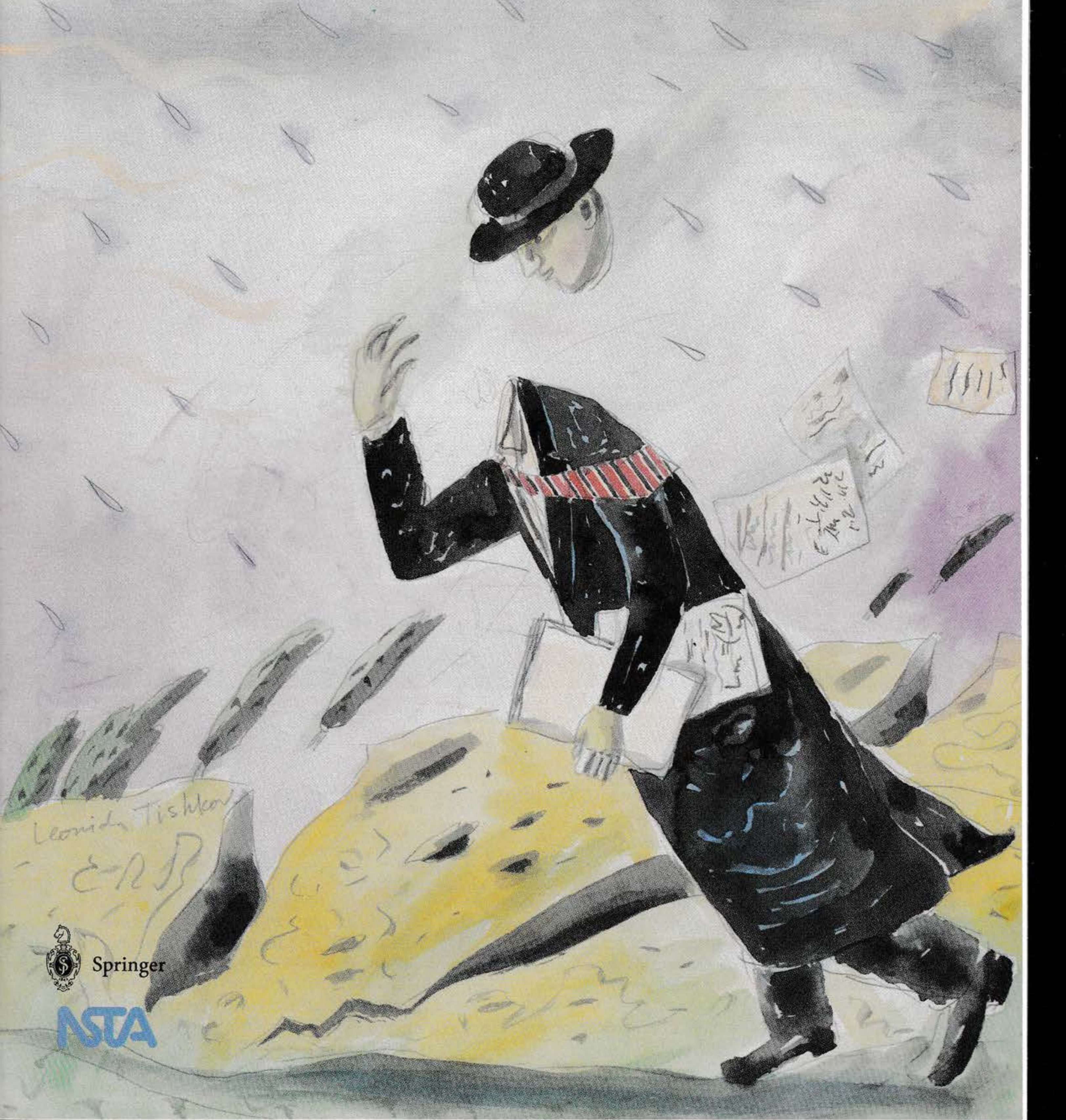
QUANTUM

JANUARY/FEBRUARY 2000

US $7.50/CAN $10.50



Springer

NSTA

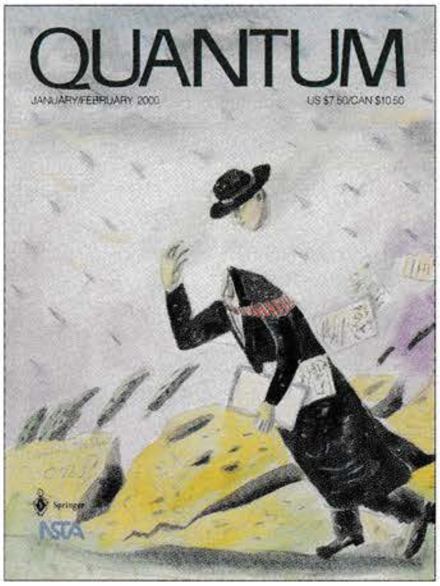
Oil on canvas, 18 1/4 × 21 3/4 inches, Collection of the Salvador Dali Museum, St. Petersburg, Florida

Disappearing Bust of Voltaire (1941) by Salvador Dali

AT FIRST GLANCE, THE PAINTING ABOVE SEEMS to be nothing more than a surreal gathering of exotically costumed individuals gathered at a desert locale. However, those familiar with the works of Dali know that his paintings should never be taken at face value. After closer inspection of the above image, the bust of Voltaire enters the room, emerging from the interplay of light and shadow. Training your eye to see beyond the obvious is talent that both artists and scientists seek to develop. This skill allows the artist to add nuance to a stretched canvas while allowing the scientist to discern detail in the canvas of the night sky. To learn more about how the brain processes the world that enters through our eyes, turn to "The Eye and the Sky" on page 10.

QUANTUM

JANUARY/FEBRUARY 2000 VOLUME 10, NUMBER 3



Cover art by Leonid Tishkov

It's important not to panic during earthquakes, and the more you understand about what's happening the less likely you are to lose your head. Fortunately, large earthquakes are rare events, thanks in part to the fluid nature of the forces at work. However, to play it safe, turn to "Fluids and Fault Lines" on page 4 to ground yourself in the fundamental physics behind earthquakes.

FEATURES

DEPARTMENTS

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.

NSTA *Quantum* (ISSN 1048-8820) is published bimonthly by the National Science Teachers Association in cooperation with Springer-Verlag New York, Inc. Volume 10 (6 issues) will be published in 1999–2000. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by Quantum Bureau (Moscow, Russia), as well as original material in English. **Editorial offices:** NSTA, 1840 Wilson Boulevard, Arlington VA 22201-3000; telephone: (703) 243-7100; electronic mail: quantum@nsta.org. **Production offices:** Springer-Verlag New York, Inc., 175 Fifth Avenue, New York NY 10010-7858.

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

Subscription Information:
North America: Student rate: $18; Personal rate (nonstudent): $25. This rate is available to individual subscribers for personal use only from Springer-Verlag New York, Inc., when paid by personal check or charge. Subscriptions are entered with prepayment only. Institutional rate: $49. Single Issue Price: $7.50. Rates include postage and handling. (Canadian customers please add 7% GST to subscription price. Springer-Verlag GST registration number is 123394918.) Subscriptions begin with next published issue (backstarts may be requested). Bulk rates for students are available. Mail order and payment to: Springer-Verlag New York, Inc., Journal Fulfillment Services Department, PO Box 2485, Secaucus, NJ 07094-2485, USA. Telephone: 1(800) SPRINGER; fax: (201) 348-4505; e-mail: custserv@springer-ny.com.

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

Advertising:
Representatives: (Washington) Paul Kuntzler (703) 243-7100; (New York) Jay Feinman (212) 460-1682; and G. Probst, Springer-Verlag GmbH & Co. KG, D-14191 Berlin, Germany, telephone 49 (0) 30-827 87-0, telex 185 411.

Printed on acid-free paper.

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

QUANTUM

THE MAGAZINE OF MATH AND SCIENCE

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

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

BRAINTEASERS

Just for the fun of it!

B281

Five at a time. Bill placed 9 empty glasses on a table and asked John to turn them upside down, turning 5 glasses at a time. Solve this problem using as few operations as possible.

B282

Failing inversion. Having solved the previous problem, John tried to invert the 9 glasses by turning 6 glasses at a time; however, he failed. Do you think it is possible?

B283

Strive for perfection. Can a number consisting of 300 ones and some number of zeros be a perfect square?

B284

Straight to the point. A line l and a point A are given on a plane. Using a compass and a straightedge, construct a perpendicular to l through point A, drawing not more than three lines and circles (the desired perpendicular counts as the last line drawn). Consider two cases, when point A does not lie on line l and when it does.

B285

Snow how? Since ice melts at 0 °C, how can streets in the winter remain covered with snow when the temperature is above zero?

ANSWERS, HINTS & SOLUTIONS ON PAGE 50

Art by Pavel Chernusky

Fluids and fault lines

Why large earthquakes are rather rare

by G. Golitsyn

THE WORLD WE LIVE IN IS very complicated and diverse, particularly in the details, which vary in time and space and are distributed in essentially a random manner. Our lives depend on many factors in the surrounding world: weather, climate, rainfall, flood, and drought. Many regions suffer from earthquakes, blizzards, hurricanes, and typhoons. These phenomena share a common feature known to everyone from experience: the stronger the disturbance, the more seldom it occurs. In other words, large deviations from the norm are rare events. This clearly reflects the *statistical* nature of our world, which is ruled by chance even while governed by physical laws. Certainly the conservation laws for energy and linear and angular momentum are always valid everywhere—one must only understand where and how to apply them. That is the main task of science, to provide a deeper understanding of the world around us.

Various technical applications of science require knowledge of the laws of fluid flow (by *fluid* we mean a liquid or gas). These laws help us understand how airplanes fly and ships move and how water flows in a pipe. Let's take a look at the factors that determine the basic features of a flow and its intensity and variations in time and space. We'll see why a large and interesting event requires a longer waiting period (expectation time), and how long one must wait (on average) to observe such an event.

Droplets and tubes

Suppose a body is immersed in some medium—say, in air or water. According to Archimedes' principle, the body will be acted upon by a buoyant force equal to the weight of the medium that is displaced by the body. It appears as if the acceleration due to gravity has changed by a factor of $(\rho_b - \rho_m)/\rho_b$, where ρ_b and ρ_m are the densities of the body and medium:

$$g' = g\frac{\rho_b - \rho_m}{\rho_b}.$$

In addition, every medium resists the motion of external bodies in it. Let's consider the motion of "large" and "small" bodies in air—for example, the fall of a drop of spray or the descent of a parachutist.

Using dimensional analysis, one can easily show that the resistive force acting on a large body of characteristic length r moving with speed u in a medium with density ρ_m must depend on these factors as follows:

$$F_a \sim \rho_m u^2 r^2.$$

Indeed, $(\text{kg/m}^3)(\text{m}^2/\text{s}^2)\text{m}^2 = \text{N}$. This is called the *aerodynamic drag force*. For small bodies the resistive force results from the mutual friction of layers of the surrounding medium. It is proportional to the first power of the speed u and the characteristic length r. Evidently, the proportionality factor must have dimensions of $\text{kg}/(\text{m}\cdot\text{s})$. It is known as the *viscosity coefficient* μ of the medium. Thus the viscous drag (Stokes') force is

$$F_\mu \sim \mu u r.$$

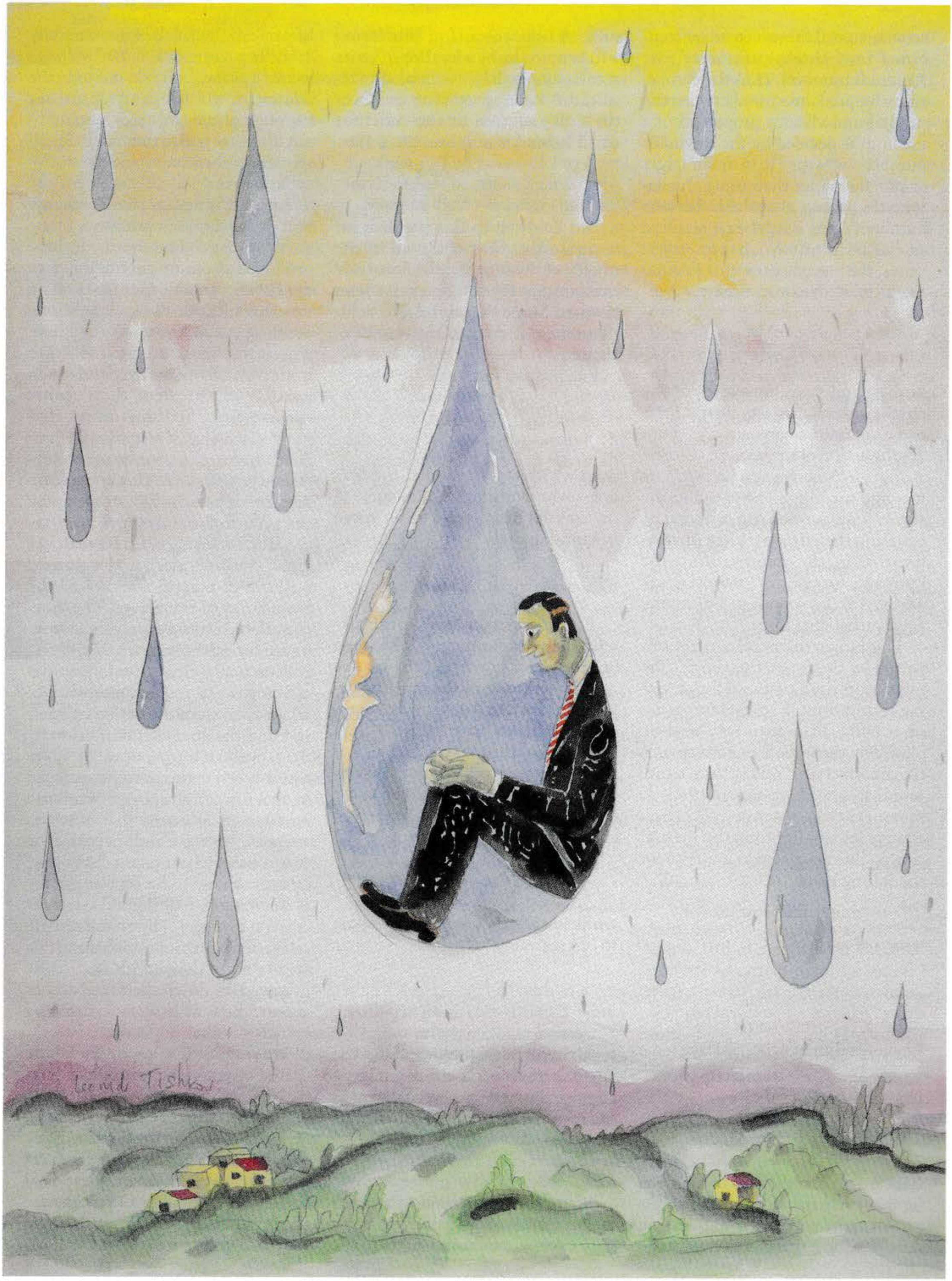
The ratio of these two drag forces is the Reynolds number:

$$Re = \frac{F_a}{F_\mu} \sim \frac{ur}{\mu/\rho_m} = \frac{ur}{\nu}.$$

For brevity here we introduce a new factor, known as the *kinematic viscosity* of the medium ν. For example, for air under standard conditions, $\nu \sim 10^{-5}$ m²/s.

Now we see that the predominance of either force (inertial or viscous) depends not only on the size of the body (whether it is large or small) but also on its speed and the kinematic viscosity of the medium it is

Art by Leonid Tishkov



moving in. All these factors are combined in a single parameter (the *Reynolds number*). Thus the competition for predominance between the inertial and viscous properties of a medium is decided by the Reynolds number—whether it is much larger or much smaller than unity. An interesting thing about this famous number is that it can be represented as the ratio of two characteristic times, the viscous time $\tau_v = r^2/\nu$ and the inertial (dynamic) time $\tau_i = r/u$:

$$Re = \frac{ur}{\nu} = \frac{r^2/\nu}{r/u} = \frac{\tau_v}{\tau_i}.$$

Both τ_v and τ_i are measures of the characteristic time required for a body to attain a constant speed under the action of the respective force.

Now let's consider a body falling in some medium. The speed of the steady-state motion can be found by equating the effective force of gravity (that is, the weight minus the buoyant force) mg', where $m = 4\pi r^3\rho_b/3$ and $g' = g(\rho_b - \rho_m)/\rho_b$, to the appropriate drag force F_μ or F_a.

The larger the acceleration imparted by the force, the shorter the relaxation time required for attaining equilibrium. If several forces act on a body, the main role will be played by the force that corresponds to the shortest relaxation time. Small Reynolds numbers ($Re \le 1$) mean that the viscous relaxation time is far smaller than the inertial time τ_v. In this case we can write for the falling body

$$mg' \sim \mu ur \sim \nu\rho_m ur,$$

which gives

$$u \sim g'\tau_v \frac{\rho_b}{\rho_m}.$$

In a similar way, for large Reynolds numbers ($Re \gg 1$) we have

$$mg' \sim \rho_m u^2 r^2,$$

and so

$$u \sim g' \frac{r}{u}\frac{\rho_b}{\rho_m} = g'\tau_i \frac{\rho_b}{\rho_m}.$$

Let's consider the flow of a viscous liquid of density ρ in a tube of radius r and length l under a driving pressure difference Δp between the ends of the tube. A unit volume of this liquid will be acted on by a force $\Delta p/l$, so its acceleration will be $a = \Delta p/\rho l$. As a result, the mean speed (averaged over the cross section of the tube) for small Reynolds numbers ($Re \le 1$) is

$$u_{\text{mean}} \sim a\tau_v \sim \frac{\Delta p}{\rho l}\frac{r^2}{\nu}.$$

The solution to this problem for narrow tubes was obtained in the middle of the nineteenth century by the famous French scientist Jean Leonard Marie Poiseuille. His solution differs from ours by a factor of about 1.

In contrast, for $Re \gg 1$ we get

$$u \sim a\tau_i \sim \frac{\Delta p}{\rho l}\frac{r}{u},$$

from which it follows that the drag is again proportional to the square of the speed, but now the mean speed is approximately

$$u_{\text{mean}} \sim \sqrt{\frac{\Delta p}{\rho l}}\, r.$$

This formula agrees beautifully with many experiments. It has long been used for practical calculations in connection with the pipeline delivery of various substances necessary for modern living.

Note that the mass flow rate G in the two cases has a different dependence on the radius at a given value of the pressure head $\Delta p/l$. In the viscous regime the mass flow is

$$G = \rho\pi r^2 u_{\text{mean}} \sim \frac{\Delta p}{l} r^4,$$

while in the nonlinear (with respect to Δp) regime

$$G = \rho\pi r^2 u_{\text{mean}} \sim \sqrt{\frac{\Delta p}{l}}\, r^{5/2}.$$

Comparing the two formulas, we see that the relative efficiency of the pipeline transport in the second case compared to the first case falls off as the pressure head or the tube radius is increased.

Turbulence

There are various energy sources in the world whose power changes only over periods comparable to the lifetime of the Earth (approximately 4.5 billion years or $1.5 \cdot 10^{17}$ sec). For example, solar energy is not only the source of all life on our planet (through photosynthesis by chlorophyll), but it is also the "fuel" for all motion in the oceans and the atmosphere.

Another huge source of energy that drives different processes in the Earth's crust and interior is the heat produced there by the radioactive decay of various elements. The mantle of the Earth extends from a depth of 3,000 km up to the crust, whose thickness is only 20–70 km under the continents and even smaller (5 km) near the oceanic ridges from which the crust originates. Heating of the mantle produces *convective motion* of the substance it is made of. This convection displaces the lithospheric plates of the crust in a nonuniform way by several centimeters per year. As slow as this shift is, it causes a buildup of elastic stresses at the boundaries of the plates. These accumulated stresses are partially relieved by earthquakes.

One can write an equation describing processes that alter the energy of a system over time by multiplying the equation of motion by the speed. The product of force times speed is the power developed by that force. If the power developed by a system is counterbalanced (on average over time and space) by the power of an extraneous source (say, by solar energy), the kinetic energy of the system will remain constant (also on average). The characteristic relaxation times for establishing this steady state depend on the forces involved (in much the same way as we saw previously when estimating velocities of falling bodies).

We start with a description of the spatial structure of a *turbulent* (that is, irregular) flow on a small scale, where the structure doesn't depend on the chosen direction and location.

More than 70 years ago the English scientist Richardson questioned whether it is possible to characterize the wind by a speed. He explained that the wind varies stochastically in time and space at any place on the Earth. He also proposed a qualitative description of turbulence as a process in which the main flow is unstable. The unstable flow decays and produces smaller vortices, which in turn generate even smaller vortices—and so on down to the very smallest scales. The smallest vortices *dissipate* by viscous damping because they are characterized by rather small Reynolds numbers.

In 1941 Andrei Kolmogorov (who, incidentally, was the founder of the Russian version of *Quantum*) wrote a paper describing the structure of turbulent flow. At virtually the same time, his postgraduate student Alexander Obukhov published a paper in which he obtained the so-called spatial spectrum of turbulence, among other interesting results. At the time, Kolmogorov did not know of Richardson's work, but he was well aware of the obstacles facing the creation of a theory of turbulence, and he provided some quantitative tools to overcome them. For example, to get around the wind speed problem he suggested considering as a parameter the mean square difference of the components of the velocities at two points separated by a distance r. In this case, the slow changes produced on larger scales by the *anisotropic* large vortices generated as a result of the instability of the main flow simply cancel out and do not have to be looked at. Kolmogorov proposed a quantitative description of the vortex fragmentation process that had been described qualitatively by Richardson.

If the development of the instability of the main flow is maintained at all times by some energy source (as, for example, the overall circulation of the atmosphere is maintained by solar radiation distributed nonuniformly over the Earth's surface), there must be a constant flow of energy from the large vortices to the small ones, where the turbulent energy is converted into heat on account of the viscosity of the fluid. This energy flow (or the rate of change of kinetic energy per unit mass per unit time) is usually measured in J/(kg · sec). The mean square difference of the moduli of the speed at two points separated by a distance r may be regarded as the relative kinetic energy of fluid particles of unit mass separated by the distance r. The cases of large and small r must be treated separately.

Suppose that r is large compared to the distances over which the viscous forces act. This distance is called *Kolmogorov's microscopic scale* $l_K = (\nu^3/\varepsilon)^{1/4}$ (prove to yourself that this formula is consistent with dimensional analysis). Then the Reynolds number is also large, and, by exploiting the analogy between the force and energy characteristics of a system, we can use the corresponding formula containing the dynamical time $\tau_i = r/u$ to obtain the famous Kolmogorov 2/3 law:

$$u^2 \sim (\varepsilon r)^{2/3}.$$

Obukhov found a formula for the spatial spectral density of kinetic energy in turbulent flow. Essentially it works as follows. The stochastic speed distribution can be represented as a sum of spatial harmonics (sinusoids) of various amplitudes and wavelengths, just as an arbitrary electrical signal can be decomposed into a sum of sinusoidal functions of time with different amplitudes and frequencies. The function that gives the contribution ("weight") of each sinusoid to the waveform is called the *spectrum*. The spectral description of any phenomenon is very effective for practical purposes.

Obukhov found that turbulence has a spatial spectrum

$$E(k) \sim \varepsilon^{2/3}k^{-5/3}, \; k = 2\pi/\lambda,$$

where λ is the wavelength of a spatial harmonic. This formula works in the atmosphere, in the oceans, in large wind tunnels, in stellar atmospheres, and even in the interstellar gas in our galaxy. Its correctness is confirmed by many direct measurements carried out by the world scientific community. There are excellent reasons why this theory is considered to be one of the most outstanding achievements of twentieth-century hydrodynamics.

On the other extreme, for small scales characterized by $r < l_K = (\nu^3/\varepsilon)^{1/4}$ the Reynolds number is not large, and we should therefore use τ_ν, the viscous time (because $\tau_\nu \ll \tau_i$). Thus,

$$u^2 \sim \varepsilon r^2/\nu.$$

At this scale one can do an exact calculation, which gives a numerical factor of 1/3 on the right-hand side of this formula.

Convection, which is an important class of motions in nature and technology, occurs when a light liquid is under a heavy one. This situation is typical in a pan of water set on a stove, which heats water from the bottom. Another example is cooling of water from the top. Heating from the bottom takes place in the atmosphere when soil is warmed by solar radiation. The resulting convection can be observed as a shimmering in the air over a plowed field, for example. Cooling from the top occurs in all natural reservoirs, where the thermal energy making up the heat of evaporation is taken from the top layer of water. The heat flux q_h removed from (or supplied to) the liquid is related to kinetic energy dissipation by the formula

$$\varepsilon = \frac{\alpha g q_h}{\rho c_p},$$

where α is the coefficient of volume thermal expansion of the liquid, g is the acceleration due to gravity, and c_p is the specific heat at constant pressure.

Of course, convection is distinct from locally homogenous and isotropic turbulence in that there is a preferred vertical direction (associated with the gravitational acceleration **g**). However, for a rough estimation of the convection rate one can use the formulas derived for turbulence. Of course, the numerical coefficients should be "customized" for convection instead of turbulence.

For practical purposes it is important to know the convection rate in viscous liquids, and so the corresponding formula has been tested in many experiments and numerical and analytical calculations. The expression obtained for the mean speed is

$$u \sim 0.1\sqrt{\frac{\varepsilon}{\nu}}\,r = 0.1\sqrt{\frac{\alpha g q_{\rm h}}{\rho \nu c_{\rm p}}}\,r.$$

Geophysicists have obtained the following parameters of the Earth's mantle: $\alpha \cong 4 \cdot 10^{-4}$ K^{-1}, $\rho \cong 3 \cdot 10^3$ kg/m^3, $c_{\rm p} \cong 3 \cdot 10^2$ J/kg $\cdot$ K, and $\nu \cong 10^{19}$ m^2/s. By inserting the mean value of the geothermal heat flux $q_{\rm h} = 0.08$ W/m^2 and the mantle thickness $r \cong 3000$ km into the above formula, we get a value of about 5 cm/year for the convective speed of the Earth's mantle. The actual speed of the lithospheric plates, as measured by navigation satellites, lies in the range of 1–10 cm/year.

The value 5 cm/year may seem to be extremely small, but with 52 weeks in a year, this rate is 1 mm/week. Incidentally, that is about the same as the growth rate of human fingernails, so we are quite familiar with such "slow" motion. (I believe that the first person to note this fact was the contemporary English geophysicist D. Mackenzie.)

In addition to gravitation, there is another factor that affects convection on the Earth. It is related to the *Coriolis force*, which results from the rotation of the Earth about its axis. This force always acts perpendicular to the velocity, so it can't perform any work and doesn't affect the energy of convection. Therefore, the Earth's spinning does not modify the formula for the energy dissipation rate. The Coriolis force introduces a new time scale $\tau_\omega = (2\omega \sin\theta)^{-1}$, where ω is the angular speed of rotation and θ is the angle between the axis of rotation and the local horizontal (for the Earth it is simply the latitude). In the middle latitudes $1/\tau_\omega = 2\omega\sin\theta \cong 10^{-4}$ s^{-1}.

The ratio of the inertial force to the Coriolis force is known as the *Rossby number*, $Ro = \tau_\omega/\tau_i$. It was named after a Swedish meteorologist who in 1940 introduced this ratio into science. In the case of large-scale atmospheric or oceanic motion, this number is far smaller than unity. For example, $Ro \cong 0.1$ for $r = 1000$ km and $u \sim 10$ m/sec. The Coriolis force is counterbalanced by a horizontal atmospheric pressure gradient. This explains the empirical rule long used by meteorologists: if you turn your back to the wind, the region of low pressure will be to your left and the high-pressure region to your right. Of course, in the southern hemisphere it is just the opposite.

Since the characteristic time τ_ω is much less than the inertial time, we can obtain the speed of convective flow in the rotating fluid:

$$u^2 \sim \varepsilon\tau_\omega, \text{ or } u \sim \sqrt{\varepsilon\tau_\omega}.$$

Many experiments (my own included) have yielded a value of 1.7 for the proportionality coefficient in this formula. When applied to the Earth's liquid core, the formula gives a speed of about 5 cm/year, which is sufficient to generate and maintain the *geomagnetic field*. It also yields speeds of 40–50 m/sec for hurricanes and typhoons, which agrees with meteorological evidence. Recently, convection in a rotating liquid was intensively studied by oceanographers in an attempt to describe the sinking of water at the edges of the ice sheets at high latitudes (the main process that "ventilates" the water in the depths of the ocean).

In the mid-1960s, A. M. Obukhov, then director of the Institute of Atmospheric Physics of the Russian Academy of Sciences, suggested I should search for data on atmospheric motion on other planets. I spent several years analyzing astronomical observations and simulating some features of the dynamics of the Martian atmosphere. This work culminated in a similarity theory for the circulation of planetary atmospheres. The theory gave reasonable estimates of the wind speed and the temperature difference driving the winds on the Earth: 12 m/sec and 45 K, respectively. It also gave predictions for these parameters on Venus, Mars, and the Saturnian moon Titan (on which the mass of the atmospheric column is 11 times greater than that on the Earth). The resulting formula for the mean wind speed is far from trivial:

$$u = a\sigma^{1/16}\, q^{7/16}\, c_{\rm p}^{-1/4}\, r^{1/2}\, m^{-1/2},$$

where a is a dimensionless factor (0.6 for the Earth); $\sigma = 5.67 \cdot 10^{-8}$ W/m$^2 \cdot$ K^4 is the Stefan-Boltzmann constant (which appears in the formula for thermal radiation $q = \sigma T^4$); m is the mass of the atmospheric column (10^4 kg/m^2); $q = q_\odot(1 - A)/4$ ($q = 240$ W/m^2) is the density of solar energy received by the planet on average over its surface, which has a reflectivity A ($A = 0.3$); and $q_\odot$ is the solar constant (1368 W/m^2). The numbers in parentheses are the corresponding values for the Earth.

This formula is too complicated for even a simplified analysis. Several years passed before the author guessed that the total kinetic energy of the atmosphere could be described by the formula

$$E = \frac{1}{2} m \cdot 4\pi r^2 u^2 = 2\pi a^2 Q\tau_e \approx 2Q\tau_e,$$

where Q is the total energy from solar radiation received over the planetary disk and $\tau_e = r/c$ is the ratio of the radius of the planet to the speed of sound, which is the damping time of pressure and density disturbances on a planetary scale. The ratio of the damping time τ_e to the inertial time τ_i is the *Mach number* $Ma = u/c$. Since $Ma \ll 1$, the damping time τ_e is the smallest time parameter.

In general, winds arise because planets are (approximately) spherical. They therefore have day and night and high and low latitudes, so that different parts of the planet are heated differently. This nonuniform heating is the primary cause of winds. For Venus the formula for the mean speed yields $u \cong 1$ m/sec. This estimate has been confirmed by direct measurements made by space probes in the lower hemisphere of that planet. For Mars the theoretical estimates of the wind speed turn out to be three times higher than the experimental data shows. This discrepancy presumably originates from the thin and transparent character of the Martian atmosphere. Under such conditions the main role in the heat exchange between the surface and atmosphere of the planet is played by thermal radiation instead of atmospheric dynamics.

Curiously, the reasoning that yielded the formula for the energy E can also be used to describe stochastic processes involving objects and events, that is, to determine the number of events occurring per unit time as a function of their mean intensity.

Earthquakes

Let's analyze the number of earthquakes occurring on the planet (fortunately, this number is small, and a strong earthquake is an extremely rare event). An earthquake is a very complicated process in its structure and nature. It was only 30 years ago that a more or less precise (about 20%) quantitative measure of earthquake intensity was proposed. It is based on the energy of the radiated waves detected by seismographic stations, which for the past 20 years have been organized into the Worldwide Standardized Seismographic Network (WWSSN). This parameter is the *seismic moment* $M = \mu_s Ss$, where μ_s is the shear modulus of the rock ruptured in the earthquake, s is the mean displacement of adjacent blocks of the crust during the earthquake, and S is the contact area along which this displacement occurs. The dimensions of M are N · m (newton meters), which means that it has dimensions of work or energy. Earthquakes relieve the stress $\Delta\sigma$ accumulated during displacement of the lithospheric plates. It is noteworthy that for different earthquakes $\Delta\sigma$ varies only slightly about its mean value (40 atm = 4 MPa = $4 \cdot 10^6$ N/m^2 = $4 \cdot 10^6$ J/m^3), although M can differ by many orders. This feature makes it possible to characterize every earthquake by its own linear scale L_m, area scale $S_m = L_m^2$, or volume scale $V_m = L_m^3$:

$$L_m = \left(\frac{M}{\Delta\sigma}\right)^{1/3}, \quad S_m = \left(\frac{M}{\Delta\sigma}\right)^{2/3},$$

$$V_m = \frac{M}{\Delta\sigma}.$$

The parameters L_m and S_m are useful in estimating the length and area of the rupture zone formed in the earthquake, while knowledge of the shear modulus enables one to determine the mean displacement s. For example, the most severe earthquake in the twentieth century occurred in Chile (May, 1960), where the rupture of the crust extended as far as 800 km and the displacement was more than 20 m!

Using dimensional analysis, we may write

$$M = a_s P\tau(\geq M),$$

where a_s is a constant factor, P is the power fed into the system (in our case of global statistics it is the total value of the geothermal power $P = 4 \cdot 10^{13}$ W, with a mean energy flux density of 0.08 W/m^2), and $\tau(\geq M)$ is the mean expectation time for an earthquake event with a seismic moment equal to or greater than M. This is precisely the distribution of earthquake events observed in the thin crust near the oceanic ridges, where the crust originates and is only about 5 km thick.

Seismologists usually write the distribution law for the mean occurrence rate of earthquakes as

$$N(\geq M) \sim \frac{P}{M^n},$$

where $N(\geq M) = 1/\tau(\geq M)$. According to carefully compiled statistics on earthquake events, the exponent n has a value of 1.05 for some sets of data and 0.94 for others. In any case, it is very close to unity. Bear in mind here that we know only the power of the energy source and are attempting to find a relationship between the intensity of an earthquake and its expectation time (or occurrence rate).

However, the overwhelming majority of earthquakes occur far from the oceanic ridges, so only a small percentage of all earthquakes (only about 50 events in the period from 1977 to 1992) obey the above distribution law. For most earthquakes, with seismic moments M less than 10^{21} N · m, the exponent n is less than 1. The experimental values of n compiled by different authors vary somewhat, but they all can be described by the value $n = 0.66 \pm 0.03$.

Now recall our scales of length, area, and volume (see formulas above). The latter formula was suggested in 1956 by the Japanese seismologist Chuji Tsuboi. It gives the volume in which the relief of mechanical stresses occurs. The heat flux, which is the ultimate source of the stresses in the crust, is delivered to an area of the crust S_m of thickness h, and thus the affected volume is $hS_m = h(M/\Delta\sigma)^{2/3}$. Therefore,

$$\frac{MN(\geq M)}{V_m} \approx a_s \frac{P}{hS_m},$$

which gives

$$N(\geq M) \cong 0.4PM^{-2/3}h^{-1}(\Delta\sigma)^{-1/3}.$$

The coefficient $a_s \cong 0.4$ in this formula was determined by comparing with the data in the catalog of global earthquakes.

This formula, which I published in 1996, not only explains the nature of the exponent $0.66 \pm 0.03 \cong 2/3$ but also reveals the factors that are conducive to earthquakes. For example, the thinner the crust, the smaller the mean expectation time for earthquakes of a given intensity. The differences in the experimental values of n are explained by the different degrees of rupturing of the crust in different earthquakes. When $n = 1$, the entire crust of thickness h is ruptured. When $n \cong 2/3$, only a partial rupture occurs in the crust, over an area S such that $\sqrt{S} < h$. Therefore earthquakes in the thin oceanic crust have a distribution with $n \cong 1$, while the far greater number of earthquakes that occur in the thick crust have a distribution described by $n \cong 2/3$. The power-law dependence of the earthquake occurrence rate on intensity, with an exponent of about 2/3, is named after the American seismologists Beno Gutenberg and Charles Francis Richter, who discovered it in 1941.

We could look at many more examples of the application of dimensional and similarity analyses, extract the characteristic times of other processes, and search for more analogies in events of absolutely different physical natures, but the universe won't fit into a short article!

Carl Friedrich Gauss

Nothing can be considered completed if anything remains to be done.

—Gauss

by S. Gindikin

IN THE NEW CENTURY, Gauss' scientific interests shifted away from pure mathematics. Periodically, he returned to it, each time achieving excellent results, worthy of a genius. In 1812, he published a work on the hypergeometric function. This function depends on three parameters. By assigning them different values, a great many of the functions that occur in mathematical physics can be obtained. A lot of credit for the geometric interpretation of complex numbers must go to Gauss. His works in geometry will be described below. However, mathematics was never again his main occupation. Here is a telling circumstance: in 1801, Gauss stopped keeping his diary regularly, although certain notes were made up to 1814. We rarely realize how short the "mathematical life" of Gauss was—less than 10 years, and the greater part of this time was spent on works that remained unknown to his contemporaries (elliptic functions).

Minor planets

We now talk about a new passion of Gauss. There has been a lot of discussion as to the reasons for which Gauss began to study astronomy.

First of all, it must be taken into account that since the time of Kepler, Galileo, and Newton, astronomy was the most impressive application of mathematics. This tradition was continued by Euler, d'Alembert, Clairaut, Lagrange, and Laplace. By predicting and explaining celestial phenomena, mathematicians felt as if they had been let in on the secrets of the Universe. Gauss, who was interested in practical computations, couldn't fail to try his power in this traditional field.

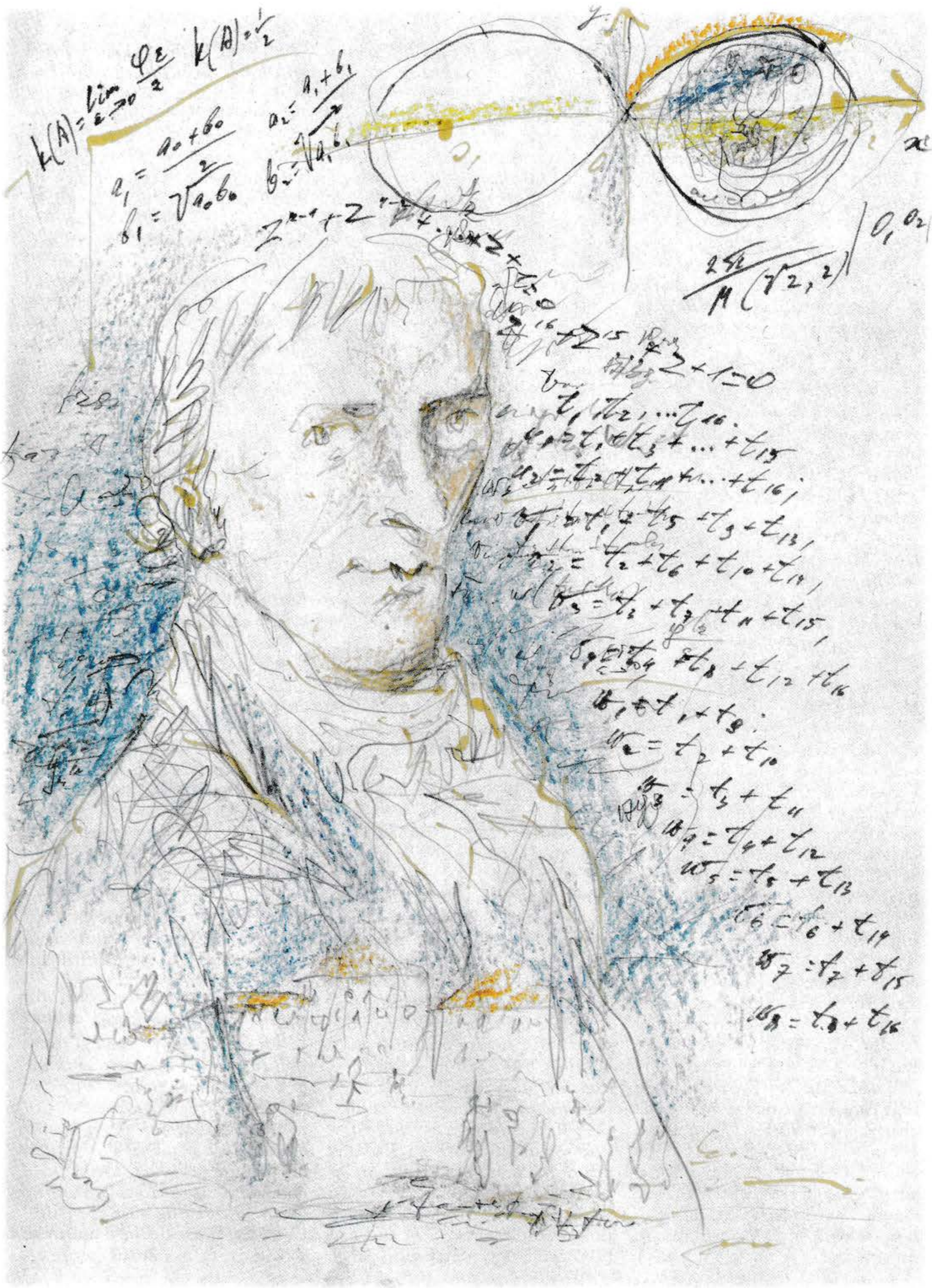
However, there were other, quite practical reasons. Gauss held the modest post of privatdocent in Brunswick, earning only 6 thalers a month. The stipend of 400 thalers from the Duke did not improve his material condition to a degree sufficient to support a family, and Gauss was thinking of getting married. It was difficult to obtain a professorship in mathematics, and, in addition, Gauss was not inclined to teaching. The growing number of observatories made a career as an astronomer possible.

Gauss began to study astronomy while still in Göttingen. Some observations were made in Brunswick—he even spent a part of the Duke's stipend on a sextant. Meanwhile, Gauss sought a suitable computational problem, while keeping busy solving minor problems. For example, he published a simple method for calculating the date of Easter and other cyclic feasts (at that time, very unwieldy methods were used for this purpose). An idea for a worthy problem occurred in 1801 under the following circumstances.

On the 1st of January 1801, the astronomer Piazzi, who was making up a catalog of stars, found an unknown star of the 8th magnitude. He observed it for 40 days and then addressed prominent astronomers, asking them to continue the observations. For various reasons, his request was not fulfilled. In June, this information reached Zach, who published the only astronomic journal at that time. Zach assumed that this was a planet that was suspected to exist between Mars and Jupiter. Zach's hypothesis seemed very likely and the "lost" planet had to be found quickly. For this purpose, its orbit was to be computed. However, it was beyond the possibilities of astronomers to compute an elliptic orbit knowing only the 9-degree arc

Art by Vadim Ivanyuk

This article is continued from the November/December 1999 issue of *Quantum*, pp. 14–19.



that had been observed by Piazzi. In September of 1801, Gauss abandoned all his affairs and got down to computations. He finished them in November, and the results were published in the December issue of Zach's journal. At night from the 31st of December to the 1st of January—exactly a year after Piazzi's observations—the renowned German astronomer Olbers found the planet on the basis of these calculations; the new planet was named Ceres. This discovery created quite a sensation.

On the 25th of March 1802, Olbers discovered yet another planet, called Pallas. Gauss quickly computed its orbit, showing that it also was located between Mars and Jupiter. The effectiveness of Gauss' computational methods became apparent.

Thus, Gauss gained recognition. One of the signs of this recognition was his election as a corresponding member of the St. Petersburg Academy of Sciences. Soon he was invited to take the post of Director of the St. Petersburg observatory. Gauss wrote that he felt flattered to be invited to the city where Euler had worked, and seriously considered accepting this offer. In his letters, Gauss wrote that the weather was often bad in St. Petersburg, so he would have time for his own studies in addition to making astronomical observations. He also wrote that the 1000 rubles he was offered were more than the 400 thalers which he got in Brunswick, but that the cost of living was higher in St. Petersburg.

At the same time, Olbers made efforts to keep Gauss in Germany. As early as in 1802, Olbers asked the curator of the University of Göttingen to invite Gauss to take the post of Director of the newly founded observatory. Olbers wrote that Gauss had an aversion to the chair of mathematics. Gauss agreed but moved to Göttingen only at the end of 1807. In the meantime, Gauss got married. In 1806, the Duke died. Gauss was sincerely attached to him, and now nothing held him in Brunswick.

Life in Göttingen was not easy for Gauss. In 1809, after the birth of a son, his wife died, and then the child died also. In addition, Napoleon levied a heavy tribute on Göttingen. Gauss himself had to pay an excessive tax of 2000 francs. Olbers and Laplace tried to pay this tax for Gauss, but he proudly refused. However, another anonymous benefactor paid the tax, and this time, there was no one to return the money to (later, it became known that this benefactor was the elector of Mainz, a friend of Goethe). Gauss wrote, between remarks on the theory of elliptic functions, that death was better than such a life. People didn't appreciate his work and considered him an eccentric at best. Olbers tried to comfort Gauss, saying that one shouldn't expect understanding but should pity people and serve them.

In 1809, the famous *Theoria Motus Corporum Coelestium* (Theory of Motion of Celestial Bodies Revolving around the Sun by Conic Sections) was published. This book was finished as early as in 1807. The delay in printing was partly due to the apprehension of the publisher that there would be no demand for the book in German, and Gauss refused to publish his book in French for patriotic considerations. The compromise settlement was to publish it in Latin. This was the only Gauss' book on astronomy (besides this book, he also published several papers).

In this book, Gauss presented his method for calculating orbits. In order to demonstrate the effectiveness of his method, he repeated the calculation of the orbit of the comet of the year 1769, which had been calculated earlier by Euler. Euler spent three days of intensive work on this job and lost his eyesight as a result. Gauss carried out the job in only an hour. Among other things, the book presented the least-squares method, which remains one of the most widespread methods for the analysis of observations. Gauss claimed that he had known this method since 1794 and had been using it systematically since 1802. (The least-squares method was published by Legendre two years before the *Theoria Motus Corporum Coelestium.*)

In 1810, Gauss received many honors: he won the prize of the Paris Academy of Sciences and the gold medal of the Royal Society of London, and he was also elected to membership in several academies.

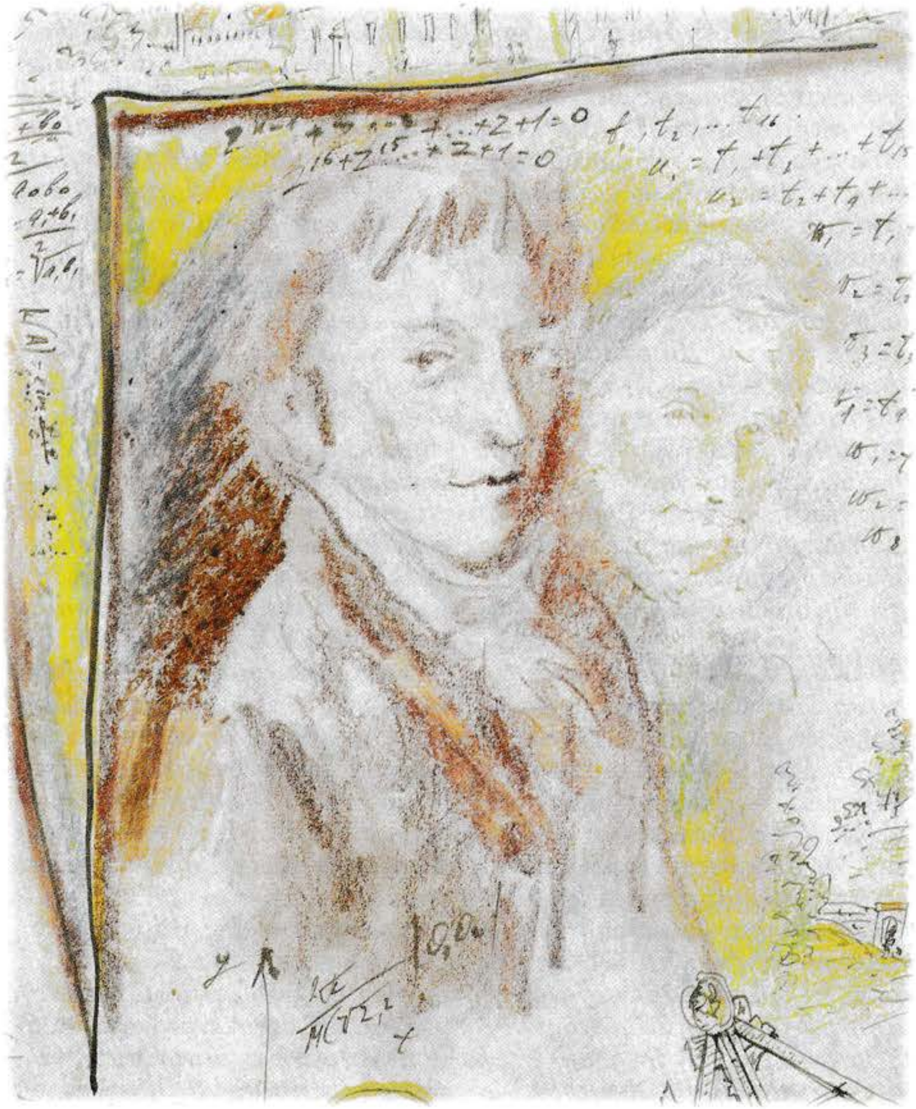
In 1804, the Paris Academy of Sciences chose as a topic for the great prize (a gold medal of 1 kg): the theory of perturbations of Pallas. The deadline was twice postponed (eventually to 1816) in order for Gauss to be able to present his work. Although Gauss' student Nicolai ("a youth tireless in calculations") helped him, the calculations still weren't finished because of Gauss' depressed state of mind.

Gauss continued his regular studies in astronomy almost until his death. The famous comet of 1812 (which "portended" the Moscow fire) was observed on the basis of Gauss' calculations. On the 28th of August, Gauss observed an eclipse. Gauss had many student astronomers (Schumacher, Herling, Nicolai, and Struve). The prominent German geometers Möbius and Staudt studied astronomy rather than geometry under Gauss. Gauss also corresponded with many astronomers, read papers and books on astronomy, and wrote reviews. We know a lot about his studies in mathematics from his letters to astronomers. The image of Gauss the astronomer was quite different from that of an inaccessible hermit, a notion that existed among mathematicians at that time.

Geodesy (Earth measurement)

In 1820, Gauss turned his attention to geodesy. As early as in the beginning of the century, he tried to use the measurements of an arc of a meridian carried out by French geodesists to set a standard for length measurements (the meter). He planned to use it to calculate the Earth's flattening near the poles. However, the arc turned out to be too short. Gauss dreamed of performing measurements of a sufficiently long arc of a meridian, but was able to get down to this work only in 1820. Although the measurements took as long as two decades, Gauss was unable to fully implement his idea. Investigations concerning the analysis of observations obtained in connection with geodesy were of major importance (his main publications on the least-squares method date to this time). Various geometric results were obtained in connection with the necessity of performing measurements on the surface of an ellipsoid.

In the 1820s a move to Berlin was discussed, where Gauss was to head an Institute. The most prominent young mathematicians were to be invited there, including Jacobi and Abel. The negotiations dragged on for four years—there were disagreements over whether Gauss was to give lectures and how much he was to be paid (1200 or 2000 thalers a year). Negotiations were unsuccessful. However, all was not lost—Gauss' salary in Göttingen was raised to what he was supposed to receive in Berlin.



Intrinsic geometry of surfaces

Geodesy was responsible for the fact that mathematics again became Gauss' main occupation for a short period of time. In 1816, he considered a generalization of the main problem of cartography—mapping one surface onto another so that the image was similar to the original in the smallest details. Gauss advised Schumacher to choose this question when announcing a competition for the prize of the Copenhagen scientific society. The competition was announced in 1822. The same year, Gauss presented a memoir in which he introduced characteristics that permitted a complete solution to the problem, particular cases of which had been analyzed by Euler and Lagrange (the mapping of a sphere or a surface of revolution onto a plane). Gauss gave a detailed description of the conclusions from his theory for numerous particular cases, some of which arose in geodesy.

In 1828, Gauss published his basic memoir on geometry called *General Investigations on Curved Surfaces.* It was devoted to the *intrinsic geometry of a surface*, that is to everything that is related to the surface itself rather than to its position in space.

Broadly speaking, the intrinsic geometry of a surface describes its properties that can be found when one "remains on the surface itself." One can measure distances on surfaces by stretching a thread so that it remains entirely on the surface. Such curves are called *geodesic arcs* (they are analogs of straight lines on a plane). One can analyze angles between geodesics and study the properties of geodesic triangles and polygons. If we warp a surface (thinking of it as a film which can neither be stretched nor torn), the distances between its points will remain unchanged: geodesics will remain geodesics, and so on.

It turns out that one can find out whether a surface is curved or not without leaving it. A "truly curved" surface cannot be developed onto a plane. Gauss introduced a numerical measure of the curvature of a surface.

Consider a neighborhood of a point A with an area of ε. Consider a normal (perpendicular) to the surface, of unit length, at each point of this neighborhood. For a plane, all the normals are parallel, while for a curved surface they diverge. Translate all these normals so that they start from the same point. Then, the endpoints fill a domain on the unit sphere. Let $\varphi(\varepsilon)$ be the (oriented) area of this domain. Then

$$\lim_{\varepsilon\to 0}\frac{\varphi(\varepsilon)}{\varepsilon}$$

is a measure of curvature of the surface at point A. It turns out that $k(A)$ remains the same under any warping of the surface. A part of a surface can be developed onto a plane if and only if $k(A) = 0$ for all its points. The measure of curvature is related to the sum of the angles of a geodesic triangle.

Gauss studied surfaces of constant curvature. The sphere is a surface of a constant positive curvature (for all its points, $k(A) = 1/R$, where R is the radius). In his draft notes, Gauss mentioned a surface of a constant negative curvature. Later, such a surface was called *pseudosphere,* and Beltrami proved that its intrinsic geometry was exactly the geometry of Lobachevsky.

Non-Euclidean geometry

There is evidence that Gauss was interested in the parallel postulate as early as in 1792. In Göttingen, he often discussed this problem with Farkas Bolyai, a student from Hungary. We know from Gauss' letter of 1799 to Bolyai that he clearly realized that there were numerous propositions which implied the fifth postulate of Euclid. Gauss wrote: *"I achieved much of what would be taken for a proof by the majority of people."* And then: *"However, the road I have chosen leads to doubt about the truth of geometry rather than to the desired destination."* There was only one step from this point to the realization of the possibility of constructing a non-Euclidean geometry. However, this step was not taken, although this phrase is often taken as evidence that Gauss had come to non-Euclidean geometry as early as 1799.

The following words of Gauss must be taken into account. He wrote that he could not devote enough time to this problem. It should be noted that the problem of parallels was not mentioned in his diary; evidently, it was never the focus of Gauss' interests. In 1804, Gauss refuted the attempts of Bolyai to prove the parallel postulate. His letter ended in the following words: *"However, I still hope that sometime, even before my death, these obstacles will be overcome."* These words give evidence that Gauss hoped that a proof would be found.

Here is some more evidence. *"In the theory of parallels we haven't yet left Euclid behind. This is a disgraceful part of mathematics that must take quite a different form sooner or later"* (1813). *"We haven't gotten beyond the point reached by Euclid 2000 years ago"* (1816). However, in the same year of 1816, Gauss wrote about *"a gap which cannot be filled,"* and in 1817, he wrote in a letter to Olbers: *"My conviction that the "necessity" of our geometry cannot be proved, at least by the human mind and for the human mind, becomes increasingly strong. It may happen that in another life we would arrive at a different view on the nature of space, but now this view is beyond our understanding. Until that time, geometry must be ranked with mechanics rather than with arithmetic, which is purely an* a priori *science."*

About the same time, similar thoughts about the impossibility of proving the fifth postulate were expressed by Schweikart, a lawyer from Königsberg. He assumed that an "astral geometry" existed along with the Euclidean geometry, in which the fifth postulate didn't hold. Herling, a student of Gauss who worked in Königsberg, wrote to Gauss about Schweikart and sent his note. In his reply, Gauss said: *"Almost everything is copied from my heart."* The work of Schweikart was continued by his nephew, Taurinus, with whom Gauss exchanged several letters beginning in 1824.

In his letters, Gauss stressed that his opinions were private and must not be made public in any case. He didn't believe that his ideas could be understood and was afraid of attracting the attention of a crowd of laymen. Gauss experienced tough years and he highly appreciated the opportunity of working calmly. He warned Herling, who was only going to mention that the parallel postulate may be wrong: *"Wasps whose nest you are going to destroy will rise around your head."* Gradually, Gauss arrived at the conclusion that he must write down his results, but not publish them. He wrote: *"It seems that it will take a long time to prepare my extensive investigations of this question for publication. It is quite likely that I won't resolve upon publishing them at all, because I am afraid of the shouts of Boeotians*[1] *which they will raise if I make all my opinions public"* (a letter to Bessel, 1829). In May, 1831, Gauss began a systematic description of his studies: *"It has been several weeks that I have been committing to paper the results of my work on this subject. This work was partly done 40 years ago, but it has never been put on paper; thus, I had to reconstruct the entire reasoning three or four times. However, I wouldn't like it if my work died with me."* (A letter to Schumacher)

However, in 1832, Gauss received from Farkas Bolyai a small paper by his son Janos called "Appendix" (this title was explained by the fact that the paper was published as an appendix to a large book by his father). *"My son appreciates your opinion more than the opinion of the whole of Europe."* The contents of the paper amazed Gauss: it presented in a complete and systematic form the construction of a non-Euclidean geometry. This paper was quite different from fragmented remarks and conjectures of Schweikart and Taurinus. Gauss himself was going to give such an outline of the theory in the near future. He wrote to Herling: *"I found all my ideas and results presented very elegantly, though they are exposited in a very condensed form, which makes them difficult to understand for those who don't work in this field... I think that this young geometer Bolyai is a genius."* Here is what he wrote to the father: *"... the entire contents of this work, the way your son went about it and the results he obtained are almost the same as my own, which I partially*

[1]Traditionally in ancient Greece, the citizen of Boeotia were known for their weak intellectual ability.

obtained 30 or 35 years ago. I am really very surprised. I have committed to paper some of my results but did not intend to publish them during my lifetime; I just didn't want these results to perish after my death. For this reason I am staggered by what has happened—it frees me from this work. I am also very glad that it was the son of my old friend who anticipated my work." However, Janos Bolyai didn't receive any public support from Gauss. It seems that simultaneously Gauss stopped his regular notes on non-Euclidean geometry, although fragmented notes of the middle 1840s do exist.

In 1841, Gauss got to know the German edition of Lobachevsky's work (his first publications date to 1829). As usual, Gauss was interested in other works of this author, but expressed his opinion only in correspondence with close friends. However, in 1842, Gauss recommended that Lobachevsky be elected a correspondent member of the Royal Scientific Society of Göttingen as "one of the best mathematicians of the Russian state." Gauss himself informed Lobachevsky of his election, but neither in the recommendation, nor in the diploma that was given to Lobachevsky, was there a mention of non-Euclidean geometry.

Gauss' works on non-Euclidean geometry became known only when his archive was published after his death. Thus, by refusing to publish his great discovery, Gauss provided himself with an opportunity to work calmly. This fact raised an unremitting discussion about whether such a position is acceptable or not.

It must be noted that Gauss was interested not only in the purely logical question of the possibility of proving the fifth postulate. He also tried to establish the place of geometry in the natural sciences and the question of the true geometry of the physical world (see his opinion of 1817 above). He discussed an opportunity for an astronomic verification and was interested in Lobachevsky's ideas concerning this issue. In his geodetic work, Gauss couldn't resist the temptation of measuring the sum of the angles of the triangle formed by the summits of three mountains: Hohenhagen, Brocken, and Inselberg. The deviation from 2π didn't exceed 0.2'.

Electrodynamics and geomagnetism

By the end of the 1820s, Gauss, who was already over 50, began to seek new fields for his research. This is evidenced by two articles of 1829 and 1830. The first of them was devoted to general principles of mechanics (Gauss' principle of least constraint); the other was concerned with capillary phenomena. Gauss had decided to study physics, but his particular interests had not been decided yet. In 1831, he tried to do research in crystallography. This was a very difficult year for Gauss: his second wife died, and he developed bad insomnia. The same year, Wilhelm Weber, the 27-year old physicist who had been invited at Gauss' initiative, came to Göttingen. Gauss made his acquaintance in 1828 at Humboldt's home. Although Gauss, then 54, was a very reserved man, he nevertheless found in Weber a companion such as he had never before met for doing research.

Klein wrote that *"the internal difference of these two people manifested itself even in their appearance. Gauss was a thickset man, a true representative of Lower Saxony, taciturn and reserved. Small, graceful, and lively Weber was his antithesis. He was very talkative and amiable, which betrayed a native Saxon. Indeed, he came from Wittenberg, this country of 'quintessential Saxons.' The Göttingen monument to Gauss and Weber makes this difference less noticeable; even in height they seem closer than in real life."*

Gauss and Weber were interested in electrodynamics and geomagnetism. Their activity wasn't purely theoretical; it also brought some practical results. In 1833, they invented electromagnetic telegraphy (this event is memorialized in the monument). The first telegraph connected the observatory and the physics institute. However, they couldn't introduce the device for large-scale use because of financial reasons.

While studying magnetism, Gauss arrived at the conclusion that systems of physical units must be constructed on the basis of several independent units, and all other units must be expressed in terms of the basic ones.

The study of terrestrial magnetism was based both on observations made at the magnetic observatory in Göttingen and on data collected in various countries by the *Union for Observing Terrestrial Magnetism,* which was founded by Humboldt after his return from South America. At the same time, Gauss created one of the most important divisions of mathematical physics—potential theory.

The common work of Gauss and Weber was interrupted in 1843, when Weber was expelled from Göttingen with six other professors for signing a letter to the King, which pointed to violations of the constitution by the King (Gauss didn't sign this letter). Weber returned to Göttingen only in 1849, when Gauss was already 72.

* * *

We conclude our story with the following words of Klein: *"Gauss reminds me of the highest summit of the Bavarian ridge as it stands before the eyes of an observer looking from the North. In this mountain range, isolated summits become higher and higher from East to West, reaching the maximum height in a mighty central giant. Breaking steeply, it is replaced by a lowland of a new form with spurs spreading deeply into it, and streams that flow down from this giant bring moisture and life."*

The eye and the sky

The art of seeing faint objects

by V. Surdin

A HUMAN BEING HAS EIGHT senses: sight, hearing, smell, taste, touch, hot and cold, pain, and a sense of the position and motion of the body. Undoubtedly, the most important and wonderful sense is sight, which provides the brain with more than 90% of the information it receives about the surrounding world. For astronomers vision is an indispensable tool for studying the Universe. When making observations with the naked eye or with a telescope, one should take into consideration the specific features of the construction and visual performance of the eye. Only with such knowledge can one discern such almost invisible objects as small lunar craters, faint stars, remote comets, and galaxies.

The first encounter with a telescope is often disappointing to the amateur astronomer. "I can't make out any details," says the novice first looking at Mars or even Jupiter. In contrast, the experienced observer can map the details of the planet with the same telescope. The reason is that the astronomers have trained their eyes (as have, say, biologists, who discern the tiny details of the living cells under a microscope). You can also master this skill by learning the special features of our vision and by observing the heavenly bodies on a regular basis.

How the eye is constructed

The construction of the eye is shown in figure 1. At first glance, it looks very similar to a photographic camera. The objective of the eye (the cornea, iris, and crystalline lens) is similar to a classical objective, with its system of lenses and a diaphragm (iris). This natural objective is exceptionally wide-angle: the visual field of the human eye extends to almost 180° in the horizontal and 140° in the vertical. Of course, the quality of such a simple objective cannot be very high: a sharp image is projected only on the center of the visual field, while the image is severely degraded at the periphery.

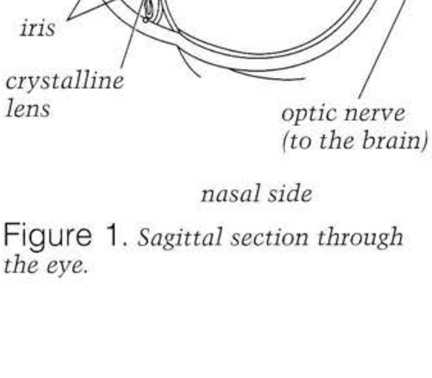


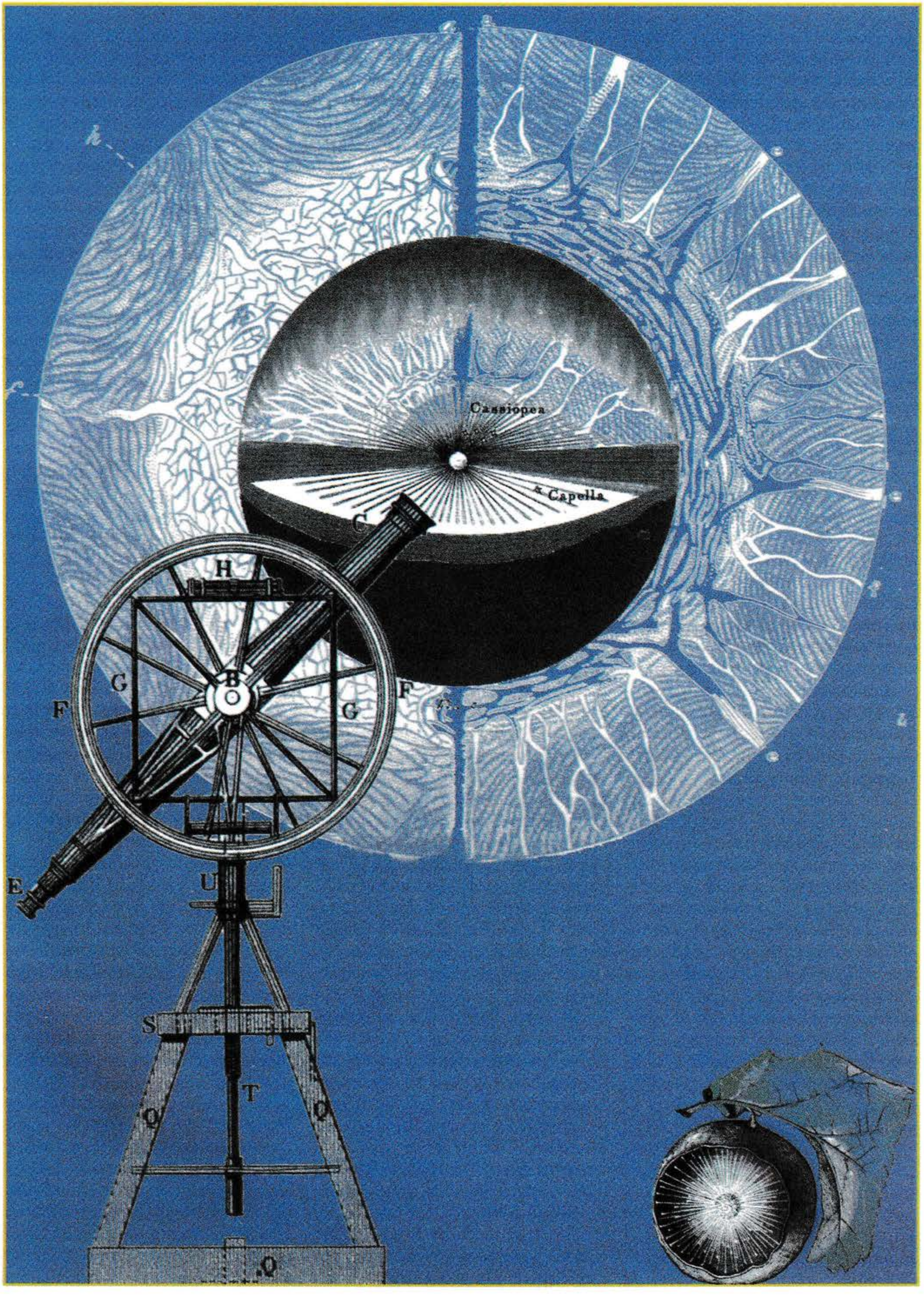
Figure 1. *Sagittal section through the eye.*

The central opening of the iris is called the pupil. It plays the same role as the aperture of a camera: it covers the edge of the lens to a greater or lesser degree. The eye can adjust the pupil diameter over a range from 2 to 8 mm. In this way it controls the luminous flux passing through the crystalline lens to the light-sensitive retina. This is where all simple analogies with a camera come to an end, because the retina has nothing in common with photographic film.

In design, the retina is far closer to the photodetector of a modern video camera, which consists of an array of elementary semiconductor photodetectors, called pixels (from "picture elements"). In the eye their role is played by light-sensitive cells of two types: rods and cones. These cells differ not only in shape, but in function. The retina of a human eye contains about 7 million cones and 120 million rods.

The rods work at very weak illumination and provide twilight and

Art by Vera Khlebnikova


Cassiopea
Capella
C
H
B
G
G
F
F
E
U
S
T
Q
Q
Q

night vision. However, they do not distinguish color (thus the saying, "all cats are gray at night"). The cones, on the other hand, can work only in broad daylight but are sensitive to color. This property is explained by the fact that there are three types of cones in a jumbled arrangement. Some of them are sensitive to blue, some to green, and others to red-orange light. Working together, they can convey precise information about the hues in the image. This works something like the picture tube of a color TV set, which incorporates three monochromatic kinescopes in one tube: red, green, and blue.

Resolving power of the eye

The size of the smallest details that can be distinguished by the eye is determined by its angular resolving power. When we cannot distinguish the tiny details in a picture, we try to bring it nearer to the eyes. The linear size does not change during this procedure, but the angular size α increases (figure 2). The smallest angle at which the eye can distinguish two points is called the (angular) resolving power or simply the resolution. In the center of the visual field (that is, at the place where the surface density of cones is highest) the normal eye has an angular resolution of about 1′ (one minute of arc). The period at the end of this sentence subtends an angle of this size if the journal is held at arm's length. Note that when viewing at arm's length, you can easily distinguish a comma from a period. However, if you place the page three times farther away, you cannot tell a comma from a period. Try it with the following string of commas and periods: (,.,..,,..,.,,.,).

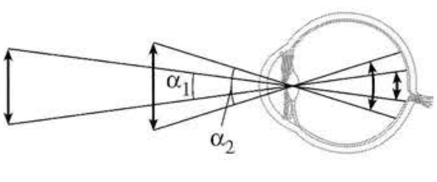


Figure 2. *The angular size of an object and the sharpness of its image depend on distance.*

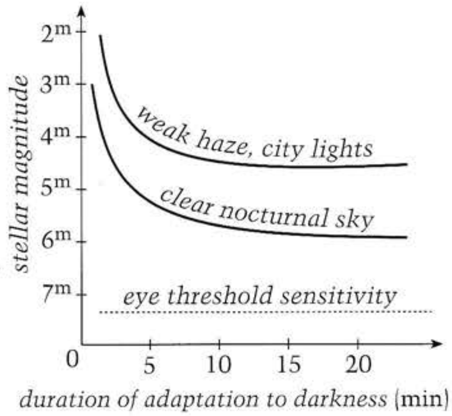


Figure 3. *Stellar magnitudes of the faintest stars that can be seen by the human eye, as a function of the duration of adaptation to darkness. The threshold sensitivity of the eye cannot be attained in an astronomical observation, because the night sky is not absolutely dark on account of illumination by atmospheric radiation, space dust, and even by distant stars.*

There is another wonderful feature of our vision—the ability to distinguish rapid movements and changes in an image. This is referred to as time resolution. For example, in some music video clips the picture changes at a rate of 5–7 frames per second. At this rate we can see each individual picture. However, when the rate of change is increased to 15–20 frames per second, the different pictures merge into one, and the individual frames cannot be discerned. The reason is that the time resolution of the human eye is only about 0.1 sec. This physiological feature accounts for the cinema effect: when the frames are changed at a rate of 24 per second, we do not see the frames flashing but perceive them as a smooth, uninterrupted movie.

The eye can perceive differences in illumination of adjacent objects only if they are greater than 2%, independent of the absolute brightness of the objects. In this process the eye operates as a comparator that evaluates not the difference but the ratio of the luminosities. This is why astronomers use the concept of stellar magnitude. When two stars differ in stellar magnitude by 1^m, it means that the luminous fluxes from them differ by a factor of 2.512... An experienced astronomer can discern differences in the brightness of adjacent stars as small as 0.02^m. This is almost exactly equal to the value of 2% that characterizes the threshold of sensitivity of the human eye.

Let's continue the analogy with electronic devices. A video camera has a system that controls its sensitivity (automatic gain control). The eye also has a similar system. In darkness its sensitivity increases by thousands of times, while in broad daylight it decreases accordingly. However, these changes do not occur as rapidly as in a video camera. Therefore, when we enter a brightly illuminated room, our eyes are dazzled for several seconds by the light, and when we enter a dark room it takes a while before we can see anything. The complete adaptation of the eyes to darkness takes about half an hour. After this adaptation period the sensitivity of our eyes is maximal (figure 3).

Some peculiarities of vision

The eye is connected with the brain by a bundle of nerve fibers, but the number of fibers is far less than the number of light-sensitive cells in the retina. This means that the eye does some preliminary processing of the visual information and sends a simplified image to the brain. For example, it tries to "recognize" the contours of geometrical figures familiar to the brain in what at first glance may look like chaotically scattered spots. Sometimes this property helps astronomers (it is easier to remember the outlines of constellations), although in some cases this feature of visual perception is misleading. An example is the "discovery" of canals on Mars or the apparent observation of strands or rings of stars.

We should remember that perception of visual information is a subtle process intimately connected to specific features of human psychology. Knowing these features, one can easily deceive those parts of the brain that analyze the two-dimensional image fixed by the retina and recreate the original three-dimensional picture. Specialists know many of the "optical illusions" exploited in the tricks of the stage magician. No less important is the reverse skill—to represent the 3D structure of a real object in a 2D drawing. An extreme example of the manipulation of visual perception is to draw "impossible" objects. The Dutch printmaker M. C. Escher was a master of this.

The skill of transforming a 2D image into a 3D model of real life is what children learn to do from the first days of life by touching objects that they see. In this sense, vision is sometimes described as an acquired skill. When translating the 2D images into 3D ones, our brains use a special "dictionary" of possible models stored in memory. This dictionary is highly individual and is compiled in accordance with specific life experiences, including one's profession. Indeed, every person will complete the Droodles, those fragmentary sketches that pose riddles, in a unique way. And having seen something unusual in the sky, each of us will fill in the details in our own way and will reach our own conclusions as to the nature of the event. Only experience in astronomical observations will make us more or less confident in our interpretation of what we see in the firmament.

The flying Moon

On windy and cloudy moonlit nights you have no doubt witnessed a surprising effect—the "flight" of the Moon. When the Moon is seen through gaps in the clouds, its apparent sideways motion will catch your eye, and you will involuntarily turn your head to follow its rapid movement. After a few seconds you realize that the Moon isn't moving, it is the clouds that are moving across the sky. If the cloud cover is dense and the gaps far between, you may not see the Moon again for a long time. Then you may be left with the certainty that you saw some bright, round object flying rapidly across the clouds. There are cases of UFO reports based on this phenomenon. How do physiologists explain the flying moon effect?

As we have said, the eye does not send all of the visual information to the brain, but processes it to select the most important part. How does the eye know what is important and what can be ignored? In the course of biological evolution, every human organ has acquired traits which are conducive to survival. The eye is no exception. It has acquired the ability to select that part of the visual information that deals with potential prey or dangerous predators. Both predator and prey are usually moving—attacking or running away. No wonder that our visual system learned to detect motion as the most important thing. This system is very efficient in detecting even slight changes in the visual field.

The visual signal undergoes complex transformations on its way from the retina to the cerebral cortex. On the whole, it is greatly simplified and carries only the data of vital importance, the lack of which could mean death for a human being. Information on displacement relative to the retina is not only preserved, it is augmented, while other data are discarded. Immobile objects, whose images do not move on the retina, are not even noticed by the eye after a while. Therefore, in order to maintain a picture of the immobile background, the eye must continually perform microscopic movements of 2–3 minutes of arc, thereby producing artificial displacements of the object on the retina. This is the reason that when looking at stars we sometimes see them jump—in reality, these jumps result from microscopic motions of the eyes.

Another important feature of the eyes originates from the nonuniform distribution of the light-sensitive cells in the retina. The region of sharpest vision is located in the center of the retina and has an angular size of only a few degrees, while the resolving power of other parts of our visual field is low.

Now imagine that you are a hunter looking over the motionless landscape, hoping to spot a prey. The clever prey may sense the danger and react in a very effective way: it simply "freezes," making it harder to spot.

Consider another possibility: the prey doesn't see the hunter and continues to graze calmly. There are two modes of action for the eye: to fix the view on the landscape or on the moving prey. If the eyes are fixed on the landscape, the image of the prey is the only irritating stimulus moving on the retina. In this case the corresponding visual signal is weak, because it originates from only a small part of the retina. In addition, the quality of the image is low, since it quickly leaves the area of sharpest vision on the retina.

Evolution has chosen the other way. Our eyes are fixed on the moving prey; its image is locked persistently at the center of the retina, in the area of sharpest vision. In this mode of vision, the landscape is moving on the retina, producing a powerful visual volley of nerve impulses to the brain and stimulating attention. This is a standard situation, and the brain knows how to interpret it: in reality, the small fixed image at the center of the visual field is a moving object, while the surrounding moving image is the motionless background landscape.

Now what happens when we look at the night sky and see the motionless Moon (or a bright star or a planet) occasionally peeking through gaps in a bank of fast-moving clouds? The eye (or, more precisely, the brain) responds to this picture in the usual way—it interprets the clouds as the stationary background and the Moon as moving. It takes some time and conscious effort to suppress the illusion, to make the Moon stop and the clouds move. This is no problem for those who are accustomed to viewing astronomical phenomena.

Interestingly, this same manner of detecting the moving object is the working principle of electronic image stabilization (EIS) systems in portable video cameras. If the picture is shifted as a whole, the EIS system returns it to the frame boundaries on the assumption that the shift was caused by a jerk of the operator's hand. In contrast, if only a small part of the picture is moving, the EIS system doesn't interfere, on the assumption that this shift represents the motion of an individual object. Therefore, when a video camera is aimed at a cloudy sky with the Moon behind, it will try to stop the clouds and let the Moon move—just as our brain does.

Performance of peripheral vision

To best see an object, we look at it directly. The image is projected onto the center of the retina, where the dense array of cones provides chromatic and very sharp vision. Everything that lies outside the central part of the retina is blurred and deprived of bright colors. However, the peripheral vision has its own virtues.

First, there are almost none of the very light-sensitive rods in the center of the retina. Recall that it is the rods that provide twilight vision. Although there are plenty of cones in the central retina, their sensitivity is lower than that of the rods, because their spectral range is limited to a certain color. In addition, the cone-filled central part of the retina is not large: its angular size is only a few degrees. In contrast, the periphery of the retina contains many rods, which are very sensitive to light but cannot distinguish colors. Therefore at night our peripheral vision will detect dim objects that can't be seen when looked at directly.

The peripheral vision is often used by astronomers to view faint stars or nebulas that are too dim to be seen when looked at directly. You may try this on your own. Use an astronomical chart to determine the location of the Andromeda Nebula (figure 4). Most probably you will not see it at the location indicated (especially in a city sky). However, if you shift your eyes slightly away from the target, you will see an oval spot resembling the flame of a

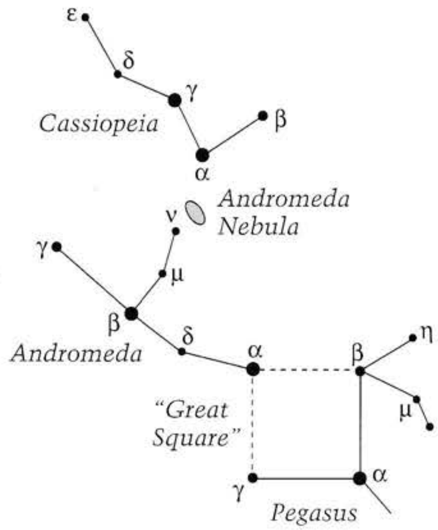


Figure 4. *Find the Andromeda Nebula with the help of your peripheral vision. Fix your eyes on the star Andromedae or slightly to the left of it.*

candle—this is the famous Andromeda Nebula, a giant stellar system similar to our Milky Way Galaxy. It is not surprising that it is difficult to discern with the naked eye. Indeed, the wonder is that we can see it at all (if only with the help of our peripheral vision), since this galaxy is hundreds of times farther away than the stars making up the familiar constellations.

Thus peripheral vision is a valuable tool for a stargazer. If you learn the skill of using it, it will serve you well. However, this physiological phenomenon may play a trick on the uninitiated. Catching sight of a bright object at the periphery of your visual field, you turn your head to examine it carefully, but you do not see anything. Sometimes such events produce legends about UFOs.

There is another important feature of peripheral vision: it has better time resolution than central vision. Test it on your own: look at a fluorescent lamp first directly, then with your peripheral vision. In the first case you will see a steady light, but in the second case you can see the flickering that results from the alternating nature of the power supply to the lamp (50 or 60 Hz), which produces flickering at twice the frequency. A similar experiment can be carried out with a TV screen: when you look at it directly, you do not see the rapid changes of the frame. However, you can easily observe the flashing of the frames with your peripheral vision.

Presumably, this feature of our vision was developed in the evolutionary period when people were hunted by many dangerous predators which attacked from behind or from the side. It is therefore important that the peripheral vision of a prey should work very quickly and report danger to the brain without delay. It is not important that the sharpness of the peripheral vision is poor. It makes no difference for the prey what beast is attacking—a striped tiger or a spotted leopard. It is of vital importance only to detect the attack quickly and get away fast. In modern times the quick reaction provided by peripheral vision is important for drivers and pedestrians. Therefore, train your peripheral vision and keep in mind its distinctive characteristics when observing various optical phenomena—both celestial and terrestrial. ◙

Quantum on vision and optical illusions:

B. Fabrikant, "Through the Glass Brightly," September/October 1990, pp. 34–38.

M. Berkinblit and E. Glagoleva, "Mathematics in Living Organisms," November/December 1992, pp. 34–39.

V. M. Bolotovsky, "What's That You See," March/April 1993, pp. 5–8.

P. Bliokh, "What Little Stars Do and the Big Old Planets Don't," March/April 1994, pp. 22–27.

V. Surdin, "Optics for a Stargazer," September/October 1994, pp. 18–21.

D. Tarasov and L. Tarasov, "The Play of Light," May/June 1996, pp. 10–13.

A. Eisenkraft and L. D. Kirkpatrick, "Color Creation," May/June 1997, pp. 36–39; November/December 1997, pp. 32–33.

A. Mitrofanov, "Can You See the Magnetic Field," July/August 1997, pp. 18–22.

V. Novoseltzev, "Visionary Science," May/June 1998, pp. 21–25.

HOW DO YOU FIGURE?

Challenges

Math

M281

Take a walk. Professor Zlata and her assistant Bruce live not far from each other and like to stroll in the evening from one house to the other and back, traveling the same route several times. One day they each left home at the same time. The first time they encountered each other was at a distance of 55 m from the professor's house, and the second time, at a distance of 85 m from the assistant's house. There is a newsstand located 25 m from the assistant's house, and there is an ice cream stand near the professor's house. It is known that after each had left home, the professor and the assistant passed the nearer stand at exactly the same time. What is the distance between the two stands?

M282

X-ponent. Solve the equation

$$x^x = \frac{1}{\sqrt{2}}.$$

M283

Don't forget your calculus. Prove that the inequality $\sin x + \arcsin x > 2x$ holds for all values of x such that $0 < x \le 1$.

M284

Locus-Pocus. A chord AB is drawn in circle O, of radius r. Points P and Q are taken on its extension beyond points A and B, respectively, such that $AP = BQ$. As P and Q vary along line AB, they determine two pairs of tangents to circle O. These four tangents, in turn, determine four new points of intersection with each other. Find the locus of all such points of intersection.

M285

At an intersection. Side BC of triangle ABC has length a, and the opposite angle has degree-measure α. The line passing through the midpoint D of BC and the center of the circle inscribed in the triangle intersects AB and AC at points M and P, respectively. Find the area of the (nonconvex) quadrilateral $BMPC$.

Physics

P281

Trapped ball. A small, heavy ball flies through a hole into the smooth interior of a sphere of the same mass. The ball passes along a line lying a distance $R/2$ from the center of the sphere, where R is the radius of the sphere. After the ball is trapped, the hole is closed automatically. Assuming that the collisions of the ball with the wall of the sphere are perfectly elastic, find the trajectories of the ball and the center of the sphere in the reference frame in which the sphere was initially at rest. Determine the parameters of these trajectories and mark the points on them where collisions occur. (B.Bukhovtsev)

P282

Just a bubble. The outer radius of a soap bubble is R, and thickness of its wall is h. What is the air pressure inside the bubble? What is the pressure inside the soap film? Consider the film to be thin ($h \ll R$). The pressure of the air surrounding the bubble is P_0. The surface tension of the soap solution is σ.

P283

What's passing? Two capacitors with capacitances of 2 F and 3 F are connected in series and hooked up to a battery with emf $\mathcal{E}$ = 120 V, whose midpoint is grounded (fig. 1). The wire connecting the capacitors can also be grounded with a switch S. Find the charges q_1, q_2, and q_3 that

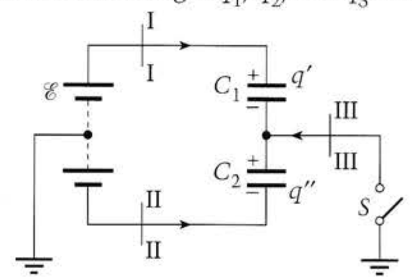


Figure 1

pass in the indicated directions through the cross sections I–I, II–II, and III–III, respectively, after the switch is closed. (I. Slobodetskii)

P284

All aglow. Two identical coils each with inductance L are wound on a circular (toroidal) ferromagnetic core having a very large magnetic permeability. They are connected

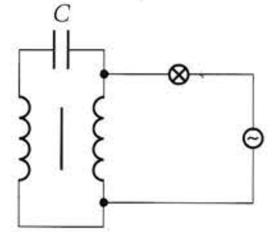


Figure 2

in series with a capacitor of capacitance C (fig. 2). You examine the circuit using a test lamp and a vari-

CONTINUED ON PAGE 27

Continued fractions

When close enough is good enough

by Y. Nesterenko and E. Nikishin

SOMETIMES A FIELD OF RESEARCH that seems promising and important falls short of expectations; the number of researchers working in the field declines, and it loses its importance. One such area of mathematical research is the theory of continued fractions. Until the end of the nineteenth century, continued fractions of various kinds often appeared in mathematical works. Many important theorems on continued fractions were proved in the nineteenth century and a little earlier. There was hope that a complete understanding of the structure of continued fractions could lead to new results in number theory and mathematical analysis. These expectations were only partially fulfilled. Later the development of powerful new methods, especially in number theory, caused the study of continued fractions to be largely abandoned.

However, just as the North Pole always attracts some visitors, problems involving continued fractions, which often have a simple formulation and seem easy, entice mathematicians to explore the nature of these strange objects. Many important problems of modern mathematics and physics lead to objects similar to continued fractions. For this reason, methods developed for continued fractions prove to be useful for many other problems. In this article, we consider only problems related to number theory, where the theory of continued fractions originated.

Calendar, toothed wheels, and a piece of history

How many days are in a year? Everybody knows that a common year has 365 days and a leap year has 366. Leap years are those divisible by 4, for example, 1904, 1908, ..., 1980, 1984, ... 1996. However, 1800, 1900, 2100, and 2200 are not leap years, while 2000 and 2400 are. Why is this?

The explanation is rather simple. The Earth rotates uniformly about its axis, making one revolution per day. It takes Earth 365.24219878... days to make one revolution around the Sun. This period is called the year. The extra 0.242... days may seem small. However, if we set the year to 365 days, the error will accumulate. In ancient times, when the length of the year was known only approximately, the accumulated error could be rather large. For example, in 46 BC in ancient Rome, the lag became as much as 90 days.

To find a law for the alternation of common and leap years, we must choose the length of the cycle q (after which the sequence of common and leap years repeats) and the number of leap years in the cycle, p. Let us write the length of the year as $365 + \alpha$, where $\alpha = 0.242...$. Positive integers p and q must be chosen so that the quantity

$$\beta = q\alpha - p$$

is as small as possible without having p and q be very large. Actually, only q needs to be chosen, and then p will be the integer closest to $q\alpha$. Then $365q + p$ days will pass in q years. It follows from the equation $365q + p = q(365 + \alpha) - \beta$ that in $365q + p$ days the Earth will make

$$q - \frac{\beta}{365+\alpha} \cong q$$

revolutions around the Sun. A one-day error accumulates only in $1/\beta$ such cycles or in q/β years. At present, we use the Gregorian calendar, in which $q = 400$. Of these 400 years, 303 are common and 97 are leap years. Leap years are those divisible by 4 with the exception of centennial years which are not leap years unless exactly divisible by 400. Thus, the average length of the year according to the Gregorian calendar

Art by Sergey Ivanov


q
p
S. Ivanov

is 365.242500. Such an approximation is quite satisfactory: a one-day error accumulates only in about 3300 years. If we took $q = 128$ and $p = 31$, an even better approximation could be obtained. However, a calendar with 128-year cycles would be less convenient.

Here we face an important mathematical problem: *given a number α, find sufficiently small integers p and q such that the number*

$$b = q\alpha - p$$

is as small as possible.

A similar problem arises in designing gears. For transferring the rotational motion from one gear wheel to another, one must cut q teeth on the first wheel and p teeth on the second, so that the ratio p/q is as close as possible to the given number ω (ω is the desired ratio of rotational velocities). It is clear that for reasons of economy, p and q should be chosen as small as possible.

There are many other examples that lead to the problem of the best approximation of real numbers by a rational. Examples include the musical scale, applications in computational mathematics, and theoretical problems of celestial mechanics.

Continued fractions provide a tool for finding the best (in a certain sense) rational approximations. Continued fractions have been applied for computational purposes for a very long time. As early as in 1572, the Italian mathematician and engineer R. Bombelli (1526?–1572) used them to compute $\sqrt{13}$. Later, the Englishman W. Brouncker (1620–1684) used them to refine the value of π. The prominent physicist, astronomer, and mathematician C. Huygens (the inventor of the pendulum clock) was the first to describe the sense in which continued fractions give the best approximation of real numbers. The great Euler (1707–1783) proved certain theorems concerning continued fractions and found a continued fraction for the number e (the base of natural logarithms). After Euler, numerous mathematicians made contributions to the theory of continued fractions, so that it is difficult to list them all. The works of the prominent Russian mathematician P.L. Chebyshev (1821–1894) initiated the active development of a theory dealing with functions defined by continued fractions.

Euclid's algorithm and the continued fraction expansion of a number

Let p and q be two positive integers. Performing a succession of divisions with remainder, we have:

$$\begin{aligned}
&p = a_0 q + q_1,\ 0 < q_1 < q,\\
&q = a_1 q_1 + q_2,\ 0 < q_2 < q_1,\\
&q_1 = a_2 q_2 + q_3,\ 0 < q_3 < q_2,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&q_{k-2} = a_{k-1} q_{k-1} + q_k,\ 0 < q_k < q_{k-1},\\
&q_{k-1} = a_k q_k.
\end{aligned}$$

Since the sequence $q_1, q_2, \ldots, q_k$ is a strictly decreasing sequence of nonnegative integers, one of them must eventually be 0 (in fact, it would be q_{k+1} in the notation used above), and so the sequence of equations above must be finite. This process is known as *Euclid's algorithm*. It can be proved that the number q_k obtained is the greatest common divisor of p and q. We won't use this fact in this article, and thus we omit its proof.

The above relations give

$$\frac{p}{q} = a_0 + \frac{q_1}{q} = a_0 + \cfrac{1}{a_1 + \cfrac{q_2}{q_1}} = \ldots$$

$$= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{a_k}}}.$$

The following process yields the same result. Let α be a real number. Set

$$\alpha = a_0 + \frac{1}{\alpha_1},$$

where a_0 is an integer and $\alpha_1 > 1$. If α_1 is not an integer, we continue the process by representing α_1 as $\alpha_1 = a_1 + 1/\alpha_2$, where a_1 is a positive integer and $\alpha_2 > 1$. Thus,

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\alpha_2}}.$$

If α_2 is not an integer, the process can be continued. Thus at the kth step we have the relation

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{\alpha_k}}} \qquad (1')$$

The "multi-story" notation (1′) of the fraction is very inconvenient, so we will use the following, more compact notation:

$$\alpha = a_0 + \frac{1|}{|a_1} + \frac{1|}{|a_2} + \ldots + \frac{1|}{|a_{k-1}} + \frac{1|}{|\alpha_k}, \quad (1)$$

where $\alpha_k > 1$.

It is easy to prove that if $\alpha = p/q$ is a rational number, then for some k, α_k will be an integer. It is also clear that if, for some k, α_k is an integer, then α is rational. However, if α is irrational, this process never terminates, and an infinite *continued fraction* is obtained:

$$\alpha = a_0 + \frac{1|}{|a_1} + \frac{1|}{|a_2} + \ldots + \frac{1|}{|a_{k-1}} + \frac{1|}{|\alpha_k} + \ldots \quad (2)$$

The equal sign here is provisional, since it is not clear what the expression on the right means. To give it meaning, consider finite sums of the form

$$\pi_k = a_0 + \cfrac{1}{a_1 + \cdots + \cfrac{1}{a_k}},$$

which are called *convergents* of the continued fraction (2). We define the *right-hand side* of (2) *as the limit of the convergents*:

$$\lim_{k\to\infty} \pi_k.$$

The following set of problems presents the properties that underlie the elementary theory of continued fractions; in particular, equality (2) is established in them. (Note: You won't need to know calculus to do most of these problems. Proof by mathematical induction is sufficient.)

Basic properties of continued fractions

1. Let $\pi_k = p_k/q_k$ be a convergent of the continued fraction (2). Prove the following recurrence relations:

$p_0 = a_0,\ p_1 = a_0 a_1 + 1,\ \ldots,\ p_{k+1}$

$$= a_{k+1}p_k + p_{k-1}, \dots$$
$$q_0 = 1,\ q_1 = a_1, \dots,\ q_{k+1} = a_{k+1}q_k + q_{k-1}, \dots$$
for $k = 1, 2, 3, \dots$

Hint. Use mathematical induction.

2. Prove the following relations:

(i) $q_n p_{n-1} - p_n q_{n-1} = (-1)^n,\ n \ge 1;$

(ii) $q_k p_{k-2} - p_k q_{k-2} = (-1)^{k-1},\ k \ge 2;$

(iii) $\pi_{n-1} - \pi_n = \dfrac{(-1)^n}{q_n q_{n-1}},\ n \ge 1;$

(iv) $\pi_{k-2} - \pi_k = \dfrac{(-1)^{k-1} a_k}{q_k q_{k-2}},\ k \ge 2.$

Hint. Use prob. 1 and mathematical induction to prove (i) and (ii) and use (i) and (ii) to prove (iii) and (iv).

3. Prove that the convergents $\pi_0, \pi_2, \pi_4, \dots$ are monotonically increasing:

$$\pi_0 < \pi_2 < \pi_4 < \pi_6 < \dots,$$

and that the convergents $\pi_1, \pi_3, \pi_5, \dots$ are monotonically decreasing:

$$\pi_1 > \pi_3 > \pi_5 > \dots,$$

Hint. See (iii) and (iv) above.

4. Let α_k the numbers be defined by (1). Prove the relation

$$\alpha = \frac{p_{k-1}\alpha_k + p_{k-2}}{q_{k-1}\alpha_k + q_{k-2}},\ k \ge 2.$$

Hint. Use mathematical induction.

5. Prove the inequality

$$\frac{1}{2q_{n+1}} < |q_n\alpha - p_n| \le \frac{1}{q_{n+1}}.$$

Hint. Use prob. 4 with $R = n$, and prob. 1.

6. Prove that

$$\lim_{k\to\infty} \pi_k = \alpha.$$

Hint. In fact, you have only to show that q_n grows to infinity. Then the result of prob. 5 is equivalent to the definition of limit.

We can see from prob. 1–6 that the convergents π_n give a rather good approximation of α. Thus, it follows from the estimates in prob. 5 and the inequality $q_{n+1} > q_n$ that if α is irrational, then the inequality

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}$$

Table 1

a		a_0	a_1	a_2	...	a_{n-1}	a_k	a_{k+1}	a_{k+1}	...
p	1	a_0	p_1	p_2	...	p_{k-1}	p_k	...		...
q	0	1	q_1	q_2	...	q_{k-1}	q_k	...		...

holds for all convergents $\pi_n = p_k/q_k$ of the continued fraction for α.

Numerical examples

The continued fraction (2) for α is sometimes written as $[a_0; a_1, a_2, \dots, a_k, \dots]$. It can be proved that, for any sequence of integers $a_0, a_1, a_2, \dots$ ($a_0 \ge 0$ and $a_j \ge 1$ for $j \ge 1$), there exists a unique positive number α such that

$$\alpha = [a_0; a_1, a_2, \dots].$$

If the sequence $a_0, a_1, a_2, \dots$ is known, then the sequence of convergents $\pi_k = p_k/q_k$ can be easily calculated by using the relations obtained in prob. 1 to fill in the table 1.

Example 1. Find the continued fraction expansion of $\sqrt{2}$. We have

$$a_0 = 1,$$
$$\alpha_1 = \frac{1}{\sqrt{2}-1} = \sqrt{2}+1,\ a_1 = 2,$$
$$\alpha_2 = \frac{1}{\sqrt{2}-1} = \sqrt{2}+1,\ a_2 = 2.$$

It is clear that in the sequence α_j, $j = 1, 2, \dots$, all elements are equal to $\sqrt{2} + 1$, and in the sequence a_j, $j = 1, 2, \dots$, all elements are equal to 2. Thus,

$$\sqrt{2} = [1; 2, 2, 2, \dots].$$

Filling in table 2, we calculate a few first convergents for $\sqrt{2}$.

Thus, we obtain a sequence of convergents 1/1, 3/2, 7/5, 17/12, 41/29, 99/70, ...

It follows from prob. 5 that

$$\left|\sqrt{2} - \frac{99}{70}\right| < \frac{1}{70\cdot 169} < 10^{-4}.$$

Table 2

a		1	2	2	2	2	2	2	...
p	1	1	3	7	17	41	99		...
q	0	1	2	5	12	29	70	169	...

Thus, the rational fraction 99/70 approximates $\sqrt{2}$ with an error less than 0.0001.

Example 2. Consider the periodic continued fraction

$$[2; 1, 1, 1, 4, 1, 1, 1, 4, 1, 1, 1, 4, \dots].$$

This is written in a "shorthand" notation as $[2; \overline{1, 1, 1, 4}]$. The number corresponding to this fraction can be found as follows. Call this number α. We have the equation (why?)

$$\alpha = 2 + \frac{1|}{|1} + \frac{1|}{|1} + \frac{1|}{|1} + \frac{1|}{|2+\alpha},$$

or, after simple manipulations,

$$\alpha = \frac{21+8\alpha}{8+3\alpha}.$$

From this, we find that $\alpha^2 = 7$, and, since $\alpha > 2$, we have $\alpha = \sqrt{7}$.

It is easy to see that this method can be used to calculate the value of any periodic continued fraction.

Best approximations and continued fractions

The best approximation to a number α is a fraction p/q ($q > 0$) such that

$$|q'\alpha - p'| > |q\alpha - p|$$

for all 1 $q' < q$ and all p'. The following theorem holds.

Theorem. *Any best approximation of α is a convergent of α; for $k > 1$. The converse is also true; that is, any convergent p_k/q_k (for $k > 1$) is a best approximation for α.*

The proof of this theorem is not difficult, and it is left for the reader. If you cannot do it yourself, the proof can be found in S. Lang's book, *Introduction to Diophantine Approximations,* Addison-Wesley, Reading, Mass. (1966).

Equivalence of numbers

Two numbers α and β are called equivalent if

$$\alpha = \frac{a\beta + b}{c\beta + d},$$

where a, b, c, and d are integers, and $ad - bc = 1$ or -1.

We denote by σ and σ_k (where k is an integer) the operations defined as follows:

$$\sigma\alpha = \frac{1}{\alpha}, \ \sigma_k\alpha = k + \alpha.$$

It is clear that by applying σ and σ_k to any number α, we obtain an equivalent number. It is also clear that the operations

$$\sigma\sigma_k\alpha = \frac{1}{\alpha + k},$$

$$\sigma_k\sigma_\alpha = k + \frac{1}{\alpha} = \frac{k\alpha + 1}{\alpha}$$

also yield numbers equivalent to the given number α.

Problems

7. Prove that if α and β are equivalent, then there exist integers k_1, k_2, ..., k_j such that

$$\beta = \sigma_{k_1}\sigma\sigma_{k_2}\sigma\ldots\sigma_{k_j}\alpha.$$

8. Assume that α and β have the same continued fraction expansion beginning with certain numbers k and s, respectively; that is,

$$\alpha = r_0 + \frac{1|}{|r_1} + \frac{1|}{|r_2} + \ldots + \frac{1|}{|r_k} + \frac{1|}{|d_1} + \frac{1|}{|d_2} + \ldots,$$

$$\beta = h_0 + \frac{1|}{|h_1} + \frac{1|}{|h_2} + \ldots + \frac{1|}{|h_s} + \frac{1|}{|d_1} + \frac{1|}{|d_2} + \ldots. \quad (3)$$

Prove that α and β are equivalent.

9. Prove that if α and β are equivalent, their continued fractions have the form (3).

Quadratic surds

Irrational numbers that are roots of quadratic equations with integer coefficients are called *quadratic surds.* They can be written as

$$\pm\frac{P + \sqrt{D}}{Q},$$

where P, Q, and D are integers, and D is not a perfect square. For example, the numbers $\sqrt{2}$, $\sqrt{7}$, $1 + \sqrt{2}$, and $(5 - \sqrt{7})/2$ belong to this category. In 1770, Lagrange proved the following theorem.

Theorem. *Quadratic surds and only quadratic surds have periodic continued fractions.*

As in the case with periodic decimal expansions (of rational numbers), the period may begin at a certain place within the expansion.

The fact that the value of a periodic continued fraction is a quadratic surd is proved in the same way as in Example 2. The proof of the inverse proposition is much more difficult.

We will prove it only for one important particular case: for reduced surds. A quadratic surd α is called *reduced* if $\alpha > 1$ and the second root α' of the quadratic equation for α (the so-called *conjugate* of α) satisfies the inequalities

$$-1 < \alpha' < 0.$$

Problems

10. Prove the following: If α is a reduced quadratic surd, then

$$\alpha = \frac{P + \sqrt{D}}{Q}, \quad (4)$$

where

$$0 < P < \sqrt{D}. \quad (5)$$

In addition, $P^2 - D$ is divisible by Q.

11. If α is a reduced quadratic surd and

$$\alpha = a_0 + \frac{1}{\alpha_1},$$

where $a_0 = |\alpha|$, then α_1 is also a reduced quadratic surd.

It follows from problem 11 that all numbers α_n in the sequence defined by the equations

$$\alpha_n = a_n + \frac{1}{\alpha_{n+1}}, \ a_n = [\alpha_k]$$

will also be reduced. Here $[x]$ denotes the integer part of x, i.e., the greatest integer that does not exceed x. Furthermore, they will all be represented by formula (4) with the same D. Now, it follows from inequality (5) that this sequence contains only a finite number of different numbers, that is, for certain numbers n and m,

$$\alpha_n = \alpha_m.$$

Then $\alpha_{n+1} = \alpha_{m+1}$, $\alpha_{n+2} = \alpha_{m+2}$, and so on. Thus, the sequence $\{\alpha_n\}$ is periodic. The theorem is proved for reduced surds.

If α is an arbitrary quadratic surd, then the sequence $\{\alpha_n\}$ must contain a reduced number. From this fact the proof of the theorem in the general case follows without difficulty. To prove this fact, α_n can be expressed explicitly in terms of prob. 4 and of the numerators and denominators of the convergents of α. The fact that the convergents of α tend to α prob. 5 must also be used. With these hints the interested reader should be able to complete the proof.

It can be proved that the period of the continued fractions for reduced quadratic surds starts at the very beginning of the expansion. Such continued fractions are called *purely periodic.* The inverse assertion is also true. This fact was first proved in 1828 by the French mathematician Évariste Galois (who was still in high school at the time).

Periods of continued fractions of quadratic surds

The properties of the periods are not only very elegant in themselves, but are also useful for solving certain equations. Let's establish some of them.

Consider the number

$$\alpha = \frac{\sqrt{7} + 2}{3}.$$

It is easy to show that it is reduced. Simple manipulations yield the following continued fraction expansion with a period of length 4:

$$\alpha = \left[\overline{1;\ 1,\ 1,\ 4}\right].$$

The conjugate of α is

$$\alpha' = \frac{-\sqrt{7} + 2}{3}.$$

Expanding the number

$$-\frac{1}{\alpha'} = \frac{3}{\sqrt{7}-2} = \sqrt{7} + 2$$

in a continued fraction, we obtain

$$-\frac{1}{\alpha'} = \left[\overline{4;\ 1,\ 1,\ 1}\right].$$

The period of our fraction has inverted. Is this accidental?

Problem 12. If a quadratic surd α is represented by a purely periodic continued fraction

$$\left[\overline{a_0; a_1 \ldots, a_n}\right],$$

then the number $-1/\alpha'$, where α' is the conjugate of α, is represented by the purely periodic continued fraction

$$\left[\overline{a_n; a_{n-1} \ldots, a_0}\right]$$

with the period equal to that of α but written in the reverse order.

If D is a positive integer but not a perfect square, and $a_0 = [\sqrt{D}]$, then $\alpha = a_0 + \sqrt{D}$ is a reduced number ($\alpha' = a_0 - \sqrt{D}$ and $-1 < a_0 - \sqrt{D} < 0$). Therefore,

$$\sqrt{D} + a_0 = \left[\overline{2a_0; a_1, \ldots, a_n}\right]$$

and

$$\sqrt{D} = \left[a_0; \overline{a_1; \ldots, a_n, 2a_0}\right]. \qquad (6)$$

Using the assertion of problem 12, it is easy to prove the following fact.

Problem 13. If D is not a perfect square, then the continued fraction of D has the form of equation (6), where $a_0 = [\sqrt{D}]$ and the part $a_1, a_2, \ldots, a_n$ of the period is symmetric.

The Pell equation

In the 3rd century BC, the great Greek scientist Archimedes formulated his famous *cattle problem.* We will not state it completely here (it would take more than a whole page). It is sufficient to note that for its solution, it is necessary to introduce 10 variables obeying 7 linear and two quadratic equations. After transformation and elimination of variables, the problem reduces to the equation

$$x^2 - 4729494y^2 = 1, \qquad (^*)$$

which must be solved in integers. Archimedes and his contemporaries could not solve this problem.

In general, the equation

$$x^2 - Dy^2 = 1, \qquad (7)$$

where D is a positive integer[1] that is not a perfect square, is called the Pell equation. It is a *Diophantine equation;* that is, an algebraic equation with integer coefficients that must be solved in integers.

How can the Pell equation be solved? The first idea is a straightforward search: we can successively substitute the numbers $x = 1, 2, 3, \ldots$ in the formula

$$y = \sqrt{(x^2-1)/D}$$

until the radicand becomes a perfect square. However, the following example demonstrates that this approach is impracticable. The equation

$$x^2 - 991y^2 = 1$$

has integer solutions (x_0, y_0), but the smallest x_0 among them is

379,516,400,906,811,930,638,014,896,080.

Even the fastest computer cannot find this solution by a simple search!

Continued fractions provide a convenient instrument for solving the Pell equation. Here, we describe the algorithm, but don't give its proof.

For any positive integer D that is not a perfect square, equation (7) has an infinite number of positive integer solutions. All of them can be found by the formula

$$x + y\sqrt{D} = (x_0 + y_0\sqrt{D})^k,\ k = 1, 2, \ldots,$$

where (x_0, y_0) is the solution with the smallest value of y. To find the least solution (x_0, y_0), we can expand $\sqrt{D}$ as a continued fraction. If

$$\sqrt{D} = \left[a_0;\ \overline{a_1, \ldots, a_n, 2a_0}\right]$$

and p_n/q_n is the nth convergent of $\sqrt{D}$, then

$$p_n^2 - Dq_n^2 = (-1)^{n-1}. \qquad (8)$$

If n is odd (the period is even), then

$$p_n = x_0,\ q_n = y_0$$

is the least solution to the Pell equation. If the period is odd (n is even), then the least solution is found from the formula

$$x_0 + y_0\sqrt{D} = \left(p_n + q_n\sqrt{D}\right)^2.$$

Example 3. Consider the equation

$$x^2 - 7y^2 = 1.$$

From the expansion $\sqrt{7} = [2; \overline{1, 1, 1, 4}]$, we find

$$\frac{p_3}{q_3} = 2 + \frac{1|}{|1} + \frac{1|}{|1} + \frac{1|}{|1} = \frac{8}{3}.$$

Since the period is even, the least solution is $x_0 = 8$, $y_0 = 3$; any other integer solution is obtained by the formula

$$x + y\sqrt{7} = \left(8 + 3\sqrt{7}\right)^k.$$

For $k = 2$, we have $x = 127$, $y = 48$.

Example 4. Consider the equation

$$x^2 - 13y^2 = 1.$$

From the expansion $\sqrt{13} = [3; \overline{1, 1, 1, 1, 6}]$, we find

$$\frac{p_4}{q_4} = 3 + \frac{1|}{|1} + \frac{1|}{|1} + \frac{1|}{|1} + \frac{1|}{|1} = \frac{18}{5}.$$

The period is odd; therefore, the least solution is found from the formula

$$x_0 + y_0\sqrt{13} = \left(18 + 5\sqrt{13}\right)^2 = 649 + 180\sqrt{13};$$

CONTINUED ON PAGE 51

[1] If $D = m^2$, $m \in \mathbf{N}$, then equation (7) has no integer solutions (why not?).

CONTINUED FROM PAGE 21

able-frequency ac power generator. How does the brightness of the lamp vary with frequency? What will happen if the leads of one of the coils are interchanged? (A.Zilberman)

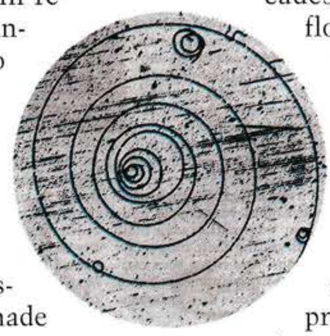
P285

Inner circle. If a capillary tube is viewed from the side, its apparent inner radius is r. What is the true value of the inner radius? The refractive index of the glass is n. The outer diameter of the capillary tube is much greater than its inner diameter.

SOLUTIONS ON PAGE 46

Do you know atoms ar

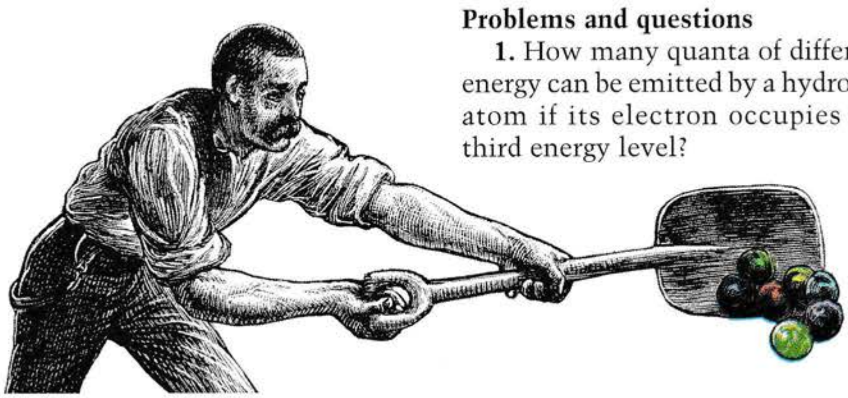
THE IDEA OF THE ATOMIC structure of matter undoubtedly sprang from the age-old tendency of human beings to impose some order on the surrounding world. The search for an eternal and immutable substance of which all objects are made began in remote antiquity and has continued for centuries, right up to the present day. Although there is still no "final solution" to this problem, some marvelous discoveries have occurred along the way. The atom and its nucleus have turned out to have a composite nature. The nucleus is made up of particles that cannot live for long outside of it. Other discoveries include radioactivity, transformations of one kind of elementary particle into another, chain and thermonuclear reactions, and many other phenomena. The last few decades have brought a flood of discoveries that have radically altered the scientific views on the structure of matter and have raised an enormous number of new problems to study.



The nature of physical experiments has changed dramatically. Nowadays an experiment can require the coordinated efforts of thousands of people. The practical applications of the methods of atomic and nuclear physics have turned out to be wonderfully versatile.

In this short column we can only sketch the broad outlines of the intricate world hidden in the smallest particles of matter.

"...these primitive particles... are incomparably harder than any porous bodies compounded of them; even so very hard, as never to wear away, or break in pieces."
—Isaac Newton



Problems and questions

1. How many quanta of different energy can be emitted by a hydrogen atom if its electron occupies the third energy level?

2. In what way is the tendency to minimize potential energy manifested in the electron shell of the atom?

3. Is there a relationship between the frequency of revolution of an electron about the nucleus and the frequency of emitted light?

4. Bombardment of boron atoms with fast protons in a Wilson cloud chamber produces three almost identical tracks of particles moving in different directions. What are these particles?

5. Why are all types of radioactivity not accompanied by changes in the chemical properties of the substance?

6. In what cases can the activity of a radioactive substance be considered constant?

7. Which is longer: three half-lives or two mean lifetimes of the nuclei of the same radioactive element?

8. The energy of alpha particles emitted by a radioactive substance can assume only discrete values. What conclusions can be drawn about the possible values of the energy of the nucleus?

9. Why can't the alpha particles emitted by radioactive substances cause nuclear reactions in heavy elements?

10. Why do the alpha particles emitted by a given sort of nuclei all have the same energy, while the energies of beta particles emitted by nuclei of a given sort are different?

11. If nucleons can attract each other, why haven't all the nuclei fused into a single huge supernucleus?

12. Why aren't the elements oc-

s and their nuclei?

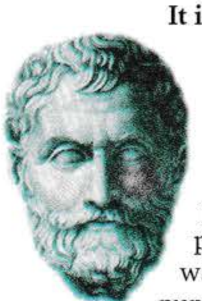
cupying the middle and end of the Periodic Table suitable for use as neutron moderators?

13. The rest mass of a nucleus is always smaller than the sum of the rest masses of its constituent nucleons. Can we say that conservation of mass is violated in the formation of nuclei?

Microexperiment

Heat an iron nail "white-hot" with a gas burner. Can you heat a piece of glass to produce similar luminance?

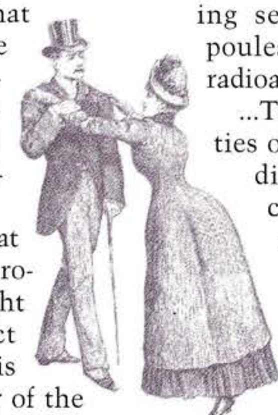
It is interesting that...



... Thales of Miletus, the founder of Antic philosophy and science, traced all the diversity of matter and phenomena to one primary substance: water. Anaximenes, his pupil and a member of the Miletus philosophic school, gave the role of the primary substance to air, the compression and rarefaction of which gives rise to all matter. By contrast, a contemporary of Thales', Heraclitus of Ephesus, preferred fire, by which he also meant soul and intelligence.

... The planetary model of the atom, which has been associated with Rutherford's name after his experiments, was devised theoretically as early as 1901 by the French physicist Jean Baptiste Perrin, famous for his experimental study of Brownian motion. The title of his paper was "The Nuclear-Planetary Structure of the Atom."

... As early as 1815, William Prout, a medical doctor from Edinburgh, advanced the hypothesis that all chemical elements were composed of hydrogen atoms. And in 1911, Rutherford proposed that nuclei were made up of alpha particles.



... Rutherford supposed that the charge of a nucleus is proportional to the atomic weight of the element. The correct idea that the nuclear charge is proportional to the number of the element in the Periodic Table was advanced by the Dutch amateur physicist Van der Brook. Rutherford was skeptical of this hypothesis and regarded it as an amusing but unsubstantiated idea.

...If Enrico Fermi had managed to completely explain the results of his experiments on artificial radioactivity, humanity would have known that it was in principle possible to construct an atomic bomb as early as 1934. Rutherford, who was still living at the time, categorically refuted the possibility of making practical use of nuclear energy.

...The methods of nuclear physics are successfully used in criminology, where they make it possible to examine substances with a mass of less than 10^{-10} g. For example, a person can be identified from a tiny piece of a hair.

...To warm the interior of Lunokhod (the first lunar rover) and maintain its operation for many months on the surface of the Moon, it was equipped with a heating unit containing sealed ampoules filled with radioactive substances.

...The intrinsic radioactivities of men and women are different because they contain different concentrations of the radioactive isotope potassium-40.

—A. Leonovich

Quantum on atoms and nuclei:

S. L. Glashow, "The Elementary Particles," September/October 1990, pp. 49–51

M. Digilov, "A Strange Box and a Stubborn Brit," March/April 1991, pp. 26–27.

S. R. Filonovich, "The Power of Likeness," September/October 1991, pp. 23–27.

I. Lalayants and A. Milovanova, "Physics Fights Frauds," January/February 1993, pp. 11–16.

A. Eisenkraft and L. D. Kirkpatrick, "Focusing Fields," January/February 1996, pp. 30–31; July/August 1996, pp. 32–33.

H. D. Schreiber, "The Name Game of the Elements," September/October 1996, pp. 24–30.

Y. Bruck, M. Zelnikov, and A. Stasenko, "Wobbling Nuclear Drops," January/February 1997, pp. 12–17.

A. Korzhuyev, "Bohr's Quantum Leap," January/February 1999, pp. 42–46.

PHYSICS CONTEST

Tunnel trouble

by Arthur Eisenkraft and Larry D. Kirkpatrick

I wonder if I should fall right through the earth! How funny it'll seem to come out among the people that walk with their heads downwards.

—Lewis Carroll,
Alice in Wonderland

HOW DO YOU LIKE THEM apples? The Garden of Eden variety brought about knowledge and led to banishment. Newton's variety extended enlightenment and led to a revolution in science and thought that echoed through politics, poetry, logic, and philosophy.

Newton was certainly not the first person to see an apple fall from a tree. He may, however, have been the first to imagine the apple and the Moon to be one and the same. The apple falls to the ground. This is a good observation, but nothing particularly special. To declare that the Moon also falls to the ground, when everyone knows it does not get any closer to the Earth, requires genius. Newton had been thinking of the gravitational force that attracts the apple to the Earth. To Newton, the Moon was merely a much larger apple that is very much further from the Earth. The Moon also falls to the Earth. It is the tangential velocity of the Moon that prevents it from getting any closer to the Earth. It is the gravitational force that holds the Moon in orbit.

Newton had begun to think of a gravitational force of attraction between the Earth and the apple. He then realized that the force is between any two masses and decreases as the square of the distance between them.

$$F = \frac{Gm_1m_2}{r^2}$$

The proportionality constant G was not determined experimentally until a hundred years later by Cavendish in 1798. Cavendish hung pairs of 15 kilogram and 125 kilogram masses and observed the tiny attraction between them. The force of attraction is quite small indeed but it was enough to twist a tiny wire, and Cavendish was able to measure that twist and determine that $G = 6.67 \cdot 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2$.

Returning to Newton, one can't help but be blown away by Newton's insistence that we cannot be cajoled into agreeing that the distance between two spheres should be measured from the center of one sphere to the center of the other. Newton invented integral calculus to prove that the small attractions of each piece of Earth on the apple are equivalent to the attraction of the entire mass of the Earth if the entire mass was located at its center, one Earth radius from the apple.

The success of this treatment can be seen in the two calculations of the Moon's acceleration. In the first calculation, we look at the approximately circular orbit of the Moon about the Earth. The Moon's period is 27.3 days and its distance to the Earth is 60 Earth radii. One Earth radius $R_E = 6.37 \cdot 10^6$ meters. The centripetal acceleration of the Moon can then be calculated:

$$a = \frac{v^2}{R} = \frac{\left(\frac{2\pi R}{T}\right)^2}{R} = \frac{4\pi^2 R}{T^2}$$

$$= 0.0029\ \text{m/s}^2 .$$

An inverse square relationship for gravity would predict that the acceleration of the Moon would be $(60)^2$ or 3600 times less than that of the apple.

$$(9.8\ \text{m/s}^2)/3600 = 0.0027\ \text{m/s}^2$$

When Newton arrived at this inverse-square conclusion, it is said that he "could hear God thinking." Newton showed us that the apple is like the Moon and, simultaneously, that the Earth is like the heavens. This law of gravitation describes the planets about the Sun, the Sun about the Galaxy, and the dance of all clusters of galaxies in our Universe.

The gravitational force allows us to calculate the orbit of a satellite in low Earth orbit. The idea of a satellite orbiting the Earth first appears in Newton's landmark *Principia*, published in 1686.

$$F = \frac{GM_E m}{R_E^2} = \frac{mv^2}{R_E} = \frac{4\pi^2 m R_E}{T^2}$$

Art by Tomas Bunk


ALICE'S ADVENTURES IN WONDERLAND
DRINK ME
EAT ME!
NEWTON
PHILOSOPHIÆ NATURALIS PRINCIPIA
TOM BUNK

Therefore, the orbital period is given by

$$T = 2\pi\sqrt{\frac{R^3}{GM_E}},$$

which yields a period of 88 minutes.

A more sophisticated problem is to calculate the period of an apple that travels through a hole created along a diameter of the Earth. Since on its travel through the tunnel, the apple experiences a force due to each part of the Earth, some of the mass of the Earth will be pulling inward and some will be pulling outward. A useful observation is that the force at any position is equivalent to that of all of the enclosed mass only, as if that mass was located at the Earth's center. The mass in the external shell has no contribution to the force. At an arbitrary point a distance r from the center, the attractive enclosed mass is then

$$M' = \rho V = \rho\frac{4\pi r^3}{3}$$

where ρ is the density of the Earth. Assuming that the density is constant and that there is no friction in our tunnel, we can solve for the force on the apple at any point in the tunnel.

$$F = \frac{G\rho 4\pi r^3 m}{r^2} = \left(\frac{4\pi G\rho m}{3}\right)r = kr$$

When the r is toward the right, the force is toward the left, and the correct form of this equation is

$$F = -kr$$

We recognize this as the equation of a mass on a spring and as the signature of simple harmonic motion. The apple will oscillate back and forth through the Earth. The period of oscillation T is given by the equation

$$T = 2\pi\sqrt{\frac{m}{k}} = 2\pi\sqrt{\frac{3m}{4\pi G\rho m}} = \sqrt{\frac{3\pi}{G\rho}}$$

Assuming that the density of the Earth is $5.5 \cdot 10^3$ kg/m^3, we get a period of 84 minutes. This is very close to the period of an orbiting satellite. Should it be?

This month's contest problem mixes some classic problems with some new twists.

A. Reprinted in Halliday, Resnick, and Walker is a 1946 Moscow Olympiad problem. A spherical hollow is made in a lead sphere of radius R such that its surface touches the outside surface of the lead sphere and passes through the lead sphere's center. (See figure 1.) The mass of the sphere before hollowing was M. With what force will the lead sphere attract a small sphere of mass m, which lies at a distance d from the center of the lead sphere on the straight line connecting the centers of the spheres and of the hollow?

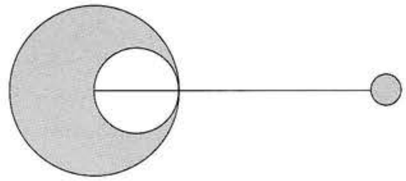


Figure 1

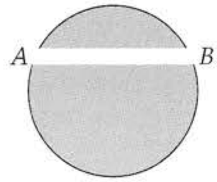


Figure 2

B. A tunnel is drilled along a chord of the Earth connecting points A and B. (See figure 2.) Calculate the period for an apple to travel from A to B. Comment on the feasibility of such a tunnel for global travel.

C. Does a straight tunnel provide for the fastest journey from A to B? If not, can you find a tunnel of two straight segments that requires a smaller time?

Image charge

In the July/August 1999 issue of *Quantum* we asked our readers to use the method of image charges and the uniqueness theorem to find the force of attraction between a charge q and a grounded metal surface. Part A asked readers to find the force for the case when the charge was equidistant from two large metal plates forming a right-angle corner as shown in figure 3. Using the results for a charge in front of an infinite metal plate calculated in the article and our knowledge of images formed by plane mirrors, we can guess that the image charges are those shown in figure 3. You can easily verify that the electrostatic potential is equal to zero along each of the two metal surfaces and at infinity.

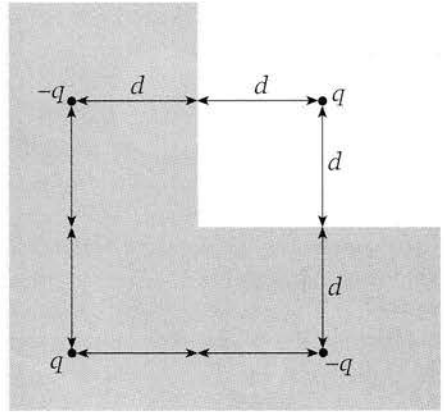


Figure 3

The electrostatic force felt by the charge is the same as that exerted by the three image charges. The two negative image charges attract the original charge toward the intersection of the two metal plates with a force

$$F_{neg} = 2\frac{kq^2}{(2d)^2}\frac{1}{\sqrt{2}} = \frac{1}{2\sqrt{2}}\frac{kq^2}{d^2},$$

while the positive image charge repels the image charge with a force

$$F_{pos} = \frac{kq^2}{\left(2\sqrt{2}d\right)^2} = \frac{1}{8}\frac{kq^2}{d^2}.$$

Therefore the net attractive force toward the corner is

$$F = \left(\frac{1}{2\sqrt{2}} - \frac{1}{8}\right)\frac{kq^2}{d^2} \approx 0.23\frac{kq^2}{d^2}.$$

Note that this is very close to the force obtained for the infinite metal sheet.

In part B we asked for the force when the charge q is located a distance d from the center of a grounded metal ball with a radius $c < d$. Let's choose an image charge Q located a distance $D < c$ from the center of the

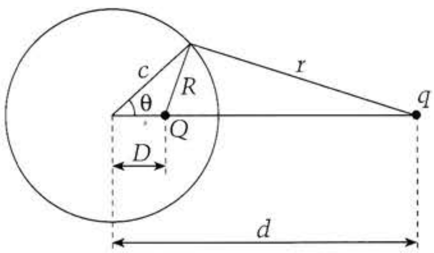


Figure 4

ball. As shown in figure 4, let's denote the distances from the real charge and the image charge to a point on the surface of the ball as r and R, respectively. The electrostatic potential at this point is given by

$$V = \frac{kq}{r} + \frac{kQ}{R}$$

where

$$r^2 = d^2 - 2dc\cos\theta + c^2$$

and

$$R^2 = D^2 - 2Dc\cos\theta + c^2.$$

Since the electrostatic potential is zero at all points on the surface of the ball,

$$Qr = -qR.$$

We now square both sides of this relationship, plug in the values of r and R, and group terms in powers of $\cos\theta$ on each side to obtain

$$Q^2(d^2 + c^2) - 2dcQ^2\cos\theta$$
$$= q^2(D^2 + c^2) - 2Dcq^2\cos\theta.$$

Because this relationship must be valid for all values of $\cos\theta$, the coefficients of each power of $\cos\theta$ must be equal on the two sides of this equation. This yields

$$Q^2 d = q^2 D$$

and

$$Q^2(d^2 + c^2) = q^2(D^2 + c^2).$$

We now solve both equations for the ratio Q^2/q^2 and equate them to obtain

$$\frac{D}{d} = \frac{D^2 + c^2}{d^2 + c^2}.$$

This is a quadratic equation in D that has two roots:

$$D = d,\ c^2/d.$$

The first root corresponds to the case where the real charge and the image charge are superimposed and the potential is zero everywhere. We are interested in the second root, for which

$$Q = -\frac{D}{c}q = -\frac{c}{d}q.$$

(Note that the location of the image is not the same as that for an object outside a spherical convex mirror.)

Using this image charge and image distance, the attractive force on the charge outside a grounded conducting sphere is

$$F = kq^2\left(\frac{cd}{d^2 - c^2}\right).$$

It is interesting to check the limiting cases. As the charge approaches the surface of the sphere, d approaches c and force becomes very large. And as the charge is moved to very large distances, the force decreases to zero. Both of these behaviors are expected and agree with the case for the infinite conducting plane.

DIGIT DEMOGRAPHICS

Repartitioning the world

Population and the powers of two

by V. Arnold

THE FIRST DIGIT OF THE number 2^n is equal to 1 about six times more often than it is equal to 9. The first digits of the populations and surface areas of the countries of the world are distributed in the same way. The explanation of this fact suggested below has led to many mathematical hypotheses, some of which have been proved and some of which are only supported by computer experiments and still await rigorous proof.

Powers of two

The sequence of the first digits of the numbers 2^n ($n = 0, 1, 2, \ldots$) begins with

$$1, 2, 4, 8, 1, 3, 6, 1, 2, 5, 1, \ldots$$

It can be verified by continuing the calculations that ones constitute about 30% of this sequence (while nines make up less than 5%). The same distribution is obtained for the sequence of the first digits of the numbers 3^n, and, in general, for almost any geometric progression. (Progressions with the ratios 10, $\sqrt{10}$, and, in general, $10^{p/q}$, where p and q are integers, are obvious exceptions.)

The proof of this surprising proposition was obtained by H. Weyl almost 100 years ago. He proved even more. Recall that any real number c can be uniquely represented as the sum of an integer (called the *integer part* of c) and a *fractional part* $\{c\}$ belonging to the interval $[0, 1)$.

Theorem. *Let x be an irrational number. Then the sequence of the fractional parts, $\{nx\}$ of the numbers nx ($n = 0, 1, 2, \ldots$) is uniformly distributed on the interval $(0, 1)$.*

This means that $1/N$ times the number of values of n ($0 \le n < N$) for which the fractional part of nx falls into any fixed segment of length a tends to a as N increases to infinity.

In other words, consider the motion of a point along the circumference of a circle where at integer moments of time (n) the point jumps ahead by an angle $2\pi x$ which is incommensurable with 2π (see fig. 1). The theorem asserts that the time spent by the moving point on any arc of the circle is asymptotically proportional (at large observation times) to the length of the arc (and depends neither on the position of the arc on the circle, nor on the initial point, nor even on the value of the angle).

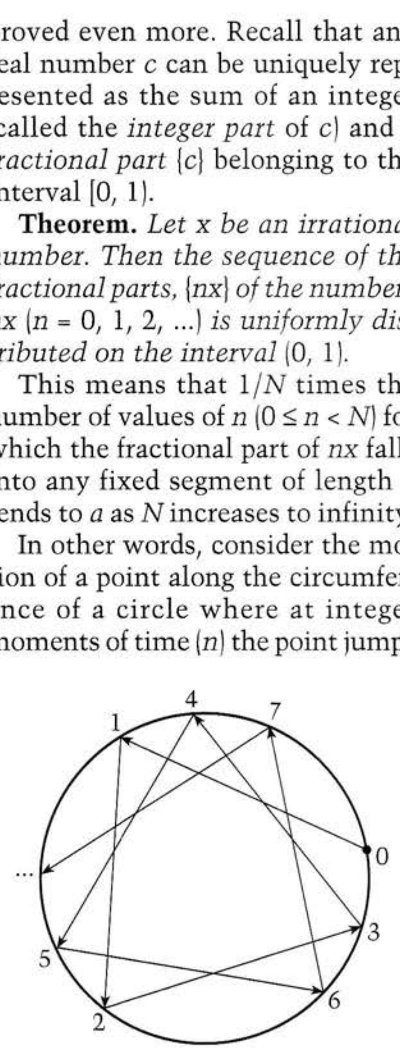


Figure 1. *The trajectory of a point when rotated again and again through an angle incommensurable with 2π.*

The distribution of the first digits of the numbers 2^n can now be obtained as follows. Consider the sequence of numbers $\log_{10}(2^n) = nx$. The number $x = \log_{10}2$ is irrational. By Weyl's theorem, the sequence of the fractional parts of nx is uniformly distributed on $(0, 1)$.

The first digit, i, of 2^n is determined by the interval between the numbers $\log_{10}(i + 1)$ and $\log_{10}i$ in which the fractional part of the number $\log_{10}2^n$ falls. By the theorem, the fraction of the numbers 2^n beginning with the digit i ($= 1, 2, \ldots, 9$) is equal to $p_i = \log_{10}(i + 1) - \log_{10}i$. For example, for the first digit $i = 1$, this fraction is $\log_{10}2 = 0.301$ (the fact that this number is close to 3/10 reflects the fact that $2^{10} = 1024$ is close to $1000 = 10^3$). This is why the percentage of ones among the first digits of the numbers 2^n is about 30%. The fractions made up by the different digits (in percent) are given in table 1.

Art by Vera Khlebnikova

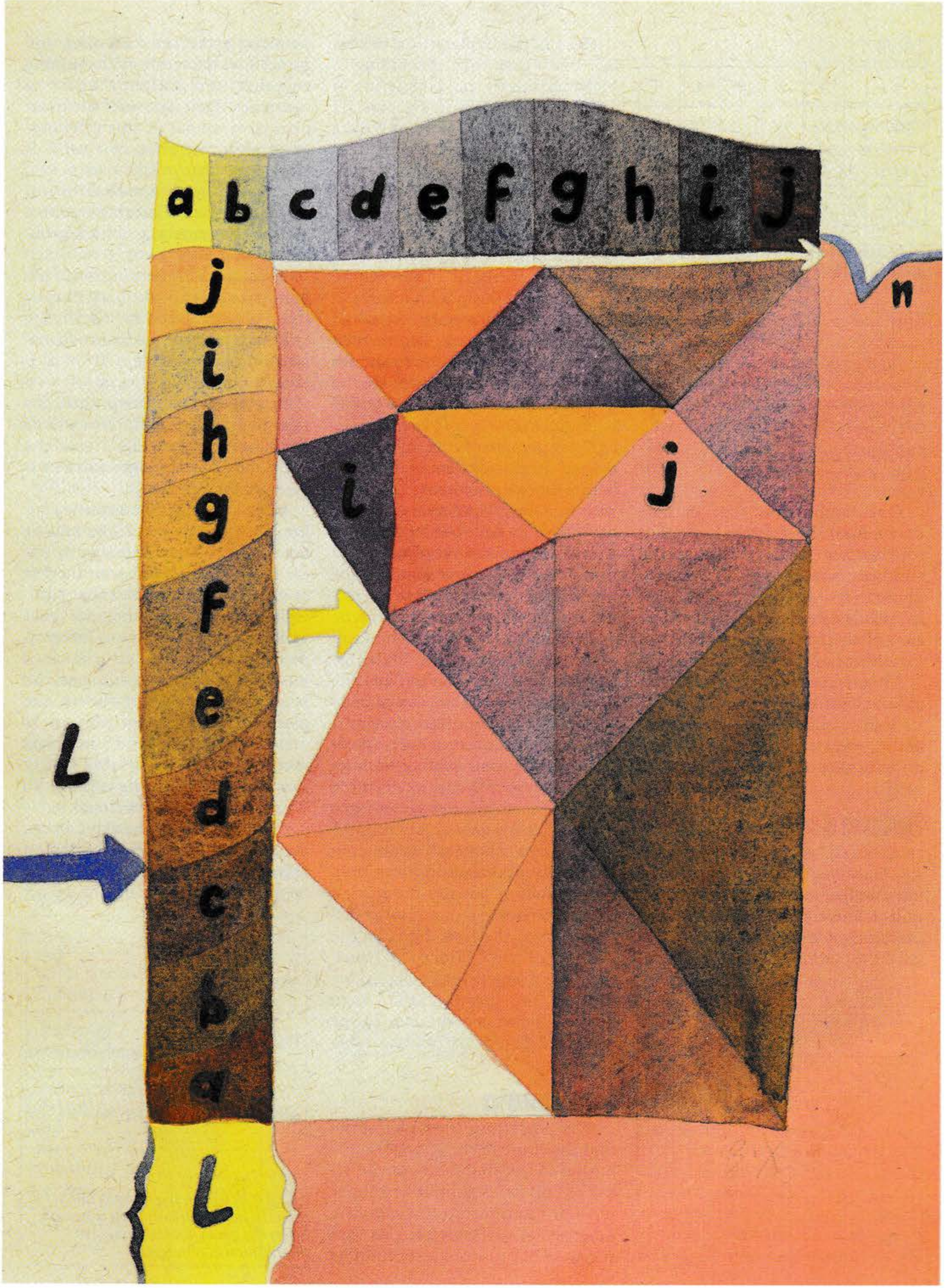

a b c d e f g h i j
j i h g f e d c b a
n
i
j
L
L

Table 1

i	1	2	3	4	5
$100 \times p_i$	30	17	12	10	8
i	6	7	8	9	
$100 \times p_i$	7	6	5	5	

Nines are about six times less common than ones (see fig. 2).

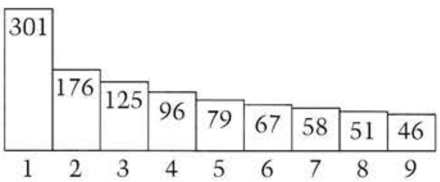


Figure 2. *Distribution of the first digits of powers of two.*

One conclusion that can be drawn from the above is important for further considerations: the strangely nonuniform distribution of the first digits of the numbers 2^n is explained by the uniform distribution of the fractional parts of the logarithms of these numbers.

This conclusion explains the similar distribution of the first digits of many different sequences (e.g., of the geometric progressions 2^n and 3^n, but other kinds of sequences as well).

Populations of countries

About 20 years ago, N.N. Konstantinov called to my attention the fact that the first digits of the population of different countries are distributed in the same way as the first digits of the powers of two (see table 2).

Table 2

first digit	1	2	3	4	5
number of countries (%), 1995	29	21	10	11	6
first digit	6	7	8	9	
number of countries (%), 1995	6	8	3	6	

Here is the explanation to this fact that I gave then. According to Malthus' theory, the population of each country grows in a geometric progression. It follows from Weyl's theorem (see the previous section) that the first digits of the population of a given country are distributed in the same way as the first digits of the powers of two (see fig. 2). According to the ergodic theorem (or, more correctly, the ergodic principle), the mean with respect to time can be replaced by the mean with respect to space, i.e., the distribution over different countries for one year must be identical to the distribution in one country in different years.

To verify this conjecture, I considered the distribution of the number of pages in the books of my library, the lengths of rivers, and the heights of mountains. In all these cases, the proportion of ones and nines among the first digits of the numbers obtained proved to be practically equal: $p_i 1/9$. Books, rivers, and mountains do not grow in geometric progression, so Malthus' theory is not applicable. Thus the difference between the statistics of the first digits of numbers representing population and, say, river lengths provides a kind of indirect confirmation of Malthus' formula (by which population grows in a geometric progression).

However, about 10 years ago, M.B. Sevryuk found that not only the population, but also the areas of countries obey the same law of distribution of the first digits as the powers of two. Since Malthus' theory is not apparently applicable to areas, the question arises as to how this behavior of the areas can be explained. Let me try to answer this question.

Areas of countries

The preceding examples show that the cause of the strange distribution of the first digits of the areas of countries should be sought in their relationship to geometric progressions. History tells us that the areas of countries (especially of empires) sometimes increase and sometimes decrease, by the joining together and splitting apart of countries. First, consider the most primitive model of this phenomenon. Assume that in a unit of time a country splits into two equal parts with a probability of 0.5 or joins with another country having the same area with a probability of 0.5.

Theorem. *The distribution of the fractional parts of the logarithm of the area occupied by such a country at time* n *tends to a uniform distribution on the interval* (0, 1) *as* n *tends to infinity.*

In other words, the probability that the first digit of the area will be 1 tends to $\log_{10}2 \cong 0.301, \ldots$, and the probability that this digit will be 9 is approximately 0.046.

Indeed, consider the sequence $l_n = \log_{10}S(n)$, where $S(n)$ is the area at time n. At the next moment, $n + 1$, the point l_n shifts either to the left or right by $\log_{10}2$ with equal probability (certainly the choice of what to do—merge or divide—is independent of the choices at other moments of time). According to the laws of the probability theory, the distribution of l_n for large n is mainly concentrated on a long segment (of order $\sqrt{n}$) and is gently sloping and symmetric (see fig. 3). When we pass to fractional parts (i.e., when we wind the axis l on the circle l mod 1), this distribution on l gives a nearly uniform distribution on the circle (for large n). The details

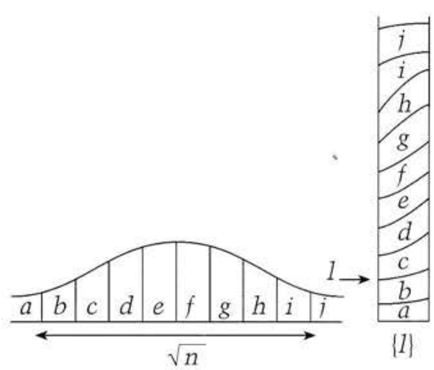


Figure 3. *When the straight line with a gently sloping distribution is wound up on the circle, a nearly uniform distribution is obtained.*

of the proof are left for the reader to complete. It is important that the sequence of fractional parts of the numbers $m\log_{10}2$ is distributed uniformly.

There are numerous more complex models for the repartitioning of the world that produce the same effect in numerical experiments. It is likely that the uniformity of the distribution of the fractional parts of the logarithms of the areas of countries in the limit can be rigorously proved for whole classes of such models. Here are some examples.

1. Initially there are k countries, with the areas $S_1, ..., S_k$. At each succeeding moment, one (randomly chosen) country divides in two with a 50% probability or unites with another (randomly chosen) country with a 50% probability. Of course, the choices made at different times are assumed to be independent, and any two countries are equally likely to be chosen.

The computations performed by M.V. Khesina (the University of Toronto, June, 1997) show that for $S_i = i$, $k = 100$, the distribution of the first digits of the areas is practically the same as the distribution of the first digits of powers of two after only 100 steps.

2. If we allow division into unequal parts with a certain law of distribution of the parts (e.g., uniform), we obtain the same result.

3. In models that allow only for neighbors to unite, the same distribution of the first digits is reached. For example, in one of F. Aicardi's models (Trieste, June, 1997), countries are represented by arcs of a circle, and areas are represented by the lengths of these arcs. A distribution that is almost the same as the distribution of the first digits of powers of two is very soon reached.

Table 3

	A	B	C
initial number of countries, n	100	62	100
number of iteration steps, T	200	150	200
average number of countries at time T	98	114	898
probability of division, p	0.5	0.6	0.5
part of area detached, α	0.5	0.5	0.3

Table 4

	1	2	3	4	5	6	7	8	9
A	0.297	0.183	0.123	0.107	0.073	0.058	0.065	0.046	0.049
B	0.309	0.180	0.123	0.106	0.069	0,059	0.058	0.050	0.045
C	0.294	0.181	0.111	0.091	0,084	0.077	0.059	0.052	0.048
D	0.301	0.176	0.125	0.096	0.079	0.067	0.058	0.051	0.046

4. In another of Aicardi's models, the world is represented by a graph that describes the decomposition of a sphere into triangles (whose n vertices represent n countries with an area randomly distributed on the interval $(1, n)$). The graph is constructed beginning with the icosahedron by iterations of the following operation: a triangular face is randomly chosen, a vertex is added at its center, and it is connected with all three vertices of this face.

In this model, the repartitioning of the world is organized as follows. At each moment in time a vertex i is randomly chosen and then the number of countries is increased with probability p or decreased with a probability $1 - p$. In the first case, a triangular face containing the vertex i is randomly chosen and a new vertex is created at the center of the face. This vertex is connected to the three vertices of the face, and the newly created country acquires a fraction α of the area of the country i (see fig. 4).

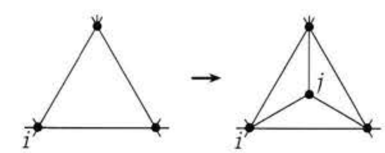


Figure 4. *Separation of a new country j from the country i.*

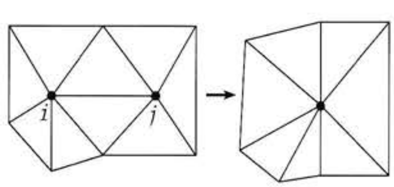


Figure 5. *Federation of two countries, j and i.*

In the second case, a vertex j adjacent to the vertex i is randomly chosen and the countries i and j are united; in the process, the edge ij and the two triangular faces divided by this edge are deleted (see fig. 5). Table 3 shows the values of the parameters in three experiments, A, B, and C. Table 4 shows the average values of the proportion of ones, ... nines among the first digits of the areas of the countries obtained after 50 repetitions of the experiment with different initial values. The last row (D) of this table shows the frequency of the first digits of powers of two.

It would be interesting not only to prove a general theorem that specified the domain of applicability of the uniform distribution of the fractional parts of logarithms but also to check whether, for example, the size of companies or their revenues obey this law.

The occurrence of the strange distribution of the first digits in many different situations has been discussed in numerous papers. However, I have never run across any mathematical theorems or hypotheses (like the ones discussed in this paper) that explained the inevitability of this distribution (except, of course, for Weyl's theorem).

Fuel economy on the Moon

by A. Stasenko

"*DRILLING THROUGH the Moon with a huge drill'—that was the subject of Mr. Lund's speech! The time is nigh when the Moon will be decorated by a hole. This hole will belong to Great Britain.*" (Anton Chekhov)

It is impossible that such a colossal project would not produce something useful. Let's simply drop some object (say, a household iron) into this shaft, which passes directly through the Moon's center (figure 1). Near the surface (point N) the force acting on the iron will be mg_M, where m is the mass of the iron and g_M is the acceleration due to gravity at the Moon's surface (approximately 1/6 of that at the Earth's surface). When the iron passes through the center of the Moon (point O), the force of gravity will be zero (if the Moon has spherical symmetry), because at that point the elementary gravitational forces generated by all the mass elements of the Moon counterbalance one another. Therefore, at that point the iron's speed is no longer increasing, and after passing through the center of the Moon the iron will begin to slow down. Intuition tells us that the iron will come to a stop at point S, which is diametrically opposite to point N. Indeed, since the Moon has no atmosphere and the iron doesn't contact the walls of the shaft, its motion is frictionless. Therefore, the iron will not lose energy, and it can repeat its fall from the point S if it is not caught there. It will pass unimpeded through the center of the Moon and on to the surface at point N—and so on. This oscillatory motion could go on forever.

As we have said, the acceleration due to gravity has different values at the Moon's center (zero) and at its surface (g_M). Therefore, the acceleration due to gravity is some function of the distance from the center of the Moon. The simplest dependence would be linear (keep in mind that this is still a hypothesis):

$$g_y = -g_M \frac{y}{R_M}. \qquad (1)$$

Now let's hurl the same iron along the lunar surface (in other words, horizontally or tangentially) in such a way that it could reach point S again (figure 2). To do this we must impart to the iron the orbital speed v_1. Recall that this value can be found by equating the acceleration due to gravity (which is constant for all points on the spherical surface) to the centripetal acceleration $g_M = v_1^2/R_M$, whence

$$v_1 = \sqrt{g_M R_M}. \qquad (2)$$

This orbital speed is also called the first cosmic speed and carries the subscript 1. Note that although the value of the iron's speed is the same at any point on the circular trajectory, its motion is nevertheless accelerated, because the velocity vector is constantly changing direction (rotating).

At some time t the iron will be located at point C on the circle, with angular coordinate θ (the polar angle). Its distance from the starting point N (corresponding to $t = 0$, $\theta = 0$) measured along the arc NC is $v_1 t$. Let's project all kinematic parameters of the iron at this point (its radius vector $\mathbf{R}_M$, velocity $\mathbf{v}_1$, and acceleration $\mathbf{g}_M$) onto the shaft axis SN (that is, the y-axis, figure 2):

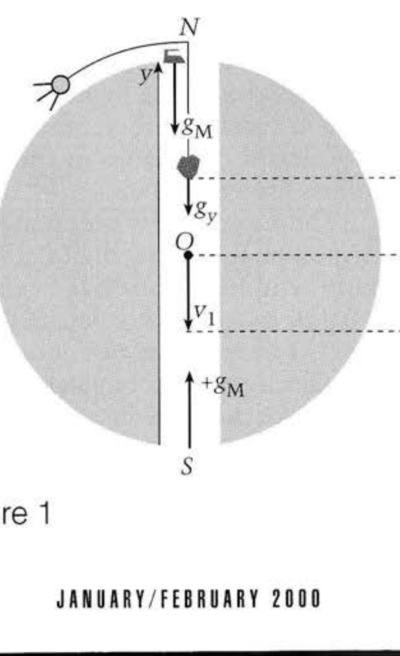


Figure 1

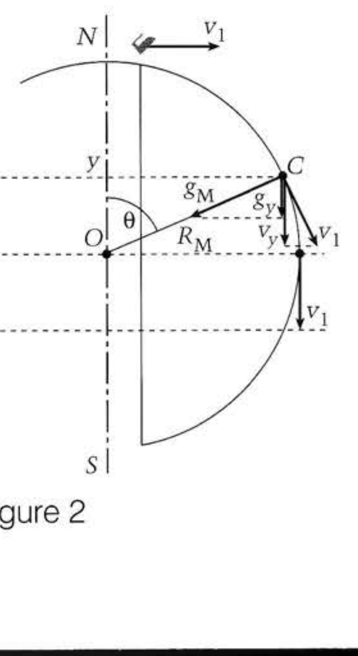
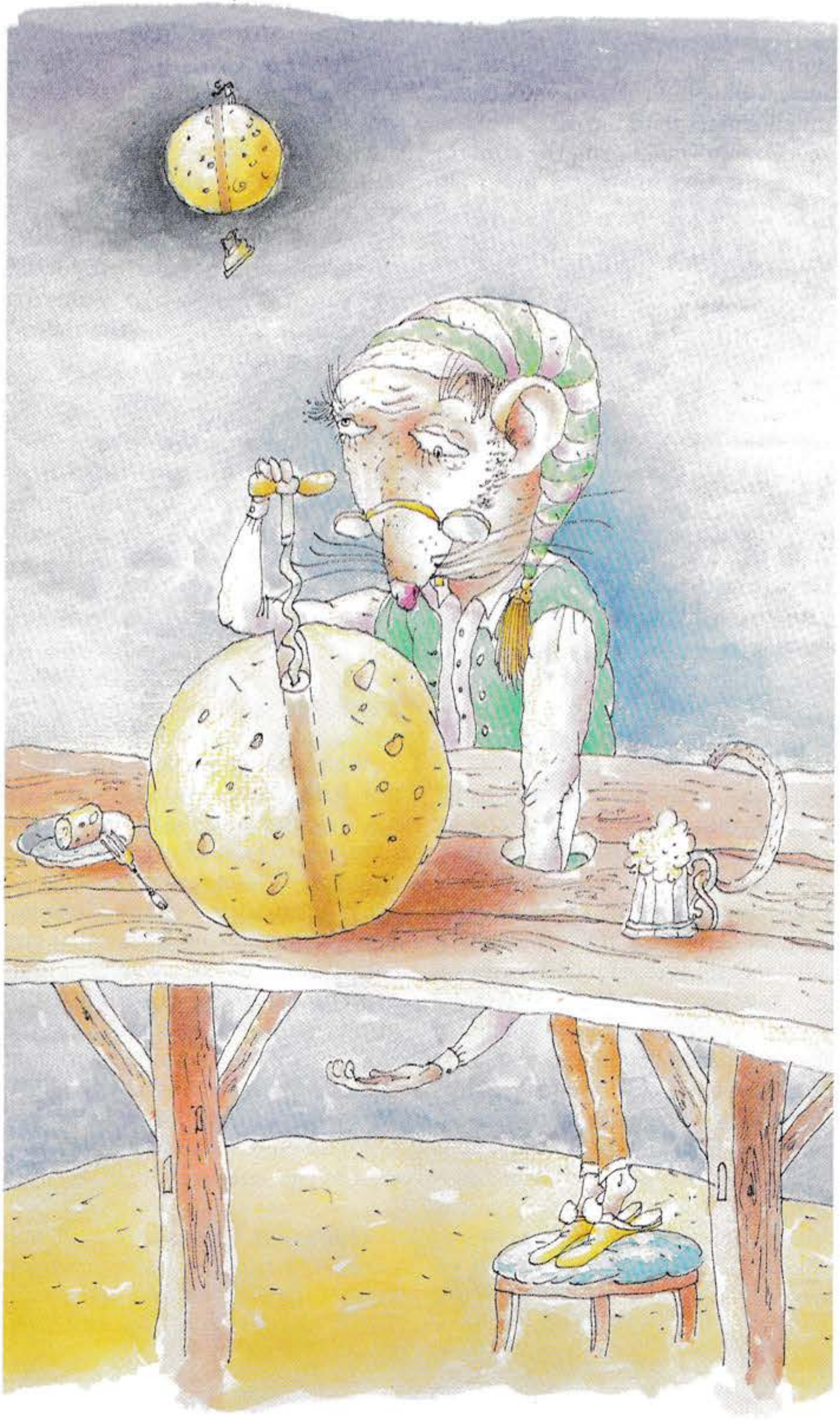


Figure 2



Art by Vadim Ivanyuk

$$y = R_{\mathrm{M}} \cos \theta, \qquad (3)$$

$$v_y = -v_1 \sin \theta, \qquad (4)$$

$$g_y = -g_{\mathrm{M}} \cos \theta. \qquad (5)$$

Note that the central angle θ is the length of the circular arc divided by its radius: $\theta = v_1 t/R_{\mathrm{M}}$. Equations (3) and (5) yield a formula for g_y which coincides exactly with (1). Again, this is not a proof of formula (1), but it does hint at the actual behavior of the force on a body inside the lunar tunnel.

Clearly, if the projections of the acceleration, velocity, and position of an orbiting iron onto a diameter of the Moon are all exactly equal to the corresponding values for an identical iron moving along the same diameter, then both the orbiting and falling (oscillating) irons will get to point S simultaneously. Then they will also return to the starting point N at the same instant. This reasoning is correct, provided that formula (1) is true. In reality, this linear dependence is correct only in the case of a uniform planet with a constant density throughout its volume.

In the last 300 years this subject has been discussed in many classics of the exact sciences. A modern student knows from elementary science classes that a ball of mass M and radius r attracts a body of mass m located at its surface with the Newtonian force of gravity $F = GMm/r^2$, as if the entire mass of the ball were concentrated at its center. Plugging $M = (4/3)\pi r^3 \rho_0$ into Newton's formula (ρ_0 is the density of the uniform ball) yields

$$F = mg = mG(4/3)\pi \rho_0 r^3/r^2,$$

which gives

$$g(r) = G(4/3)\pi \rho_0 r.$$

This means that the acceleration is indeed proportional to the distance from the center. A very important feature in this reasoning is that the mass of a spherical layer lying on a surface r = const does not affect the force of gravity (or the acceleration due to gravity) in the region enclosed by this surface. This result is nontrivial. In fact, Newton held off publishing the *Principia* for nine years until he could prove it.

Now formula (3), written in the form

$$y = R_{\mathrm{M}} \cos (v_1 t/R_{\mathrm{M}}),$$

is the "timetable" for the motion of the iron in the lunar shaft. This formula describes a harmonic oscillation (figure 3, solid curve). According to (4), the speed will vary as

$$v_y = -v_1 \sin (v_1 t/R_{\mathrm{M}}),$$

reaching maximal values at those times when $y = 0$, that is, at the center of the Moon (figure 3, dashed curve).

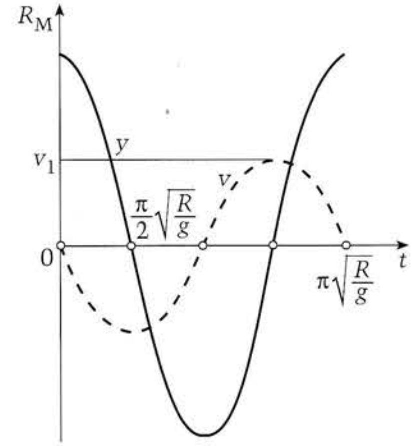


Figure 3

This last formula shows that at the center of the Moon, the iron will reach the first cosmic speed (the orbital speed of a lunar satellite). At this point in the exposition the practical mind of a bright student may flash on an idea: Why should the iron be allowed to swing idly from pole to pole? The price of fuel on the Moon is astronomical, so why not save some with the help of the gravitational field?

To harness the Moon's gravity for useful work we replace the iron by a very big rock (VBR) and attach a weightless inelastic cord of length R_M to it. The other end of the cord will be hooked to a spacecraft or a satellite to be launched. While falling into the shaft, the VBR will accelerate the spacecraft in the horizontal direction on the Moon's surface (presumably without friction). When the VBR passes the center of the Moon, the spacecraft will be just over the shaft. At this precise moment we unhook it from the cord, because both bodies will have attained the first cosmic speed v_1. The space vehicle will then orbit about the Moon, while the VBR will proceed on its way to the opposite pole S. At that pole another space vehicle could be hooked on. Thus the oscillating VBR could launch one satellite after another instead of aimlessly rocking in the shaft. Such a project would provide a great economy of fuel, which is of extreme importance on the Moon.

We have tacitly made a very important assumption. We have assumed that the mass of the VBR is much greater than the masses of the accelerated vehicle and the cord. In short, the VBR must be a VERY big rock.

Let's take things one step further. Equations (3) and (4) yield

$$\cos\theta = \frac{y}{R_M}, \ \sin\theta = -\frac{v_y}{v_1}.$$

We square each equation and add the resulting equations together:

$$\left(\frac{v_y}{v_1}\right)^2 + \left(\frac{y}{R_M}\right)^2 = \sin^2\theta + \cos^2\theta = 1.$$

Dividing this equation by 2 and plugging in the expression for v_1 from formula (2), we get

$$\frac{v_y^2}{2} + \frac{1}{2}\frac{g_M y^2}{R_M} = \frac{v_1^2}{2} = \frac{g_M R_M}{2}. \quad (6)$$

Doesn't this look familiar? Of course, this is the good old energy conservation law for a unit mass. The first term on the left-hand side is the kinetic energy of the falling body at distance y from the center. The second term depends only on the location of the body, so it must be its potential energy. The sum of the two terms is constant (equal to the right-hand side of equation (6)). This constant equals either the kinetic energy at the center of the Moon (where the potential energy is zero) or the potential energy on the Moon's surface at the entrance to the shaft (where the speed and kinetic energy are zero).

Thus potential energy of a body with mass m located at distance y from the center of the Moon is written in the form $(1/2)mg_M y^2/R_M = (1/2)mg_y y$. Note that the energy conservation law for a body thrown vertically upward above the Earth's surface to an altitude y looks like

$$\frac{mv^2}{2} + mg_0 y = \frac{mv_0^2}{2} = mg_0 y_{max},$$

where v_0 is the initial speed imparted to the body and y_{max} is its maximum altitude. In this case the potential energy is $mg_0 y$, and it is equal to the work that must be performed to raise the body to an altitude y. In contrast, when a body falls into a long shaft, it is acted upon by a variable force of gravity (1), so the potential energy in this case is one-half the usual product of mg_y times y, because in geometrical terms it is the area of the shaded triangle in figure 4.

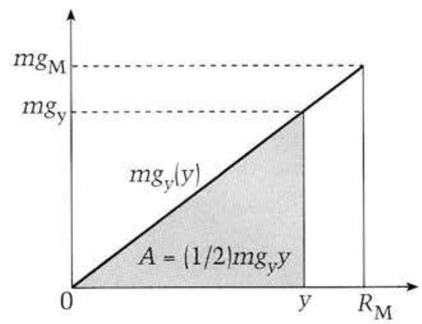


Figure 4

What should we do if the rock is not VERY big, that is, when its mass is comparable to that of the satellite and the cord? In this case the energy conservation law (6) must be modified to take into account the energy of the satellite and cord. This exercise is a good "nightcap" before going to bed. Keep in mind that as the rock falls into the shaft, an ever-increasing fraction of the cord will be inside the shaft. The weight of this part of the cord will also contribute to the acceleration of the satellite.

Quantum on space travel and gravitational catapults:

Y. Osipov, "Catch as Catch Can," January/February 1992, pp. 38–43.

A. Byalko, "A Flight to the Sun," November/December 1996, pp. 16–20.

V. Surdin, "Swinging from Star to Star," March/April 1997, pp. 4–8.

V. Mozhaev, "In the Planetary Net," January/February 1998, pp. 4–8.

A. Stasenko, "From the Edge of the Universe to Tartarus," March/April, 1996, pp. 4–8.

Bulletin Board

Olympiad honors

Competing against teams representing 81 countries, a team of six American high school students won six medals at the 40th International Mathematical Olympiad (IMO) held in Bucharest, Romania, this past summer.

This year's team was chosen solely on the basis of the 28th annual USA Mathematical Olympiad held last April. Team members included gold-medalist Reid Barton from Arlington, Massachusetts, gold-medalist Paul Valiant from Milton, Massachusetts, silver medalist Gabriel Carroll from Oakland, California, silver medalist Po-Shen Loh from Madison, Wisconsin, silver medalist Melanie Wood from Indianapolis, Indiana, and bronze medalist Lawrence Detlor from Brooklyn, New York.

U.S. team leader Titu Andreescu, director of the American Mathematics Competitions, asserted: "The competition is getting tougher and tougher. There are more and more strong countries. We consider this year's participation a success. Two of our students received a gold medal, and the two first-time IMO participants on our team received a silver and bronze medal. It appears that next year's U.S. team will include four gold medalists, two from 1998 and two from 1999. We will have a stronger team, and all six team members will go for the gold."

The following is a list of the top 10 teams and their scores (out of a possible 252 points).

China (182), Russia (182), Vietnam (177), Romania (173), Bulgaria (170), Belarus (167), Korea (164), Iran (159), Taiwan (153), USA (150).

From right to left: Titu Andreescu (USA team leader), Reid Barton, Melanie Wood, Po-Shen Loh, Gabriel Carroll, Paul Valiant, Lawrence Detlor, Walter Mientka (USA observer).

Here is a sample question from this year's olympiad:

Determine all functions f from R to R such that

$$f(x - f(y)) = f(f(y)) + xf(y) + f(x) - 1$$

for all x and y in R.

Getting to the square root of the problem

Depending on your world view—base ten or binary—there were actually two correct answers to this month's contest. An asterisk denotes those who chose the binary approach to solving the CyberTeaser. This month's winners are:

Maxim Bachmutsky (Kfar-Saba, Israel)
Fred Witkowski (Bydgoszcz, Poland)
Bruno Konder (Rio de Janeiro, Brazil)
Shvachko Nikita (San Carlos, California)
Michael L. Marfil (Camalig, Albay, Philippines)*
Nick Fonarov (Staten Island, New York)
Jerold Lewandowski (Troy, New York)
Lena Oleynikova (Waldorf, Maryland)*
Sergio Moya (Culiacans, Sin, Mexico)
Igor Astapov (Kingston, Ontario, Canada)

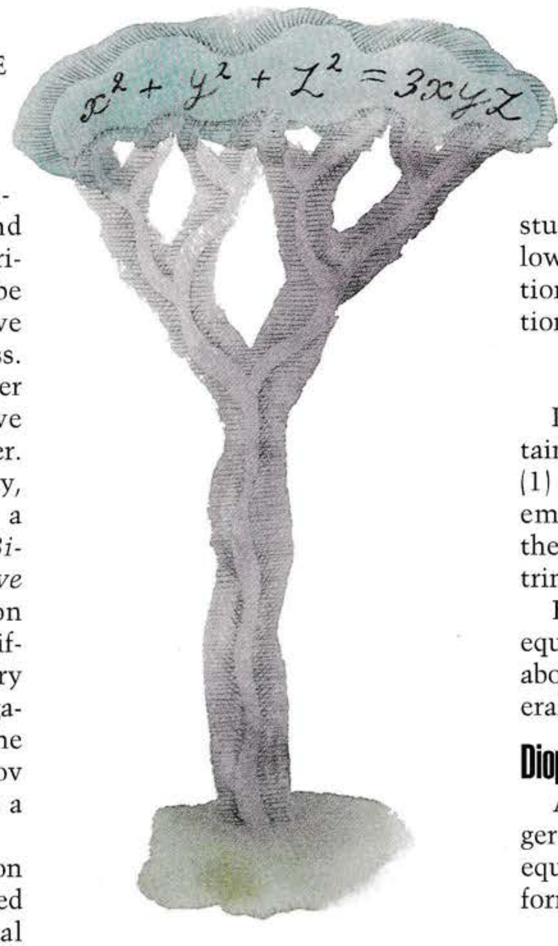
Congratulations to our winners, who will receive a *Quantum* button and a copy of this issue.

Everyone who submitted a correct answer before it was posted at our website was eligible to win a copy of our brainteasers collection *Quantum Quandaries*. Visit http://www.nsta.org/quantum to find out who won the book, and while you're there, try your hand at our new CyberTeaser!

AT THE
BLACKBOARD II

The Markov equation

by M. Krein



IN THIS PAPER I OUTLINE the history of an equation in integer variables (Diophantine equation) and give its solution. The solution uses only the simplest properties of integers and "Vieta's theorem"[1] for quadratic trinomials. It can therefore easily be understood by students who have taken an elementary algebra class. Our solution is based on a number of statements whose proofs have been left as problems for the reader.

In 1879, at Petersburg University, a young man of 23 defended a master's dissertation entitled *Binary Quadratic Forms with Positive Determinants.* This dissertation gave the solutions of some very difficult problems of number theory and started a new line of investigation in that field. The author of the dissertation was Andrei A. Markov (1856–1922), who later became a prominent mathematician.

The main part of the dissertation was based on two articles published in the well-known mathematical journal *Mathematische Annalen* in 1879 and 1880. However, it was only after more than 30 years had passed that these works of Markov's were "discovered." In 1913, the prominent German mathematician Georg Frobenius (1849–1917) published a memoir called *On Markov's Numbers.* He wrote in the preface that even though Markov's investigations are "extraordinarily important and remarkable," they appear to have remained little-known. Frobenius attributed this to the complexity of the presentation (Markov made systematic use of continued fractions, which were unpopular in this field at that time[2]).

In this article I will not try to present even a simplified account of Markov's profound investigations. It happens, however, that in the course of his studies Markov considered the following auxiliary Diophantine equation (now called the Markov equation):

$$x^2 + y^2 + z^2 = 3xyz. \qquad (1)$$

It is wonderful that Markov obtained all the solutions to equation (1) using only the methods of elementary mathematics (among them Vieta's theorem for quadratic trinomials played a central role).

Before presenting a solution to equation (1), let me say a few words about Diophantine equations in general.

Diophantine equations

A Diophantine equation for integer-valued variables $x, y, \ldots, w$ is an equation that can be reduced to the form

$$P(x, y, \ldots, w) = 0,$$

where P is a polynomial in the given variables with integer coefficients.

Sometimes, rather simple problems lead to Diophantine equations. For example, the problem of finding ways to pay a sum of n kopecks with coins of value 1, 2, 3, and 5 kopecks (all of which existed in the old Soviet Union) leads to the Diophantine equation

$$x + 2y + 3z + 5w = n.$$

The problem of constructing a right triangle with pairwise com-

[1]This theorem says that the sum of the roots of the quadratic equation $ax^2 + bx + c = 0$ is $-b/a$, and the product of the roots is c/a. There is an important generalization to equations of higher degree, and the theorem is often (but not in American textbooks) attributed to the French algebraist François Viète or Vieta.

[2]There is an article on continued fractions in this issue of *Quantum.*

mensurable sides arose as early as in ancient Babylon. That the sides are commensurable means that there exists a scale in which the legs and the hypotenuse are expressed by integer numbers x, y, and z. Then,

$$x^2 + y^2 = z^2.$$

Thus the Babylonian problem reduces to the problem of constructing all triples of positive integers x, y, and z satisfying the above equation. The Pythagoreans knew a method for constructing all solutions to this equation. It is possible that this method had been found even earlier, in ancient Babylon or India. In any case, solutions (x, y, z) to the equation $x^2 + y^2 = z^2$ are called *Pythagorean triples*[3].

The problem of finding solutions of a Diophantine equation (even if the equation itself looks simple) is often difficult. There is no uniform method (no *algorithm*) for finding out whether a Diophantine equation has integer solutions or not. Analysis of particular Diophantine equations continues to this day; in recent years, considerable progress has been made with the help of the complicated formalism of modern algebraic geometry. However, the Markov equation can be solved by elementary means.

Family tree of the Markov equation

An ordered triple of integers (a, b, c) is called a solution to a Diophantine equation in the unknowns x, y, z if this equation is converted into a valid numerical equality by the substitution $x = a$, $y = b$, and $z = c$. The numbers a, b, and c of the solution (a, b, c) will be called the *coordinates* of the solution. For the Markov equation (1) we will consider only nonzero solutions (it is easy to see that if one of the coordinates of a solution to (1) is zero, all the other coordinates must be zero as well).

The left-hand side of (1) is positive for any solution (a, b, c); thus, either all a, b, and c are positive or two of them are negative. In the latter case, $(|a|, |b|, |c|)$ is a solution to (1) with positive coordinates. Conversely, if (a, b, c) is a solution with positive coordinates, then by changing the sign of two arbitrary coordinates, we obtain another solution. For this reason, without loss of generality, we will consider only solutions with positive coordinates.

It follows from the symmetry of equation (1) that if (a, b, c) is a solution, then

$$(a, b, c), (c, a, b), (b, c, a),$$
$$(b, a, c), (a, c, b), (c, b, a)$$

are solutions as well. That is, triples obtained from a solution (a, b, c) by various permutations of the coordinates are also solutions.

Thus we may consider all six solutions obtained from each other by permutations to be a single solution; that is, only the values of the coordinates are essential, and their order may be neglected.

The Markov equation (1) has the easily guessed solution $(1, 1, 1)$. Let us now see how, knowing one solution, we can find other solutions. If (a, b, c) is a solution to the Markov equation, then a is a root of the quadratic equation

$$F_a(x) = x^2 + b^2 + c^2 - 3bcx = 0.$$

By Vieta's theorem, this equation has another root $x = a'$ such that

$$a + a' = 3bc, \; aa' = b^2 + c^2. \qquad (2)$$

It is clear (from the second condition above) that $a' > 0$ and that (a', b, c) is also a solution to equation (1). It is called the *adjacent solution with respect to the coordinate a*. Clearly, if (a', b, c) is the adjacent solution to (a, b, c), then (a, b, c) is the adjacent solution to (a', b, c) with respect to the coordinate a'.

Similarly, we can obtain solutions adjacent with respect to the coordinates b and c.

Let us find the solution adjacent to the solution $(1, 1, 1)$. To do this we must solve the quadratic equation

$$x^2 + 1^2 + 1^2 - 3 \cdot 1 \cdot 1 \cdot x = 0.$$

In addition to the root $x = 1$, this equation has the root $x = 2$. Thus we have obtained another solution $(2, 1, 1)$. These two solutions, $(1, 1, 1)$ and $(2, 1, 1)$, play an important role. Following Markov, we shall call them the *singular solutions*.

The singular solutions are the only ones with two coordinates equal.

Problem 1. Prove that a solution (a, b, c) to the Markov equation is a singular solution if and only if two of its coordinates are equal.

The first singular solution, $(1, 1, 1)$, has only one adjacent solution. The second singular solution has two adjacent solutions: $(1, 1, 1)$ and $(2, 5, 1)$. The latter solution is adjacent to $(2, 1, 1)$ with respect to the coordinate 1 and is obtained from the equation

$$2^2 + y^2 + 1^2 = 3 \cdot 2 \cdot y \cdot 1.$$

In turn, the solution $(2, 5, 1)$ has three adjacent solutions: one of them is, naturally $(2, 1, 1)$, and the other two are $(13, 5, 1)$ and $(2, 5, 29)$. Generally, every nonsingular solution (a, b, c) generates three adjacent solutions

$$(a', b, c), (a, b', c), (a, b, c'),$$

where (compare with (2))

$$a' = 3bc - a, \; b' = 3ac - b,$$
$$c' = 3ab - c.$$

Problem 2. Prove that if the solution (a, b, c) is nonsingular, one of its adjacent solutions has a smaller maximum coordinate, and the other two have a greater maximum coordinate.

Markov's theorem. *Any solution to equation (1) can be linked by a chain of adjacent solutions to the singular solution* $(1, 1, 1)$.

Proof. Let (a, b, c) be a nonsingular solution to equation (1). Then it has an adjacent solution (a_1, b_1, c_1) with a smaller maximum coordinate (problem 2). If this solution is also nonsingular, it has an adjacent solution (a_2, b_2, c_2) with a still smaller maximum coordinate, and so on. But positive integers cannot form an infinite decreasing sequence. Thus, this process inevitably terminates. The terminal solution (a_n, b_n, c_n) will have equal coordinates; that is,

[3]Pythagorean triples were discussed in the article "As easy as (a, b, c)?" in *Quantum*, January/February 1999.

the terminal solution is singular (Problem 1). If this solution is (1, 1, 1), then the proposition is demonstrated; otherwise, it is (2, 1, 1), which is known to have the adjacent solution (1, 1, 1). Thus the theorem is proved.

It follows from Markov's theorem that, starting from the singular solution (1, 1, 1) and successively going to adjacent solutions with a greater maximum coordinate, we can obtain all solutions. In the process, we obtain a *family tree* (see figure 1).

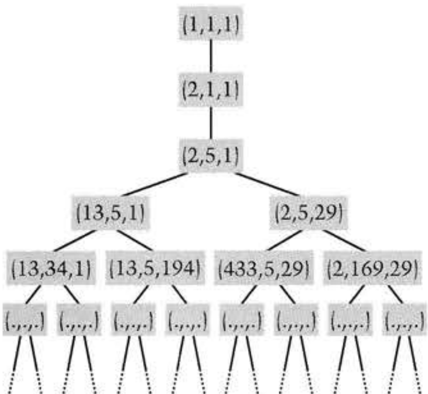


Figure 1

Using this tree, given $N \geq 1$, we can find, in a finite number of steps, all the solutions to the Markov equation whose coordinates do not exceed N.

Problem 3. Prove that coordinates of any solution to the Markov equation are pairwise relatively prime.

The exclusivity of the Markov equation

Consider the following problem, which may seem rather strange at first glance: *if the sum of the squares of three positive integers is divisible by their product, what values may the quotient have?*

This problem is equivalent to the following: for which positive integers k does the Diophantine equation

$$X^2 + Y^2 + Z^2 = kXYZ \qquad (3)$$

have a nonzero solution? For $k = 3$, this equation is the Markov equation. It is easily seen that equation (3) has solutions for $k = 1$; for example, (3, 3, 3). Hurwitz and Frobenius analyzed equation (3) and proved that it has solutions only for $k = 3$ and $k = 1$. This result can be obtained by elementary means.

First, consider the case $k = 1$. It turns out that the search for solutions in this case reduces to solving the Markov equation.

Problem 4. Let A, B, and C be positive integers. Then the remainder upon division of $A^2 + B^2 + C^2$ by 3 equals the number of numbers among A, B, and C that are not divisible by 3 if there are fewer than three of these; otherwise, the remainder is zero.

Problem 5. Prove that all of the solutions to the equation

$$X^2 + Y^2 + Z^2 = XYZ \qquad (5)$$

are given by the formulas

$$A = 3a,\ B = 3b,\ C = 3c,$$

where (a, b, c) is an arbitrary solution to the Markov equation

$$x^2 + y^2 + z^2 = 3xyz. \qquad (6)$$

Now let's turn to the case $k = 2$.

Problem 6. Let A, B, and C be positive integers. Prove that the remainder upon division of $A^2 + B^2 + C^2$ by 4 equals the number of odd numbers among A, B, and C.

Problem 7. Prove that equation (3) has no solutions for $k = 2$.

Theorem. *Equation (3) has a nonzero solution only for $k = 1$ and $k = 3$.*

Proof. For $k = 1$ the solutions are obtained as described in Problem 5. For $k = 2$, equation (3) has no solutions, as was proved in Problem 7. Consider the case $k > 3$.

Suppose that equation (3) has a solution (a, b, c) for a certain $k > 3$. We prove that its coordinates a, b, and c must be pairwise distinct. For example, let $b = c$. Then $a^2 = kab^2 - 2b^2 = (ka - 2)b^2$; thus, $a = bd$, where d is an integer. Therefore, we have $b^2d^2 = (kbd - 2)b^2$, $d^2 = kbd - 2$, and $2 = d(kb - d)$. Thus 2 is divisible by d, and therefore $d = 1$ or $d = 2$. In both cases, $kb = 3$, which contradicts the condition $k > 3$.

Thus, for $k > 3$, any solution to equation (3) has pairwise distinct coordinates. Without loss of generality, we may assume that $a > b > c$. Using the quadratic trinomial

$$P(x) = x^2 + b^2 + c^2 - kxbc,$$

we can find for the solution (a, b, c) an adjacent solution (a', b, c) with respect to the coordinate a. Since

$$P(b) = 2b^2 + c^2 - kb^2c < 3b^2 - kb^2c \leq 3b^2 - kb^2 < 0,$$

we see that b lies between the roots a and a' of the polynomial (x); that is, $a > b > a'$. Therefore, the maximum coordinate of the solution (a', b, c) is less than that of the solution (a, b, c). Thus, for every solution (a, b, c), we can find a solution (a_1, b_1, c_1) with a smaller maximum coordinate. This construction can be repeated to obtain a solution (a_2, b_2, c_2) with a still smaller maximum coordinate. Since the coordinates of every solution are pairwise distinct, this process can be repeated infinitely to obtain an infinite sequence of solutions to equation (3) with a decreasing maximum coordinate. However, this is impossible, since the coordinates are positive integers. Thus the theorem is proved.

Corollary. *For any solution (a, b, c) to the Markov equation, the numbers a, b, and c are pairwise relatively prime.*

Proof. Suppose that a and b have a common divisor $d > 1$. By virtue of equation (1), d is also a divisor of c. Therefore, there exist numbers X, Y, and Z such that $a = dX$, $b = dY$, and $c = dZ$. By virtue of (1), we have $X^2 + Y^2 + Z^2 = 3dXYZ$, which contradicts the theorem just proved.

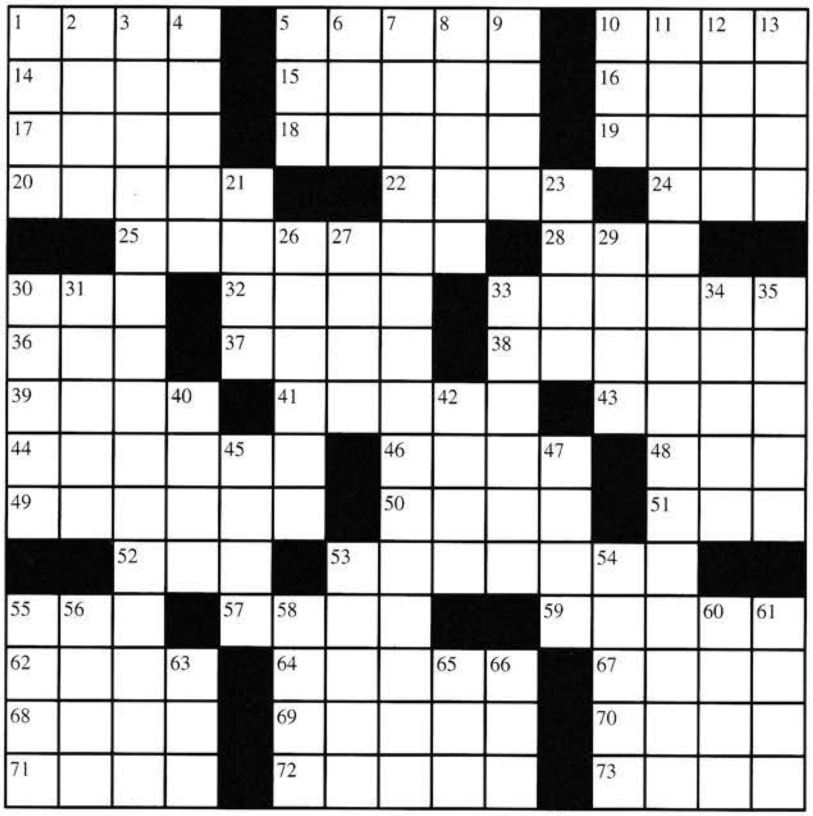
The following equation is a straightforward generalization of the Markov equation for the case of n variables $(n > 3)$:

$$x_1^2 + x_2^2 + \ldots + x_1^2 = nx_1x_2\ldots x_n. \qquad (7)$$

It is easy to see that some of the above propositions can be generalized for this case (there exists a singular solution $x_1 = 1, x_2 = 1, \ldots, x_n = 1$; for any solution, there exist adjacent solutions, and so on). However, I don't know of any place where a complete theory of equation (7) has been elaborated. The construction of such a theory could be a subject for a small independent research project.

Criss × cross science

by David R. Martin



ACROSS

1 Shoo
5 Drenches
10 1936 medicine Nobelist Sir Henry ___
14 Melody
15 Burgundy grape
16 Indium arsenide
17 Heart graphs
18 Aquatic plant
19 Ova
20 Planet 9
22 Arm bone
24 Employ
25 Arch plate
28 ___ de mer
30 Bible book: abbr.
32 Concert halls
33 Growing in pairs
36 Patriotic group: abbr.
37 Olfactory sensation
38 County near Belfast
39 Pollster ___ Roper (1900-1971)
41 Alloy used in watches
43 Swedish botanist ___ Afzelius (1750-1837)
44 Orbital extremum
46 Zeus' shield (alt. sp.)
48 Anger
49 Kaph followers
50 White: comb. form
51 Swiss river
52 3818 (in base 16)
53 Element 67
55 Trig. function
57 Onde de ___
59 Distance/time
62 Invertebrate hair
64 Nimble
67 Type of engine
68 Caked deposit
69 Despiser
70 Element 10
71 ___ black (carbon pigment)
72 Two dimensional quantum particle
73 1948 Chemistry Nobelist ___ Tiselius

DOWN

1 ___-down transformer
2 Cuprous chloride
3 $I\omega$
4 Seed coat
5 Health resort
6 Mixture of hydrocarbons
7 ω
8 Australian tree-dweller
9 British gun
10 Threading device
11 Astronomer's measurement
12 Trails
13 Exist (in France)
21 Melville opus
23 60 Coulombs: abbr.
26 Murphy and Van Halen
27 Ponce de ___
29 Pilaster at the end of a wall
30 Type of gas
31 Oceanic tunicate
33 Element 56
34 Pope's crown
35 Type of wheat
40 Molding shape
42 Gebang palm fiber
45 60,844 (in base 16)
47 Winter sport equipment
53 Golfer Ben
54 "Once ___ time ..."
55 Cesium chloride
56 10^{12}: pref.
58 Laugh sound
60 English college
61 Finished
63 Glycolysis chemical: abbr.
65 Physicist ___ Szilard
66 Sea eagle

SOLUTION IN THE NEXT ISSUE

SOLUTION TO THE NOVEMBER/DECEMBER PUZZLE

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

ANSWERS, HINTS & SOLUTIONS

Math

M281

The ratios of the distances traveled by the professor and her assistant at each encounter (or in any interval of time) is constant, and is equal to the ratio of their rates of walking. We can use this relationship to solve this problem. But what distances have they traveled at each encounter? They must meet "head-on" at the first encounter. But analyzing the second encounter is more complicated.

Let S be the distance between the two houses. We can distinguish three cases.

(a) At the time of the second encounter, each had traveled the entire route once and had started back. The ratios we mentioned give

$$\frac{55}{S-55}=\frac{S+85}{2S-85}.$$

Now the proportion $a/b = c/d$ is equivalent to the proportion $a/(a+b) = c/(c+d)$ (the reader can verify this, for example, by looking at the cross products). Applied to the present case, this means that we have $55/S = (S+85)/35$, or $165 = S + 85$. But then $S = 80$, and the second encounter could not have occurred 85 m. from the assistant's house. This case is not possible.

(b) The professor reaches the assistant's house and turns back, overtaking the assistant before he reaches her house. In this case the proportion gives

$$\frac{55}{S-55}=\frac{S+85}{85}.$$

This leads to a rather complicated quadratic equation.

However, we can also reason as follows. In this case, while the assistant walks 85 m, the professor walks twice that distance plus a bit more. So her speed must have been more than twice that of the assistant, and the distance traveled by the assistant at the time of the first encounter would have been less than half of the distance traveled by the professor; that is, less than $55/2 = 27.5$ m. Then the distance between their houses would be less than $55 + 27.5 = 82.5 < 85$ m, and this case is also not possible.

The same conclusion would have been reached by solving the complicated quadratic equation.

(c) The assistant reaches the professor's house and turns back, overtaking the professor before she reaches his house. In this case the proportion gives

$$\frac{55}{S-55}=\frac{S-85}{2S-85}.$$

Again, we use equivalent proportions. The reader is invited to check that $a/b = c/d$ is equivalent to $a/b = (c-2a)/(d-2b)$. It follows that each of the fractions in our proportion is equal to $(S-195)/25$. Therefore, by the time the assistant had traveled 25 m, the professor had traveled $S - 95$ m, and was at a distance of 195 m from the assistant's house. Thus the distance between the two stands is $195 - 25 = 170$ m.

M282

We take the base-2 logarithm of each member of the equation to be solved:

$$x\log_2 x=-\frac{1}{2},$$

Let us write a possible solution in the form

$$x=\frac{1}{2^m}.$$

Then the above equation becomes

$$2^{-m}\cdot(-m)=-2^{-1},$$

which simplifies to

$$m=2^{m-1}.$$

Some astute guessing shows that this last equation has two roots: $m_1 = 1$ and $m_2 = 2$. The respective values of x are $x_1 = 1/2$ and $x_2 = 1/4$.

Using some calculus, we can now prove that this original equation cannot have more than two roots. We can write the original equation in the form $x\cdot\ln x = a$, where $a = -(1/2)\ln 2$. Assume that the function $y = x\cdot\ln x$ takes some value a at three different points. Then its derivative, $y' = 1 + \ln x$, must vanish at at least two points, which is impossible. Thus the number of roots cannot be more than two, and we have found all the solutions.

Answer: $x_1 = 1/2$ and $x_2 = 1/4$.

M283

Figure 1 shows curves $y = \sin x$, $x = \sin y$, and $y = x$. On these curves, points $A(x_0, \sin x_0)$, $B(x_0, x_0)$, $C(\sin x_0, x_0)$, and $D(x_0, \arcsin x_0)$ are marked. It is clear that points C and D lie on the same curve $y = \arcsin x$, and that curves OA and OC are symmetric about the bisector OB of the angle between the axes. Now the derivative of $y = \sin x$ (curve OA) is $y' = \cos x \le 1$. Hence the slope of OA cannot exceed 1 at any point. It follows that the slope of curve OC (which is symmetric to OA with respect to line OB) cannot be less than

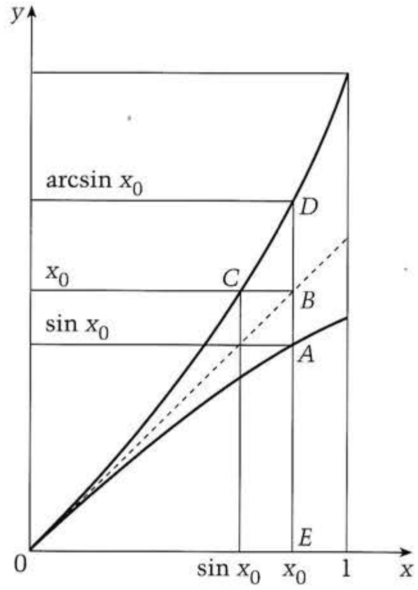


Figure 1

1 at any point. Therefore the slope of the segment CD is greater than 45°. Consequently, $BD > BC$ and $BD - AB = BD - BC$ is positive. Thus we have

$$\begin{aligned}\sin x_0 + \arcsin x_0 &= EA + ED\\ &= (EB - AB) + (EB + BD)\\ &= 2EB + (BD - AB) > 2EB = 2x_0.\end{aligned}$$

M284

Consider two cases (figure 2).

(1) Pairs of tangents that are symmetric about perpendicular OC (where C is the midpoint of AB) will intersect on this perpendicular (by symmetry). Conversely, for any point of the line OC not on diameter DE, one can draw a pair of tangents that meet line AB at points P and Q such that $PA = QB$. Thus a part of the desired locus consists of those points of the perpendicular OC which lie outside the circle. If AB is a diameter, this is the entire locus.

2) Consider an asymmetric pair of tangents LQM and PKM. We will show that the line connecting their points of tangency passes through C. Indeed, connect O with P and Q, and draw radii to the points of tangency. We obtain a pair of congruent triangles POK and QOL ($\angle PKO = \angle QLO = 90°$, $OK = OL$, and $PO = QO$). Since $\angle OKP = \angle OCP = 90°$, points P, O, C, and K lie on the same circle with diameter PO. Therefore, $\angle PCK = \angle POK$. Similarly, points L, O, C, and Q lie on the circle with diameter OQ, so $\angle LCQ = \angle LOQ$. Thus $\angle LCQ = \angle PCK$, and points K, L, and C lie on the same line. This proves that any chord that connects the points of tangency passes through the midpoint of chord AB.

Next we drop a perpendicular MN onto the extension of OC. Since MO subtends right angles at N, K, and L, all the points M, N, K, O, and L lie on the circle with diameter OM. Drawing this circle, we see that $\angle KNO = \angle OKL$, since they are subtended by equal chords $OL = OK$. Thus triangles OCK and OKN are similar. It follows that $ON = OK^2/OC$, a ratio which does not depend on the position of points P and Q. Thus for any pair of points P and Q, segment ON is of constant length, and the desired part of the locus is a line parallel to AB. In more advanced work, this line is called the *polar line* of point C.

Second solution. The line obtained in item (2) can be obtained in another way. Notice that $POQM$ is an inscribed quadrilateral, since $\angle OPM + \angle OQM = \angle OQL + \angle OQM = 180°$. Therefore $\angle OMP = \angle OQP = \angle OPC$ and $\angle OMN = \angle OMP + \angle PMN = \angle OPC + \angle QPM = \angle OPK$. These equalities imply the similarity of the following pairs of triangles: OMK and OPC, and OMN and OPK. Therefore, $OK/OM = OC/OP$ and $ON/OM = OK/OP$. Dividing the second equality by the first, we have $ON/OK = OK/OC$. Thus we have once more $ON = OK^2/OC$, which again does not depend on the positions of P and Q.

Answer: A line parallel to AB, plus the part of line OC lying outside the circle (where O is the center of the circle and C is the midpoint of chord AB).

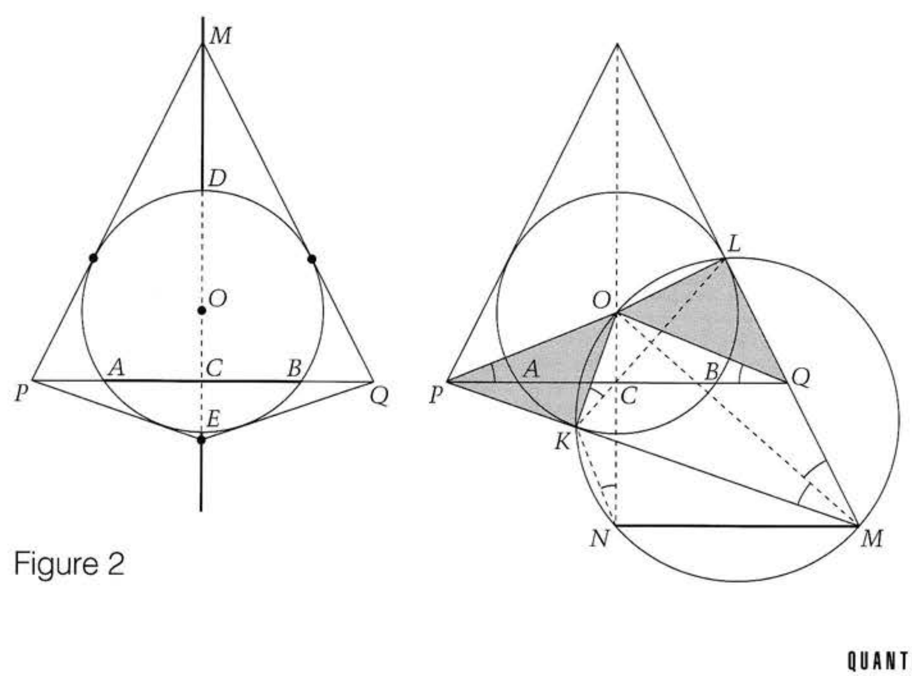


Figure 2

M285

Since D is the midpoint of BC, it follows that $S_{BMD} = S_{DMC}$ and $S_{PBD} = S_{PCD}$. Hence we have

$$\begin{aligned}S_{BMPC} &= S_{BMD} + S_{PDC}\\ &= S_{CMD} + S_{PDB} = S_{BPMC}.\end{aligned}$$

Now we recall the theorem that the area of a quadrilateral (convex or non-convex) is given by half the product of its diagonals and the sine of the angle they form. Thus

$$S_{BMPC} = S_{BPMC} = \frac{1}{2} BM \cdot CP \sin\alpha.$$

Let us set $\angle ABC = 2\beta$, $\angle ACB = 2\gamma$, and $\angle PDC = \phi$. Using the law of sines in triangle BMD, we find that

$$\frac{BM}{\sin\angle BDP} = \frac{BM}{\sin\phi} = \frac{a/2}{\sin\angle BMD},$$

or

$$BM = \frac{a\sin\phi}{2\sin(\phi - 2\beta)}.$$

Similarly, from the law of sines in triangle PCD, we have

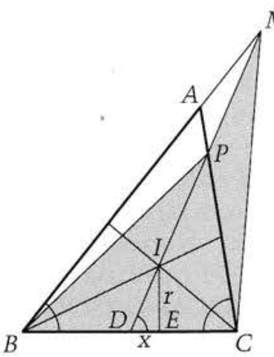


Figure 3

$$CP = \frac{a\sin\phi}{2\sin(\phi + 2\gamma)}.$$

Substituting these expressions in the formula for the area of quadrilateral $BMPC$, we obtain:

$$S_{BMPC} = \frac{a^2\sin^2\phi}{8\sin(\phi - 2\beta)\sin(\phi + 2\gamma)}. \quad (*)$$

Furthermore, from the formula for the sine of the sum of two angles, we have

$$\frac{\sin(\phi - 2\beta)}{\sin\phi} = \cos 2\beta - \sin 2\beta \cdot \cot\phi \quad (1)$$

and

$$\frac{\sin(\phi + 2\gamma)}{\sin\phi} = \sin 2\gamma \cdot \cot\phi + \cos 2\gamma. \quad (2)$$

We must now express $\cot\phi$ in terms of β and γ. In figure 3, I is the center of the circle inscribed in triangle ABC, x is the projection of ID onto BC, and r is the radius of the inscribed circle. Without loss of generality, we assume that $\beta \le \gamma$. We have:

$$\cot\beta = \frac{\frac{a}{2} + x}{r}, \quad \cot\gamma = \frac{\frac{a}{2} - x}{r}, \quad \cot\phi = \frac{x}{r}.$$

Now we subtract the second equation from the first to obtain

$$2\cot\phi = \cot\beta - \cot\gamma. \quad (3)$$

We use these relations to transform equations (1) and (2).

We obtain:

$$\frac{\sin(\phi - 2\beta)}{\sin\phi} = \cos 2\beta - \sin 2\beta \cdot \cot\phi$$

$$= \sin^2\beta\Big(\big(\cot^2\beta - 1\big) - \cot\beta\big(\cot\beta - \cot\gamma\big)\Big)$$

$$= \sin^2\beta\big(\cot\beta \cdot \cot\gamma - 1\big)$$

$$= \frac{\sin\beta\big(\cos\beta \cdot \cos\gamma - \sin\beta \cdot \sin\gamma\big)}{\sin\gamma}$$

$$= \frac{\sin\beta}{\sin\gamma}\cos(\beta + \gamma) = \frac{\sin\beta}{\sin\gamma}\sin\frac{\alpha}{2}$$

Similarly,

$$\frac{\sin(2\gamma + \phi)}{\sin\phi} = \frac{\sin\gamma}{\sin\beta}\sin\frac{\alpha}{2} \quad (2')$$

Substituting (1′) and (2′) for the corresponding expressions in (*), we find that the desired area is

$$\frac{a^2}{4}\cot\frac{\alpha}{2}.$$

Physics

P281

Let the ball fly through the hole with a velocity **v** in the reference frame in which the sphere was at rest before the first collision. Evidently, all subsequent motion will take place on the plane in which the center of the sphere and the vector **v** lie.

Since the masses of the ball and sphere are equal, the center of mass (CM) of the system will move with velocity **v**/2 along the straight line that lies in the same plane at the distance $R/4$ from the center of the sphere. Since no external force acts on the system, the motion of the CM will not change regardless of any collisions occurring inside the sphere.

From the statement of the problem, the collisions are perfectly elastic, and the interior surface of the sphere is smooth. Therefore, at a point of collision the tangential component of the ball's velocity remains constant, while the normal component changes sign after each impact. Therefore, a collision doesn't affect the magnitude of the ball's velocity relative to the sphere (it remains always equal to v), and the angles of incidence and reflection are equal. The ball's trajectory relative to the sphere follows an inscribed equilateral triangle with a side equal to $\sqrt{3}R$ (the red triangle in figure 4). The period between two successive collisions is $\sqrt{3}R/v$.

The CM of the system moves relative to the sphere with one-half the velocity, and its trajectory is also an equilateral triangle, but the length of its sides is only $\sqrt{3}R/2$ (the black triangle in figure 4).

Consider a uniformly moving reference frame in which the CM is at rest instead of the sphere. In this new reference frame the velocity of the ball and the center of the sphere are equal in magnitude (both are equal to $v/2$), and the motion takes place along the trajectories shown by the corresponding colors in figures 5 and 6.

Finally, we should return to the initial reference frame, in which the sphere was at rest before the first collision. In this frame the ball initially moves with velocity **v**, and the CM moves with a constant velocity **v**/2 (figure 7). To perform this trans-

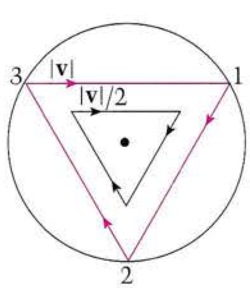


Figure 4

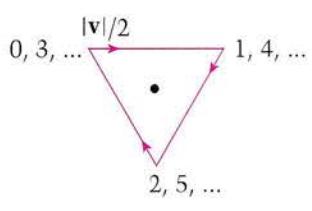


Figure 5

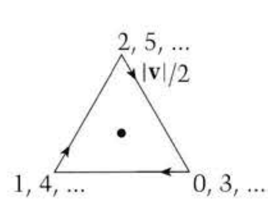


Figure 6

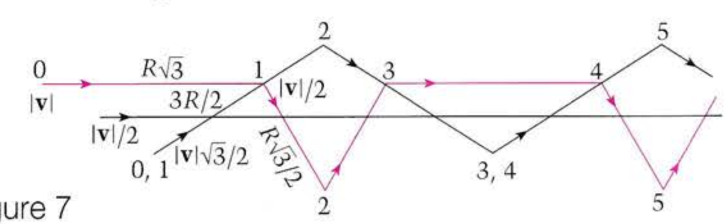


Figure 7

formation, we must add a vector $\mathbf{v}/2$ to all the velocity vectors drawn in figures 5 and 6. We find that before the first collision (or after the virtual "zeroth" collision and then after the 3rd, 6th,... $3n$th collisions) the ball moves with a speed v which is directed parallel to the velocity (and trajectory) of the CM. During the time until the next collision the ball travels a distance $\sqrt{3}R$, while the sphere does not move at all.

After the first (and 4th, 7th, ... $(3n+1)$st) collision the velocity of the ball is directed at an angle of $(-\pi/3)$ to the CM velocity. Its speed is equal to $v/2$. Until the second collision the ball travels a distance of $\sqrt{3}R/2$. Correspondingly, the velocity of the sphere is directed at an angle of $\pi/6$ to the same vector, and its speed is $\sqrt{3}\,v/2$. On this leg the sphere travels a distance $3R/2$.

After the second (and 5th, 8th, ... $(3n+2)$nd) collision the ball's velocity is directed at an angle of $\pi/3$ to the velocity of the CM. Both vectors have the same speed, equal to $v/2$. The distance traveled by the ball is the same as before, that is, $\sqrt{3}R/2$. The inclination of the velocity of the sphere on this leg is $-\pi/6$, and the speed and distance traveled are $\sqrt{3}\,v/2$ and $3R/2$, respectively.

Subsequently the process is repeated: the sphere is at rest between the third and the fourth collisions, while the ball is again moving with velocity $\mathbf{v}$, and so on.

P282

The well-known spherical shape of a soap bubble results from surface tension. Indeed, if the potential energy of the soap film in the gravitational field is negligible in comparison with its surface energy (this is correct for a thin film), the observed geometry of a soap bubble is determined by the surface energy alone. Since the volume of air trapped inside the bubble is constant, the bubble should have the minimal surface area for a constant volume, that is, it must be a ball. Therefore, in equilibrium the soap film forms a spherical surface.

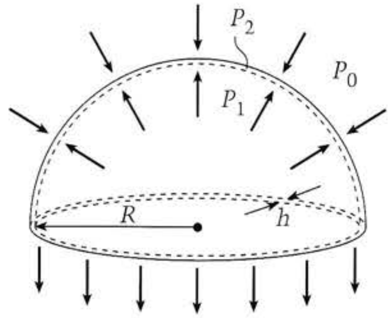


Figure 8

Let's mentally divide the spherical shell into two equal halves and consider one of them (figure 8). Our aim is to formulate equilibrium conditions for the external and internal spherical surfaces of the upper hemisphere.

The external spherical surface of radius R is affected by the downward force of the external pressure $F_1 = P_0\pi R^2$(calculate it on your own), the downward force of surface tension $F_2 = \sigma \cdot 2\pi R$, and the counterbalancing upward force of the compressed liquid (the curved liquid film) $F_3 = P_2\pi R^2$. In equilibrium,

$$F_1 + F_2 - F_3 = 0,$$

or

$$P_0\pi R^2 + 2\pi R\sigma = P_2\pi R^2,$$

which gives us the pressure inside the film:

$$P_2 = P_0 + \frac{2\sigma}{R}.$$

The internal spherical surface (of radius $R - h$) is affected by the downward force of the compressed liquid $F_4 = P_2\pi(R-h)^2$, the downward force of surface tension $F_5 = \sigma \cdot 2\pi(R-h)$, and the upward force of air pressure inside the bubble $F_6 = P_1\pi(R-h)^2$. The equilibrium condition for the internal hemisphere is

$$F_4 + F_5 - F_6 = 0,$$

which gives us the air pressure inside the bubble:

$$P_1 = P_2 + \frac{2\sigma}{R-h},$$

or

$$P_1 = P_0 + 2\sigma\left(\frac{1}{R} + \frac{1}{R-h}\right).$$

P283

Since the capacitors are connected in series, the charges on them are equal before the switch is closed:

$$q = \frac{C_1C_2}{C_1+C_2}\mathscr{E}.$$

After the switch was closed, the voltage across each capacitor becomes $\mathscr{E}/2$, so their charges are

$$q' = \frac{1}{2}C_1\mathscr{E} \text{ and } q'' = \frac{1}{2}C_2\mathscr{E}.$$

Therefore, the following charges flow in the indicated directions through the cross sections I–I and II–II:

$$q_1 = q' - q = \frac{C_1(C_1-C_2)}{2(C_1+C_2)}\mathscr{E}$$
$$= -2.4 \cdot 10^{-5}\text{ C}$$

and

$$q_2 = q - q'' = \frac{C_2(C_1-C_2)}{2(C_1+C_2)}\mathscr{E}$$
$$= -3.6 \cdot 10^{-5}\text{ C}.$$

The total charge on the interconnected plates of the capacitors was zero before the switch was closed, while after the closing it becomes $q'' - q'$. Clearly this charge flows through the cross section III–III, that is

$$q_3 = q'' - q' = \frac{1}{2}(C_2 - C_1)\mathscr{E}$$
$$= 6 \cdot 10^{-5}\text{ C}.$$

P284

When the leads are connected in one of the possible ways, the voltage across the capacitor will always remain zero. In that case the current in the left part of the circuit is zero, and all the current flows through the right coil. Therefore, the lamp will become dim monotonically as the frequency of the alternating current

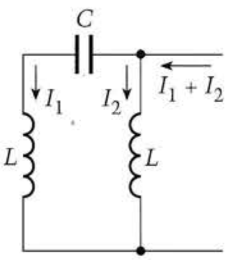


Figure 9

rises, because the inductive impedance of a coil is proportional to the frequency.

If we interchange the leads of either coil, the voltages across the coils remain the same, but now they add instead of subtract. We analyze the resulting two-pole circuit in the following way (figure 9). We apply an ac voltage $V(t) = V_0 \cos \omega t$ to its poles and determine the current supplied by the source. The voltage drop across one of the coils equals the voltage of the source. Since the magnetic fluxes threading both coils are identical, the voltage drop across the second coil will also be equal to the source voltage. Therefore, twice the voltage is applied to the capacitor, and the current $I_2 = -2V_0\omega C \sin \omega t$. The field threading the turns of each coil is determined by the difference of their currents (this follows from the directions chosen for the currents in figure 9):

$$L(I_1' - I_2') = V_0 \cos \omega t.$$

Integration of this differential equation yields

$$I_1 - I_2 = \frac{V_0}{\omega L} \sin \omega t.$$

The total current is

$$I_1 + I_2 = (I_1 - I_2) + 2I_2$$
$$= V_0\left(\frac{1}{\omega L} - 4\omega C\right)\sin \omega t,$$

and the impedance assumes

$$Z = \frac{1}{\frac{1}{\omega L} - 4\omega C} = \frac{\omega L}{1 - 4\omega^2 LC}.$$

Such a dependence of the impedance Z on the frequency is characteristic of a parallel circuit with a coil of inductance L and a capacitor of capacitance $4C$. Evidently, at low frequencies the two-terminal network behaves like a coil, while at higher frequencies it looks like a capacitor. At the resonance frequency $\omega_r = 1/(2\sqrt{LC})$ the impedance of the two-terminal network assumes very high values. Therefore, the lamp is very dim. At higher frequencies the lamp will increase in brightness.

P285

Denote by OC the line connecting the eye with the center of the tube. Point A is the end point of the inner diameter of the tube (figure 10). The distance from the virtual image of point A to axis OC equals the apparent radius r of the inner channel.

One of the laws of geometrical optics states that all the rays diverging from point A at small angles will travel after refraction in such a way that they appear to come from a virtual point source A^*. Since we are not interested in the exact position of this point, but only in its distance from OC, it will be sufficient to consider the trajectory of only one ray, ABD, which travels parallel to OC after emerging from the tube. Indeed, the virtual image of point A will lie somewhere on the extension of this ray (figure 10).

The distance between the rays BD and OC is r, and the distance we are looking for is $AO = r_0$. Since the outer diameter of the tube is much greater than its inner diameter, the angle of incidence α will be small. From triangle OEB we obtain $r = R\alpha n$ (R is the outer diameter of the tube), and from triangle OAB we obtain $r_0 = R\alpha$. Thus $r_0 = r/n$.

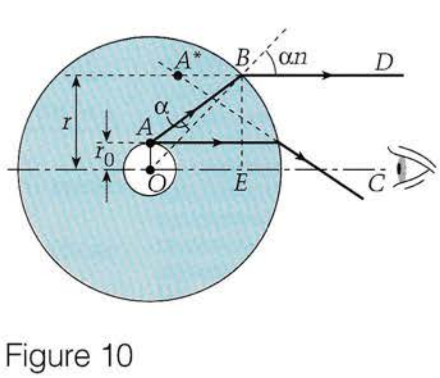


Figure 10

Brainteasers

B281

The minimum possible number of operations is 3. For example, we can turn glasses 1, 2, 3, 4, and 5 first; then turn glasses 1, 2, 3, 6, and 7; and, finally, turn glasses 1, 2, 3, 8, and 9. Two moves cannot be enough, because if only two moves are made, some glass must be turned exactly twice, and so will not remain upside down.

B282

It is impossible to turn 9 glasses by turning 6 glasses at a time. Indeed, every glass must be turned an odd number of times. Since there are 9 glasses, the total number of operations must be odd. On the other hand, if we turn an even number of glasses each time (6), the total number of operations is also even. This contradiction proves our assertion.

B283

This number is divisible by 3, but not by 9 (since the sum of its digits is divisible by 3, but not by 9). Therefore it cannot be a perfect square.

B284

One possible construction is illustrated in figures 11 and 12. In the first case (fig. 11), we draw two arbitrary circles centered on the given line and passing through point A and find the second point of intersection of these circles (A'). Line AA' is the

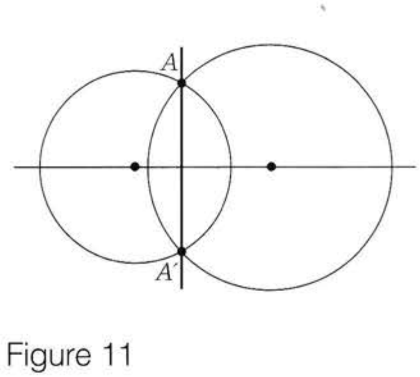


Figure 11

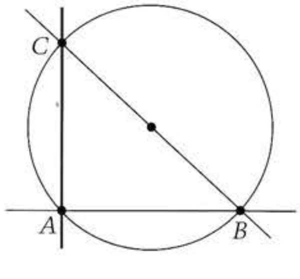


Figure 12

desired perpendicular. In the second case (fig. 12), we draw an arbitrary circle whose center lies off the given line and which passes through point A; the other point of intersection of this circle with the given line is a point B. Then we draw the diameter BC. Line AC is the desired perpendicular.

B285

Snow is a very poor conductor of heat, and its specific heat of fusion is very large. It therefore melts very slowly and can last for a long time at temperatures not much above 0 °C.

Kaleidoscope

1. Three.

2. When an excited atom emits a photon, its potential energy decreases.

3. No.

4. Alpha particles:

$$^{11}_{5}\mathrm{B} + {}^{1}_{1}\mathrm{p} = 3\,{}^{4}_{2}\mathrm{He}$$

5. The charge of a nucleus determines the chemical properties of the atom. For example, the nuclear charge is not changed during gamma emission.

6. When the period of observation is small compared to the half-life of the radioactive substance.

7. Three half-lives.

8. The energy of a nucleus can assume only discrete values.

9. The energy of an alpha particle is not sufficient to overcome the repulsion force of a heavy nucleus, which carries a large electric charge.

10. In addition to electrons, neutrinos are also emitted during beta decay. They carry away part of the energy, the amount being widely variable.

11. Even in the first transuranic elements the Coulomb repulsion of the constituent protons leads to instability of the nuclei.

12. The energy imparted to an atom in a collision with a neutron is higher for lighter atoms.

13. No. The missing mass is carried away by gamma rays emitted during the formation of the nucleus.

Microexperiment

The valence electrons in metals are easily transferred to an excited state by absorbing thermal energy. By the same token, they are ready to return to the ground state, giving up the extra energy in the form of photons (light). In glass, however, all the electrons are tightly bound to the nuclei of the atoms, and it is therefore very difficult to change their energy state. Much higher temperatures are needed to obtain visible luminescence of a glass. ◘

CONTINUED FROM PAGE 27

that is, $x_0 = 649$, $y_0 = 180$. The general solution is

$$x + y\sqrt{13} = \left(649 + 180\sqrt{13}\right)^k.$$

Problem 14. Find the least solution to the equation

$$x^2 - 61y^2 = 1.$$

Answer: $y_0 = 226{,}153{,}980$.

Note that the name of the English mathematician J. Pell (1610–1685) was related to equation (6) as a result of an error on Euler's part. Before Pell, this equation was studied by his compatriots J. Wallis and W. Brouncker, and by the French mathematician P. Fermat. As to the Archimedes cattle problem, it was solved only in 1880. The least solution to equation (*) contains 41 digits, and the total number of cattle is an astonishingly large number, on the order of $10^{206,545}$.

The equation $x^2 + y^2 = p$

It can be proved that *if* p *is an odd prime of the form* $4k + 1$, *then the period of the continued fraction for* $\sqrt{p}$ *is odd.* It follows from prob. 13 that the expansion is

$$\sqrt{p} = \left[a_0; \overline{a_1, \ldots, a_m, a_m, \ldots, a_1, 2a_0}\right].$$

Denote by α_{m+1} the number having the following continued fraction expansion:

$$\alpha_{m+1} = \left[\overline{a_m; a_{m-1}, \ldots, a_1, 2a_0, a_1, \ldots, a_m}\right].$$

Since this fraction is purely periodic, α_{m+1} is a reduced number, and by virtue of prob. 10

$$\alpha_{m+1} = \frac{A + \sqrt{p}}{B},$$

where $A > 0$ and $B > 0$. It follows from the assertion of problem 12 that $-1/\alpha'_{m+1}$ has the same continued fraction expansion as α_{m+1}. It follows from the uniqueness of the continued fraction that

$$\alpha_{m+1} = -\frac{1}{\alpha'_{m+1}};$$

in other words, $\alpha_{m+1} \cdot \alpha'_{m+1} = -1$. Since

$$\alpha'_{m+1} = \frac{A - \sqrt{p}}{B},$$

then

$$\frac{A^2 - p}{B^2} = -1$$

or $A^2 + B^2 = p$.

This reasoning gives an algorithm for finding the integer solution to the equation

$$x^2 + y^2 = p. \qquad (9)$$

It can be proved that *such a solution is unique (up to the interchange of x and y) and that equation (9) has no integer solutions for primes p of the form* $4k + 3$.

Example 5. Find the integer solution to the equation

$$x^2 + y^2 = 1009.$$

The continued fraction expansion of $\sqrt{1009}$ is

$$\sqrt{1009} = [31; \overline{1, 3, 3, 1, 62}].$$

Therefore,

$$\alpha_1 = \frac{31 + \sqrt{1009}}{48}, \; \alpha_2 = \frac{17 + \sqrt{1009}}{48},$$

$$\alpha_3 = \frac{28+\sqrt{1009}}{15}.$$

Thus, the pair of numbers $x = 28$, $y = 15$ is the desired solution. This method of solving equation (9) was invented by the French mathematician Legendre (1808).

Problem 15. Find the integer solution to the equation

$$x^2 + y^2 = 1129.$$

The equation $x^2 - Dy^2 = -1$

If the period, $n + 1$, of a continued fraction representing $\sqrt{D}$ is odd, then equation (8) implies that the numbers

$$x_0 = p_n,\ y_0 = q_n$$

are solutions to the Diophantine equation

$$x^2 - Dy^2 = -1. \qquad (10)$$

All solutions to this equation can be obtained from the formula

$$x + y\sqrt{D} = \left(x_0 + y_0\sqrt{D}\right)^{2k+1},$$

$$k = 0, 1, 2, \ldots .$$

For example, the solutions to

$$x^2 - 13y^2 = -1$$

are given by the formula

$$x + y\sqrt{13} = \left(18 + 5\sqrt{13}\right)^{2k+1},$$
$$k = 0, 1, 2, \ldots .$$

It can be proved that *if the period of the continued fraction expansion of $\sqrt{D}$ is even, then equation (10) has no integer solutions.*

Length of the period of the continued fraction of $\sqrt{D}$

The speed of the algorithms described above depends on the length of the period of the continued fraction of $\sqrt{D}$. However, very little is known about this dependence. The period length is very peculiar. For example, for $D = 986{,}045 = 5199991$, the expansion

$$\sqrt{986045} = \left[992;\ \overline{1, 495, 2, 495, 1, 1984}\right]$$

has a period of length 6. However, the period length of the expansion of $\sqrt{20\,989}$ is 205.

It is known that for all D, the period length doesn't exceed

$$4\sqrt{D}\ln D.$$

On the other hand, it can be proved that numbers $D = 5^{2k+1}$ have periods not less than

$$\frac{1}{3}\sqrt{D}(\ln D)^{-1}.$$

Thus the period increases rapidly with k. A vast amount of numerical material supports the hypothesis that there exist infinitely many numbers D which are not divisible by the square of any integer and for which the period length of the continued fraction for $\sqrt{D}$ is greater than $D^{1/2-\varepsilon}$ for any fixed positive ε.

The fact that the period of the continued fraction of $\sqrt{p}$ is odd for primes p of the form $4k + 1$ was proved by Legendre in 1785. Later, the German mathematician Dirichlet proved a similar theorem for $D = p \cdot q$, where p and q are primes satisfying certain conditions. Only quite recently, in 1980, the American mathematician Lagarias found an algorithm for determining whether equation (10) has integer solutions or not (or, equivalently, whether the period of the continued fraction of $\sqrt{D}$ is odd) in about $(\log D)^{5+\varepsilon}$ operations.

Continued fractions for particular numbers

We already know that quadratic surds and only quadratic surds have expansions in periodic continued fractions. Thus, the structure of the continued fractions of quadratic surds is rather well defined. It is natural to wonder what other classes of surds have well-structured expansions in continued fractions. However, we don't have a satisfactory answer. For example, it is not yet known whether the partial quotients $a_0, a_1, a_2, \ldots, a_k, \ldots$ of the continued fraction expansion of $\sqrt[3]{2}$,

$$\sqrt[3]{2} = a_0 + \frac{1|}{|a_1} + \frac{1|}{|a_2} + \ldots,$$

are bounded or not.

Several thousand of the first numbers $a_0, a_1, \ldots$ were calculated by computer. Here is the beginning of the sequence:

$\sqrt[3]{2}$ = [1; 3, 1, 5, 1, 1, 4, 1, 1, 8, 1, 14, 1, 10, 2, 1, 4, 12, 2, 3, 2, 1, 3, 4, 1, 1, 2, 14, 3, 12, 1, 15, 3, 1, 4, 534, 1, 1, 5, 1, 1, 121, 1, 2, 2, 4, 10, 3, 2, 2, 41, 1, 1, 1, 3, 7, 2, 2, 9, 4, 1, 3, 7, 6, 1, 1, 2, 9, 2, 3, 3, 1, 1, 69, 1, 12, ...].

It is seen from this expansion that the sequence $a_0, a_1, \ldots$ behaves like a bounded sequence, and only isolated overshoots (for example, 534 or 121) disturb the picture.

Moreover, sufficient information about the continued fraction expansion does not exist for even one algebraic number that is not a quadratic surd (an algebraic number is one that is a root of a polynomial with integer coefficients).

Of particular interest are the continued fraction expansions of certain classical constants. There are very few of these for which the continued fraction expansion is known.

L. Euler found the following continued fraction expansion for e

$$e = 2 + \frac{1|}{|1} + \frac{1|}{|2} + \frac{1|}{|1} + \frac{1|}{|1} + \frac{1|}{|4} + \ldots, \qquad (11)$$

where the partial quotients $a_0, a_1, \ldots$ are given by the formula

$$a_0 = 2,\ a_1 = 1,\ a_2 = 2,\ a_3 = 1,\ a_4 = 1,$$
$$a_5 = 4, \ldots,$$
$$a_{3m} = a_{3m-2} = 1,\ a_{3m-1} = 2m,$$
$$(m = 1, 2, \ldots).$$

Recall that

$$e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{\infty}\frac{1}{k!}$$

$$= 2.718281828459045\ldots$$

Although not elementary, the derivation of formula (11) is not difficult. A similar "good" expansion of π in a continued fraction is not known.

In this article, we have discussed only a small fraction of the problems in which continued fractions are useful. They have a very wide range of applications, especially if functions defined by continued fractions are considered. This is a very rich and complex branch of mathematics that is still far from being completed.

Cantor Cheese

by Don Piele

WELCOME TO A NEW COLUMN THAT WILL explore problems in Informatics—the science of applying computer algorithms to model and manage the world we live in. Informatics is a term commonly used in Europe to represent what is called Computer Science in the United States. However, Informatics is actually a broader term that has been adopted by fields outside of computer science. Many areas of science are now using computer algorithms to model and manage their world. Bioinformatics is an excellent example. It refers to the fast growing area in molecular biology of using computers to build models of molecules and to write computer programs for high speed genomic mapping and sequencing.

One of the goals of this column is to seek out and present interesting computer algorithms that are simple to understand and write. Another goal is to get young readers hooked on programming as a problem solving tool of the 21st Century. Being able to develop algorithms is similar to being able to construct good mathematical proofs. Both skills require the ability to take a set of givens, sequence together small logical steps, and end up with the desired goal. The best way to learn both programming and mathematics is to "just do it." Of course it helps a great deal to be curious about a problem and want to know what makes it tick. Hopefully, the problems presented will pique your curiosity.

I will often include graphics, because pictures are fun to make and inherently interesting. I will be working with a high level programming tool called *Mathematica*. This general purpose computer-programming environment integrates several features into a unified framework: numerical and symbolic calculations, functional, procedural, rule-based, and graphics programming, as well as animations. With *Mathematica*, how we think about a problem and how we code it are reasonably close. Young people, with no preconceived notion of what a programming language is, should find it natural. At least that's the outcome I'm shooting for. So let the games begin.

Recursion

Recursion is an indispensable tool in Informatics. Its well-known mathematical cousin is called induction. Proofs in mathematics by induction require one to prove two steps: problemSolved[1], and problemSolved[n] given you have problemSolved[$n-1$]. If you can prove these two steps, then problemSolved[n] is true for all n. In informatics, we perform a similar process. We first define problemSolved[1]. Then, assuming problemSolved[k], $k < n$ is done, we define problemSolved[n]. Let's look at a simple example of a recursive definition to see how it works. Our task is to define SUM[n] recursively as the sum of the positive integers from 1 to n.

Recursive definition of the SUM of positive integers

```
(* define SUM from 1 up to 1 *)
SUM[1] = 1;
(* define SUM[n] in terms of SUM[n - 1]
and n *)
SUM[n_] := SUM[n - 1] + n ;
(*compute SUM[100]*)
SUM[100]
5050
SUM[255]
32640
SUM[256]
$RecursionLimit::reclim : Recursion depth
of 256 exceeded.
32895 + Hold[SUM[2 - 1]]
```

If you try to compute SUM[256] in *Mathematica* you will not get an answer, because you have run into the default recursion limit. This is a service to you in case you define a recursive function badly and it runs on indefinitely. This can easily happen, for example, if you forget to assign a value to SUM[1]. By default, *Mathematica* will go back 255 steps in a recursive definition but not more. You can change the recursion limit in *Mathematica* by giving the system variable $RecursionLimit a new value.

```
$RecursionLimit = Infinity;
SUM[1000]
500500
```

Many efficiencies can be built into recursive definitions. For example, you can remember what you have computed so you don't waste time doing it again. (This, by the way, is the secret behind dynamic programming—a topic for another column.) Now let's turn to graphics designs done with recursion.

Recursive designs

Recently I was reading the book "Computers and the Imagination," by Clifford A. Pickover, when I came across an intriguing graphic with the name Cantor Cheese. Being from Wisconsin, cheese will always catch my attention. You may recall the famous Cantor set, described by Cantor (1845–1918), that is constructed by taking the unit interval [0, 1] and removing its open middle third interval (1/3, 2/3). This leaves two smaller intervals {[0, 1/3], [2/3, 1]}, each one third the length of the original interval. Continue deleting the middle one third from these intervals and repeat the process. Let Cantor[n] be the set of 2^{n-1} closed intervals remaining after n extractions of the middle one third. The Cantor set is equal to the Intersection of Cantor[n] for all n.

Cantor Cheese is made in a similar way, beginning with a unit disk and removing everything *except* two interior subdisks. An example is shown in figure 1 where each of the interior subdisks is shrunk in radius by half. Let's see how this graphic was created using recursion.

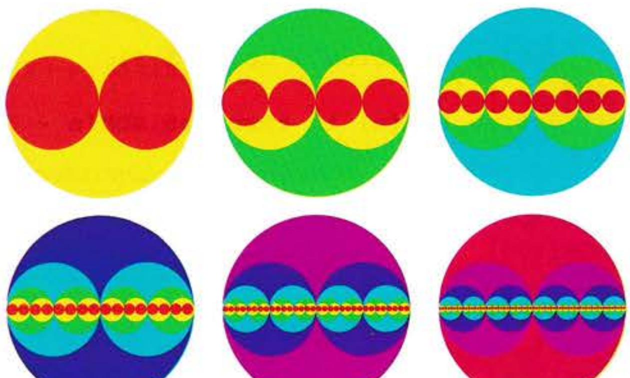


Figure 1

First a bit of housekeeping. The following command in *Mathematica* makes sure that circles are drawn correctly and that we reserve a region in the plane for our cheese.

```
SetOptions[Graphics, AspectRatio →
Automatic, PlotRange → {(-1), 1}];
```

First we need to know how to make a simple disk of cheese with center {x, y} and radius r. This is done with the built-in Disk[{x,y} ,r] command that draws a disk with center {x, y} and radius r. Hue[0] is the color red.

```
CantorCheese[x_, y_, r_, 0, shrink_] :=
{Hue[0], Disk[{x, y}, r]}
```

The recursive step defines CantorCheese at step n in terms of the CantorCheese you know how to make at step $n - 1$. The shrink factor is under your control and should be ≤ 1/2. You should be curious about why CantorCheese[$x - r/2$, y, shrink*r, $n - 1$,shrink] and CantorCheese[$x + r/2$, y, shrink*r, $n - 1$,shrink] place the next disks on the a axes, one to the left and one to the right. Here is the recursive definition. Examine it carefully.

```
CantorCheese[x_, y_, r_, n_, shrink_]
:= {{Hue[ .16 n], Disk[{x, y}, r]},
CantorCheese[x - r/2, y, shrink*r, n - 1,
shrink], CantorCheese[x + r/2, y,
shrink*r, n - 1, shrink]}
```

```
Show[Graphics[CantorCheese[0, 0, 1, 6,
.5]]]
```

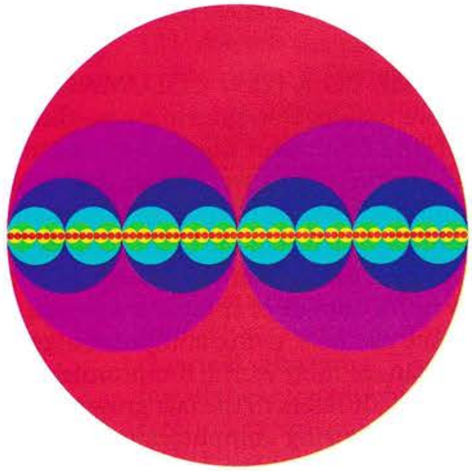


Figure 2

Here is CantorCheese being made at stage 4 with a shrink factor of 1/3.

```
Show[Graphics[CantorCheese[0, 0, 1, 4,
1. /3]]]
```

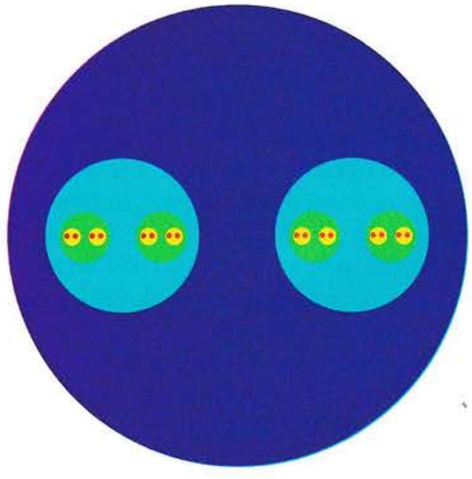


Figure 3

Sierpinski triangle

The Sierpinski triangle has a recursive construction similar to Cantor Cheese. The first graphic below shows the construction of an equilateral triangle using the built-in *Mathematica* command Polygon. Given a list of points, Polygon builds the polygon that connects these points. Now divide up the triangle into four equal subtriangles and extract the middle one. Continue this process recursively on the remaining triangles as we did in the construction of Cantor Cheese. The Sierpinski triangle is the outcome when we set shrink = 1/2.

```
SetOptions[Graphics, AspectRatio → Automatic, PlotRange → {0, 1}];
SierpinskiTriangle[x_, y_, r_, 0, shrink_] := {Hue[0], Polygon[{{x, y}, {x + r, y}, {x + .5 r, y + √3./2 r}}]};
SierpinskiTriangle[x_, y_, r_, n_, shrink_] := {{Hue[ .17 n], Polygon[{{x, y}, {x + r, y}, {x + .5 r, y + √3./2 r}}]},
SierpinskiTriangle[ x, y, shrink*r, n - 1,shrink],
SierpinskiTriangle[x + shrink*r, y, shrink*r, n - 1, shrink],
SierpinskiTriangle[ x + shrink*r/2, y + shrink*√3 r/2. , shrink*r, n - 1, shrink]}
Show[Graphics[SierpinskiTriangle[0, 0, 1, 5, .5]]]
```

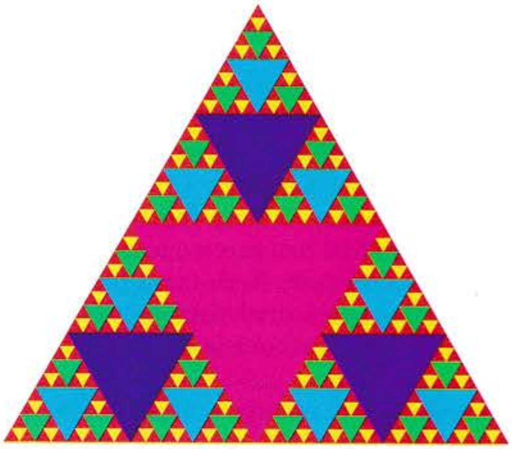


Figure 4

Here are the stages in the construction of the Siepinski triangle.

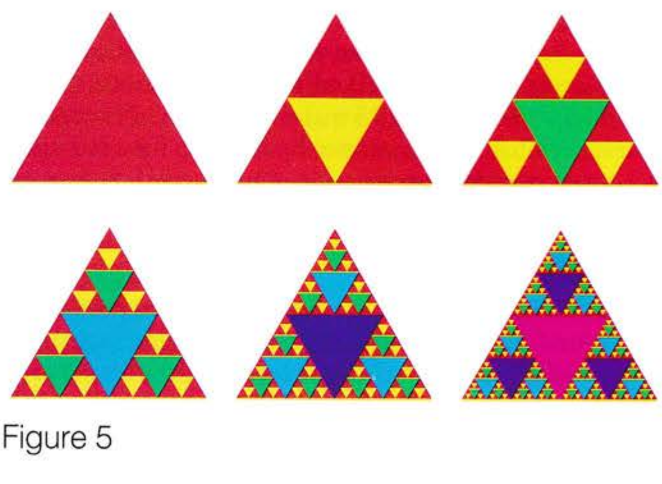


Figure 5

Your turn

Now it is your turn to try your hand at building something recursively. Programming is not a spectator sport, and to learn anything you must just do it. Try building the following Cantor-Cheese-like object below by dividing a square into four equal parts and throwing away two of the squares. Continue the same division with the subsquares. Here is what the stages should look like:

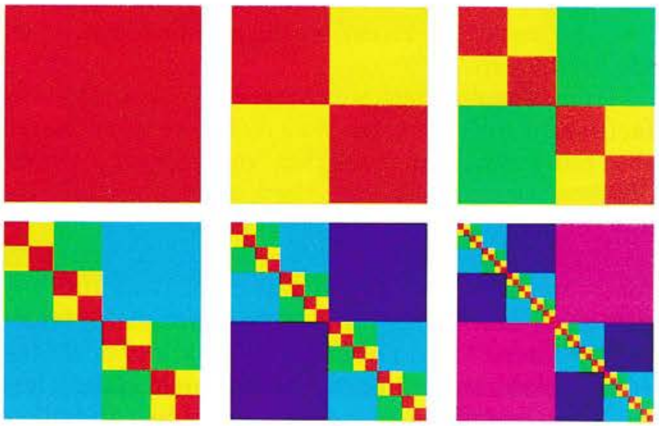


Figure 6

Web page support

Waiting for two months to see a solution is not necessary today thanks to the Internet. Therefore, all solutions to the problems presented in this column are available by going to the Informatics web page: http://www.usaco.org/informatics/.

1999 International Olympiad in Informatics

The eleventh IOI is now history, and I have just returned from the site of this competition in Antalya, Turkey. USA team members, Daniel Wright from Colorado, Ben Mathews from Texas, Percy Liang from Arizona, and David Cheng from Delaware, are back home after a week-long event held on the shores of the Mediterranean. Fortunately, each member of the team managed to score high enough to receive a medal. A total of 257 students from 65 countries competed for 22 gold, 42 silver, and 64 bronze medals. Out of 600 points possible the median score was 135, rather low by previous IOI standards. Our youngest team members, David and Percy, who still have a year of eligibility, received bronze medals. Our retiring seniors, Ben and Daniel, who are back in their freshman dorms at Cal Tech and Stanford, received silver medals.

The top score of 480 points went to Hong Chen from China, for which he received the gold first place trophy. Second place was shared by Mathijs Vogelzang of the Netherlands and Roman Pastoukhov of the Russian Federation. We were fortunate to have had Mathijs join us at our training camp last summer, so watching him receive the first gold medal ever for the Netherlands was a special treat.

Looking back over the past week, there are many memorable moments. Here are some of them.

Team Leader Rob Kolstad and I met up in Chicago with three of the team members, while Deputy Team Leader Brian Dean from MIT escorted David Cheng on a flight from New York. We all met up in Istanbul and took a short one-hour flight together to Antalya, Turkey's principal holiday destination. As soon as we left the airport, we were met by our student guide who escorted us by bus to Sirene City Resort, an impressive five star vacation retreat on the sunny shores of the Mediterranean.

That evening when we entered the outdoor dining facility for the first time, we saw before us an incredible buffet of salads, cheeses, main dishes, meats, chicken, fish, breads, and then, if you possibly had any room left, desserts without end. Occasionally at an IOI, you will hear someone complain about the food—but not this year. It was clear from our very first meal that there would be only accolades for the food at Sirene City. The same wide variety of choices was available morning, noon, and night. Great food is clearly one of the advantages of holding an IOI at a five star resort.

Sunday

The opening ceremony was held at the beautiful Talya Hotel in downtown Antalya on a bluff overlooking the Mediterranean Sea. Professor Namik Kemal Pak, Director of Tübitak, relayed a message sent by the President of Turkey, Süleyman Demirel. Professor Göktürk Uçoluk followed with a brief journey through the history of computer science, ending with a string and ball demonstration of a constant time algorithm for finding the shortest path between two nodes. Native dancers performed a marriage dance that began with a camel ride for the bride and a shave for the groom.

A tour of the Antalya Kaleiçi Marina and old city began near the elegant, fluted minaret of the Yivli Manareli Mosque built in the 13th century. The day ended with a look at the ancient Roman statues housed at the Antalya Museum. The museum is rich with relics from the Roman and Byzantine periods.

That evening, students and team leaders separated. We were housed in separate quarters in order to ensure there would be no contact between the team leaders and the students once the problem selection process had begun. The general assembly began the question selection process at 9:00 p.m. It took longer than expected, and translations from English into each native language began in the wee hours of the next day. Some countries were still translating when the sun came up the next morning, making for a long and tiring night.

Yes, the sun did come up each morning, and furthermore, it stayed visible all day. We were in sun bathers' heaven all week. The vacationers to this area, mostly Germans, lounged by the two huge outdoor pools or soaked up the sun on the beach. For as far as the eye could see, beach umbrellas lined the coast in front of resort after resort, with people dedicated to one task—getting baked.

Monday

The students were up early, had breakfast, and were led into a huge convention room that was partitioned off and filled with 300 computers, all networked together. They spent the next five hours trying to solve three tough problems by writing programs in Turbo Pascal or Turbo C/C++. The team leaders were in a completely separate location and never saw the students in action. The IOI competition has never been a spectator sport. Once the clock began, many team leaders headed off to their rooms or to the pool for some well-deserved rest.

Tuesday

This was a day of relaxation that began with a visit to a carpet and jewelry mall. We watched women making the carpets and were told they can do this job for only two to four hours a day because of the intensity of the work. The beautiful silk rugs with the highest density of knots were truly works of art. We were led into a large room and seated against the wall, and then the show began. Our "Master of Rugs" gave the history of each carpet, and attendants rolled them out as he snapped his fingers. The smaller rugs were sent spinning airborne with dramatic flair. Once the show was over, many salesmen came out of the woodwork looking for prey. If they spotted a likely buyer, he or she was invited to a private room where the real bargaining would begin. Both team leaders, Rob and Brian, who never dreamed they would come home with a rug, did. Of course, they each "really got a good deal."

The rest of the day was spent at Kemer for lunch and then on to Phaselis on the west coast of Antalya. Here we toured an impressive site of Roman ruins, walking under an aqueduct and down a main street of gray-white marble blocks with bath houses on either side. A visit to a Roman amphitheater finished the tour, and we loaded up for a high-speed bus ride back to the hotel. It appeared that the bus drivers were very eager to be done for the day. Kim, the team leader from Holland, who was sitting in front, walked back and announced, in disbelief, "We're going 95 miles per hour." Enroute we marvelled that we never once had to stop for a red light. It turned out that we were in a highspeed, police escorted, motorcade.

Wednesday

The second round of competition was a repeat of the first round. Again, several delegation leaders were up all night making their translations. After the competition ended, everyone headed for some R&R beside the pools or on the sandy beach.

Thursday

Best described as an antiquities and nature tour, Thursday began with a short bus ride to Perge, a beautifully complete Roman city. We learned that Appollonius of Perge, known as the Great Geometer, introduced the terms Ellipse, Hyperbola, and Parabola. Next we visited the two beautiful waterfalls at Kursunlu Falls. The tour ended with a visit to the Aspendos Amphitheater, one of the largest and best preserved theaters of antiquity. It was built during the rein of Marcus Aurelius (2nd century BC) and can hold 20,000 people.

We returned to the resort hotel and our final evening general assembly. First, the cut off scores for the medals were quickly decided. Then we began discussing another recommendation that came from the New Environments Committee. This was a plan to create an IOI Scientific Committee (ISC) that would assist each country's Scientific Committee in the formulation and review of problems and test data. It was explained that as the competition advanced in complexity, we should try to provide a level of continuity from competition to competition, so that each country need not start from ground zero. As it was explained by our team leader Rob, "It is far better to have cooperation between countries and help raise the level of all IOI's than to have a competition for the dubious title of "Best IOI." In this way the last IOI will always be the "Best IOI," since it will be constantly improving. The IOI Scientific Committee would provide another level of review to ensure that the competition problems and test data are consistently of high quality. Another recommendation was to create an IOI software team (IST). It would be responsible for the creation, maintenance, and distribution of evaluation software. Both recommendations were approved by the General Assembly. It was clear to me that the General Assembly was receptive to new ideas that are needed, well documented, and properly presented.

Tobias Thierer of Germany (center) joins US team members (left to right) Percy Liang, David Cheng, Ben Mathews, Daniel Wright for a meal.

Friday

The closing ceremonies were held at the Dedeman Hotel in Antalya. Those responsible for major portions of the IOI were given the honor of handing out the awards. When their names were called, the audience responded with warm applause to express their thanks for a job well done. The Deputy Prime Minister of Turkey was in attendance and helped distribute the gold medals. Finally, the orange and white IOI flag was handed over by Göktürk Uçoluk of Turkey to Zide Du of China. Zide then invited everyone to come to the 12th IOI in Beijing, China, September 23–30, 2000.

A final dinner followed in the dining hall of the hotel overlooking the blue waters of the Mediterranean Sea. This was my last chance to photograph the teams and team leaders together, and I took advantage of it. During the week I had taken well over 250 digital photographs which I had loaded onto my laptop. I was planning to select the best and make them available on our website when I return.

The next day we flew from Antalya to Istanbul and then headed back to the United States. As usual, we seemed to have been gone much longer than a week, since we had done and seen so much. The 11th IOI was another fantastic experience for students, team leaders, and me. The entire Turkish organizing committee deserves our most sincere congratulations for a truly impressive last IOI of the millennium.

Finally...

To participate fully in this column, you will need to have access to *Mathematica*. Readers who are students in any school or college may purchase the student edition of *Mathematica*—which is a complete version—at the bargain price of $139. Go to the Wolfram website for details at http://www.wolfram.com/products/student/.

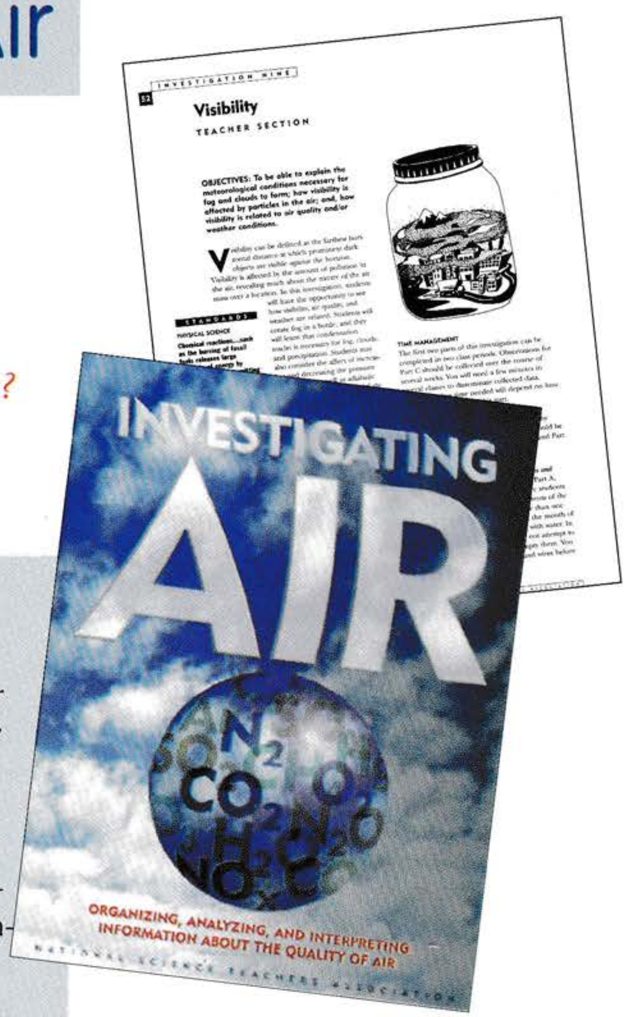
If you are interested in participating in the USA Computing Olympiad, go to http:// www.usaco.org. To find solutions to all problems used in this column, go to http://www.usaco.org/Informatics/.

Pictures of IOI'99 are available on the USACO web site at: http://www.uwp.edu/academic/mathematics/usaco/1999/ioi/ioi/htm.

The complete competition results and the questions used at IOI'99 are available at: http://www.ioi99.org.tr/ ◘

Investigating Air

What is air made of?
How does air work?
Do humans impact air quality?
How does air quality impact humans?
Do we have a right to clean air?



Investigating Air provides the investigations and information that lead students to the answers to these and other questions. Students use data generated by researchers, and generate their own data, in situations analogous to real air quality issues—increases in atmospheric CO_2, acid deposition, ozone issues, and visibility. This resource culminates in a decision-making investigation that builds on the science and equity knowledge students have developed.
(grades 9–12, 1998, 78 pp.)

Produced by the National Science Teachers Association with support from the US Environmental Protection Agency.

#PB148X **$15.95**

Table of Contents

To order, call 1–800–722–NSTA

NSTA Pub Sales, 1840 Wilson Blvd.
Arlington, VA 22201–3000
www.nsta.org/scistore

P999Q2