2
CHAPTER

Similarity and Dissimilarity

The Universe is built on a plan the profound symmetry of which is somehow present in the inner structure of our intellect.

—PAUL VALERY

The concepts of similarity and dissimilarity have long played an important role in human affairs and nature. We welcome similarity, but are often more powerfully attracted by the *dis*similar. Mass attracts mass, but electrons are attracted by their opposite numbers: antielectrons, or *positrons*. In fact, electrons and positrons can consummate their love in a small bang with an energy of a million electron volts. (But more often than not, the dissimilar evokes discrimination. For a mild example, think of the Russian word for *German*: *nyemets*, meaning "the mute one" or "not one of us." The reader will have no difficulty finding more partisan instances.)

In the sciences, similarity has both mystified and illuminated. Why are all electrons similar to each other—in fact, for all we know, *identical*? Here is one of the great unsolved riddles that Nature likes to tease us with. Once we know the answer and why the electron and other "elementary" particles have precisely the masses, charges, and spins they have, we shall know a lot more about the environment we inhabit (and inhibit). In this chapter, before indulging further in *self*-similarity, we shall touch upon some of the uses to which the ideas of similarity and *dis*similarity have been put in physics, psychophysics, biology, geology (mountaineering?), and other fields in which *scaling* is a crucial concept.

More Than One Scale

Measurement is usually thought of as an unambiguous, if imprecise, process. A soccer field has an area of roughly 50 meters times 100 meters. Thus, 5000 square

meters is the area of the field as far as the soccer player is concerned, or the real estate agent who sold the land.

But there is another area associated with a soccer field or any other lawn or meadow: the area important to the little bug that stalks up and down the grass blades. This area, corresponding to the total surface area of all the grass blades, is much larger than the soccer-playing area of the field, perhaps by a factor of 100. This larger area is also the relevant area for the sun's photons that are absorbed by the chlorophyll in the grass to convert carbon dioxide in the air to carbohydrates and oxygen.

Thus, for a soccer field, the question *What is its area?* has at least *two* true answers; the field is characterized by two area *scales* that differ by a very large factor. In other situations, measurements can lead to *many* answers. For example, the boundary between two European countries typically depends on the scale used in its determination. Thus, on a globe of the world, the length of the border between Spain and Andorra (or Austria and Liechtenstein, if shown) is considerably shorter than that determined from a map of Europe, which in turn is shorter than the border length obtained from a map of the Pyrenees (or the Alps).

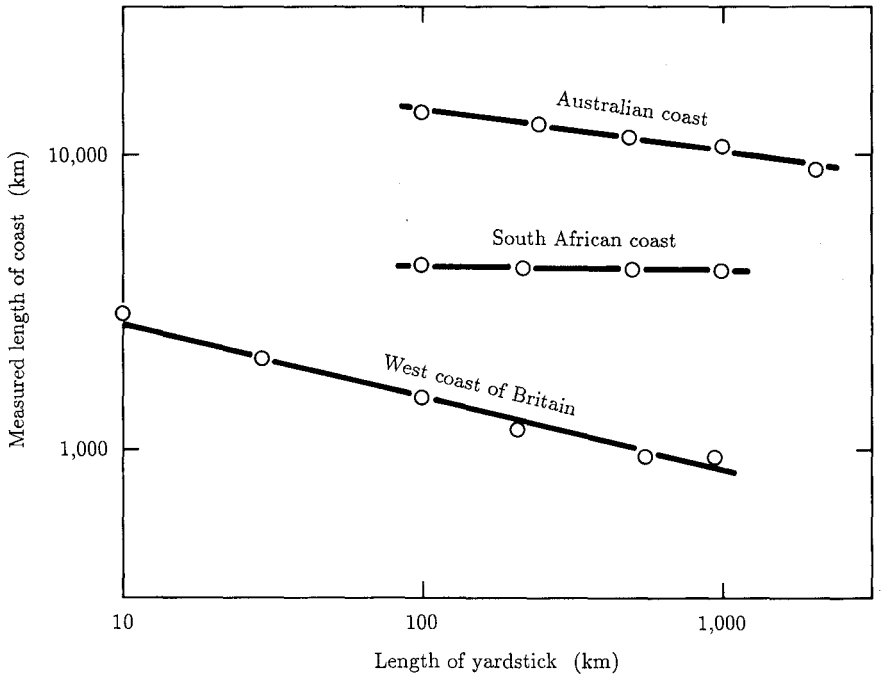


Figure 1 Measured lengths increase as the length of the yardstick is reduced.

Still longer lengths are obtained from more detailed maps, showing just the region in question, and from hiking maps. And actually walking (swimming or climbing) along the border will reveal an even longer length (see Figure 1). Thus, there is no *one* border length—there are *many*. In contemporary parlance, the border, like the fractal Koch curve discussed on pages 7–8 in Chapter 1, is said to have *many length scales*, an important concept in self-similarity and fractals.

In physics, there are numerous phenomena that are said to be "true on all scales," such as the Heisenberg uncertainty relation, to which no exception has been found over vast ranges of the variables involved (such as energy versus time, or momentum versus position). But even when the size ranges are limited, as in galaxy clusters (by the size of the universe) or the magnetic domains in a piece of iron near the transition point to ferromagnetism (by the size of the magnet), the concept *true on all scales* is an important postulate in analyzing otherwise often obscure observations.

To Scale or Not to Scale: A Bit of Biology and Astrophysics

Elephants and hippopotamus have grown clumsy as well as big, and the elk is of necessity less graceful than the gazelle.
—D'ARCY THOMPSON

Ironically, Galileo (see Figure 2), who discovered the scaling law for falling objects and thereby inaugurated modern experimental science, was also the one who noticed that some laws of physics (and biology) are *not* unchanged under changes of scale. In reflecting about the strength of bones, he argued that an animal twice as long, wide, and tall will weigh 8 times more. But, he pointed out, bones that are twice as wide have only four times the cross section and can support only *four* times the weight. Thus, to support the full weight, bone width must be scaled up by a factor greater than 2. This deviation from simple similarity introduces a natural scale in the design of animals, land-bound or aquatic: at some roughly predictable size, the bones become larger than the rest of the animal, and scaling (and the hypothetical beast) break down; see the essay by J. B. S. Haldane (1892–1964) *On Being the Right Size* [Hal 28].

Another instance of scaling in biology is the energy dissipation of warm-blooded animals as a function of their weight or mass (see Figure 3). One would naively expect the energy dissipation P as measured by daily caloric consumption to be proportional to the animal's surface area, which, for "similar" animals, is roughly proportional to the two-thirds power of its volume or mass m: $P \sim m^{2/3}$.

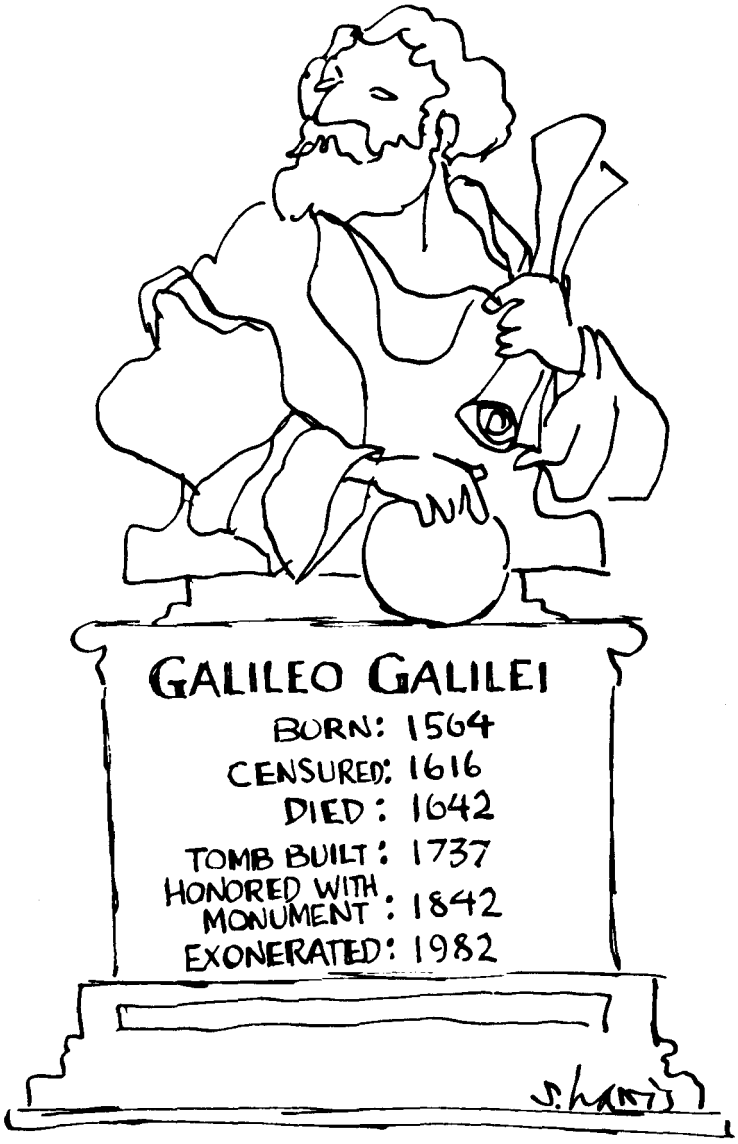


Figure 2 Galileo Galilei Vindicated [Har 77]. (© 1991 Sidney Harris)

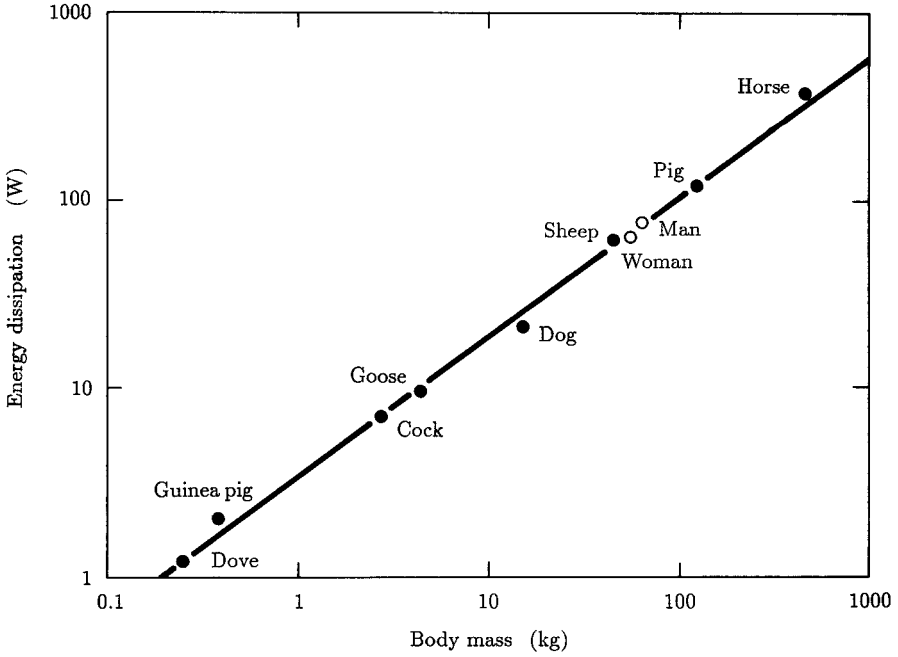


Figure 3 Energy dissipation of warm-blooded animals as a function of their weight.

In fact, the slope in Figure 3 is tantalizingly close to $\frac{2}{3}$, yet there is a small but systematic deviation from the expected slope: larger animals dissipate more power than the relation $P \sim m^{2/3}$ would predict. Actually, the data for a wide variety of species, including *Homo sapiens*, are much better fitted by an exponent of $\frac{3}{4}$. Why? This is a good question that merits further study. Is it that larger animals are less energy-efficient? Are they pushing at the same constraint on size that finally did the dinosaurs in? It seems that *people* are getting taller and bigger by the decade and they should perhaps be careful lest they join the extinct mammoth in oblivion.

A similar scaling failure occurs in photography, commonly referred to as *reciprocity failure*. It was discovered by the German astrophysicist (in fact, the founder of astrophysics) Karl Schwarzschild (1873–1916).[1] Schwarzschild, in

1. Schwarzschild was barely 16 when he published his first papers (on the orbits of double stars). As early as 1899 he developed theories of the curvature of space, and in 1916 he gave the first exact solution of Einstein's general relativity equations, predicting the existence of *black holes* once a star shrinks below the "Schwarzschild radius", a characteristic length at which gravity overpowers all other forces. Thus, gravity limits the mass of both stars and animals.

cataloging the brightness of stars, found that a star half as bright as a reference star required *more* than twice the exposure time to blacken the photographic plate to the same degree as the reference. To achieve a given degree of blackening at low brightness b, the required exposure time t is *not* proportional to the reciprocal of the brightness (as is true at higher brightnesses and shorter exposure times). Rather, Schwarzschild found, $t^p \sim b^{-1}$, where the *Schwarzschild exponent* p is less than unity.

Thus, when it gets dark (or when you stop down the lens of your camera too much), you should expose longer than you might think. In color photography, since the different colors show different reciprocity failures, the color balance may change at low illumination unless corrected by special filters.

In spite of these failures, there is much to be gained by scaling—from fish to physics.

Similarity in Physics: Some Astounding Consequences

In physics, similarity arguments have carried us quite far. But even on an elementary level, similarity reasoning can help a great deal. Think of a physical system whose potential energy U is a homogeneous function of degree k of the spatial coordinates r_m:

$$U(\alpha r_1, \alpha r_2, \ldots) = \alpha^k U(r_1, r_2, \ldots) \tag{1}$$

If we scale all spatial coordinates by a factor α and time by a factor β, then velocities are changed by a factor α/β and kinetic energy changes by α^2/β^2. Now, if α^2/β^2 equals the factor α^k for the potential energy U, then the *Lagrangian* of the physical system is multiplied by a constant factor α^k and the equations of motion are unaltered. The resulting paths of all mass points ("particles") remain similar to the original ones; the only change is a change of scales [LL 76].

Time intervals along the new path scale as $\beta = \alpha^{1-k/2}$. Energies, of course, vary as α^k, and angular momenta, having the same dimension as Planck's quantum of action (energy × time), are multiplied by $\alpha^{1+k/2}$.

What can we conclude from all this? Let us look at a linear oscillator, whose potential energy is a homogeneous *quadratic* function, that is, a function for which the exponent k in equation 1 equals 2. A pendulum swinging with very small amplitudes is a simple example of a linear oscillator. Does the period of oscillation depend on the amplitude of the oscillation? We could, of course, solve the equation of motion and see that it does not. But that is unnecessary just to answer the question. As we just noted, times scale as $\alpha^{1-k/2}$ which, for $k = 2$, equals α^0. Thus, we see immediately that all times, including the period of oscillation, are unaffected: the natural frequency of a *linear* oscillator, adhering strictly to a quadratic potential energy function, is independent of its amplitude

or energy. (In quantum mechanics this fact is reflected in the constant spacing, $h\nu$, between adjacent energy levels.)

What about a *non*linear oscillator with a cubic-law restoring force, that is, a quartic (fourth-power, $k = 4$) homogeneous potential? Now we cannot solve the equation of motion with a simple trigonometric function. But similarity tells us that times must scale as $\alpha^{1-k/2} = \alpha^{-1}$, that is, that frequencies are proportional to α: the higher the energy of the oscillator, the higher its resonance frequency, as might be expected for a spring with increasing stiffness. More precisely, the resonance frequency of such a nonlinear oscillator scales as the fourth root of its energy. (Quantum mechanically, $E_n \sim n^{4/3}$.)

In a *uniform* force field, the potential energy is a homogeneous *linear* function of the spatial coordinates; that is, $k = 1$. As a consequence, times scale as $\alpha^{1-k/2} = \alpha^{1/2}$, as indeed they do, a fact that an early practitioner of scaling, Galileo, discovered a long time ago in Pisa: he had to scale 4 times as many steps on the Leaning Tower to double the fall time (a fine tale, albeit apocryphal).

In Newtonian attraction, the potential energy is *inversely* proportional to distance. Thus, $k = -1$ and, for circular orbits around a massive center, we must expect times to scale as $\alpha^{1-k/2} = \alpha^{3/2}$. In other words, the square of a planetary orbital period is proportional to the cube of its size. And so we have just rediscovered a special case of Kepler's immortal third law of celestial mechanics—without solving a single integral!

Isaac Newton, in his *Principia*, even considered more general "planetary" laws. For a circular orbit of radius r and period τ, he deduced, from an assumed scaling relation $\tau \sim r^n$, the following law for the gravitational potential: $U \sim r^{2-2n}$. The same result follows directly from our similarity principle. For $n = \frac{3}{2}$, we are back to $U \sim r^{-1}$ and the real world of falling apples and orbiting moons. In fact, it was Newton's sudden inspiration[2] that the gravitational pull the earth exerted on an apple was $3600 = 60^2$ times larger than the pull it exerted on the moon (60 times more distant from the earth's center) that led him to formulate his universal law of gravitation: gravitational force must be proportional to the reciprocal of the distance squared.

For $U \sim r^{-2}$, a possible orbit is a logarithmic spiral: $r(\phi) = r_0 e^{\gamma\phi}$ in polar coordinates, a self-similar object! See pages 89–92 in Chapter 3 and Figure 4 for an artistic elaboration of the logarithmic spiral. What does scaling tell us about velocities and timing for *this* motion as a "planet" spirals into (or away from) its sun?

For $U \sim r^{-3}$, a cardioid $r = r_0(1 + \sin \phi(t))$ is a possible motion. What can we say about the angle $\phi(t)$?

Exploiting similarity, we can even prove the *virial theorem*, which relates *average* potential energy $\overline{U}$ to the *average* kinetic energy $\overline{T}$ for bounded motions. Since the kinetic energy T is a homogeneous *quadratic* function of the velocities

2. Intriguingly, the German word for inspiration is *Einfall*, spelled like *ein Fall* (a fall).

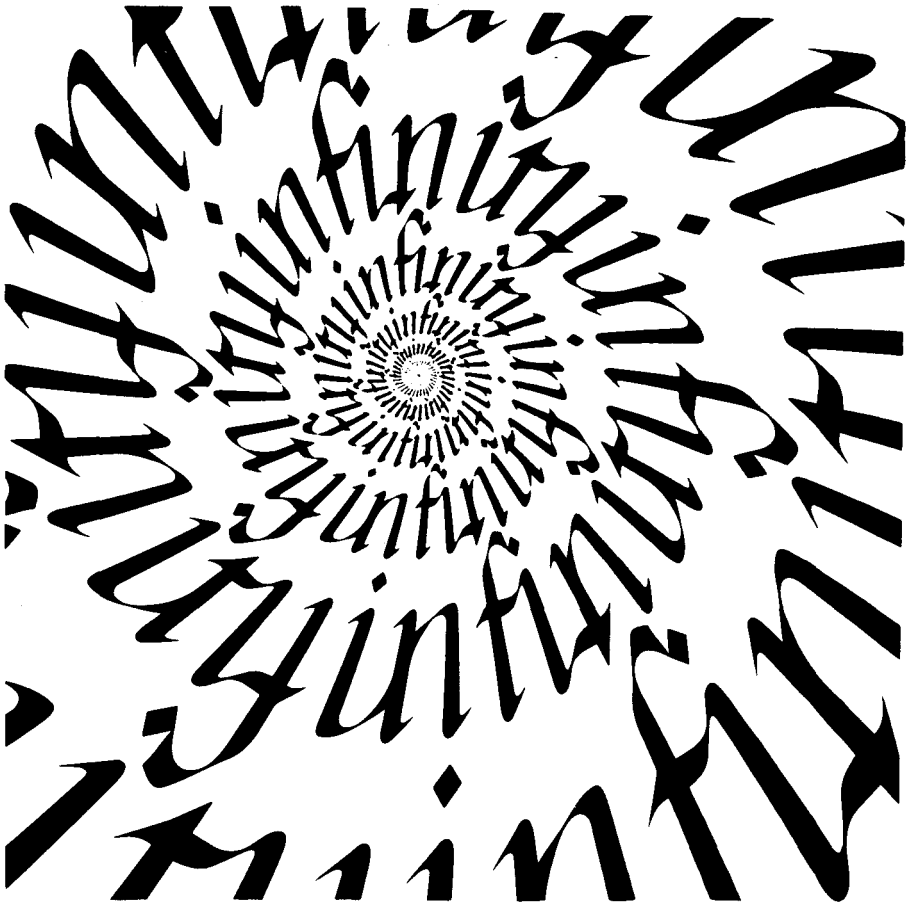


Figure 4 Logarithmic spiral to infinity.

and the potential energy U is a homogeneous function of degree k of the spatial coordinates, it follows (almost) immediately from equation 1 that $2\overline{T} = k\overline{U}$.

For the linear oscillator ($k = 2$) we recover the well-known equality between average kinetic and potential energies: $\overline{T} = \overline{U}$.

Similarity in Concert Halls, Microwaves, and Hydrodynamics

Similarity transformations have been particularly fruitful in hydrodynamics and other difficult fields. Already in the nineteenth century, Sophus Lie (1842–1899)

and later George David Birkhoff (1884–1944) had looked for transformation groups that leave given partial differential equations, and thus their solutions, invariant. Such solutions are called *similarity solutions*.

Suppose the following limit of a solution $\phi(x, y)$ exists:

$$\lim_{\varepsilon \to 0} \varepsilon^a \phi(\varepsilon^b x, \varepsilon^c y) = \Phi(x, y) \tag{2}$$

Then the similarity solution, $\Phi(x, y)$, obeys the following scaling law:

$$\Phi(x, y) = \lambda^a \Phi(\lambda^b x, \lambda^c y)$$

which follows immediately from equation 2 and is, in fact, a generalization of equation 1.

But scaling is not always *that* easy. A good method of designing concert halls and opera houses, for example, is to build *scale models* first and study sound transmission in *them*, instead of in the finished hall.[3] Linear dimensions, wavelengths, and frequencies scale easily. In a one-tenth scale model, for example, all linear dimensions are one-tenth of those in the real hall. Thus, since sound diffraction from hard surfaces depends only on the ratio of the wavelength to the linear dimensions of the scattering surfaces, wavelengths should likewise be scaled down by a factor of 10. For a fixed sound velocity ($c = 343$ m/s in dry air at room temperature), this means that frequencies should be scaled *up* by a factor of 10. Travel times are, of course, 10 times shorter in the scale model. Thus, times scale inversely with respect to frequencies—a good thing, because frequencies are measured in reciprocal seconds. But sound absorption (by people, for example) is more difficult to scale. Nevertheless, special materials (known as "instant people") have been invented that mimic human absorption at upscale frequencies.

Absorption, friction, and other energy loss mechanisms generally cause scaling difficulties. For example, a microwave cavity (a hollow metallic resonator) scaled down in size by a factor of 10 has its resonance frequencies scaled up by the same factor of 10. However, the minimum *bandwidths* of its resonant modes (as determined by skin-effect losses) will go up by a factor of $10^{3/2} \approx 32$, because the small penetration depth of electromagnetic fields in conductors (the finite penetration depth causes the skin effect) is proportional to $f^{-1/2}$, not f^{-1}. (The *relative* bandwidth of an electromagnetic cavity resonance is given approximately by the ratio of two volumes: the inner surface area of the cavity times the skin depth—i.e., the volume where the energy losses occur—divided by the total

3. But the alternative approach, *first* building the full-scale hall and *then* the scale models, has also been tried—with disastrous consequences requiring expensive "remodeling" (a euphemism if there ever was one).

volume of the cavity, or the volume where the energy is stored. At least *some* things about microwaves seem safe and simple.

Another instance where friction causes endless scaling problems is in shipbuilding model basins. A battleship scaled down in size by a factor of 50 experiences rather ill-mannered drag forces because drag is caused by viscous boundary layers, which, like the skin effect, follow a different scale.

How *much* size can affect scaling is illustrated perhaps best by the following eyewitness observation. A large ocean liner, "steaming" into New York harbor during a tugboat strike, had to shut off its engines *miles* ahead of its berth and then drift with the slowing tide up the Hudson River to arrive exactly at its pier just as the tide began to turn. This is no mean feat, because stopping distances of ocean liners, without external assist, are rather larger than even those of titanic trucks on glare ice.

By contrast, stopping distances on microscopic scales in anything but frictionless fluids (superfluids) are *so* short that any particles suspended in a liquid seem devoid of inertia. This was, in fact, the experience of the present investigator while observing hydrodynamic streaming in a model of the inner ear under a microscope: the moment the sound was turned off, the streaming motion stopped instantly, as if everything were massless. Such is "life at low Reynolds numbers" as described in an engaging article of that name by Edward Purcell [Pur 77]. Reynolds numbers are only one kind in a long list of dimensionless numbers that reflect the importance and difficulty of scaling in hydrodynamics [McG 71].

Scaling in Psychology

Whereas measurement in classical physics is a well-understood process, relating an observed quantity to a well-defined unit, the situation in psychology was not so clear-cut until the physiologist E. H. Weber (1795–1878)—brother of the physicist Wilhelm Weber (1804–1891), collaborator and son-in-law of C. F. Gauss—made careful studies of the sensations of sound and touch, thereby laying the foundations of a new science, the science of sensations. According to Weber's law, an increase in stimulus necessary to elicit a just noticeable increase in sensation is not a fixed quantity, but depends on the ratio of increase to the original stimulus. Later, the physicist and philosopher G. T. Fechner (1801–1887) restated Weber's law (as the Weber-Fechner law) and specified its domain of validity.[4] Modern psychologists, and particularly S. S. Stevens, have succeeded in introducing measurement methods into psychology that are nearly as unambiguous

4. Fechner also fathomed experimental aesthetics by measuring which shapes and dimensions are most pleasing. He may have been the first to conduct a public opinion poll (to discover which of two Holbein paintings was preferred by most people).

as objective measurements in physics [Ste 69]. The new discipline has therefore rightly earned the designation *psychophysics*, of which psychoacoustics is a special branch, as is psychovisual research.

One of Steven's great contributions was the introduction of *ratio scales* for subjective variables (like loudness and brightness) and the discovery of simple *power-law* relations between these subjective variables and corresponding physical quantities (like energy flux or intensity).

For example, for a sound to double in loudness L, its intensity I has to be multiplied by a factor of 10; this is true over much of the intensity range that the human ear can perceive without pain (a range exceeding 12 orders of magnitude at mid-range audio frequencies). Thus, because $\log_{10} 2 \approx 0.3$, we have the following power law for loudness as a function of acoustic intensity:

$$L \sim I^{0.3} \tag{3}$$

Someone who has not participated in a psychoacoustic scaling test might object that "loudness doubling" is not a well-defined concept. But surprisingly, the random scatter encountered in such tests is remarkably small even between different listeners.

The exponents found in psychophysical power laws, such as the value 0.3 in equation 3, are not universal but are specific to the sense modality studied (subjective brightness, perceived weight, or apparent length, for example) and have been analyzed in great detail by psychophysicists.[5] One important research question concerns the *transitivity* of these exponents when comparing loudness with weight and weight with brightness, for example, and what it might reveal about brain functions.

If we replace the sound intensity I in equation 3 by the sound pressure p, then, because intensity is proportional to pressure *squared*, we have

$$L \sim p^{0.6}$$

Interestingly, the exponent 0.6 can be derived from an exponent of 0.5 found at a more fundamental level, the Fourier-like "critical" frequency-band decomposition of sounds in the inner ear. The exponent 0.5, in turn, turns our attention in the direction of statistical analysis and uncertainty, resulting in the following simplified model of loudness perception. If loudness were perceived as the mean rate of nerve pulses traveling along the acoustic nerve up to higher auditory centers in the brain, and if these pulses were a modulated Poisson process whose mean rate was proportional to the sound pressure p, then the uncertainty of the number of pulses in a given time interval (100 ms, say) would be proportional to $p^{0.5}$. Since many ratio scales in psychophysics are found to be directly related

5. Thus, *universality*, so beloved by physicists, is lacking in psychophysics.

to perceptual uncertainties ("just noticeable differences"), the observed power law for subjective loudness *versus* physical intensity would then indeed be predicted by such a statistical model of neural firing rates.

In reality, loudness perception is more complicated, but the observed power laws and their exponents have yielded important clues and steered researchers in the right direction.

Acousticians, Alchemy, and Concert Halls

Concert halls are built to transmit pleasing sounds from performing musicians to attentive listeners (while keeping everybody dry and comfortable at the same time). Thus, nothing seems more apropos for an acoustical scientist than to measure the "frequency response" of a concert hall between the stage and various points in the audience area. Here *frequency response* means the effectiveness with which various frequencies (musical pitches) are transmitted between two distant points in the hall. Figure 5 shows a typical sample of such a frequency response on a logarithmic scale versus frequency.

The many ups and downs of such a response, even over narrow frequency intervals, are immediately apparent, as they were to Edward C. Wente (1889–

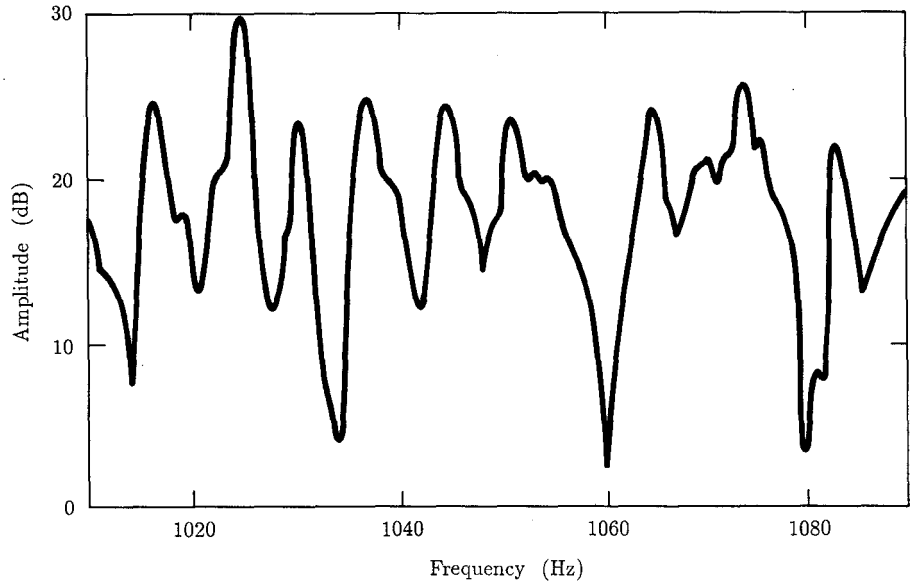


Figure 5 Sound transmission between two points in a concert hall as a function of frequency. Note large statistical fluctuations.

1972), inventor of the condenser microphone, who first published such a response in 1935. As a communications expert, Wente wondered (in print) why people could actually enjoy music in concert halls, given such large response fluctuations—fluctuations that exceed even those of the cheapest loudspeaker. The answer is that, when listening to speech or music, the ear unconsciously "switches" to a *short*-time spectral analysis, which, in accordance with Heisenberg's uncertainty principle, does not resolve fluctuations on such a fine frequency scale as shown in Figure 5.

Oblivious to the profound perceptual insignificance of these high-resolution frequency responses, acousticians around the world kept measuring them with great abandon. Worse, before too long, various supposedly objective criteria for acoustical quality were concocted from these responses. Of course, every time a new concert hall "came on line" (was inaugurated, to use a more archaic term), the extant criterion had to be modified to make the new hall's response characteristics conform with its perceived musical quality. These "scientific" attempts have been compared, quite appropriately I think, to tea-leaf reading and alchemy (although one should not unduly belittle the latter).

When the author, then a young student at Göttingen, heard about these activities—still going strong in 1954—he thought that maybe these frequency responses were nothing but *noise* in the frequency domain (technically: the modulus of a complex Gaussian process, resulting from the random interference of many overlapping room resonances). If this was indeed so, then practically all frequency responses in large halls should be *similar* to each other, characterized by a *single scaling parameter*: the reverberation time of the hall.

This turned out to be the case, as further measurements soon revealed. For example, the average frequency spacing (in hertz) of the response maxima was found to be about 4 divided by the reverberation time (in seconds), in excellent agreement with the theoretical prediction [Schr 54, or, in the most recent English translation, Schr 87]. Thus, people have been measuring nothing but reverberation time all these years, and in a very complicated and roundabout way at that. Ironically, the "new" criteria that had been distilled from frequency responses were supposed to *supplement* reverberation time, which had been found wanting in its predictive power of acoustical quality. Thus, a little insight and a good similarity argument liberated a lot of manpower from a useless pursuit.

High-resolution frequency responses became important later in solving the problem of stability of public-address systems, for which acoustic feedback ("howling") always threatens to become a problem. By shifting all frequency components of a speech signal by the average spacing between the room's response maxima and minima (a few hertz, according to the aforementioned statistical theory), the stability can be considerably increased—a howling success [Schr 64].

More generally (and perhaps more important), the theory of randomly interfering coherent waves has assumed a central role in the analysis of hologram laser speckles and electromagnetic multipath propagation; think of mobile cellular telephones in cars and the cordless handset at home, a marvelous invention,

especially when it comes with the matchless sound quality of a good "corded" phone.

Preference and Dissimilarity: Concert Halls Revisited

In this section we shall make a brief call on a problem in psychological scaling with which the author, although not a psychologist himself, is somewhat familiar: the ranking of concert hall acoustics.

A deeply entrenched procedure for getting a reading on the acoustical quality of a concert hall (or opera house) is to collect comments from listeners, musicians, conductors, and music critics. These subjective ratings are then correlated with various architectural and physical characteristics of the hall (such as width of the listening area, reverberation time, and frequency response). From these correlations a mathematical formula is then constructed for predicting acoustic quality on the basis of measurable, objective parameters; see, for example, Beranek's book *Music, Acoustics and Architecture* [Ber 62].

Typical responses elicited from German-speaking music lovers to characterize the acoustics of concert halls include such lovely locutions as *glasklar, jämmerlich, krankhaft, ruinös, unheimlich,* and—last but not least—*wunderbar.* To translate these high-sounding words into basic English would be sheer waste because they are not only ill defined but nearly meaningless in any language.

At Lincoln Center for the Performing Arts in New York City, Philharmonic Hall (now Avery Fisher Hall) had been designed on the basis of the aforementioned approach, and thus it is unsurprising that it required a major acoustic rescue effort. When the author was confronted with this, he and his colleagues recognized that as a first order of business, more reliable methods for the subjective (and objective) evaluation of concert halls had to be developed.

New objective measurement techniques [ASSW 66] revealed that the overhead acoustic panels ("clouds"; see Figure 6) did not reflect low-frequency components (especially from the cellos) with sufficient strength into the main audience area [SASW 66]. This was partly the result of poor *scaling*: to properly reflect musical notes of different wavelengths from an acoustic panel (not a panel of listeners), the panel's geometric dimensions must be at least comparable in size with the longest wavelength present in the sound. In actual fact, they were much too small, a failure that evoked both seering sarcasm and much mirth.[6]

6. The maestro of the acclaimed Cleveland Orchestra, the none-too-reticent George Szell, was so enraged by the whole debacle that he dubbed the panels "schwangere Frösche mit beleuchtetem Bauchnabel" (pregnant frogs with illuminated navels—on account of their double-duty function as lighting fixtures). A contemporary *New Yorker* cartoon showed two ladies walking in the foyer under the Lipchitz sculpture (vaguely reminiscent of suspended acoustic panels) and remarking, "No wonder the acoustics is so bad in there; it's all hanging out here!"

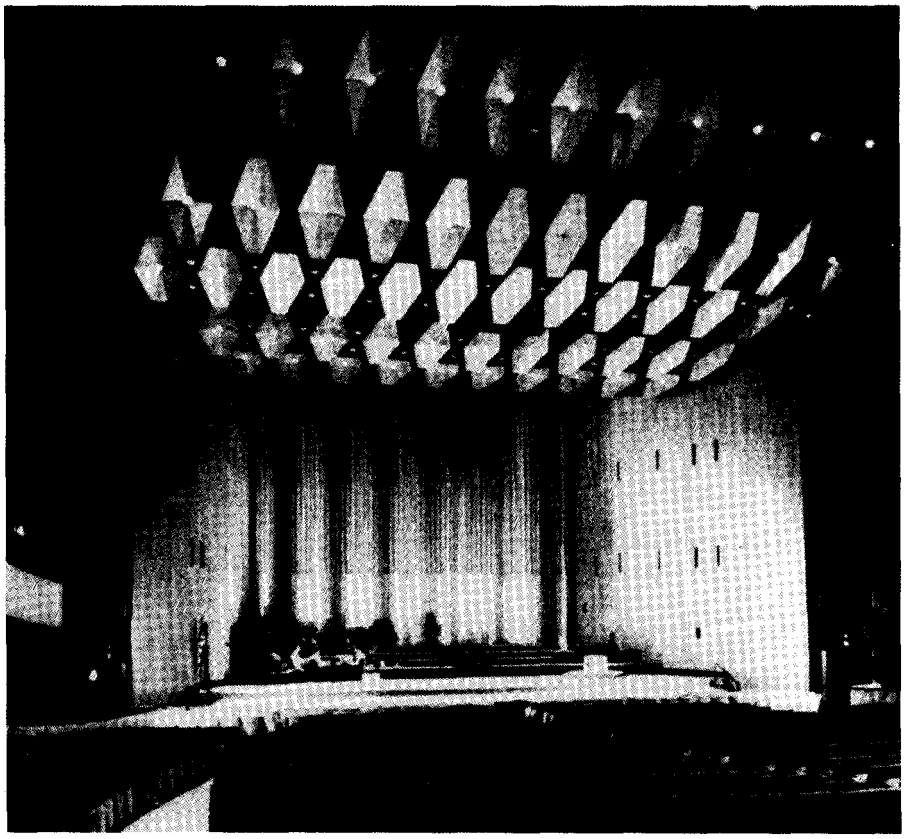


Figure 6 The acoustic panels ("clouds") in Philharmonic Hall, Lincoln Center for the Performing Arts, New York.

To put the subjective evaluation of concert hall quality on a firmer basis, the author suggested eschewing the use of any ill-defined epithets, such as those quoted in this section, and restricting listeners' responses strictly to an expression of acoustic *preference* between two halls or a degree of *dissimilarity* that they perceived. To preserve individual differences in musical preference, these responses were not simply averaged but were analyzed by modern *multidimensional scaling* algorithms [SGS 74]. With a sufficient number of responses—even just binary responses as in the case of two-valued preference judgments—these algorithms are capable of constructing a well-defined Euclidean space, usually of two or three dimensions, in which Euclidean distances are closely proportional to the perceived dissimilarities or differences in preference.

Figure 7 shows an example of a so-constructed *preference space*. The different symbols T_1, Q_3, and so forth, represent different concert halls and recording

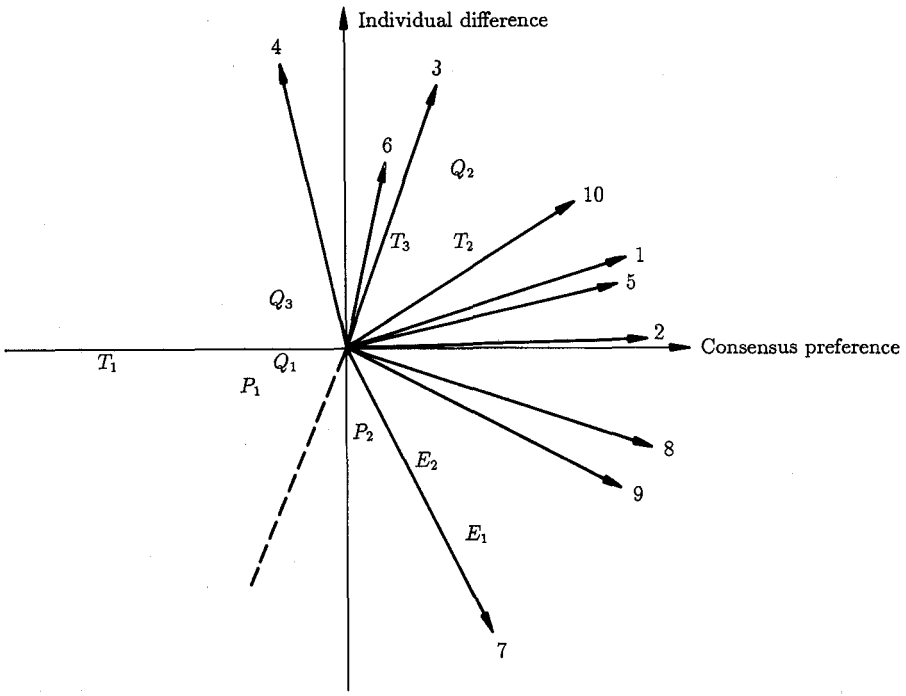


Figure 7 Preference space for concert halls.

locations in these halls (e.g., Q_3 is the third location in concert hall Q). The numbered arrows are unit vectors representing the 10 listeners who participated in this preference test. (The fact that some arrows appear shorter than others indicates that they have nonnegligible components in the third dimension used in the analysis, which is not shown here.)

For each pair of the 10 test conditions, for example T_3 and E_2, each of the 10 listeners states which of the two conditions he prefers. The accumulated preference data (a total of 450 paired comparisons) is subjected to multidimensional scaling by a metric linear factor analysis [Sla 60].

A computer program that implements this factor analysis iteratively changes the position in the preference space of each test condition, for example T_1, and the direction of each listener vector in a three-dimensional Euclidean space until the normal projections for all 10 test conditions on each of the 10 listener vectors agree, as closely as possible, with the preference data. Thus, Figure 7 tells us, for example, that listener 3 prefers test condition T_1 least and Q_2 most.

For an exact representation, a 10-dimensional Euclidean space would generally be required. In fact, almost 90 percent of the total variance in the data is accounted for by the two first dimensions of the preference space.

Since almost all listener's arrows in Figure 7 point into the right half plane, the abscissa could be labeled "consensus preference." Indeed, if some architectural modification of a hall would move the position in the preference space for a given location in that hall to the right, *all* listeners, except one (listener 4), would respond with a higher preference score for that seat.

By contrast, about half the arrows point into the upper half plane and about half into the lower. Thus, the ordinate strongly reflects "individual differences" in musical tastes—a personal dimension that should be honored in the design of spaces for the enjoyment of music.

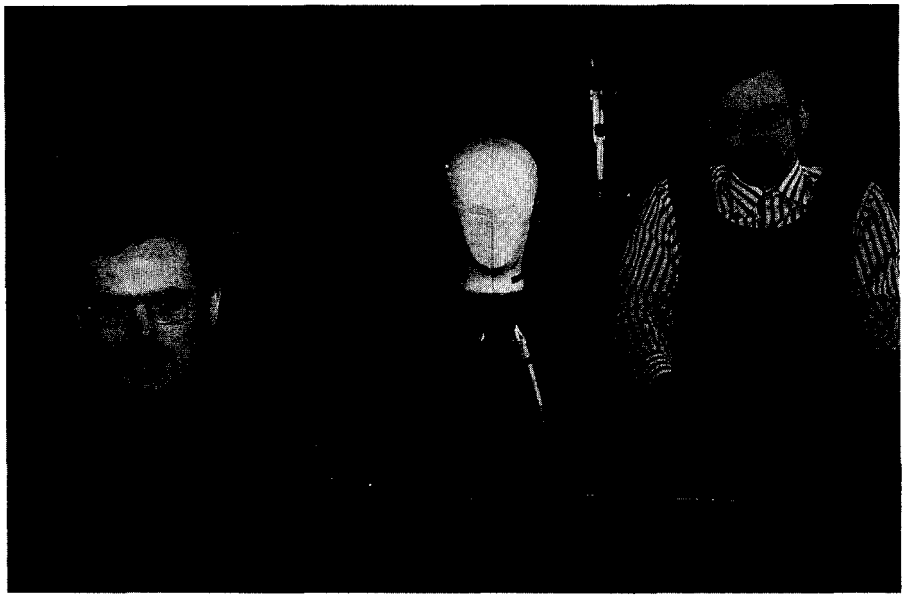
Subjective tests in which judgments of *dissimilarity* were elicited from the listeners gave very similar results. Thus, our confidence in *multi*dimensional scaling methods for constructing perceptual spaces was further strengthened. The success is doubtless due to two ingredients (or, rather, omissions):

1 Avoiding the use of ill-defined terms or empty words to describe acoustic quality

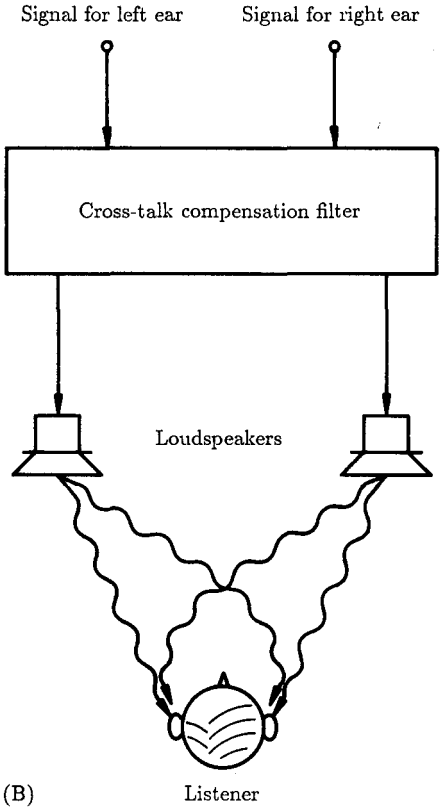
2 Not forcing musical tastes into a one-dimensional Procrustean bed (Procrustes would have loved the very thought of one-dimensional cots)

In conclusion, we should mention one "technical detail" without which the cited results could not have been achieved. Comparisons between the (often subtle) acoustic differences prevailing in different halls are notoriously difficult. Listening experiences separated by days or weeks, based on different musical programs, executed (good word!) by different orchestras, are highly unreliable. Thus, much would be gained by the possibility of *instantaneous* comparisons between different halls. This requirement was realized in the aforementioned study by an ingenious method that allowed faithful reproduction of music recorded in different halls at different times, using a tape recording of Mozart's *Jupiter* Symphony (and representatives of other musical styles) played by the English Chamber Orchestra in an anechoic environment and kindly made available to the author.

Figure 8A shows three crucial collaborators in this project. The reproduction method [Schr 70] is based on the fact that most people have just two ears. By properly preprocessing the two audio signals from the dummy head shown in Figure 8A, it is possible to transfer these two signals, via two loudspeakers, to the eardrums of a listener seated at some distance in front of the loudspeakers (see Figure 8B). The preprocessing, a kind of inverse filtering, compensates for the *cross talk* from each of the loudspeakers to the "wrong" ear. Since the sound transfer matrix between loudspeakers and ears is nonsingular for proper loudspeaker placement in an anechoic listening room, a physically realizable inverse exists, which is then incorporated into appropriate cross-talk compensation filters. The transfer matrix and its inverse depend upon the geometry and sound diffraction around the listener's head, which is measured for a "standard" head



(A)



(B)

Figure 8 (A) Three crucial collaborators in concert hall measurements. (B) Listening to loudspeakers with compensated cross-talk between the two speakers.

shape. (For once, success depends not so much on a head's inner workings but on its outer shape.)

To test this method of sound reproduction, sound waves arriving laterally from an angle of 90° were simulated using two loudspeakers located at $\pm 22.5°$. The simulation turned out to be so realistic that many listeners turned their heads 90° to locate a third (nonexistent) sound source. (Of course, when they turn their heads, the effect disappears because the geometry is changed.)

When this reproduction method was first applied to concert halls, the effect was truly overwhelming: at a flick of a switch the listener could "transport" himself from the Vienna Musikvereinssaal, say, to the Amsterdam Concertgebouw and back again. These instantaneous comparisons finally made reliable quality judgments possible that had eluded acousticians for so long.

The preference coordinates obtained in these tests were correlated with various physical parameters, revealing a consistent absence of strong *laterally* traveling sound waves in the less preferred halls. Thus, good acoustics—given proper reverberation time, frequency balance, and absence of disturbing echoes—is mediated by the presence of strong lateral sound waves that give rise to a preferred "stereophonic" sound. In old-style, narrow and high halls, such lateral sound is naturally provided by the architecture. By contrast, in many a modern, fan-shaped hall with a low ceiling, a "monophonic" sound, arriving in the symmetry plane through the listener's head, predominates, giving rise to an undesirable sensation of detachment from the music.

This, in a nutshell, is the reason why many modern halls project such a poor musical sound. The stimulating role that the concepts of (dis)similarity and multidimensional scaling [She 62] have played in the clarification process can hardly be overemphasized.

To increase the amount of laterally traveling sound in a modern hall, highly efficient sound-scattering surfaces have been invented recently [Schr 90]. These "reflection phase gratings," to use the physicist's term, are based on number-theoretic principles (primitive polynomials over finite fields, quadratic residues, and discrete logarithms) and have the remarkable property of scattering nearly equal acoustic (or radar) intensities into all directions. Such broadly scattering surfaces are now also being introduced into recording studios, churches, and even individual living rooms—and who knows where else.

These sound-dispersing structures should also find good use in noise abatement, because a dispersed (and therefore weakened) noise is more easily "masked" (rendered inaudible) by other sounds.

3 CHAPTER

Self-Similarity—Discrete, Continuous, Strict, and Otherwise

Big fleas have little fleas upon their backs
to bite them,
and little fleas have lesser fleas, and so ad infinitum

—SWIFT,
Poems II.651 (1733)

We say that an object—a geometric figure, for example—is invariant with scaling, or *self-similar*, for short, if it is reproduced by magnifying some portion of it.

Self-similarity comes in many different shapes and forms. Some of it continuous, some discrete; some is accurate over many powers of 10 and some over less than a factor of 10. We find examples of self-similarity in our daily surroundings or deeply hidden in the behavior of physical or biological systems. Some self-similarity is fully deterministic, some is only probabilistic. A few cases of self-similarity are mathematically exact; however, most instances in the real world are only asymptotically self-similar or just approximately so. Cantor sets and Weierstrass functions (see pages 96–98 and Chapter 7) are two well-known delegates from mathematics at this reunion. Brownian motion represents both the physical sciences and probabilistic self-similarity. And Bach's tempered 12-tone scale shows the importance of self-similarity in music.

A well-known example of discrete, albeit limited, self-similarity is a *set* of Chinese boxes or Russian dolls—the (usually wooden) kind where a large doll (discreetly) hides a *similar* smaller one inside its "body" and the smaller doll hides a similar third one and so forth for two or three more "generations." If we had a doubly infinite number of dolls, both ever smaller and ever larger dolls, then, provided the scaling ratio of doll sizes between successive generations was

constant, we would have a set with *exact* discrete self-similarity. But such a set could, of course, exist only in our imagination; real dolls have finite measurements: they must be both larger than single atoms and smaller than the full universe. (For Charles Addams's vision of incipient self-similarity, see his posthumously published cartoon given in Figure 1.)

Another example of discrete, if severely limited, self-similarity can be discovered on some product labels. Think of a beer-bottle that shows the same beer bottle on its label, which shows the same beer bottle on *its* label.

Or take a look at the cover of Paul Halmos's book *Naive Set Theory* [Hal 74], which shows the cover of the book, which shows the cover of the book, which shows the cover of the book without showing the cover. Of course, printing costs (not to mention other constraints) put an early and abrupt end to this progression of ever smaller images. A cheaper way to get many more scaled-down images is to stand between two almost parallel mirrors such as those found in clothing stores (see Figure 2). But of course, the high-order reflections are

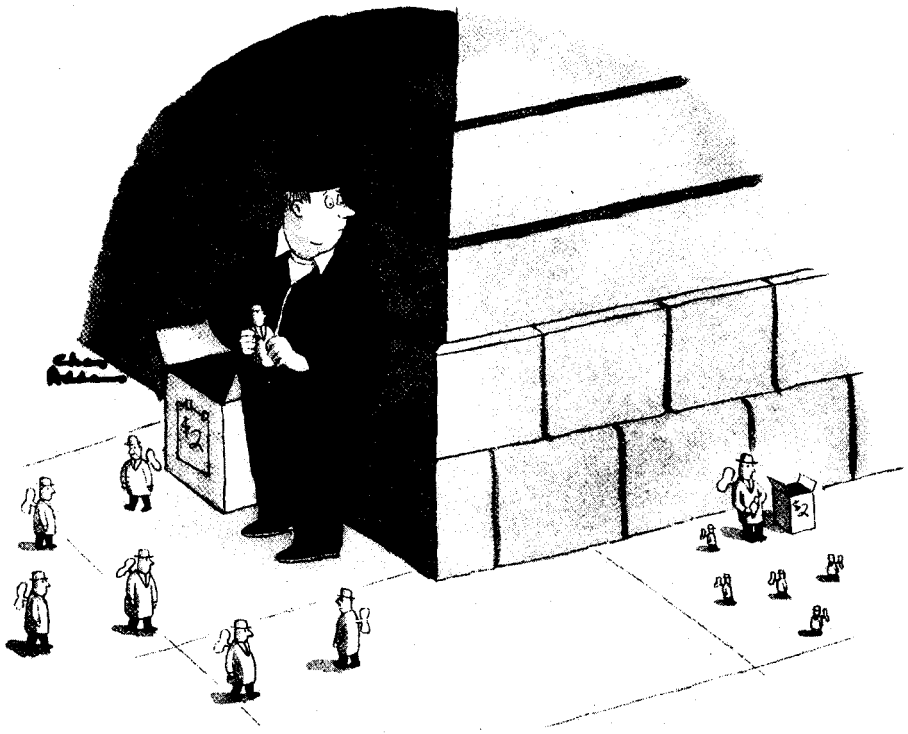


Figure 1 Self-similarity. (Drawing by Chas. Adams; © 1987 The New Yorker Magazine, Inc.)

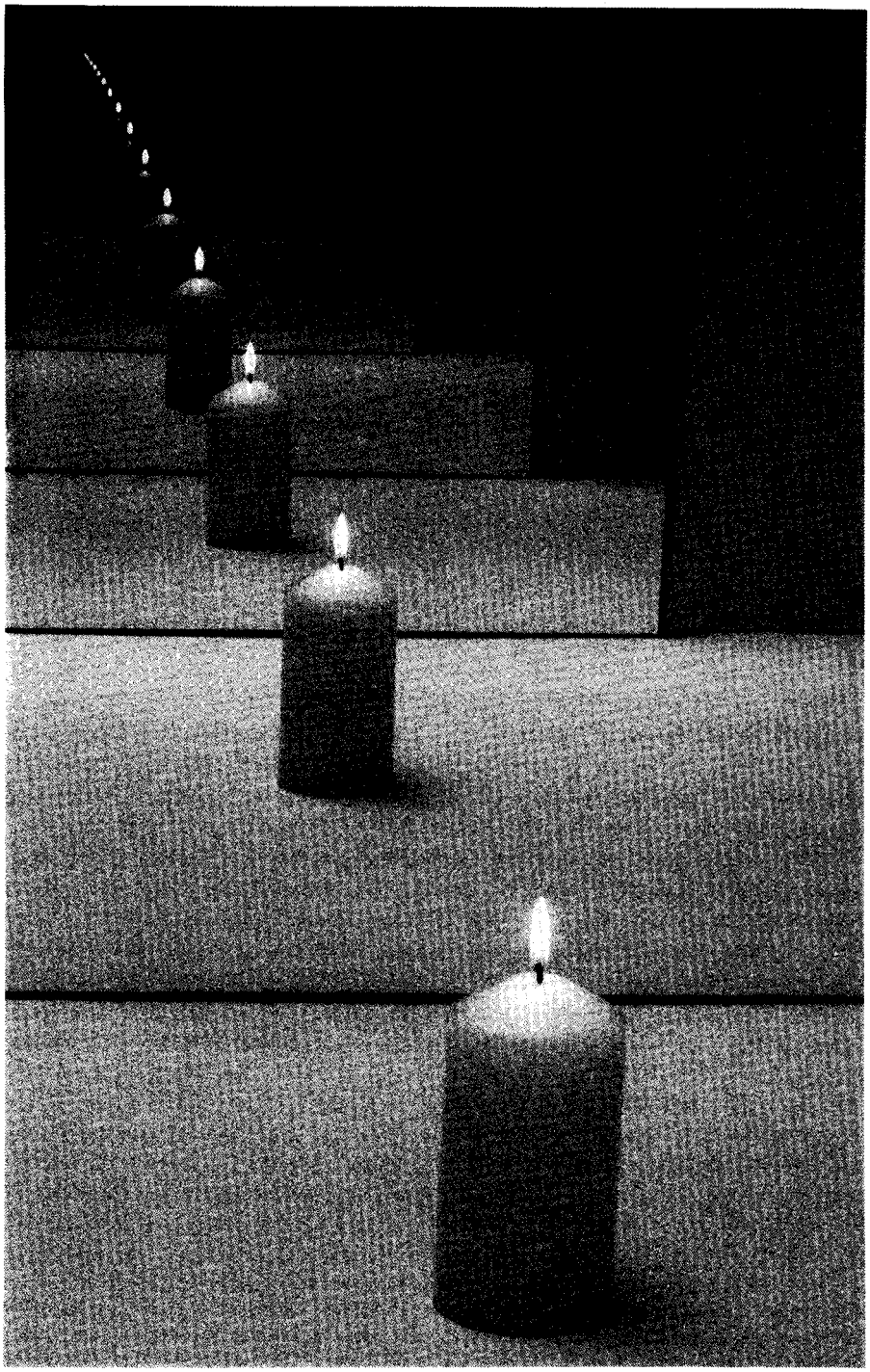


Figure 2 A long row of candles: self-similarity induced by parallel mirrors.

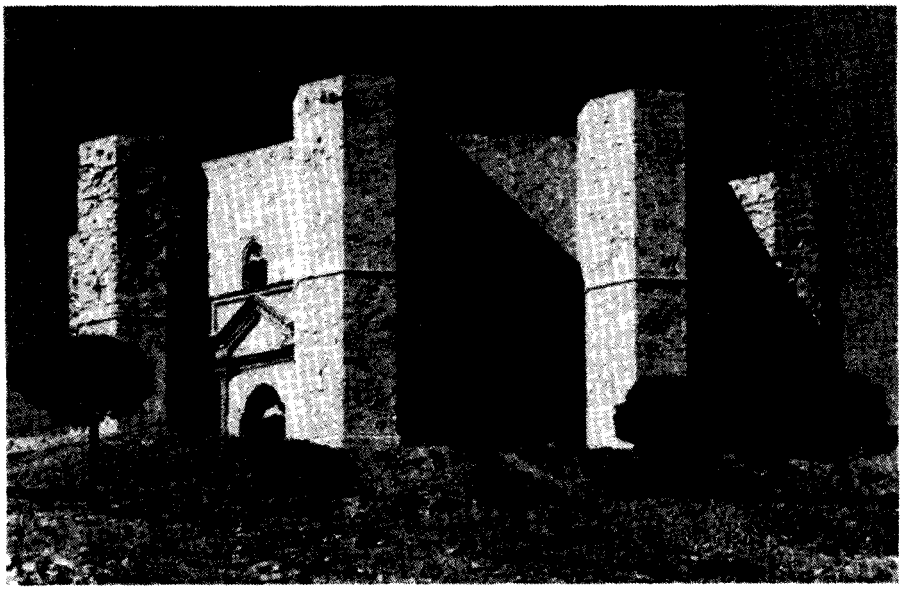


Figure 3 Castel del Monte, an emperor's early attempt at self-similarity.

distorted and weakened, because no real mirror is perfectly planar (or 100 percent reflective).

The result of an early attempt at self-similarity in architecture is the monumental Castel del Monte (Figure 3), designed and built by Holy Roman Emperor—also King of Sicily, Germany, and Jerusalem—Frederick II (1194–1250), the great falconer, rare mathematician (among medieval emperors, anyhow), and last but not least, irrepressible castle builder. The basic shape of the castle is a regular octagon, fortified by eight mighty towers, again shaped like regular octagons. (These towers were devised for the easy release and retrieval of hunting falcons.)

Ironically, Frederick himself was born in a tent, hastily erected in the market square of the little town of Iesi in the Italian Marches near Ancona on the Adriatic Sea.[1] Frederick, who of course wrote and illustrated the book on falconry [Fri 1240], was graced with one of the most stupendous minds of his time (hence

1. Frederick's birth in a tent may have been in*tent*ional rather than unexpected because his mother, the Norman queen Constance, had taken "forever" to become expectant with the sorely awaited future sovereign, the prospective Hohenstaufen ruler. Thus, to forestall any suspicion of double-dealing, the precious child was delivered in the most public of places, with no double walls and no false bottoms to conceal a surrogate mother. As an adult, Frederick revisited Iesi, which he called his "Bethlehem" (he was born during Christmas).

his Latin epithet *stupor mundi*). He grew up in Palermo and Apulia, speaking and reading mostly Arabic, Greek, and Latin. He later introduced—by way of poetry—the Italian *volgare* spoken by the people at the imperial court, thereby strengthening its linguistic links with Italy and elevating the vulgar language to official status. Frederick was also a friend of Fibonacci (Leonardo Pisano), furnishing the latter with algebraic problems whose solutions became part of the history of mathematics.

In music, too, self-similarity abounds. Adjacent notes of a well-tempered scale are in a constant frequency ratio ($2^{1/12}$ for an octave divided into 12 semitones). Thus, because of the inverse proportionality between resonant frequency and the length of a tube, clusters of organ pipes exhibit self-similarity. Less pleasing (and more dangerous) but self-similar nevertheless is the perspective of railroad tracks and ties receding to a distant vanishing point (in Figure 10, on page 93, this self-similarity is exploited in a proof of the formula for the sum of an infinite geometric sequence).

Other examples of discrete self-similarity are "log-periodic" antennas (see Figure 4), which cover a wide spectrum of wavelengths in many discrete steps. Note that both the lengths of the adjacent dipoles and their spacings are scaled by the same similarity factor. Thus, except for end effects, these antennas cover a wide range of wavelengths with nearly the same sensitivity and spatial resolution. TV antennas adorning our roofs are but poor cousins of the antenna shown in the illustration; but they, too, must capture many different channels with equal gain and clarity.[2]

Another kind of wideband "antenna," albeit for sound, is the *basilar membrane*, the frequency analyzer in our inner ears. Different frequencies excite different places along the basilar membrane. This resonating membrane therefore effects a mapping from frequency to place. To cover the enormous frequency range of human hearing, from 20 Hz to 20,000 Hz, without unduly compressing the space available for the important low and middle frequencies, the ear must map frequencies on a *logarithmic* scale. In fact, above about 600 Hz, constant frequency *ratios* correspond to constant shifts in the locations of the resonances along the basilar membrane. In the frequency range from 600 Hz to 20,000 Hz, frequency ratios and places (i.e., the locations of the resonance) scale almost perfectly, the scaling factor being 5 mm along the basilar membrane per octave.

There is another good reason for this logarithmic mapping of frequency to place. It means that the *relative* change with place of the parameters (mass density, stiffness) controlling resonance frequency is constant along the basilar membrane. The basilar membrane therefore behaves like an exponential acoustic horn, such as the horn in a woofer loudspeaker, for example, thus minimizing the reflection

2. Unfortunately, the abovementioned end effects that limit exact self-similarity often manifest themselves at the expense of public television channels occupying the less sensitive band edge (such as channel 13 for the VHF band in the United States).

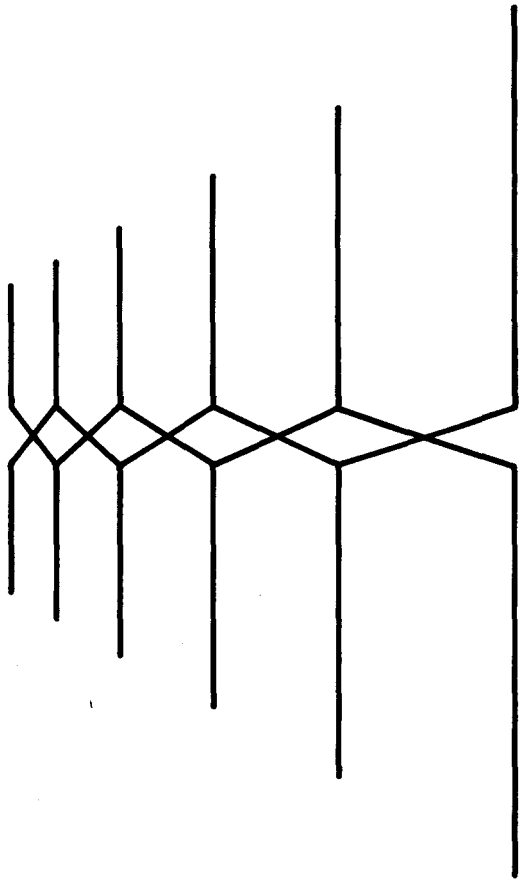


Figure 4 Self-similar TV antenna.

of acoustic energy for a given total length and frequency range. This must have been an important design consideration in the evolution of our ears to a highly sensitive acoustic receptor. (The healthy ear is almost able to detect the Brownian motion of the air molecules!)

For the physicist, the fact that the mechanical parameters change relatively slowly along the length of the basilar membrane—indeed, as slowly as possible for a given total length and frequency range—means that he can analyze wave motion on the basilar membrane in terms of the convenient Wentzel-Kramers-Brillouin approximation [ZLP 76]. This useful method was invented by Joseph Liouville (1809–1882) and rediscovered by the three named authors, who introduced it into quantum mechanics.

Below 600 Hz, the mapping between place and frequency itself (not frequency ratio) is linear. Otherwise, the five octaves below 600 Hz would be given the same space on the basilar membrane as the five octaves above 600 Hz. Since the density of auditory neurons along the basilar membrane is approximately constant, the neural representation of the high frequencies would thus suffer relative to the low frequencies.

The response of the basilar membrane shows another interesting scaling behavior that obtains for its entire frequency range: local wave *velocity* is proportional to local resonance *frequency*. The constant of proportionality is a characteristic length that equals about 1 mm. As a consequence the *delays* of acoustic signals along the basilar membrane are *inversely* proportional to place of detection. This scaling behavior leads to a simple integrable mathematical model of the basilar membrane [Schr 73].

Hierarchical structures, such as phylogenetic trees, for example, often show self-similarity, as do mathematical Cayley trees (also called *Bethe lattices* in physics). A Cayley tree is defined as a graph without loops in which each node has the same number of branches (namely, two, in Figure 5). The self-similarity of such graphs is not necessarily manifest in their geometric representation, but is seen

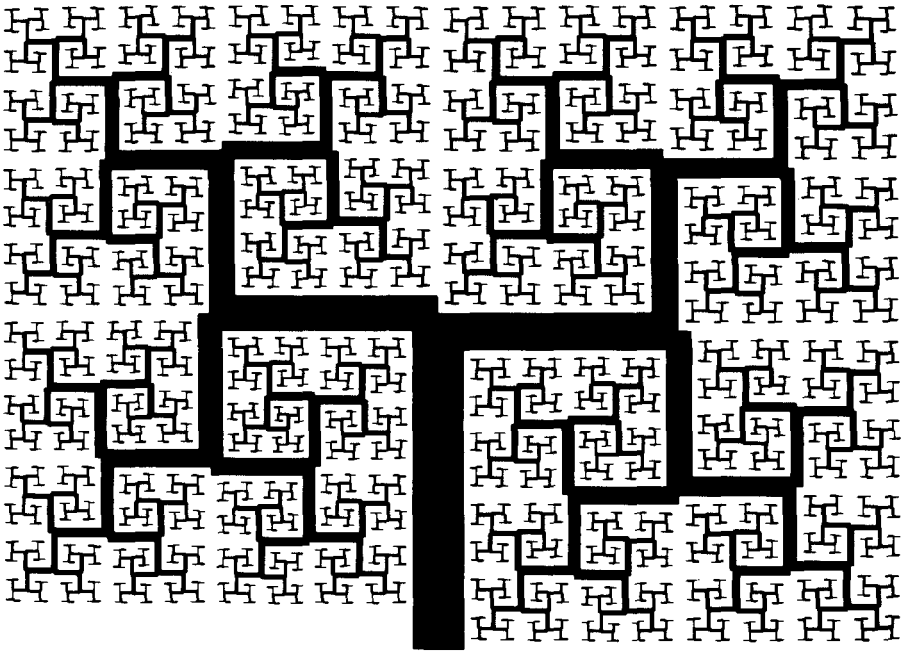


Figure 5 Artistic Cayley tree with 1:2 branching ratio [Man 83].

in their connectivity or topology. Bethe lattices, although highly unphysical. often afford the only exactly solvable models in difficult situations such as Anderson localization and percolation (important concepts in contemporary physics; see Chapter 16).

Figure 6 shows another example, a binary-code tree. It is interesting to note that if we define the distance between two "leaves" (endpoints) by how many generations we have to go back to find a common ancestor, then the space so generated is *ultrametric*. For phylogenetic trees, this means that either the three distances between three existing species are all equal or two are equal and the third is smaller. In other words, all triangles in such a space are either equilateral or short-base isosceles, with interesting consequences in Hensel codes and error-free computing [Schr 90].

If one can identify an ultrametric space in a given problem, there is usually a hierarchical structure lurking behind it that holds the key to better understanding; such structures are found in problems from taxonomy to statistical physics and

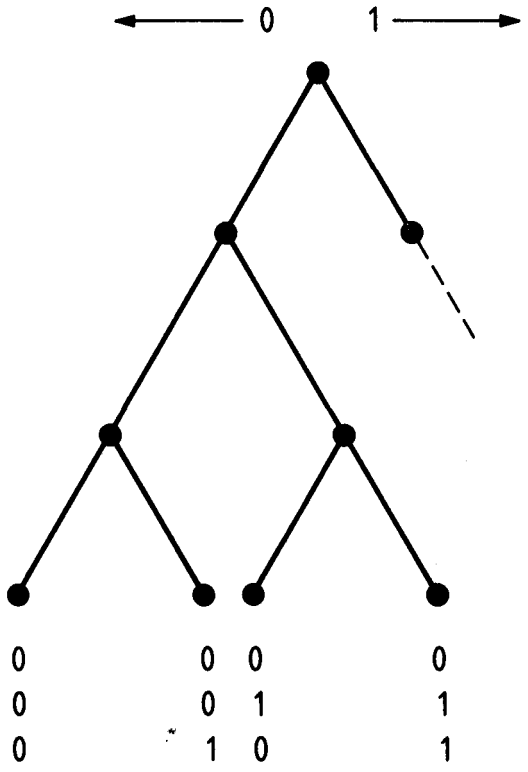


Figure 6 Binary-code tree: a self-similar hierarchical structure.

optimization theory. For an excellent overview of ultrametricity for the mathematical amateur, see the paper by Ramal, Toulouse, and Virasoro [RTV 86].

The Logarithmic Spiral, Cutting Knives, and Wideband Antennas

A charmingly simple example of a self-similar object, with a practical application to boot, is the *logarithmic spiral*—well known from high school mathematics. In polar coordinates (r, ϕ) we have $r(\phi) = r_0 \exp(k\phi)$, where $r_0 > 0$ and k are constants. Scaling the radius vector r, that is, the size of the spiral, by a factor s results in the same spiral rotated by a constant angle $(\log s)/k$. Since the angle ϕ is defined only modulo 2π, scaling factors equal to $s = \exp(2\pi mk)$, where m is an integer, leave the infinite spiral invariant—that is, the logarithmic spiral is self-similar, with a similarity factor $s = \exp(2\pi|k|)$. If we disregard rotations, the logarithmic spiral is self-similar for *any* real scaling factor.

The self-similarity of the logarithmic spiral has several interesting consequences and useful applications. For one, the direction of the tangent to the spiral depends only on the angle ϕ and not on which branch of the spiral one considers (see Figure 7). This follows directly from the scaling invariance but can also be verified, of course, by a more circuitous calculation. Furthermore, since the rotated logarithmic spiral is similar to itself, the angle β between the radius vector and

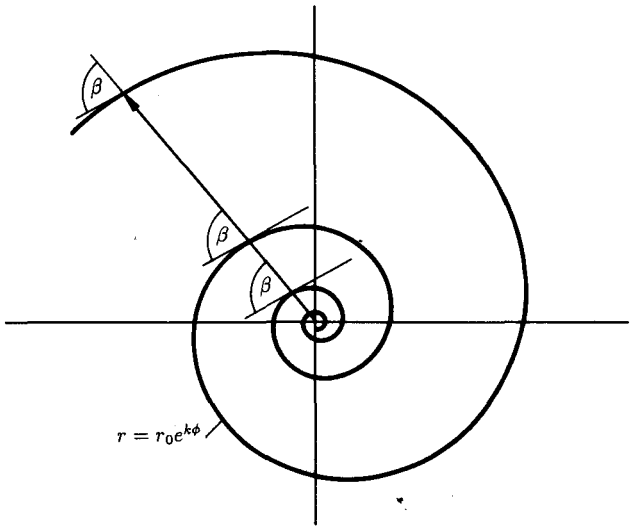


Figure 7 Logarithmic spiral: a *smooth* self-similar curve.

the tangent at any point must be the same for the entire spiral. A little thought (or a small sketch) shows that cot β must equal $(dr/d\phi)/r = d(\log r)/d\phi = k$. So *that* is the geometric meaning of k: it is the cotangent of the constant angle between the radius vector and the tangent at any point. (For $k = 0$, β equals $\pi/2$ and the spiral has degenerated into a circle.)

The just-stated property of the logarithmic spiral has an interesting application. Suppose you had to design a cutting tool with a rotating knife in the form of a disk. Which form should the knife have so that the cutting angle is constant, independent of the angle of rotation of the knife? Why, the perimeter of the knife should follow a logarithmic spiral! Since the perimeter of the knife is necessarily a single-valued function of the rotation angle, it must, of course, "jump back," by an amount $r_0 \exp(2\pi|k|)$, at some angle.

But things can be even more exciting. As we discovered, the logarithmic spiral is self-similar with *arbitrary* scale factors if we disregard rotation. This means that, if we do not care about rotations, the logarithmic spiral has *no* length scale—contrary to the impression that its mathematical formula, $r(\phi) = r_0 \exp(k\phi)$, imparts because the factor r_0 seems to imply a length scale. However, r_0 can be absorbed into a rotation, as can be seen by writing $r(\phi) = \exp\{k[\phi + (\log r_0)/k]\}$, where $(\log r_0)/k$ is just a constant angle.

Once we have something that is scale-free, we should find any number of useful applications. People are forever baffled by problems engendered by changes in scale or size.[3] Suppose we could construct scale-free footwear, shoes that fit all sizes—what a boon! (But to whom?) For stockings, of course, limited scale-free-ness has in fact been achieved: stretch hose that fit all sizes (or perhaps none).

Another field where scale-free-ness is urgently desired is the design of transmitting and receiving antennas for communication systems that have to cover a wide range of wavelengths, such as the log-periodic antenna shown in Figure 4, which is self-similar at a set of discrete wavelengths. If only such antennas were equally efficient at a *continuous* range of wavelengths! Well, suppose we use circularly polarized waves; then a rotation of the antenna would have no effect on the antenna's gain and directivity. And if we gave the antenna the form of a logarithmic spiral (ideally, of thin superconducting wire), then it would work equally well for all wavelengths within any desired range [CM 90]. Antennas exploiting this enticing principle do in fact exist. They look like tapered bedsprings.

But nature, too, has exploited the self-similarity of the logarithmic spiral. In the chambered nautilus (see Figure 8), each chamber is an upscaled replica of the preceding chamber with a constant scale factor. As a result, the nautilus will grow along a logarithmic spiral. And not just nature: Even artists have been

3. Think of the *two* holes the thoughtful Sir Isaac (Newton) is said to have cut in his door to accommodate his two different-size pets.

Figure 8 The chambered snail *Nautilus* follows a logarithmic spiral in its self-similar design. (© Edward Weston; from Center for Creative Photography, University of Arizona.)

inspired by spirals. Color Plate 4 shows an infinity of logarithmic spirals in the colors of the rainbow infinitely intertwined.

A logarithmic spiral with a specific value of the scale factor occurs in the following geometric problem. Take a rectangle with sides $a > b$ and cut off a square of side b from one side of the rectangle (see Figure 9). From the remaining rectangle cut off a square of side $a - b$, as shown in Figure 9. For the construction to work as shown, $a - b$ must be smaller than b; that is, a must be smaller than $2b$. At the second stage of construction, the inequality is $2a > 3b$, and at the $(2n)$th stage, the condition that cutting off a square will result in a rectangle whose longer side is the shorter side of the preceding rectangle is $F_{2n-1}a > F_{2n}b$, where the F_k are the Fibonacci numbers. Similarly, $F_{2n}a < F_{2n+1}b$ must hold. For the construction to carry through for arbitrarily large n, the side ratio b/a of the initial rectangle must equal the limit as $n \to \infty$ of F_{2n-1}/F_{2n}, that is, the golden mean $\gamma = (\sqrt{5} - 1)/2$.

The result of this construction is a limitless spiral of ever smaller squares with a scaling factor equal to the golden mean. The logarithmic spiral with $k = -(\pi/2) \log \gamma$, passing through successive cutoff points, is also shown in the illustration. The vanishing point of the squares and origin of the spiral is given by the common intersection of the diagonals of the rectangles.

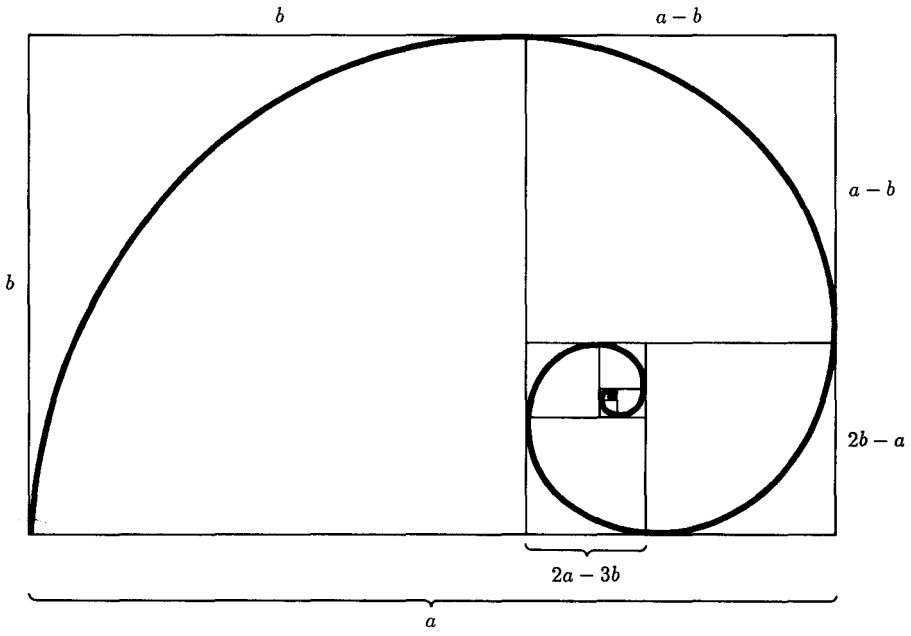


Figure 9 A self-similar succession of golden-mean rectangles.

There is a charming connection between our amputating of squares from residual rectangles and continued fractions. Observe that the continued fraction for the golden mean is $1/(a_1 + 1/(a_2 + \cdots))$, usually abbreviated as $[a_1, a_2, \ldots]$, with $a_1 = a_2 = \cdots = 1$. Thus $\gamma = [1, 1, \ldots]$. Suppose we want to be able to cut off *two* squares at each stage so that the longer side of the remaining rectangle equals the shorter side of the preceding rectangle. What is the appropriate side ratio b/a of the initial (and all subsequent) rectangles? A little experimentation will show that b/a must equal $\sqrt{2} - 1$, which has the continued fraction expansion $[2, 2, 2, \ldots]$. In general, the nth term a_n in the continued fraction of b/a tells us how many squares we must cut off at the nth stage. Thus, self-similar cascades and logarithmic spirals emerge for initial rectangles whose side ratio b/a equals a periodic continued fraction with period length 1. The resulting irrational numbers $[n, n, \ldots]$ with $n > 1$, which I have called *silver means*, also play a role in the construction of quasiperiodic lattices (see Chapter 14) and the modeling of quasicrystals (see Chapter 13).

By the way, the logarithmic spiral, like the infinite straight line, is a specimen of a self-similar object that is *smooth*, in stark contrast to the fractals that we usually associate with self-similarity such as rocky coasts, Brownian motion, and other nondifferentiable functions.

Some Simple Cases of Self-Similarity

One of the simplest self-similar entities is a two-sided infinite geometric sequence, for example,

$$s_n = \ldots, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \ldots$$

Multiplying each member of this sequence by a factor of 2 produces

$$2s_n = \ldots, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, \ldots$$

which is of course the same sequence. The factor of 2 is called the similarity factor or *scaling factor*. Obviously, together with 2 all integer powers of 2 are also scaling factors, including $\frac{1}{4}$, 8, and $\frac{1}{128}$. Of all those scaling factors, we shall call the smallest factor whose magnitude exceeds 1 the *primitive* scaling factor. In our example, 2 is the primitive scaling factor; 4 is not primitive. (And $\frac{1}{2}$, although capable of generating all scaling factors, is excluded by our definition because it is not greater than 1.)

In many practical applications, the self-similar sequence will be only one-sided; for example,

$$1, r, r^2, r^3, \ldots$$

This self-similar sequence is illustrated geometrically in Figure 10, where it is

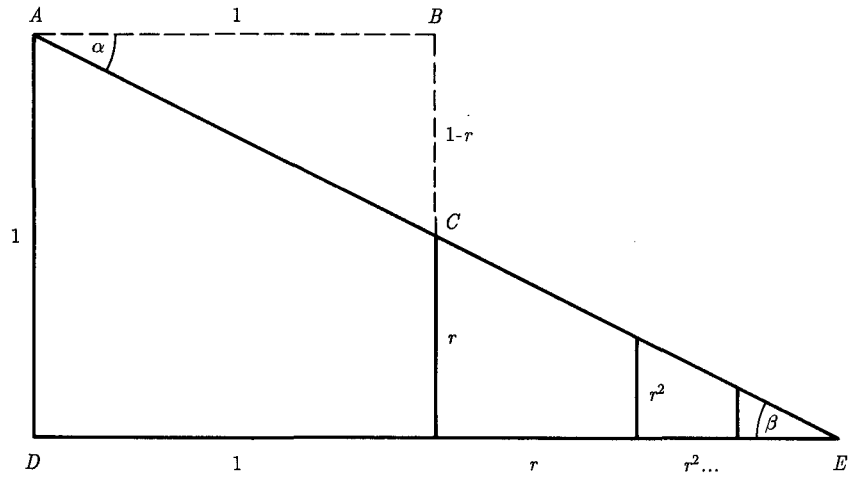


Figure 10 Proof by self-similarity [KB 88].

exploited in a lovely "look-see" proof, due to Benjamin Klein and Irl Bivens [KB 88], of the formula for the sum of geometric series:

$$1 + r + r^2 + r^3 + \cdots = \frac{1}{r - 1} \qquad (|r| < 1) \tag{1}$$

Note that the two isosceles triangles ABC and EDA are similar because the angle α equals β. Thus the ratios of corresponding sides must be the *same*. Specifically, the side ratio $\overline{DE}/\overline{DA} = 1 + r + r^2 + r^3 + \cdots$ must equal $\overline{BA}/\overline{BC} = 1/(1 - r)$. End of proof. (But how would the figure have to be drawn for $r < 0$?)

Another geometric proof of equation 1 relying on self-similarity, due to Warren Page, is shown in Figure 11 [Pag 81]. Starting with the unit square, we cut off a rectangle with side $0 < q < 1$. From the remaining rectangle we cut off a cascade of rectangles, each having an area smaller by a factor $1 - q$ than

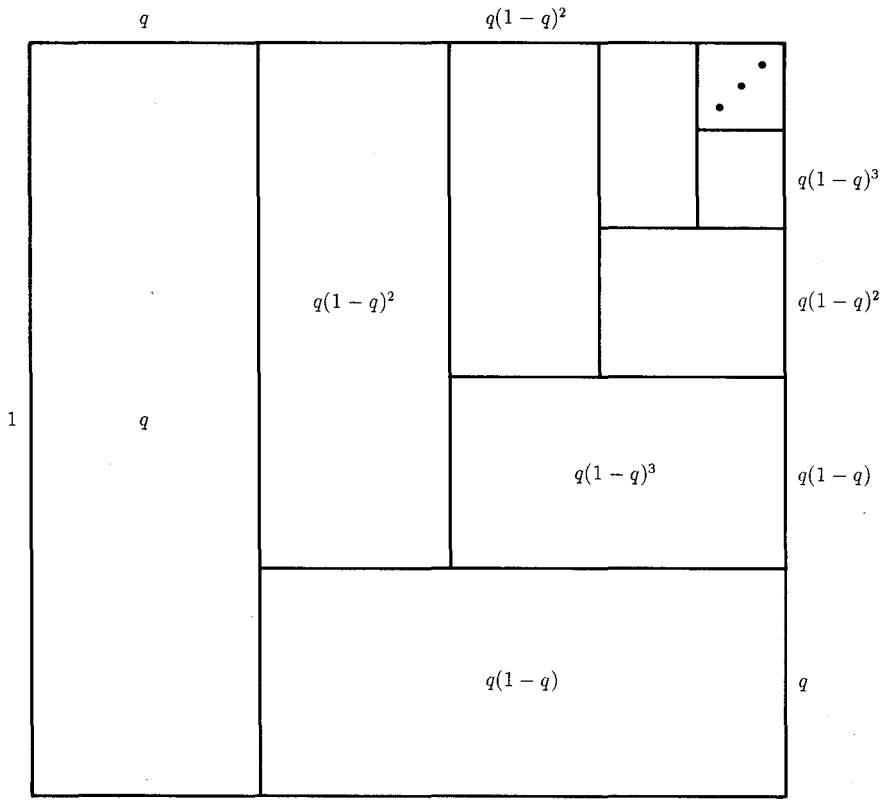


Figure 11 Another proof by self-similarity [Pag 81].

the preceding rectangle. Since all the rectangles together cover the original unit square, we have $q + q(1 - q) + q(1 - q)^2 + \cdots = 1$. Setting $1 - q = r$, we obtain equation 1.

Most readers, I am sure, are familiar with three-way light bulbs and benefit from the three different levels of brightness they offer: low, medium, and bright. Of course, most commercial bulbs accomplish this with just *two* filaments, consuming separately x watts and y watts, respectively; and $x + y$ watts when both are turned on together. For typical values, say, $x = 50$ watts and $y = 100$ watts, the bright condition (150 watts) unfortunately is not much brighter than the medium one (100 watts). It would be preferable if the three wattages formed a self-similar geometric progression; that is, for $y > x$, $(x + y)/y$ should equal y/x. The solution involves the golden mean $\gamma = (\sqrt{5} - 1)/2 = 0.618$. Indeed, for $x/y = \gamma$, $y/(x + y)$ will likewise equal γ.

Similarly, with three filaments, one can realize five different wattages that form a self-similar sequence if the third filament has a wattage $z = y/\gamma^2$. With $y = 100$ watts, for example, the self-similar wattages, rounded to the nearest integer watt, are $x = 62$, $y = 100$, $x + y = 162$, $x = 262$, and $y + z = 424$ watts.

Of course, the reader reading by the light of a multifilament bulb is not really interested in self-similar physical wattages; he is concerned with subjective *brightnesses*. Fortunately, over a fairly large brightness range, brigntness B follows a simple power law, $B \sim W^{\alpha}$, as a function of wattage W. Since power laws are themselves self-similar (see Chapter 4), our choice of filaments is still the proper one even if we desire equal *brightness* ratios.

An amusing auditory paradox, based on a finite self-similar sequence of musical notes, was devised by Roger Shepard [She 64]. The paradoxical sound consists of a superposition of 12 notes with each note an octave higher in frequency than its lower-frequency neighbor. Beginning with 10 Hz, the other 11 frequencies in the composite sound are then 20, 40, 80, 160, 320, 640, 1280, 2560, 5120, 10,240, and 20,480 Hz. Increasing all 12 frequencies by a semitone (about 6 percent) will yield a sound with frequency components at 10.6, 21.2, . . . , 10,489, and 21,698 Hz, which will of course sound a bit higher in pitch (in fact, a semitone higher) because all frequencies have been increased by a semitone.

Increasing the frequencies by another semitone will result in a sound still higher in pitch. Repeating the process a few more times will give rise to further increases in pitch. But after 12 increases in pitch, the sound will be indistinguishable from the original sound! (The 10-Hz component present in the original stimulus and the extra component at 40,960 Hz are inaudible.)

By supplying a sufficient number of (inaudible)low-frequency components, Shepard was able to generate a succession of sounds whose pitches increase forever! With a personal computer, connected to a digital-to-analog converter, such sounds are now easy to generate and I encourage interested readers to subject themselves to this weird perceptual paradox. Figure 12 shows a self-similar waveform that has a constant pitch when all frequencies are doubled.

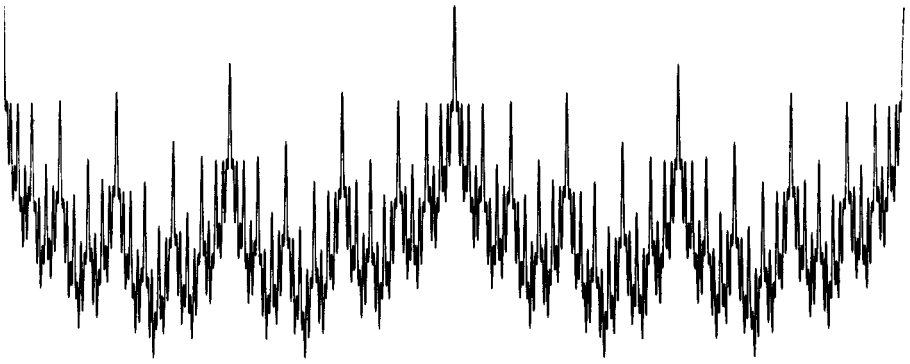


Figure 12 Paradoxical self-similar waveform. Frequency doubling leads to same pitch.

Weierstrass Functions and a Musical Paradox

The ever-ascending tones of Shepard are closely related to the nondifferentiable functions introduced by Karl Weierstrass (1815–1897) as examples of *continuous* functions that are nowhere differentiable. Such functions seemed to defy common sense and were hotly debated in the nineteenth century. A Weierstrass function is defined by

$$w(t) := \sum_{k=1}^{\infty} \alpha^k \cos(\beta^k t) \qquad \alpha \text{ real}, \ \beta \text{ odd}$$

Weierstrass showed that, for $\alpha\beta > 1 + 3\pi/2$, $w(t)$ is a continuous but nowhere differentiable function, such as the fractal Koch flake and Hilbert's space-filling curve, which we first encountered in Chapter 1. Like Cantor sets, nondifferentiable functions are a rich mine of paradoxes, such as Shepard's ever-ascending tones.

Another example of a musical chord patterned after a Weierstrass function can have the following weird property. If recorded on magnetic tape and replayed at *twice* the recording speed, the chord will not sound an octave higher in pitch, as every well-behaved recorded sound would, but a semitone *lower* [Ris 71, Ris 75, Schr 86]. How is this possible? Let us construct a finite-sum approximation to a Weierstrass function (we can omit the factors α^k that are needed to make the infinite series converge):

$$w_K(t) = \sum_{k=1}^{K} \cos(\beta^k t)$$

If we scale the time dimension t by a factor β, we obtain

$$w_K(\beta t) = \sum_{k=1}^{K} \cos(\beta^{k+1} t) = \sum_{k=2}^{K+1} \cos(\beta^k t)$$

That is, except for end effects, $w_K(\beta t)$ equals the unscaled function $w_K(t)$. Thus, $w_K(t)$ is approximately self-similar (see Figure 12, where $K = 7$ and $\beta = 2$).[4] Obviously, in the limit as $k \to \infty$, such a function cannot have a finite nonzero derivative anywhere, because derivatives change with scaling.

Now suppose we select $\beta = 2^{13/12}$ to give

$$w_K(t) = \sum_k \cos(2^{k(13/12)} t)$$

and make $w_K(t)$ audible as a sound. Then the frequencies $2^{k(13/12)}/2\pi$ have to cover only the audio frequency range (10 Hz to 20,000 Hz). Recording $w_K(t)$ on magnetic tape and playing it back at twice the speed produces

$$w_K(2t) = \sum_k \cos(2^{k(13/12)+1} t) = \sum_{k'} \cos(2^{k'(13/12)} \cdot 2^{-1/12} t)$$

where $k' = k + 1$. Now, if these summations cover the entire audio range, then, as far as the human ear is concerned,

$$w_K(2t) = w_K(2^{-1/12} t)$$

Thus, a doubling of the tape speed will produce a sound with a pitch lowered by a factor of $2^{1/12}$. In musical terms, the chord will sound one semitone *lower* rather than an octave higher. Such are the paradoxes engendered by fractals!

It is easy to program a personal computer to produce a $w_k(t)$ with 11 components comprising the frequencies from 10.0 Hz to 18,245.6 Hz. By doubling the playback speed, the sixth component, for example, will change in frequency from 427.15 Hz to 854.3 Hz. But in comparing the two chords, the human auditory system will identify the doubled sixth component at 854.3 Hz with the *nearest* component of the original chord, namely, the seventh at 905.1 Hz. Since 854.3 Hz is a semitone lower than 905.1 Hz and the same argument can be made for all frequency-doubled components, a pitch lowered by one semitone will be perceived.

4. It is interesting to note that the concept of self-similarity entered mathematics at two independent points, Cantor sets and Weierstrass functions, at about the same time and for similar reasons: to elucidate two of the foundations of mathematics, numbers and functions. Even earlier, however, Leibniz had used the concept of self-similarity ("worlds within worlds") in his *Monadology* [Lei 1714] and in his definition of a straight line.

A frequency ratio $\beta = 2^{14/12}$ will produce a lowering by a full note upon tape speed doubling, and $\beta^{15/12}$ will lower the perceived pitch by three semitones, and so on. However, for $\beta = 2^{24/12} = 4$, the percept will be ambiguous, because when the numbers 1, 4, 16, 64, . . . are doubled, the resulting sequence (2, 8, 32, . . .) could be considered either one octave lower *or* one octave higher.

All physical objects that are "self-similar" have limited self-similarity—just as there are no perfectly periodic functions, in the mathematical sense, in the real world: most oscillations have a beginning and an end (with the possible exception of our universe,[5] if it is closed and begins a new life cycle after every "big crunch"; see the admirably equationless bestseller by Stephen Hawking, *A Brief History of Time* [Haw 88]). Nevertheless, self-similarity is a useful abstraction, just as periodicity is one of the most useful concepts in the sciences, any finite extent notwithstanding.

A mathematical object that is self-similar except for an "end effect" is the sinc function

$$\operatorname{sinc} x := \frac{\sin \pi x}{\pi x}$$

which describes the wave diffraction pattern of a rectangular slit and plays an important role in interpolating sampled functions in discrete-time digital systems. (Here the slit is a rectangular "window" through which the function's spectrum is seen.)

By applying the trigonometric identity

$$\sin 2x = 2 \sin x \cos x$$

to the sinc function, we obtain

$$\operatorname{sinc} x = \cos\left(\frac{\pi x}{2}\right) \frac{\sin (\pi x/2)}{\pi x/2} = \cos\left(\frac{\pi x}{2}\right) \operatorname{sinc}\left(\frac{x}{2}\right)$$

Thus, except for the factor $\cos (\pi x/2)$, the sinc function is self-similar with a similarity factor of 2.

By repeating the factoring process, we obtain Euler's famous infinite product

$$\operatorname{sinc} x = \cos\left(\frac{\pi x}{2}\right) \cos\left(\frac{\pi x}{4}\right) \cos\left(\frac{\pi x}{8}\right) \cdots$$

5. A *universe is something that happens once in a while*—according to some of the latest physical phantasies.

The sinc function has zeros for all positive and negative integers x. The first factor in the Euler product produces the zeros of sinc x at all odd values of x. The second factor gives the zeros for x equal to twice an odd number. The third factor creates the zeros at x equal to 4 times an odd number, and so forth, giving the required zeros at all integer values of x except at $x = 0$. (Note that each nonzero integer n can be written, uniquely, in the form $n = 2^m k$, where k is an odd integer and $m \geq 0$.)

In many applications, self-similarity is only approximate; there may be statistical or deterministic "perturbations." Thus, the sequence (based on another well-known infinite product, namely, for $2/\pi$)

$$s_n = \prod_{k=1}^{n} f_k - \frac{2}{\pi}$$

with the recursion

$$f_{k+1} = \left(\frac{1 + f_k}{2}\right)^{1/2} \qquad f_0 = 0$$

is self-similar only in the limit as $n \to \infty$. For $n = 1, 2, 3, 4, \ldots$, we obtain

$$s_n = 0.070482,\ 0.01662,\ 0.004109,\ 0.0001024, \ldots$$

Its terms approach a constant scaling factor of 4. Such asymptotic self-similarities are often encountered in recursive computations.

In subsequent chapters we shall meet the other deviation from strict self-similarity: *statistical* self-similarity, in which the *statistical* laws that govern the object exhibit a scaling invariance. The object itself may change upon scaling, but its probabilistic aspects remain the same.

More Self-Similarity in Music: The Tempered Scales of Bach

The ancient Greeks, with their abundance of string instruments, discovered that dividing a string into two equal parts resulted in a pleasant musical interval, now called the *octave*. The corresponding physical frequency ratio is 2:1.

"Chopping off" one-third of a string produced another pleasant interval, the *perfect fifth*, with a frequency ratio of 3:2.

The Pythagoreans asked themselves whether an integral number of octaves could be constructed from the fifth alone by repeated application of the simple

frequency ratio 3/2. In mathematical notation, they asked for a solution of the equation

$$\left(\frac{3}{2}\right)^n = 2^m$$

in positive integers n and m. But the fundamental theorem of number theory tells us that no positive power of 3 can equal a power of 2, that is, that the equation $3^n = 2^k$ has no integer solutions for $n > 0$.

However, the Greeks were not discouraged and, by trial and error, found an excellent approximate solution:

$$\left(\frac{3}{2}\right)^{12} \approx 2^7$$

which is based on the near equality of $3^{1/19}$ and $2^{1/12}$.

A systematic way of finding such near-coincidences is to write the ratio of the logarithms of the two integer bases (2 and 3) as a continued fraction:

$$\frac{\log 2}{\log 3} = [1, 1, 1, 2, 2, \ldots]$$

where the bracket notation is a convenient way to write the continued fraction

$$\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{2+\cdots}}}}}$$

Continued fractions generally yield good rational approximations to irrational numbers; for example, $\pi \approx \frac{355}{113}$. This excellent approximation to π using not very large integers was known already to the ancient Chinese.

Breaking the foregoing continued fraction off after the fifth term (as shown) yields the excellent approximating fraction for the musical fifth:

$$\frac{\log 2}{\log 3} \approx \frac{12}{19}$$

from which follows $(\frac{3}{2})^{12} \approx 2^7$.

Another important fact here is that the exponents 12 and 7 are coprime, so that repeated application of the perfect fifth modulo the octave (the "circle

of fifths") will not be close to a previously generated frequency until the twelfth step. These 12 different frequencies within an octave are all approximate powers of the basic frequency ratio $1:2^{1/12}$, the *semitone*. Thus, there is always some value of k for which

$$\left(\frac{3}{2}\right)^k \approx 2^{r/12} \qquad r = 1, 2, 3, \ldots$$

The solution of this approximate equation is $k \equiv r/7 \equiv 7r \bmod 12$. One-third of the octave, or $2^{1/3}$ ($r = 4$, frequency ratio ≈ 1.260), for example, is equivalent (modulo the octave) to $k = 4$ fifths (frequency ratio ≈ 1.266).

The third part of the octave, is also close to the pure major third (frequency ratio 5/4). This is the lucky result of another, *independent*, number-theoretic near-coincidence, $5^3 \approx 2^7$, relating the next prime number above 3, namely 5, with the smallest prime, 2.

The fifth itself is approximated by seven semitone intervals with an accuracy of 0.1 percent: $2^{7/12} \approx 1.4983$. The resulting shortfall from the exact value 1.5 is called the *Pythagorean comma*. It is interesting to note that not only do seven semitones make one fifth, but, modulo the octave, seven fifths make one semitone.[6] This coincidence results from still another number-theoretic fluke, namely, that 7 is its own inverse modulo 12: to wit, $7 \cdot 7 = 49 \equiv 1 \bmod 12$.

To ensure that fixed-note instruments, such as the piano, can be played in many different musical keys, the frequencies of the different keys should be selected from the same basic set of frequencies. This led to the development of Bach's tempered scales, based on the semitone with a frequency ratio of $2^{1/12}$. A musical instrument tuned according to the tempered scale thus has frequencies approaching the following multiples of the lowest note:

$$1, 2^{1/12}, 2^{2/12}, 2^{3/12}, 2^{4/12}, 2^{5/12}, \ldots$$

up to some highest note.

Thus we see that the frequencies of a well-tempered instrument form a self-similar sequence, with the similarity factor $2^{1/12}$. If all these notes were sounded *simultaneously*, the instrument would produce an acoustic output (to put it mildly) that approximates a self-similar Weierstrass function. (Actual tuning of pianos differs from exact self-similarity, the tuning being somewhat "stretched" to minimize the beating of overtones, which are not precisely harmonic, as a result of the finite bending stiffness of the strings.)

For an excellent introduction to the science of musical sound see the book of that title by John R. Pierce [Pie 83].

6. Of course, seven fifths can also make one "semi*conscious*."

The Excellent Relations between the Primes 3, 5, and 7

John R. Pierce of communications satellite fame asked himself a few years ago whether one could not replace the frequency ratio 2:1 of the octave by the frequency ration 3:1, which he called the *tritave*, and design a self-similar (equal-tempered) scale that matches frequency ratios constructed from the *next* two prime numbers, namely, 5 and 7 [MPRR 88]. In other words, is there some integral root of 3, $3^{1/N}$, such that $\frac{5}{3}$ and $\frac{7}{5}$ are well approximated by integer powers of $3^{1/N}$—in analogy to the good approximations $\frac{3}{2} \approx 2^{7/12}$ and $\frac{5}{3} \approx 2^{9/12}$ for Bach's well-tempered scale?

To answer this question in a systematic way, we have to expand the ratios log 3/log 5 and log 3/log 7 into continued fractions. This yields for the first ratio

$$\frac{\log 3}{\log 5} = [1, 2, 6, \ldots]$$

Breaking the continued fraction off after the 6 gives the following rational approximation:

$$\frac{\log 3}{\log 5} \approx \frac{13}{19}$$

or $3^{1/13} \approx 5^{1/19}$. This means that the basic frequency ratio $3^{1/13} = 1.088\ldots$ is a good "semitone" for constructing a musical scale that closely matches, modulo the tritave, notes that are generated from the frequency ratio 5:3. In fact, $3^{6/13}$ equals $\frac{5}{3}$ within 0.4 percent!

But what about the frequency ratio 7:3? Continued fraction expansion of log 3/log 7 gives the excellent approximation $3^{1/13} \approx 7^{1/23}$. *Again* $3^{1/13}$ emerges as the preferred basic interval to construct the well-tempered Pierce scale—another number-theoretic fluke! In fact, the match between $\frac{7}{5}$ and a power of $3^{1/13}$, namely, $3^{4/13}$, is uncannily close: the difference is but 0.16 percent. The resulting frequency ratios again form a self-similar sequence:

$$1, 3^{1/13}, 3^{2/13}, 3^{3/13}, \ldots$$

But while the number theory of this new musical scale may be nearly perfect, its compositional value is a matter of taste and open to debate.[7]

7. This patient listener, who has served as a subject in musical tests involving the Pierce scale, showed a stubborn preference for compositions written in well-established traditional scales.

4

CHAPTER

Power Laws: Endless Sources of Self-Similarity

What really interests me is whether God had any choice in the creation of the world.

—ALBERT EINSTEIN

Similarity not only reigns in plane geometry, as in Einstein's triangles (see Chapter 1), but underlies much of algebra too. Think of a homogeneous power function such as

$$f(x) = cx^{\alpha}$$

where c and α are constants. For example, for $\alpha = 1$ we get the special case $f(x) = cx$, which, for $c < 0$, describes the restoring force of a linear spring; for $\alpha = -2$ (and c still negative) we get Newton's law of gravitational attraction $f(x) = cx^{-2}$. Such simple power laws, which abound in nature, are in fact *self-similar*: if x is rescaled (multiplied by a constant), then $f(x)$ is still proportional to x^{α}, albeit with a different constant of proportionality. As we shall see in the rest of this recitation, power laws, with integer or fractional exponents, are one of the most fertile fields and abundant sources of self-similarity.

The Sizes of Cities and Meteorites

Many objects that come in different sizes have a self-similar power-law distribution of their relative abundance over large size ranges. This is true for objects that

have grown, like cities, as well as for objects that have been shattered, like crushed stones [Mek 90]. The only prerequisite for a self-similar law to prevail in a given size range is the absence of an inherent size scale. (Of course, no earthly "city" has fewer than 1 inhabitant or more than 1 billion, and no stone on earth is smaller than an atom or larger than a continent.)

One of the best behaved shattering mechanisms occurs not on earth but in outer space: the mean frequency with which different kinds of interplanetary debris (shooting stars or meteors) slam into the earth's atmosphere is inversely proportional to the squared diameter of the projectile, and this is true over 10 orders of magnitude (see Figure 1). Whereas the space shuttle is hit at a rate of one particle every 30 microseconds (10^{12} particles per year) with a diameter under 1 micrometer (μm), meteorites of the size that created the Arizona crater, with a diameter of 100 meters or more, are expected (thank heavens!) only once every 10^4 years. And the next "shooting star" of the size that hit Sudbury, Ontario, with an "astronomical" 10,000-meter diameter, should not rock the earth for another 10^8 years.

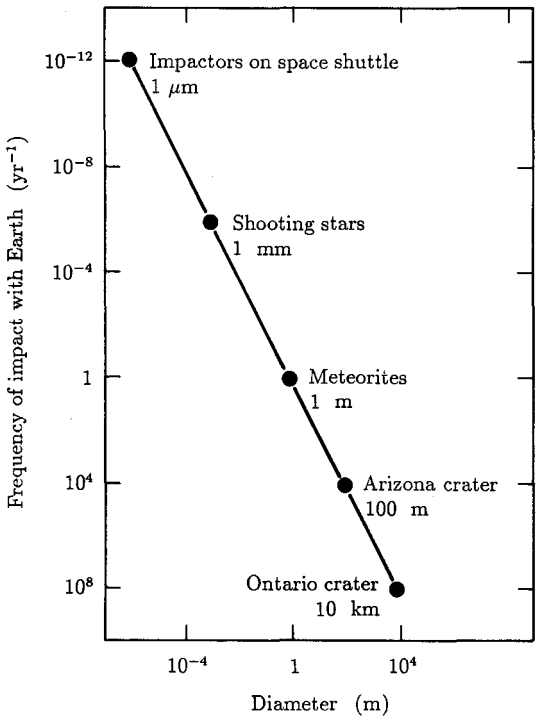


Figure 1 Frequency of meteor collisions with earth in relation to particle diameter. (After E. Schoemaker, U.S. Geological Survey.)

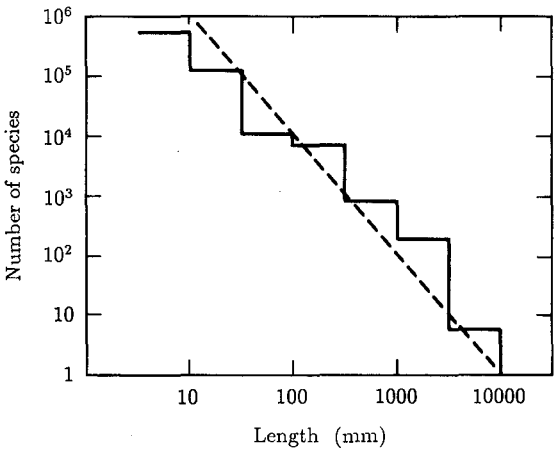


Figure 2 Number of species of terrestrial animals in relation to individuals' lengths [May 88].

Luis Alvarez, whom we already encountered chasing submarines during World War II (see pages 33–34 in Chapter 1), and his associates used the data in Figure 1 to help support his theory of the sudden disappearance of the dinosaurs 65 million years ago. According to Alvarez, the impact of a large meteorite kicked up a lot of sunlight-blocking dust, thereby depriving the dinosaurs of the greenery necessary for their survival [AAMA 82].

Speaking of dinosaurs, one is reminded of the sizes of animals and their distribution. Figure 2 shows the estimated distribution of the number of species of terrestrial animals as a function of their length. Again we find a power law with an exponent of -2 over four orders of magnitude, from 1 millimeter to 10 meters [May 88].

A Fifth Force of Attraction

One of the early laws of physics that shows self-similarity resulted from Galileo's observation that large stones and small stones dropped from the Leaning Tower of Pisa fall with (nearly) equal speed.[1] In fact, ignoring aerodynamic drag, the

1. Actually, Galileo used rolling balls on inclined plains, but the Leaning Tower story has taken on a life of its own.

falling time t is simply proportional to the square root of the dropping height h, independent of the stone's mass: $t \sim h^{1/2}$, a law that is independent of scale—or nearly so, because when one goes to astronomical heights (or even up a tall mountain), the gravitational pull of the earth is diminished. Thus, there *is* a natural length scale that limits the scale invariance or self-similarity of Galileo's $t \sim h^{1/2}$: the radius of the earth. Self-similarity is only *approximate*, a situation that we shall encounter again and again: self-similarity reigns supreme, but only over bounded domains.

In fact, even Newton's just mentioned universal law of gravitation—in full, $f(x) = GMx^{-2}$, where G is the gravitational constant and $f(x)$ is the attractive force of a mass M at distance x—is being called into question these days (even before the advent of a theory of quantum gravity, which will fuse Newton's G and Planck's h). A reexamination of the old attraction data and careful new force measurements seem to have revealed a *nonscaling* correction to Newton's law, called the "fifth force" by some imaginative minds. (The other four forces are gravity, electromagnetism, and the weak and strong nuclear interactions.) The fifth force, whatever its origin and if it is real, appears to depend on distance x as $\exp(-x/x_c) \cdot x^{-2}$, which is *not* a homogeneous power law and therefore not self-similar: it contains a characteristic cutoff length, x_c, as it must because the exponential function calls for a dimensionless argument.

What is the significance of x_c, whose order of magnitude is 100 m? Forces in modern physics are mediated by *particles*, with a rest mass equal to h/cx_c, where h is again Planck's constant and c is the velocity of light. For forces with an infinite range ($x_c = \infty$), as for electromagnetic fields, the rest mass is zero, which is believed to be the case for the electromagnetic particle, commonly called the photon. In fact, efforts to establish an upper limit for the rest mass of the photon focus on the *range* of the electromagnetic field. And the same is true for the rest mass of the neutrino—with potentially enormous consequences for the total weight (and the final fate) of our universe if the neutrino's rest mass should turn out to be nonzero.

The importance of scaling (or rather *non*scaling) with distance was never as dramatically demonstrated as when the Japanese physicist Hideki Yukawa (1907–1981), in the 1930s, concluded from the finite cutoff length of nuclear forces ($x_c \approx 10^{-14}$ m) that a particle, called the *meson*, must exist with a mass of approximately 240 electron masses. A bit later, such a particle was indeed found (in the cosmic showers of particles that "rain" down on earth), but it turned out to be just a heavy brother to the common electron, now called the *muon* (with a mass 207 times that of the electron).[2] Thus, the search for Yukawa's hypothetical meson continued until it *was* found, weighing in at 270 electron masses and baptized pi-meson or *pion*.

2. "Who ordered *that*?" as Isidor Rabi (1898–1987) asked, in an often quoted question, when this completely "unnecessary"particle was first brought to his attention.

Now what particle goes with the fifth force? With $x_c \approx 100$ m, its mass would have to be less than 10^{-13} of the electron's mass, which is already extremely small. Perhaps the apparent range of the fifth force is the result of two (or more) new forces that almost cancel each other. Or maybe there is no new force at all, as mass scaling and the latest experimental results would suggest [TKFFHKMM 89, BBFPT 89, JER 90].

Free of Natural Scales

As we said, homogeneous functions have an interesting scaling property: they reproduce themselves upon rescaling. This scaling invariance can shed light into some of the darker corners of physics, biology, and other sciences, and even illuminate our appreciation of music.

Scaling invariance results from the fact that homogeneous power laws lack natural scales; they do not harbor a characteristic unit (such as a unit length, a unit time, or a unit mass). Such laws are therefore also said to be *scale-free* or, somewhat paradoxically, "true on all scales." Of course, this is strictly true only for our mathematical models. A *real* spring will *not* expand linearly on all scales; it will eventually break, at some characteristic dilation length. And even Newton's law of gravitation, once properly quantized, will no doubt sprout a characteristic length.[3]

This concept of something happening *on all scales* (another much liked locution) is one of our central themes. In fact, it is said (and Mandelbrot was perhaps the first to say it emphatically enough) that for mountainscapes to be interesting they must have features (cliffs, crevices, peaks, and valleys) *on many length scales.* And music, written in any scale, to be appealing, had better have pitch changes on many frequency scales and rhythm changes on more than one time scale. This is in fact how the Bachs (J. S. et al.) composed their music, although they never said so.

Bach Composing on All Scales

When Johann Sebastian Bach (1685–1750) composed his *Brandenburg* Concertos he was, unwittingly no doubt, using homogeneous power functions in the selection of his notes [VC 78]. The power spectrum (the squared magnitude of the

3. Heisenberg heralded the appearance of a new constant of nature, a characteristic *length*, in the basic laws of physics more than 50 years ago, but it still has not happened. (The Planck length, $(G\hbar/c^3)^{1/2} \approx 10^{-35}$ m, which governed the "big bang" that may have created our universe, is but a derived entity.)

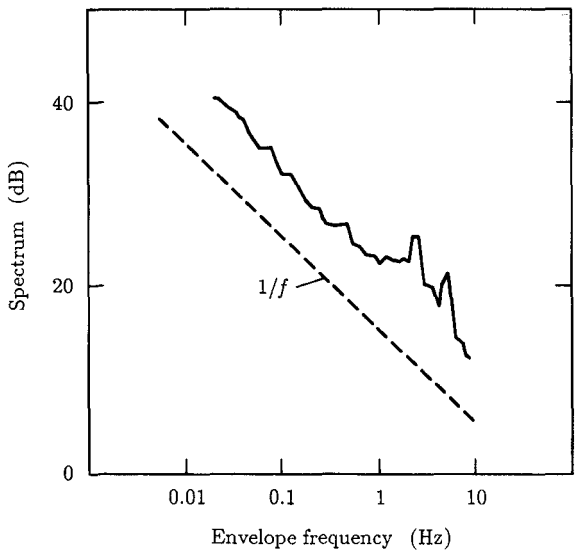


Figure 3 Spectrum of amplitude variations for Bach's First Brandenburg Concerto [VC 78].

Fourier transform) $f(x)$ of the relative frequency intervals x between successive notes can be approximated over a large range by a homogeneous power function with an exponent of -1:

$$f(x) = cx^{-1}$$

which is also called a *hyperbolic law* (because its plot looks like a hyperbola). Taking logarithms yields

$$\log f(x) = \text{const} - \log x$$

where x is in semitones. Thus, on a doubly logarithmic plot, $\log f$ versus $\log x$, the data follow a straight line with a slope of $-45°$.

Not only does the spectrum of the frequency intervals follow a homogeneous power law, but the spectrum of the *amplitudes* (instantaneous "loudness") of Bach's music follows a homogeneous power law with the same exponent (see Figure 3). The amplitude of the music is obtained by temporal smoothing[4] of

4. The smoothing, which could falsify the result, can be circumvented by taking Hilbert transforms and computing the so-called *Hilbert envelope*. The Hilbert envelope of a function is defined as the envelope of the *family* of functions generated by phase-shifting the Fourier transform of the given function through all angles from 0 to 2π.

the magnitude of the sound pressure recorded near the orchestra. Note that the time scale for these amplitude or envelope variations extends to 100 s, corresponding to 0.01 Hz.

Why is it that Bach should have chosen the simple hyperbolic power law when composing his music? Well, first one has to say that he (and countless other composers) did nothing of the kind. Composers compose to create interesting music. So the real question should perhaps be, Why does (at least some) interesting music have hyperbolic spectra for frequency intervals and amplitudes?

Birkhoff's Aesthetic Theory

A partial answer may come from the *"theory of aesthetic value"* propounded by the American mathematician George David Birkhoff (1884–1944). Birkhoff's theory, in a nutshell, says that for a work of art to be pleasing and interesting it should neither be too regular and predictable nor pack too many surprises. Translated to mathematical functions, this might be interpreted as meaning that the power spectrum of the function should behave neither like a boring "brown" noise, with a frequency dependence f^{-2}, nor like an unpredictable white noise, with a frequency dependence of f^0.

In a white-noise process, every value of the process (e.g., the successive frequencies of a melody) is completely independent of its past—it is a total surprise (see Figure 4A). By contrast, in "brown music" (a term derived from Brownian motion), only the *increments* are independent of the past, giving rise to a rather boring tune (see Figure 4B). Apparently, what most listeners like best, and not only in Bach's time, is music in which the succession of notes is neither too predictable nor too surprising—in other words, a spectrum that varies according to f^{α}, with the exponent α between 0 and -2. As Richard Voss discovered, the exponents found in most music are right near the middle of this range: $\alpha = -1$, giving rise to the hyperbolic power law f^{-1} (see Figure 4C) [VC 78]. Or, as Balthazaar van der Pol once said of Bach's music, "It is great because it is inevitable [implying $\alpha < 0$] and yet surprising [$\alpha > -2$]." (I found this quotation in Marc Kac's captivating autobiography, *Enigmas of Chance* [Kac 85].)

Figure 5B shows a sample of a noise waveform with hyperbolic power spectrum, f^{-1}. Such time functions are also called *pink* noise, because they are intermediate between brown(ian) (f^{-2}) and white (f^0) (see Figure 5C and 5A, respectively). Since the power spectrum of any noise that obeys a homogeneous power law (f^{α}) is self-similar, the underlying waveform must likewise be self-similar. In fact, if the frequency axis of the power spectrum is scaled by a factor r, then, by the law of Fourier reciprocity, the time axis of the corresponding

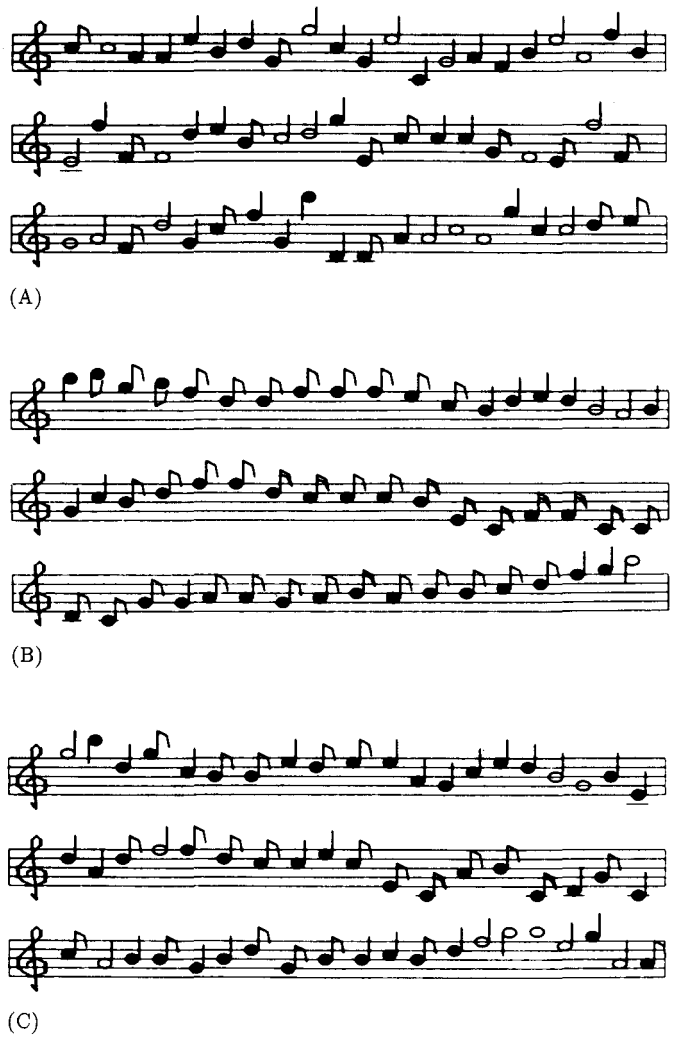


Figure 4 (A) "White" music produced from independent notes; (B) "brown" music produced from notes with independent increments in frequency; and (C) "pink" music—frequencies and durations of notes are determined by $1/f$ (pink) noise [VC 78].

waveform is scaled by $1/r$. Of course, in the case of noise (and other probabilistic phenomena), the self-similarity is only *statistical*; a magnified excerpt is not an exact, deterministic replica of the unscaled waveform.

Also, to preserve power when rescaling frequencies, amplitudes should be adjusted by a factor $r^{-\alpha/2}$. Strictly speaking, such stochastic processes are therefore

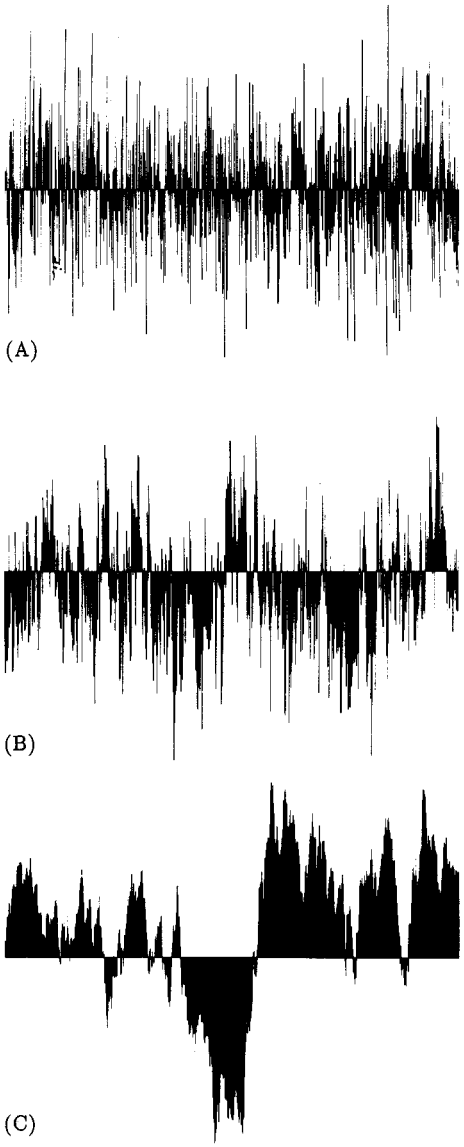


Figure 5 Sample of (A) white noise with f^0 power spectrum; (B) "pink" noise with $1/f$ power spectrum; and (C) "brown" noise with $1/f^2$ power spectrum.

self-affine—that is, they have more than one scaling factor: r for frequencies (or, equivalently, $1/r$ for times) and $r^{-\alpha/2}$ for amplitudes.

In fact, a pink (or white, or brown) noise is the very paradigm of a statistically self-similar process. Phenomena whose power spectra are homogeneous power[5] functions lack inherent time and frequency scales; they are *scale-free*. There is no characteristic time or frequency—whatever happens in one time or frequency range happens on *all* time or frequency scales. If such noises are recorded on magnetic tape and played back at various speeds, they sound the same—unlike a human voice, which sounds like the cartoon character Donald Duck when played at twice the tape speed. There are even self-similar tones that go *down* in pitch when the tape speed is doubled (see pages 96–98 in Chapter 3). We shall further explore noises of different colors in Chapter 5.

Heisenberg's Hyperbolic Uncertainty Principle

Hyperbolic laws are so widespread that they are often not even recognized as such, especially when they are written as a product that equals a constant. A case in point is Heisenberg's famous uncertainty relation of quantum mechanics:

$$\Delta q \cdot \Delta p \geq \hbar$$

where q and p are two (quantum mechanically) "canonical conjugate" variables such as position and momentum or energy and time, and where $\hbar$ is Planck's constant (divided by 2π).[6] The uncertainty principle says that the smaller the error (in the sense of a statistical standard deviation) in one variable, the greater the error in the conjugate variable. Thus, if someone wants to determine, in a *Gedanken* (thought) *experiment*, the position of, say, an electron very accurately, he has to use photons (light particles) of very short wavelengths, that is, very high momentum, which, after bouncing off the poor electron, leave it in a state

5. Note the double duty that the noun *power* is serving here: power as an exponent, as in *third power*; power as a force, as in *third-rate power* or *nuclear power* (both physical and political)

6. "What *is* this thing called 'h bar'?"—a whispered question overheard at a recent physics conference (between two metallurgists?). It seems that Planck's "quantum of action" needs more time (or energy) to penetrate universal consciousness. In fact, sad to say, this can even be said of some professional physicists. This patient chair of the Göttingen General Physics Colloquium was once witness at a talk in which the speaker (quite matter-of-factly) mentioned that a (camera) shutter chopping up a stream of Mössbauer gamma quanta in time would (of course) widen their energy spectrum. Whereupon one of the *experts* in the field (of Mössbauer spectroscopy!) objected that the shutter, moving only laterally to the quanta, could not *possibly* alter their energy. Well, as Richard Feynman said in 1967, "I think it is safe to say that no one understands quantum mechanics!"

of highly uncertain momentum. This statement is often phrased more qualitatively as "any observation disturbs the system to be observed."

But, however phrased, the uncertainty principle is nothing but a consequence of the well-known reciprocal scaling relationship predicted by Fourier transformation theory for a pair of Fourier variables. (However, the fact that Fourier transformation—and the Hilbert spaces—are the proper domains for quantum systems is anything but trivial; it is one of the deeper insights into the makings of nature so far afforded the human mind.)

The greatest possible accuracy in p and q is achieved if both variables are distributed according to Gauss's normal law of probability, in which case the equal sign in the preceding inequality holds:

$$\Delta q = \frac{\hbar}{\Delta p}$$

Although this minimal-uncertainty relation is not usually thought of as a hyperbolic law, once we have written it as such, we can ask, What is its range of validity? The answer is perhaps one of the more awesome in all of physics: although tested and retested over vast ranges of energy, time, position, and momentum, never has the slight*est* violation of $\Delta q \geq \hbar/\Delta p$ been found. There is no doubt in the mind of physicists that uncertainty, like relativity, is of an absolutely fundamental nature that admits no exceptions. Theories may come and go, but "h bar" will always be with us.

One of the fundamental consequences of uncertainty is the very size of atoms (which, without it, would collapse to an infinitesimal point). In fact, we can calculate the radius of the lightest atom, hydrogen, directly from the uncertainty relation: the potential energy U of the atom is proportional to the reciprocal of the radius. Thus, making the radius smaller increases the magnitude of the potential energy. By the virial theorem (see pages 66–68 in Chapter 2), this engenders a proportional increase in the atom's kinetic energy T, which means that the atom would have to *increase* in size. The minimum or *Bohr radius*, r_{Bohr}, is given by the uncertainty relation with the equal sign, $\Delta x \cdot \Delta p = \hbar$: we simply identify r_{Bohr} with Δx and the momentum corresponding to the kinetic energy with Δp. Using the exact relations for these two energies ($U = e^2/r4\pi\varepsilon_0$ and $T = p^2/2m_e$, where e and m_e are the charge and mass of the electron and ε_0 is the permittivity of the vacuum) yields, in one fell swoop, without the usual extensive calculations, the correct value for the Bohr radius. This radius, also called the *atomic unit of length*, equals, in terms of four fundamental physical constants,

$$r_{\text{Bohr}} = \frac{4\pi\varepsilon_0\hbar^2}{m_e e^2}$$

or $5.3 \cdot 10^{-11}$ m. Thus, the diameter of the hydrogen atom, $2r_{\text{Bohr}}$, equals approximately 10^{-10} m, or 1 angstrom (Å).

The immense range of validity of the uncertainty relation is impressively illustrated by the large *coherence length* of laser light, which is of great importance in holography and refined tests of relativity. With a relative energy uncertainty $\Delta E/E = 10^{-7}$, the "length uncertainty" Δx (i.e., the coherence length) for a helium-neon laser with a wavelength $\lambda = 630$ nm becomes $\Delta x = \lambda E/\Delta E \cdot 2\pi$, or about 1 m—a *macroscopic* length more than 10 powers of 10 larger than the Bohr radius derived from the same principle.

In Mössbauer spectroscopy, we deal with even smaller relative energy uncertainties (10^{-14} or less), yet Heisenberg's principle still holds.

While the coherence time (time uncertainty) Δt for laser light, say, 10^{-9} s, is perhaps not very long, *neutron* spectroscopy makes energy measurements with an incredibly small uncertainty of $\Delta E = 2 \cdot 10^{-36}$ watt-seconds (W · s) possible, because of the macroscopic coherence time for neutron waves of 50 seconds! Again, the hyperbolic uncertainty relation is still firmly entrenched, spanning more than 10 orders of magnitude.

Perhaps the most astounding consequence of uncertainty is the excommunication of nothingness, innocently called *vacuum*, from our worldview. A classical vacuum contains neither matter nor energy. But *zero* energy would be a precise value, and that is forbidden by uncertainty. Thus, the modern, quantum mechanical vacuum has finite energy fluctuations as dictated by Heisenberg's prescription—just as the finite size of the smallest atom in its state of lowest energy is prescribed by the same law.

The reality of vacuum fluctuations is now an integral part of quantum physics, with numerous consequences that are testable with great precision, such as the hyperfine structure of atomic spectra. There are even creditable theories of the origin of our universe as a vacuum fluctuation run amok [Haw 88]. As Thomas Cranmer, archbishop of Canterbury, wrote as early as 1550, "Naturall reason abhorreth vacuum."[7]

For all we know, the range of validity of the uncertainty relation is unlimited. On the other hand, for any homogeneous power law representing energy as a function of frequency (called a *spectrum*), there must, of course, be an upper or a lower limit (or both) beyond which the homogeneous power law cannot hold. White noise, for example, has a flat power spectrum only up to some, possibly very high, frequency. And pink noise must have an upper ("ultraviolet") *and* a lower ("infrared") transition frequency, possibly far apart, beyond which the hyperbolic law breaks down because otherwise the total power (the integral over the spectrum) would be infinite. (Of course, physicists use such laws anyhow—

7. This authentic citation, recalled in the *Oxford English Dictionary* [Bur 87], is not to be confused with a more modern quote, circulated by John Robinson Pierce on the occasion of the transistor's birth (which he so baptized in 1948): "Nature abhors vacuum tubes."

since they scale so well—and then they complain of ultraviolet and infrared "catastrophes.")

Fractional Exponents

Power laws are not restricted to integer exponents as in white, pink, and brown noises. In fact, *fractional* exponents abound in nature. After all, self-similarity prevails for integer and noninteger exponents alike. And not infrequently a fractional exponent contains an important clue to the solution of an intricate puzzle. Often such exponents seem to be the same in rather different situations (such as melting or magnetism, for instance), providing a hint of similar underlying "universal" mechanisms.

A simple example of the appearance in nature of a self-similar law with a fractional exponent is the relation between the radiation density ρ_r and matter density ρ_m in the expanding universe which prevailed shortly after its creation:

$$\rho_r \sim \rho_m^{4/3}$$

From this simple relation Alpher and Herman were able to calculate, back in 1948, the present radiation density of the universe [AH 48]. Given other, albeit defective, data, they predicted a blackbody radiation—a remnant from the big bang that gave birth to the universe and is now bathing it (like a baby?) in a "warm" background—corresponding to a temperature of about 5 kelvins (degrees centigrade above absolute zero). Later Arno Penzias and Robert Wilson, while "tuning" microwave antennas, found this cosmic background radiation with a temperature of 2.7 degrees, and received the 1978 Nobel Prize for their discovery of this ancient "footprint" of the early universe.

In subsequent chapters we shall encounter other simple power laws with fractional exponents showing scaling invariances with far-reaching consequences in a wide variety of real-world situations, from the floods of the Nile to the gambler's ruin and the distribution of galaxies in the universe. In fact, in a surprising number of instances, complicated functions of two or more variables exhibit simple power-law behavior near "critical points." Thus, the function of two variables $f(x, y)$ can very often be written in the generic form

$$f(x, y) = x^{\alpha} g(y/x^{\beta})$$

in which $f(x, y)$, has been replaced by a function of only one variable, g. For any range of the variables over which g is relatively constant, $f(x, y)$ is then approximated by a simple power law in x.

This kind of representation, in terms of power laws and their exponents, is enormously fruitful in the analysis of critical phenomena from percolation (see Chapter 15) to ferromagnetism and superconductivity.

The Peculiar Distribution of the First Digit

Power laws, or relations of the form $f(x) \sim x^{\alpha}$, lead to skewed nonuniform homogeneous distributions of the first (leftmost) digit when the self-similar data are listed numerically. For the exponent $\alpha = -1$, the probability p_m that the most significant digit equals $m > 0$ is given by

$$p_m = \log_b \left(1 + \frac{1}{m}\right)$$

where b is the base of the number system used [Pin 62, Rai 76]. For decimal data, these probabilities are approximately $p_1 = 0.301$, $p_2 = 0.176$, $p_3 = 0.125$, . . . , $p_9 = 0.046$. Note that $p_2 + p_3$ equals p_1, as it must because the digits 2 and 3 cover the same *relative* data range as the digit 1.

These probabilities p_m, which favor the digit 1 as the leftmost digit, are obtained by integrating x^{α}, with $\alpha = -1$, from m to $m + 1$. For other values of α this integration yields

$$p_m = \frac{m^{\alpha+1} - (m+1)^{\alpha+1}}{1 - b^{\alpha+1}} \qquad \alpha \neq -1$$

where b is again the base of the number system in which the self-similar data are expressed. For example, for $b = 10$ and $\alpha = -2$, $p_1 = 0.\overline{5}$, $p_2 = 0.\overline{185}$, $p_3 = 0.\overline{0925}$, . . . , $p_9 = 0.\overline{012345679}$.

Real-world data are of course never *exactly* scale-invariant, if only because of "end effects." No living village has fewer than 1 inhabitant or more than 100 million, say—except the proverbial "global village." For recent results on leading digits, see the paper by Diaconis [Dia 77].

There are of course plenty of nonscaling "data," such as telephone numbers and the digits on automobile license plates, for which our skewed distributions of digits do not hold.

Skewed distributions of *numbers* (rather than digits) are also known from continued fractions. According to Carl Friedrich Gauss (1777–1855), for most irrational numbers, the asymptotic probability that a given term a_n in the continued fraction expansion equals k is given by the following expression:

$$\text{prob}\,(a_n = k) = \log_2 \frac{(k+1)^2}{k(k+2)} \qquad n \to \infty$$

The exceptional irrational numbers, such as the golden mean $\gamma = (\sqrt{5} - 1)/2$, for which $a_n = 1$ for all n, have measure zero.

Again, much as in the case of self-similar power laws, the number 1 has the greatest probability: prob $(a_n = 1) \approx 0.415$.

In Chapters 5 and 6 we shall further explore the connection between power laws and statistics.

The Diameter Exponents of Trees, Rivers, Arteries, and Lungs

Consider a tree trunk of diameter d bifurcating into two main branches with diameters d_1 and d_2. Is there any consistent relationship between these diameters as one moves up the tree to branches bifurcating into subbranches and subbranches bifurcating into twigs and so on up to the leaf-bearing stems?

Leonardo da Vinci argued that for the sap to be able to flow unimpeded up the tree, the combined cross-sectional areas of the two main branches must equal that of the trunk [RI 57]. In other words, Leonardo believed that $d^2 = d_1^2 + d_2^2$. This claim has stood the test of time and is now enshrined in the "pipe model" of biological tree design [Zim 78]. The pipe model rests on the mental image of the sap being carried up the tree from roots to leaves by many nonbranching vessels ("pipes"), which occupy a fixed proportion of the cross section of each branch.

The same relationship, namely,

$$d^{\Delta} = d_1^{\Delta} + d_2^{\Delta} \tag{1}$$

with $\Delta = 2$, holds for the confluence of two rivers, where d, d_1, and d_2 are the river widths. In fact, the width d of a river is found proportional to the square root of the quantity of water Q transported by the river: $d \sim Q^{0.5}$ [Leo 62]. But the depth t of a river typically varies only as $Q^{0.4}$. The resulting slack is taken up by an increase in the water's velocity v, which is found to be proportional to $Q^{0.1}$. In other words, a river having two tributaries of equal size and thus carrying twice the volume of water per second is typically 1.4 times wider but only 1.3 times deeper than one of its tributaries. But its water velocity is about 1.1 times higher than that of the tributaries. Of course, 1.1 times 1.3 times 1.4 equals very nearly 2, as it should if no water is lost or is added at the confluence.

As Mandelbrot has pointed out, it is impossible to estimate the scale of a map if all river widths are shown to scale. And if the meanders of the river are also self-similar, the course of the river, too, contains no clue to the map's scale.

By contrast, converging or bifurcating roads, which, unlike rivers, possess no depth, should have widths that scale according to equation 1 with an exponent $\Delta = 1$ provided that the traffic flow in cars per lane and minute is the same on all roads. Here the traffic lanes play the same role as the sap pipes in the pipe model for trees.

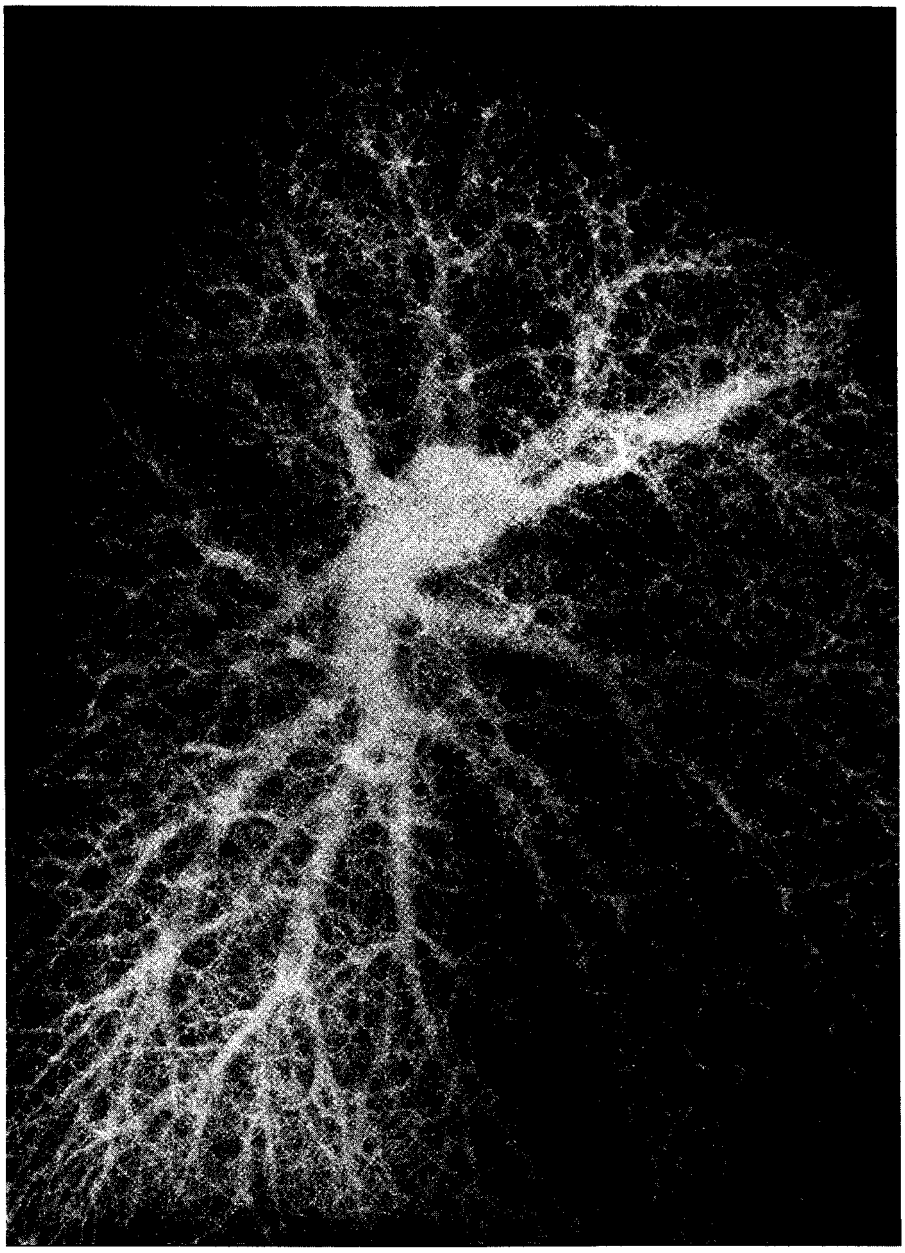


Figure 6 Self-similar bronchial tree [Com 66].

Arteries and veins in mammalian vascular systems, too, have been found to obey the scaling law (equation 1) over a range of 20 bifurcations between heart and capillaries. Estimates of the exponent Δ give values near 2.7 [ST 71]. This is a reasonable value for biological evolution to have attained, given the requirement that arteries and veins should come close to every point of the body that needs nourishment and waste disposal. But the ideal value $\Delta = 3$ for this purpose is, of course, unattainable, because a space-filling vascular system leaves too little tissue for other tasks.

By contrast, the bronchi of the lung do attain a scaling exponent very close to $\Delta = 3$, the value for a fractal that fills three-dimensional space. In fact, the bronchial tree is nearly self-similar over 15 successive bifurcations (see Figure 6 [Com 66]).

The exponent $\Delta = 3$ can be derived from assuming that the geometry of the bronchial tree is determined by the least possible resistance to airflow in the entire bronchial system [Tho 61]. This implies a fixed branching ratio of $d/d_1 = d/d_2 = 2^{1/3}$. With equation 1, the exponent Δ must therefore equal 3 [Wil 67].

However, Mandelbrot has a much more convincing argument, which does not require the branching ratio d/d_1 to be encoded genetically [Man 83]. Rather, Mandelbrot assumes a simple self-similar growth process during the prenatal stage of lung developments: "The growth starts with a bud, which grows into a pipe, which forms two buds, each of which behaves as above."

Iteration of these rules results in a self-similar tree structure for the lung. Thus, the empirically observed self-similarity is obtained not because it is optimum but as a result of the shortest growth-governing program: each step repeats the previous one on a smaller scale. The lung's geometry is therefore fully determined by two parameters: the width/length ratio of the branches and the diameter exponent Δ. In this model, the value $\Delta \approx 3$ simply results from the fact that a large number of bifurcations should be nearly space-filling without crowding each other out.