angry
55 Oscillations
56 1945 chem. Nobelist ___ Virtanen (1895–1973)

58 Cold: comb. form
59 Most rational
60 Like light
64 Cellist Pablo ___ (1876–1973)
67 Like the countryside
68 Sideways
72 Group of double salts
73 Jerks
74 Daunt (archaic)
76 Silver iodide: abbr.
77 Carpets
78 Records
79 Acted dramatically
80 ___ product (or inner product)
81 Winglike structure
82 African capital
83 Town near Jaicos
84 Employ
85 ___ acid HOOC-$(CHOH)_2$COOH
87 Harmony
88 Type of eclipse
89 Spheres
90 Italian river
91 Flat coral island
92 Shout of exultation
95 Zodiac sign
96 Trig. function
100 Hydrated amorphous silica
101 Double-helix Nobelist
105 Dammed German river
106 Calcium oxide
107 966,350 (base 16)
108 1963 Chem. Nobelist Giulio ___
109 Columbus' ship
110 About 4.9289 cm^3: abbr.
111 Hindu gateway
112 Dismiss
113 Plant part

DOWN

1 g-aminobutyric acid: abbr.
2 Actor Baldwin
3 Escape
4 Computer language
5 Far east
6 ___ majesty
7 Collection of anecdotes
8 Bank account
9 One dyne/cm^2
10 Stockholm residents
11 City in Sussex, Del.
12 Poet's to
13 Tokyo, formerly
14 An explosive
15 Up-to-date
16 Element in steel
17 Use a fishing pole
18 Anthropologist ___ Hrdlicka (1869–1934)
25 Points of minimum disturbance
28 Identity element of a set
31 Gallium arsenide
33 Portland cement clinker constituent
34 Former Czech. president
35 Speak pompously
36 One-celled animal
37 Direction finder
38 ___ circle (equator, e.g.)
39 Like the top layer of oceans
40 Neptune's moon
41 Type of box
42 Washes
43 Excrete
45 Startles
46 Crystal reflection physicist
49 Skeletal parts
50 Courtroom event
51 Spear
54 La Douce et. al.
55 Stayed in anticipation
57 Bulls and Suns, e.g.
58 Knox and Sumter
59 Declares
61 ___ of Cancer
62 Reigns
63 Happy
64 0.200 grams
65 Bird wing part
66 Saccharide
69 Anthropologist Paul ___ (1883–1959)
70 Marketplace, in ancient Greece
71 Unit of volume
73 Singer ___ Joplin
74 Son of Lot
75 1972 Chem. Nobelist Stanford ___
78 Type of engine
79 $C_{20}H_{42}$
82 Poi source
83 Common fastener
84 ___' principle (of wave propagation)
86 ___'s aldehyde test
87 Woken up
88 Computer network: abbr.
90 Betel nut palm
91 The auditory ___
92 Pediatrician Luther ___ (1855–1924)
93 Bee genus
94 ___ black (carbon pigment)
95 USA word: abbr.
96 Type of bag
97 Revise and correct
98 Hawaiian goose
99 Ore cart
102 ___ blood group
103 Sales or income follower
104 1 atm. and 0°C: abbr.

SOLUTION IN THE NEXT ISSUE

SOLUTION TO THE JULY/AUGUST PUZZLE

A	S	A	P		B	O	S	H		S	O	U	S	A
N	O	D	E		E	D	I	E		I	N	N	E	R
A	D	A	R		T	O	O	L		N	O	B	L	E
L	A	M	I	N	A	R		I	K	E		A	Y	N
			G	A	S		Q	U	O		A	R	E	A
B	A	K	E	S		S	U	M	N	E	R			
R	E	N	E		A	R	A		R	A	M	J	E	T
A	B	A		E	L	I	N	V	A	R		O	R	E
D	I	P	O	L	E		T	A	D		I	L	I	A
			L	I	T	M	U	S		A	S	T	E	R
B	O	N	D		T	E	M		G	I	O			
A	V	A		F	E	S		F	E	R	M	I	O	N
R	A	D	A	R		O	P	E	N		E	L	S	A
G	R	I	M	E		N	O	T	E		R	Y	A	N
E	Y	R	I	E		S	T	A	T		S	A	R	D

ANSWERS, HINTS & SOLUTIONS

Physics

P301

Let's denote the mass of the cord by m and the tension at points A and B by $\mathbf{T}_A$ and $\mathbf{T}_B$, respectively. Keep in mind that at every point the direction of the tension is tangential to the cord. Since the cord is at equilibrium, the total sum of the forces acting on the cord is zero:

$$\mathbf{T}_A + \mathbf{T}_B + m\mathbf{g} = 0.$$

The vector character of this equation means that the sum of the projections of these forces in any direction is also zero. In particular, for the vertical axis OY we have:

$$T_{Ay} + T_{By} - mg = 0,$$

from which we get

$$m = \frac{T_{Ay} + T_{By}}{g}.$$

Therefore, to determine the mass of the cord, we need to find T_{Ay} and T_{By}.

The segment CB of the cord is in equilibrium. This means that the sum of the projections onto the horizontal axis OX of the forces acting on the cord at points C and B is zero:

$$T_{Cx} + T_{Bx} = 0.$$

Now draw the vector force $\mathbf{T}_C$ on an arbitrary scale (figure 1a). Its direction is tangential to the cord at point C. Let's find the projection of this force onto the X-axis and draw the tangent to the cord at point B. The force $\mathbf{T}_B$ is directed along this tangent line. The next step is to draw the horizontal component BK of the force T_{Bx}. This projection is equal to $-T_{Cx}$. Drawing a vertical line at point K allows us to determine the force $\mathbf{T}_B$. Now it's easy to find the projection of the force $\mathbf{T}_B$ onto the vertical axis. It's determined by the length of the segment KE measured in the scale we chose for drawing the force $\mathbf{T}_C$. In our case $T_{By} \cong 22$ N.

In a similar way we plot the tension $\mathbf{T}_A$ at point A (figure 1b) and find its projection onto the Y-axis. In our case it is about 9 N. Therefore,

$$m = \frac{22+9}{9.8}\ \text{kg} \approx 3.2\ \text{kg}.$$

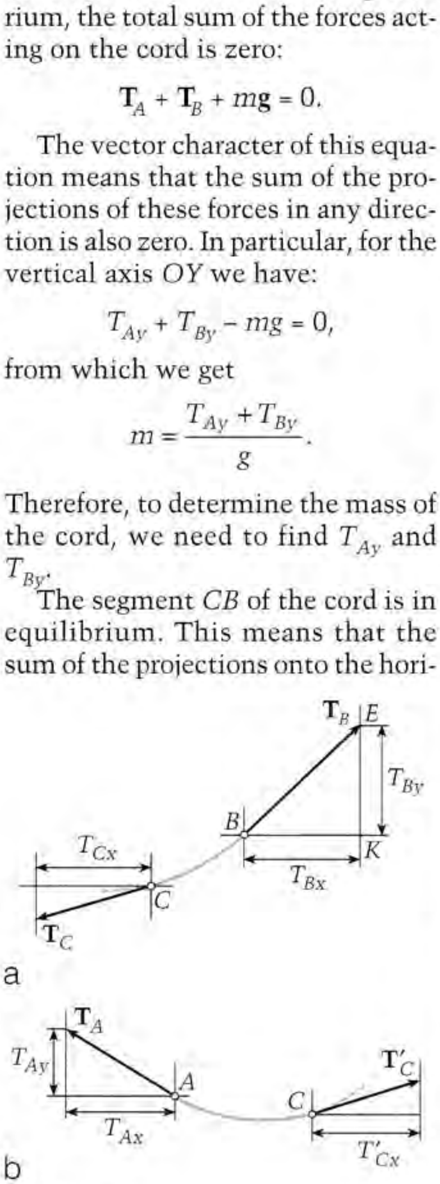


Figure 1

P302

Let's consider the process in the (P, V) plane and denote the states with temperatures T_1 and T_2 by points 1 and 2, respectively (figure 2). Now draw isothermal and adiabatic curves through these points and label the intersections 3 and 4. By the statement of the problem, the process $1 \to 2$ is possible, during which the temperature does not decrease and heat is not lost. This means that point 2 is situated to the right of the adiabatic curve that passes through point 1. In addition, the plot of an arbitrary process $1 \to 2$ satisfying the given conditions is located within the cycle $1 \to 3 \to 2 \to 4 \to 1$.

Denote by Q_{12}, Q_{132}, and Q_{142} the heat added to the gas during the processes $1 \to 2$, $1 \to 3 \to 2$, and $1 \to 4 \to 2$, respectively. Let's consider the process $1 \to 3 \to 2 \to 1$. In this process the gas acquires heat Q_{132}, then dissipates heat Q_{12}, performing work $Q_{132} - Q_{12}$ as it does so. The work is equal to the area of the figure delimited by the lines $1 \to 3$, $3 \to 2$, and $2 \to 1$. Since this area is not negative,

$$Q_{132} \geq Q_{12}.$$

In a similar way we consider the process $1 \to 2 \to 4 \to 1$. The work performed by the gas in this process is also nonnegative and equal to $Q_{12} - Q_{142}$, from which we get

$$Q_{12} \geq Q_{142}.$$

These two inequalities lead to

$$Q_{142} \leq Q_{12} \leq Q_{132}.$$

Since the process $1 \to 2$ is arbitrary (it is one of the processes that satisfies the conditions of the problem), the last inequality means that the minimum amount of heat Q_1 that can be imparted to the gas in this process is Q_{142}. The maximum heat Q_2 that can be transferred to the gas in this process is Q_{132}, by the same reasoning.

Therefore, to solve the problem, we must consider the Carnot cycle

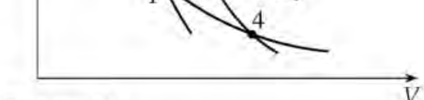


Figure 2

$1 \to 3 \to 2 \to 4 \to 1$, where the efficiency is

$$\eta = 1 - \frac{Q_{142}}{Q_{132}} = 1 - \frac{T_1}{T_2}.$$

Taking into consideration that $Q_{142} = Q_1$ and $Q_{132} = Q_2$, we obtain

$$Q_2 = Q_1 \frac{T_2}{T_1}.$$

P303

Electrolysis of water is accompanied by polarization of the electrodes, which means that the cell turns into a galvanic cell. The emf of this cell can be found by assuming that the energy released by the hydrogen combustion is not used to perform any work against external electrical forces. In accordance with Faraday's electrolysis law, to obtain hydrogen of mass m, an electric charge q must pass through the cell:

$$q = \frac{mZeN_A}{M},$$

where $Z = 1$ is the valence, $M = 1$ g/mol is the atomic mass of hydrogen, $e = 1.6 \cdot 10^{-19}$ C, and $N_A = 6.02 \cdot 10^{23}$ 1/mol is Avogadro's number. As noted above, work equal to Q must be performed in this process:

$$\mathcal{E} = \frac{Q}{q} = \frac{QM}{mZeN_A}.$$

This results in the following cell efficiency (which is equal to the efficiency of the corresponding galvanic cell and therefore of the electrolysis cell):

$$\eta = \frac{A_{\text{obtained}}}{A_{\text{consumed}}} = \frac{q\mathcal{E}}{qU} = \frac{QM}{mZeN_AU} \approx 0.75 = 75\%.$$

P304

First let's find the equilibrium position of the jumper. Since both batteries have the same emf, the voltage drop between the midpoints of the rails is zero (figure 3a). If the jumper is at rest at the very center of the rails, no current flows in it, so it does not experience a magnetic force. Therefore, this is the equilibrium position of the jumper. We'll see later that this equilibrium is stable.

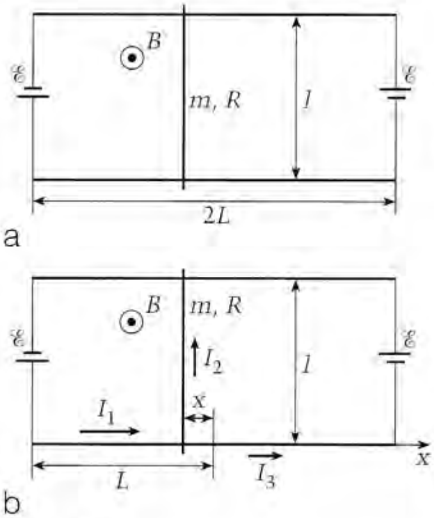


Figure 3

Note that if the jumper moves near the equilibrium position, it carries electric current generated by two sources: first, by the change in resistance in that part of the electrical circuit; and second, by the generation of electromagnetic inductance emf. According to Lentz's law, part of the magnetic force due to induced current damps the oscillations (it can be shown that this part of the magnetic force is proportional to the velocity of the jumper and is analogous to viscous friction). According to the statement of the problem, the induced emf can be neglected.

Let's fix a rectangular coordinate system to the midpoint of the lower rail and with the positive x-axis to the right (figure 3b). Consider a small shift of the jumper along the x-axis—say, to the left. If the jumper is shifted a distance x, it will carry an electric current. Denote the current flowing from the left battery to the jumper as I_1, the current carried by the jumper as I_2, and the current entering the right battery as I_3. According to Kirchhoff's first law,

$$I_1 = I_2 + I_3.$$

Now write Kirchhoff's second law for the circuit loop containing the left battery and the jumper and the loop containing the right battery and the jumper:

$$2I_1\rho(L - x) + I_2R = \mathcal{E},$$

$$2I_3\rho(L + x) - I_2R = \mathcal{E}.$$

Solving the system of three equations, we get the current flowing through the jumper:

$$I_2 = \frac{\mathcal{E}x}{\rho\left(L^2 - x^2\right) + RL} \approx \frac{\mathcal{E}x}{\rho L^2 + RL},$$

where we have dropped the second term in the parentheses due to the small amplitude of the oscillations.

Since the jumper is situated in the magnetic field, it experiences a magnetic force:

$$F_A = I_2lB = \frac{\mathcal{E}lBx}{L(\rho L + R)}.$$

This force is directed to the right and tends to restore the jumper to the equilibrium position. The equation of motion of the jumper is

$$ma_x = -\frac{\mathcal{E}lB}{L(\rho L + R)}x,$$

from which the period of small oscillations of the jumper can be obtained:

$$T = 2\pi\sqrt{\frac{mL(\rho L + R)}{\mathcal{E}lB}}.$$

P305

A beam striking the flat face of a semicylinder at point A with an angle of incidence $\alpha = 45°$ is refracted and then travels in the glass at an angle β (figure 4a) such that

$$\frac{\sin\alpha}{\sin\beta} = n,$$

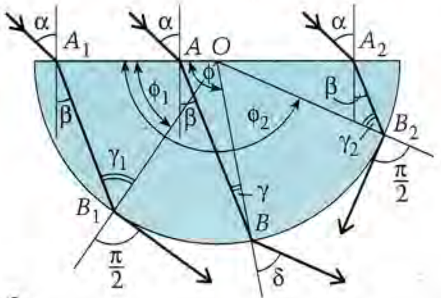
from which we get

$$\beta = \arcsin\frac{\sin\alpha}{n} = \arcsin\frac{1}{n\sqrt{2}}.$$

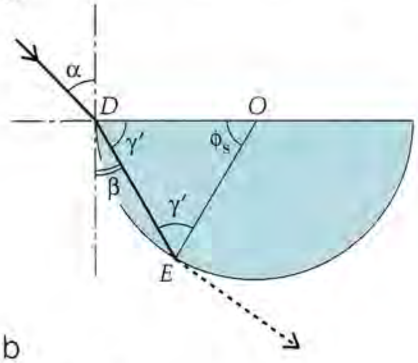
Then this beam strikes the curved surface at point B at an angle γ, which depends on the angle ϕ formed by the radius OB with the flat face of the semicylinder: $\gamma = \pi/2 + \beta - \phi$ (figure 4a). This beam emerges from the semicylinder at an angle δ, which satisfies the formula

$$\frac{\sin\gamma}{\sin\delta} = \frac{1}{n},$$

—that is,



a



b

Figure 4

$$\sin\gamma = \frac{\sin\delta}{n}.$$

The only rays that can escape the semicylinder are those that satisfy the condition $\delta \le \pi/2$ (figure 4a). Therefore, for these rays $\sin\gamma \le 1/n$. This means that the only rays emerging from the glass are those that strike the cylinder's surface between points B_1 and B_2, and for which $\phi_1 \le \phi \le \phi_2$. Since

$$\phi_1 = \frac{\pi}{2} - \gamma_1 + \beta,\ \phi_2 = \frac{\pi}{2} + \gamma_2 + \beta,$$

and

$$\gamma_1 = \gamma_2 = \arcsin\frac{1}{n},$$

we get finally

$$\frac{\pi}{2} + \arcsin\frac{1}{n\sqrt{2}} - \arcsin\frac{1}{n} \le \phi \le \arcsin\frac{1}{n} + \arcsin\frac{1}{n\sqrt{2}} + \frac{\pi}{2}.$$

Clearly the upper limit must be smaller than π. Therefore, for the values of n corresponding to $\phi_2 > \pi$, the upper limit for ϕ is π.

It's possible that for some value of n the lower limit for ϕ (that is, ϕ_1) corresponds to the shaded region, where no light ray falls. The boundary of the shade can be found from the trajectory of the ray striking the semicylinder at point D (figure 4b). Clearly the shaded region is subtended by the arc DE.

Since the triangle DEO is equilateral and has an angle at the base DE equal to

$$\gamma' = \frac{\pi - \phi_s}{2} = \frac{\pi}{2} - \beta,$$

it follows that

$$\phi_s = 2\beta = 2\arcsin\frac{1}{n\sqrt{2}}.$$

The lower boundary of the illuminated region is equal to ϕ_s at $\phi_s > \phi_1$—that is, when

$$2\arcsin\frac{1}{n\sqrt{2}} > \frac{\pi}{2} + \arcsin\frac{1}{n\sqrt{2}} - \arcsin\frac{1}{n},$$

or at

$$\arcsin\frac{1}{n\sqrt{2}} + \arcsin\frac{1}{n} > \frac{\pi}{2}.$$

Math

M301

Let d, q, and r ($0 < r < d$) be the divisor, quotient, and the remainder, respectively. Then

$$201 = dq + r. \qquad (1)$$

Since $r < d$, r must be either the first or second element of the geometric progression. Can it be the second? If it were, then we would have $dq = r^2$. Plugging this into equation (1) we get $201 = r^2 + r$. A little investigation will show that this equation has no integer solutions. Thus r cannot be second and so must be the smallest element of the geometric progression. Let the common ratio of this progression, expressed in lowest terms, be m/n. Then the number rm^2/n^2 must be an integer, so $r = kn^2$, where k is some natural number. Then d and q are equal to kmn and km^2 in some order. Therefore, equation (1) can be rewritten as

$$201 = (kmn)(km^2) + kn^2 = kn(km^3 + n). \qquad (2)$$

Thus n is a divisor of 201 and equals 1, 3, 67, or 201. The last two of these numbers give no solution. For $n = 1$, we have $201 = k^2m^3 + k$, and k is a divisor of 201. Trying all possible values, we see that this value of n doesn't work either. Therefore, $n = 3$. Dividing equation (2) by 3, we obtain the equation $67 = k^2m^3 + 3k$, whose solution is $k = 1$, $m = 4$. Thus the remainder is 9, and the divisor can be 16 or 12.

M302

Let $x = 42000$, $y = 58000$. We will show that the maximum distance is $5xy/2(x + y) = 60900$ km. After 1 km, the rear tires each have $1/x$ units of wear, where we define 1 unit to be the point at which the tire becomes unusable, and the front tires each have $1/y$ units of wear, so the total wear is $2/x + 2/y$. Since each tire can only take 1 unit of wear, the maximum wear over the entire trip is 5, so the maximum distance is at most $5/(2/x + 2/y) = 5xy/2(x + y)$.

To attain this distance, drive $xy/2(x + y)$ kilometers, then rotate the five tires cyclically, then drive $xy/2(x + y)$ kilometers, then rotate in the same order, and so on, stopping when we rotate back to the original positions.

Then each tire has been in front twice and in back twice, and its wear has been $(2/x + 2/y)(xy/2(x + y)) = 1$. (Solution by Gabriel Carroll)

M303

This situation is possible if n is odd, but not if n is even.

Suppose there are n players in the tournament. Since each player's score is an integer multiple of 1/2, and since all the scores are different, two consecutive scores must differ by at least 1/2. In particular, if we look at the scores after the first round, from lowest to highest, each differs by at least 1/2, and thus the highest score must be at least $(n - 1)/2$. This is true both for the first and the second rounds individually.

Let us refine this effort a bit. Let k be the number of points earned in the first round by the player with the lowest score (after the first round). Note that this player is destined to win the tournament. The argument of the preceding paragraph shows that the highest score in the first round cannot be less than $k + (n - 1)/2$.

The lowest possible high score in the first round is $k + (n - 1)/2$, and the lowest possible high score in the second round is $(n - 1)/2$, so the lowest total high score, the score of the winner of the tournament, cannot be less than the sum of these two, which is $k + (n - 1)$.

On the other hand, the winner of the tournament, who came in last in the first round, earned k points in that round, and could not have earned more than $(n - 1)$ points in the second round (by beating everyone else once), so he could not have earned more than $k + (n - 1)$ points.

It follows from the last two paragraphs that the winner of the tournament earned exactly $k + (n - 1)$ points. How did this happen? Well, the winner after the first tournament must have earned as little as possible (for a fixed k), which is $k + (n - 1)/2$ points, and the scores in that round must then all have differed by as little as possible, which is 1/2.

What was the sum of the points earned in the first round? On the one hand, it was just the number of games played, which is $n(n - 1)/2$. On the other hand, it was $k + (k + 1/2) + (k + 2/2) + (k + 3/2) + \dots + (k + (n - 1)/2) = nk + n(n - 1)/4$. Of course, these two computations must yield the same number, so $nk + n(n - 1)/4 = n(n - 1)/2$, or $k = (n - 1)/4$. This is certainly impossible for even n, since k must be an integer multiple of 1/2.

But is it always possible for odd values of n?

We will show that it is by an inductive construction. Let a_i be the score of the player ranked ith in the first round, and let b_i be the score of that player in the second round.

We start with three contestants. Our previous analysis shows that the scores in the first round must be bunched, and that they must sum to 3. We can achieve this by letting $a_1 = 1/2$, $a_2 = 1$, and $a_3 = 3/2$. If the players are A_1, A_2, and A_3, this means A_1 lost to A_2 and tied with A_3, while A_2 lost to A_3. Then we construct a second round by letting player i beat player j if $i < j$. This means that $b_3 = 0$, $b_2 = 1$, and $b_1 = 2$. The reader can construct the appropriate wins and losses, and check that the sums $a_i + b_i$ give the proper rank.

To construct a tournament with five players, we duplicate the 3-person tournament described above, then introduce players 1 and 2, and, for $i \geq 3$, let player i beat player 1 and draw player 2 if i is even, and vice versa if i is odd. Finally, let player 2 beat player 1. The result will be that $a_1 = 1$, $a_2 = 3/2$, $a_3 = 2$, $a_4 = 5/2$, and $a_5 = 3$.

For the second round, let player i beat player j whenever $i < j$. We then have $b_1 = 4$, $b_2 = 3$, $b_3 = 2$, $b_4 = 1$, and $b_5 = 0$, and the sums $a_i + b_i$ are in the correct order.

This construction generalizes. If we have a tournament set up for n people, we can get one for $(n + 2)$ people by setting up a first round among the 3rd, 4th, ..., $(n + 2)$nd players in which the lowest score is $(n - 1)/4$ (notice that this is an integer multiple of 1/2 when n is odd). Then, for $i \geq 3$, let player i beat player 1 and draw player 2 if i is even, or vice versa if i is odd. Finally, let player 2 beat player 1. Then, in the second round, player i beats player j whenever $i < j$; this gives $b_i = n - i$, and the scores are ranked as desired. (Solution by Gabriel Carroll)

M304

In solving this problem we'll use the following well-known fact: the distance from any vertex of a triangle to the point where its altitudes intersect (its *orthocenter*) is twice the distance from the center of the circumscribed circle to the side opposite this vertex.

This fact can be proved in many different ways. For example, let H be the point where the altitudes of triangle ABC intersect, and let O be the center of the circumscribed circle (see figure 5a). Denote by A_1, B_1, and C_1 the midpoints of the corresponding sides of triangle ABC. The triangle $A_1B_1C_1$ is similar to ABC with a similarity coefficient of

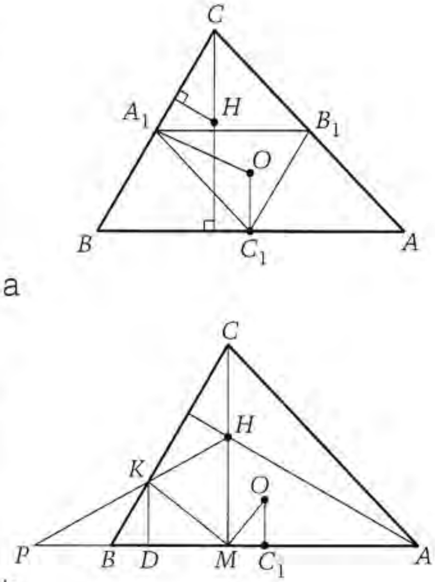


Figure 5

1/2. Now the perpendicular bisectors of the sides of triangle ABC contain the altitudes of triangle $A_1B_1C_1$, so O is in fact the orthocenter of triangle $A_1B_1C_1$. Therefore, we have $CH = 2C_1O$ (and two additional similar equalities).

Now let's return to our original problem. Let H be the point where the altitudes of the given triangle ABC intersect (figure 5b). Let line KH (which is given to be symmetric to AH about CM) intersect AB at point P, and let C_1 be the midpoint of AB. Denote by D the projection of K onto AB. We have $\angle HPM = \angle HAM = \angle HCK$ (the last equality follows from the fact that both angles are complementary to angle ABC if that angle is acute, and are complementary to the angle adjacent to ABC if it is obtuse). By the auxiliary proposition proven above, we have

$$CH = 2C_1O. \qquad (1)$$

Without loss of generality, we can assume that $AC > BC$ (the case when $AC = BC$ is easy). Then $AM > BM$. Since C_1 is the midpoint of AB, we have

$$\begin{aligned} MC_1 = BC_1 - BM &= (1/2)AB - BM \\ &= (1/2)(AM + MB) - BM \\ &= (1/2)(AM - MB). \qquad (2) \end{aligned}$$

Since $PM = AM$, we have

$$PB = PM - BM = AM - BM. \quad (3)$$

Next we show that triangles KHC and KBP are similar. Indeed: we have $\angle KPB = \angle HPM = \angle HAM = \angle HCK$ (the last pair are equal because they are both complementary to $\angle CBM$), and $\angle PKB = \angle CKH$ (they are a pair of vertical angles). Now the altitudes drawn to corresponding sides of similar triangles are proportional to those sides. Here, we see that KD is an altitude KD in triangle KBP, and DM is equal in length to the corresponding altitude in triangle KHC. So, from the similar triangles and from (3), (2), and (1), we have

$$\frac{KD}{DM} = \frac{PB}{CH} = \frac{MA - BM}{CH} = \frac{2MC_1}{2OC_1} = \frac{MC_1}{OC_1}$$

Thus the right triangles KDM and MC_1O are similar. We can now write (for the case shown in figure 1b) $\angle KMD + \angle OMC_1 = \angle KMD + \angle MKD = 90°$ and $\angle OMK = 180° - 90° = 90°$.

M305

Denote the fraction on the left-hand side of the equation by Q_n. Then

$$Q_n = \cfrac{1}{x_n + \cfrac{1}{x_n + Q_{n-1}}}. \quad (1)$$

Let us show that

$$Q_1 = \cfrac{1}{x_1 + \cfrac{1}{x_1}} \le \frac{1}{2}.$$

Indeed, it's well known that the inequality

$$x + \frac{1}{x} \ge 2 \quad (2)$$

holds for all positive x, and that equality occurs only for $x = 1$. We can write

$$Q_2 = \cfrac{1}{x_2 + \cfrac{1}{x_2 + Q_1}}.$$

Therefore,

$$\frac{1}{Q_2} = x_2 + \frac{1}{x_2 + Q_1} = (x_2 + Q_1) + \frac{1}{(x_2 + Q_1)} - Q_1.$$

(We use here the trick of adding and subtracting the quantity Q_1.) By inequality (2) we have

$$\frac{1}{Q_2} \ge 2 - Q_1 \ge \frac{3}{2}.$$

Thus,

$$Q_2 \le \frac{2}{3},$$

and the equality occurs only for $Q_1 = 1/2$ (that is, $x_1 = 1$) and $x_2 + Q_1 = 1$ (that is, $x_2 = 1/2$).

We now prove by induction that

$$Q_k \le \frac{k}{k+1},$$

and equality occurring only for $x_m = 1/m$, in which case

$$Q_m = \frac{m}{m+1}$$

for $m = 1, 2, \ldots, k$.

Proof: For $k = 1$ and 2, this inequality has already been proved (in fact, the inductive proof requires checking only $k = 1$). It remains to prove that if the inequality holds for k, it also holds for $k + 1$. It follows from equation (1) that

$$\frac{1}{Q_{k+1}} = x_{k+1} + \frac{1}{x_{k+1} + Q_k} = x_{k+1} + Q_k + \frac{1}{x_{k+1} + Q_k} - Q_k.$$

By inequality (2), we have

$$\frac{1}{Q_{k+1}} \ge 2 - Q_k,$$

and the equality occurs if

$$x_{k+1} + Q_k = 1. \quad (3)$$

By the inductive hypothesis, we have

$$Q_k \le \frac{k}{k+1}.$$

Therefore,

$$Q_{k+1} \le \frac{1}{2 - Q_k} \le \cfrac{1}{2 - \cfrac{k}{k+1}} = \frac{k+1}{k+2}.$$

Here, if

$$Q_{k+1} = \frac{k+1}{k+2},$$

then

$$Q_k = \frac{k}{k+1}$$

(which, by the inductive hypothesis, implies $x_m = 1/m$ for $m = 1, 2, \ldots, k$). In addition, we find from relation (3) that

$$x_{k+1} = 1 - Q_k = 1 - \frac{k}{k+1} = \frac{1}{k+1}.$$

The inequality is proved.

Answer: $x_1 = 1$, $x_2 = 1/2$, …, $x_n = 1/n$.

Brainteasers

B301

It's clear from the statement of the problem that Billy got more than 1/10 of all the fruit and less than 1/8. Therefore, he got 1/9 of all the fruit. So Mrs. Brown had nine grandchildren. This situation is possible, for example, if there were eight apples and ten pears in the bowl. In this case, eight children got one apple and one pear each, and one child got two pears. Other examples, with larger numbers of fruit, are also possible.

B302

To make things definite, let's say that the person who reported dancing to five songs is a boy. Then each girl danced to 3, 6, or 9 songs, so the sum of the numbers of dances declared by the girls is divisible by 3. But each time a girl danced, a boy (her partner) also danced. Hence the number of times the boys danced was also divisible by 3. But the total number of dances is $6 \cdot 3 + 1 \cdot 5 + 4 \cdot 6 + 1 \cdot 9 = 56$, and half of them, or 28, were danced by girls. Since 28 is not divisible by 3, someone must have made a mistake. (A similar argument holds if the person who declared 5 dances was a girl.)

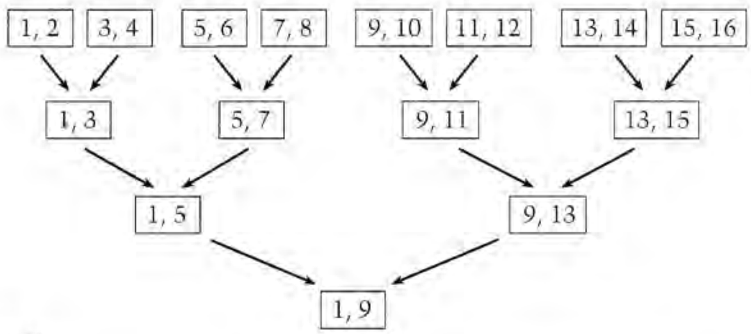


Figure 6

B303

Suppose there were no uninteresting games at all. Then the team with rank 1 played those with ranks 2, 3, 4, and 5. Therefore, these four teams didn't play each other. Thus one of these teams played three games with a weaker team before it was defeated by the champion; one played two games with a weaker team; and one played one game with a weaker team. In sum, they played 1 + 2 + 3 = 6 matches with teams ranked 6 through 16. Therefore, one of these teams played a team ranked 10 or lower, and this game was uninteresting. On the other hand, the tournament can be organized to have only one uninteresting game (see figure 6). Unfortunately this must be the championship game. (The organizers of real playoff series would certainly avoid this arrangement.)

B304

See figure 7. The folding necessary to create the cube from this pattern is left to the reader.

B305

The film reacts to red light as if it were darkness, so it cannot distinguish red from black. Therefore, the red skull on the white background will be black (in the print made from this negative), while the skull on the black flag will not be visible at all (that is, the entire flag will appear black).

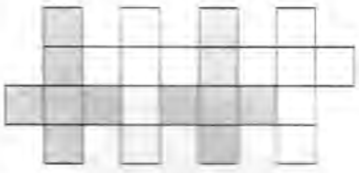


Figure 7

Fertilizer with a bang

1. Since the concentrations of the initial substances change only slightly (this is a case of "small conversion depth"), the reaction rate is constant. The time τ we seek is determined by the simple equation $w\tau = 10^{-2}n$, or

$$\tau = \frac{10^{-2}}{zn}e^{E/kT} = 7 \cdot 10^{-12} e^{E/kT}\ (\mathrm{s}).$$

Plugging in the values for the activation energy and temperature, we get

$$\tau_{273} = 7 \cdot 10^{19}\ \mathrm{s} = 2 \cdot 10^{12}\ \text{years}.$$

This is longer than the age of the Universe! It's natural to consider the gas mixture as inert at this particular temperature. At a temperature of 600 K the reaction period is $t_{600} = 7 \cdot 10^2$ s $\cong$ 10 minutes, which signifies a slow reaction. Finally, $t_{800} = 0.3$ sec, which indicates a rapid reaction.

Thus an increase in the absolute temperature of the mixture by a mere factor of two (from 273 K to 600 K) changes the rate of the chemical reactions by seventeen orders of magnitude!

2. If the graph of heat transfer is steeper than that of heat production, a random increase in temperature would result in the inequality $P_- > P_+$, so the system will be cooled and the temperature will return to the starting point. Similarly, a random temperature decrease results in the opposite inequality, so the system will be heated and again equilibrium will be restored. By contrast, if the slope of the graph for heat production is greater than that of the graph for heat transfer, a random change in temperature will increase—that is, the equilibrium point will be unstable.

3. (a) Varying the heat of combustion of a reaction changes the heat production function (the ordinate values are multiplied by a constant coefficient), while the straight line of the heat transfer function will not change. In figure 8, $Q_1 < Q_2 < Q_3$.

(b) Varying the heat transfer coefficient α will change the slope of the heat transfer line. In figure 9, $\alpha_1 > \alpha_2 > \alpha_3$.

Kaleidoscope

1. Since both objects move with the same acceleration, the distance between them will not change.

2. The weight of a person fully submerged in water is proportional to the acceleration due to gravity and to the difference in density of the person's body and the water. If we assume the properties of water

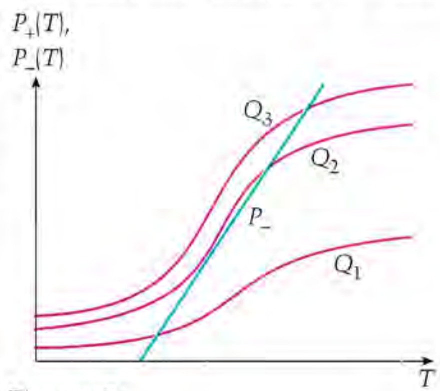


Figure 8

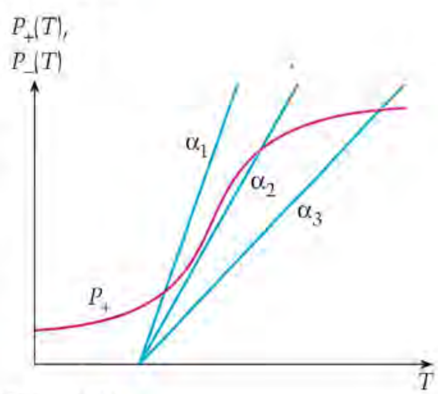


Figure 9

are the same on the Earth and Moon, it's easier to swim on the Moon, where the acceleration due to gravity is about 1/6 that on Earth.

3. Yes.

4. According to the law of universal gravitation, the attractive force decreases from mg to zero when an object moves from the Earth's surface to infinity. Therefore, in this case the object's weight decreases from $2mg$ at the Earth's surface to mg at infinity.

5. No.

6. After returning to Earth, astronauts sleep without pillows to compensate for the drastic redistribution of blood caused by the transition from weightlessness to normal gravity and thereby to provide the head with an adequate supply of oxygen. "Upside down" sleep models the conditions of weightlessness to which the astronauts were accustomed during a long-term mission.

7. Yes, because weightlessness doesn't affect the process of thermal expansion of a liquid.

8. No.

9. It is due to the spinning of the Earth about its axis of rotation.

10. In the large massive planets the force of gravity prevails over the elastic repulsive force. Therefore, it flattens any outcropping of the planetary landscape. By contrast, on asteroids and comets the force of gravity is negligible, so the shape of these small objects is determined by the processes of collision, adhesion, and destruction. This explains why small space objects have such irregular shapes.

11. The Earth is "flattened" slightly at the poles, so the path along the meridian is shorter than the equatorial route. Thus the polar traveler will be the first at the finish line.

12. The linear velocity of rotation at the equator must be equal to the escape velocity.

13. To paraphrase Richard Feynman's explanation of this phenomenon: The attraction of water and land by the Moon is balanced at the Earth's center. By contrast, the Moon more strongly attracts the water situated on the "lunar" side of the Earth than it attracts the Earth on average. Conversely, the attraction of water on the far side of the Earth is weaker than the mean attraction of the Earth. In addition, in contrast to solid rock, water can flow. The tides are caused by these two factors.

14. The tidal effect of the Moon is added to the tidal effect of the Sun.

15. In that ancient time (about two billions years ago), not only did lunar eclipses last longer, they occurred more frequently, because the lunar shadow covered a far greater area on the Earth than nowadays.

16. Due to the heterogeneity of the Sun's gravitational field, even on a spherically symmetrical planet that does not rotate, the acceleration due to gravity would be different at various points on the planet's surface.

Microexperiment

No. Your body is affected by the buoyant force, so the water is affected by your body's weight.

The WRITE stuff

by Don Piele

DURING THE THIRD WEEK IN JUNE, THE USA Computing Olympiad staff and sixteen high school participants from around the United States gathered at the University of Wisconsin—Parkside in Kenosha, Wisconsin for a week of instruction, games, cows, and competition. In the course of eight days, the finalists improved their programming skills, enhanced their frisbee throwing ability, expanded their gaming skills in Nine Men's Morris, and felt the full g-force of Raging Bull at Six Flags Great America. In the end, the strongest four computer programmers emerged as the USA 2000 Informatics Team. This team will represent the United States in the Central European Olympiad in Informatics in Romania in August, and in the International Olympiad in Informatics IOI'2000 in Beijing, China in September.

Free ride to Dairyland

To be invited to Wisconsin, high school students must show us that they have the WRITE stuff. They do this by entering the Fall, Winter, and Spring Internet competitions and the USA Open Competition in April. These competitions are offered in two divisions: Junior for beginners, and Senior for those at the top level who have a shot getting invited to Wisconsin. We look at all of the results and pick the top fifteen finalists, who are given a free ride to our eight-day training camp and informatics team selection competition.

Unfortunately, again this year, the field of competitors comprised only boys. It was a surprise that seven of the fifteen finalists came from one high school—Thomas Jefferson High School for Science and Technology in Alexandria, Virginia. The finalists from "TJ" were: Kevin Caffrey, sophomore; John Danaher, senior; Richard Eager, junior; and Vladimir Novakovski, Gregory Price, Gary Sivek, and Steven Sivek, all sophomores. Also invited to camp were: Reid Barton, Arlington, MA; Jacob Burnim, Silver Springs, MD; Adam D'Angelo, Redding, CT; Percy Liang, Phoenix, AZ; Jack Lindamood, Dallas, TX; Yuran Lu, Presque Isle, ME; Thuc Vu, Westminster, CA; and Tom Widland, Albuquerque, NM. Tomasz Czajka, a member of the Polish informatics team, joined the camp as a guest competitor.

Gender divide

What is it about competitive computer programming that makes it such a male-dominated activity? It is not something that is cultural, since all 65 countries that

Finalists of the USA Computing Olympiad. Top (left to right): John Danaher, Tom Widland, Vladimir Novakovski, Jack Lindamood, Steven Sivek, Yuran Lu, Tomasz Czajka, Percy Liang. Bottom (left to right): Jacob Burnim, Gregory Price, Kevin Caffrey, Adam D'Angelo, Richard Eager, Thuc Vu, Reid Barton, and Gary Sivek.

compete in the International Olympiad in Informatics (IOI) have teams that are almost exclusively male. There have been only a handful of exceptions, and each year they get fewer. The typical biography of a finalist goes something like this:

"My first introduction to programming was on the Apple IIe with BASIC at a summer day camp when I was eight. I still have a book which I later used to teach myself more BASIC. I remember having to skip over the sections that were QBASIC specific because all I had was GW-BASIC.

When I was in the fourth or fifth grade, I used to spend a substantial amount of time programming, mainly in QBASIC. I remember reading in *Scientific American* about a project to simulate simple societies on a computer, and trying to write a similar program myself. I got the engine working with some of the basic features, but the inadequacies of QBASIC for the type of complex program I was trying to write, combined with my own inexperience, eventually caused me to give up on continuing the project.

Since then I have learned C++ and some better programming techniques, but most of what I know about algorithms, I've learned in preparation for the USACO contests. I am not one who knows the workings of Unix or Windows inside out, or who knows how to write a good-looking web page. In fact, the only languages I know besides C/C++ are Lisp and a little Java and BASIC.

When I am not programming, I enjoy math, physics, and whatever related endeavors I can find, including as many academic competitions as I can juggle. This year, I'll be competing in the International Physics Olympiad in July. I was also invited to the math olympiad training camp. Unfortunately, because it conflicts with both the USACO and the physics olympiad, I could not go."

Common threads that run through all fifteen USACO finalists and our guest Tomasz from Poland are the following: Early introduction to programming in BASIC—mostly self taught at home; strong background in mathematics and mathematics competitions (almost all have competed in regional and national mathematics competitions, and some have reached the level of the International Mathematics Olympiad); moderate background in physics competitions including the Physics Olympiad training camp; and an early programming experience, often to create their own computer games in BASIC.

Of course, many of the competitors have other interests too, such as music, reading, or playing sports. But the following revelation by one of the finalists this year may contain the key to the passion for computers that puts them on the fast track to the USA Computing Olympiad.

"I am interested in anything involving computers, including programming, game playing, taking them apart, putting them back together, and trying to figure out why they don't work after I took them apart and put them back together."

Problem types

What kinds of problems will our team face at International programming competitions? This is a question our students need to know. Hal Burch, a staff member and graduate student in computer science from Carnegie Mellon, has classified all previous IOI problems into 16 types: dynamic programming, greedy algorithm, complete search, flood fill, shortest path, recursive search, minimum spanning tree, knapsack, computational geometry, network flow, Eulerian path, convex hull, big numbers, heuristic search, approximate search, ad hoc problems. Furthermore, the top half of this list comprises almost 80% of the problems seen at IOI.

The most valuable technique is dynamic programming. Until one understands it, however, it seems almost like magic. Exhaustive search algorithms, which try all possibilities and select the best, come with a prohibitive cost in time. Since time is a big factor in the IOI competition, finding the fastest solutions is critical. Dynamic programming is based on the idea of storing the consequences of all possible decisions to date and using this information in a systematic way to minimize the time to find a solution. No two cases seem exactly alike, so dynamic programming is best learned by studying a number of examples. It is often the case that recurrence relationships hold the information you need in dynamic programming.

We devoted a whole day to dynamic programming, during which Russ Cox, a staff member and undergraduate from Harvard, organized lectures, drills, lab problems and a set of eight different problems to be solved by the participants working together in two-person teams. Each team then presented its solution to the whole group and did an outstanding job.

"Crafting Winning Solutions" was a lecture/discussion given by Greg Galperin, a staff member from MIT who is now on grad school leave and working for a startup company. He said: "When analyzing an algorithm to figure out how long it will run for a given data set, the first rule of thumb is: modern computers can deal with 10^7 actions per second. So 5×10^7 is the number of actions that can be handled in a five-second time limit. Do the math. Plug in the actual expected and worst case-numbers."

Competition day

Two days were reserved for competition. We simulate the same environment our team will face at the International Olympiad in Informatics (IOI)—three to four problems to be solved working alone for five continuous hours. After the time was up, Russ Cox applied his grading program to each of the contestants' programs. Test data, which had been carefully chosen by the staff to test all aspects of the problem, was fed into each program, and the output was analyzed for correctness and speed. Each contestant got a printed result form showing the result on each test case, the time taken by the program, and the correct answer. After lunch, our solutions were explained by Hal Burch, and each contestant got a chance to run the test data once again against his program to verify the results.

Nine Men's Morris

Writing programs that play games is fun and very instructive. Back in the early days of the personal computer, you could buy a book such as "101 BASIC Computer Games," by David Ahl, and spend hours typing them in and running them. This is the way many of our finalists first learned how to write their own games. Before coming to camp, Rob Kolstad, the head coach for the USACO, offered the finalists a challenge: Write a program that will play the game Nine Men's Morris. (The rules for the game can be found on the Web at www.game-club.com/jav10-10/nmmhist.htm.) At camp, Rob hooked up his hardware to the computers and monitored the play between dueling programs. All of the game moves were displayed on a screen through a projection system for the enjoyment of the fans and the admiration of the proud programmers. Our Polish guest, Tomasz Czajka, wrote the winning program, which was never defeated.

Training materials

The finalists this year were the best crop ever. One reason is probably the on-line training materials that have been placed on the Web. These were the creation of Rob Kolstad, head coach and USACO spark plug, whose "Jeopardy" quiz show is another big hit at summer camp. At ace.delos.com/usacogate, anyone can submit solutions to an ever expanding set of problems and have his or her program automatically tested with the sample data sets and the results returned. All of the finalists had visited this site before coming to camp and were at various stages of completing all of the problems. Expanding these materials and adding some for beginners is a goal for 2001.

The USA Informatics Team

After the second day of competition was over and judged, the results were compiled, and all that was left for us was to skim the cream off the top. At our awards banquet, we announced the four programmers who made the USA 200 Informatics Team: Reid Barton, a home-schooled junior who just recently returned from Macedonia, where he placed first in the Balkan Olympiad in Informatics as a representative of the USA; John Danaher, a senior from Thomas Jefferson High School for Science and Technology who placed second, one point behind Reid, at the recent BOI; Percy Liang, our veteran senior from Phoenix who won a bronze medal last year on the USA IOI team in Turkey; and Gregory Price, a sophomore this year from TJHSST, who came to the camp for the first time. The alternate was Jacob Burnim, a sophomore from Silver Springs, Maryland. Besides receiving a handsome plaque, each team member received a glass milk pitcher with our cow mascot etched on the side.

Fibonacci cows

Below is the first problem of the first day of competition—usually the easiest. Of course, it has a barnyard setting, as do all of the problems created by the coaches at the USACO. That's the tradition.

Farmer John's brother, James, decided to get into the cow business. He bought a white cow his first year and a brown one the following year. Each succeeding year he duplicated his purchases of the preceding two years, buying the same number of cows, of the same colors, and in the same order. Thus, in the third year, he bought a white cow and then a brown cow; in the fourth year, a brown, then a white, and then a brown; and so on.

Your task is to tell farmer John the color of the Nth cow his brother bought. Farmer John will always ask about the Nth cow for five different values of N.

INPUT FORMAT (file FIB.IN)

Five integers on a single line ($1 \le$ integer $\le$ 2,000,000,000).

SAMPLE INPUT
4 8 12 16 20

OUTPUT FORMAT (file FIB.OUT)

SAMPLE OUTPUT
brown white brown white brown

For those who would like to solve this problem, see if your program will arrive at the correct solution for N = 2,000,000,000.

The answer is brown, and you have one second to find it.

Sponsor

Once the finalists had been selected for the USACO training camp, they leave their money at home. The roundtrip ticket to Wisconsin, room, board, awards, banquets, polo shirts, T-shirts, and a day-long excursion to Six Flags Great America were paid for by our sponsor USENIX. USENIX is the Advanced Computing Systems Association, which brings together the community of engineers, system administrators, scientists, and technicians working together on the cutting edge of the computing world (www.usenix.org).

Finally

To find out more about the USACO and this year's finalists and the final team of four going to IOI, go to our Web site at www.usaco.org and click on 2000 and Training Camp/Finals. View our training materials located at ace.delos.com/usacogate. If you think you may have the WRITE stuff, "come on down." We want you! And if you are female, we need you all the more!

Volume 10 (1999–2000)

HAPPENINGS

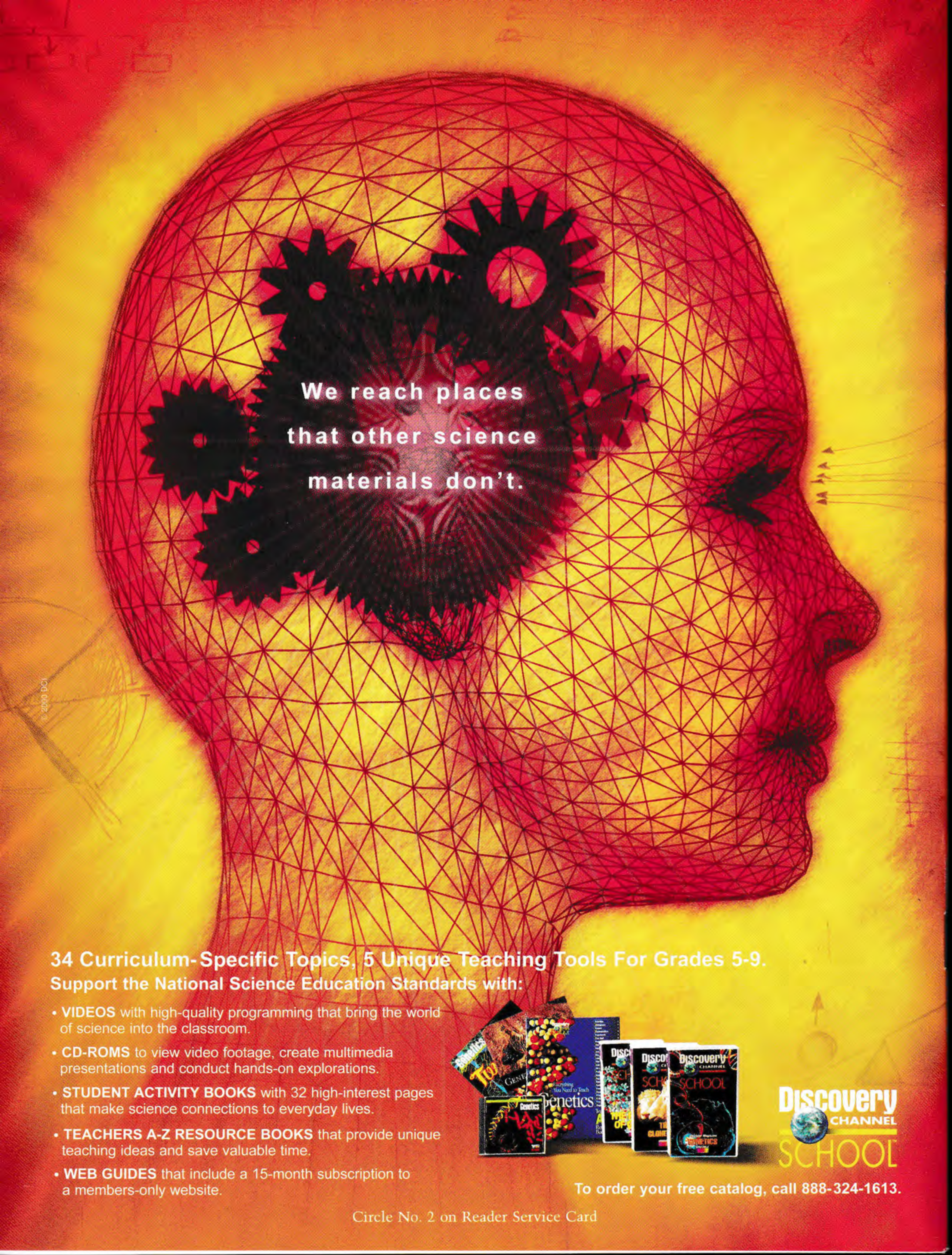
Cyberteaser winners

THE FOLLOWING ARE THE names of first ten people to submit a correct answer to this month's Cyber-Teaser—Be Fruitful and Multiply. As always, we received a cornucopia of international entries this month, but the following were the pick of the crop.

Jerold Lewandowski (Troy, New York)
Bruno Konder (Rio de Janeiro, Brasil)
Dimitrios Vardis (Ithaki, Greece)
Theo Koupelis (Wausau, Wisconsin)
Nick Fonarev (Staten Island, New York)
Christopher Franck (Playa Del Rey, California)
Gianluca Crippa (Dervio, Italy)
Jacopo De Simoi (Treviso, Italy)
Michael Brill (Morrisville, Pennsylvania)
Marco Devigili (Verona, Italy)

Our congratulations to the winners, who will receive a copy of this issue of *Quantum* and the coveted *Quantum* button. Everyone who submitted a correct answer (up to the time the answer is posted on the web) is entered into a drawing for a copy of *Quantum Quandaries*, a collection of 100 *Quantum* brainteasers. Our thanks to everyone who submitted an answer—right or wrong. You will find our next CyberTeaser at:
http://www.nsta.org/quantum.

Index of Advertisers


We reach places that other science materials don't.
34 Curriculum-Specific Topics, 5 Unique Teaching Tools For Grades 5-9.
Support the National Science Education Standards with:
• VIDEOS with high-quality programming that bring the world of science into the classroom.
• CD-ROMS to view video footage, create multimedia presentations and conduct hands-on explorations.
• STUDENT ACTIVITY BOOKS with 32 high-interest pages that make science connections to everyday lives.
• TEACHERS A-Z RESOURCE BOOKS that provide unique teaching ideas and save valuable time.
• WEB GUIDES that include a 15-month subscription to a members-only website.
Discovery CHANNEL SCHOOL
To order your free catalog, call 888-324-1613.
Circle No. 2 on Reader Service Card