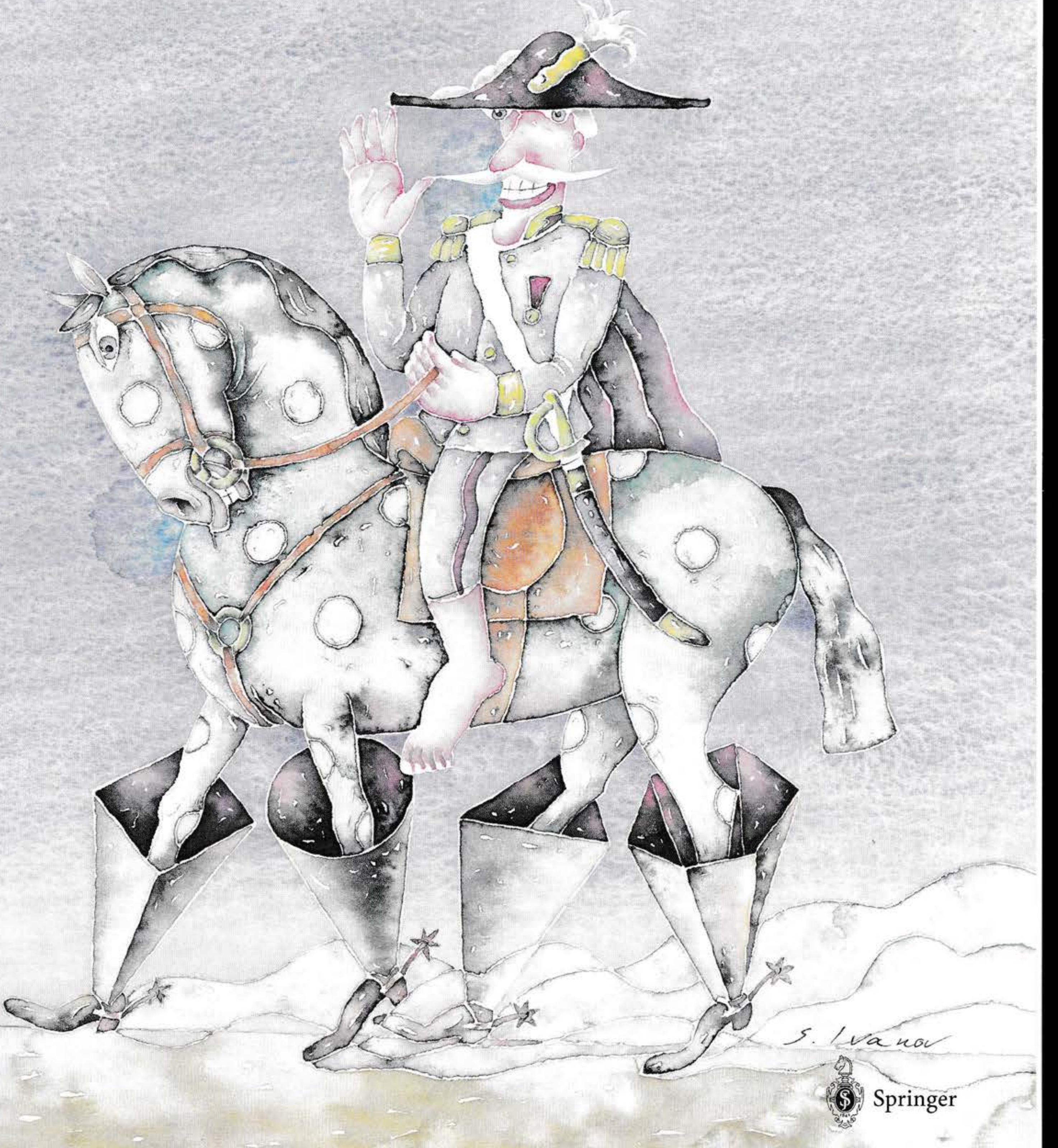
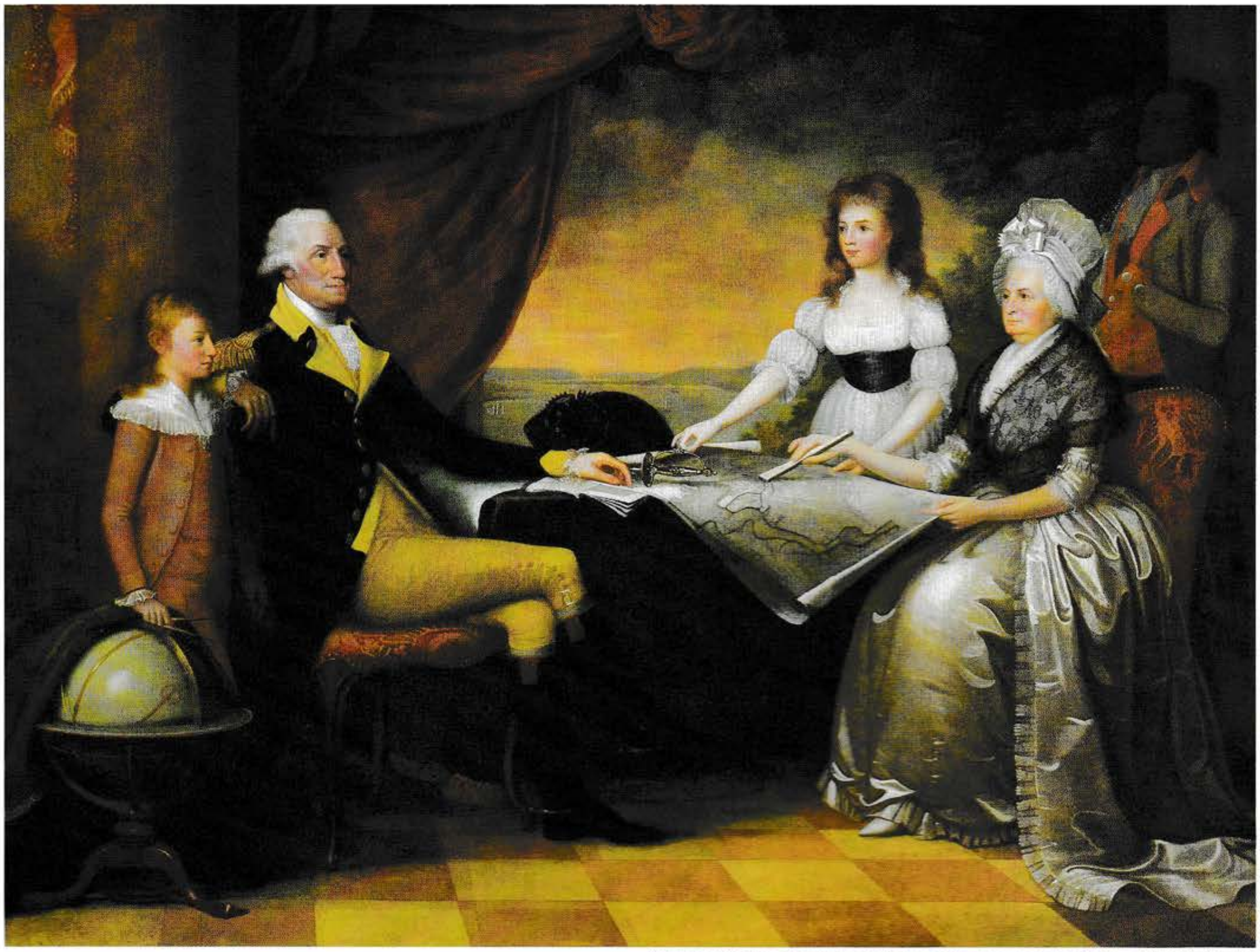
QUANTUM

MARCH/APRIL 1997

US $7.50/CAN $10.50



Springer



The Washington Family (1789–96) by Edward Savage

BEFORE HE WAS A SOLDIER, GENTLEMAN FARMER, and president of the United States, George Washington was a surveyor. At the age of 16 he served as an assistant to the official surveyor of Prince William County, Virginia, and kept a diary of his experiences. He records the discomforts of sleeping under "one thread Bear blanket with double its Weight of Vermin such as Lice Fleas & c" and describes an encounter with an Indian war party bearing a scalp. A year later Washington was appointed the official surveyor of Culpeper County, which gave him additional experience in dealing with the challenges of the wilderness.

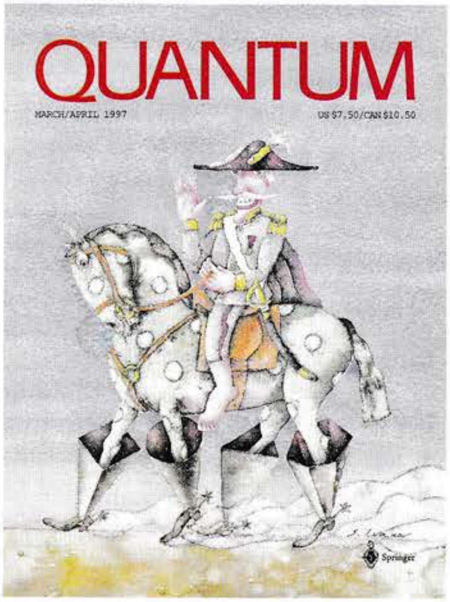
Many years later, as he and his family gathered for a portrait, symbols of his maturity and his youth intermix. The map spread on the table, the globe off to the left, reflect the public servant devoted to international matters as well as the private citizen who amassed a great deal of land. His adopted son holds a compass, as if his father had just quizzed him about the number of acres at Mount Vernon.

The globe contains a curious contrivance, perhaps for determining the position of the Sun or other heavenly body. It most assuredly showed latitude and longitude to a reasonable degree of precision. Did Washington hear stories, as a young surveyor, of how difficult it once was to determine one's longitude accurately? And how the invention of the marine chronometer in 1835 (when George was just three) revolutionized navigation?

It was arguably the most important scientific problem of the time, and you can read about it in "The Long Road to Longitude," which begins on page 42.

QUANTUM

MARCH/APRIL 1997 VOLUME 7, NUMBER 4



Cover art by Sergey Ivanov

The jaunty horseman on our cover is confident that his horse is comfortably shod. Even though each boot is a different shape, he knows that the volumes of the boots are all the same.

How does he know? He knows because he's Francesco Bonaventura Cavalieri, and in 1629 he developed a method for determining volume without the use of integrals. (His method involved "indivisibles.") Cavalieri was a student of Galileo and, true to his name, he chivalrously delayed publishing his result for six years because his mentor was planning a work on the same subject.

Read about Cavalieri's principle in the Kaleidoscope, right in the middle of this issue.

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

Creating scientist-citizens

How many scientists does it take . . . ?

EVERY PLACE I TURN, SCIENCE and technology seem to be making the news. Exciting astronomy and astrophysics show up almost weekly in the *New York Times*. The *Washington Post* and *Time* magazine regularly report breakthroughs in biology and the health sciences. Stunning photos from the Hubble Space Telescope regularly adorn magazines and even serve as wall-adorning posters.

In vitro fertilizations bring the miracle of life to otherwise childless couples. Premature babies now regularly overcome overwhelming odds. Television shows dramatic video from within living bodies. Faster and faster computers enrich (and complicate) our lives. DNA testing and other laboratory techniques have revolutionized law enforcement and judicial proceedings. Microprocessors enrich our homes, vehicles, shopping malls, factories, and offices.

Within the science community itself, the exponentially increasing size of journals reflects the incredible growth of research activities. Physics sees the merging of particle physics and cosmology, a dramatic blending of the very smallest and the very largest scales of the universe. Events during the first femtosecond after the big bang explicate the ten-billion-year history of the cosmos.

While science is thriving, while technology is remaking the economy of the United States and of much of the world, why then do we hear so much about the sad state of science education in our nation? While science and technology are omnipresent in our lives, why do policy makers talk so much about "scientific illiteracy"? In my own discipline of physics, why do many of our most committed education researchers now tell us that our students do not understand what we think we are teaching them, even if they manage to earn passing grades?

As a scientist, I am thrilled about the excitement across all fields of science. As a citizen, I am pleased that research and technology have redefined and strengthened our economy. As a teacher, however, I am concerned that most people do not begin to understand the science and the technology on which we increasingly rely.

Not everyone wants to become a scientist. Many don't have either the interest or the perseverance to learn much about the quantitative aspects of science. Nevertheless, the community of scientists, science teachers, and prospective scientists needs to recognize our continuing responsibility to nurture scientific interests in all of our citizens.

The gap between a limited number of scientists creating exciting scientific and technical breakthroughs and a growing number of citizens who are unaware of the science and technology that underlie their daily lives is already wide enough. Those on both sides of the divide must assure that the gap doesn't become a chasm.

Even with a limited number of scientists, science will continue to thrive. But without many more scientist-citizens, our society will be increasingly alienated from the science and technology forces that define our future. And a society estranged from its science will have neither good science nor an auspicious future.

—Bernard V. Khoury

Bernard V. Khoury *is the Executive Officer of the American Association of Physics Teachers (AAPT).*

QUANTUM

THE MAGAZINE OF MATH AND SCIENCE

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

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

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

Advertising:

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

Printed on acid-free paper.

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

Subscription Information:

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

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

Swinging from star to star

How to accelerate a spacecraft far into the cosmos

by Vladimir Surdin

THE POSSIBILITY OF INTERstellar flight is a very enticing problem. If we could solve it, entire vistas would open up. We could study the surfaces of different kind of stars, find new planetary systems, and even (dare we say it?) come into contact with extraterrestrial civilizations. In this article we won't examine every conceivable solution to this problem. Many clever projects have been proposed in recent years, ranging from nuclear and photon-powered spacecraft (for example, the British "Daedalus" project) to a design involving a solar sail. Instead we'll look at one comparatively new idea that in principle makes it possible to send a large number of small automated probes to various stars in our Galaxy.

Slingshot effect

The trajectories of interplanetary flights are often planned so that the spacecraft pass near the planet not only to study it, but also to accelerate the spacecraft by means of the planet's gravitation. A change in the flight trajectory due to the action of a planet's gravitational field is usually referred to as a "slingshot effect." This effect was repeatedly used during the flight of the Voyager II spacecraft on its Earth–Jupiter–Saturn–Uranus–Neptune itinerary. To investigate the Sun's polar regions, the Ulysses spacecraft was paradoxically launched toward Jupiter (it isn't a trivial thing to get near the Sun—see "A Flight to the Sun" in the November/December 1996 issue). To minimize fuel consumption in its mission to Jupiter, the Galileo spacecraft flew to Venus first. Plans call for the spacecraft to fly near every intervening planet to use its gravitation to accelerate and point the spacecraft in the proper direction.

The mechanics of this effect can easily be understood by the following simple example. If a heavy ball rolling on a table collides with a light ball rolling in the opposite direction, its velocity will hardly change, while the lighter ball will bounce away with increased velocity. (See figure 1. Solve this problem yourself, using the laws of conservation of kinetic energy and of momentum.) A similar phenomenon occurs during the "gravitational collision" of a heavy planet with a light spacecraft. The only difference is that the collision of solid bodies is almost instantaneous, while the gravitational encounter goes on for a long time. However, the laws of mechanics are the same in both cases.

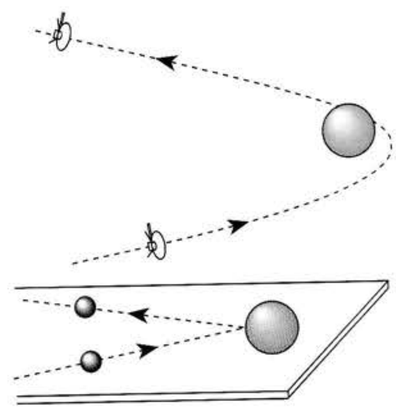


Figure 1

Thus the outcome is the same: after the spacecraft flies around the planet, its velocity increases. But how much?

Figure 2 will help us solve this problem. We use the simple rule of velocity addition in two inertial reference frames. From the viewpoint of a distant observer (who watches the flight while sitting comfortably on, say, the Sun!), the planet moves with a velocity v_p, while the spacecraft moves with a velocity v_s. The directions of these velocities form an angle α. From the viewpoint of an observer on the planet, the spacecraft approaches with a velocity v_+ and flies off with the same velocity—only its direction differs from the previous one by some angle β. In

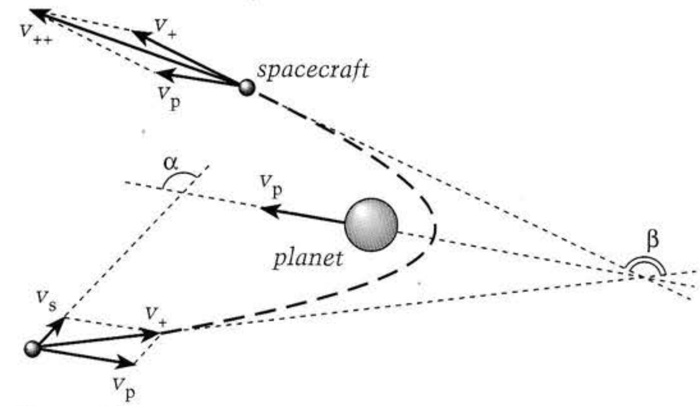


Figure 2

our reasoning we assume that the planetary system of reference is practically inertial—indeed, the perturbation of the planetary motion caused by the nearby flight of a spacecraft is inconceivably small. Relative to the distant observer, the receding space vehicle has a velocity v_{++}. Using the parallelogram rule, we get

$$v_+^2 = v_s^2 + v_p^2 - 2v_s v_p \cos \alpha$$

and

$$v_{++}^2 = v_+^2 + v_p^2 - 2v + v_n \sin (\beta/2).$$

It's clear at once that when the planet and spacecraft move toward each other (that is, in opposite directions—$\alpha \cong 180°$), and if their meeting causes the spacecraft to head in the opposite direction ($\beta \cong 180°$), the spacecraft's velocity will be increased by twice the planet's velocity. This is the most efficient case (fig. 3). However, to make the maneuver reasonably effective as a velocity enhancer, both angles should be in the range $\alpha, \beta \gtrsim 90°$. Of course, the spacecraft must fly over the planet's surface without crashing into it, so the minimum distance to the planet's center must be more than its radius. The laws of celestial mechanics say that in this case the maximum receding velocity of the spacecraft relative to the planet after their meeting will be $v_{max} = 0.46v_{esc}$, where v_{esc} is the escape velocity at the planet's surface. If the planet revolves about a star, the corresponding value will be $v_{max} \cong v_{orb} + 0.5v_{esc}$, where v_{orb} is the orbital velocity of the planet.

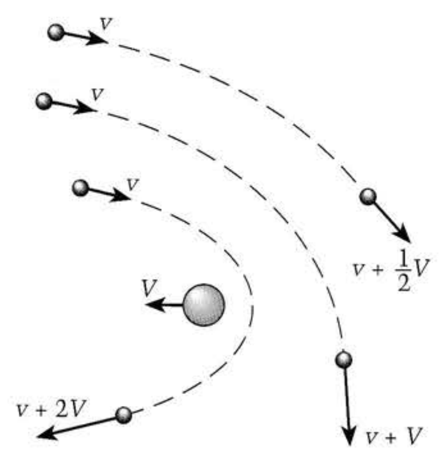


Figure 3

To make it possible for the spacecraft to leave the Solar System after approaching a planet, the condition $v_{max} \geq \sqrt{2}\, v_{orb}$ must be met—that is, the spacecraft's velocity must be greater than the velocity needed to escape from the planet's orbit. Table 1 shows that not all the planets in the Solar System can be effective "boosters"—only the giant planets can kick the spacecraft out of the Solar System with a single slingshot maneuver. The last column in the table shows the resulting velocity of a spacecraft leaving our Solar System after such a slingshot maneuver.

Stellar catapult

Now we know that by choosing a particular near-planet trajectory, we can impart extra velocity to a spacecraft without fuel consumption and thus send it out of the Solar System. Clearly this extra energy is taken from the mechanical energy of the planet. Can we come up with a similar mechanism to accelerate spacecraft in the vast expanses of the Galaxy? After all, the stars are moving too, so a slingshot maneuver near them could increase the velocity of our interstellar probe.

Astronomers know that the characteristic velocities of stellar motion are in the range of 10–20 km/s for young stars and 250–300 km/s for the oldest ones. So each meeting with a star along a proper trajectory will add hundreds if not thousands of kilometers per second to the spacecraft's speed. As in the case with a planet, the limit of an effective maneuver is determined by the value v_{max}, which is much larger for stars than for planets (see table 2). However, despite a wide variety of masses and sizes, ordinary stars have similar values of v_{max}, approximately equal to that of our Sun. This value isn't particularly large (~ 300 km/s), so such stars won't be the focus of our interest. Of particular importance for slingshot maneuvers are old compact stars—white dwarfs, neutron stars, and perhaps black holes (although the last-named won't be considered

Table 1

Planet	Orbital velocity v_{orb} (km/s)	Escape velocity at the surface v_{esc} (km/s)	Velocity required to leave solar system Δv (km/s)
Mercury	48	4.2	—
Venus	35	10	—
Earth	30	11	—
Mars	24	5	—
Jupiter	13	60	36
Saturn	10	36	22
Uranus	6.8	21	13
Neptune	5.4	24	14
Pluto	4.7	1?	—

Planetary data related to perturbation maneuvers performed in their vicinity.

Table 2

Parameter	Common star (Sun)	White dwarf	Neutron star	
			$R_{min} = R.$	$R_{min} = 50R.$
Mass of star $M.$	1	0.7	2	2
Radius of star $R.$	1	0.01	20 km	20 km
Escape velocity at the surface v_{esc} (km/s)	617	5,165	$15 \cdot 10^4$	$23 \cdot 10^3$
Maximum velocity of probe v_{max} (km/s)	309	2,583	$8 \cdot 10^4$	$12 \cdot 10^3$
Tidal acceleration a (m/s^2) for $\Delta r = 1$ m	10^{-6}	0.5	$7 \cdot 10^7$	530

Maximum speed of a probe and tidal acceleration for flybys of stars of various types.

here because very little is known about them).

The luminosity of most compact stars is not high, and their electromagnetic radiation offers no great danger for the spacecraft flying near them. However, there is a physical phenomenon that can severely restrict the very possibility of performing a slingshot maneuver, and this is particularly important for a flight in the vicinity of a neutron star. I'm talking about the tidal effect of a star's gravitational field, which tries to impart a differential acceleration $a = 2GM.\Delta r/R^3$ to parts of the spacecraft, where G is the gravitational constant, $M.$ is the star's mass, R is the distance from the spacecraft to the star's center, and Δr is the distance between the spacecraft's parts. Table 2 shows that the tidal forces near a neutron star are very strong, so they may be dangerous for the integrity of the spacecraft. Nowadays you can find electrical and mechanical devices that can withstand tidal accelerations up to 10^6 m/s^2, so the minimum distances between the interstellar probe and the surface of a neutron star presented in table 2 seems quite reasonable.

We should note that in contrast to common stars, whose mass is approximately proportional to their radius and consequently does not have much effect on v_{max}, the radii of neutron stars and white dwarfs decrease as their masses increase. This feature dramatically affects the value of v_{max}. Table 2 shows only the average values for these strange stars, which may differ from the characteristics of individual stars by a factor of 3 to 4.

How long will the acceleration last?

The strategy for accelerating interstellar probes is clear: as the spacecraft approaches the next star-accelerator, the probe's automatic pilot probe must choose from among the nearest stars one moving toward the probe and then adjust the approach to the first star so that the slingshot maneuver will send the probe off to the second star. If the path correction occurs far enough from the flyby point, the amount of fuel consumed will be negligible. Maybe other methods of flight control will have been found by then—using interstellar magnetic fields, perhaps, or radiation pressure—so that no fuel will be needed at all.

Art by Sergey Ivanov

Now let's estimate the time necessary for a probe to acquire the velocity v_{max} if it had been launched from the Solar System with a velocity v_0. For simplicity we'll assume a homogeneous distribution of stars in space with an average distance l between them. All the stars have the same velocity σ, which is chaotically oriented. If every "collision" with a star results in an increase in the probe's velocity of $\Delta v = \sigma$, the probe needs to perform $N = (v_{max} - v_0)/\sigma$ slingshot maneuvers, which will require a time

$$t = \sum_{k=0}^{N-1} \frac{l}{v_0 + k\sigma}.$$

For a rough estimate we can replace this sum with $t \sim l/v_0$ when $v \ll \sigma$, and with an integral for $v_0 > \sigma$, which yields $t \sim (l/\sigma) \ln (v_{max}/v_0)$. We see that the dependence of the accelerating period on the initial and final velocities is weak: for $v_0 = 100$ km/s and 300 km/s $\le v_{max} \le 10^5$ km/s, we have $1 \le \ln (v_{max}/v_0) \le 7$. Thus in all cases the formula

$$t \cong 2l\left(\frac{1}{v_0} + \frac{1}{\sigma}\right).$$

gives a rather accurate estimate. It uses three values. What are they?

Escaping the Solar System

What velocity does our probe need to leave the Solar System? Up to now only slingshot maneuvers near the giant planets have been used to accelerate the Pioneer and Voyager spacecraft into interstellar space. The velocities of these space probes relative to the Sun were about 20 km/s. In principle, a complicated slingshot maneuver in the gravitational fields of Jupiter or Saturn could accelerate a probe to a speed of almost 100 km/s. However, the planets must be in a certain configuration for this to happen.

Another method of entering interstellar space that seems quite possible today is the acceleration of small probes with electromagnetic mass accelerators. Such accelerators were developed within the framework of the Strategic Defense Initiative (SDI). Laboratory versions of such devices have accelerated a mass of 10 g to a velocity of 10 km/s. It's expected that a large-scale electromagnetic accelerator could impart a velocity in the range of 20–40 km/s to an object with a mass of about 1 kg. A further increase in velocity requires a drastic increase in the size of the accelerator ($\gtrsim$ 1 km), which is considered unacceptable for the SDI program. However, it may not be an obstacle in a project to launch interstellar probes. So one might hope that an electromagnetic accelerator will be constructed in space—one that could accelerate small probes to velocities of at least 100 km/s and could be pointed in any direction at any time, without regard for how the planets are configured.

The best location

Let's say that an interstellar probe has left the Solar System with a velocity v_0 = 100 km/s. The acceleration time depends on the type of star chosen for the maneuver. Table 3 shows the times it takes to accelerate the probe to maximum velocity. It should be noted that the concentration of stars is related to the average interstellar distance by the simple formula $l = n^{-1/3}$.

As we see, near the Sun the acceleration time is measured in the hundreds of thousands of years no matter what star "population" is used. However, if we were located at the center of a globular star cluster, this time would decrease to a few thousand years, and at the center of our Galaxy this period is a few hundred years. For a planetary system located at the center of our Galaxy, it would be reasonable to launch probes with a velocity v_0 = 300–400 km/s. In this case the probe could be accelerated by white dwarfs to a velocity of 5,000 km/s in only 100 years, and by neutron stars to a velocity of 100,000 km/s in a mere 300 years (provided, of course, the probe could withstand the tremendous tidal forces in the vicinity of the neutron star). Even more elaborate variants of gravitational acceleration for interstellar probes are possible in principle. For example, astrophysicists know of double star systems consisting of a neutron star and a white dwarf. These compact stars orbit with a velocity of more than 1,000 km/s. An approach to one of the components of such a system would increase a probe's velocity by 2,000 km/s! These systems are generally found in the depths of globular star clusters. (In fact, the central regions of these star clusters are extremely attractive places for civilizations that dare to take the first steps on the road to space colonization.)

Invisible space scouts

Nowadays we have at our disposal very compact and energy-efficient information devices. Microtransducers and microprocessors can be found in the most unexpected places—telephone receivers, notebooks, ballpoint pens, and greeting cards. And we are on the threshold of a new era in microsurgery, when tiny devices floating in the bloodstream will gather information for diagnosis and provide treatment.

Table 3

Location	Star cluster	n (ps^{-3})	σ (km/s)	t (year)	v_{max} (km/s)
Near the Sun	disc stars	0.1	45	10^5	400
	halo stars	0.005	250	$2 \cdot 10^5$	400
	white dwarfs	0.05	50	$2 \cdot 10^5$	5,000
	neutron stars	(10^{-3})	(100)	$4 \cdot 10^5$	10^5
Star cluster	common stars	$4 \cdot 10^4$	20	$4 \cdot 10^3$	400
	white dwarfs	($5 \cdot 10^4$)	20	$3 \cdot 10^3$	5,000
	neutron stars	(10^3)	20	10^4	10^5
Center of the galaxy (R ~ 1 pc)	white dwarfs	(10^6)	250	300	5,000
	neutron stars	(10^4)	250	10^3	10^5

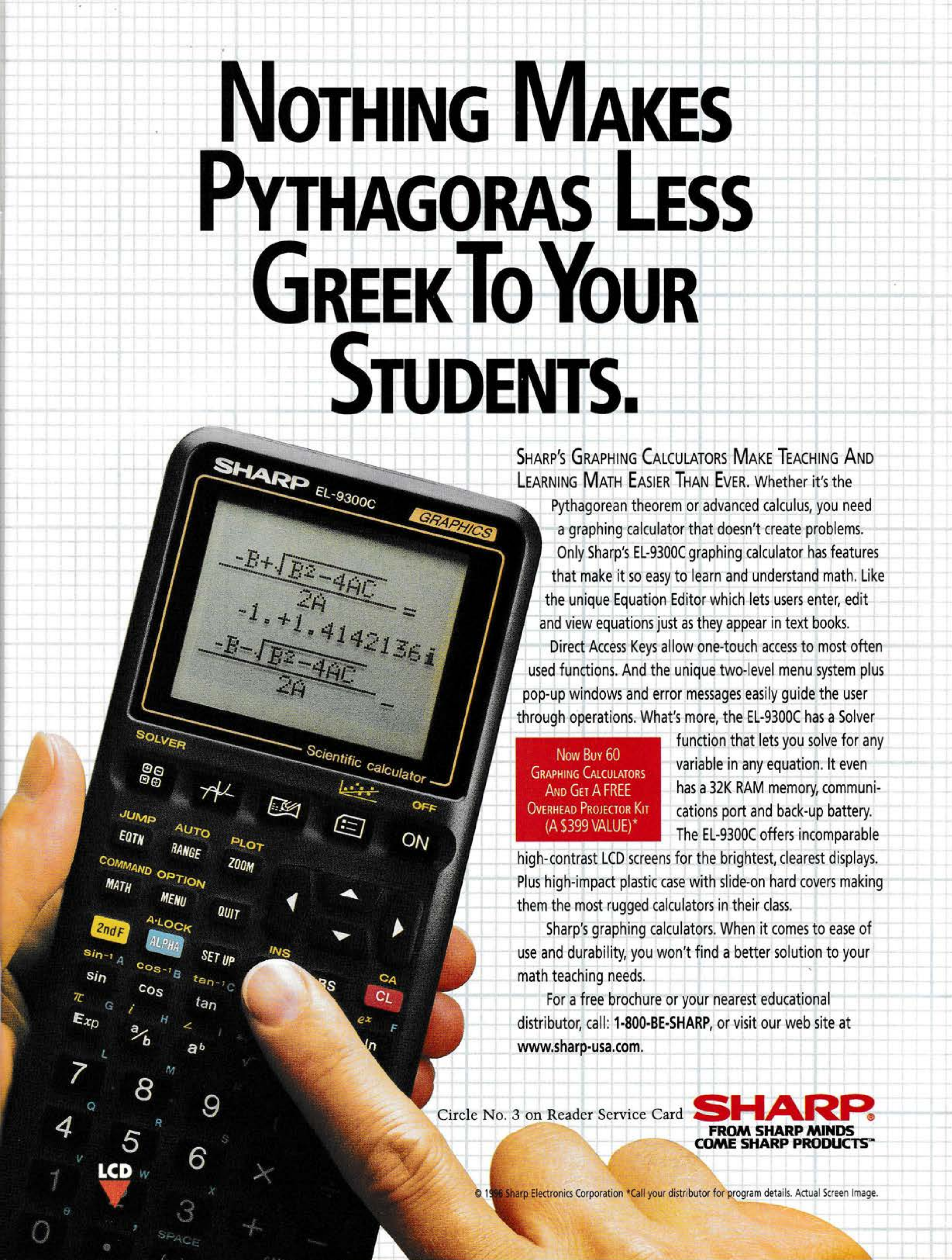
Characteristic acceleration time of a spacecraft launched from various sites and using various star clusters in the galaxy. The initial velocity v_0 = 100 km/s. Figures in parentheses are theoretical values; all others are based on observation.

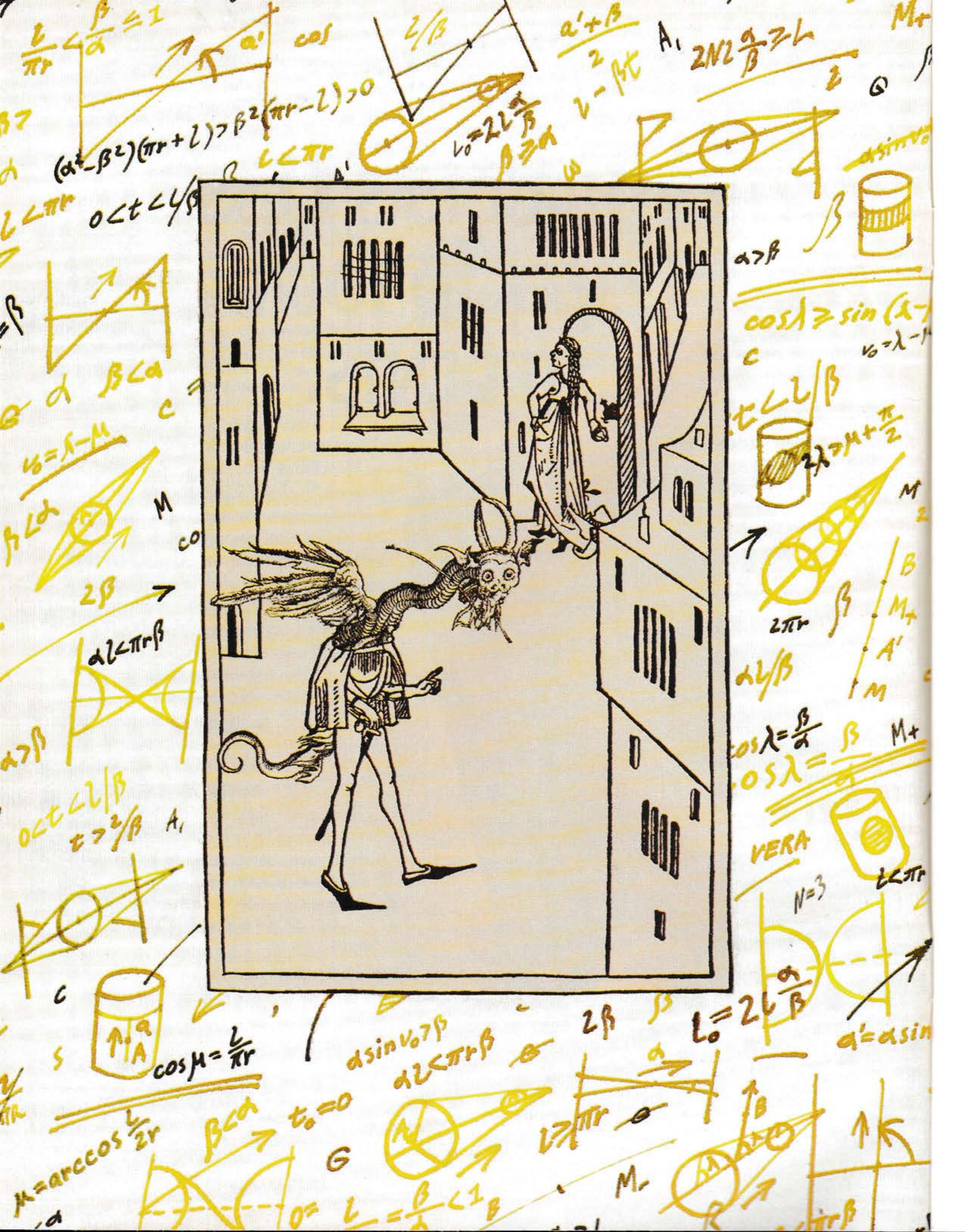
Space research will possibly take a similar route. Today it's hard to imagine that an interstellar spacecraft like the proposed Daedalus, requiring a kilometer-long rocket with a nuclear engine, might actually be built someday. Such a monster could be assembled only in Earth orbit, and the necessary work would take at least 15 years. For a flight involving only a small crew, it would be necessary to spend almost the entire store of nuclear fuel available on our planet. Only then could this rocket be accelerated to a velocity of a few thousand kilometers per second. A trip to the nearest star would take about a hundred years. A vast amount of effort and money would be spent on a single expedition, whose cost would be numbered in the trillions of dollars. This project is hardly more than a dream.

A more promising method of studying the distant cosmos would be to build many similar, relatively inexpensive microprobes, about 1 m in size and 10–100 kg in mass. Given the rapid rate of miniaturization, they might be even smaller. This is the only approach to galactic research that would not exhaust the energy and material resources of the planet. Since they are small, the microprobes could enter regions of relatively dense interstellar and interplanetary matter and could approach very compact and massive objects.

The strategy of using microprobes in space research must remain the subject of another article. Such a discussion would include the potential of optical communication as the most favorable for the vast distances involved, and also the possibility of returning the probes to their launch site.

If similar probes launched from other planetary systems have entered our Solar System, we have no means of detecting them at present. No doubt the same would be true for our future space probes and for any intelligent life forms we would hope to encounter. So our proposed method of galactic reconnaissance seems the safest and most responsible for the life forms on this planet. ◘


NOTHING MAKES PYTHAGORAS LESS GREEK TO YOUR STUDENTS.
SHARP EL-9300C
GRAPHICS
SHARP'S GRAPHING CALCULATORS MAKE TEACHING AND LEARNING MATH EASIER THAN EVER. Whether it's the Pythagorean theorem or advanced calculus, you need a graphing calculator that doesn't create problems.
Only Sharp's EL-9300C graphing calculator has features that make it so easy to learn and understand math. Like the unique Equation Editor which lets users enter, edit and view equations just as they appear in text books.
Direct Access Keys allow one-touch access to most often used functions. And the unique two-level menu system plus pop-up windows and error messages easily guide the user through operations. What's more, the EL-9300C has a Solver function that lets you solve for any variable in any equation. It even has a 32K RAM memory, communications port and back-up battery.
NOW BUY 60 GRAPHING CALCULATORS AND GET A FREE OVERHEAD PROJECTOR KIT (A $399 VALUE)*
The EL-9300C offers incomparable high-contrast LCD screens for the brightest, clearest displays. Plus high-impact plastic case with slide-on hard covers making them the most rugged calculators in their class.
Sharp's graphing calculators. When it comes to ease of use and durability, you won't find a better solution to your math teaching needs.
For a free brochure or your nearest educational distributor, call: 1-800-BE-SHARP, or visit our web site at www.sharp-usa.com.
Circle No. 3 on Reader Service Card
SHARP®
FROM SHARP MINDS COME SHARP PRODUCTS™
© 1996 Sharp Electronics Corporation *Call your distributor for program details. Actual Screen Image.



$(\alpha^2-\beta^2)(\pi r+l)>\beta^2(\pi r-l)>0$
$0<t<l/\beta$
$l<\pi r$
$\beta<\alpha$
$v_0=\lambda-\mu$
2β
$\alpha l<\pi r\beta$
$\alpha>\beta$
$0<t<l/\beta$
$t>l/\beta$
$\cos\mu=\frac{l}{\pi r}$
$\mu=\arccos\frac{l}{2r}$
$t_0=0$
$\alpha\sin v_0>\beta$
$\alpha l<\pi r\beta$
$L_0=2l\frac{\alpha}{\beta}$
$\cos\lambda\geq\sin(\lambda-\mu)$
$2\lambda>\mu+\frac{\pi}{2}$
$2\pi r$
$\cos\lambda=\frac{\beta}{\alpha}$
VERA
$N=3$
$l<\pi r$
$\frac{\alpha'+\beta}{2}$
$2Nl\frac{\alpha}{\beta}\geq L$

Desperately seeking Susan on a cylinder

A geometric approach to problems of search and detection

by A. Chkhartishvili and E. Shikin

EVEN THE MOST ABSTRACT mathematical problems often have easily recognized and traceable roots in our everyday life. Sometimes the circumstances that give rise to these problems are rather far removed from mathematics.

This article will examine one such class of problems: search and detection.

We hardly need to point out the importance of search problems in our lives. In fact, it's all but impossible to pinpoint when these problems began to attract attention.

You can easily imagine one of our most distant ancestors creeping carefully into a cave, torch in hand. It was vitally important to make sure that this potential abode be free of dangerous inhabitants such as wild animals or poisonous plants, snakes, or spiders. If they happened to have settled there first, our ancestor would have to find them and get rid of them.

A somewhat modernized version of this prehistoric story is the dynamic game "Beauty and the Beast," in which the Beast has to catch the Beauty in a dark room. The room has an arbitrary shape, known to both players (maybe because of several small openings high in the walls, letting in some light). The Beast, who is assumed to be extremely intelligent, moves at a constant speed and can instantly change direction; the Beauty enjoys absolute freedom of movement. Capture occurs if the distance between the Beauty and the Beast is less than a given value.

Here's another example. Imagine an ancient castle surrounded by a picturesque forest. To prevent his enemies from sneaking into the castle, its owner, a rich and noble duke, has ordered that a path be cut around the castle, so that his faithful knights could keep watch along it (see figure 1). The duke must decide how many knights he needs to be sure that no enemy scout can cross the path unnoticed. We'll answer this question below.

As you can see, the participants in these search problems pursue different goals: some of them search, while others try to hide or to escape. The problems we'll investigate in this article will involve only one searching object—the "seeker"—and, generally speaking, an arbitrary number of objects that are sought. The behavior of the objects sought may also vary. So we need to distinguish between searches involving stationary objects and those involving moving objects. In the second case, the most interesting problems are those in which the objects sought try to avoid capture and therefore take all available information about their opponent into consideration when they move.

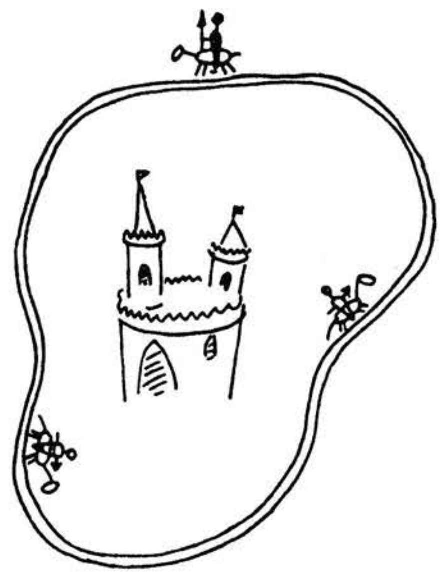


Figure 1

Search sets—that is, sets on which the search process takes place—come in the widest variety imaginable. One of the simplest is the infinite round cylinder, which

Art by Vera Khlebnikova

will serve as the location for most of the basic events below.

In this article we propose a geometric approach to search problems that uses certain auxiliary sets whose forms and locations change with time. It will be convenient to start a discussion of their nature, origins, and useful properties with the planar case (that is, when all the objects move on a single plane).

Simple search on the plane

Let A and B be two point objects on the plane. We assume that they can move with constant velocities α and β, respectively ($\alpha > \beta$), and that there are no other restrictions on their behavior. We say that the object B is detected by the object A, if at some moment the distance between them is less than or equal to a constant positive number l. We'll call A the *searching object* and B the *fleeing object*.

Information available to the objects. We assume that both objects know the shape of the search set and the values of all the parameters α, β, and l. In addition, the fleeing object B knows the future trajectory of the searching object A as well as A's position on the search set at any moment. On the other hand, the searching object A knows nothing about the location of object B up to the very moment of possible detection.

If an object's velocity is constant, one usually says that its motion is "simple."

The searching object A carries with it an imaginary circle of radius l, and it stays at the center of this circle at all times (fig. 2). If the fleeing object B falls into this *l-circle of detection*, it means that A has successfully completed its task (detection). Clearly A tries to find B, which in turn tries to avoid capture.

Prohibited sets. As long as A is stationary, the l-circle of detection is

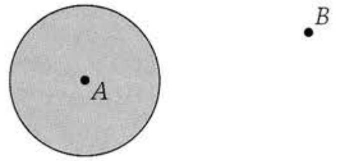


Figure 2

a prohibited set for B. But as soon as A starts to move (with velocity α), the set prohibited for B starts to grow. We'll show how this happens.

Suppose that A travels along a straight line L. Then at every moment, near line L in the direction of travel of object A, a set arises around L on which B must not appear or it will be detected as A advances along this line. Let's construct this set, taking $t_0 = 0$ as the moment when A starts moving. Suppose A starts at a point A_0 and reaches a point A_t after t units of time (fig. 3). In that time, B can travel at most a distance of βt in any direction. So if B is caught in a circle of radius $l - \beta t$ around A, it will be detected. So this circle is part of B's prohibited set. It's clear that if in the beginning B is in this circle, then at the moment t, when A comes to A_t, it will enter the l-circle of detection.

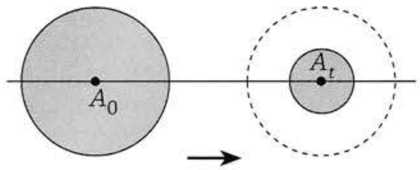


Figure 3

This reasoning is valid for all t from 0 to l/β. Therefore, all the circles drawn in this manner (they are lined up along L, and their radii decrease in the direction of A's movement) are prohibited for B. Hence their union is prohibited for B as well. This union is bounded by an arc of a circle ω of radius l and two line segments tangent to this circle, drawn from the point that lies on L at a distance $\alpha l/\beta$ from the center of the circle (fig. 4). We'll call this

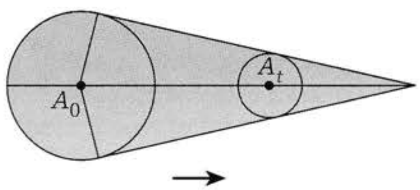


Figure 4

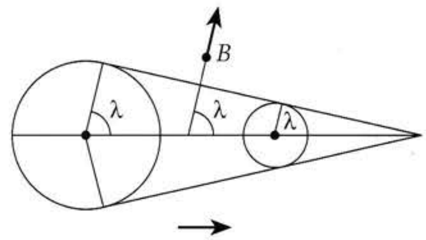


Figure 5

figure the *warning area*. The angle λ between L and ω's radii drawn to the contact points, is determined by the following equation:

$$\cos\lambda = \frac{\beta}{\alpha}.$$

The reader is invited to show that if B is outside the warning area, it can flee from A if it moves at an angle λ to A's trajectory L (or, equivalently, if it moves perpendicular to the straight part of the warning area's border) (fig. 5).

There is another set of points that B cannot enter: the area B doesn't have time to enter after the searching object A has left it along with its l-circle of detection. Its structure is quite similar to that of the warning area: it's the union of circles with centers on A's trajectory and gradually decreasing radii (starting with l—the radius of the l-circle of detection). We'll call this set the *residual area*. Note that after a time equal to l/β after object A starts to move, the residual area acquires the shape shown in figure 6.

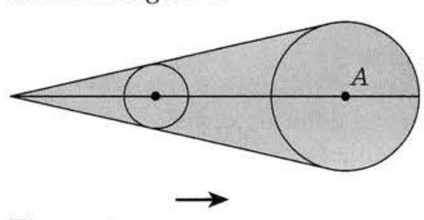


Figure 6

We'll call the union of the warning and residual areas the *tracing area* (fig. 7).

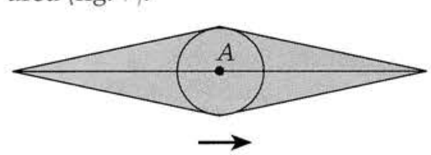


Figure 7

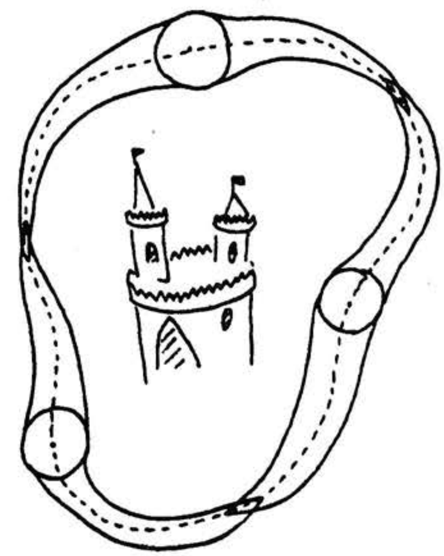


Figure 8

Properties of the tracing area (in the case where the trajectory of the searching object A is a straight line):

1. The length of the segment cut by tracing the area of the trajectory equals

$$L_0 = 2l\frac{\alpha}{\beta}.$$

2. The shape of the tracing area depends neither on the direction of movement nor on the moment of time $t > l/\beta$ under consideration.

3. The straight line L is an axis of symmetry of the tracing area, and the center of the l-circle of detection is its center of symmetry.

Let's return to the knights patrolling the guard path around the castle. Suppose their speeds are identical and equal to α and that $\beta < \alpha$ is the greatest speed of a villain who tries to sneak into the castle (he has to cross the path to do so). We'll also consider that a knight is able to recognize a spy at a distance less than or equal to l.

When knights are riding along the path, each of them carries his own tracing area along with him. These areas are determined by the numbers α, β, l and the trajectory—that is, the guard path, whose length is L. Each tracing area cuts an L_0-long segment from the path. Thus the number N of knights needed to protect the castle from enemy intrusion must satisfy the inequality

$$2Nl\frac{\alpha}{\beta} \geq L$$

(see figure 8, where $N = 3$).

Search on the infinite cylinder

We define an infinite cylinder as a set of points in space that are equidistant from a given straight line—the *axis of the cylinder*. A plane perpendicular to the axis intersects the cylinder in a so-called *directing circle*. Its radius does not depend on the cutting plane and is called the radius of the cylinder. Any plane containing the cylinder's axis cuts it along two straight lines—called *linear elements of the cylinder*.

Let C be an infinite cylinder of radius r. Consider an infinite $2\pi r$-wide strip Π on the plane (that is, the part of the plane bounded by two parallel straight lines). Clearly, it's possible to wind this strip Π around the cylinder so that the straight lines that define Π coincide with each other and with a linear element of the cylinder. Or, if we start with the cylinder, we can cut it along one of its linear elements and "unwrap" it to obtain a $2\pi r$-wide strip (fig. 9).

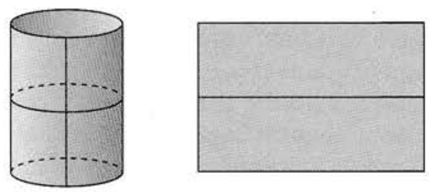


Figure 9

Plane development of the cylinder. Now imagine that one of the linear elements and one of the directing circles of a cylinder are covered with a paint that never dries. Now let the cylinder roll over the plane uniformly in a straight line. After each full revolution of the cylinder, the painted linear element will leave a straight line painted on the plane; and the distances between these traces will be equal to $2\pi r$—the circumference of a directing circle. The trace left by a directing circle will also be a straight line (fig.10).

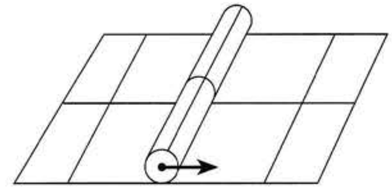


Figure 10

Reversing this procedure, we can roll the entire plane onto the cylinder (fig. 11). What will the lines on the plane look like in this *plane development*?

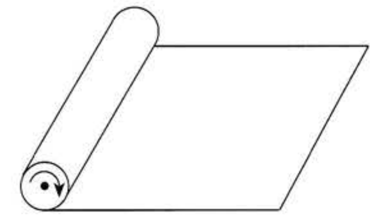


Figure 11

Depending on the angle between the line and the direction of the rolling, we can assign these lines to three classes:

1. Lines that are parallel to this direction (each of them rolls onto one of the directing circles of the cylinder (fig.12).

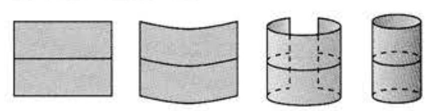


Figure 12

2. Lines that are perpendicular to this direction (each of them turns into a linear element of the cylinder (fig.13).

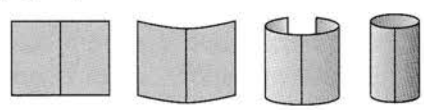


Figure 13

3. Lines that form an acute angle with this direction (they turn into curves on the cylinder, called *corkscrew lines* (fig.14).

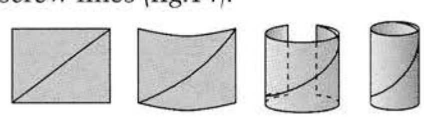


Figure 14

It's interesting to see how the corkscrew lines look on the plane

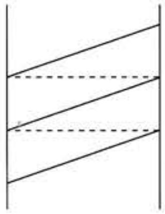


Figure 15

development of the cylinder.

Let Γ be a corkscrew line on a cylinder C. Cut C along a linear element and develop it on the plane. Figure 15 represents Γ's image after this procedure. Note that this is not a continuous line.

Directing circles, linear elements, and corkscrew lines are the analogues of the straight lines on the plane: they have the property of being the shortest distance between any two points. Lines that have this property are called *geodesics*.

Let's consider all possible geodesics emerging from an arbitrary point on the cylinder C and mark off segments that are $l < \pi r$ units long. The sum of all these segments is called the *geodesic circle of radius l* (fig.16).

Two point objects—the searching object A and the fleeing object B—move on a cylinder C of radius r. We assume that their scalar velocities α and β (respectively) are constant and that $\alpha > \beta$. B is considered to be found if it enters the geodesic circle with radius l and center at A.

Suppose that at first B is not close to A and that the latter knows which half of the cylinder B is in.

We'll show that if the parameters of the problem satisfy certain conditions, then there exists a corkscrew line such that if the searching object A moves along it, it will necessarily find the fleeing object B.

It's clear that when A moves along a corkscrew line, a tracing area appears around it on the cylinder,

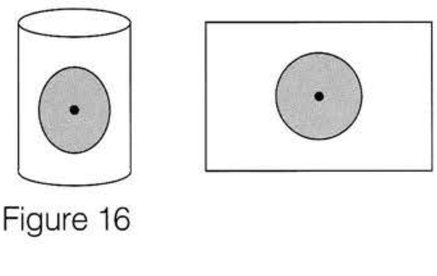


Figure 16

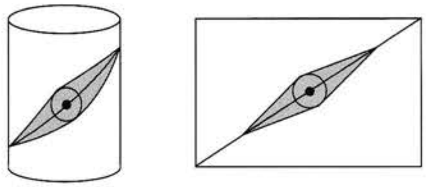


Figure 17

which is a union of geodesic circles. If we develop the cylinder on the plane, the corkscrew line will change into a straight line, and the geodesic circles into ordinary circles, and so the tracing area on the cylinder will develop into an ordinary tracing area on the plane (fig. 17).

Let's look at how the situation changes when we change the angle between A's trajectory and the directing circle (that is, the angle of inclination). Suppose that the following inequality is satisfied:

$$\alpha l < \pi r \beta.$$

Let object A begin moving along a directing circle of the cylinder. Then the tracing area contains a cylindrical strip of nonzero width, which is prohibited for B (fig. 18a). When A moves along a corkscrew line with a small angle of inclination, the prohibited area will still encircle the cylinder, and the warning and residual areas will overlap (fig. 18b). This will happen until the warning and residual areas just touch (this is ensured by the third property governing the tracing area for a straight line on the plane). We can calculate (fig. 18c) that this occurs when the angle of inclination between the corkscrew line and the directing circle is equal to

$$\nu_0 = \lambda - \mu,$$

where λ is defined as before and

$$\mu = \arccos \frac{l}{2\pi r}.$$

If the angle ν_0 increases further, the picture changes abruptly: the tracing area will contain no cylindrical strip (fig. 18d and 18e). The object in figure 18f is moving along the linear element of the cylinder.

When A moves along a geodesic on the cylinder, the tracing area does not change its shape, just as when A moved with a constant velocity along a straight line on the plane. Thus, moving along the corkscrew line Γ_0 inclined at an angle ν_0 to the directing circle, the position of A will not allow B to jump over to the other half of the cylinder, and if the projection of A's speed on the cylinder's axis is greater than B's velocity, $\alpha \sin \nu_0 > \beta$, then A will catch B.

Since $\cos \lambda = \beta/\alpha$ and $\nu_0 = \lambda - \mu$, the last inequality can be written as

$$2\lambda > \mu + \frac{\pi}{2}. \qquad (1)$$

So, if formula (1) is satisfied, detection is possible and can be achieved by A's moving along the corkscrew line Γ_0, which intersects

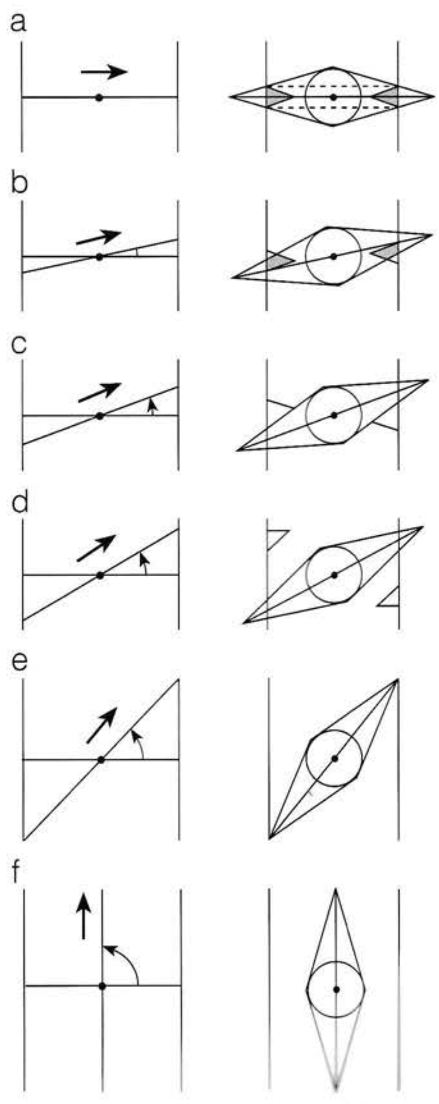


Figure 18

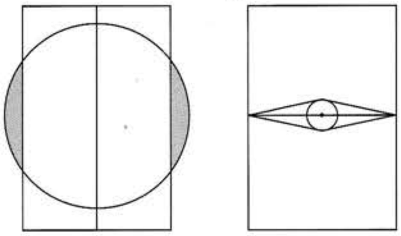


Figure 19

the directing circle at an angle ν_0.

If $l \geq \pi r$, then A may just move along a linear element, catching B because of its greater speed. In the case when $\alpha l < \pi r\beta$, there is practically no chance for successful detection (fig. 19).

Two detection problems on the infinite cylinder

Applying the notion of tracing areas to a search problem allows us not only to find essential relations between parameters that are sufficient for a successful search (detection), but also to determine the necessary trajectories.

A. The unlimited case. It was shown above that object B, which moves with a constant scalar velocity β, can be detected on an infinite cylinder C of radius r by object A, which moves with a constant scalar velocity α, provided that inequality (1) is satisfied and if it is known which half of the cylinder B occupies initially. In fact, if inequality (1) is satisfied, this second bit of information is superfluous. Let's see why.

In accordance with its resources, the searching object A at first chooses for its trajectory the corkscrew line Γ_0, which intersects the directing circle at an angle

$$\nu_0 = \lambda - \mu,$$

where λ and μ are determined by the relations

$$\cos\lambda = \frac{\beta}{\alpha}, \quad \cos\mu = \frac{l}{\pi r}.$$

Independently of the direction of A's movement along this line, the vertical component of its velocity (for definiteness we orient the cylinder

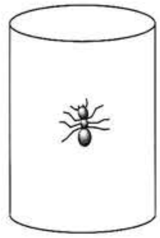


Figure 20

C as shown in figure 20) is constant and equals

$$\alpha' = \alpha \sin \nu_0.$$

It's clear that the most advantageous way for B to escape is to move along a linear element of the cylinder. So, to show that it doesn't matter which half of the cylinder B is on, it's convenient to consider the movement of both objects as projected on the vertical axis.

Let objects A' and B move along a (vertical) line with constant scalar velocities α' and β, respectively, $\beta < \alpha'$. We'll describe a strategy for A' that gives it an opportunity to get as close as l to B to the given distance l. Imagine two new objects M_+ and M_- that move from A''s initial position up and down, respectively, with velocities

$$\frac{\alpha' + \beta}{2}$$

(fig. 21). Object A' starts to move upward along the straight line and, catching up with M_+ (because of its greater speed), it immediately turns in the opposite direction and moves downward. When it catches up with M_-, it once again changes direction and chases after the object M_+, and so on. Clearly, by acting in this way A will sooner or later end up at a distance less than l from B.

It's not hard to see that A''s movement along the cylinder's axis, described above, corresponds to A's movement along the chosen corkscrew line Γ_0.

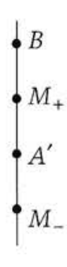


Figure 21

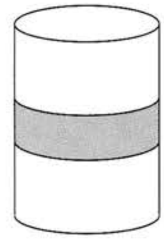


Figure 22

B. The limited case. Suppose at the initial moment the searching object A knows that the fleeing object B is standing somewhere in the cylindrical belt G bounded by two directing circles (fig. 22). If the searching object is far enough from this *uncertainty belt* G, then as time passes, the width of G will grow (at the rate of 2β).

Assuming that the parameters in the problem comply with inequality (1), we'll describe a trajectory of the searching object A that will necessarily allow it to find the fleeing object B.

There are three possibilities. At the initial moment $t_0 = 0$, the l-circle of detection around A could

(1) lie outside the belt G;

(2) belong to the belt G completely;

(3) belong to the belt G partially.

In the first two cases the searching trajectory is constructed according to a common rule. At first, object A moves along the linear element of cylinder C toward the nearest border circle of the belt G, until the distance from A to this circle is less than or equal to $a = l \sin \mu$ (fig. 23). At this moment, A decides to change the linear trajectory for the

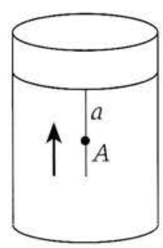


Figure 23

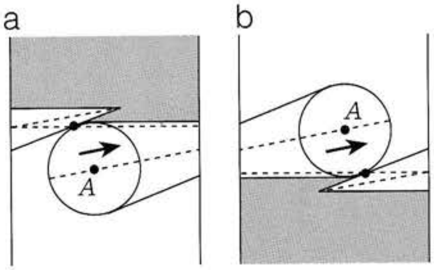


Figure 24

corkscrew line Γ_0 (see figure 24a, which shows the position of the warning area and the border circle at this time). After this, A moves along Γ_0 until the distance from A to the farther border is less than or equal to $l \sin \mu$. Figure 24b depicts the position of the residual area and the second boundary circle at the final moment of the search game.

Since we are interested only in finding sufficient conditions for successful detection, we can just reduce the third case to one the of previous two.

Problems of combing, patrolling, and deterrence

In considering the movement of arbitrary objects A and B on the surface of an infinite cylinder, we can't be sure that the parameters α, β, l, and r satisfy inequality (1), let alone the condition $l \geq \pi r$. Although an area prohibited for B still arises, it can not be used to solve the detection problem. However, there are many other interesting problems, rich with applications, that could be solved with much easier conditions imposed on their parameters. We'll pose these problems and point out the relations between the parameters that allow for them to be solved.

A. Combing. Let's assume that

$$l < \pi r,\ \beta < \alpha,\ \cos \lambda < \sin(\lambda - \mu).$$

In this case, moving along the corkscrew line Γ_0 inclined at an angle $\nu_0 = \lambda - \mu$ to the directing circle of the cylinder, A continuously encloses the cylinder with its tracing area. But since the vertical component of its velocity is less than or equal to β, A can only push B away, and only if it knows which part of the cylinder

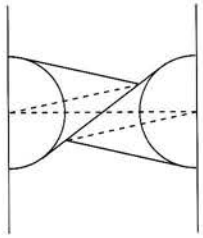


Figure 25

B was in originally. Figure 25 illustrates this situation (it shows the cylinder's development, and the linear element by which the cut was made is drawn through the center of the detection circle).

If, on the other hand,

$$l \geq \pi r,\ \beta \geq \alpha,$$

then A's detection circle will itself enclose the cylinder, and if in addition A knows which part of the cylinder B is in, it will be able to push B away.

B. Patrolling. Let's assume that

$$\frac{l}{\pi r} = \frac{\beta}{\alpha} < 1.$$

In this case the best trajectory for A is a directing circle. Then the warning and residual areas will have a common vertex, so A, patrolling the cylinder, will cover the neck of the cylinder for object B. (fig. 26).

C. Deterrence. Let's assume that

$$\frac{l}{\pi r} < \frac{\beta}{\alpha} \leq 1.$$

In this case A is not able to prevent B from moving along the cylinder. However, traveling in front of B along the corkscrew line inclined at an angle $\mu - \lambda$ to the directing circle, A will deter B's movements (fig. 27).

D. (Failure.) If, finally, we assume that

$$l < \pi r,\ \beta > \alpha,$$

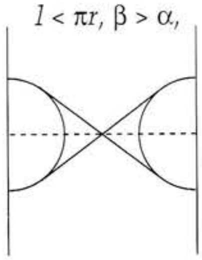


Figure 26

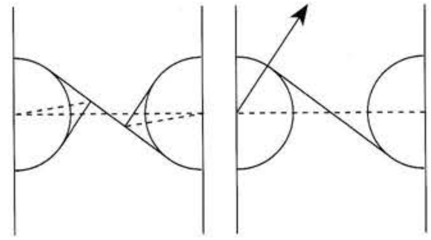


Figure 27

then A lacks the resources even to deter B.

So we see that the relations between α, β, l, and r became weaker and weaker as we move from the detection problem to the ousting ("combing") problem, and further to the patrolling and deterrence problems.

General remark. The last three problems may strike you as rather simple (we not only posed them, we also pointed out the relations between the parameters that ensure a resolution and the strategies that lead to success for the objects). But that's because we could base our solutions on our preceding analysis of the more difficult detection problem.

Just for the fun of it!

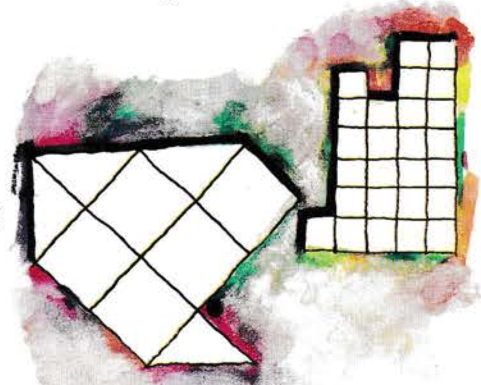
B196

Intimidating equation. Solve this equation:

$$10 - 9(9 - 8(8 - 7(7 - 6(6 - 5(5 - 4(4 - 3(3 - 2(2 - x)))))))) = x.$$

B197

Ball in a jar. Sally and her brothers Jesse and Raphael were playing with a ball that was 4 cm in diameter. They dropped it in a cylindrical jar whose mouth had a diameter of 5 cm. They managed to get the ball out without turning the jar upside down. How did they do it?



B198

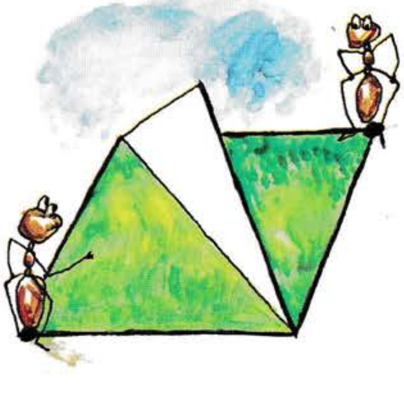
Straight line. Can you draw a single straight line that divides the areas of both figures at the right in half?

B199

Boiled oil. One morning a laboratory assistant at the United Science Institute weighed an open vessel of boiling oil on a very precise scale. Before going home in the evening she weighed the oil again, after it had cooled. The result was surprisingly different! What happened?

B200

Traveling ants. Two ants stand at opposite corners of a 1-meter square. A barrier was placed between them in the form of half a 1-meter square attached along the diagonal of the first square, as shown in the picture. One ant wants to walk to the other. How long is the shortest path?



Art by Pavel Chernusky

ANSWERS, HINTS & SOLUTIONS ON PAGE 62

Whistling in space

It's not an idle pastime—it's a phenomenon that allows us to study the near-Earth region without leaving the ground

by Pavel Bliokh

THE SPACE AROUND EARTH isn't as "empty" as you might think. There you can find neutral particles (atoms and molecules of atmospheric gases) as well as free charges—the electrons and ions forming the galactic plasma. Even though we live almost exclusively at the bottom of an ocean of air, it's important for us to know what's going on in the other layers of the atmosphere, because there is a close connection between the process occurring in outer space and the conditions of life here on Earth.

Cosmic particles and electric and magnetic fields at high altitudes are routinely monitored with devices installed on satellites and rockets. However, there are other ways of doing this. It turns out that much information on atmospheric properties hundreds and even thousands of kilometers from Earth can be obtained without leaving the lab. And these "Earth-bound space studies" can be conducted with very simple methods. The equipment needed costs about as much as a radio or TV set. But to take advantage of these tempting possibilities, you first need to learn how to hear the cosmic "whistling" (electromagnetic signals coming from outer space). And that's what this article is all about.

Radio waves in a neutral gas

Radio waves travel in a vacuum at the speed of light $c = 3 \cdot 10^8$ m/s. They are electromagnetic oscillations with a frequencies F ranging from several hertz to thousands of gigahertz (1 GHz = 10^9 Hz). By way of comparison, a conventional electronic device operates at frequencies from hundreds of kilohertz (kHz) to hundreds of megahertz (MHz). In this article we're interested in frequencies in the range of a few kilohertz ($f = 10^3$–10^4 Hz).

The propagation of radio waves in a medium differs from that in a vacuum. A medium always contains electrically charged particles, which can be in either a "fixed" (or bound) state (that is, electrons in neutral atoms) or a "free" state (electrons in plasma). In addition to electrons, which are carriers of negative charge, the medium contains positively charged particles (ions). In the frequency range of interest to us here, the role of ions is rather small, because they have a very large mass (compared to the electron) and thus oscillate with a very small amplitude. But electrons, oscillating in the radio wave's electric field, themselves become sources of secondary radio waves at the same frequency.

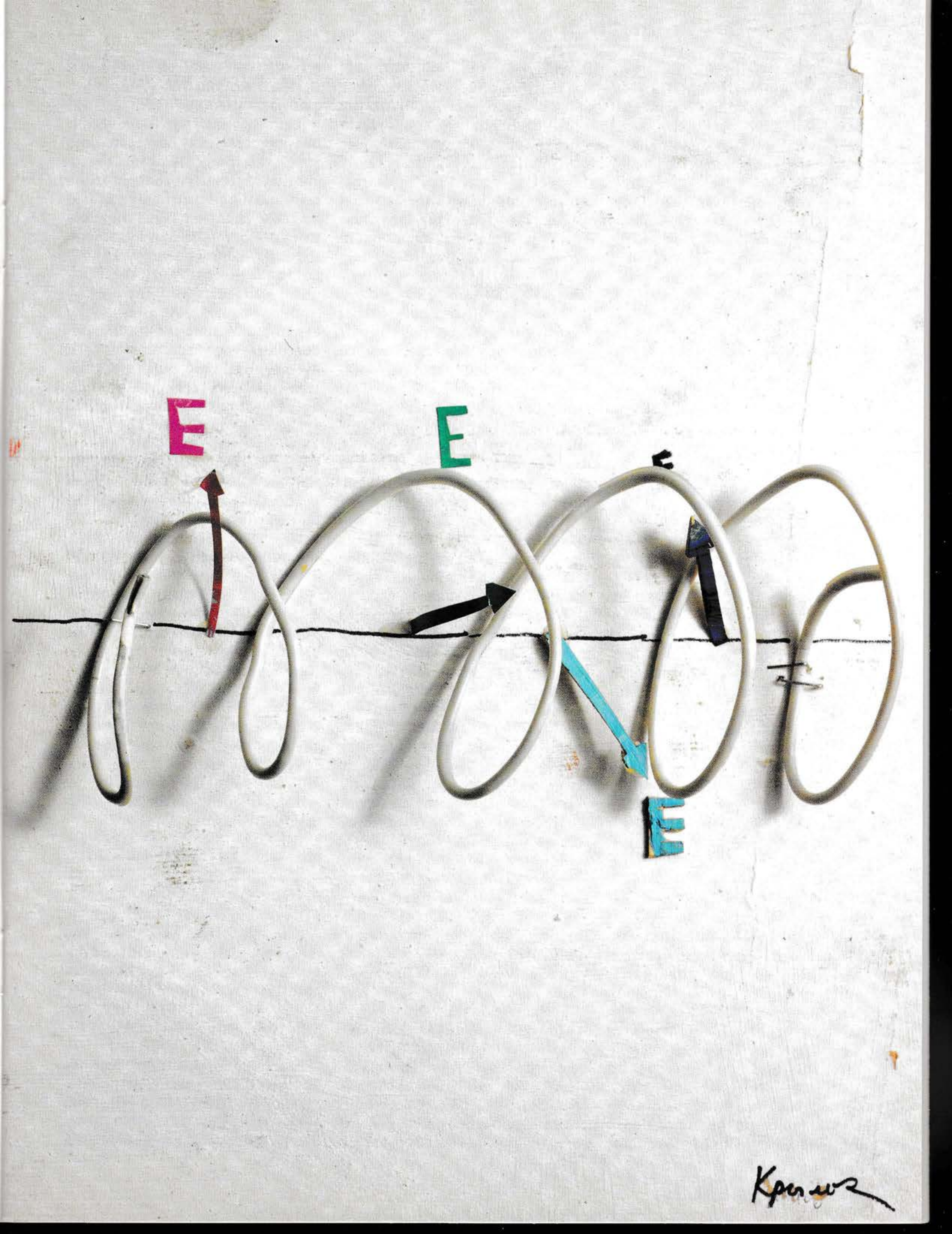
Thus the resulting wave is generated in a medium that spreads with another velocity. The change in the wave's velocity is explained by the refractive index n, which shows the factor by which the velocity of the electromagnetic wave in the medium is less than that in vacuum:

$$v_p = \frac{c}{n(\omega)}. \tag{1}$$

The meaning of the subscript "p" will be discussed below. Denoting the refractive index by $n(\omega)$,[1] we emphasize the fact that in general the refractive index of a given medium may differ at various frequencies. The dependence of n on ω is called the *dispersion*, and it becomes particularly noticeable when there is resonance—that is, when the radio wave's frequency ω approximately equals the natural oscillation frequency of electrons ω_0. If the frequency difference is large enough—for example, $\omega \ll \omega_0$—the dispersion is very small. In this article the nature of the function $n(\omega)$ plays a decisive role, so as a first step we need to estimate the natural frequencies of electron oscillation in a neutral gas and in plasma.

[1]The factor ω is called the angular frequency of oscillation. It's measured in rad $\cdot$ s^{-1}. The relationship between ω and f is very straightforward: $\omega = 2\pi f$.

Art by Dmitry Krymov


E
E
E

Let's start with a neutral gas, where electrons are bound within atoms and molecules. The precise calculations of their natural frequencies are performed by quantum mechanical methods, but we can obtain reasonable estimates by analogy with the simplest oscillatory system—the free pendulum. Recall the famous formula for the (natural) frequency of its oscillations:

$$\omega_0 = \sqrt{\frac{g}{l}}, \qquad (2)$$

where g is the acceleration due to gravity and l is the pendulum's length. Multiplying the numerator and denominator of the radicand (that is, the quantity under the radical sign) by the pendulum's mass m, we get

$$\omega_0 = \sqrt{\frac{F_g}{ml}}. \qquad (3)$$

This formula deals with the gravitational force $F_g = mg$ instead of g. Now we can modify it to describe the oscillations of electrons. We merely replace the gravitational force F_g with the electrostatic force F_e that "holds" electrons in the atom. According to Coulomb's law, this is equal to $F_e = e^2/4\pi\varepsilon_0 a^2$, where $\varepsilon_0 = 9 \cdot 10^{-12}$ C^2/N-m^2 is the permittivity of free space, $e = 1.6 \cdot 10^{-19}$ C is the electron's charge, and $a \cong 10^{-10}$ m is the size of the atom (which here plays the role of the pendulum's length). Inserting $l = a$ into equation (3) gives us

$$\omega_0 \cong \sqrt{\frac{e^2}{\varepsilon_0 m a^3}}. \qquad (4)$$

Taking into consideration the electron's mass $m = 9 \cdot 10^{-31}$ kg, we get $\omega_0 \cong 10^{16}$ rad $\cdot$ s^{-1}, which corresponds roughly to the frequency range of visible light and exceeds by far the frequency of the radio waves we are interested in. Therefore, the inequality $\omega \ll \omega_0$ is correct with a high degree of reliability, which means that the neutral gas has virtually no affect on the propagation of radio waves. Even in the lower, most dense layers of the atmosphere near the Earth's surface, the refractive index n differs from 1 only in the fourth decimal place, and in outer space the effect of the neutral gas is still smaller. It is manifested only indirectly, when "free" (plasma) electrons collide with their neutral counterparts. These collisions cause a damping of the oscillations, but in our case we can neglect it.

Radio waves in plasma

Under the influence of external factors (for instance, radiation or collisions), one or a few electrons can be knocked out of atoms, resulting in "free" electrons and ions. This ionized gas is what scientists call *plasma*. Plasma usually contains neutral particles as well, but the fewer of them there are, the more clearly the peculiar properties of plasma manifest themselves.

These is practically no natural plasma in the lower layers of the atmosphere, because the ionizing solar radiation (ultraviolet rays and X rays) is much weaker here. At an altitude of about 50 km, the ionizing effect of solar radiation begins to increase. This marks the start of the ionosphere—that is, the Earth's plasma shell. At altitudes of 300 to 400 km, the density of electrons and ions reaches a maximum and then slowly decreases, even though the intensity of ionizing factors increases with altitude. This is because the density of the air is very small at these altitudes, and although the atmosphere is almost completely ionized, the number of "free" electrons and ions is nevertheless small. In the outermost regions of space, the degree of ionization is very high, and according to modern estimates, ~99.9% of the visible universe exists in the plasma state.

Perhaps you've noticed that, when I mention plasma electrons, I keep putting the word "free" in quotes. Here's why. Even though the electrons are not bound to their atoms, they still interact with other electrons and ions by means of electrical fields. These forces operate at great distances, and it is these forces that are responsible for specific plasma oscillations.

Let's assume that the electron density increases by chance in some small volume. This means that a surplus of negative charges arises here, and the electric field generated by these charges pushes the electrons out of this region. Eventually the extra charge is dissipated away, but in the process the electrons acquire a velocity and move away from one another. As a result, the electron density in the volume under consideration becomes less than average, while the density of positive ions doesn't change. The shortage of electrons is equivalent to the appearance of a positive charge and an electric field that draws the escaping electrons back again. But as they move back, the electrons again acquire velocity and pass over the equilibrium position because of their inertia, resulting in an accumulation of negative charge, and the cycle begins anew.

To estimate the frequency of this oscillatory motion we again use equation (4), but instead of the atom's size a we insert the average (equilibrium) distance between electrons. Let the average number of electrons per unit volume (1 m^3) be N_0. Then the distance between them is $N_0^{-1/3}$ (on average, of course). Assuming $a = N_0^{-1/3}$ in equation (4), we get the formula for the natural frequency of electron oscillations in plasma:

$$\omega_p \cong \sqrt{\frac{e^2 N_0}{\varepsilon_0 m}}. \qquad (5)$$

Measuring the frequency in hertz and the electron density in m^{-3}, and inserting into equation (5) the numerical data for ε_0, e, and m, we obtain a simple formula:

$$f_p\,[\mathrm{Hz}] \cong 9\sqrt{N_0\,[\mathrm{m}^{-3}]}.$$

The maximum density of electrons is about 10^{12} m^{-3}, so the natural frequency $f_p \sim 10$ MHz lies in the radio wave range. This means that the refractive index of the ionosphere can deviate substantially from 1, and at $\omega \cong \omega_p$ a strong dispersion arises. Indeed, the formula of the refractive index looks like this:

$$n = \sqrt{1 - \frac{\omega_p^2}{\omega^2}}. \qquad (6)$$

Let's look at the main features of this equation (which I've simply given you without deducing it step by step). If we increase the frequency of the radio waves, starting from the range $\omega < \omega_p$ to the values $\omega > \omega_p$, we can see that the properties of the plasma change drastically at resonance—that is, when $\omega = \omega_p$. When $\omega < \omega_p$, the radicand in equation (6) becomes negative and the refractive index is imaginary. This means that such low-frequency waves cannot propagate in plasma. Conversely, when $\omega > \omega_p$, then $n < 1$, and it approaches 1 as the frequency goes to infinity. This tendency of $n(\omega)$ to approach 1 as $\omega \to \infty$ is a characteristic property of any medium and not exclusively of plasma. It's explained by the fact that, due to inertia, electrons cannot oscillate at an infinitely high frequency. Thus secondary waves of extremely high frequencies are not generated in the medium, and the primary electromagnetic wave passes through the medium as if it were propagating in a vacuum.

When $n < 1$, equation (1) gives $v_p > c$, which means that the wave travels in plasma with *super-relativistic velocity*. It might seem that this relationship violates the basic tenet of Einstein's theory of relativity, which says that no event (physical body, field perturbation, or signal) can move with a velocity greater than the speed of light c. But no violation has actually occurred. The velocity calculated according to equation (1) has to do with a wave of a certain frequency. Such a wave is an infinite sinusoid, which cannot itself transmit a signal, because its shape is stable at all times.

In order to transmit a signal, one must use not one frequency but a group of frequencies from which it is possible to form a wave of the required shape. The rate of propagation of the entire group of waves is different from the velocity of a single wave and is determined by the equation

$$v_g = \frac{c}{n + \omega \dfrac{dn}{d\omega}}. \qquad (7)$$

To distinguish between these types of velocity, the velocity determined by equation (1) is called the *phase velocity* (thus the subscript "p" introduced above), while the velocity determined by equation (2) is called the *group velocity* (which explains the subscript "g" in equation (7)). In a vacuum, $n = 1$ and $dn/d\omega = 0$, so $v_p = v_g = c$. By calculating $dn/d\omega$ using equation (6), we can show that, in plasma,

$$v_p \cdot v_g = c^2.$$

Since $v_p > c$, then $v_g < c$, which means that radio signals propagate in plasma with a velocity less than c, which corresponds to the aforementioned tenet of the theory of relativity: as ω approaches ω_p from the higher frequencies, the phase velocity becomes infinitely large ($v_p \to \infty$) and the group velocity tends to zero ($v_g \to 0$). This means that signals with $\omega \le \omega_p$ do not exist (that is, do not propagate) in plasma.

Radio waves in magnetically active plasma

Magnetic fields exist everywhere in the cosmos. They are generated by electric currents (streams of charged particles) and by permanently magnetized heavenly bodies (Earth is one such). The basic magnetic field in the ionosphere is a geomagnetic field. It is this field that deflects the magnetic needle of a compass.

Now we want to clarify how equation (6) must be modified if plasma is immersed in a constant magnetic field (such plasma is called *magnetically active*). As we've seen, the dispersion properties of a medium—that is, the nature of the function $n(\omega)$—is closely linked with the natural frequencies of electron oscillations. In the absence of a magnetic field, plasma electrons move (oscillate) identically in every direction. The natural frequency of these isotopic oscillations (that is, oscillations that are independent of the direction of the velocity) is determined by equation (5). When plasma is placed in a magnetic field $\mathbf{B}_0$, the character of the electron movement will change drastically: a strong dependence on the velocity's direction $\mathbf{v}$ arises.

We recall that a magnetic field does not affect stationary charges or charges that move along the magnetic lines of force $\mathbf{B}_0$. However, charges that move perpendicular to $\mathbf{B}_0$ are affected by the Lorentz force in the direction perpendicular to $\mathbf{v}$ and $\mathbf{B}_0$. This force is equal to

$$F_m = ev_\perp B_0, \qquad (8)$$

where $v_\perp$ is the projection of the velocity vector $\mathbf{v}$ on the plane perpendicular to $\mathbf{B}_0$.

Let's decompose an arbitrary electron velocity $\mathbf{v}$ into its longitudinal and transverse components: $\mathbf{v} = \mathbf{v}_\parallel + \mathbf{v}_\perp$. The movement along $\mathbf{B}_0$ is not affected by the magnetic field, so the longitudinal velocity component $\mathbf{v}_\parallel$ doesn't "feel" the presence of $\mathbf{B}_0$. Thus longitudinal electron oscillations are characterized by the same frequency ω_p that had been obtained earlier (equation (5)).

The motion of electrons in the transverse plane, however, is quite different. Because the Lorentz force (equation (8)) is perpendicular to $\mathbf{v}_\perp$, the velocity doesn't change in magnitude, although the trajectory curves constantly. As a result, an electron revolves in a circle of radius ρ with angular frequency ω_m. Thus $v_\perp = \rho\omega_m$, and the force

$$F_m = e\rho\omega_m B_0.$$

Referring back to the initial equation (3) and substituting ω_0 for ω_m, l for ρ, and F_g for F_m, we get

$$\omega_m = \sqrt{\frac{F_m}{m\rho}} = \sqrt{\frac{e\omega_m B_0}{m}}.$$

Solving this equation yields

$$\omega_m = \frac{eB_0}{m}. \qquad (9)$$

This value is known as the *gyromagnetic frequency* or *Larmor precession frequency* of electrons. In the ionosphere, $B_0 \cong 4 \cdot 10^{-5}$ T and $\omega_m \sim 10^6\ s^{-1}$. This frequency belongs to the radio wave range, so Earth's magnetic field exerts a significant influence on radio signals propagating in the ionosphere if their frequency ω is close to ω_m.

Note that equations (5) and (9), which give the natural frequencies of electron oscillations in plasma, contain equal signs, unlike equation (4). This is because a strict calculation of the frequencies ω_p and ω_m yield the same result.

Now we need to write the formulas for the refractive index in magnetically active plasma. This is not an easy task, because the velocity of radio wave propagation depends on its direction relative to $\mathbf{B}_0$. In addition, the structure of the electric field in the wave (that is, its *polarization*) must also be taken into account. Here we're restricting ourselves to the simple case of longitudinal propagation along a magnetic line of force $\mathbf{B}_0$. Even in this case, though, there are two equations (not one) for the refractive indices:

$$n_1 = \sqrt{1 - \frac{\omega_p^2}{\omega(\omega - \omega_m)}}, \qquad (10)$$

$$n_2 = \sqrt{1 - \frac{\omega_p^2}{\omega(\omega + \omega_m)}}. \qquad (11)$$

(To avoid a complicated digression, I've simply given you the equations without showing how they are derived.)

Ordinary and extraordinary waves

The resonance condition formulated above as the coincidence of radio wave frequency with the natural frequency of electron oscillations (in the case where $\omega = \omega_m$) is a necessary but not a sufficient condition for resonance in a magnetic field. It's also necessary that the structure of the wave's electric field (polarization) correspond to the character of the electron motion. As the electrons revolve around the magnetic lines of force $\mathbf{B}_0$ (fig. 1), the electric field of the radio wave must also rotate. However, the electric field can rotate in one or the other direction depending on the mode of radio wave generation. The resonance at frequency $\omega = \omega_m$ occurs only in the case when the direction of revolution of the electrons coincides with direction of rotation of the electric field. As equation (10) shows, the refractive index $n_1(\omega)$ tends to infinity at resonance. In reality the increase of n_1 is limited, and we would find its maximum if we take into account the collision of electrons with other particles. We then find that a wave with a refractive index n_1 has some very odd properties, and for that reason it is referred to as *extraordinary*.

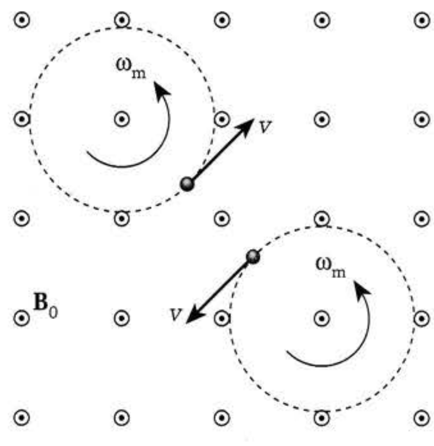


Figure 1

Electrons in a magnetic field, which revolve in the same direction for any initial velocity. The magnetic field $\boldsymbol{B}_0$ points toward the reader, perpendicular to the plane of the page.

In the opposite case, when the wave's electric field rotates in the direction counter to that of the electron's revolution, nothing unusual can be expected at $\omega = \omega_m$. Indeed, equation (11) for n_2 supports this conclusion. Generally, the properties of the wave n_2 are very similar to those of a wave propagating in plasma without a magnetic field. Thus it's called an *ordinary* wave.

The particular features of n_1 and n_2 waves manifest themselves clearly. Let's recall that in plasma without a magnetic field, a radio wave with frequency $\omega < \omega_p$ cannot propagate due to the negative radicand in equation (6). An ordinary wave n_2 shares the same property, although its low-frequency limit is determined by a slightly modified inequality: $\omega(\omega + \omega_m) < \omega_p$.

Extraordinary waves behave quite differently. If $\omega < \omega_m$, the second term in the radicand equation (10) becomes positive, so *at low frequencies no constraints on movement are imposed on an extraordinary wave*. It can easily be seen that the propagation of radio waves in the kilohertz range is limited by the following strong inequalities: $\omega \ll \omega_m$ and $\omega \ll \omega_p^2/\omega_m$. At low frequencies, equation (10) for n_1 can be simplified—the 1 in the radicand can be thrown out, which gives us

$$n_1 \cong \frac{\omega_p}{\sqrt{\omega\omega_m}}. \qquad (12)$$

We recall that the refractive index in magnetic free plasma is always less than 1. Here, though, $n_1 \gg 1$, due to the aforementioned inequalities. This means that the wave has a very small velocity compared to c ($v_p = c/n_1 \ll c$). To get a sense of its structure, draw the strength of the electric (or magnetic) field as an arrow $\mathbf{E}$ and follow its motion. If we made an animated film of our sketches, we would see that the tail end of the arrow moves with velocity v_p along the line of force $\mathbf{B}_0$, while its pointed end revolves around $\mathbf{B}_0$ with an angular velocity ω, and the vector $\mathbf{E}$ always stays perpendicular to $\mathbf{B}_0$ (fig. 2).

Simultaneous translation and rotation result in a helical trajectory of the arrowhead of vector $\mathbf{E}$. Thus its name—a helical wave, or simply helicon. There are other synonyms in the literature: whistling wave, whistle, or whistler. These words have nothing to do with the structure of the electric field—they describe the peculiar natural phenomena that result from the propagation of helical waves. These phenomena are called *whistling*

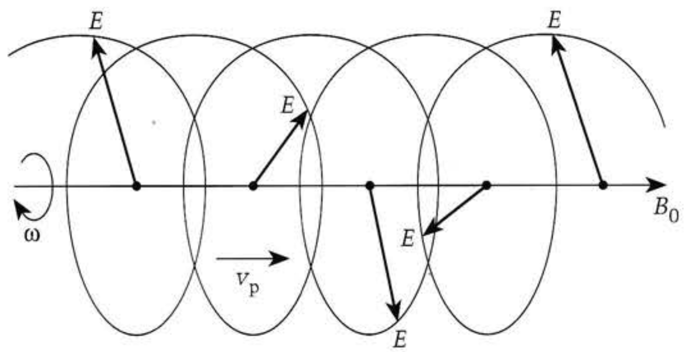


Figure 2

Helical wave propagating in plasma along a constant magnetic field $\mathbf{B}_0$ with a phase velocity $\mathbf{v}_p$. The arrowhead of vector $\mathbf{E}$ circumscribes a helix rotating with angular velocity ω.

atmospherics, and they are used in the near-Earth research mentioned at the outset of this article.

Whistling atmospherics

Atmospherics are natural electrical discharges in the atmosphere, generally induced by lightning. No doubt you've already encountered them, when you turned on an AM radio during a thunderstorm. The crackling noise is caused by atmospherics. The nature of this interference is well understood. In addition to a flash of light and a blast of sound, lightning also produces powerful electromagnetic radiation in a broad frequency range. This radiation was first detected in 1895, when Alexander Popov constructed his "thunderstorm detector."[2] The first receivers had a limited range of detection, but with the advent of the vacuum tube it became possible to detect lightning discharges at great distances. And so in 1919 the first reports of peculiar radio signals, dubbed "whistling atmospherics," began to appear.

Atmospherics are radio impulses of quickly changing frequency in the kilohertz range. A receiver for detecting them can be a simple low-frequency amplifier (even without a rectifier!). The amplified atmospherics are perceived as a kind of whistle, which explains their name.

The characteristic period of the frequency change in one pulse is between a fraction of second and one second, while the oscillation periods $T = 1/f$ are much shorter—of the order of 10^{-3}–10^{-4} s. So we can speak of an "instantaneous frequency" $f(t)$ for a given moment t. We'll see below that the dependence of frequency on time contains information about the properties of cosmic plasma thousands of kilometers from Earth.

As a rule, signals are not detected singly but rather as a series of pulses, one after the other, with intervals in the second range. Soon after the discovery of whistling atmospherics it became clear that the entire series was generated by a single atmospheric discharge and thus consisted of repeated echo signals. But how can we explain such long delays (of the order of a second or more)? There simply are no such distances on Earth! Even an around-the-world echo of a radio wave traveling at the speed of light returns in about 0.13 s. A persuasive hypothesis was offered in the early 1950s. It proposed that the pulses propagate from the lightning to the detector not along the Earth's surface but through outer space, along a line of force of Earth's magnetic field between two magnetically conjugated points[3]) (fig. 3). If the discharge occurred near the receiver, the primary atmospheric is recorded first (it travels along the Earth's circumference and is heard as a brief crack called a *spheric*. Later the receiver detects a whistle that traveled along the magnetic line of force to the opposite hemisphere and returned by the same route after reflection from Earth (a so-called "late whistler"). When the lightning and the receiver are located in different hemispheres, spherics are absent, and the first recorded signals are whistling atmospherics ("early whistlers"). In both cases, a repeatedly reverberated echo is heard (up to 20 repetitions) with an interval ratio of 2:4:6... for late whistlers and 1:3:5 ... for early whistlers (fig. 3).

The hypothesis that the signals travel along geomagnetic lines of force also explains such features of whistlers as the increase in the detection delay with the geographic (or more precisely, magnetic) latitude of the observation site and the fact that whistlers rarely occur in the low latitudes. Experiments conducted in 1958 with artificial sources of electromagnetic radiation provided valuable support for this hypothesis. Signals transmitted at a frequency of 15.5 kHz were received in the opposite hemisphere with a delay of $t \sim 0.7$ s. The whistlers recorded at these conjugated points at the same frequency were

[2]See S. M. Rytov's article "From the Prehistory of Radio" in the May/June 1990 issue of *Quantum*.

[3]Magnetically conjugated points are the points on the Earth's surface that lie on the same magnetic line of force.

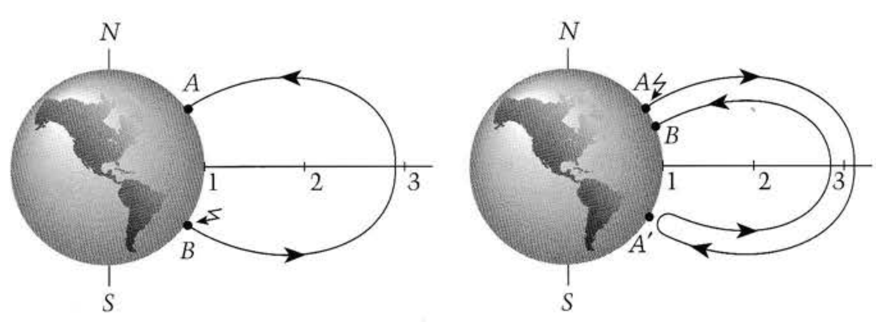


Figure 3

Diagrams illustrating the mechanism of generation of early (left) and late (right) whistlers. A is the location of the lightning discharge, B is the receiver, and A′ is the location of the signal's reflection. Distances are measured in Earth radii.

also delayed by the same value. The common delay corresponded to the calculated length of the geomagnetic line of force.

The observed delay can be explained with the help of previously obtained results. We recall that a signal propagates in plasma with an group velocity v_g given by equation (7). Using the simplified equation for the whistler wave refractive index (equation (12)), we can show that

$$\omega \frac{dn}{d\omega} = -\frac{n_1}{2}$$

and

$$v_g = 2v_p = \frac{2c}{n_1} = \frac{2c\sqrt{\omega\omega_m}}{\omega_p}.$$

The length of the track L and the delay t are linked by the usual relationship $L = v_g t$. It should be noted that the values of ω_m and ω_p vary along the track, as does the group velocity v_g. Therefore, in the aforementioned equation we must take some average value for v_g. We can use the following values for magnetic field intensity and electron concentration, which are characteristic for altitudes of several thousand kilometers: $B_0 \cong 4 \cdot 10^{-5}$ T, $N_0 \sim 10^9$ m^{-3}. For a frequency $\omega \sim 10^4$ s^{-1} we get $n_1 \sim 10$ and $v_e \sim 6 \cdot 10^7$ m $\cdot$ s^{-1}. Since the length of the magnetic line of force between the conjugated points in the our experiment was $L \sim 40{,}000$ km, the calculated value of the group velocity corresponds to a delay $t = L/v_g \sim 0.67$ s, which is in a good agreement with experimental values. This supports the hypothesis that whistlers are in fact the helical waves we considered above.

It's not too difficult to calculate also the relationship between the delay and the signal frequency. To this end we insert the previously derived value $v_g(\omega)$ into the equation $t = L/v_g$ and obtain $t = L\omega_p/2c\sqrt{\omega\omega_m}$. This equation is usually written as

$$t = \frac{D}{\sqrt{f}}, \qquad (13)$$

where

$$D = \frac{L\omega_p}{2c\sqrt{2\pi\omega_m}}.$$

The factor D is not influenced by frequency but does depend on plasma parameters and the length of the whistler's track. This coefficient is called the whistler's *dispersion*. It is found experimentally by analyzing the function $f(t)$—that is, the spectrogram of the whistlers (fig. 4). According to the value of D that is found, we can estimate the electron concentration and the magnetic field strength.

There is no doubt about the connection between whistling atmospherics and lightning discharges, but there is one apparent contradiction that should be examined. At any given moment, about 2,000 thunderstorms are raging in the Earth's atmosphere. These storms produce nearly 100 lightning flashes per minute on average. Yet the rate of detection of whistler in the temperate latitudes is only a few events per minute. Why are these numbers so drastically different? The answer is that a receiver located at a particular site doesn't "hear" most of the whistlers. To detect a whistler, a receiver must be located either near the lightning discharge or near its magnetically conjugated point. However, thunderstorms are distributed very nonuniformly in the Earth's atmosphere. They mostly occur in the equatorial regions, but in these regions whistlers are not observed due to the unsuitable geometry of the geomagnetic field (the lines of forces are located too close to Earth). There is another reason for the low rate of detected whistlers. Whistlers are generated only when the electromagnetic pulse is located near the magnetic line of force (when it "sticks" to it). Some special conditions are necessary for this—conditions that are not always met for any given lightning discharge. "Sticking" occurs when the plasma contains fluctuations running parallel to the magnetic field. However, even in homogeneous magnetically active plasma the direction of energy flow of the helical wave approaches the direction of $\mathbf{B}_0$. Calculations show that the limiting angle of deflection of the group velocity from this direction is about 20° when the frequency ω is much less than the gyrofrequency ω_m. Nevertheless, this condition cannot

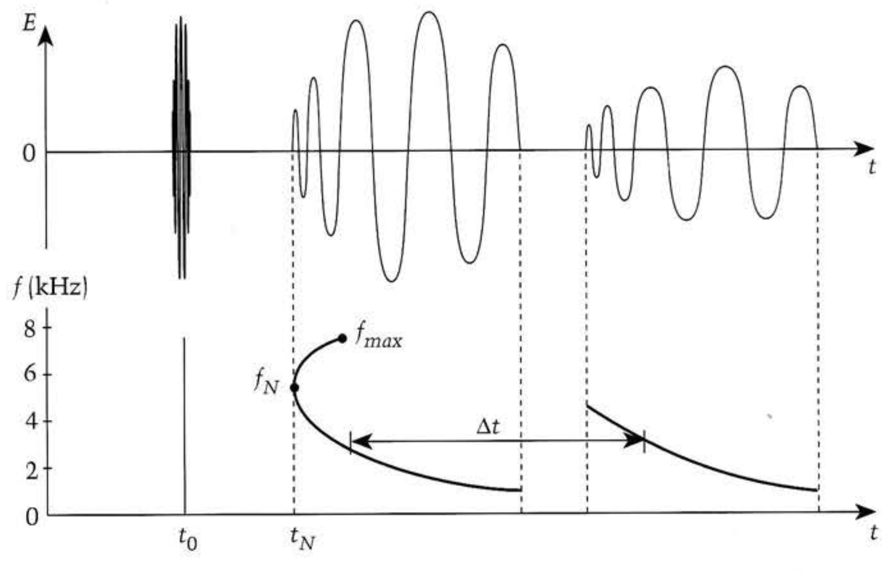


Figure 4

Whistling atmospherics (above) and its spectrogram (below). At a moment t_0 the lightning discharge simultaneously excites all the frequencies (in the low frequency range), which is shown by a vertical line in the lower graph. At a moment t_n the whistler's "nose" has arrived. After a certain delay Δt the signal is detected, after being reflected from the magnetically conjugated point on the opposite hemisphere.

ensure by itself the propagation of a whistler along the geomagnetic line of force. When whistling atmospherics are detected by a satellite, their number is much larger compared to terrestrial observations, because the apparatus on the satellite detects all the signals and not just those that were trapped in the waveguide.

Measuring space plasma characteristics

A geomagnetic line of force (the carrier of whistling atmospherics) can extend far from Earth—up to tens of thousands of kilometers, depending on the geomagnetic latitude of the observation point. This means that variations in a radio signal contain information about the characteristics of the plasma at very high altitudes. By calculating the signal delay in heterogeneous plasma, we can show that the dispersion D is integrated (accumulated) along the entire track of the pulse, and the contribution of those portions of the trajectory with small B_0 is particularly large. This becomes clear if we remember that, in the formula for D, the frequency ω_m stands in the radicand's denominator, so

$$D \sim \frac{1}{\sqrt{B_0}}.$$

Although it's not possible to determine the plasma density at various altitudes directly according to the dispersion D (since only an integral concentration along the entire track can enter into the equation), this relationship can be used to check one or another model describing the dependence of N_0 on altitude. It is particularly important that, due to the drastic decrease in B_0 with distance from Earth, the main contribution to dispersion is made by values of N_0 near the apogee of the trajectory (that is, at the highest altitudes).

Whistling atmospherics have played and continue to play a notable role in investigations of near-Earth space. Data obtained from such research led to a revision of views about how far Earth's ionosphere extended (the consensus had been a few thousand kilometers at most). The conclusion that high concentrations of electrons exist at great distances, which resulted from the analysis of whistlers, was directly confirmed later by the measurements made on board rockets and satellites.

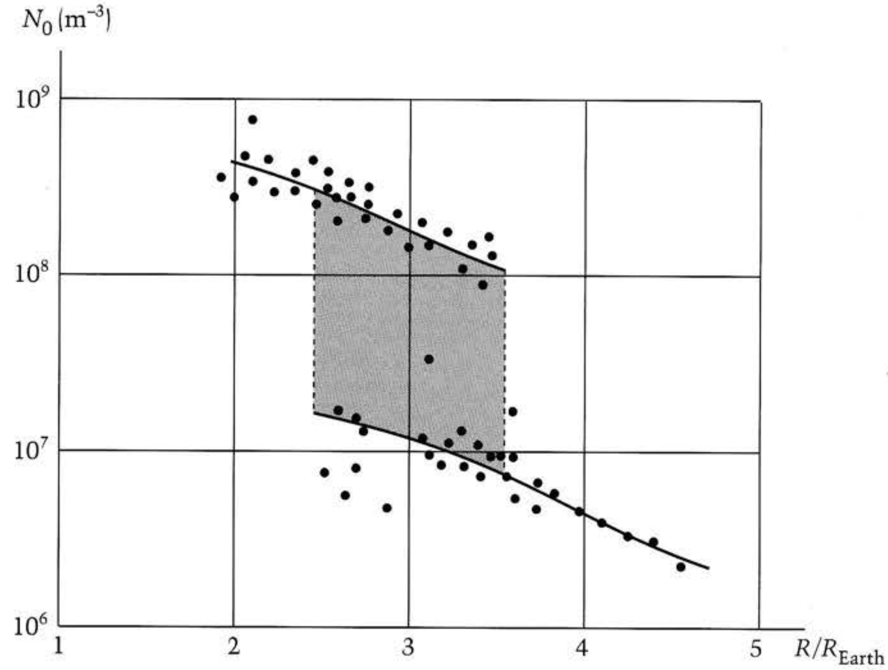


Figure 5

A drastic change in the electron concentration ("knee") at the boundary of the inner layer of the magnetosphere (shaded region). The data points were obtained by means of whistlers. The distance from Earth is measured in Earth radii.

The history of the discovery of the so-called "knee" (a drastic decrease in electron density at altitudes of 15,000–25,000 kilometers—see figure 5) is another example of fruitful research involving whistlers. The boundary of the inner region of the magnetosphere, full of relatively dense plasma ($N_0 \geq 10^8$ m^{-3}) and rotating together with Earth, is located at these altitudes. Evidence for the existence of such a vast plasma shell around Earth was obtained from the measurements made by a Soviet rocket in 1959 and by the American satellite Explorer I in 1963. The results obtained with the whistler technique also confirmed the existence of the "knee" and make it possible to record regularly the shell's boundary with an accuracy of 0.1 Earth radius.

I have illustrated the usefulness of whistlers by the example of electron concentration measurements. In doing so, I used equation (13) which is valid, as you recall, in a limited range of frequencies—about 1 to 7 kHz. At higher frequencies the condition $\omega \ll \omega_m$ is violated (we used it when we deduced the approximate equation (12) for $n_1(\omega)$, while the lower frequency limit is due to the fact that we neglect the motion of ions at the very outset.

If the frequency range for recording whistlers is made broader, new possibilities arise. I'll just make note of a few. Often whistlers are observed that have a minimal arrival time at a certain frequency. These are called "nose" whistlers (at the frequency f_n in figure 4). Two branches appear in the spectrogram at higher frequencies: an increase with an abrupt stop at some frequency f_{max}, and a decrease that enters the region we have been analyzing. In this frequency range, where $\omega \leq \omega_m$, the influence of the magnetic field is manifested

CONTINUED ON PAGE 37

$x \to \frac{1}{x}$
$y = \frac{1}{x}$
$a < x < b;$
$0 < y < \frac{1}{x}$
y
0
1
b
X
+
y
0
b
1
-
X

Squaring the hyperbola

A different approach to logarithms and exponents

by Andrey Yegorov

IN SCHOOL, THE EXPONENTIAL function $x \to a^x$ is usually introduced after several generalizations of the operation of raising numbers to powers. First, powers with natural exponents are defined, then powers with rational exponents, and finally, powers with irrational exponents. At this point, the logarithm is defined as a function that is the inverse to the exponential function.

This article will take the opposite approach: we'll start with a definition of the logarithm, and then we'll proceed to its inverse function, "the exponent." The definition we are going to give brings to light many of the basic properties of these functions and allows one to find estimates that are useful in physics. It also illustrates a frequently used mathematical method for producing new functions from those already known.

We'll be applying this method to the function $x \to 1/x$, whose graph $y = 1/x$, as you know, is a hyperbole.

DEFINITION. Let b be a positive number. Denote by $\ln b$ the real number whose absolute value is equal to the area of a figure bounded by the graph $y = 1/x$, the positive x-axis ($y = 0$), and the lines $x = 1$ and $x = b$. We take the sign of this area to be positive when $b > 1$ (fig. 1a) and negative when $b < 1$ (fig. 1b). If $b = 1$, we put $\ln b = 0$. We call the function $b \to \ln b$ the *natural logarithm*.

Art by Vasiliy Vlasov

For those who are familiar with the notion of integrals, I'll rewrite our definition as follows:

$$\ln b \overset{\text{def}}{=} \int_1^b \frac{1}{x}\,dx, \quad \text{if } b > 0.$$

Now, if you are eager to follow all of our reasoning in all its rigorous detail and solve all the problems (which are an important part of this article), you'll need a solid understanding of real numbers and their properties. However, the article is written in such a way as to make all the basic properties of logarithms and exponents understandable to anyone familiar with only the "naive" concepts of limit, real numbers, continuity, and so on, but not familiar with all the details involved in a rigorous definition of these concepts. At least, that was my aim.

Area

One of the notions that require a rigorous definition is that of area. Indeed, we must explain what we mean by the phrase "the area of a curvilinear trapezoid" (see figure 1). In school one hears about the area of a polygon, or the area of a circle and its parts, but our "trapezoid" is bounded by a hyperbola on one side.

Here I'll confine ourselves to the remark that area is a function definable for a rather large class of figures (this class includes all the polygons, all the convex bounded figures, and our "trapezoid" as well). This function must satisfy the following conditions:

1. The area of any figure is a positive number.
2. Equal figures have equal areas.
3. If we cut a figure into two parts

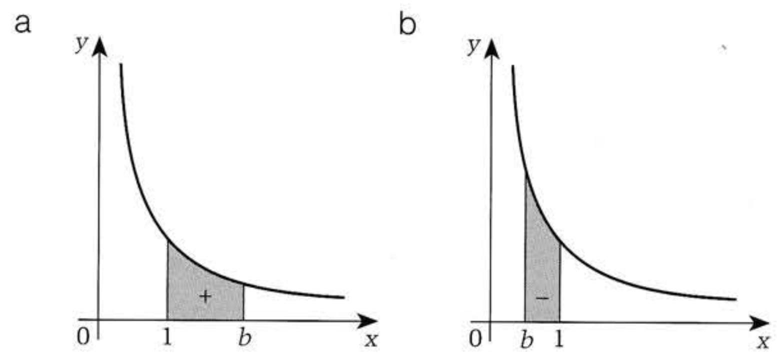


Figure 1

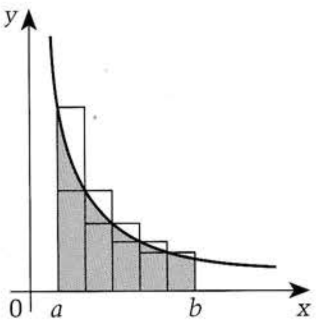


Figure 2

such that the area is defined for both of them, then the sum of these areas is equal to the area of original figure.

4. A rectangle with sides a and b has area ab.

I won't describe the class of figures for which such a function S exists, but I'll show that the area of a "curvilinear trapezoid" $a < x < b$, $0 < y < 1/x$ (fig. 2) is uniquely defined by the above conditions 1–4.

To this end, we'll divide the segment $[a, b]$ into n equal parts and construct two "stairways"—that is, steplike figures composed of rectangles with bases $(b - a)/n$ on the positive x-axis. One of these figures will contain the "trapezoid," while the other is contained in it (fig. 2). Let S_n' and S_n'' be the areas of these stairways. These areas are well defined by conditions 3 and 4. It's clear that $S_n' < S < S_n''$.

On the other hand, we can show that

$$S_n' - S_n'' = \frac{b-a}{n}\left(\frac{1}{a} - \frac{1}{b}\right). \qquad (1)$$

Indeed, if we let $(b - a)/n = k$, then this difference is the sum of the differences

$$k\left(\frac{1}{a} - \frac{1}{a+k}\right) + k\left(\frac{1}{a+2k} - \frac{1}{a+k}\right) + k\left(\frac{1}{a+3k} - \frac{1}{a+2k}\right) + \cdots + k\left(\frac{1}{a+nk} - \frac{1}{a+(n-1)k}\right).$$

This sum "telescopes" (most of the terms cancel out), and the result follows from the fact that $a + nk = b$.

Equation (1) shows that when n is big enough, the difference between S_n' and S_n'' is arbitrarily small. So there exists only one number S lying between S_n' and S_n'' for all n (it also follows from this that both sequences S_n' and S_n'' approach S as n increases—that is, S is their common limit).

This clarifies the concept of the area of the "curvilinear trapezoid" and along with it our definition of the function $y = \ln x$.

The natural logarithm

The fundamental property of the natural logarithm is expressed by the formula

$$\ln x_1x_2 = \ln x_1 + \ln x_2 \qquad (2)$$

(for $x_1 > 0$, $x_2 > 0$). This means that the natural logarithm of a product is equal to the sum of the natural logarithms of the factors.

Before we prove equation (2), let's establish one important property of curvilinear trapezoids formed by the function $y = 1/x$. Let $S[A, B]$ denote the area of a curvilinear trapezoid with vertices A and B (fig. 3). Then, if $b > a > 0$ and k is an arbitrary positive number, we have (fig. 4)

$$S[a, b] = S[ka, kb]. \qquad (3)$$

To prove this statement, consider the transformation of the plane that maps the point (x, y) into the point $(kx, y/k)$. (You can think of this as a combination of two mappings, the first expanding all the distances from the positive y-axis by the factor k, and the second constricting all the distances from positive x-axis by the same factor (fig. 4)). We can see that this transformation turns the trapezoid over the segment $a \le x \le b$ into that over $ka \le x \le kb]$ (fig. 5). In fact, if the point (x, y) belongs to the first trapezoid, then $a < x < b$ and $0 < xy \le 1$. But this means that $ka < kx < kb$ and $0 < (kx)y/k \le 1$—that is, the point $(kx, y/k)$ belongs to the second trapezoid. Conversely, if $(kx, y/k)$ is a point of the second trapezoid, then the point (x, y) belongs to the first.

Note that this mapping does not change the area of the trapezoid (or any other figure). Indeed, since rectangles whose sides are parallel to the axes do not change their areas (their bases are multiplied by k and their heights by $1/k$), the areas of the stairway figures associated with a curvilinear trapezoid do not change either. Thus the areas of curvilinear trapezoids will remain the same.

Therefore,

$$S[a, b] = S[ka, kb].[1]$$

Now it's not difficult to prove that $S[a, b] = \ln b - \ln a$. Figure 6 shows the different cases, depending on the sign of the area representing $\ln a$ and $\ln b$. But then

$$\ln b - \ln a = \ln kb - \ln ka. \qquad (3')$$

[1]That is, $\int_a^b \frac{1}{x}dx = \int_{ka}^{kb} \frac{1}{x}dx.$

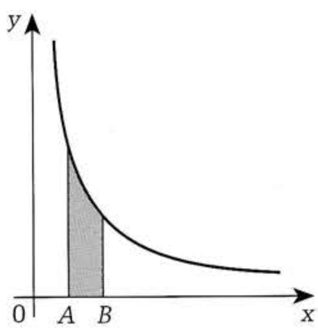


Figure 3

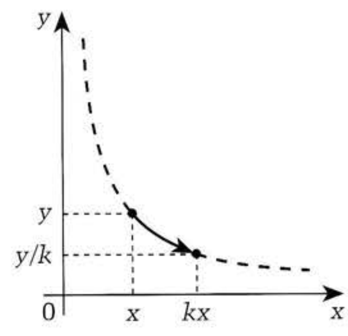


Figure 4

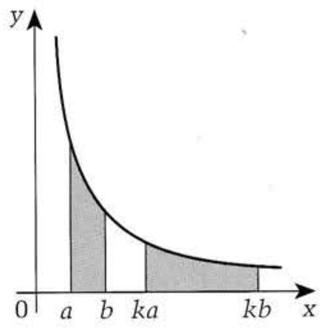


Figure 5

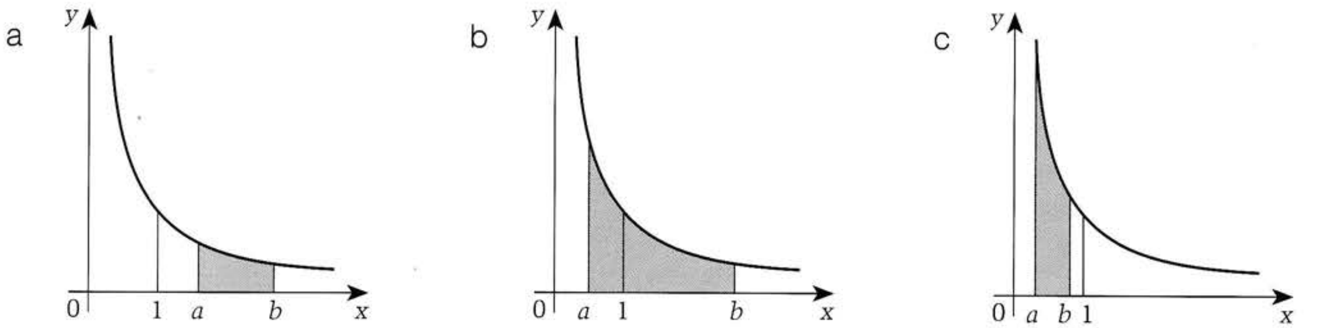


Figure 6

Although this formula has been proved so far only for $b > a$, it is true for all positive numbers a and b, since for $a > b$ we have

$$\ln a - \ln b = \ln ka - \ln kb,$$

which is equivalent to equation (3′).

Now the fundamental property of the logarithm follows directly from equation (3′). It suffices to take $b = x_2$, $a = 1$, $k = x_1$. In particular, when $x_2 = 1/x_1 = x$, we get

$$\ln x = -\ln\frac{1}{x}. \qquad (4)$$

The following equations can be derived without much trouble from the fundamental property (equation (2)):

$$\ln x_1 x_2 \ldots x_n = \ln x_1 + \ln x_2 + \ldots + \ln x_n, \quad (5)$$

$$\ln\frac{x_1}{x_2} = \ln x_1 - \ln x_2, \qquad (6)$$

We'll be using these equations below.

Graph of $x \longrightarrow \ln x$

Equation (2) allows us to learn more about the behavior of the function $y = \ln x$. First, let's check that $\ln x$ grows indefinitely as x increases. In fact, since $\ln 2 > 0$, and by virtue of equation (5), $\ln 2^n = \ln (2 \cdot 2 \cdot \ldots \cdot 2) = \ln 2 + \ln 2 + \cdot + \ln 2 = n \ln 2$, then $\ln 2^n$ grows without limit as n increases, and this implies unlimited growth of $y = \ln x$. (Indeed, $\ln x > n \ln 2$ for $x > 2^n$—I leave it to the reader to prove that $\ln x$ increases monotonically.)

Now let's examine the logarithm's behavior as x approaches 0. Since

$$\ln\frac{1}{2^n} = -\ln 2^n = -n\ln 2,$$

we see that $\ln x < -n \ln 2$ for $0 < x < 1/2^n$—that is, if x is small enough, the logarithm can be arbitrarily large negative number.

Now we're ready to draw an approximate graph of $y = \ln x$ (fig. 7).

The exponent

We can prove that each real number is a value of the function we examined above. In addition, this function takes each value only once—that is, for each real x there is only one solution y of the equation $x = \ln y$. The real number y that satisfies this equation is denoted by

$$y = \exp x.$$

Thus we have obtained a new function $x \to \exp x$ that is inverse to $x \to \ln x$. It's called the *exponent*. Note that the very definition of the exponent implies the identity

$$\exp (\ln x) = \ln (\exp x) = x. \qquad (7)$$

The graph of the function $y = \exp x$ is symmetric to the graph of $y = \ln x$ with respect to the line $y = x$. Indeed, since the equality $y = \exp x$ is equivalent to $\ln y = x$, we can obtain the exponent's graph from that of the logarithm by means of a transformation that maps the point (x, y) into the point (y, x). And this mapping is simply the symmetry with respect to the line $y = x$ (fig. 8).[2]

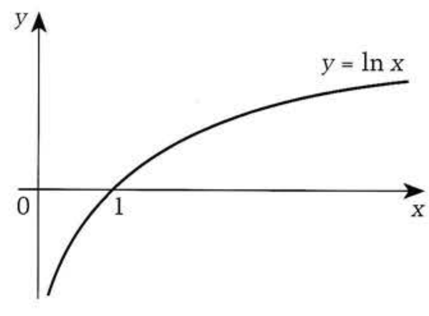


Figure 7

Thus the exponent is an increasing function, defined on the number line $-\infty < x < +\infty$ and taking positive values. In addition, it takes arbitrarily large values as x increases and tends to zero as x approaches $-\infty$.

Its fundamental property

The fundamental property of the exponent is expressed by the following equation:

$$\exp (x_1 + x_2) = \exp x_1 \cdot \exp x_2. \quad (8)$$

To prove equation (8) we'll use the fundamental property of the natural logarithm and equation (7). Since $\ln y_1 y_2 = \ln y_1 + \ln y_2$, then for $y_1 = \exp x_1$, $y_2 = \exp x_2$, we have $\ln y_1 y_2 = x_1 + x_2$—that is, $y_1 \cdot y_2 = \exp (x_1 + x_2)$, or $\exp x_1 \cdot \exp x_2 = \exp (x_1 + x_2)$. From the fundamental property of the exponent (or from the corresponding property of the logarithm), we can derive the equalities

$$\exp(-x) = \frac{1}{\exp x}$$

[2]To verify that this figure is correct, prove that $\ln x < x$ for all $x > 0$.

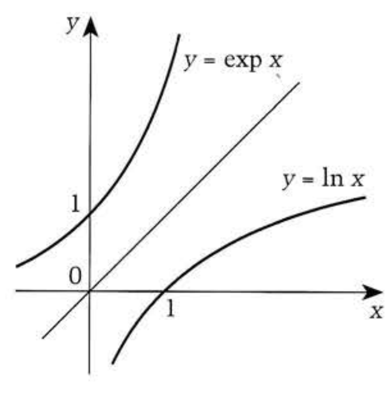


Figure 8

and

$$\exp(x_1 + x_2 + \ldots + x_n) \\ = \exp x_1 \cdot \exp x_2 \cdot \ldots \cdot \exp x_n.$$

We'll use the symbol e for the number exp 1. In other words, e is the solution of the equation $\ln e = 1$.

Using the fundamental property of the exponent, we'll prove that for all rational $x = m/n$, $\exp x = e^x$. First of all, for natural m,

$$\exp m = \exp(1 + 1 + \ldots + 1) \\ = \exp 1 \cdot \exp 1 \cdot \exp 1 = e^m$$

and

$$\exp(-m) = \frac{1}{\exp m} = \frac{1}{e^m} = e^{-m}.$$

So $\exp m = e^m$ for every whole number m. Moreover, for natural numbers n,

$$\left(\exp\frac{1}{n}\right)^n = \exp\left(\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}\right) \\ = \exp 1 = e$$

—that is,

$$\exp\frac{1}{n} = \sqrt[n]{e}.$$

For any rational $x = m/n$ (where $n > 0$, and n and m are integers) we get

$$\exp\frac{m}{n} = \left(\exp\frac{1}{n}\right)^m = \left(\sqrt[n]{e}\right)^m = e^{m/n}$$

—that is, $\exp x = e^x$ for all rational x.

If α is an irrational number, it's most convenient to regard the equation $e^\alpha = \exp \alpha$ as a definition of e^α. So, for all x,

$$\exp x = e^x. \qquad (9)$$

This equation, together with the equation (7), means that for any positive x, its logarithm is such a number that e raised to this power gives x:

$$e^{\ln x} = \exp \ln x = x.$$

Arbitrary bases

Now we can define $y = a^x$ for any $a > 0$, $a \neq 1$, and all x, setting

$$a^x = \exp(x \ln a) = e^{x \ln a},$$

as well as the logarithm of $x > 0$ with base a, setting

$$\log_a x = \frac{\ln x}{\ln a}.$$

One can show that these functions satisfy the same fundamental properties:

$$a^{x_1 + x_2} = a^{x_1} \cdot a^{x_2},$$
$$\log_a x_1 x_2 = \log_a x_1 + \log_a x_2.$$

(for $x_1 > 0$, $x_2 > 0$), and that

$$a^{\log_a x} = x.$$

We leave it to the reader to check that the definitions given above correspond to the traditional ones. In what follows, however, we'll primarily be interested in those properties of logarithms and exponent for which it is essential that the base is e and not any other number.

Derivatives

Up to this point we've been discussing properties of logarithms and exponential functions that you might already know from your high school studies. But now we're going to look at properties connected with the rates of variation of these functions. These properties are very important for physics. Let's begin with some definitions. We'll call the difference $\Delta f = f(x_1) - f(x_0)$ the increment of the function $x \to f(x)$ on the segment $[x_0, x_1]$, and the difference $\Delta x = x_1 - x_0$ the increment of the argument. It's natural to call the quotient

$$\frac{\Delta f}{\Delta x} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

the mean velocity of variation of the function $x \to f(x)$ on the segment $[x_0, x_1]$. Indeed, if x is time and $f(x)$ is the distance traveled by a moving body up to the moment x, then $\Delta f/\Delta x$ is the average speed during the interval $[x_0, x_1]$. Suppose that x_0 is fixed and x_1 approaches x_0. If the quotient $\Delta f/\Delta x$ tends to some limit, we'll call this limit the derivative of function $x \to f(x)$ at the point x_0 and denote it by $f'(x_0)$. (We can regard this as the instantaneous rate of the function's change at the moment x_0.) The derivative describes the function's behavior near x_0.

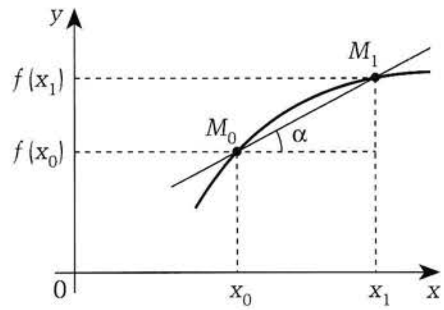


Figure 9

The derivative has a simple geometric meaning. Let M_0 be a point on the graph $y = f(x)$, corresponding to $x = x_0$—that is, the point with coordinates $(x_0, f(x_0))$; M_1 is the point $(x_1, f(x_1))$. Draw the line M_0M_1. The slope of this line—that is, the tangent of its inclination to the positive x-axis (fig. 9)—equals

$$\tan\alpha = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

As x_1 approaches x_0, the secant line tends to a limiting position—the tangent drawn to the curve $y = f(x)$ at the point x_0. Thus the derivative at $x = x_0$ is equal to the slope of the tangent to the graph $x \to f(x)$ drawn through the point $(x_0, f(x_0))$.

Let's try to calculate the derivative of the function $f(x) = \ln x$. Its increment on the segment $[x_0, x_1]$ equals

$$\ln x_1 - \ln x_0 = \ln\frac{x_1}{x_0} \\ = \ln\left(1 + \frac{x_1 - x_0}{x_0}\right).$$

Therefore,

$$\frac{\ln x_1 - \ln x_0}{x_1 - x_0} = \frac{1}{x_0}\frac{\ln(1+h)}{h}, \qquad (10)$$

where

$$h = \frac{x_1 - x_0}{x_0}.$$

As x_1 tends to x_0, h approaches 0. So we just need to find how the quotient $[\ln(1+h)]/h$ behaves for small h.

The definition of the logarithm given above will soon allow us to make some simple evaluations, from which it will follow that $[\ln(1+h)]/h$ approaches 1 as h tends to zero—that is,

$$\ln(1+h) \cong h \text{ (for small } h). \quad (11)$$

This equation is very important for making estimates. We can rewrite it as

$$\exp h \cong 1 + h \text{ (for small } h). \quad (12)$$

From these we obtain basic equations for the derivatives of logarithms and exponents:

$$(\ln x)' = \frac{1}{x} \quad (13)$$

and

$$(\exp x)' = \exp x. \quad (14)$$

Evaluating the natural logarithm near 1

Here is one of the most important properties of the logarithm:

$$x - \frac{x^2}{x+1} = \frac{x}{x+1} < \ln(1+x) < x. \quad (15)$$

In order to prove these inequalities for $x > 0$, we need only compare the area of the curvilinear trapezoid $ABCD$ with the areas of rectangles $AD'CB$ and $ADC'B$ (fig. 10). The trapezoid's area is $\ln(1+x)$, while the area of $AD'CB$ is equal to

$$AB \cdot BC = x \cdot \frac{1}{x+1},$$

and the area of $ADC'B$ is $AB \cdot AD = x \cdot 1 = x$. Readers are invited to check this estimate for $-1 < x < 0$ on their own.

Inequalities (15) allow us to evaluate the average rate at which a logarithm changes (see equation (10)). Let $x = (x_1 - x_0)/x_0$ in equation (15). We get

$$\frac{(x_1 - x_0)/x_0}{(x_1 - x_0)/x_0 + 1} = \frac{x_1 - x_0}{x_1} < \ln\frac{1 + x_1 - x_0}{x_0} < \frac{x_1 - x_0}{x_0},$$

or

$$\frac{1}{x_1} < \frac{\ln(1 + (x_1 - x_0)/x_0)}{x_1 - x_0} < \frac{1}{x_0}$$

(for $x_1 > x_0$). When x_1 tends to x_0, this rate converges to $1/x_0$, thus proving equation (13) for the derivative of a logarithm. So we have shown that there exists a tangent to the logarithm's graph at every point, and that its slope is equal to $1/x_0$ for $x = x_0$.

To find the slope of a tangent drawn to the exponent's graph, we recall that this graph is symmetric to the logarithm's with respect to the line $y = x$. As you can see in figure 11, $\alpha_0 + \beta_0 = \pi/2$, and since α_0 is the inclination of the tangent drawn to the logarithm's graph at the point (x_0, y_0), we have $\tan\alpha_0 = 1/x_0$. Finally we obtain the following:

$$\tan\beta_0 = \tan\left(\frac{\pi}{2} - \alpha_0\right) = \frac{1}{\tan\alpha_0} = x_0 = e^{y_0}$$

—that is, the slope of a tangent drawn to the exponent's graph for each x equals the exponent's value at x, thus proving equation (14).

For arbitrary bases, without much difficulty we obtain

$$(\log_a x)' = \frac{1}{x}\ln a, \quad (13')$$

$$(a^x)' = a^x \ln a. \quad (14')$$

Note that both these equations are valid for natural logarithms and exponential functions with base e. For other exponential functions, the instantaneous rate of change is proportional to (not equal to) the value at point x_0.

Consider the bounds of equation (15) once more. We see at once that the approximate equality (11) and, therefore, its equivalent (12), follow directly from them. In fact, when $|x| \ll 1$, the quantity x^2 is negligibly small compared to x. For example when $|h| \ll 0.1$, the relative error of equations (11) and (12) is not greater then 1%—that is, the difference between the quotient of their left and right parts and 1 is not greater than one hundredth.

We can also derive from the same bounds of equation (15) the following remarkable equation for the number e:

$$e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n$$

and more generally, the following equation for the exponent:

$$\exp x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n \quad (16)$$

(see problem 1 below).

Series representation of e^x

Equation (16) is rather cumbersome to use in calculations, because we must take very big n if we want to achieve good precision. In this section, we'll look at another expression for the exponent, one that represents it as the sum of an infinite series:

$$e^x = 1 + x + \frac{x^2}{1\cdot 2} + \cdots + \frac{x^k}{1\cdot 2\cdot \ldots \cdot k} + \cdots, \quad (17)$$

just as the geometric progression $1 + x + x^2 + \ldots$, $|x| < 1$, represents the function $y = 1/(1 - x)$.

This decomposition is convenient for two reasons. First, if you want to calculate the exponent, you

CONTINUED ON PAGE 56

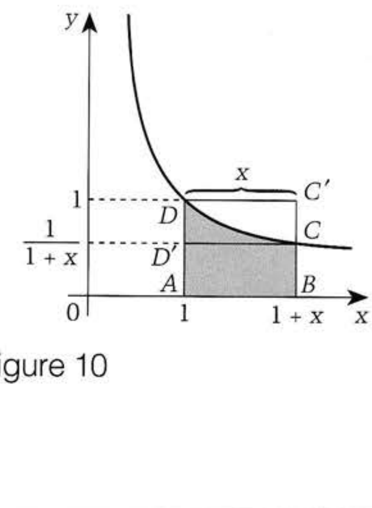


Figure 10

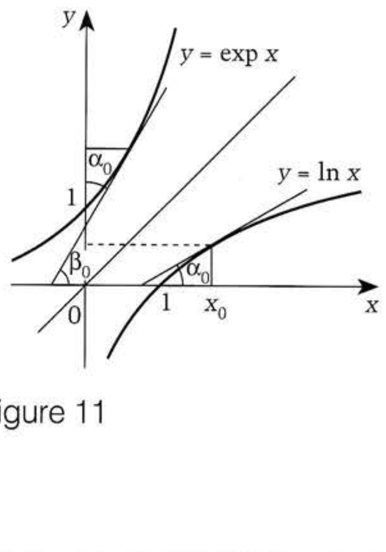


Figure 11

KALEIDOS

Volumes witho

Call in the Cavalie

N THIS ARTICLE WE'LL DERIVE several formulas for finding volumes without the use of integrals. We'll rely heavily on a principle formulated by a pupil of Galileo in 1629.

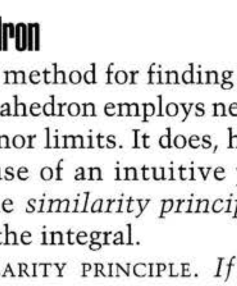
Tetrahedron

This method for finding the area of a tetrahedron employs neither integrals nor limits. It does, however, make use of an intuitive proposition, the *similarity principle*, that masks the integral.

SIMILARITY PRINCIPLE. *If all the edges of a tetrahedron are multiplied by the factor k, its volume will be multiplied by the factor k^3.*

In particular, if all the edges are multiplied by 2, the volume will be multiplied by 8.

Before we prove the formula for the volume of a tetrahedron, let's recall two formulas for the volume of a trihedral prism. Suppose that S is the area of its base and h is its altitude (fig. 1). Then its volume is



Figuure 1

$V = Sh$. We can also express the volume of the prism using the area of its lateral face Q and the distance d from this face to the opposite edge: $V = {}^1/_2 Qd$ (fig. 2).

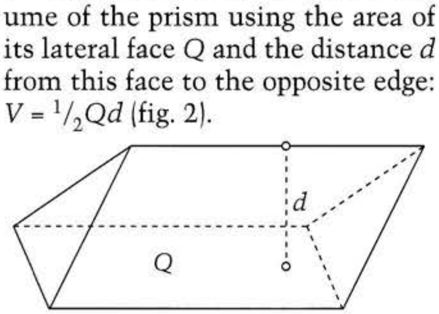


Figure 2

Now let's look at the tetrahedron $ABCD$, in which the area of ABC is S and the altitude drawn to this face is h (fig. 3). Let V be the volume of the tetrahedron. Denote the midpoints of the tetrahedron's edges by the letters K, L, M, N, P, T as in figure 3.

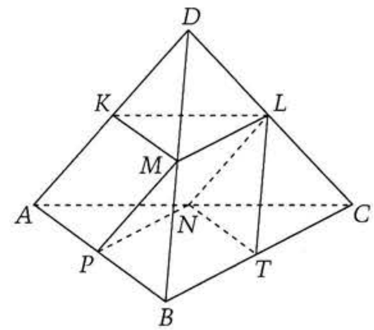


Figure 3

Decompose the tetrahedron into four polyhedrons: two tetrahedrons $DKML$ and $LNTC$ whose edges are half those of the of the original tetrahedron, and two trihedral prisms $APNKML$ and $PMBNLT$. According to the similarity principle, the volumes of the tetrahedrons $DKML$ and $LNTC$ are both $V/8$. Let's express the volume of each prism using their corresponding formulas. The volume of the first prism is

$$\frac{S}{4}\cdot\frac{h}{2}=\frac{1}{8}Sh,$$

and that of the second is

$$\frac{1}{2}\cdot\frac{S}{2}\cdot\frac{h}{2}=\frac{1}{8}Sh.$$

We obtain the following equation for the volume V:

$$V=\frac{V}{4}+\frac{1}{4}Sh.$$

From this equation we obtain the formula we seek:

$$V=\frac{1}{3}Sh.$$

Sphere

To derive the formula for the volume of a sphere, we'll use another well-known principle.

CAVALIERI'S PRINCIPLE. *Two bodies lying between two parallel planes have the same volume if their cross sections cut off by any plane parallel to these planes have equal areas.*

Consider four solids: a sphere of radius R, a cylinder circumscribed around the sphere, and two perpendicular cones. The vertices of the cones coincide with center of the sphere, and their bases coincide with those of the cylinder. Figure 4 shows a section of this solid, formed by a plane passing through the axis of the cylinder. We will use this picture to illustrate how the formula for the sphere's volume can be derived from Cavalieri's principle, if we know the formulas for the volume of a cylinder and a cone.

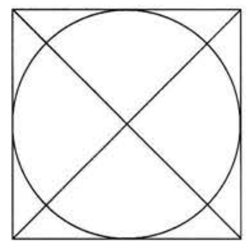


Figure 4

Let's consider the sphere and the cylinder separately, and think of both cones as removed from the cylinder. (The sections of these two bodies are drawn separately in the figure 5. The sphere lies between the bases of the cylinder.)

Take any plane parallel to the bases of the cylinder such that the

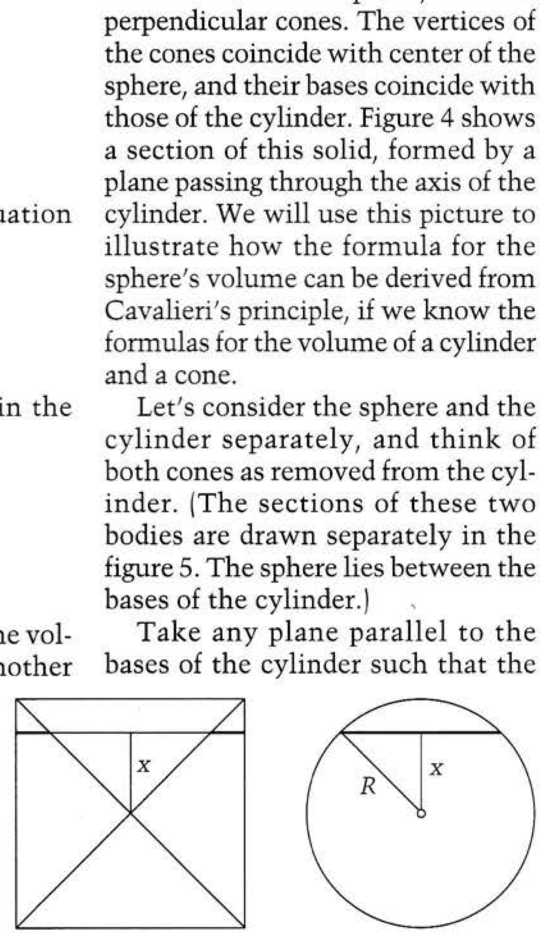


Figure 5

hout integrals

valieri principle!

distance between it and the sphere's center is x. This plane cuts a circle of radius $\sqrt{R^2 - x^2}$ from the sphere. The area of this circle is $\pi(R^2 - x^2)$. The intersection of the plane and the second body is a round ring with external radius R and internal radius x. It's clear that the area of the ring is also equal to $\pi(R^2 - x^2)$. According to Cavalieri's principle, we can equate the volume of the sphere to that of the cylinder minus the volumes of the two cones:

$$V = 2\pi R^3 - \frac{2}{3}\pi R^3 = \frac{4}{3}\pi R^3.$$

Intersecting cylinders

Let's solve the following problem:

Calculate the volume of the common part of three infinite cylinders of radius R such that their axes lie in one plane, pass through one point, and form equal angles with each other (fig. 6).

Solution. Consider a sphere of radius R whose center is at the point where the axes of the cylinders intersect (fig. 7). Draw an arbitrary plane parallel to the one where the axes lie. The section of the common part formed by this plane is a regular hexagon circumscribed around a circle that is a section of the sphere. The ratio of the areas of these sections does not depend on the plane chosen and is equal to $2\sqrt{3}/\pi$. Thus the ratio of the volumes of the common part and the sphere is the same. So the volume we seek is $(8\sqrt{3}/3)R^3$.

This notion allows one to solve a similar problem, when the number of cylinders is not three but an arbitrary integer n (that is, the axes of the cylinders lie in one plane, pass through one point and the angle between adjacent axes is π/n).

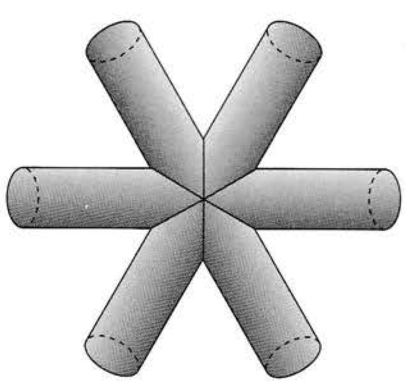


Figure 6

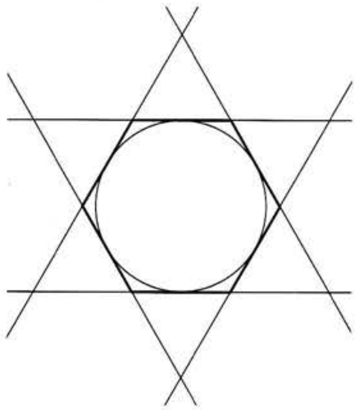


Figure 7

A rotation problem

Cavalieri's principle can help us solve the following problem as well:

Given an isosceles triangle with base a and altitude h drawn to the base. Let the altitude belong to the line l. The line m intersects l and lies in the plane perpendicular to that of the triangle, so that the angle between m and l is α. Find the volume of the body that arises as a result of the triangle's rotation about m (fig. 8).

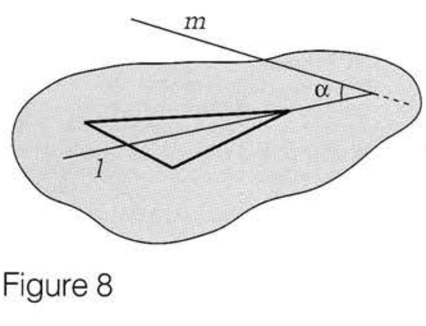


Figure 8

Using Cavalieri's principle, we can show that the volume we seek is equal to that of a cone with a circular base of radius a and altitude $h \cos \alpha$. To check this, let's project the altitude of the triangle onto line m. We'll obtain a segment of length $h \cos \alpha$. Construct an isosceles triangle in which this segment is the altitude and the base is equal to a. (The vertex opposite the base in this triangle is the projection of the corresponding vertex of the original triangle.) If the constructed triangle is rotated about m, the outcome will be the aforementioned cone. A plane perpendicular to m cuts the solid whose volume we seek along a ring, while a cone cuts it along a circle. Now it suffices to show that the areas of these figures are equal, and we find

$$V = \frac{\pi a^2 h \cos \alpha}{12}.$$

A distance problem

And, to conclude the article, here's a simple problem whose solution is based on common sense:

A plane convex figure is given, with area S and perimeter L. Find the volume of the body consisting of all points in space whose distance from the figure is less than or equal to a given positive number d.

This problem is easy to solve when the given figure is a plane polygon. In this case, the volume we seek consists of two prisms of height d and bases of area S, a number of half-cylinders of radius d and total height L, and several segments of a sphere of radius d whose union is the whole sphere (these segments adjoin the vertices of the polygon and their surfaces consist of spherical 2-gons and two semicircles). Thus for a polygon we can write the following formula for the volume:

$$V = 2Sd + \frac{1}{2}\pi d^2 L + \frac{4}{3}\pi d^3.$$

In fact, it can be shown that the same formula holds for an arbitrary convex figure.

—I. F. Sharygin

Mars or bust!

"I've always wanted to see a Martian," said Michael. "Where are they, Dad? You promised."

"There they are," said Dad, and he shifted Michael on his shoulder and pointed straight down.

—Ray Bradbury, The Martian Chronicles

by Arthur Eisenkraft and Larry D. Kirkpatrick

HAVE YOU EVER WANTED TO go to Mars? Mars is the next frontier. Those of you who are too young to have watched the Herculean efforts to send the first humans to the Moon may be able to participate in the next big space exploration. You may be an astronaut, an engineer, or a computer analyst helping with the mission. Thousands of people will be required. Recently the public's interest in Mars was heightened by NASA's announcement that scientists may have found evidence for the existence of primitive life on ancient Mars.

Sending humans to Mars will require a lot of preparation. The work has already begun. Besides the crucial work on studying how humans live in space for long periods of time, two recent launches have sent satellites to our nearest planetary neighbor.

On November 7, 1996, NASA launched the Mars Global Surveyor (MGS), which will reach Mars on September, 12, 1997, to begin a two-year survey of the atmosphere and surface of Mars from orbit. The MGS will require 309 days to make its journey. (You can learn more about the Mars Global Surveyor mission on the Web at http://mgs-www.jpl.nasa.gov.)

On December 4, 1996, NASA launched Mars Pathfinder on a trajectory so that it will land on the surface of Mars on Independence Day, a trip of 212 days. Mars Pathfinder contains a microrover that will be used to deploy scientific instruments and explore the terrain around the landing site. (You can learn more about Mars Pathfinder at http://mpfwww.jpl.nasa.gov.)

It's interesting that the satellite that was launched last arrives at Mars first. We can get some understanding of this by looking at a simplified orbital problem.

Let's assume that we have a satellite in a circular orbit around the Sun with a radius equal to the average radius of the Earth's orbit. Let's fire rockets in the forward direction tangent to the orbit. If we increase the speed of the satellite by the correct amount, the satellite will be placed into an elliptical orbit that has its greatest distance from the Sun equal to the average radius of Mars's orbit. If Earth and Mars are in the proper relative positions, this would allow the satellite to orbit or land on Mars. In our calculations, we neglect the gravitational effects of Mars and the Earth and consider only the Sun's gravity.

We can find the required speed of the satellite using conservation of energy

$$\frac{1}{2}mv_E^2 - \frac{Gmm_S}{r_E} = \frac{1}{2}mv_M^2 - \frac{Gmm_S}{r_M}$$

and conservation of angular momentum

$$mv_E r_E = mv_M r_M,$$

where m and m_S are the masses of the satellite and the Sun, respectively; r_E and r_M are the orbital radii of Earth and Mars, respectively; and v_E and v_M are the orbital radii of Earth and Mars. Note that these two velocities occur at the ends of the ellipse—that is, when $r = r_E$ and $r = r_M$, respectively, and that the velocities are perpendicular to the radii. Solving these two equations for v_E, we obtain

$$v_E = v_0\sqrt{\frac{2r_M}{r_E + r_M}},$$

where

$$v_0 = \sqrt{\frac{Gm_S}{r_E}}$$

is the orbital speed of the Earth. Using $r_M = 1.53\ r_E$, we find that

Art by Tomas Bunk


PATHFINDER 1
MARS
1
TOM BUNK

$v_E = 1.10\ v_0$. Knowing that $G = 6.67 \cdot 10^{-11}\ N \cdot m^2/kg^2$, $m_S = 2.0 \cdot 10^{30}$ kg, and $r_E = 1.5 \cdot 10^{11}$ m, we obtain numerical values of $v_0 = 29.7$ km/s and $v_E = 32.7$ km/s. Therefore, we must increase the satellite's forward speed by 3.0 km/s.

We can do a similar calculation at the other end of the ellipse to find out how much we need to accelerate the satellite to match the orbital speed of Mars. Conservation of angular momentum tells us that $v_M = 21.4$ km/s and that Mars's orbital speed is 24.1 km/s. Therefore, the satellite needs to speed up by 2.7 km/s.

Because Kepler's laws are applicable for any object orbiting the Sun, we can use Kepler's third law to find out how long it takes the satellite to reach Mars. Let's compare the circular orbit of the Earth to this transfer ellipse connecting Earth and Mars with a major axis equal to $r_E + r_M$:

$$\left(\frac{T_t}{T_E}\right)^2 = \left(\frac{r_E + r_M}{2r_E}\right)^3.$$

Therefore, $T_t = 1.42\ T_E = 1.42$ years. Because the satellite only executes one half of the elliptical orbit, the time is 0.71 years = 260 days. Longer and shorter periods can be obtained by using different ways of leaving Earth orbit and entering Mars orbit. For a better calculation, we also need to take into account the gravitational fields of Mars and Earth.

Our problem this month is based on a problem that appeared on the second exam used to select the members of the 1995 US Physics Team, which won four gold medals and one silver medal at the International Physics Olympiad in Australia.

Let's assume that the Mars Global Surveyor is in a circular orbit about Mars at the designed height of 367 km above the surface. We also assume that we can neglect the effects of Mars's atmosphere and that Mars has a radius $R = 3{,}400$ km and a surface gravity $g = 3.72\ m/s^2$.

A. Find the speed of the satellite in its circular orbit about Mars in terms of the values given above.

Although the Global Surveyor is not designed for this purpose, let's assume that we want to send the satellite down to the Martian surface. The satellite could reach the surface by firing its rocket engines for a short period of time. We will consider two special cases.

B. In the first method, the retrorockets are fired at point X tangent to the orbit to slow the satellite. The circular path becomes an elliptical path that brings the satellite to a landing strip on the Martian surface at point A on the side opposite to point X, as shown in figure 1.

(i) Determine the speed of the satellite immediately after the retrorockets have been fired.

(ii) Determine the speed of the satellite as it reaches Mars's surface at point A.

C. In the second method the rockets are fired at point X perpendicular to the orbit, giving the satellite a momentum directed toward Mars. The circular path becomes an elliptical path that brings the satellite to a landing strip on the lunar surface at point B one quarter of the way around Mars, as shown in figure 2.

(i) Determine the speed of the satellite as it reaches Mars's surface at point B.

(ii) Determine the velocity of the satellite immediately after the rockets have been fired.

D. How do the magnitudes of the changes in velocity at point X compare for the two methods?

E. How do the speeds of the satellite at Mars's surface compare for the two methods?

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

The bombs bursting in air

In the September/October issue of *Quantum* we asked our readers to analyze the paths of fragments emitted in a fireworks display. Specifically, we asked at what time is the frequency of fragments hitting the ground the greatest? To simplify the problem, we assumed that there was no air resistance and that the explosion was isotropic. We will follow the solution given a few years ago by Tainan Wang, an undergraduate student at SUNY Stony Brook and former member of the Chinese Olympiad Team.

Assuming that all the fragments have the same speed relative to the center of mass, then the fragments will form a sphere with the center of mass as its center. The radius of this sphere will increase in proportion to the time, as the entire sphere descends toward the ground with acceleration g.

Figure 3 shows the expanding sphere at times t and $t + \Delta t$. During the time interval Δt, all fragments

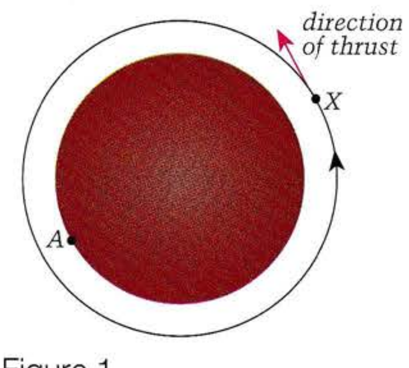


Figure 1

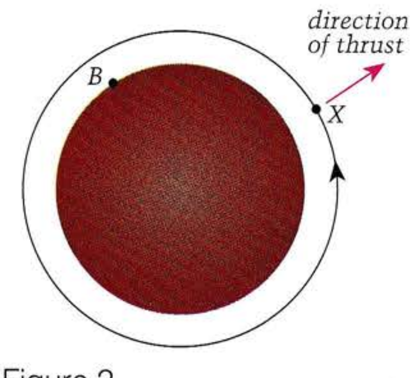


Figure 2

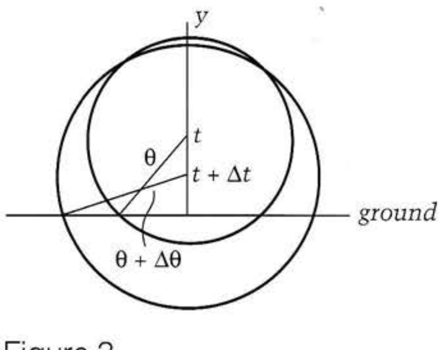


Figure 3

within $\Delta\theta$ have hit the ground.

The motion of the fragments is described by the following simultaneous equations:

$$y(t) = h - \frac{1}{2}gt^2 + v_0 t,$$
$$r(t) = v_1 t,$$

where v_0 is the initial velocity of the center of mass and v_1 is the speed of the expanding sphere.

Since

$$r\cos\theta = y,$$
$$\cos\theta = \frac{h + v_0 t - \frac{1}{2}gt^2}{v_1 t}.$$

the fragments within $\Delta\theta$ can be found by comparing the area of this surface fragment with the total surface area of the sphere. The area of the surface fragment is the product of the circumference of the surface and the arc length of that fragment $(r\Delta\theta)$, assuming that Δt is very small.

The number of fragments is therefore

$$\Delta n = \frac{(2\pi r\sin\theta)(r\Delta\theta)}{4\pi r^2}N$$
$$= \frac{1}{2}N\sin\theta\Delta\theta,$$

where N is the total number of fragments.

The frequency can now be found:

$$f = \frac{dn}{dt} = \lim_{\Delta t\to 0}\frac{\Delta n}{\Delta t}$$
$$= \lim_{\Delta t\to 0}\frac{N}{2}\sin\theta\frac{\Delta\theta}{\Delta t}$$
$$= \frac{1}{2}N\sin\theta\frac{d\theta}{dt}$$
$$= -\frac{1}{2}N\frac{d\cos\theta}{dt}$$
$$= \frac{1}{2}N\left(\frac{g}{2v_1} + \frac{h}{v_1}\frac{1}{t^2}\right).$$

The graph of this equation is shown in figure 4. In the figure, t_1 is the time at which the first fragment hits the ground and is the time where the frequency of particles hitting the ground is the greatest.

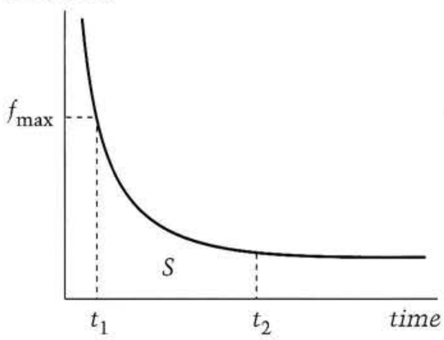


Figure 4

As a check on our work, the area S beneath the curve must equal the total number of particles N:

$$\int_{t_1}^{t_2} f\,dt = \int dn = \int_0^{\pi}\frac{N}{2}\sin\theta d\theta = N.$$

Does this make physical sense? Imagine a limiting case where the velocity of the fragments is very, very large. We can see that the particles that were shot downward will have a small difference in their vertical velocities and will all hit the ground at almost the same time. The particles that were shot straight up will also have a small difference in vertical velocities. This small difference will lead to a large lag time due to the long time in flight.

We also asked the more provocative question about whether it makes a difference if we learn to solve projectile problems using sports and rescue planes or morter shells and bombs. We now wonder why most of our readers did not bother to express their opinion on this pedagogical/social question. ◘

"WHISTLING IN SPACE" CONTINUED FROM PAGE 25

very strongly, and whistling atmospherics play the role of natural "magnetometers."

At very low frequencies, so-called "ionic whistlers" can be observed. Their spectrogram contains information about the masses of ions, which makes it possible to do a chemical analysis of cosmic gas at distances of tens of thousands of kilometers from Earth! Not only that, the strict formulas for $n_1(\omega)$ include the dependence on electron temperature, so whistlers can also be used as remote "thermometers"...

Isn't it wonderful that, with the help of physics, we can penetrate far into outer space without ever leaving the planet! This is not to take anything away from those who build rockets and the sophisticated equipment they carry—they are monuments to human ingenuity. But it's significant that nature itself has provided us with the possibility of carrying out remote measurements of the cosmic phenomena. It has laid thousand-kilometer-long waveguides along the geomagnetic lines of force running from one hemisphere to the other in outer space. They have existed as long as the Earth has, but it wasn't until the advent of radio that it became possible to use such a wonderful "instrument." ◘

Challenges in physics and math

Math

M196

Not 1997. Find a positive decimal number (which may not be an integer) that will increase by a factor of 1996 if you exchange the digits in the first and fifth decimal places. (D. Averiyanov)

M197

Switch and square. Find all pairs of natural numbers x and y such that both $x^2 + 3y$ and $y^2 + 3x$ are squares of whole numbers. (I. Sharygin)

M198

Conservative estimate. Find a relation between a, b and c if $a = x + 1/x$, $b = y + 1/y$, $c = xy + 1/xy$ (where x, y are variables). The relation must not contain any radicals.

M199

Hands-on politics. In the parliament of Illyria, each member slapped the face of exactly three other members of this prominent body. Several parliamentary committees are to be organized so that each member of the parliament works in one (and only one) of them. To avoid conflicts within a committee, it's necessary to fill it with deputies who have never slapped one another in the face. Prove that if the number of committees is greater than or equal to 7, this condition can always be met, but that if there are fewer than seven committees, the task may sometimes prove impossible. (A. Belov)

M200

Circular reasoning. A trapezoid $ABCD$ (in which AD and BC are bases) is inscribed in a circle. Its diagonals intersect at point M. Let a straight line perpendicular to the bases of $ABCD$ meet BC at K and meet the circle at L (where L is the point of intersection for which M lies on line segment KL). Let $MK = a$ and $LM = b$. Express in terms of a and b the radius of the circle tangent to segments AM and BM, and also tangent internally to the circle circumscribed about $ABCD$. (I. Sharygin)

Physics

P196

Elbow in motion. Two rigid rods of length L_1 and L_2 are connected by an articulated joint at point A. Their free ends move away from each other uniformly with the corresponding velocities v_1 and v_2 directed along the same line (fig. 1). Find the acceleration of point A at the moment the rods make an angle of 90°. The rods are moving in the same plane. (B. Bukhovtsev)

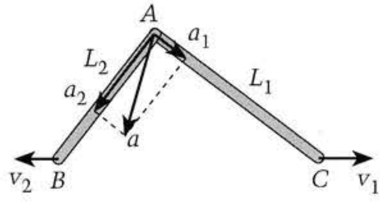


Figure 1

P197

Heating a planet. The planet "E" is very similar to Earth, but on this planet the anti-ecologists won the last planetary parliamentary elections. They built huge nuclear power stations all over the planet, including the seas and oceans. When they are in operation, 1,000 W of power is dissipated per square meter of the planet's surface. How soon after energy production begins will the atmospheric temperature rise by 1 K? Assume that the planet itself is not heated and that energy is dissipated into space at a constant rate. (S. Varlamov)

P198

Discharge in a gas. The dependence of the current I flowing in a gas discharge tube on the voltage V supplied between the tube's electrodes for the case of a non–self-maintained gaseous discharge is given in figure 2. A tube with a series load resistance $R = 3 \cdot 10^8$ Ω is connected to a source with a constant emf $\mathscr{E} = 6$ kV. Find the steady-state current flowing in the tube and

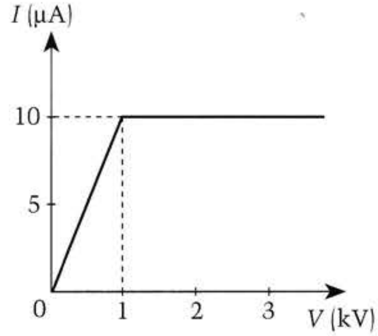


Figure 2

the voltage drop across the tube. Neglect the internal resistance of the emf source. (V. Mozhayev)

P199

Magnetized spiral. A plane helix with a large number of turns and an external radius R (fig. 3) is placed in a homogeneous magnetic field that is normal to the helix's plane and varies according to the formula $B = B_0 \cos \omega t$. Find the emf induced

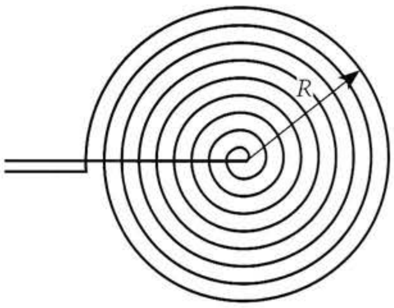


Figure 3

in the spiral. The distance between adjacent turns is constant. (I. Slobodetsky)

P200

Photojournalism. A page from a newspaper was photographed twice with a camera whose objective has a focal length of 50 mm. The first photo was made at the minimum distance (for this objective) $a = 0.5$ m. Before the second photo was taken, from the new minimum distance, a spacer ring of thickness $h = 25$ mm was attached to the camera. Find the ratio of the sizes of the images on the film for these two cases. (V. Deryabkin)

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

AT THE
BLACKBOARD I

A planetary air brake

Viscous drag and the slowing of the Earth

D. C. Agrawal and V. J. Menon

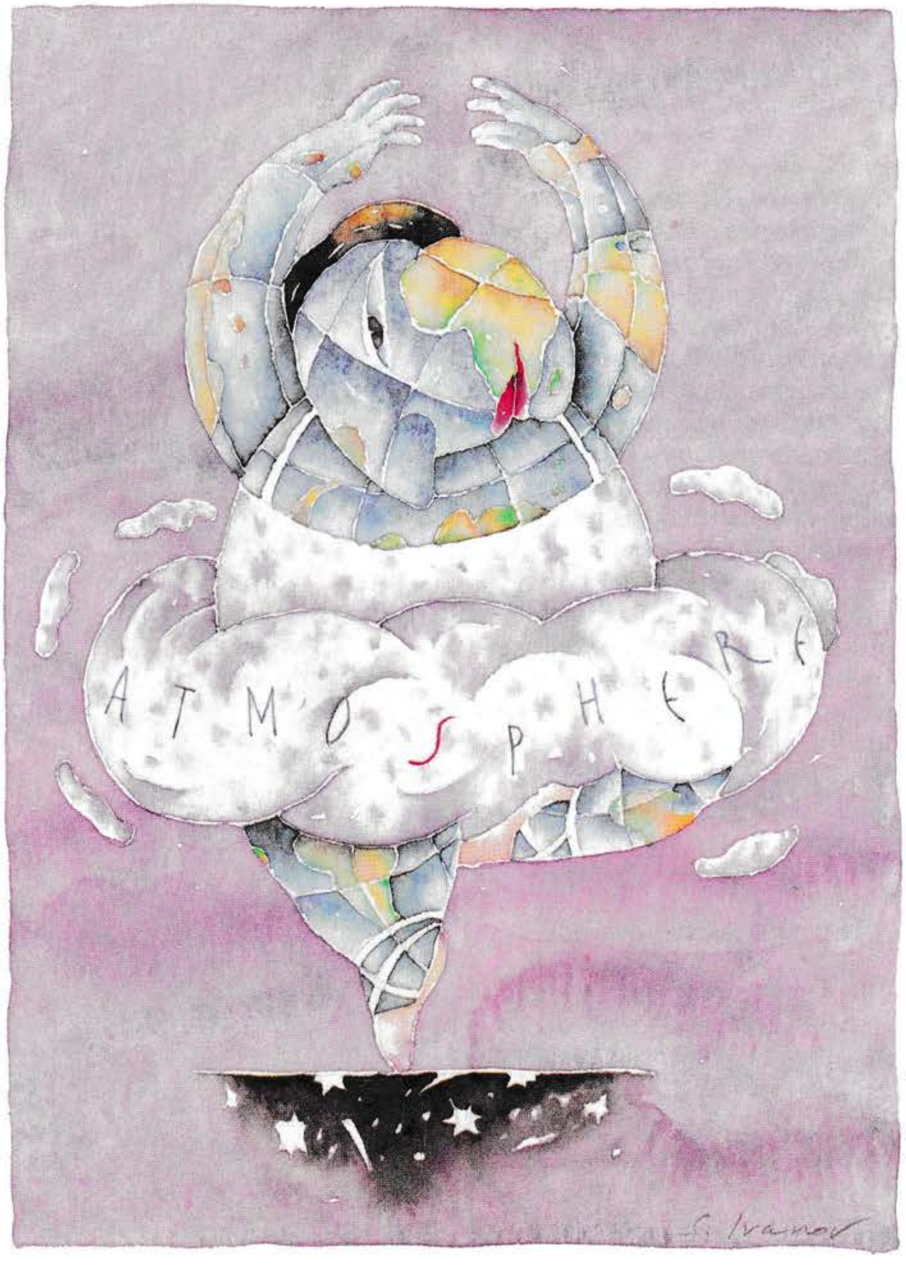
LET'S SEE IF WE CAN CALCULATE the slowing of the Earth's rotation due to the viscous drag of its atmosphere. Most standard textbooks quote Newton's formula for the viscous force as

$$F = -\eta A \frac{dv}{dx}, \qquad (1)$$

where η is the coefficient of viscosity, A is the area of the surface under consideration, v is the streamline velocity of the layer at a height x, and dv/dx is the velocity gradient at that height. Let's first consider a laboratory example before extending these ideas to the Earth's rotation.

A laboratory example

Consider a tub containing water of depth h. Let's move a rectangular piece of glass horizontally with velocity v while keeping its lower surface just in contact with the upper surface of the water. The no-slip condition between the solid and the liquid requires that the water in contact with the glass move with the same velocity v. Because of the viscosity of the water, successive layers at progressively increasing depths will acquire smaller and smaller velocities until the velocity becomes zero at the bottom. Assuming that the motion of the plate causes



Art by Sergey Ivanov

only streamline flow, the velocity gradient becomes $dv/dx = v/h$, where the vertical coordinate x has been measured from the bottom of the tub. Therefore, the *external* force required to keep the glass plate moving at this uniform velocity becomes, by virtue of equation (1),

$$F = \eta A \frac{v}{h},$$

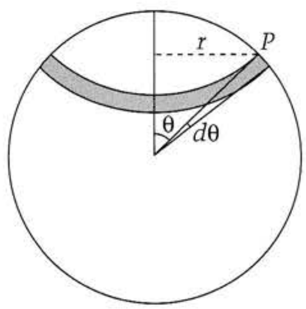
where A is the area of the plate.

Some interesting aspects of this equation are worth emphasizing. The force is inversely proportional to the depth of the water, which implies that a larger force is required when the water is shallow than when it is deep. This idea can be roughly generalized to the motion of a boat in a still pond provided we ignore the extra resistance offered by the water due to a sizable portion of the boat being submerged.

The Earth's rotation

It is well known that the Earth has a rotational kinetic energy due to its spin about its axis. This energy is about $2.2 \cdot 10^{29}$ J. The period of this rotation is experimentally known to decrease at a rate of 1 ms per century. This implies a continuous loss of the rotational kinetic energy of $1.1 \cdot 10^{12}$ W. The major reason for this loss is attributed to the tidal friction in the Earth–Moon system, but there should also be a contribution from the viscous drag on the Earth from the presence of its atmosphere.

Let's estimate the effect of the atmosphere using a simple model with minimal mathematical labor. Assuming that the Earth has a radius R and an angular velocity ω about the North–South axis, a point P at a latitude θ (see the figure) will have a linear velocity $v = \omega r = \omega R \sin\theta$. The point P is situated on a narrow circular strip of radius $r = R \sin\theta$, width $Rd\theta$, and area $dA = 2\pi R^2 \sin\theta\, d\theta$. The atmospheric layer in contact with this strip moves with velocity v, but the higher layers have gradually decreasing velocities, vanishing at



some height h above the ground. Since the velocity gradient is v/h, the viscous force on the strip under consideration is

$$dF = \eta\left(2\pi R^2 \sin\theta\, d\theta\right)\frac{v}{h}.$$

The power consumed by the viscous drag on the strip is then

$$dP = v\,dF = \eta\left(2\pi R^2 \sin\theta\, d\theta\right)\frac{v^2}{h}.$$

The net power lost by the Earth due to the viscous drag is given by the integral

$$P = \int_0^{\pi} dP = \frac{8\pi\eta\omega^2 R^4}{3h}.$$

Putting in the average numerical values

$$R = 6.37 \cdot 10^6 \text{ m},$$
$$\eta = 18.1 \cdot 10^{-5} \text{ poise},$$
$$\omega = 7.29 \cdot 10^{-5} \text{ s}^{-1},$$
$$h = 10^5 \text{ m},$$

we obtain

$$P = 1.3 \cdot 10^{10} \text{ W},$$

which is only about 1% of the power lost due to the tidal friction. Of course, the value we obtained is to be regarded only as a representative number illustrating the direct application of Newton's formula. There are further complications, such as the change of η with altitude, uncertainty in selecting h, and the presence of water vapor and dust in the atmosphere. ◘

D. C. Agrawal *is a reader in the Department of Farm Engineering at Banaras Hindu University in Varanasi, India.* **V. J. Menon** *is a UGC Research Scientist in the Department of Physics at the same institution.*

The long road to longitude

How we finally became "coordinated"

by A. A. Mikhailov

THE POSITION OF A POINT on the Earth's surface is determined by two coordinates: the geographic latitude and longitude. Do you know where these concepts came from? In the second century A.D. the Greek astronomer and geographer Claudius Ptolemaeus (Ptolemy) introduced the concepts of length and width to measure the extent of the countries bordering on the Mediterranean Sea, which stretched from east to west. Measurements along the length of the Mediterranean he called "length," and those in the transverse direction he called "width." Later these concepts, which could be applied to any object, were modified to describe the position of an object on the Earth's surface. So special terms were coined: latitude and longitude. Although these notions, clearly illustrated on a globe, are known to all, many are unaware of their strict definitions.

It's often said that the latitude ϕ is the angular distance of a particular point from the Earth's equator, and that the longitude λ is the dihedral angle between the plane of the given meridian and the one conventionally designated as the zero (or prime) meridian. Thus the network of meridians and the parallels of latitude drawn on the globe's surface make it possible to indicate the geographic coordinates of any place on the Earth. Naturally a knowledge of coordinates is necessary. But how do we determine our own geographic coordinates? We do it by performing some astronomical observations. Let's assume we've calculated the latitude and longitude of our location. Can we be sure that the point on the globe with these coordinate is the actual place where we are on Earth? The answer is an emphatic "no"—even if we know our coordinates with the utmost precision.

This is because the Earth is not a true sphere, while the globe—a simplified and greatly reduced model of the Earth—is. It is the irregular (nonspherical) shape of the Earth that causes the deviation between the geographic coordinates obtained by astronomical observations and those given on a map or globe for a particular location.

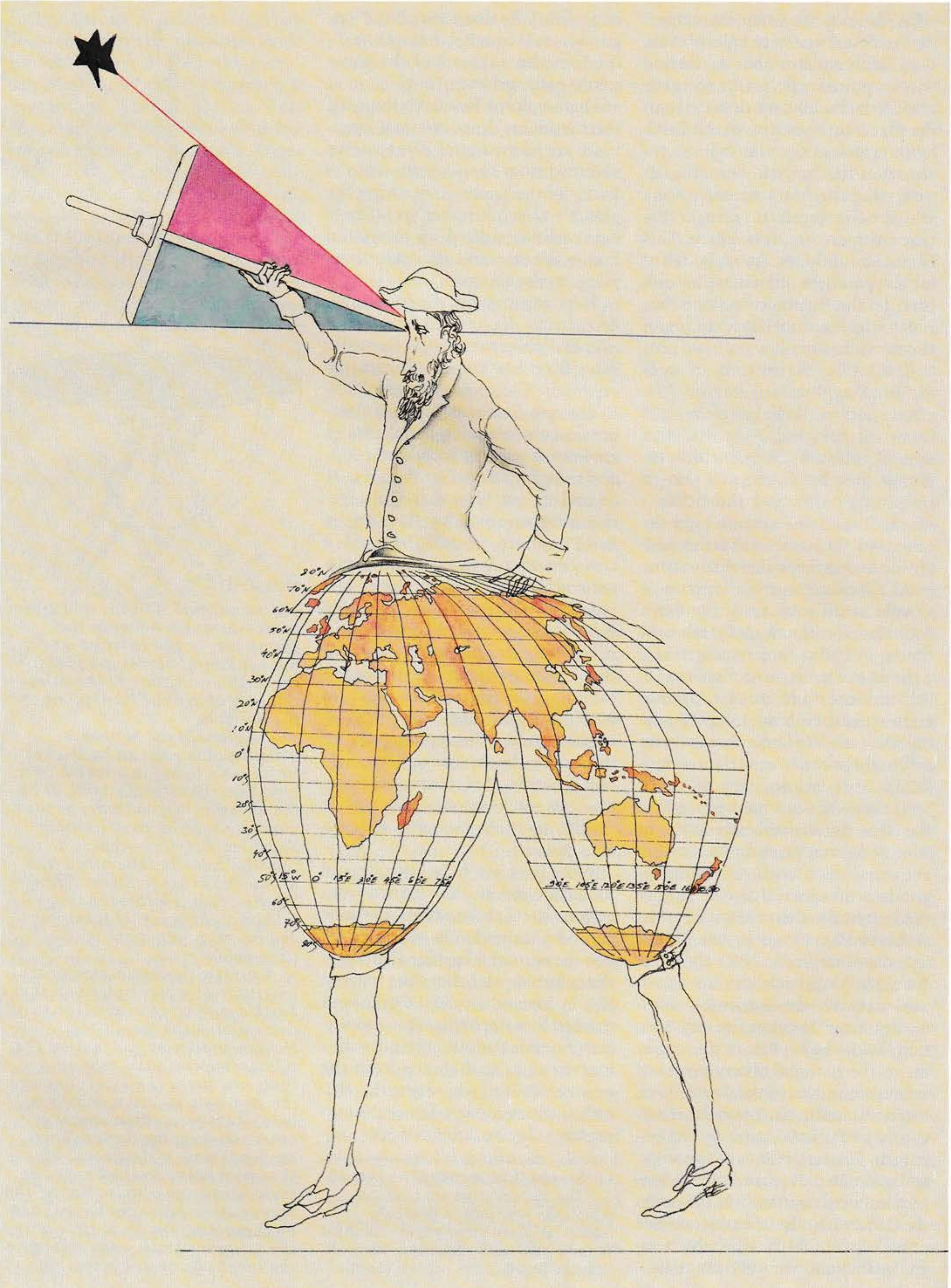
If the Earth were strictly a sphere, any plumb line dropped from the Earth's surface would pass through the Earth's center, and the equator and meridians would be circles of the same radius, equal to that of the planet. In that case, the geographic latitude could be measured as the meridianal arc drawn from the equatorial plane to the given site, and the geographic longitude as the equatorial arc stretching from the conventional prime meridian to that of the given point.

The true shape of the Earth (*geoid*) is rather complicated, although it is close to that of an ellipsoid of revolution, flattened in the polar regions. This shape results from the heterogeneous distribution of mass inside the planet and on its surface—in particular, from the existence of continents with high mountains and oceans with deep hollows in their troughs.

Due to the ellipsoidal shape of the Earth, a plumb line doesn't necessarily pass through the center of the planet, and the irregularities of the geoid shape results in additional lateral deviations of the plumb line, so it might not even intersect the Earth's axis. These factors lead to the so-called plumb line deviations—that is, the "anomalous" (from the spherical point of view) variations in its direction for different locations on the Earth's surface. Taking into account that the practical method of determining latitude and longitude uses the plumb line as a basic direction to orient the astronomical tools, the discrepancy between the true geographic coordinates of a point on the Earth's surface and that shown on a globe becomes quite understandable.

Now it's time to strictly formulate the definitions of latitude and longitude. Latitude is the angle between the local plumb line and the equatorial plane. Notice I didn't say

Art by Yury Vashchenko


80°N
70°N
60°N
50°N
40°N
30°N
20°N
10°N
0°
10°S
20°S
30°S
40°S
50°S
60°S
70°S
80°S
15°W 0° 15°E 30°E 45°E 60°E 75°E
90°E 105°E 120°E 135°E 150°E

"the plane of the Earth's equator." Due to deviations of the plumb line, the Earth's equator (that is, the line whose points all have zero geographic latitude) does deviate from the plane cross section of the Earth by the value of the plumb line's inclination, reaching 10″ or more (particularly in mountainous regions), which corresponds to a linear distance of approximately 300 m. This is true not only for the equator but for the parallels of latitude as well (that is, the lines of the same latitude), which are not plane curves for the same reasons.

The strict definition of geographic longitude is as follows: longitude is the dihedral angle formed by two planes parallel to the Earth's axis of rotation, one of which includes the plumb line at a chosen reference point and the other a plumb line in the location to be determined. Rather than measure longitude in degrees, we could use time units. The necessary calculation is simple: a full rotation of the Earth about its axis (that is, 360°) takes 24 hours, so 1 hour corresponds to 15°; 1 minute of time corresponds to 15′ ("15 minutes") of arc; 1 second of time corresponds to 15″ ("15 seconds") of arc. We know that noon (or midnight) occurs at different moments at locations with different longitudes. It's not hard to see that the time difference between noontime at a given place and noontime at the reference location (whose longitude is taken to be zero) is simply the longitude of that place.

According to an international agreement ratified in 1884, the reference (zero) longitude corresponds to the location of the Greenwich observatory near London, which was founded in 1675. Before the agreement, the national observatories of various countries served as reference locations: in Russia, it was the observatory at Pulkovo, near St. Petersburg; in France, the Paris observatory; and so on. In previous centuries, longitudes were counted from Ferro in the Canary Islands—the westernmost point of the Old World (which included the continents of Europe, Asia, and Africa). At that time all the longitudes were counted in the same direction (the values had the same sign)—eastward from Ferro.

Due to plumb line deviations, all the meridians (lines of equal longitude) are likewise not plane curves obtained from a cross section of the Earth by the planes containing the planet's axis, but rather are slightly curved lines shifted up to several hundred meters to either side of the plane cross sections.

How accurate are the geographic coordinates determined from astronomical observations? It basically depends on how these measurements are made. The most precisely determined coordinates are those of astronomical observatories: their latitudes are known with an accuracy of 0″.1, and their longitudes to within 0.01 second of time. Note that 1″ of a meridian arc corresponds to 31 m on the Earth's surface, while 0″.1 second of time corresponds to 46 m at the equator (and about half that in the middle latitudes). Thus the aforementioned values correspond approximately to 3 m on the Earth's surface. Now we see that it is not enough to say that the Paris observatory is located at 48°50′11″ North latitude, $0^h9^m20^s.93$ East longitude. We need to indicate the spot on the observatory's grounds that has these coordinates! (In this case, it's the location of a device for measuring the coordinates of heavenly bodies.)

Now let's see how latitude and longitude can be determined from astronomical observations. Latitude can be obtained very simply: you just measure the angular altitude of the celestial pole[1] above the horizon (fig. 1). Since the celestial pole isn't marked by a star or planet in the sky, astronomers use either a star with a known angular distance from the pole or the Sun, whose angular altitude is given for any day of the current year in the Ephemerides.[2]

[1]See the "Commentary" on page 46.

[2]The Ephemerides are tables in which the positions of heavenly bodies are given for any given moment of time (sometimes for many years into the future).

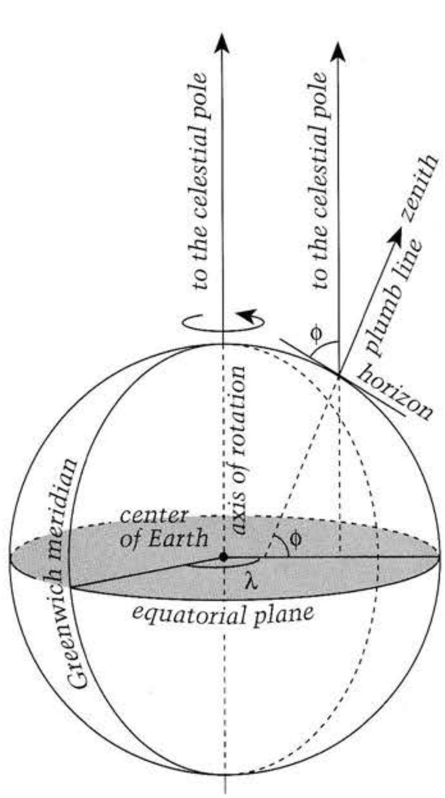


Figure 1

The geographic latitude of a given point (the angle between the plumb line at the location and at the equatorial plane) is equal to the angular altitude of the celestial pole above the local plane of the horizon.

The plane of the horizon is determined in practice by the free surface of a liquid in a vessel, by a bubble in the tube of a level, or by the direction perpendicular to the plumb line at the given location. The other direction—that is, to the celestial pole—is determined from astronomical observations. The celestial pole is located in the middle of the points of the upper and lower culminations of any circumpolar star.

A precise determination of the longitude at a given place on the Earth's surface is equivalent to determining the local time. Having chosen by convention the reference location of zero longitude, one notes the local time at the reference location, which is being kept by a chronometer or other timekeeping device. Then the local time is determined for the given point on the Earth's surface from an observation of the Sun or other heavenly body. Then, by comparing the local time with the reference time, one can obtain the longitude of this place.

The altitude measurement is performed when the heavenly body passes through the celestial meridian (that is, when this altitude is maximum). Special devices were invented for this purpose that can measure angles with a high degree of accuracy. The first primitive angular device, called a Jacob's staff, was used in the Middle Ages, and in the 1730s the sextant was invented in England (fig. 2), which is still used for measurements at sea or in the air. As the visual sightings are being made, the sextant is held in the hand to lessen the effects of pitching and rolling. Usually the image of the Sun (or a star at night) is lined up with the horizon in the eyepiece. With some experience one can measure the angular altitude above the horizon with an accuracy of one minute of arc or less. When measuring the angles on land, where it's possible to fix a device firmly on a tripod, one uses more precise instruments—for example, a goniometer. It can measure angular distance with an accuracy of 1″ or better.

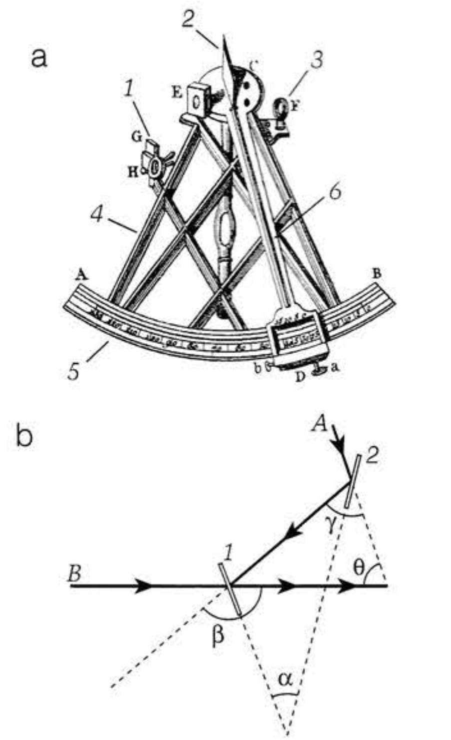


Figure 2

(a) The images of two objects whose angular distance is to be measured are superimposed in the sight (3) by means of two mirrors (1, 2). Mirror 1, attached to a stationary frame (4), is covered with silver only to half its height; the other half is transparent. The frame ends with a limb (5), which is a circular arc of 60° (this explains the name of the device). Mirror 2 is attached to the movable part of the frame—the so-called alidade (6), which can turn about the axis passing through the limb's center perpendicular to its plane. (b) This figure shows how the limb helps one find the angle α formed by the two mirrors. This angle is related to the angular distance θ between the observed objects by the formula $\theta = \beta - \gamma = 2(\beta/2 - \gamma/2) = 2\alpha$.

In principle, it's very easy to determine longitude: the difference between the longitudes of two points is equal to the difference in the local times at the same instant. However, before the invention of radio it was very difficult in practice to determine longitude. Indeed, with no means of direct communication, how could one know the precise time at a given moment in Greenwich (the reference location for counting longitudes) if one was hundreds or thousands of kilometers away? In the old days, one approach was to use an astronomical phenomenon visible simultaneously from both of these places and which occurred at a certain Greenwich time (known beforehand)—for example, eclipses of the Moon or Jupiter's moons. (The possibility using Jovian eclipses to find longitudes was suggested by Galileo, who discovered Jupiter's moons in 1610.)

Unfortunately, these events do not occur instantaneously, but can last for a number of minutes, so they can be recorded only with a corresponding error. One minute of time corresponds to 28 km at the equator, so this method of longitude determination has an intrinsic error of hundreds of kilometers. There were other drawbacks: it was almost impossible to observe the eclipses of Jupiter's satellites from the deck of a rolling ship; days would pass before the next eclipse occurred; and the planet couldn't be seen for several months in the year. Lunar eclipses are far rarer events, occurring two times a year at most. Travelers would have to wait days or even weeks before they could make several observations (and one needed to make several to obtain control values and improve the accuracy of the measurements). Clearly it wasn't possible to use such a method of longitude measurement for navigation at sea.

In the 16th century another way of determining the local time on the standard meridian, the so-called "lunar displacement" method, was found, although its practical application wasn't possible until the invention of the sextant. The Moon makes its way around the Earth like the hand of a clock moving on the stellar "dial." However, this hand rotates very slowly—at 1/55 the rate of the hour hand of an ordinary clock. Now, obviously one can determine time by looking only at the hour hand, but due to its slow movement such "measurements" will be very approximate. Since the lunar "hand" moves even more slowly, one can't expect to obtain precise data from the lunar clock.

But the situation isn't as bad as it seems, because the "face" of this clock was created with great accuracy: the positions of the stars are known, so the "divisions" of this face are very fine and precise. One need only determine the position of the Moon relative to the stars very accurately by means of precise goniometric (angle-measuring) devices. During the course of an hour the Moon shifts relative to the stars by a distance equal to its diameter, which is visible from the Earth at an angle of about 30′. Fixing the Moon's position relative to the stars with an accuracy of 1′ results in an accuracy in time measurement of two minutes.

To apply this method, one must know the minute details of the Moon's motion. Only then can the apparent lunar position provide the local time at the standard meridian and thus the difference between the local time at the observer's location and at Greenwich (that is, the longitude). This

Commentary

IF YOU OBSERVE THE NIGHT SKY over the course of several hours, you can see that the constellations change their position relative to the horizon. For example, in the eastern part of the celestial sphere, the stars ascend above the horizon and move to the right. In the northern part of the sky, most stars trace their daily concentric circles. We may say that the entire firmament (the celestial sphere) revolves about some line called the celestial axis. The points where the celestial axis intersects the celestial sphere are called the celestial poles (North and South). The North pole is the one where an observer on the outer side of the firmament watches the celestial sphere rotate in the clockwise direction. Of course, this rotation is only an illusion—in reality, it's the Earth that rotates about its axis in the counterclockwise direction. It's clear that the celestial axis is parallel to the planet's axis.

In addition to the celestial axis and poles, there are a number of characteristic points and lines on the celestial sphere (fig. A). The plane *SWNE* is the plane of the horizon—that is, the plane tangent to the Earth's sphere at the point *C* where an observer is located. The line *NS* is the midday line (at midday vertical objects cast shadows along this line). A plumb line drawn through point *C* intersects the celestial sphere at the zenith point *Z*. The plane that passes through the points *S*, *Z*, *P* (the celestial North Pole) and *N* is the celestial meridian plane. The celestial equator is the line where the celestial sphere intersects the plane that is perpendicular to the celestial axis and passes through the observer (point *C*).

Each star crosses the celestial meridian twice a day (these "culminations" are their corresponding positions). When a star is at its upper culmination, it assumes the highest position above the horizon. Correspondingly, the lower culmination is the lowest position of a star relative to the horizon (fig. B).

Both the lower and upper culminations can be observed for stars located near a celestial pole. Rising and setting stars, on the other hand, have only an upper culmination (the lower one occurs below the horizon). The moment of the Sun's upper culmination is called true midday, and its lower culmination is true midnight.

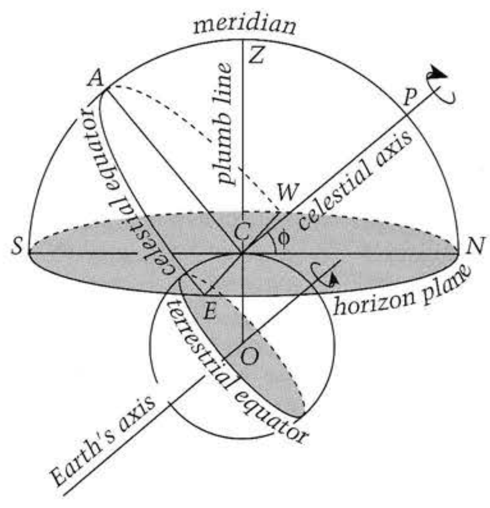


Figure A

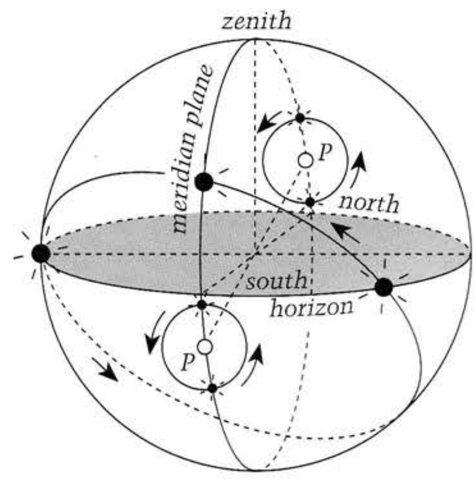


Figure B

method imposed heavy demands on celestial mechanics, which developed the theory of the rather complicated motion of the Moon. The accuracy of the lunar displacement method wasn't markedly better than that of the previous one, but it could be applied at any time when the Moon was visible, so it's clear why it was very popular among travelers, especially sailors.

However, the lunar displacement method didn't satisfy the old salts entirely, because of its complicated calculations and poor accuracy. Long voyages of sailing vessels depended greatly on the force and direction of the wind, and very often the ships were out of sight of land for long periods. So it was extremely important to know the ship's location (that is, the latitude and longitude). In 1713 the British government, taking a keen interest in secure navigation (at that time Great Britain had the largest and most developed fleet of ships in the world), offered a huge prize of £20,000 for the invention of a reliable method to determine longitude with an accuracy of 1/2 degree. A portion of this prize was awarded (posthumously, alas) to the German astronomer Tobias Mayer for his tables of lunar motion, which made it possible to improve precision when measuring longitude by the lunar displacement method. Half of the prize was given to the English clock maker and inventor John Harrison, who in 1735 had constructed the marine chronometer.[3]

The marine chronometer is a spring-driven clock much like a modern pocket watch. The chronometer had three hands—hour, minute, and second. The second hand would jump forward every half-second with an audible click that could be heard several meters away. What made the chronometer so convenient was that it could be transported at sea without affecting its accuracy. It was mounted in a box with a gimbaled suspension along two perpendicular axes, so that the clock maintained its level position even if the box was tipped.

Soon this method of transporting chronometers became widespread for work on land. Scientists undertook expeditions to determine the longitudes of populated areas for map-making purposes and to refine the longitudes of the various national observatories. For example, the first director of the Pulkovo observatory, V. Y. Struve, organized a special expedition in 1843, equipped with 60 chronometers, to determine the difference in longitude between Pulkovo and Greenwich.

The problem of longitude measurement was solved in a quite unexpected way when the telegraph was invented in the middle of the last century. Now it became possible to transmit Greenwich mean time (GMT) and thus determine the difference between local time and GMT with the previously unachievable precision of fractions of a second. Of course, the given location would need a telegraph link for this to be possible, which made the method useless for ships at sea. Nevertheless, telegraph cables laid on the ocean floor connected the continents and made it possible to determine the longitude of Washington's observatory very accurately. In Russia, the standard time was that of St. Petersburg (strictly speaking, Pulkovo mean time). It was transmitted by wire from the Pulkovo observatory to the central telegraph office in St. Petersburg, and from there to telegraph and railway stations all over the country. At noon Pulkovo time, a gun fired from the Peter and Paul Fortress.

As you may guess, the invention of the wireless telegraph (that is, radio—see the May 1990 issue of *Quantum*, p. 39) provided the final and universal solution to the old longitude problem. As early as 1921 the radio station Novaya Gollandia in St. Petersburg began transmitting (several times a day) rhythmic time signals consisting of 61 ticks per minute, resulting in a time interval between ticks of 1/60 s. By timing the ticks of one's chronometer with the ticks transmitted by radio, one could tune one's timepiece to Greenwich mean time with an accuracy of several hundredths of a second.

The procedure wasn't perfect, though. The faulty operation of an observatory's clock and other technical difficulties in broadcasting the radio signals might cause an error of a fraction of second. However, this error could be corrected later, after the signals were received at the observatory and the times of current astronomical events were recorded. Special astronomical bulletins were published monthly for this purpose, giving the precise moments of transmission of past signals and thus providing a way to correct a chronometer later. Of course, such high accuracy was necessary only for determining longitude with the utmost precision. For practical navigation, where an accuracy of 1 s is more than adequate, such published corrections had no value at all. Nowadays, due to the invention of quartz clocks and then molecular and atomic clocks,[4] which keep time for many months with an accuracy of several thousandths of a second, corrections of the transmitted radio signals have been practically reduced to zero. At long last the age-old problem of finding the geographic longitude was solved with the utmost precision.

I'll conclude our account with an example. Imagine a person who knows the astronomical tricks for finding one's geographic coordinates. Send that person to an arbitrary place—a desert island, the mountains, an uncharted wilderness—anywhere. And to complicate the problem, let this person not know which country, or even which hemisphere, he or she is in. Let that person lose track of time, becoming unaware of what day or month it is. However, give that person a copy of the astronomical annual with logarithmic tables, a theodolite or goniometer, a chronometer showing an arbitrary time, and a radio receiver (for hearing the time ticks only). All that person will need is to observe the stars on a clear night and the Sun the next morning in order to determine (after the necessary calculations) the month, day, latitude, and longitude. Looking at a map, the person could point out her or his location and the route to the nearest town or any other place. In other words, an educated person can never be lost on planet Earth. ◘

[3]For a fascinating account of John Harrison's achievement, see *Longitude* by Dava Sobel (New York: Walker and Company, 1995).—*Ed.*

[4]The role of the pendulum in a quartz clock is played by a slice of quartz cut in a certain way. Electromagnetic oscillations are excited in this crystal, the frequency of which do not vary for a long period of time. The operation of molecular and atomic clocks is based on the natural oscillations of certain molecules and atoms.

AT THE BLACKBOARD II

Three physicists and one log

Then, we were young and strong. Now, we know physics . . .

by Roman Vinokur

ONCE UPON A TIME THERE were three physicists carrying logs to build a house (fig. 1). Their shoulders ached, but nobody complained. However, physicists are physicists, even if they're doing a backbreaking job. One of them raised an interesting question: which position under the log—front, middle, or rear—bears the brunt? They stopped to draw the statics diagram (fig. 2) on the ground. The log was considered a straight, uniform beam. Three supports (the shoulders of the carriers) were assumed to be equally spaced (a distance L apart), and the middle support was exactly under the beam's center of gravity. The log was assumed to be horizontal, so that the forces F_1, F_2, and F_3 at the supports are vertical.

The fundamental equations for static equilibrium could be written

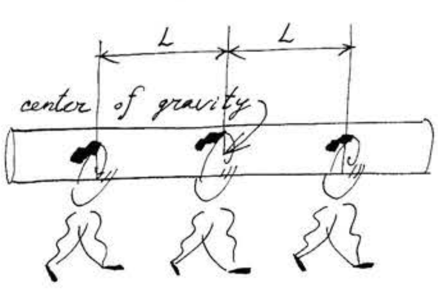


Figure 1

Three physicists carrying a log. They look similar, but they aren't (see below).

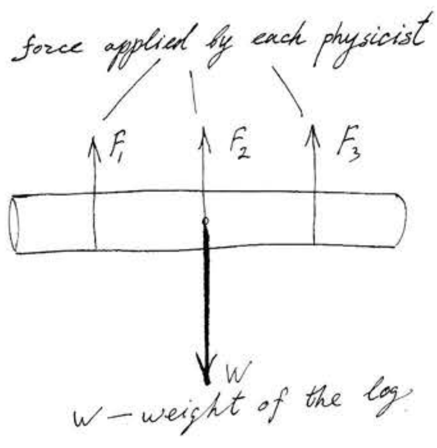


Figure 2

Statics diagram representing a statically indeterminate problem: three unknowns and only two equations.

in the form

$$F_1L = F_3L$$

(because the torques about the center of the log are equal), or

$$F_1 = F_3 \qquad (1)$$

and

$$F_1 + F_2 + F_3 = W \qquad (2)$$

(the weight W of the beam equals the resultant of the forces F_1, F_2, and F_3). The physicists arrived at two linear equations (1) and (2) with three unknowns. The number of possible solutions in such a case is infinite. For example, $F_1 = F_3 = 0$, $F_2 = W$ (that is, one person works hard, the others rest), or $F_1 = F_3 = W/2$, $F_2 = 0$ (where the previous situation is reversed). Of course, the "fair" solution $F_1 = F_2 = F_3 = W/3$ also holds true.

Later the physicists learned, after looking in some textbooks, that such a problem is called *statically indeterminate*. I should also mention that the physicists were quite young. In fact, they were just first-year students of physics, and the author was one of them.

"Well," said Jeff, "I know what to do. Let's consider the log as being cut in half. Now we have two similar logs. Let Roman and Peter support the outer ends, and I'll stay in the middle, carrying the inner ends" (fig. 3).

"The problem has only one solu-

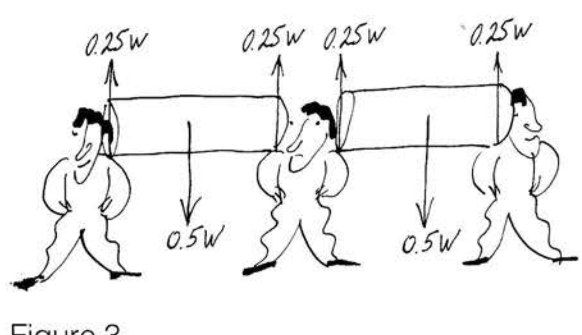


Figure 3

Jeff's model: a log consisting of two similar pieces.

Art by Sergey Ivanov

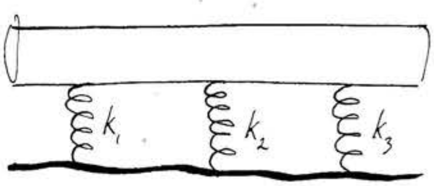


Figure 4

Roman's model: springs are put in play instead of people.

tion," Jeff continued. "Each of you two creates a force of $W/4$, and I have to withstand a load of $2(W/4) = W/2$, because I carry two ends simultaneously. Am I right? Certainly! So the only solution is $F_1 = F_3 = W/4$, $F_2 = W/2$. The guy in the middle has to work harder. Oh, well—I guess the 'middle way' isn't always the best!"

"Give me a break!" Peter objected. "I've been both in the middle and in the rear, and I felt no difference. I think your model may be flawed, Jeff! Maybe it's not such a good idea to 'split' the log. What do you think, Roman?"

"I'm not sure," I said. "But I think we need to derive some additional equations that allow for the deformation of the log and the carriers. It won't be easy . . ."

Here's what we managed to come up with.

Since wood is stiffer than a human body, we assumed that the log is absolutely rigid (it undergoes no deformation at all). On the other hand, the carriers were modeled as simple springs with spring constants k_1, k_2, and k_3 (fig. 4). From the viewpoint of mathematical physics, the model strikes me as pretty true to life: the stronger the carrier, the bigger his spring constant.

If all the springs were initially the same height, their deformations x_1, x_2, x_3 after loading are linearly related:

$$x_2 = \frac{x_1 + x_3}{2} \qquad (3)$$

(see figure 5). Such a simple relationship holds because the log is considered straight and rigid, and

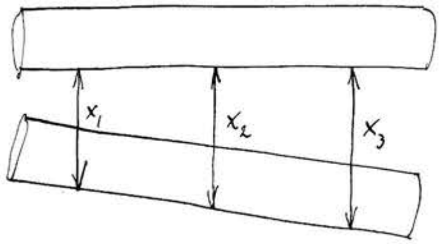


Figure 5

Linear relationship of spring deformation.

the middle support is the same distance from the front and rear supports. The deformations x_1, x_2, and x_3 were assumed to be small to prevent the log from tilting significantly. In this case, the horizontal components of the forces are relatively insignificant and can be neglected without grave consequences.

Using the well-known spring equation $F_i = k_i x_i$ (where $i = 1, 2, 3$) and equation (3), we got

$$\frac{F_2}{k_2} = \frac{F_1/k_1 + F_3/k_3}{2}. \qquad (4)$$

After solving the system of three linear equations (1), (2), and (4), we arrived at the final solution:

$$F_1 = F_3 = \frac{2k_1k_3}{4k_1k_3 + k_2(k_1 + k_3)} W,$$
$$F_2 = \frac{k_2(k_1 + k_3)}{4k_1k_3 + k_2(k_1 + k_3)} W. \qquad (5)$$

The expressions obtained looked a bit cumbersome and so in need of further analysis. As a start, we briefly reviewed the simplest situations.

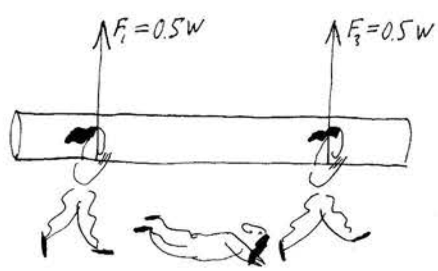


Figure 6

Loss of middle carrier.

If the springs are identical ($k_1 = k_2 = k_3$), our equations give $F_1 = F_2 = F_3 = W/3$. Thus the load is equally distributed among "uniform" carriers. But what if one of the carriers has fallen?

Let's consider the situation when the middle carrier falls out of the picture (fig. 6). Now $k_1 = k_3$, $k_2 = 0$. It follows from equations (5) that $F_2 = 0$ and $F_1 = F_3 = W/2$. (By the way, in the case of two carriers, the problem is statically determined and can be solved without using equation (4).) So the load on the front and rear carriers becomes 50% greater than in the three-carrier case.

If the front carrier falls (fig. 7), so that $k_2 = k_1$ and $k_3 = 0$, the result obtained from equations (5) seems paradoxical: $F_1 = F_3 = 0$, $F_2 = W$. Heaven help the person in the middle! All the weight is now on his shoulders. If the rear carrier tries to help by pushing up that end of the log, a torque is created, rotating the log clockwise. So the log would move out of the equilibrium position . . .

Naturally, a similar scenario unfolds if the rear carrier falls down.

At this point Peter and Jeff recalled that they were not "uniform." Jeff was twice as strong as each friend (he was a weightlifting champ).

"Okay!" Peter said. "Let one of the spring constants be twice that of the other two. You're a giant of a man, Jeff, and we both admire you greatly!"

We analyzed two cases: (1) with Jeff in the middle and (2) with Jeff in the rear. In the first case, $k_1 = k_3$ and $k_2 = 2k_1$. Applying these relationships in equations (5) we got

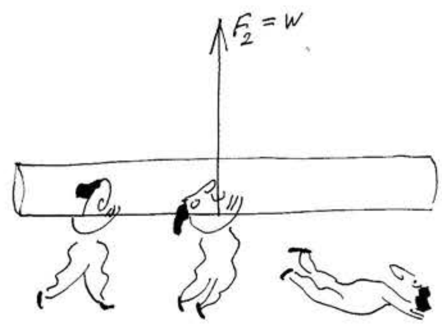


Figure 7

Loss of front carrier.

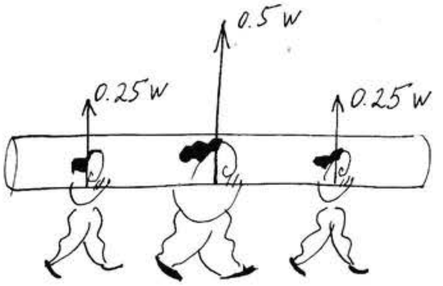


Figure 8
Jeff in the middle: he's twice as strong and bears twice the burden.

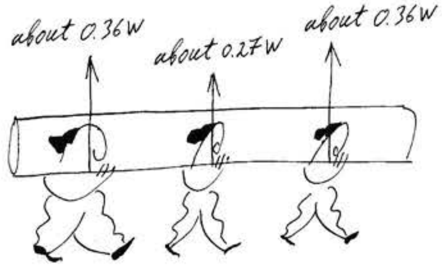


Figure 9
Peter in the middle now: he feels no difference (but Jeff and I did).

$F_1 = F_3 = W/4$, $F_2 = W/2$ (fig. 8).

"Hey, that looks familiar!" Jeff exclaimed, smiling. "My initial model proved to be incorrect, but I'm still in the middle, bearing the brunt!"

"Well, not only you," said Peter. "Look, in the second case (fig. 9) we have $k_2 = k_3$ and $k_1 = 2k_3$. As a result, $F_1 = F_3 = {}^4/_{11}W$, and $F_2 = {}^3/_{11}W$. So when I moved from the rear to the middle, I felt no significant difference, because ${}^3/_{11}W \cong 0.27W$ is only a little bigger than $0.25W$. Jeff experienced a more significant change: ${}^4/_{11}W \cong 0.36W$ rather than $0.5W$. So, we were both right, Jeff! That's why you felt a difference between the rear and middle positions and I didn't. Once again we see the importance of learning physics!"

"And the importance of having good friends," I said.

This happened long ago. Now we're able to solve any statically indeterminate problems, but as for carrying logs . . .

Roman Vinokur *is a physicist specializing in mechanics. A graduate of the Moscow Physico-Technical Institute, he works as a test engineer at Valeo Engine Cooling in Jamestown, New York. (This article is based on an actual occurrence.)*

MATH INVESTIGATIONS

The equalizer of a triangle

A clever line that does double duty

by George Berzsenyi

LET l BE A LINE IN THE PLANE of $\triangle ABC$. We will say that l is an *equalizer* of $\triangle ABC$ if it cuts the triangle into two parts whose areas and perimeters are equal. For instance, if in the figure shown below,

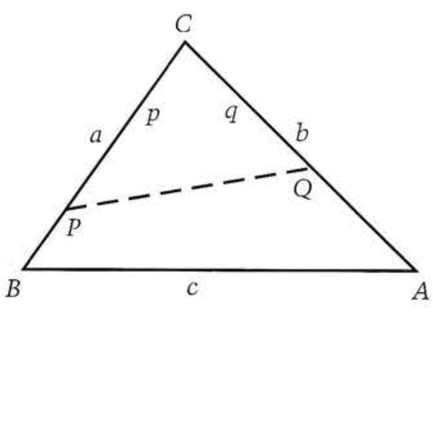
$$\text{area}(CPQ) = \text{area}(PQAB)$$

and

$$CP + PQ + QC = PQ + QA + AB + BP,$$

then we say that PQ is an equalizer of $\triangle ABC$. My first challenge to my readers is: **Prove that there is an equalizer for every triangle**. I recommend a carefully executed continuity argument to accomplish this task.

One can also observe that some triangles can have as many as three equalizers. However, on the basis of some computer experiments conducted by my friend and former colleague Professor Emeritus Herb Bailey, it seems that there are no triangles which have exactly two equalizers. Hence, my next challenge is: **Prove or disprove this claim**. Moreover, I hereby challenge my readers to **prove or disprove that no triangle can have more than three equalizers**. On



the basis of the figure above, it may seem that some triangles can have as many as six equalizers. For example, if $CP < CA$ and $CQ < CB$, then one can reverse P and Q (that is, reflect them in the bisector of $\angle BCA$), and the same may be done when P and Q are located on the other two pairs of sides. However, I strongly believe that three is the upper limit for the number of equalizers.

To analyze the situation via further computer experiments, let a, b, c be the lengths of sides BC, CA, AB, respectively, and assume that $a \le b \le c = 1$. Then there is a one-to-one correspondence between these "scaled" triangles and the points (a, b) of the plane that lie in the region bounded by the lines $x + y = 1$, $x = 1$, and $y = x$. Consequently, by introducing a fine mesh in the x–y-plane, one can plot those points whose corresponding triangles have one, three, or any other number of equalizers.

To simplify our further discussion, let $PC = p$ and $QC = q$. Then the requirements for PQ to be an equalizer can be expressed by the conditions

$$pq = \frac{1}{2}ab$$

and

$$p + q = \frac{1}{2}(a + b + c),$$

where the first condition comes from $\frac{1}{2}pq \sin C = \frac{1}{2}(\frac{1}{2}ab \sin C)$ and the second one from $c + (a - p) + PQ + (b - q) = p + q + PQ$. Therefore, one may recognize p and q as the two solutions of the quadratic equation

$$r^2 - \frac{1}{2}(a + b + c)r + \frac{1}{2}ab = 0.$$

Letting $s = (a + b + c)/2$, it follows that

$$p, q = \frac{1}{2}\left(s \pm \sqrt{s^2 - 2ab}\right).$$

Clearly, one must have $s^2 \ge 2ab$, as well as $p \le a$ and $q \le b$. Similar analysis applies to the other potential equalizers as well.

The present investigation was prompted by a nice problem posed in the 1988–89 Scottish Mathematical Challenge. For more information about this wonderful program the reader is referred to my "Problems, Puzzles, and Paradoxes" column in the Spring 1997 issue of *Consortium*.

In closing the present column I wish to point out that there is also a three-dimensional analogue of the equalizer. Hence my final challenge is: **Prove that for any tetrahedron there exists a plane which cuts the tetrahedron into two parts with equal volumes and surface areas**. Surely, many of the questions asked about the equalizers of the triangle have their appropriate analogues for the tetrahedron, and should keep my readers busy. Please send your findings to me c/o *Quantum*, 1840 Wilson Boulevard, Arlington VA 22201-3000. Perhaps they will generate further discussion in a future column.

Stirring up bubbles

Vapor cones and vortices in a boiling liquid

by T. Polyakova, V. Zablotsky, and O. Tsyganenko

THERE'S NOTHING EXOTIC about a boiling liquid. You encounter this phenomenon just about every day. It may seem that nothing new or surprising could occur when ordinary water boils. But actually this is a rather complicated and multifarious process, one that is still not fully understood.

This article deals with boiling water that is rotating. We'll begin with the description of a very simple experiment anyone can do at home. Heat up some water in a cylindrical vessel with a radius of about 10 cm and a height of 25–30 cm. Just when the water begins to boil, stir it vigorously to set it in rapid rotation. The surface of the water will take on the shape of a paraboloid of revolution, and the angular velocity will decrease gradually as time passes due to the friction between the water and the vessel wall. However, if we apply heat in a small area at the center of the vessel's bottom, we'll see a very strange picture. Boiling will take place only at the center of the bottom, and a large number of bubbles will quickly ascend along the axis of revolution. Then a column of vapor bursts through the surface of the water, and the characteristic noise and splashing of a boiling liquid occur. Immediately after this the water level near the cylinder's wall drops, and the velocity of rotation increases. Then the angular velocity decreases and, after a period of 1–2 s, the cycle is repeated. The time dependence of the angular velocity of rotation of a boiling liquid is indicated by curve A in figure 1. The other curve B shows the angular velocity of a nonboiling liquid.

If you tried to do this experiment yourself and failed, don't give up. Try it again, changing the intensity of the heat or altering the water level in the vessel. This phenomenon is very sensitive to the heating conditions. The experiment is easier to perform if the heat is supplied by a gas burner.

At least two questions arise when we compare curves A and B. First, why does the boiling liquid have an oscillating angular velocity in contrast to the nonboiling one? Second, what does the period of oscillation of the angular velocity depend on? To answer these questions and understand the nature of the observed phenomenon, let's first examine the main features of the boiling process in a liquid heated from below.

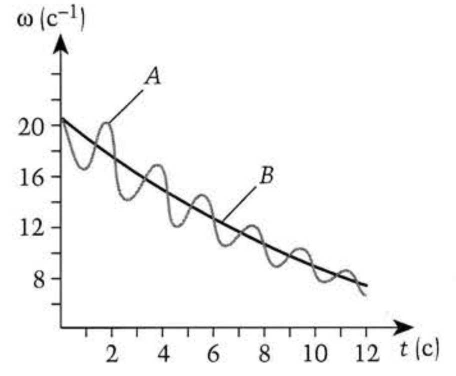


Figure 1

Boiling is the process of intense vapor formation, characterized by the continuous generation and growth of bubbles inside the liquid, which ascend to the surface due to the buoyant force. A very important parameter that affects how the boiling proceeds is the so-called *thermal head* $\Delta T = T_1 - T_s$, where T_1 is the heater's surface temperature and T_s is the liquid's boiling point. There are three types of boiling, which depend on the value of ΔT: nucleate, transitional, and film boiling. In addition, if all parts of the liquid have the same temperature T_0, which is equal to the boiling point, such boiling is called *saturated*. When $T_0 < T_s$, and the boiling takes place only near the heater, this is nonsaturated (or underheated) boiling.

Nucleate boiling is usually divided into four basic stages. For water being heated in a metal vessel, the first stage is observed at ΔT = 10–16 K, which is referred to as the region of separate bubbles. This stage is characterized by the existence of individual active centers of bubble generation. Bubbles of vapor sticking to the bottom of the vessel do not interact with the other

Art by Vera Khlebnikova

bubbles heading upward. Every active center is surrounded by its own "zone of influence" whose radius r is equal to the diameter of the bubbles when they break away from the bottom (d_0). Similarly, the influence zone of the rising bubbles is a sphere of diameter $2d_0$. The number of active centers on the vessel's bottom increases with ΔT, while the distance between the neighboring centers decreases. When the average distance between adjacent centers approaches (approximately) $2d_0$, the second stage of boiling begins.

In this stage, continuous chains of bubbles (rather than individual ones) are generated at some active centers—that is, the vapor columns arise from the interaction of individual bubbles. A further increase in ΔT causes bubbles to merge not only within the same column but in adjacent centers as well. Several columns produce a unified structure called the "vapor cone," which marks the beginning of the third stage of boiling. Now a large amount of vapor is rising from the heated surface. Usually the vapor cones are connected to this surface by a large number of vapor stems. When a vapor cone grows large enough, it breaks away from the bottom and rises.

The fourth stage of boiling begins at $\Delta T = 22$ K for water and lasts until *burn-out* occurs. At this stage the cone's stems merge to produce a vapor cloud that touches the heating surface. Thus at this stage of boiling there are some regions of film boiling.

Now let's return to the enigmatic oscillations of the angular velocity of boiling water. Let the liquid rotate as a whole (that is, like a solid body) with an angular velocity ω. It's easy to show that in this case the free surface is described by the following equation:

$$z - z_0 = \frac{\omega^2}{2g} r^2, \qquad (1)$$

where g is the acceleration due to gravity, z is the vertical coordinate, z_0 is the value of z along the axis,

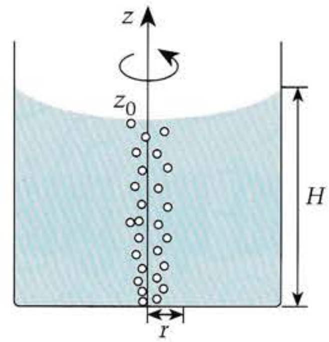


Figure 2

and r is the radius of the bubble column (fig. 2). Since the pressure of the liquid acting on the vessel's bottom is $P = gz$, equation (1) yields the dependence of the pressure at the bottom on the distance from the axis of rotation:

$$P = P_0 + \frac{\rho\omega^2 r^2}{2}, \qquad (2)$$

where P_0 is the pressure at the center of the bottom and ρ is the liquid's density. If a bubble of radius R is formed in the rotating liquid, it can grow only when the pressure inside the bubble (P_b) is larger than the sum of the external pressure (equation (2)) and the extra pressure under the curved surface of the liquid (the Laplace pressure). The latter is equal to $2\sigma/R$, where σ is the coefficient of surface tension. It follows from equation (2) that if the temperature of the bottom surface is the same throughout, the bubbles generated at the center of the bottom enjoy the best conditions for growth. In addition, the linear velocity of the rotating liquid increases at larger distances from the axis of revolution, which is also unfavorable to bubble growth. If any bubble appears at some distance from the center, the buoyant force $\rho\omega^2 rV$ (where V is the bubble's volume), which is directed toward the axis of rotation, it will move to the center.

We've come to the conclusion that in a rotating and boiling liquid all the bubbles should concentrate near the axis of rotation. Now let's examine in greater detail the condition leading to growth for a bubble located at this axis. To characterize the degree of vorticity of a fluid flow, a special physical parameter is introduced—the circulation Γ. In the vicinity of the bubbles under consideration, the circulation is not zero—it equals

$$\Gamma = 2\pi\omega R^2.$$

The presence of a nonzero circulation means that the bubble (strictly speaking, its "equator") is affected by the negative pressure from the liquid, which is related to the circulation by the formula

$$P_\Gamma = -\frac{\rho\omega^2 R^2}{2} = -\frac{\rho\Gamma^2}{8\pi^2 R^2}.$$

The condition for bubble growth is met when $P_b \geq P_1$, where

$$P_1 = P_0 + \frac{2\sigma}{R} - \frac{\rho\Gamma^2}{8\pi^2 R^2}. \qquad (3)$$

Taking different values for the circulation Γ as a parameter, we can derive the set of functions $P_1(R)$ shown in figure 3. The curves indicate that an increase in the circulation provides better conditions for bubble growth.

It's interesting that the pressure from the liquid caused by its rotation about a bubble acts counter to the Laplace pressure. For example, in rotating water the influence of the water on the boiling process is entirely counterbalanced at $\Gamma > 1.5$ cm^2/s. This means, in particular, that one can't superheat water that has vortices with this value for the circulation.

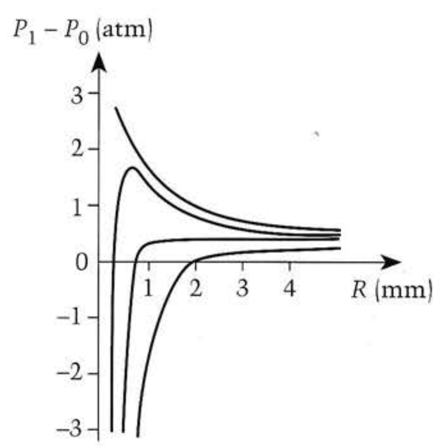


Figure 3

From our previous discussion and from figure 3, we can conclude that a vortex with an arbitrary value for the circulation can be a possible center for bubble generation. Such centers do appear in our experiment with rotating water. As the temperature of the water in the cylinder $T_0 < T_s$ (that is, we're dealing with underheated boiling), boiling centers can be formed only near the bottom, where the circulation is greatest. Since the boiling point is determined by the condition $P_b(T_s) = P_1$, then in accordance with equation (3) and figure 3 the boiling point is lower at sites where $\Gamma \neq 0$ than at sites where $\Gamma = 0$. This means that the value $(T_1 - T_s)$ can be much larger at the center of the bottom than elsewhere.

If the thermal head at the center of the bottom reaches $\Delta T \cong 16\text{–}20$ K, the second or third boiling stages begin, depending on the value of Γ. In our experiment the third stage most likely occurs. The vapor cones generated at the bottom rise along the axis of rotation and merge, producing rather large vapor cavities that ascend to the surface. When such a vapor column breaks at the center of the liquid's surface, the surrounding "cold" surface water $(T_0 < T_s)$ rushes into the hollow cavity. The water moves downward, swirling like water going down the drain in a bathtub. At this moment the angular velocity of the water's rotation increases, because the moment of inertia for the system decreases when water moves from the wall toward the axis of rotation.[1] Cold water entering from the surface is quickly heated inside the cylinder to a temperature that allows the vortex to again become the active center of boiling. Large vapor cavities form anew at the axis of rotation, which push the water toward the wall of the vessel, thereby increasing the moment of inertia for the system and decreasing the angular velocity of the water before the vapor column breaks through the surface.

Thus we can provide a very simple answer to the first question posed at the beginning of this article. The angular velocity of a nonboiling liquid decreases monotonically due to friction with the vessel's wall—just like figure skaters who don't change the position of their arms during a spin. On the other hand, the angular velocity of a boiling liquid does vary, as with spinning figure skaters who periodically spread their arms and draw them in.

It remains to explain who coordinates the motion of the figure skater's arms—or, in other words, what determines the period of oscillation of the angular velocity of the boiling liquid. According to our proposed mechanism for the oscillations, their period is equal to the sum of the time t_1 of vapor bubble ascent and the time t_2 it takes for water to drop from the surface to the bottom. It's known that large bubbles ($R > 0.1$ cm) rise in water with a velocity $v \cong 30$ cm/s. So the time it takes for a bubble to ascend is $t_1 \cong z_0/v \cong 0.7$ s. It isn't so easy to find the value of t_2, because the water moves downward along a complicated path. However, we can obtain a rough estimate using the formula $H = gt^2/2$ for free fall. At $H \cong 25\text{–}30$ cm, this equation yields $t_2 \cong 0.3$ s. Thus the period of oscillation is

$$T = t_1 + t_2 \cong 1 \text{ s},$$

which is in good agreement with experimental data.

It's also interesting to see how boiling of underheated water occurs when the circulation in the vortex is small. To do such an experiment, the water needs to be rotated rather slowly before boiling begins, and the phenomenon should be observed by illuminating the water with a lamp. If the rate of heating isn't very high and the heat source is small enough, a separate active center of bubble generation may arise near the axis of rotation. In this case, the second stage of boiling can be observed when the bubbles merge in the vertical direction, forming a vapor column. Naturally, when the vapor rises to the surface, no change in the angular velocity of the entire system occurs, because both the vapor source and the boiling mass of water are too small. However, a tiny eddy can be seen at the center of the water's surface, which casts a shadow on the bottom of the vessel. (The whirlpool produced on the surface forms a lens that scatters the light from the lamp.) This means that after the vapor column breaks through the surface, the cold water runs downward. Indeed, about 0.5–1 s after the eddy appears, the boiling center stops its activity, but in a few moments the whole process begins again.

In this article we considered only a few aspects of nucleate boiling, but they were enough to help us draw a number of important conclusions about the nature of this remarkable phenomenon. ◘

[1]You may recall the law of conservation of angular momentum from reading "A Venusian Mystery" in the July/August 1996 issue of *Quantum*.—*Ed.*

"SQUARING THE HYPERBOLA" CONTINUED FROM PAGE 31

need only take a comparatively small number of terms in the equation

$$e^x \cong 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$$

(this is important for work with computers). Second, this representation suggests a more insightful approach to the exponent—for example, it allows one to prove that e is an irrational number (see problems 11 and 12 below). We'll prove equation (17) for $x > 0$ only.

Set

$$T_n(x) = \left(1 + \frac{x}{n}\right)^n$$

and

$$S_n(x) = 1 + x + \frac{x^2}{1\cdot 2} + \cdots + \frac{x^n}{n!}.$$

We need to show that

$$\lim_{n\to\infty} S_n(x) = \exp n.$$

The binomial theorem gives us

$$\left(1+\frac{x}{n}\right)^n = 1 + n\cdot\frac{x}{n} + \frac{n(n-1)}{1\cdot 2}\cdot\frac{x^2}{n^2} + \cdots + \frac{n(n-1)(n-2)\ldots(n-k)}{k!}\cdot\frac{x^k}{n^k} + \cdots + \frac{x^n}{n^n}$$
$$= 1 + x + \frac{\left(1-\frac{1}{n}\right)}{1\cdot 2}x^2 + \cdots + \frac{\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{k!}x^k + \cdots + \frac{x^n}{n^n}.$$

All the numbers in parentheses on the right are less than one. If we replace all of them with 1, the right side will increase, and therefore

$$T_n(x) = \left(1+\frac{x}{n}\right)^n < S(n). \quad (18)$$

On the other hand, if we get rid of all but the first k terms on the right side of this inequality, we get

$$\left(1+\frac{x}{n}\right)^n > 1 + x + \left(1-\frac{1}{n}\right)\frac{x^2}{2!} + \cdots + \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{k-1}{n}\right)\frac{x^k}{k!}.$$

If k is constant and n grows to infinity, the right side of this inequality converges to $S_k(x)$, since every factor in parentheses tends to 1, while $(1 + x/n)^n$ approaches $\exp x$. Finally, we have

$$\exp x > S_k(x)$$

for all k—that is, the sequence $S_k(x)$ is bounded above and (for $x > 0$) increases. Therefore, it has a limit, so that

$$\lim_{n\to\infty} S_k(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + \cdots \le \exp x.$$

From equation (18) we find (proceeding to the limit for $n \to \infty$) that

$$\lim_{n\to\infty} S_n(x) \ge \exp x.$$

It follows that

$$\exp x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots.$$

Problems

1. (a) Prove the inequalities

$$\frac{nx}{n+x} < n\ln\left(1+\frac{x}{n}\right) < x$$

($x > -n$, where n is a natural number).
(b)Use them to verify the equation

$$\exp x = \lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n.$$

2. Find the following limits:
(a) $\lim_{n\to\infty} \sqrt[n]{n}$;
(b) $\lim_{n\to\infty} \frac{n}{2^n}$;
(c) $\lim_{n\to\infty} \frac{\ln n}{n}$;
(d) $\lim_{n\to\infty} \frac{\ln n}{\sqrt{n}}$.

3. Find the areas of the curvilinear trapezoids defined by the functions $y = a^x$ and $y = \log_a x$.

4. Find following limits:
(a) $\lim_{n\to\infty}\left(\frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n}\right)$;
(b) $\lim_{n\to\infty}\left(\frac{1}{n+x} + \frac{1}{n+2x} + \cdots + \frac{1}{n+nx}\right)$.

5. Prove that

$$\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} < \ln n < 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

6. Prove that the sequence

$$a_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln(n+1)$$

converges.

7. Prove that

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4}\cdots = \ln 2.$$

8. Find the slopes of the tangents to the curves $y = a^x$ and $y = \log_a x$.

9. Find the limit

$$\lim_{n\to\infty} n\left(\sqrt[n]{n} - 1\right).$$

10. Use the geometric definition of the logarithm to verify the inequalities

$$\frac{2x}{2+x} < \ln(1+x) < \frac{x(x+2)}{2(x+1)}$$

for $x > 0$ and use them to show that

$$x - \frac{x^2}{2} < \ln(1+x) < x - \frac{x^2}{2} + x^3.$$

11. Prove that

$$0 < e^x - \left(1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!}\right) < \frac{x^n e^x}{(n+1)!}$$

for $x > 0$.

12. Use problem 11 to prove that number e (a) is irrational, (b) is not a solution of any equation $ax^2 + bx + c = 0$ with whole coefficients ($a > 0$).

Bulletin Board

Fractal Mania

Software and book
by Phil LaPlante
McGraw-Hill, Blue Ridge Summit, PA 17294-0850
$29.95 (PC)

The beauty of fractals is in the images created. In print and on the computer, LaPlante has produced many of the basic images students usually use to study fractals and chaos theory. Test material concisely defines terms and discusses concepts of fractals and chaos theory by using examples effectively. The programming and math code used to produce the computer images are printed in hard copy. Given the proper computer tools and a sufficient background in math and computer programming, a user can manipulate the programs. This leads to an understanding of a basic concept of fractals—slight changes in initial conditions iterated many, many times lead to dramatic differences in output. This sensitive dependence on initial conditions is a theme that unites unstable systems, fractals, and chaos. Chapters in the text include "Foundations of Chaos and Fractal Theory," "Chaos and Fractals in Nature," and "Simulated Fractals and Chaos." The appendixes include an explanation of Turbo Pascal graphics.

As suggested by the author, the text and software can be used to supplement courses in Pascal programming, precalculus, geometry, and computer graphics. The software can be used effectively by a teacher with little computer or math expertise to demonstrate fractals and chaos theory. The mathematical basis for the production of the fractal images is carefully explained as is the programming code, but a user does not necessarily need to understand them to benefit from this material. Understanding the formulas can help unlock some of the mysterious beauty of the fractal images. A strength of the materials is that a user can explore the "what if" possibilities by changing the algorithms used to generate fractals. Users will certainly enjoy the many Julia sets produced.

For the science teacher, the examples are a bit weak. However, the study of chaos theory and fractals as it applies to science is a new study with great possibilities, as can be seen in the chapter "Chaos and Fractals in Nature." An explanation of self-similarity using a coastline does not clearly illustrate the concept. A wolf–caribou relationship is graphed with the aid of a Lotus 1-2-3 spreadsheet. A user can manipulate some variables, but, as in any natural system, not all variables affecting population can be factored in.

It is intriguing to use a relatively simple mathematical code to produce an image from nature—for example, a fern—but isn't this what DNA is all about? Several images of nature are produced with some success, but it takes a bit of imagination to see the images in other scenes. The problem of accurately predicting the weather is discussed, as are problems in predicting stock prices and fluid dynamics.

Programs are written using Borland's Turbo Pascal 5.5 compiler, and they require an IBM-compatible PC with an Intel 80286 processor and an EGA or VGA monitor. Users with Turbo Pascal 5.5 can modify the programs and discover that changing one line will lead to vastly different results. This provides great opportunities for interaction and hence understanding.

Fractal Mania is a concise set of supplemental text and software, which is very useful for introductory work in fractals and chaos theory.

—Carol Houck
Broward County Schools
Fort Lauderdale, Florida

Entomological CyberTeaser

Sometimes the long way around is the shortest way there. At least, that was the case with the March/April CyberTeaser (brainteaser B200 in this issue) posted at the *Quantum* Web site. Here are the first 10 respondents who submitted a correct answer:

Matthew Spencer (State College, Pennsylvania)
A. S. Sudheesh (Bangalore, India)
Robin Damion (Nottingham, England)
Bob Lind (Potosi, Wisconsin)
Simona Nikolova (Regina, Saskatchewan)
Oleg Shpyrko (Cambridge, Massachusetts)
Jonathan Devor (Jerusalem, Israel)
Peter DiFiore (Jacksonville, North Carolina)
Charles Kehoe (Wichita, Kansas)
Jim Paris (Doylestown, Pennsylvania)

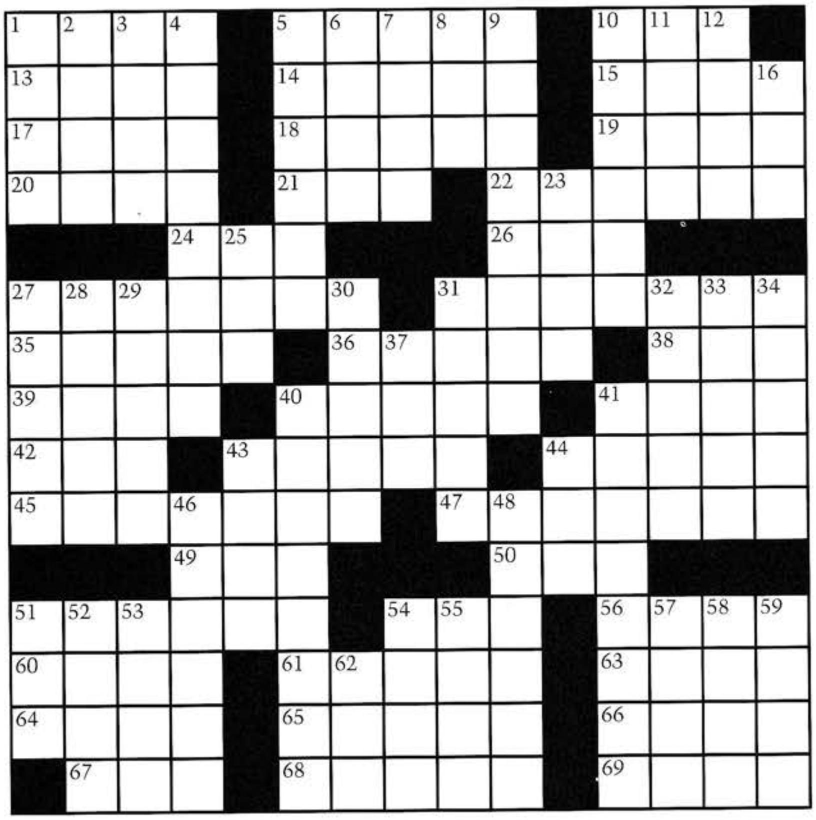
These folks will receive a *Quantum* button and a copy of this issue. All those who submitted a correct answer were eligible for a drawing to receive a copy of *Quantum Quandaries*, our collection of brainteasers.

One person almost slipped through a loophole in the statement of the problem as posted on the Web (we closed it in the print version). While most of our entrants assumed that one ant would *walk* to the other ant, this contestant took advantage of our loose wording and assumed that the ants could *fly* (and some ants do fly—for a while, anyway). But he made the fatal mistake of not explicitly stating the actual distance—he just described the path.

Let's hope our current CyberTeaser is airtight. You'll find it at **http://www.nsta.org/contest**.

Criss × cross science

by David R. Martin



Across

1 Shady mountain side
5 703,149 (in base 16)
10 47,821 (in base 16)
14 Brazilian novelist Joao Guimaraes ___
15 Cosmologist Sir Hermann ___
16 Nautical attention-getter
17 Mathematician ___ Artin
18 699,818 (in base 16)
19 Placed
20 Ordinate's companion
22 Quality of sound
24 Western Indian
25 ___-Einstein statistics
26 Worked hard
29 Geologic period
33 Happen
34 Varnish ingredient
35 Reciprocal ohm
36 The thing here
37 Type of acid
38 60,141 (in base 16)
39 Billion years
40 Type of equation
41 Aquatic plant
42 Like some iron
44 Horses
45 Middle East country
46 And so on: abbr.
47 Seed proteins
50 Cyclotron-resonance maser
55 German title
56 Fuming sulfuric acid
58 ___ Descartes
59 Leave ___ Beaver
60 Sodium: comb. form
61 Interested in
62 British archaeologist ___ Garstang
63 Biological stain
64 Anthropologist Carleton ___

Down

1 Carbamide
2 Exploding device
3 Ad phrase
4 Math. course
5 Degraded
6 Anthropologist Franz ___
7 Journalist Carr Van ___
8 Computer language
9 Simple molecular form
10 ___ series (of spectroscopy)
11 *Moby Dick* character
12 Coconut husk fiber
13 Unit of force
21 Roman road
23 Book's ID
25 Alkaline
26 Type of pole
27 Biochemist Severo ___
28 Cake topping
29 ___ oxide (CeO_2)
30 Real or virtual follower
31 In front
32 Points of minimum disturbance
34 Large thrush
37 Methyl ethyl ketone
38 Like some motors
40 Bird nostril's covering
41 10^{-18}: pref.
43 $CN_4(C_6H_5)_3CH$
44 Sunday speech
47 Alphabet run
48 Apollo's mother
49 Norse goddess
50 Obtains
51 Soviet cosmonaut ___ Gagarin
52 Gambling town
53 Aware of
54 Element 10
57 ___-tzu (Chin. philosopher)

SOLUTION IN THE NEXT ISSUE

SOLUTION TO THE JANUARY/FEBRUARY PUZZLE

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

ANSWERS, HINTS & SOLUTIONS

Math

M196

It's clear that in the number we seek all the digits to the left of the decimal point are zeros, as is the digit in the first decimal place (otherwise the number could not grow so much when we exchange the digits). Let x be the number we seek, and a the digit in the fifth decimal place—$x = 0.0{*}{*}{*}a{*}\ldots$. If we exchange the digits in the first and fifth places, we obtain the number

$$x + \frac{a}{10} - \frac{a}{10^5}.$$

Thus we obtain the equation

$$x + \frac{a}{10} - \frac{a}{10^5} = 1996x,$$

from which we get

$$x = \frac{9999a}{1995 \cdot 10^5} = \frac{3333}{665 \cdot 10^5}a = (0.0000501\ldots)a.$$

But a is the fifth decimal digit of x. If we look at how the multiplication indicated above is actually carried out, this means that the number $5a$ must end with the digit a. This is possible only if $a = 5$. Thus

$$x = \frac{3333 \cdot 5}{665 \cdot 10^5} = \frac{3333}{1330000}.$$

M197

Certainly, if $x^2 + 3y = p^2$, then $p > x$, so we can write $x^2 + 3y = (x + m)^2$ for some positive integer m. Similarly, we can write $y^2 + 3x = (y + n)^2$ for some positive integer n. From these relations we find that

$$x = \frac{2nm^2 + 3n^2}{9 - 4mn},$$
$$y = \frac{2n^2m + 3m^2}{9 - 4nm}.$$

But x, y, m, n are all natural numbers. Thus $9 - 4mn \ge 0$, $mn = 1$, or $mn = 2$. In the first case, $m = n = 1$, from which we get $x = y = 1$. In the second case, either $n = 2$, $m = 1$ (which gives us $x = 16$, $y = 11$), or $n = 1$, $m = 2$ ($x = 11$, $y = 16$).

M198

We have

$$a + b = x + y + \frac{x+y}{xy} = (x+y)\frac{xy+1}{xy}. \qquad (1)$$

Also,

$$ab = xy + \frac{1}{xy} + \frac{x}{y} + \frac{y}{x}.$$

Thus

$$ab - c = \frac{x^2 + y^2}{xy} = \frac{(x+y)^2 - 2xy}{xy} = \frac{(x+y)^2}{xy} - 2.$$

Therefore,

$$ab - c + 2 = \frac{(x+y)^2}{xy}. \qquad (2)$$

Let's square equation (1) and divide the result by equation (2). We get

$$\frac{(a+b)^2}{ab-c+2} = \frac{(xy+1)^2}{xy} = xy + \frac{1}{xy} + 2.$$

Thus

$$\frac{(a+b)^2}{ab-c+2} = c + 2.$$

or, after simplifying,

$$a^2 + b^2 + c^2 = abc + 4.$$

M199

First we'll construct a case in which six committees are not enough. Consider a group of seven deputies. We can number them 1, 2, 3, 4, 5, 6, 7. Assume that the first deputy slapped deputies 2, 3, and 4. We can write this as $1 \to 2, 3, 4$. Continuing on in this fashion, we have $2 \to 3, 4, 5$; $3 \to 4, 5, 6$; $4 \to 5, 6, 7$; $5 \to 6, 7, 1$; $6 \to 7, 1, 2$; $7 \to 1, 2, 3$.

We can express this situation graphically, as in figure 1. The vertices denote the deputies, and the arrows the relations among them. It's clear that no pair of deputies can be appointed to the same committee. Each deputy has slapped or been slapped in the face by every other deputy!

Now let's prove that seven committees are enough. We'll do it by induction on the number of deputies.

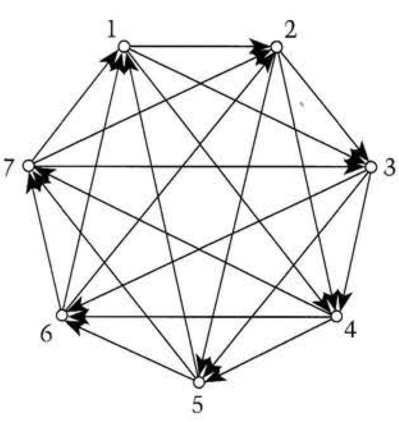


Figure 1

Suppose that it's true for any parliament of n deputies. We must show that it must be true for any parliament of $n + 1$ deputies. Take any such parliament. Then there must be a deputywho was slapped no more than three times (otherwise the number of blows received would be greater than the number of blows given). According to the inductive hypothesis, the remaining n deputies can be appointed to seven committees, and thus the condition of the problem is satisfied. (The fact that some of them perpetrated fewer than three slaps can only make our task easier.) But since the last deputy slapped three others and was in turn slapped no more than three times, there must exist a committee to which he can be appointed.

M200

Let O be the center of the circle circumscribed about $ABCD$, Q the center of the circle we seek, N the point diametrically opposite to L, and F the point where the smaller circle touches BM. Suppose that $NK = c$ and x is the radius we're looking for. We have

$$BK^2 = BK \cdot KC = NK \cdot KL = c(a + b).$$

Applying the Pythagorean theorem to triangle BKM, we get

$$BM^2 = BK^2 + KM^2 = c(a + b) + a^2. \quad (1)$$

We know that BC is parallel to MQ (since MK and MQ are bisectors of angles that are adjacent and supplementary). From this it's not hard to see that triangles BKM and MFQ are similar right triangles. Therefore, we can write $QM = (BM \cdot QF)/MK$. Substituting $QF = x$, $MK = a$, and the value of BM found in equation (1), we find

$$QM = \frac{x}{a} \cdot \sqrt{c(a+b)} + a^2. \quad (2)$$

We now write down the Pythagorean theorem for triangle QMF, using the fact that $OQ = (a + b + c)/2 - x$, $OM = |ML - OL| = |b - (a + b + c)/2| = \frac{1}{2}|a + c - b|$, and the value of QM we've found above (see equation (2)):

$$\left(\frac{a+b+c}{2} - x\right)^2 = \frac{(a-b+c)^2}{4} + \frac{x^2}{a^2}\left[c(a+b) + a^2\right]. \quad (3)$$

After a tedious but straightforward simplification of this equation, all the terms containing c^2 drop out. Then, if we collect terms with c and without c separately, we find

$$c[(a + b)x^2 + a^2x - a^2b] + a^2[(a + b)x - ab] = 0.$$

The first quadratic trinomial factors, and we get

$$c[(a + b)x - ab](x + a) + a^2[(a + b)x - ab] = 0.$$

Then the whole left-hand expression factors:

$$[(a + b)x - ab][c(x - a) + a^2] = 0.$$

Setting each factor equal to zero, we find that

$$x = \frac{ab}{a+b}.$$

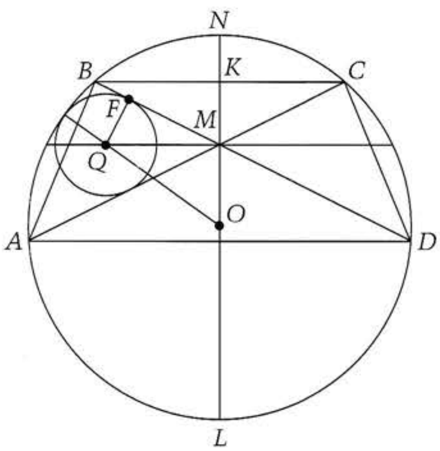


Figure 2

Physics

P196

Decompose the vector $\mathbf{a}$ of the total acceleration of point A at the moment the rods form a 90° angle into the sum of two accelerations $\mathbf{a}_1$ and $\mathbf{a}_2$ directed along the corresponding rods (see the figure accompanying the problem). To calculate $\mathbf{a}_1$ and $\mathbf{a}_2$ we use the fact that the acceleration does not vary when one inertial system of reference is substituted for another.

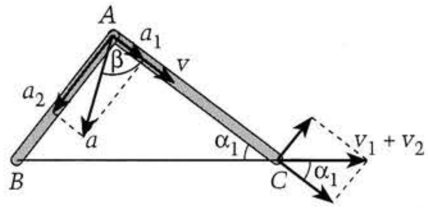


Figure 3

Let's switch to a reference frame moving to the left with velocity v_2. In this system the left-hand end of the left rod (point B) is at rest, and the right-hand end of the right rod (point C) moves to the right with a velocity $v_1 + v_2$, while point A describes a circle of radius L_2 (fig. 3). Thus the acceleration $\mathbf{a}$ can be decomposed into centripetal and tangential components. The components correspond to the accelerations $\mathbf{a}_2$ and $\mathbf{a}_1$, respectively.

The magnitude of the centripetal acceleration $a_2 = v^2/L_2$, where v is the linear velocity of point A's motion along the circle. At the moment the rods form a 90° angle, the velocity vector $\mathbf{v}$ is directed along the right-hand rod. At that very instant the right-hand end of rod AC has a velocity $v_1 + v_2$ directed to the right. Since the rod is rigid, the projections of the velocities of its ends on the rod itself must be equal (otherwise its length would vary). Thus (see figure 3)

$$v = (v_1 + v_2)\cos\alpha_1 = \frac{(v_1 + v_2)L_1}{\sqrt{L_1^2 + L_2^2}}.$$

Therefore,

$$a_2 = \frac{v^2}{L_2} = \frac{(v_1 + v_2)^2 L_1^2}{\left(L_1^2 + L_2^2\right)L_2}.$$

Similarly, switching to a reference system moving to the right with a velocity v_1, we get

$$a_1 = \frac{(v_1 + v_2)^2 L_2^2}{\left(L_1^2 + L_2^2\right)L_1}.$$

The total acceleration is

$$a=\sqrt{a_1^2+a_2^2}=\frac{(v_1+v_2)^2}{\left(L_1^2+L_2^2\right)L_1L_2}\sqrt{L_1^6+L_2^6}.$$

The angle β formed by vector **a** and the rod AC is characterized by the formula

$$\tan\beta=\frac{a_2}{a_1}=\left(\frac{L_1}{L_2}\right)^3.$$

P197

When heated, the air will expand at constant pressure, determined by the weight of the air column above it. The air is considered a mixture of two diatomic gases, oxygen and nitrogen (we neglect all other gases). The molecular heat capacity of this gas when heated at a constant pressure is $2.5R + R = 3.5R$. We need to find only one more value: the total number of moles n of gas in the atmosphere. The pressure at the planet's surface is determined by the weight of the entire atmosphere. Because the thickness of the atmosphere is far less than the radius of the planet, we can neglect the decrease in the acceleration due to gravity with altitude.

If m is the mass of the atmosphere, S is the area of the planet's surface, and M is the mean molar mass of the gas, then for pressure $P = 1$ atm we have

$$P=\frac{mg}{S}=\frac{nMg}{S}$$

and

$$n=\frac{PS}{Mg}.$$

The time τ that it takes to raise the temperature of the atmosphere 1 K can be found from the heat balance equation:

$$3.5Rn\Delta T = NS\tau,$$

which gives us

$$\tau=\frac{3.5RP\Delta T}{MgN}\cong 10^4 \text{ s}.$$

P198

Figure 4 shows schematically how the gas discharge tube is connected. Let's find the minimal value of emf that saturates the current in the tube:

$$\mathscr{E}_{\min} = V_s + I_sR,$$

where $V_s = 10^3$ V and $I_s = 10^{-5}$ A (see the figure accompanying the problem). Inserting these numerical values we get $\mathscr{E}_{\min} = 4$ kV.

In our case, the emf of the source $\mathscr{E} > \mathscr{E}_{\min}$. Thus the current in the tube equals the saturation current. The voltage V across the tube is defined by the condition $\mathscr{E} = V + I_sR$, from which we obtain

$$V = \mathscr{E} - I_sR = 3 \text{ kV}.$$

The current in the tube could be determined in another way. Let's assume that the current in the circuit is less than the saturated current. In this case, Ohm's law can be written as

$$\mathscr{E} = I(R_i + R),$$

where R_i is the tube's internal resistance. It can be determined by the slope of the linear portion of the volt–ampere curve for the tube (see the figure accompanying the problem):

$$R_i=\frac{V_s}{I_s}=10^8\ \Omega.$$

Thus

$$I=\frac{\mathscr{E}}{R_i+R}=15\ \mu\text{A}.$$

This value is larger than I_s. Therefore, the initial assumption that $I < I_s$ was wrong. This means that the current flowing in the tube is saturated.

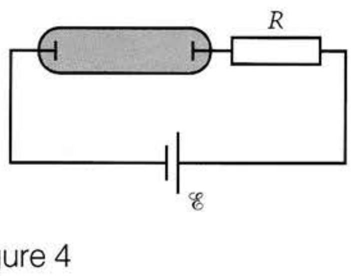


Figure 4

P199

The emf induced in a turn of radius r_i is equal to

$$\mathscr{E}_i=-\frac{\Delta\Phi_i}{\Delta t},$$

where $\Delta\Phi_i$ is the change in the magnetic flux through the area encircled by a turn during a short time interval Δt:

$$\begin{aligned}\Delta\Phi_i &= S\Delta B\\ &=\pi r_i^2B_0\left(\cos[\omega(t+\Delta t)]-\cos\omega t\right)\\ &=-2\pi r_i^2B_0\left[\sin\omega\left(t+\frac{\Delta t}{2}\right)\sin\omega\frac{\Delta t}{2}\right]\\ &\cong -\pi r_i^2B_0\omega\Delta t\sin\omega t.\end{aligned}$$

Since Δt is small,

$$\sin\omega\left(t+\frac{\Delta t}{2}\right)\cong\sin\omega t,$$

$$\sin\omega\frac{\Delta t}{2}\cong\omega\frac{\Delta t}{2}.$$

The total emf induced in the entire helix is the sum of all the electromotive forces induced in the turns:

$$\begin{aligned}\mathscr{E}=\sum\mathscr{E}_i&=\sum-\frac{\Delta\Phi_i}{\Delta t}\\ &=B_0\omega\sin\omega t\sum\pi r_i^2.\end{aligned}$$

The value πr_i^2 can be considered to be the volume of a cylinder of unit height having a base of area πr_i^2. Then $\sum\pi r_i^2$ is the sum of the volumes of such cylinders. Since the difference between the radii of adjacent cylinders is small (the number of turns is large), the total volume of all these cylinders is approximately equal to the volume of a cone of height n and base area πR^2 (fig. 5):

$$\sum\pi r_i^2=V=\frac{1}{3}\pi R^2 n.$$

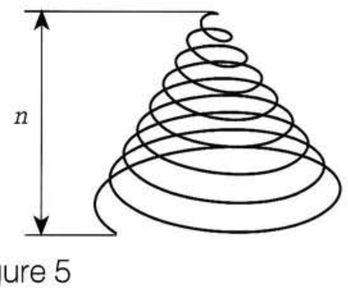


Figure 5

Thus the total emf induced in the helix is

$$\mathscr{E} = \frac{1}{3}\pi R^2 n B_0 \omega \sin \omega t.$$

P200

First we apply the lens formula

$$\frac{1}{d} + \frac{1}{f} = \frac{1}{F},$$

where d is the distance from the photographed object to the camera's objective, f is the distance from objective to the image, and F is the focal length of the objective (fig. 6).

When the newspaper is photographed in the first instance (without the spacer ring) where $d_1 = a$, a sharp image will be formed at a distance

$$f_1 = \frac{aF}{a - F}.$$

In this case the linear size of the image (see figure 6) is

$$l_1 = L\frac{f_1}{a} = L\frac{F}{a - F}.$$

When a spacer ring of thickness h is used to increase the distance between the film and the objective, a sharp image will be formed at a distance

$$f_2 = f_1 + h = \frac{aF}{a - F} + h$$

from the objective. In this case, the minimum possible distance between the newspaper and the objective is

$$d_2 = \frac{f_2 F}{f_2 - F} = \frac{[aF + h(a - F)]F}{F^2 + h(a - F)}.$$

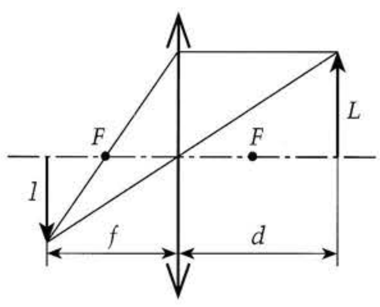


Figure 6

The corresponding size of the image is

$$l_2 = L\frac{f_2}{d_2} = L\left(\frac{F}{a - F} + \frac{h}{F}\right).$$

Thus the ratio of the sizes obtained on the film in the two cases is

$$\frac{l_2}{l_1} = \frac{h(a - F)}{F^2} + 1 = 5.5.$$

Brainteasers

B196

We could expand out and find that $x = 1$. Or we could guess at the beginning that $x = 1$ (it's easy to see that this works), then note that the equation is linear and cannot have any more solutions.

B197

The children poured water into the jar. The ball floated up and they took it out.

B198

The answer to the question is yes. There is a result in geometry that guarantees that for any two figures on the plane, there is a line that divides both in half. If you don't believe this theorem, look at figure 7.

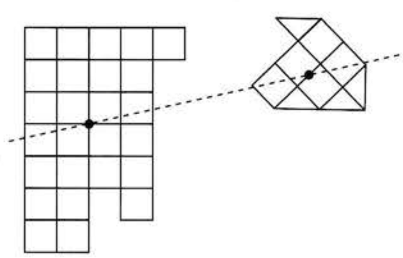


Figure 7

B199

The cooled oil has less volume than the heated oil. The "vacant" space had been occupied by air. So in the evening the scales showed a higher value than in the morning.

B200

Suppose one ant, in walking toward the other, tries to go toward the barrier rather than around it (along the edges of the square). Certainly, its path should be symmetric with respect to the barrier: if it follows two different paths, then one must be shorter than the other, and the longer path wastes time. Also, when it gets to the base of the barrier, the shortest way over is a path perpendicular to an edge that is not on the ground. If we fold the barrier flat against the original square, we get the diagram in figure 8. Since $ABCD$ is a square, we are comparing $a + b$ to $b + c$. Since $a > c$, the path along the edge of the original square is the shortest.

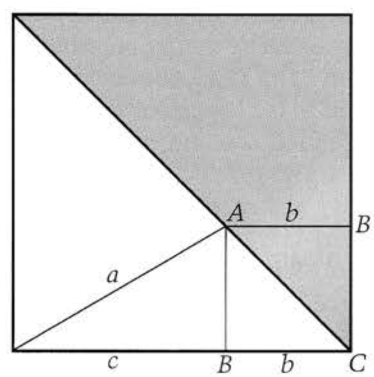


Figure 8

Index of Advertisers

Bottling milk

Let them count the whey

by Dr. Mu

WELCOME BACK TO Cowculations, the column devoted to problems best solved with a computer algorithm.

It's getting harder these days to make ends meet by selling milk in bulk 10-gallon cans to the creamery. The price of milk last year took a big tumble, with the corresponding devastating effect on our income. Every year for the past 10 years we've been paid a smaller percentage of the retail cost for a quart milk. It's down now to about 34% of what you pay in the store. Since the price of milk has not gone up noticeably, the small Wisconsin dairy farmer is being squeezed. Naturally, I am familiar with the feeling.

Farmer Paul, my boss and a member of the Wisconsin Institute for New Entrepreneurs, has decided to do something drastic to raise the family income. He's going to distribute our milk himself. He plans to put the milk in bottles, as he did years ago when he started farming, and deliver it door to door. All those beautiful glass bottles he carefully stacked away in crates 40 years ago will spring back into service once again. In the good old days we had long-necked glass milk bottles. The cream that collected at the top was poured off and put on your hot cereal. And

Art by Mark Brenneman

across the side in red, was printed our proud name—Progressive Dairy.

The other day, Farmer Paul took a 10-gallon can of milk and started pouring it into some bottles he had lying around the barn. (Remember: 2 pints = 1 quart, and of course 4 quarts = 1 gallon.) When he was done, he had the distribution of milk-filled bottles shown in the following table:

Size	2 gal	1 gal	1/2 gal	quart	pint	1/2 pint
Number	1	3	2	13	3	6

Hmmm, I thought, this is interesting. How many ways is it possible for Farmer Paul to distribute a 10-gallon can of milk among the bottle sizes that we have displayed above? Two distributions are the same if they fill exactly the same number of bottles of each size. I want to know the number of unique ways. But if you can solve the problem for 10 gallons, surely it would be just as easy to solve it for 20 or 30 or even *N* gallons. So that's it—that's your "Challenge Outta Wisconsin," or COW, for this issue:

COW 3. Find an efficient algorithm to cowculate the number of ways of pouring *N* gallons of milk into bottles of the sizes found in farmer Paul's milk cooler.

You can e-mail your cowculations to me, Dr. Mu, at drmu@cs.uwp.edu.

Our milk is exceptional, of course, and we guarantee:

You can whip our cream,
But you can't beat our milk.

—Dr. Mu

Solution to COW 2a

In the last issue, we introduced the idea of a superprime. A superprime is a prime number that remains a prime number when any number of digits are deleted from the right-hand side of the number. For example, 59393 is a superprime, because 59393, 5939, 593, 59, and 5 are all prime numbers.

Farmer Paul needs all the superprime numbers he can find to identify his growing herd of superprime cows. Your job was to find all the superprime numbers, if there is only a finite number, otherwise devise an algorithm to generate a generous supply.

Mathematica solution to COW 2a

We begin with the list of 1-digit primes which, of course, are all superprimes:

```
superPrime[1]={2,3,5,7}
{2, 3, 5, 7}
```

The 2-digit superprimes are constructed from the 1-digit superprimes by multiplying each digit by 10 and adding one of the four odd single digits {1, 3, 7, 9}. All primes larger than 5 must end with one of these digits. We add all possible combination of elements from **superPrime[1]*10** with {1, 3, 7, 9} in Mathematica as follows:

```
Outer[Plus,superPrime[1]*10,
{1,3,7,9}]
{{21, 23, 27, 29}, {31, 33,
37, 39}, {51, 53, 57, 59},
{71, 73, 77, 79}}
```

The output of the most recent computation is denoted by %. We then flatten this output to one list:

```
Flatten[%]
{21, 23, 27, 29, 31, 33, 37,
39, 51, 53, 57, 59, 71, 73,
77, 79}
```

These are all the possible 2-digit superprimes. Now we select out those that test positive for being prime (in other words, **PrimeQ[x]** is true):

```
Select[%,PrimeQ]
{23, 29, 31, 37, 53, 59,
71, 73, 79}
```

Voila! We have the 2-digit superprimes. Use the same technique to generate the 3-digit superprimes. This suggests a recursive relationship between **superPrime[n]** and **superPrime[n-1]**. Here is the definition that combines into one function what we developed step by step above:

```
superPrime[1]={2,3,5,7};
superPrime[n_]:=
Select[Flatten[Outer[Plus,
  superPrime[n-1]*10,
  {1,3,7,9}]],PrimeQ]
```

Now let's try out our new function:

```
superPrime[3]
{233, 239, 293, 311, 313,
317, 373, 379, 593, 599,
719, 733, 739, 797}
```

OK, let's print them all in a table:

```
Table[superPrime[n],{n,1,8}]
{{2, 3, 5, 7}, {23, 29, 31,
37, 53, 59, 71, 73, 79},
{233, 239, 293, 311, 313,
317, 373, 379, 593, 599,
719, 733, 739, 797}, {2333,
2339, 2393, 2399, 2939,
3119, 3137, 3733, 3739,
3793, 3797, 5939, 7193,
7331, 7333, 7393}, {23333,
23339, 23399, 23993, 29399,
31193, 31379, 37337, 37339,
37397, 59393, 59399, 71933,
73331, 73939}, {233993,
239933, 293999, 373379,
373393, 593933, 593993,
719333, 739391, 739393,
739397, 739399}, {2339933,
2399333, 2939999, 3733799,
5939333, 7393913, 7393931,
7393933}, {23399339,
29399999, 37337999,
59393339, 73939133}}
```

It turns out that there are no superprimes with more than eight digits:

```
superPrime[9]
{}
```

The **primeQ** function in Mathematica 3.0 is very fast and known to be correct for any $n < 2.5 \cdot 10^{10}$, which is far beyond the number of all the cows that ever lived. For those programming in C or Pascal, you'll need to write your own **primeQ** function.

Solutions in C

Kenny Brazier, a 10th grader at Pinewood School in Woodland Park, Colorado, submitted the following **primeQ** function, called **chkPrime**, in C. It's good to $n \le 2{,}147{,}483{,}647 = 2^{31} - 1$:

```
int chkPrime(long n)
{
  long i;

  for (i = 3; i*i <= n;i += 2)
    if ((n % i) == NULL)
  return 0;
return 1;
}
```

His complete short and sweet program can be seen on the Cowculations Web page given at the end of this column.

Superprime graph

How many superprimes are there for each value of *n*? We will cowculate that number and graph the results. In Mathematica, **Length** is used to measure the size of a list. For example, there are 8 superprimes with 7 digits:

```
Length[superPrime[7]]
```

```
8
```

Figure 1 is a graph of the number of superprimes for each digit size from 1 to 8:

```
ListPlot[Table[Length
[superPrime[n]],{n,1,8}],
PlotJoined->True,
AxesLabel->
{"digits","superprimes"}]
```

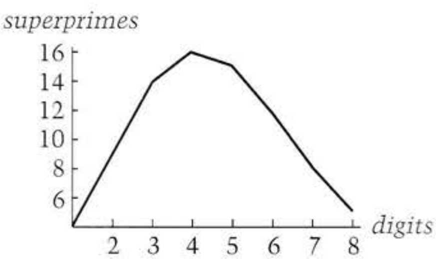


Figure 1

Solution to COW 2b

There is more than one way to slice a prime beef, and some of my left-footed bovine companions let me know about it. They asked, "What about primes that remain prime when any digit on the *left* is cut off?" They were called repusprimes, and COW 2b asked you to generate them.

Mathematica solution to COW 2b

Begin with the set of 1-digit primes:

```
repusPrime[1]={2,3,5,7}
```

```
{2, 3, 5, 7}
```

As we did before, add any digit from 1 to 9 on the left:

```
Flatten[Outer[Plus,10*Range[9],
repusPrime[1]]]
```

```
{12, 13, 15, 17, 22, 23,
25, 27, 32, 33, 35, 37, 42,
43, 45, 47, 52, 53, 55, 57,
62, 63, 65, 67, 72, 73, 75,
77, 82, 83, 85, 87, 92, 93,
95, 97}
```

(Note: I decided not to add a zero in front of a digit to avoid getting a number such as 007.) Now select out the primes:

```
Select[%,PrimeQ]
```

```
{13, 17, 23, 37, 43, 47,
53, 67, 73, 83, 97}
```

As before, we define the repusPrime function recursively:

```
repusPrime[n_]:=
Select[Flatten[Outer
[Plus,10^(n-1)*Range[9],
repusPrime[n-1]]],
PrimeQ]
```

Let's try it out:

```
repusPrime[3]
```

```
{113, 137, 167, 173, 197,
223, 283, 313, 317, 337,
347, 353, 367, 373, 383,
397, 443, 467, 523, 547,
613, 617, 643, 647, 653,
673, 683, 743, 773, 797,
823, 853, 883, 937, 947,
953, 967, 983, 997}
```

What happens at 8 digits?

```
Length[repusPrime[8]]
```

```
521
```

Unlike superprimes, repusprimes are growing. Let's check it out by graphing the number for each digit size:

```
ListPlot[Table[Length
[repusPrime[n]],{n,1,8}],
PlotJoined->True,
AxesLabel->
{"digits","repusprimes"}]
```

Figure 2 shows the plot of this function. It appears as if the left-handed farmers are going to have all the repusprimes they will ever need.

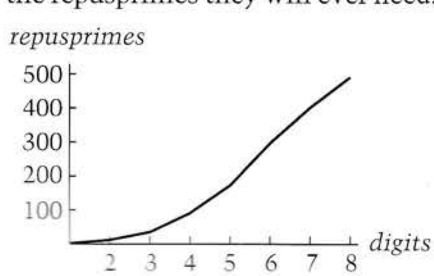


Figure 2

Other solutions

Benjamin Mathews, a 10th grader at St. Mark's School of Texas in Dallas, submitted an impressive, high-precision C solution to both the superprime and repusprime problems. It's posted along with Kenny's solution on the Cowculations Web page at http://usaco.uwp.edu/cowculations.

A big superprime "thank you" goes to two other cowhands for submitting Pascal solutions: Po-Shen Loh, a 9th grader at James Madison Memorial High School in Madison, Wisconsin; and Noam Zeilberger, a 10th grader at West Windsor-Plainsboro High School in West Windsor, New Jersey.

Finally, it was pointed out by Ben Mathews that 1997 is a repusprime year. Officially, 1 is not a prime, which is the only thing that prevents 1997 from being a superprime year. But in Wisconsin, nothing could prevent 1997 from being a Super Bowl Champs year.

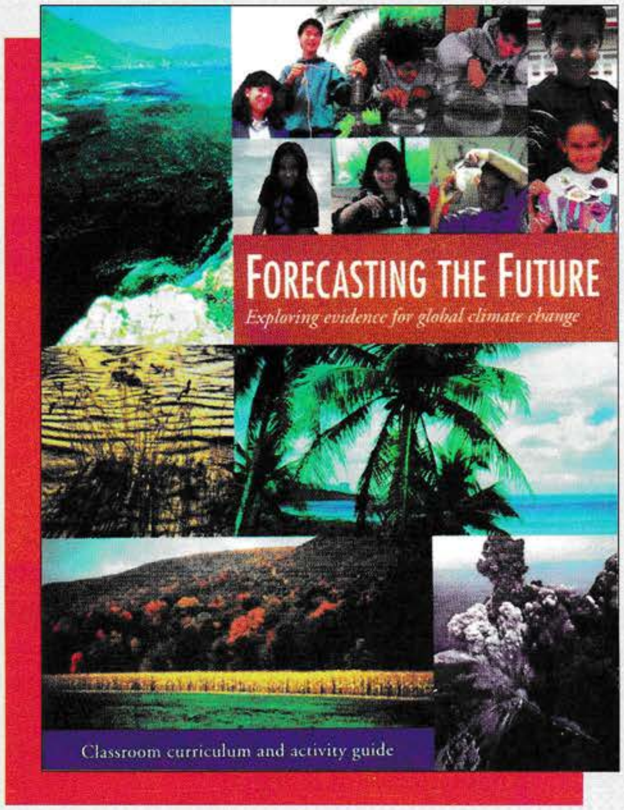
How to be a QUANTUM author

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

Send your inquiries to:

Managing Editor
Quantum
1840 Wilson Boulevard
Arlington VA 22201-3000

What does the future hold for Earth's climate?



Earth's climate has developed over billions of years—enabling plant and animal life to evolve and thrive. How does climate affect life on Earth? How does climate change? Do humans have an impact on it? What is the evidence?

Cut through the hype and confusion to address the vital subject of climate change in a scientific manner with *Forecasting the Future*. This new publication from NSTA and the Scripps Institution of Oceanography offers educators and students a multi-science route for exploring all aspects of climate and climate change. Fifty pages of background information and 14 classroom activities help students understand climate change through animal biology, chemistry, geology, meteorology, physics, and plant biology. The concepts are tied together by a sophisticated timeline that shows climatic changes, major extinctions, and other key events in Earth's history.

Explore the evidence for global climate change with *Forecasting the Future*—order your copy today!

Written by the Education Department of the Stephen Birch Aquarium-Museum, Scripps Institution of Oceanography

Grades 6–10, 1996, 160 pp.

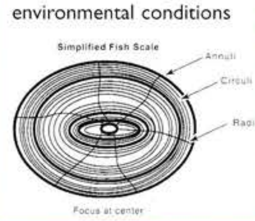
#PB118X $21.95

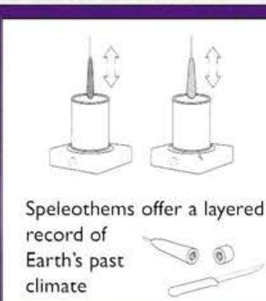
Highlights include:

- beautifully illustrated narrative of the components of climate
- 14 activities and more than 40 extension activities
- activities from six disciplines: chemistry, geology, physics, meteorology, plant and animal biology
- innovative timeline of Earth's natural history
- glossary of terms
- extensive annotated bibliography
- procedural drawings that illustrate activities
- time estimates and a teachers section for each activity

Learn about Earth's natural climate cycle and the effect humans have on global climate with *Forecasting the Future*—new from NSTA.

Fish scales can be used as indicators of changing environmental conditions





Speleothems offer a layered record of Earth's past climate

To order, call 1-800-722-NSTA