

Symmetry of Nature and Nature of Symmetry

1

1.1 What Is Symmetry That We Should Be Mindful of It?

Our immediate sense of symmetry comes from looking at objects around us. It may well be that the idea of symmetry is very primitive and comes naturally to the human mind. Perhaps the human mind can grasp it internally all by itself. But we shall leave these questions to the philosopher and to the artist. Instead, let us for a moment turn experimentalist and consider a sphere. Then we will be left in no doubt that we are in the presence of a perfect symmetry. We may view the sphere actively by turning it around every which way we like and find that it looks the same. We may view it passively by keeping the sphere fixed but shifting ourselves around it and find again that it looks just the same. It is this unchanging aspect of sameness against a changing viewpoint that symmetry is all about. But then we have to get sophisticated. We have to abstract the general idea of symmetry and make it free from this static and rather limited visual setting. This we must do and in doing so we will see more, and not less than the artist can, for all his sensitivity and imagination, ever hope to see. There is much more subtlety familiar in the world of physics than meets the eye. However, we will continue to use the same word for it: symmetry.

Symmetry suggests a sense of balance and proportion, of pattern and regularity, of harmony and beauty, and finally of purity and perfection. These synonyms just about sum up all our subjective reactions to the symmetries that abound in Nature, with her myriads of inanimate objects and life forms — the celestial spheres of the sun, the moon and the planets, the hexagonal snowflake with its six-fold symmetry, the five-fold symmetry of the starfish and of many a wild flower, the bilateral symmetry of the butterfly with its outstretched wings and of the man in his poise (Fig. 1.1). One even speaks of the fearful symmetry of the tiger. Examples will fill volumes. And as life imitates Art and Nature, we find something of it reflected in the art forms created by man — be it sculpture, architecture, painting, poetry or music. It is true though that in most of these cases the symmetry is only approximate. As a matter of fact the ancient Greeks used to intentionally

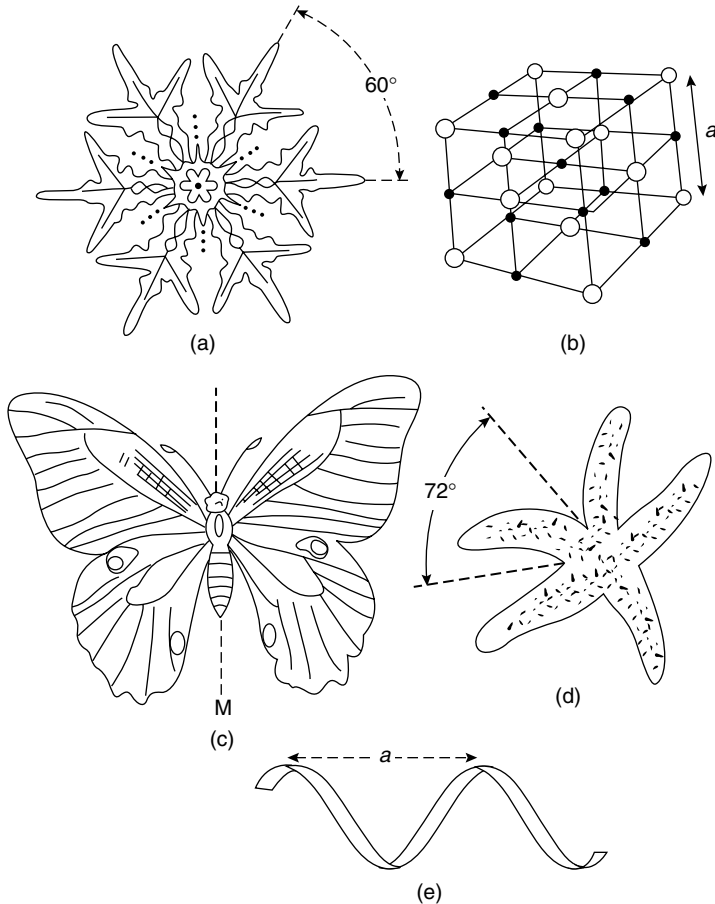


Figure 1.1: (a) Snowflake with six-fold axis; (b) crystal of common salt; (c) butterfly with bilateral symmetry; (d) starfish with five-fold axis; (e) right-handed helix.

and secretly introduce some degree of asymmetry in their otherwise symmetric designs. (After all, there is no perfect beauty that has not in it a certain *strangeness of proportion*). The fact remains, however, that the human mind is absolutely fascinated by symmetry. In physics, the term symmetry takes on an objective meaning which is much deeper and far more precise, almost more austere than our vague feelings of it can command. Let us get acquainted with it.

Now, we can hardly do better than just repeat the definition of symmetry given by the great German mathematician Hermann Weyl — a thing is symmetrical if there is something you can do to it so that after you have finished doing it, it looks the same as it did before. This is an operational definition — it can decide. The ‘thing’ here is the *object* of interest. What you do to it is called the *Symmetry operation* or *transformation*. And ‘looks the same’ is yet another

name for *invariance*. The ‘look’ itself is some discernible property of the object that remains invariant. Thus, there has to be an object with a discernible property that remains invariant under the action of the ‘group’ of symmetry transformations. Now, the point of all this is that the object itself can be just about anything. It depends on our interest and on the level or the depth of our enquiry. At its simplest, the object may be a mere geometrical figure (a hexagon, a helix or a lattice), or the geometrical shape of a material body (a snowflake, a screw or a crystal of common salt) (Fig. 1.1). The symmetry operations involved here are purely geometric in nature — rotation by $360/6 = 60$ degrees or multiples of it about the six-fold axis of rotation, mirror reflection in the plane of the bilateral symmetry, translation in space by a repeat distance, or combinations of these (Fig. 1.1). The object and its transform must be superposable if the symmetry is true. (This is obviously not so for a screw, or a helix. Although the screw is intrinsically identical with its mirror image, the two are not superposable. We will return to this interesting case later). But at its subtlest the object can be a mathematical entity, a (differential) equation expressing a physical law. Now, how do you rotate, reflect or translate an equation anyway? Well, we really do not do so literally. We perform these transformations passively on the independent variables, *i.e.*, the space-time coordinates occurring in the equation accompanied then by suitable transformations on the dependent variables. The invariance then is the invariance of the *form* of the equation under these symmetry transformations. More properly, it is called covariance. Thus, for instance, an expression $x^2 + y^2 + z^2$ is invariant under any rotation of the Cartesian coordinate system (x, y, z) with its origin fixed at $x = 0$, $y = 0$ and $z = 0$. It just becomes $x'^2 + y'^2 + z'^2$, where the primed quantities are the coordinates of the same point, but with respect to the rotated (primed) coordinate system (x', y', z') . Similarly, the wave equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} - \frac{1}{c^2} \cdot \frac{\partial^2 \phi}{\partial t^2} = 0$$

keeps its form under the above symmetry transformation, and additionally under translation in space and in time. Just replace the unprimed quantities by the primed quantities. In particular $\phi(x, y, z, t)$ becomes $\phi'(x', y', z', t')$ and is numerically equal to it. If you take ϕ to be the pressure or the density, then the wave equation begins to describe a sound wave propagating in a medium such as air or water which is homogeneous (translationally invariant in space), isotropic (rotationally invariant) and unchanging in time (translationally invariant in time). In fact this equation has a much higher symmetry and it can describe the widest range of wave phenomena that occur in Nature. These symmetries of the medium (and the medium may well be vacuum as in the case of light) *almost uniquely* fix the form of this equation. Such is the restrictive power of symmetry.

One speaks of the symmetry of a particular law. Thus, we have the spherical symmetry of the Coulomb law of electrostatic attraction between a negatively

charged electron and a positively charged nucleus of an atom. The Coulomb potential energy varies as the inverse of their distance apart, independent of the direction. The force on the electron, being the gradient of potential, of course, varies as the inverse of the square of this distance and is directed radially inward. But for an atom embedded in a molecule or a solid, the potential law governing the motion of the electron has the symmetry of its environment which is necessarily lower than the spherical symmetry of the free atom. Much of the chemistry of molecules and the physics of solids depend on these environmental symmetries.

As we probe matter deeper, we uncover special laws that govern the goings-on at the nuclear and the subnuclear level — the domain of the elementary particles and the fundamental interactions between them. Here we encounter yet another kind of symmetry different from the space-time symmetries described above. These are the so-called *local gauge symmetries* that seem to be at the very heart of the nature of things. It is already present in the interaction of light with charged particles, where it was discovered first. But, of this more later.

A note of caution at this stage is in order. The symmetry of a law as expressed by a symmetric equation does not necessarily lead to symmetric phenomena resulting from it. The states of a system, the processes or the events represent the allowed solutions of the governing (differential) equation. But a particular solution gets selected by the initial conditions that can be imposed at will. These conditions need not have the symmetry of the system. And so it happens that the law of gravitational attraction between the earth and the sun is spherically symmetric, and yet the orbit of the earth round the sun is an ellipse — a foreshortened circle, with the sun at one of its foci. The same is true of a man-made satellite orbiting the earth. Its orbit depends on its height and the velocity at the time of its injection into orbit. A symmetry operation will not leave the particular orbit invariant but carry it into another, albeit allowed orbit. Thus, in general the particular solutions (realizations) or the events or physical conditions themselves are not invariant. What is indeed invariant is the governing equation that fixes only the correlations between the successive events. The idea that the state of a system can have a symmetry lower than that of the governing law takes on a deep physical significance as we will see when we discuss the phenomenon of *spontaneous symmetry breaking*, which is the most symmetrical way of breaking the symmetry. In this we may catch a glimpse of the act of creation whereby Nature seems to have generated the observed diversity of fundamental laws as a result of a descent from the most symmetric, possibly a ‘*grand-unified*’ law of interactions.

The all pervasive nature of symmetry is in itself a sufficiently strong reason for us to be mindful of it. But the most compelling reason of all is that symmetry is a great ordering principle and we can make it work for us. We will now demonstrate this power with the help of some simple and some not-so-simple examples.

To start with, symmetry simplifies things. Suppose you are asked to draw a butterfly with its out-stretched wings. Now, all that you really have to do is to

draw only the left, or the right half of the butterfly, preferably on a tracing sheet. The other half is related to it by mirror reflection. It is more of the same. You can simply fold the sheet along the median line of bilateral symmetry and re-trace over your half-drawing. That is all. The reflection symmetry has halved your work, or very nearly so. In general, an n -fold symmetry divides your work by n . This is really a common trick and we should imagine that the makers of patterns use it all the time. This is, however, a trivial example.

A highly non-trivial example of reduction of a problem by symmetry is provided by the case of a hydrogen atom. Here we have an electron bound to the nucleus (proton) by the attractive Coulomb potential which is spherically symmetric. In order to appreciate reasonably well the promised reduction of the problem, we have to describe the atom properly. It is now well known that in the domain of the very small, and that is where the atoms belong, the proper theoretical framework is that of *Quantum Mechanics*, and not the *classical (Newtonian) mechanics* that describes our sensible world of middle dimensions so well (see Appendices A and B). Thus, we have to abandon the classical view of the hydrogen atom as a miniature solar system with sharply defined orbits for the planetary electron. We have, instead, an all pervasive waviness associated with the motion of the electron. We can picture the state of the electron as a fuzzy cloud around the nucleus, with the proviso that the density of the cloud at a point gives the probability (density) of finding the point-like electron at that point. (This replacement of the classical certainty of sharply determined orbits by the quantum uncertainty of dicey probabilities of being found somewhere is most disturbing. It was so to Einstein himself who was, ironically, one of the founders of this 'plutonic' republic of Quantum Mechanics, but never quite belonged there as a citizen. Quantum Mechanics is today the established *framework* theory for everything in the physical universe. Its predictions differ from those of classical mechanics and the difference gets more and more pronounced as we go deeper into the domain of the small). To get these probabilities one has to solve a certain wave equation, the *Schrödinger equation*, for the wave function ψ , which is complex in general. The probability is then simply $|\psi|^2$, the square of its absolute magnitude. All that is important for our discussion is to note that ψ has both radial as well as angular dependence. The spherical symmetry of the Coulomb potential now helps us factor out the angular dependence and determine it completely without having to solve the Schrödinger equation. The spherical symmetry by itself determines the allowed values of the angular momentum ℓ ($= 0, 1, 2, \dots$) and its component m ($= -\ell, -\ell + 1, \dots, \ell - 1, \ell$) along a chosen direction in units of Planck's constant h divided by 2π . These are the labels, called *quantum numbers* that symmetry provides to specify completely the angular aspect of the state of the system. This is no mean reduction of the problem. In fact one can do better than this. In addition to the spherical symmetry, the Coulomb law has yet another 'dynamical' symmetry following from a certain special value of a parameter in its form, namely that the force involves the square of the reciprocal of the distance,

and not any other power such as the cube or the fourth power, and so on. (This is, of course, a rather hidden dynamical symmetry and is for the preoccupied eyes of the mathematical physicists only). Properly treated, this symmetry solves the remaining radial problem too and provides yet another label, the principal quantum number n ($= 1, 2, \dots$) that fixes the allowed electronic energies.

That there is something special about the inverse square law which singles it out from among all possible central forces, can be seen from the following fact. Consider the motion of the earth around the sun, or better still the motion of a man-made satellite around the earth. The orbits are elliptical as we know. But the real point, which hardly ever gets emphasized, is that the orbit closes upon itself! This will not be the case if you deviate ever so slightly from the inverse square law. For small deviations the orbit will still be close to being an ellipse but the ellipse will slowly precess or turn around the focus. The motion of perihelion (the point of closest approach to the sun) of the orbit of the planet Mercury around the sun may be viewed as due to small deviation from this dynamical symmetry of the inverse-square law caused by Einstein's general relativistic corrections to Newton's law of gravitation.

Now we turn to another aspect of this great ordering principle, namely, that symmetry classifies things. All classification is based on identification of a set of common characteristics. Thus we have the classification of the animal kingdom into vertebrates and invertebrates depending on the presence or the absence of the vertebral column. The *periodic table* of elements prepared by the great Russian chemist Mendeleev is a classic example of classification. The most striking and rigorous example of classification by symmetry is the grouping of crystalline forms of solids. A crystal is a periodic arrangement of atoms in space. It can have spatial symmetries of discrete translation, discrete rotation and reflection and, of course, combinations of these. Symmetry considerations have led to the remarkable result that only a finite number of distinct groupings of these symmetry elements are possible. These are the celebrated 230 space groups of crystallography! Any of the nearly countless varieties of crystals, no matter how complex, must belong to one of these groups. We must hasten to add, however, that the crystals belonging to a given space group are certainly not identical, no more than all the vertebrates in the animal kingdom are identical. Finding the space group of a crystal is the first step towards understanding its molecular structure.

A much more profound example of symmetry-based classification in physics is the classification of identical particles as *fermions* (after the great Italian physicist Enrico Fermi) and as *bosons* (after the great Indian physicist S.N. Bose). Here the symmetry is with respect to permutation, or more simply, reshuffling. Let us understand this. Consider a set of particles located arbitrarily in space. Let the particles be identical in all respects, *i.e.*, the same mass, the same charge, and so on. You may think of a pack of cards, somewhat unusual in the sense that all the cards are alike — only queens of diamond, say. Now it is clear that any

permutation of these identical particles (the same as reshuffling of the identical cards) will leave our system unchanged. After all, a permutation involves just pairwise interchanges, and interchanging identical objects changes nothing. But not quite. The different permuted configurations are undoubtedly identical but they are distinguishable all the same. The reason for this is that nothing prevents you from keeping track of these identical particles as these are being moved around to their new permuted locations. This knowledge is sufficient to distinguish between the different permuted configurations even though the objects being permuted are identical. Thus the identical particles are distinguishable even if only by virtue of being initially located differently. You may wonder if this knowledge is of any consequence and if this distinction between identity and indistinguishability is not mere nitpicking. Classically, you are right. But as we have noted earlier, the correct framework for dealing with microscopic particles is quantum mechanics. And, most importantly, quantum mechanics does not allow sharply defined trajectories. It replaces them with an irreducible fuzziness. Therefore, even in principle, we really cannot keep track of our identical particles in the process of permuting them as we did before. This idea of indistinguishability is brought home rather forcefully if you consider, *e.g.*, a pair of algebraic equations $x^2 + y^2 = 13$ and $x + y = 5$. These two equations are left invariant if we interchange x and y and hence permutation symmetric. Now, you can readily solve these two equations. You get either $x = 3$ and $y = 2$, or $x = 2$ and $y = 3$. Thus all you can say is that one of them equals 2 and the other equals 3, but which one is which you cannot say even in principle. So is the case with our identical particles. We can only say how many are there at a given point of space (*i.e.*, the occupancy) but it is meaningless to ask which ones. This *indistinguishability* when treated properly leads to the great divide of identical particles into two classes — the fermions (*e.g.*, electrons, protons, neutrons, neutrinos, etc.) and bosons (photons, mesons, etc.). Identical fermions, electrons, say, exclude each other in that not more than one can occupy the same state. This is the Fermi statistics — kind of negative feedback at work. In contrast to this, any number of identical bosons, photons say, are allowed to occupy the same state. This is the Bose Statistics. In fact, bosons tend to clump together, a kind of positive feedback. What determines whether a given set of identical particles will be fermions or bosons requires deeper analysis of relativistic invariance. It is beyond our scope to go into that. But the result is simple. It turns out that a particle can have an intrinsic angular momentum called spin. You may roughly picture it as a spinning top much the same way as the earth spins about its own axis in addition to orbiting around the sun. The spin angular momentum is immutable (you cannot stop it spinning). It is quantized in multiples of $\hbar/2\pi$, denote by slashed \hbar . Now the rule is that particles with integral spin ($0, \hbar, 2\hbar, \dots$) are bosons and those with half odd-integral spin ($\hbar/2, 3\hbar/2, 5\hbar/2, \dots$) are fermions. This connection between spin and statistics has been one of the marvels of the symmetry principles in physics. The fact that two electrons (fermions with spin half) cannot simultaneously occupy the

same point of space with their spins pointing in the same direction (*i.e.*, cannot be in the same state) is responsible for the stability of all matter, and for the fortunate circumstance that your hands do not go through the table which they might be resting on. For, in doing so the electrons in your hand must go through the electrons in the table which is clearly forbidden. The clumping tendency of photons (bosons with spin unity), on the other hand, makes it possible for any number of them to condense into a given state — *Bose condensation* (See Chapter 4). This is what makes laser beams so coherent (See Chapter 2 on Lasers). Similarly, superfluidity of ^4He (the isotope of helium with total spin zero), namely that it can flow through the finest capillaries without any viscosity, is due to the same Bose condensation of these atoms in the lowest energy state at low temperatures close to absolute zero. ^3He , the fermionic isotope, on the other hand, behaves differently even though chemically the two isotopes are identical.

Elementary particle physics abounds in examples of order brought about by classification of the zoo of particles based on certain postulated, rather abstract and well concealed symmetries without knowledge of the details of the underlying laws (see Chapter 9).

Symmetry is also highly restrictive. It limits the possibilities allowed without detailed knowledge of the system. The classic example is the forbidden five-fold axis of rotational symmetry in a crystal. The only allowed ones are the two-fold, three-fold, four-fold and the six-fold axes. The compatibility of the rotational and the translational symmetries rules out the five-fold axis as also the higher order axes of rotation. The five-fold axis is also conspicuous by its absence on the floor designs, or the tiling of a plane called tessellation. You see the square, the equilateral triangular and the regular hexagonal motifs, but never a regular repeating pentagonal pattern with the five-fold symmetry. However, individual molecules and other objects can and in fact do have the five-fold axis. Just think of a pentagram or the starfish. It is an interesting thought that living organisms like the starfish may adopt the five-fold symmetry as a natural defence against the deadly ‘capture’ by the rigid crystalline formation.

The really restrictive power of symmetry in physics derives from the overriding conservation laws that it imposes — the conservation of energy, momentum, angular momentum and charge. We will return to this when we discuss this connection between invariance and conservation laws. Processes violating these are simply forbidden.

Symmetry is at its most powerful when it predicts. Let us illustrate this with an example from solid geometry. Suppose you are interested in regular convex polyhedra (poly = many, hedra = faces). A regular polyhedron is a volume bounded by plane faces which are identical regular polygons. A simple cube (the common dice, for earth) is one such polyhedron. It has six faces that are square (you can call it a regular hexahedron). There are other regular polyhedra, namely the tetrahedron (for fire) with four equilateral triangular faces, the octahedron (for air) with eight

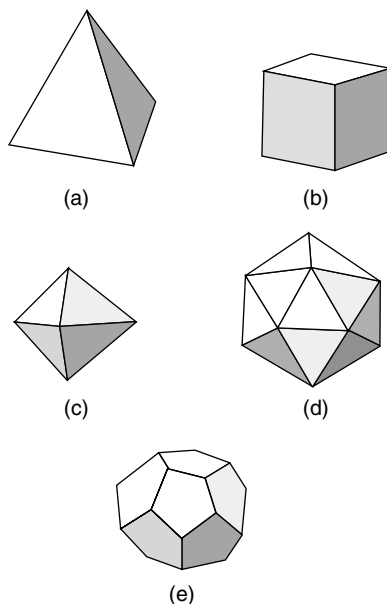


Figure 1.2: Five platonic solids: (a) tetrahedron; (b) cube; (c) octahedron; (d) dodecahedron; (e) icosahedron;

equilateral triangular faces, the dodecahedron (for quintessence) with twelve regular pentagonal faces and finally the mysterious icosahedron (for water) with twenty equilateral triangular faces (Fig. 1.2). These are the so-called Platonic Solids contemplated by the Greek Pythagoreans. The question is if there are more. Well, the answer is a definite no. Symmetry forbids any other occurrence. This is a restrictive aspect of symmetry. The predictive aspect is just the flip side of the coin. If there are intelligent inhabitants in some distant galaxy interested in these exotic dice-forms, we can predict that they will find just these five and no more.

But the real predictive power of symmetry is seen in particle physics. The basic idea is just this. Having identified or guessed the symmetry of the governing law, the processes, or the states or the particles related by the symmetry operations are all treated *at par*, *i.e.*, equally allowed and intrinsically the same. Thus, if you find one, the others, the missing ones are predicted. This is, for example, how the short-lived particle called Ω^- was predicted by Gell-Mann in 1962, and later confirmed in 1964 as the missing member of the family of ten objects (resonances) predicted on the basis of a postulated symmetry $SU(3)$. This was a historic triumph of symmetry in physics.

There are two other aspects of symmetry with far-reaching consequences. These are its unifying and creative powers. We will return to this point later.

There is an ingenious way crystallographers use the power of symmetry constructively. Suppose you need to know the structure of a complex molecule. It may

be a protein with some hundred thousand atoms, or a fragment of DNA. These are very important but complex molecules. Proteins are the building blocks of cells and enzymes, while DNA (Deoxyribonucleic acid) carries the genetic information for making these proteins. Now, you cannot use ordinary light to probe these. Its wavelength of several thousand Angstroms ($1 \text{ \AA} = 10^{-8} \text{ cm}$) is much too large to reveal the finer molecular details on the scale of a few Angstroms. We must use X-rays with wavelengths of about an Angstrom or so. If you shine X-rays on a sample containing these molecules, placed and oriented randomly, the scattered waves of X-rays will interfere randomly to produce a mere smudge on a photographic plate. If, however, you could somehow arrange the molecules periodically in space, that is to say if you could crystallize the substance, the waves scattered from the molecules would interfere constructively in certain well-defined directions and thus produce a systematic pattern of bright sharp spots (the diffraction pattern) on the plate. This is like making Fourier series analysis of a periodic function. One can invert this to get at not only the periodic structure of the crystal lattice but also the structure of the molecules making it up! (One only hopes that the imposed crystalline arrangement has not done too much violence to the molecule whose structure we were interested in). This is why crystallographers-turned-molecular biologists round the world are preoccupied with crystallizing these substances. At this point, we should note that the crystalline order as a necessary condition for getting sharp X-ray spots has been called into question recently with the discovery of the so-called *quasicrystals* by D. Shechtman, I. Blech, D. Gratias and J. W. Cahn (1984). The first quasicrystal was an alloy, $\text{Al}_{14}\text{Mn}_{86}$, *i.e.*, 14 atomic per cent aluminum and 86 atomic per cent manganese. Since then many more have been found. These materials show sharp diffraction spots like any other good crystal but the arrangement of spots has a five-fold symmetry which is, of course, forbidden in the real space crystal lattice. The conclusion is that the conventional crystalline order is not necessary for sharp spots in X-ray diffraction. A two-dimensional quasicrystal is exemplified by the so-called Penrose aperiodic tiling of a plane with motifs of two rhombuses fitted as pieces of a jigsaw puzzle (Fig. 1.3). The smaller and the larger rhombuses have angles 72 degrees and 108 degrees, and 36 degrees and 144 degrees, respectively, and their areas and numbers are in the golden ratio $= (1 + \sqrt{5})/2$. This is the intellectual property of the Oxford mathematician Roger Penrose, who constructed it for play. The tiling has no translational symmetry of the conventional crystals and yet would give a sharp diffraction pattern. It is now known that quasicrystals may be viewed as a projection of ordinary-crystalline order from hypothetical higher dimensional spaces.

Our discussion of symmetry so far has been rather discursive. But, as we have remarked repeatedly, symmetry is a very precise concept. The proper language for a systematic study of symmetry is that of *group theory*, which is a highly developed branch of mathematics. The basic idea is simplicity itself. Identify *all* the symmetry operations that leave a given object invariant. Call them A, B, C, \dots . This is then

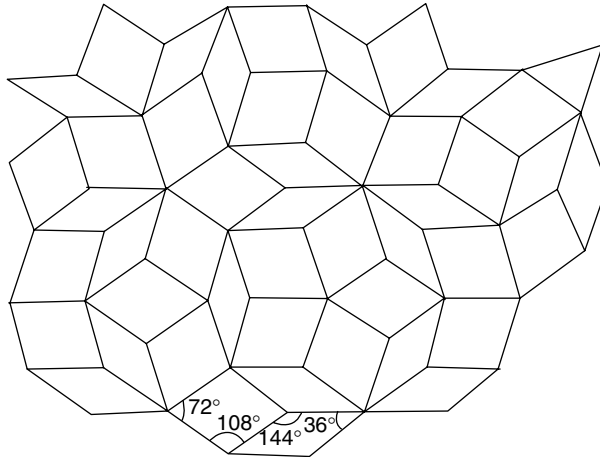


Figure 1.3: Penrose aperiodic tiling of a plane.

an exhaustive list. It is clear from the very definition of symmetry that the successive applications of any two operations, first A and then B , say, will also leave the object invariant. Therefore, the combined operation must also be one of the symmetry elements we have listed exhaustively above. Let it be C . Then we can write $C = BA$. Mark the order of A and B in BA . It means A operates first, followed by B and the result is the same as C . This is a kind of multiplication, composition or successive operation, that gives the interlocking of the various symmetry operations. We say that the symmetry operations are closed under this multiplication. Next, we note that doing nothing at all to the object is also a symmetry operation because it trivially leaves it invariant. In fact it leaves it alone! We denote this trivial symmetry operation of ‘doing nothing’ by E (This is a fairly standard notation). Finally, we note that reversing a symmetry operation is also a symmetry operation — it restores *status quo ante*. Remember that the reverse of a clockwise rotation by an angle θ is an anticlockwise rotation by the same angle θ about the same axis. In obvious notation, we denote the reverse (more properly called inverse) of A by A^{-1} . It is now clear that $E = A^{-1}A$. (That is applying a symmetry operation followed by its inverse amounts to doing nothing). We are all set now. A set of elements having a law of multiplication (successive operations) under which the set is closed, with an identity (doing nothing) and where each element has a unique inverse is called a *group*. The symmetry operations then form a group. We are now compelled by the sheer logic of it. The inner structure of the symmetry group is given completely by enumerating the results of all pairwise multiplications, *e.g.*, $C = BA$. Constructing a multiplication table is like finger-printing the symmetry. Identical multiplication tables imply identical symmetry structures no matter how physically different the objects themselves may be. It is all very nice, but what can we do with all this, you may ask. Well, you can do a lot. An example will help

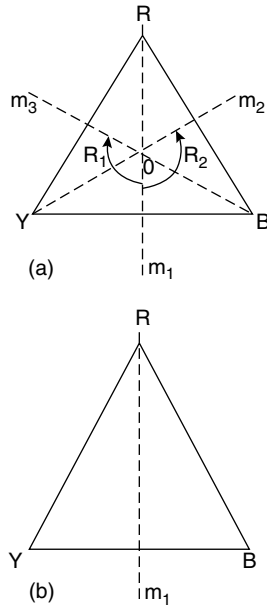


Figure 1.4: Symmetry elements of (a) equilateral triangle; (b) isosceles triangle.

illustrate the point. Suppose the symmetry of the physical law in question turns out to be the symmetry of an equilateral triangle living on a plane (Fig. 1.4).

The symmetry operations are then E (identity), R_1 and R_2 (clockwise and anti-clockwise rotations by 120 degrees, respectively, about the three-fold symmetry axis) and the three reflections m_1 , m_2 and m_3 in the three mirror-lines (medians). One can readily construct the multiplication table. Thus verify, for example, $R_1R_2 = E$, $m_1m_2 = R_2$, $m_2m_1 = R_1$, and so on. Remember that we have completely specified the symmetry structure of the physical law that governs our physical system. The latter may be a molecule with an atom literally at the center of an equilateral triangular environment formed by three other identical atoms. (Situations analogous to but more complicated than this are very common in chemistry, *e.g.*, an atom, or rather a doubly charged ion of copper at the center of a regular octahedron formed by the negatively charged atoms of oxygen in copper sulphate). Let us assume now that this system can exist in one of the three states, or *linear combinations of these*, which are permuted among themselves under the symmetry operations. Of course, we are assuming here that it is meaningful to speak of such linear superpositions. This is indeed the basic structure underlying quantum mechanics (see Appendix B). Thus we can identify the three vertices of our equilateral triangle with these three states. We provocatively label them R (for red), Y (for yellow) and B (for blue). It is easily verified that our symmetry operations indeed permute them in all possible ways. (There are six ways in which three objects can be permuted and the number of elements in our symmetry group is also six). While the symmetry operations do

permute the three states among themselves, they do not mix them indiscriminately. Indeed, they split the possible linear combinations into two sets (called multiplets more properly) such that only members of the same multiplet mix among themselves. One multiplet, call it $W = R + Y + B$ has only one member (a singlet). The other has two members, $Z_1 = R - 2Y + B$ and $Z_2 = R - B$. Now Z_1 and Z_2 mix freely under our symmetry operations and, therefore, they are intrinsically the same — they differ according to our viewpoint only. Thus, for instance, they should have the same energy (or mass). We say that the multiplet is two-fold degenerate. Their energy, however, must be in general different from that of W with which they do not mix under symmetry. In this simple case we could write down this multiplet structure by mere inspection. In general one has to use the multiplication table in a systematic way. It is called the representation theory of groups. In our example W and $\{Z_1, Z_2\}$ provide, respectively, one- and two-dimensional representations. We can go further and lower the symmetry to that of an isosceles triangle by pulling one of the vertices out (Fig. 1.4b). Our symmetry group now will consist of only two elements $\{E, m_1\}$. It is a sub-group of the earlier larger group. The result is that the doublet is further split into two singlets. We now have three non-degenerate (unequal) levels.

This splitting or reduction of degenerate multiplets with the progressive lowering, or descent, of symmetry is well known and well studied in chemistry and solid state physics, where the symmetry is mostly geometrical and known from structure. The situation is quite different in elementary particle physics where the symmetry is rather abstract and not directly accessible. Here symmetry takes on a creative role. This is made possible by the fact that a given group uniquely specifies the possible multiplet structures it can support. Thus one can postulate a symmetry and then work out the multiplet structures it implies and compare with the observed families of closely related particles. This is the idea underlying the unending quest for symmetries, *e.g.*, $SU(2)$, $SU(3)$ and so on. One is limited only by his ingenuity and insight. Thus $SU(3)$ (special unitary group of rotations in a three-dimensional complex space) has a multiplet with eight members (eight-dimensional representation) and one with ten members (ten-dimensional representation) that fitted so well the observed families of eight baryons and ten hyperons — behold the ‘unreasonable’ effectiveness of symmetry in physics!

Finally, a remark on the group multiplication. Note, that in our example we had $m_1 m_2 = R_2$ and $m_2 m_1 = R_1$. Thus unlike ordinary multiplication of numbers, the order in the applications of symmetry operations is important. We say that m_1 and m_2 do not commute. Such a symmetry group is said to be non-abelian. The important group of rotations in three-dimensional space $SO(3)$ is non-abelian. The corresponding group of rotations in a plane is Abelian. An amusing demonstration of this is the following. You fly out of the North Pole down the zero-degree longitude through Greenwich to the equator. You will be over the Atlantic, south of Ghana. This is a rotation by 90 degrees about the east-west axis. Now you turn and follow

the equator to longitude 90 degrees east. You should be over the Indian ocean east of Sumatra. This amounts to a rotation by 90 degrees about the north-south axis. Now, you perform these operations in the reverse order. Start out at the North Pole and turn by 90 degrees eastwards. But since you are right on the axis of rotation, you just stay put. Next fly down the zero degree longitude till you reach equator, and thus you end up over the Atlantic, south of Ghana, thousands of kilometers away from your earlier destination in the Indian ocean. It turns out that most of the symmetries in Physics are non-Abelian, and that makes it richer. Abelian symmetry gives only non-degenerate one-dimensional or single-state multiplets.

1.2 Space-Time Symmetries: Invariance and the Great Conservation Laws

Objects are located in space. They endure in time. This is true of all events and processes, of beings and becomings, that ultimately involve the elementary particles and their interactions that make up the world of physics. Admittedly, this is a highly reductionist viewpoint but you can hardly fault it. It seems reasonable, therefore, that the study of symmetries of objects and phenomena must be preceded by a proper study of the symmetries that this background space-time continuum may have. For obvious reasons we will call these the framework symmetries. These symmetries must be established as facts of experience, no matter how compelling *a priori* they may appear to be. To the best of our knowledge, then, the following symmetries are true.

Space is homogeneous. That is to say that the absolute position of an object is irrelevant. What it operationally means is that if we perform an experiment at a location and then repeat the same experiment somewhere else, in outer space, say, the results will be identical — translationally invariant in space. By the ‘same experiment’ we mean that all conditions relevant to the experiment must be reproduced exactly. Thus, if the change in earth’s gravity in going out there is relevant, then the earth must be transported along with the apparatus. One may argue that this claim is then vacuous inasmuch as any discrepancy between the results of the two experiments can always be blamed on something that may have escaped our attention, to wit our altered position with respect to the distant stars! Now, this is perverse because it is possible to isolate our experiment far enough to any desired degree of accuracy by including larger and larger regions of space as part of our experimental set-up and, because one can assume reasonably that all effects are essentially local in nature. Ghosts are not admitted! In any case, there is nothing to suggest violation of this translational symmetry.

Next comes isotropy of space, or the irrelevance of absolute direction. Operationally, it means that if we perform a certain experiment and then rotate our entire setup to a new orientation and repeat the same experiment, the results will

be identical. We can re-word all our earlier provisions and arguments in support of this. So far there is no empirical evidence in support of a preferred direction in space. Thus isotropy of space is a good symmetry.

There is an interesting connection between these two symmetries. Isotropy (relative to every point of space) implies homogeneity but not *vice versa*. This is readily proved. Let P_1 and P_2 be two points in space. Draw a sphere passing through P_1 and P_2 , with center, say, at O . You can draw any number of such spheres. Now, viewed from O , P_1 and P_2 are related by isotropy and, therefore, are equivalent. You can repeat this process till you cover the entire space and thus establish homogeneity of space.

Next comes homogeneity of time, or time-translation symmetry. There is no absolute origin of time. If you perform an experiment now and repeat the same experiment at a later date, the results will be identical. Indeed, without these symmetries the universe will hardly be comprehensible. We should perhaps mention here that there is evidence that the universe is finite, though unbounded, and that it had a beginning some 15 billion years ago — the *Big Bang*. We hope that we are at a sufficient remove from this boundary (though there is actually none) and initial conditions to ignore these symmetry breaking effects here and now.

Finally, to these irrelevancies, namely those of absolute position, absolute direction and absolute time, we add the irrelevance of absolute rest, or of absolute uniform motion. Consider two unaccelerated platforms in uniform relative motion, that is to say that one platform is moving with a constant velocity as seen by an observer who is stationary on the other platform. Now, if we perform an experiment on one of these platforms and then repeat the same experiment on the other, the results should be identical. Thus no *local* experiment, *i.e.*, without reference to the other platform, will detect any effect that can distinguish between these two unaccelerated platforms — there is no absolute uniform motion. This equivalence of unaccelerated platforms is the great symmetry expressed by the *principle of relativity* and was a wonderful achievement of Galileo. Acceleration is, on the other hand, absolute and can be detected locally by an accelerometer — a mass attached to one end of a spring, the other end of which is fixed to the platform. (In all these discussions, we will ignore the presence of gravitation). This is quite consistent with our every day experience. We are hardly aware of the velocity with which the lift, by which we may be traveling, is moving except at the times of start and stop, *i.e.*, when there is acceleration or deceleration.

A platform is, more formally, a set of points at rest relative to one another. It is convenient to introduce a rectangular coordinate system (x, y, z) at rest with respect to these points. One may also assume a clock, an atomic clock, say, attached to every point of this set. The identical clocks may be synchronized by exchanging light signals. Thus, if A and B are two points and if t_1 is the time at which a light signal is sent out from point A , and if t_2 is the time at which the signal is received at and reflected by the point at B , and finally if t_3 is the time at which the signal is received

back at the point A , then the clocks at A and B are synchronized if $t_2 - t_1 = t_3 - t_2$. Note that this is purely by symmetry and does *not* require knowledge of the speed of light. Thus an elementary event is completely located by giving spatial coordinates (x, y, z) and the time of its occurrence t , read out by the clock at the point (x, y, z) . Such an unaccelerated platform equipped with the markers $(x, y, z; t)$ is called a Galilean frame of reference S , say. Another Galilean frame S' , say, will have a primed space-time coordinate system $(x', y', z'; t')$. Now the relativistic invariance asserts that the laws expressed in terms of the primed and the unprimed space-time coordinates should have the same form. The question now is how the primed and the unprimed space-time coordinates of the same event are related. The relativity of motion encountered in everyday life, also called Galilean relativity, would suggest the following answer. Time intervals are absolute. So are the space intervals. This means that, if $(x_1, y_1, z_1; t_1)$ and $(x_2, y_2, z_2; t_2)$ are the space-time coordinates of two events observed in a Galilean frame S , and $(x'_1, y'_1, z'_1; t'_1)$ and $(x'_2, y'_2, z'_2; t'_2)$ are those for the same two events but in another Galilean frame S' , then the time interval $t_{12} = (t_1 - t_2) = t'_{12} = (t'_1 - t'_2)$ and the space interval squared $r_{12}^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = (r'_{12})^2 = (x'_1 - x'_2)^2 + (y'_1 - y'_2)^2 + (z'_1 - z'_2)^2$.

This leads to the rules of vector addition of velocities and displacements well known from our high-school days. The symmetry operations here are the familiar translation and rotation (re-orientation) in space, 'boosting' to a relatively uniformly moving frame, and time translation. Galilean relativity is, however, based on our common experience with slow objects moving at small velocities, *e.g.*, the speed limit of about 100 kilometers per hour on national highways. Compare this with the speed of light, 1080 million kilometers per hour in vacuum. Can we extrapolate our tardy experience to such high velocities? Let us see. In Galilean relativity the speed of light in vacuum would depend on the relative velocity of the source of light and the observer. One can then, in principle, chase light and even outrun it. Or one can run just fast enough to keep pace, bringing light to a relative standstill. This is true, for instance, in the case of sound. But sound propagates only in a medium, *e.g.*, air. Light, however, can propagate in vacuum. Is vacuum too filled with an all pervasive medium — the 'aether' as was indeed thought for quite some time? This hypothetical medium, the aether, could then provide the preferred frame of reference at absolute rest, making thus the different Galilean frames moving relative to it in principle non-equivalent. Even the most careful laboratory measurements and astronomical observations have, however, failed to detect this aethereal medium. Einstein did not like this loss of symmetry anyway. The point is that light is an electromagnetic wave whose propagation in vacuum relative to a Galilean frame is described by a wave equation of the type we wrote down in the last section. Notice that the speed of light ' c ' occurs explicitly in this equation. The invariance (or rather covariance) of this equation with change from one Galilean frame to another then demands the invariance of the speed of light. Thus, we have the fundamental postulate of the absolute constancy of the speed of light for all

Galilean frames of reference — the basis of *Einstein's special theory of relativity*. The changes from one frame of reference to the other are the symmetry operations that leave the speed of light unchanged. It is clear that for this to be so, the notion of *absolute* time interval t_{12} as separate from that of the absolute space interval r_{12} between the two events labeled 1 and 2 inherent in the Galilean relativity must be abandoned. Einstein's relativity replaces these two with a single *absolute* invariant interval s_{12} between the two events, given by $s_{12}^2 = r_{12}^2 - c^2 t_{12}^2$. Three-dimensional Euclidean space and time $(x, y, z; t)$ are replaced by a four-dimensional space-time (the *Minkowski world*) that treats time t as just another co-ordinate to label the events, *at par* with space coordinates (x, y, z) (Fig. 1.5).

An event is now located at a world-point (x, y, z, t) — the semicolon that set time apart from space has now been replaced by a common comma. The transformation from (x, y, z, t) to (x', y', z', t') is now the symmetry operation of displacement and rotation (Lorentz transformation) in this four dimensional world, keeping in mind,

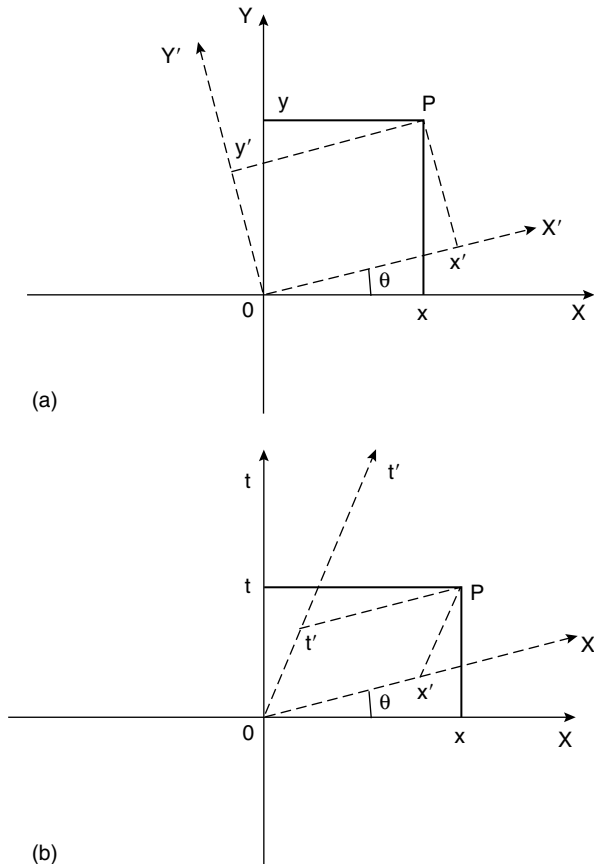


Figure 1.5: Rotation in (a) ordinary space; (b) Minkowski space-time.

however, the technical point about the minus sign that occurs in $s_{12}^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 - c^2(t_1 - t_2)^2$. Einstein's special relativistic space-time symmetry now demands that the laws of physics be invariant under the Lorentz transformation from one Galilean frame (unaccelerated or *inertial frame*) to another. This replaces the old Galilean invariance with its absolute time intervals. Indeed, the Maxwell wave equation for the propagation of an electromagnetic disturbance in vacuum has the Lorentz invariance, but *not* the Galilean invariance. One may mathematically absorb the minus sign by defining an imaginary time $\tau = it$, with $i = \sqrt{-1}$ and treat time formally completely *at par* with space. But it is actually better to leave it as such, as a gentle reminder that time is, after all, qualitatively different from space. The negative sign implies that the interval s_{12} can vanish without the two events coinciding in space-time, *i.e.*, $s_{12} = 0$ but $r_{12} \neq 0$, $t_{12} \neq 0$. We can even have s_{12} negative. (We say that the Minkowski world has an indefinite metric).

The geometry of this four-dimensional world has important and interesting physical consequences. The well advertised popular effects — the variation of mass with velocity, the equivalence of mass and energy, the Lorentz contraction and time dilation, all belong here. The speed of light in vacuum is the limiting speed that cannot be exceeded. Our main concern here is, however, only the symmetry aspect of relativity — the great framework symmetry. Let us note one highly counter-intuitive aspect of it because it has a deep significance for our discussion of invariance and conservation law later. Since it is only the interval s_{12} that remains invariant from the unprimed frame to the primed one, it is clear that we can have t_{12} zero but t'_{12} non-zero. That is to say that in the unprimed frame the two events are simultaneous, but in the primed frame they are not. This is the relativity of simultaneity that totally demolishes the notion of absolute time interval. (Incidentally, one may have an uneasy feeling, when simultaneous events in one Galilean frame appear non-simultaneous in the other, about what happens to their chronological order of occurrence — which is older of the two. Well, relativity does allow a certain amount of play in this game of courtesy, but there is an absolute past and an absolute future even here consistent with notions of cause and effect).

We now turn to the deep connection between these relativistic space-time symmetries (invariances) and the conservation laws. Inasmuch as these symmetries are the framework symmetries to which all the basic laws of physics are subject, we will call the corresponding conservation laws the Great Conservation laws. Consider a physical process written schematically as $x + y \rightarrow z + w$. A quantity is said to be conserved if its total value for the reactants $x + y$ is the same as its total value for the products $z + w$ of the process as observed in a given Galilean frame. Thus we speak of conservation of energy, linear momentum and of angular momentum. It turns out that the conservation of energy follows from the invariance with respect to translation in time. The conservation of linear momentum follows from the invariance with respect to translation in space (homogeneity of space). The conservation of angular momentum follows from the invariance with respect to rotation

in space (isotropy of space). A proper discussion of conservation of these quantities (and even their definition in general) as a consequence of the invariances requires the introduction of ‘action’ and ‘action principle.’ This is beyond our scope. The important point to note is that this connection between invariance and the conservation law is not restricted to any specific dynamical laws such as Newton’s laws of motion. The connection is purely kinematic. For the specific case of mechanical systems where, for instance, momentum is mass times velocity, one may derive conservation of linear momentum by applying Newton’s three laws of motion. And so on for energy and angular momentum. But the connection is really much more general. After all, there are non-mechanical objects, light for instance, that also carry energy, momentum and angular momentum. We should note in passing that just as isotropy of space implies (but is not implied by) homogeneity, conservation of angular momentum implies conservation of linear momentum, but not *vice versa*.

Much of the restrictive and predictive power of these symmetries comes from the associated conservation laws. The striking example is radioactivity (β -decay) in which a neutron was thought to decay into an electron, a proton and something else. The electric charge is conserved as required by another invariance called *global gauge invariance* to be discussed later. However, a careful reckoning of energy and momentum of the system before and after the reaction led to an imbalance. Thus a new particle was suspected as a decay product that carries the missing energy and momentum. It was predicted to be neutral and to have zero rest mass. Also, recalling that the neutron, the electron and the proton all carry spin half (angular momentum $\hbar/2$), conservation of angular momentum required the then unknown particle to carry spin half. All this was confirmed happily later. This is the now well known but elusive elementary particle, the electronic anti-neutrino denoted by $\bar{\nu}_e$, and the corrected process reads $n \rightarrow e^- + p + \bar{\nu}_e$. These particles are now routinely and abundantly produced in laboratories, in nuclear reactors as well as accelerators.

The full power of these great framework symmetries is realized only when these are combined with the great framework theory — Quantum Mechanics (Appendix B). But this will take us very far afield. We will be content with just mentioning it. In addition to these continuous symmetries, there are discrete space-time symmetries too. One of them is the symmetry under space reflection, also called the mirror symmetry or parity. This produces enigmatic effects in ordinary laboratory physics and chemistry as also in the extraordinary processes involving elementary particles. We will take this up next.

1.3 Reflection Symmetry

We have spoken of objects having bilateral symmetry, also called the left-right symmetry. A butterfly with outstretched wings or a maple leaf for example. When an object is reflected in a mirror, the left and the right sides of it get interchanged.

Thus, an object having bilateral symmetry is by definition superposable on its mirror image. The mirror is just an optical device that enables us to visualize the result of reflection of objects in space through a plane. For these reasons the terms bilateral symmetry, left-right symmetry, mirror symmetry and the symmetry under space reflection are all used interchangeably. In physics, handedness is often referred to as *chirality*.

There is something that sets this symmetry apart from the rest that we have discussed so far. As noted above it is a discrete symmetry unlike the continuous symmetry of rotation or translation, say. Changes caused by continuous symmetry operations can be made arbitrarily small. Not so with discrete ones. You reflect or you don't: The excluded middle — there is nothing in between. Also, unlike these, it is a *non-performable* symmetry operation. Space reflection involves turning the object inside out laterally, an operation we can hardly perform continuously. But we can and we do visualize it by the optical trick of reflecting it in a mirror. Having visualized it so, nothing prevents us from making a physical copy of the image, using silly putty, say, which can then be tested for superposability on our object. This is the operational meaning of reflection symmetry as applied to shapes of material objects or geometrical figures. The non-performable nature of reflection symmetry conceals an important aspect of it that we will try to uncover now. Consider an arbitrarily shaped object and its reflection in a mirror. A rather handy example would be, well, your right hand itself. Its mirror image is constructed by translating every point on the hand to a point on the other side of the mirror, along the line perpendicular to the plane of the mirror and equidistant from it (Fig. 1.6a).

Now, it is clear that this image (ideally your left hand) is not superposable on your right hand. Such an asymmetric object is called 'handed,' and with very good reason. Have you ever tried your left-hand glove on your right hand? And yet nothing else is more like my right hand than its mirror image, that is my left hand. The reason that the two cannot be superposed is an inconvenient circumstance of life, namely, that the hand is a three-dimensional object and so is our physical world (space) in which cooped up we live. In a world of higher dimensions, the right hand could have been turned around by a temporary excursion into the extra dimensions, and thus superposed on the left hand. This can be demonstrated quite easily with an example taken from the world of lower dimensions — of a two-dimensional object, a flatlander living on a plane which is embedded in our familiar three-dimensional space. Thus, the symbol "Om" in Fig. 1.6b can be superposed on its mirror image by simply folding the paper along the mirror line M. The act of folding involves a temporary lift or escape into the third dimension coming out of the plane of the paper. Science fiction is full of such excursions into the extra dimension — the "tesseract" in "A Wrinkle in Time" by Madeleine L'Engle is a fascinating case in point. These extra dimensions, somewhat curled up and rather inaccessible, are also the subject of serious thought by the physicists of our times. But we are digressing. All this suggests that an object and its mirror image are intrinsically the same. To

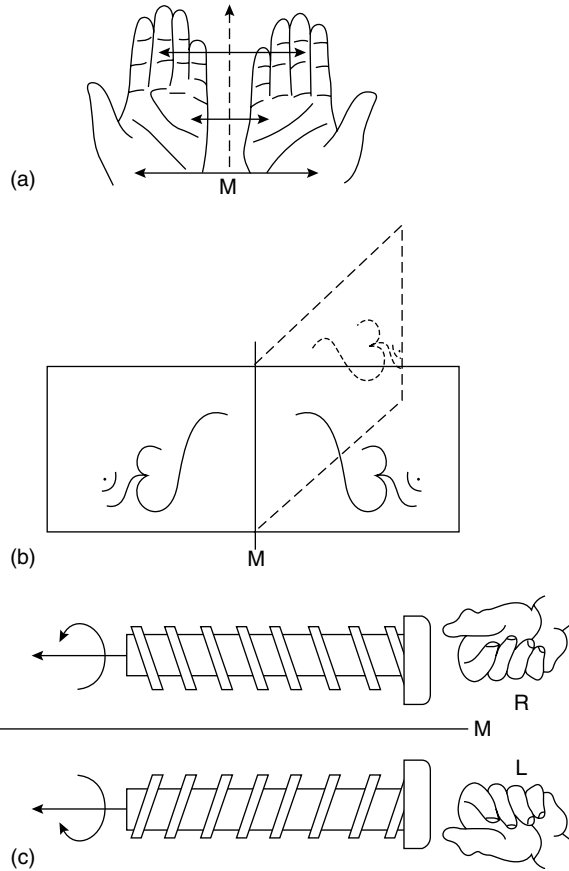


Figure 1.6: (a) Mirror reflection; (b) continuous reflection via third dimension; (c) right-handed screw reflected as left-handed screw.

emphasize this we call them a pair of *enantiomorphs* or antipodes — an impressive name for ordinary mirror images.

What is in a hand that makes it so ‘handed’? To understand handedness more thoroughly we can think of a screw — the common screw that we use to fasten things together, and without which much of our civilized world would simply come apart. The common screw is nothing but a helical ridge, called thread, cut into the surface of a cylinder, or a cone if it is a tapered screw. When the screw is turned as indicated by the circular arrow (Fig. 1.6c), it advances (or recedes) along its axis as indicated by the linear arrow. This makes the screw a machine that converts a rotary motion about its axis into a translational motion along that axis. You may think of the Archimedes Screw that was used by the Egyptians to raise the waters of the Nile, and is still in use for similar purposes. It should be clear that there are two and only two classes of screws possible. These correspond to the two possible relations between the circular and the linear arrows. Let us give them

names. Suppose that you clasp the screw in your right hand with your fingers pointing in the direction of the circular arrow while your thumb stays parallel to the axis of the screw. Now, if your thumb points in the direction of the linear arrow, then the screw is said to be right-handed. If, on the other hand, it points in the opposite direction, then the screw is said to be left-handed (Fig. 1.6c). It is easy to see that the mirror image of a right-handed screw is a left-handed screw. The two form a pair of enantiomorphs. That there are two classes of screws, *i.e.*, the two-ness of it, is an absolute fact. But defining them as the left- and the right-handed screw is, of course, a matter of convention — a very useful convention though, which is followed uniformly all over the civilized world. This has been made possible by the intimate contacts we have had over centuries of togetherness, and not a little by our admirable practice of shaking hands. But a convention all the same. The non-triviality of this is brought home by the following thought provoking circumstance. Suppose we establish radio-contact with some advanced civilization in a galaxy far away. Such an eventuality can not be ruled out, thanks to the project SETI (Search for Extra-Terrestrial Intelligence) mounted by some serious-minded people. Now we should have no difficulty convincing our distant correspondents that these are the two classes of screws possible. But try hard as we may, we will not be able to explain to them what we mean by the right-handed screw. This is the famous problem of Ozma (named after the mythical prince Ozma in Lyman Frank Baum's classic "Wonderful Wizard of OZ"). The *Ozma problem*, suggested by Martin Gardner, is a deep problem of communication theory, and its solution involves deeper understanding of symmetry in physics. We will return to it briefly later.

From the reflection symmetry of geometrical shapes let us now pass to the real question. Are the various laws of physics symmetric with respect to space reflection? To fix ideas consider a simple molecule CH_4 , the molecule of methane (marsh gas found commonly in marshy lands). The molecule consists of a carbon atom surrounded by four equidistant hydrogen atoms arranged at the vertices of a regular tetrahedron (Fig. 1.7).

The molecule is clearly reflection symmetric, *i.e.*, it is superposable on its image. We can break this reflection symmetry by replacing the four hydrogen atoms by four different atoms (or groups of atoms) X , Y , Z and W , say. Thus, for example if $X = \text{H}$, $Y = \text{CH}_3$, $Z = \text{C}_2\text{H}_5$ and $W = \text{OH}$, we get a molecule of butyl alcohol. Numerous other examples are possible. Chemists refer to such a molecule as having an *asymmetric carbon atom*, and the pair of enantiomers are called *stereoisomers*. Such a molecule is handed because it is no longer superposable on its mirror image. Thus for example, if you look down the XO direction, the atoms Y , Z and W will be seen as arranged either clockwise or anticlockwise. Now suppose we synthesize this molecule in the laboratory starting from the elements C, H and O. The question is which one of the pair of stereoisomers we will get. The answer is simply this. If the laws governing the chemical reaction are symmetric with respect to space reflection,

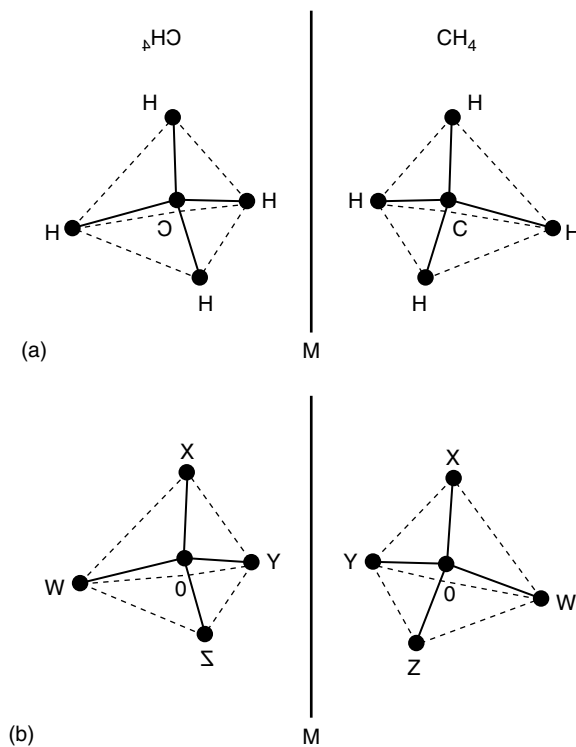


Figure 1.7: Mirror reflection of molecules: (a) CH_4 with symmetric carbon atom; (b) OXYZW with asymmetric carbon atom, hence optically active.

then the probabilities of getting the two stereoisomers are strictly equal. Therefore, at the end of the reaction we will get a mixture of the two in equal proportions. Chemists call this a racemic mixture. The mixture will have no *net* handedness. A law is said to be reflection symmetric, if a process or phenomenon and its mirror image are *equally* allowed by that law. Experimental evidence strongly suggests that the laws of physics that govern processes at low energies, like chemical reactions, are indeed reflection symmetric. Thus the molecules of butyl alcohol in our example and the molecule of its mirror image will have the same physical and chemical properties, *e.g.*, the same boiling point, the same freezing point, the same density and, of course, the same molecular weight. Next we will demonstrate the predictive power of this symmetry of the physical law — we will predict optical activity. Consider again our handed molecule with the asymmetric carbon atom. A molecule of sugar is perhaps a more pleasing example. Sugar molecules are also handed but a bit more complex. Now, we can hardly experiment on a single sugar molecule. So consider trillions of these identical sugar molecules — a solution of the sugar molecules in water, for example. The water molecules (H_2O) are mirror symmetric and, therefore, any handedness at all will be due only to the sugar molecules. We can and we will ignore

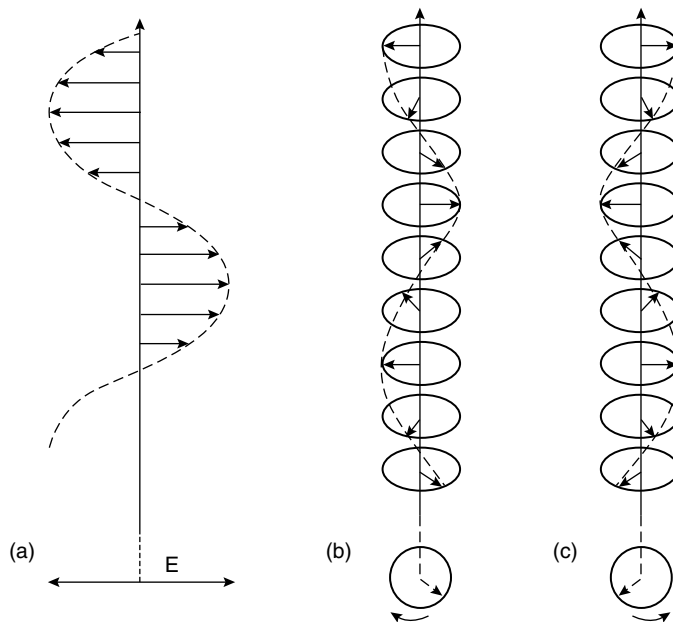


Figure 1.8: (a) Plane-polarized light; (b) left-circularly polarized light; (c) right-circularly polarized light.

the solvent (water) completely in what follows. In the aqueous solution the sugar molecules are located and oriented randomly. This, however, does not neutralize or average out their handedness. After all, a chest full of left gloves can hardly be confused with a chest full of right gloves, no matter how randomly the gloves are placed. We will now send a beam of light from a laser, say, through our sugar solution. But first let us remind ourselves of some elementary facts about light.

Light is a transverse electromagnetic wave. The electric and the magnetic fields oscillate sinusoidally in time and space with a given frequency and wavelength. They are perpendicular to each other and to the direction of propagation of the wave (hence transverse). Light can be circularly polarized. Here the tip of the electric vector describes a helix with its axis along the direction of propagation. It may be left- or right-circularly polarized according as to whether the helix is left- or right-handed (Fig. 1.8).

Light can also be linearly (or plane) polarized if the oscillating electric vector lies in a plane containing the direction of propagation. Finally, we note that a linearly (plane) polarized light may be viewed as a vector addition of the two oppositely circularly polarized light waves of the same frequency and wavelength. We are all set now. Let the beam of light passing through our sample of handed sugar solution be circularly polarized. The mirror image of this process will be an oppositely circularly polarized light passing through a sugar solution of opposite handedness. Given the reflection symmetry of the governing law, the two enantiomorphic processes must

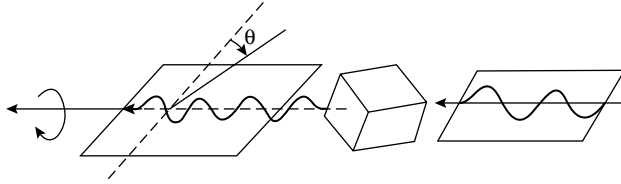


Figure 1.9: Rotation of plane of polarization of light by optically active medium.

be equally and identically allowed. In particular the speed of light in the two cases must be exactly the same. But what if we keep our sugar solution the same and only reverse the sense of circular polarization of light. Well, this is not symmetry related to the earlier situation, and there is no sufficient reason to expect the speed of light to remain the same — it will in general be different. Thus the speed of light in a handed medium depends on the sense of circular polarization of the light! This effect can be made more spectacular by taking our light to be plane polarized. Recall that it may be viewed as a superposition of two oppositely circularly polarized light waves. Now that these two components must travel with different speeds, they will get out of phase as they traverse the handed medium. This results in the twisting of the plane of polarization of the light relative to that of the incident light (Fig. 1.9). This twisting or rotation of the plane of polarization is called optical activity. It is perhaps the most dramatic manifestation of handedness of the medium. The substance (sugar in our case) is said to be dextrorotary (dextro = right) if the plane of polarization twists as a right handed screw. It is said to be levorotary (levo = left) if it twists as a left-handed screw. A racemic mixture of the two will leave the plane of polarization unchanged. (Somewhat confusingly, the opposite convention is also in use).

Let us re-emphasize that optical activity is due to the handedness of the substance. The law itself is even-handed, *i.e.*, reflection symmetric. This is expressed perhaps most forcefully by the famous example of a milk drinking kitten of the Looking Glass world. Milk contains asymmetric molecules of sugar, proteins and fats. So does, of course, the body of the cat. And conventional cats love conventional milk. Reflection symmetry now demands that the reflected kitten love the reflected milk just as much, and fare just as well in all respects.

The living world is, however, far from being racemic. Thus, practically all the 20 odd amino acids that make up the proteins of the living cells are left-handed. The proteins in the living cells, in turn, have a helical backbone which is almost always right-handed. Each of the sugar phosphate chains in the double helix of the information bearing molecule DNA is a right-handed (double) helix, and a man has typically 10^{11} kilometers of it. Left amino acids are common and are assimilated by our body, but the right amino acids are rare and filtered out by our kidneys. The nicotine commonly found in tobacco is known to be harmful but its reflected stereoisomer is rare and much less offensive. Limone, in perfumes, has the pleasant

orange scent — its enantiomer smells like turpentine. The same is true of other biochemicals such as the lactic acid found in milk, or the table sugar (sucrose) found in sugarcane, etc. The sugar D-glucose is found throughout the animal kingdom but its mirror image L-sugar is unknown except in laboratory synthesis. (Handedness of drug molecules poses a serious problem — one has only to recall the tragedy of the thalidomide babies with birth defects caused by the wrong handedness of the drug molecules involved. Preparing *optically pure* compounds, *i.e.*, those of a given handedness, is, however, difficult and expensive).

There are indeed few exceptions to the rule that anything from lactic acid to the double helical DNA, having handedness, will occur biologically in only one form. Indeed if we let a colony of bacteria feed on a *racemic* (optically inactive) mixture of (L) and right (R) sugars, the bacteria would feed preferentially on L-sugars, and then leave the mixture right-handed and optically active. The question now is, do we understand this dominance of handedness, or shall we say high-handedness of the living world when the governing laws themselves are so just and even-handed? Well, not quite. But a highly plausible answer is something like this. The observed handedness of the living matter may be the result of a fantastic amplification of an initial chance asymmetry, ever so slight. This is made possible by the positive feedback inherent in the process of multiplication (reproduction) by self-replication that is all pervasive in the animate world. To see this clearly let us simplify things to the absurd limit and consider the first (single) helical strand of the DNA molecule ever formed. We know that it is potentially equally likely to be right- or left-handed. But once formed it has got to be just one of them. So let it be right-handed. Now this single right-handed strand proliferates or multiplies by self-replication. It acts as a template and makes a copy of itself which is now necessarily right-handed. The process gets repeated over and over again. This is the positive feedback at work that may lead to the necessary amplification of an initial chance event over the aeons of chemical and biological evolution, and thus produce the handed life as we know it today. This is made all the more plausible by the observation that the inorganic world, by contrast, seems to be quite racemic. Consider the mineral quartz for example. It is one of the crystalline forms of silicon dioxide (SiO_2), the common silica sand. The basic unit here is SiO_2 which by itself is mirror symmetric, making quartz optically inactive when dissolved. But in a quartz crystal the units are arranged in the form of parallel helices which can be either left- or right-handed, making quartz optically active. In Nature both the forms occur with equal frequency.

Does this not go against the laws of thermodynamics, the entropy principle, that makes states of equal energy equally probable? The left- and the right-handed strands are, of course, energetically equivalent, being related by reflection symmetry. Well, the point is that the thermodynamic statement is about a system in thermal equilibrium. But the living state is far from equilibrium. It is self-organized and maintained at the cost of 'freely' available energy that comes eventually from the

sun. Once the cell is dead, the right-handed helices of the DNA molecule will begin to flip their handedness, and gradually tend to the racemic state as dictated by thermodynamics. Indeed, the rate of racemization can be, and has been, used for the dating of dead cells older than 40,000 years or so, much better than the conventional dating based on the decay of ^{14}C , a radioactive isotope of carbon. Louis Pasteur regarded handedness as a sign of life. Racemization signaled death.

Are the fundamental laws of physics all strictly symmetric under space reflection? Is the antipodal world of the *Looking Glass* just as legal as our conventional world? We now know that the answer to this question is a definite *no*. There are fundamental processes such as the β -decay (radioactivity) controlled by the so-called *weak interaction* that break this reflection symmetry. There is a screw at the very heart of Nature. To see this we have to get more sophisticated. We have seen how to reflect geometrical figures and shapes of material objects. But how do we reflect magnetism? Take a bar magnet with the poles N and S marked on its ends. Its mirror reflection will be just another bar magnet with the letters N and S laterally inverted (Fig. 1.10).

But this is a naive reflection of the body of the magnet. It hardly addresses the real question of how to reflect the magnetism of it. The magnetic field of the magnet may be regarded as due to an electric current circulating in a loop around the body of the magnet as indicated by the circular arrow (Fig. 1.10). Remember Ampère's Law! (Incidentally, when Ernst Mach learnt of the sideways deflection

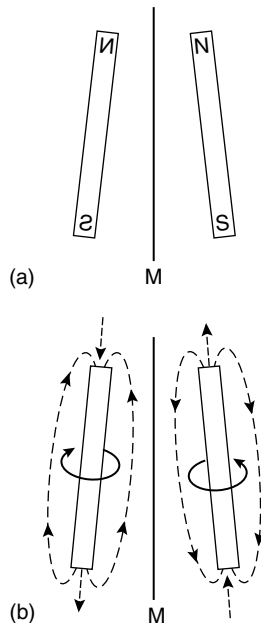


Figure 1.10: Mirror (M) reflection of (a) body of a magnet; (b) magnetism.

of a compass needle when placed below and parallel to a current carrying wire, he was shocked out of his wits as he thought it to be violating the left-right symmetry. With our picture of the magnet, now we see that Mach's shock was a false alarm as there is no such symmetry in this situation to start with). It should be clear now that when reflected, the sense of the circular arrow will reverse. And so will the polarity of the magnet. With this we are all set to describe the phenomenon that shook the world of physics — the fall of parity. Parity is yet another name for reflection symmetry. (Note that space reflection actually means inversion through the origin — that is letting (x, y, z) go to $(-x, -y, -z)$. In three-dimensional space, however, the inversion through the origin and the reflection in a mirror are related through a mere rotation by 180 degrees about an axis perpendicular to the mirror plane. And, of course, the rotation symmetry is not in doubt).

Take an atom of cobalt, the isotope ^{60}Co to be precise. The nucleus of ^{60}Co has a spin. It is like a spinning top. This makes it a tiny magnet with the magnetic poles on the spin axis. As before, this is equivalent to having a circulating current loop indicated by the circular arrow. We now apply a magnetic field. The nuclear magnet will align parallel to this field just as a compass needle aligns parallel to earth's magnetic field. The ^{60}Co is a radioactive nucleus. It decays by emitting, among other things, electrons, the β -rays, in all directions. The question is whether they come out equally in all directions, or there are some preferred directions. We can know this by placing detectors all around our sample and counting the number of electrons coming out in a given direction in a given interval of time. In particular let us compare the number of electrons shot out of the north pole (N) parallel to the field with the number shot out of the south pole (S) antiparallel to the field. If we perform this experiment as Chien-Shiung Wu did in 1957, we will find that more electrons are shot out of the south pole than out of the north pole. Is this consistent with the reflection symmetry of the underlying law? To answer this all we need to do is to look at the process reflected in a mirror (Fig. 1.11).

Everything looks the same except that the sense of the circular arrow and, therefore, the polarity, is reversed. Thus in the mirror world, more electrons would come out of the north pole antiparallel to the field than out of the south pole. The reflection symmetry is violated! (This violation of parity was predicted by the two Chinese-American physicists T. D. Lee and C. N. Yang in 1956 on theoretical grounds. It was confirmed by C. S. Wu in 1957. The same year Lee and Yang won the Nobel Prize in Physics). In fact the nuclear spin of the ^{60}Co nucleus (the circular arrow) and the preferred direction of electron emission define a left-handed screw. Nature is weakly left-handed after all! This is more than a mere convention. One could perhaps use the ^{60}Co decay to communicate to our distant correspondent the meaning of the left and the right, and thus solve the Ozma problem. But not quite, as we will presently see.

We can go further and, as it were, pinpoint the screw by looking at the full β -decay reaction: $^{60}\text{Co} \rightarrow ^{60}\text{Ni} + e^- + \bar{\nu}_e$. The neutrino (or rather anti-neutrino $\bar{\nu}_e$), as we have noted earlier, is an elusive particle with zero rest mass. This particle has

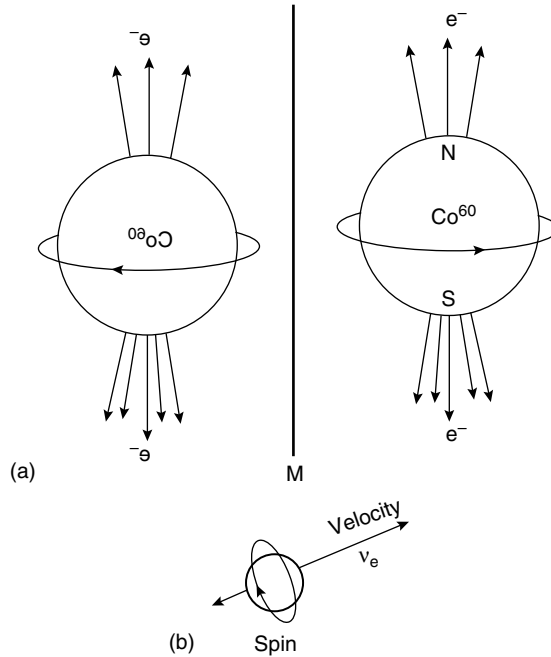


Figure 1.11: (a) Parity violation in β -decay of cobalt-60 nucleus; (b) left-handed neutrino.

a zero charge and relativity requires it to move with the speed of light. The neutrino has spin one-half and this is important for us, since relativity demands that the spin of this massless particle be either parallel or antiparallel to its velocity! Now the spin (the circular arrow) and the velocity (the linear arrow) form a screw or a helix that can be either right-handed or left-handed. In Nature we find only left-handed neutrinos (and right-handed antineutrinos) (Fig. 1.11). Here lies the screw at the heart of Nature! All reactions involving these handed objects violate parity.

Besides *parity* (\mathcal{P}), there are two other discrete symmetries — *charge conjugation* (\mathcal{C}) and *time reversal* (\mathcal{T}). The symmetry operation of *charge conjugation* (\mathcal{C}) replaces a particle with its antiparticle, denoted by an overhead bar. A particle and its antiparticle have the same mass but equal and opposite electric charges, among other things. Thus we speak of the anti-electron e^+ (commonly called positron), antiproton (\bar{p}), antineutron (\bar{n}), antineutrino ($\bar{\nu}$), and so on. The photon (γ) is its own antiparticle. Charge conjugation symmetry demands invariance of physical laws under the operation \mathcal{C} . Thus a reaction $X + Y \rightarrow Z + W$ and its conjugate $\bar{X} + \bar{Y} \rightarrow \bar{Z} + \bar{W}$ should proceed at the same rate. The antiworld is as allowed as our conventional world. And yet we see more electrons around than positrons, more protons than antiprotons and so on. The asymmetry seems to be of a cosmological origin, not fully understood at present. One thing should, however, be clear. We can hardly expect particles and antiparticles to co-exist in close proximity. They

would annihilate immediately producing a flash of radiation, *e.g.*, $e^- + e^+ \rightarrow \gamma + \gamma$. This positron annihilation is used in solid state physics to study electrons in metals.

Time-reversal symmetry demands invariance of the law under the operation of time reversal (\mathcal{T}). Thus, if we take a movie of a process and then re-run the reel backwards, what we observe will be an equally allowed process. The reaction $X + Y \rightarrow Z + W$ is as legal as the time-reversed reaction $Z + W \rightarrow X + Y$. The time-reversal operation (\mathcal{T}) requires reversing all velocities and spins in detail and interchanging past and future. Thus, at the level of elementary processes, there is no arrow of time. Microscopically, every process is reversible. (But how do we reconcile this microscopic reversibility with the all too common irreversibility of processes in complex systems — the irreversibility at the macroscopic level? What about ageing for instance? There is a thermodynamic arrow of time no doubt. The connection between the time-reversal symmetry of the microscopic laws and the observed asymmetry of complex processes has been and continues to be a subject of much debate. We will not pursue this matter here any further).

Like parity (\mathcal{P}), the time-reversal (\mathcal{T}) and the charge-conjugation (\mathcal{C}) symmetries are also approximate. There are subnuclear reactions in which \mathcal{C} and \mathcal{T} are individually violated. But amazingly, the combined action of these approximate symmetry operations (in any order) is an exact symmetry of nature with no violation known. Thus, if in any process we replace all particles by their respective antiparticles, reflect the resulting process in a mirror, and then reverse all velocities and interchange past and future, we will get an equally allowed process. This celebrated CPT theorem expresses a deep symmetry of Nature. “All Hell will break loose” if CPT invariance is ever found to be violated.

Finally, what about the Ozma problem? We now know why it is not sufficient just to ask our otherworldly correspondent to repeat the ^{60}Co experiment, as he (or she) may belong to the antiworld (of antimatter). There will always be an ambiguity inasmuch as both a right-handed helix of matter and a left-handed helix of antimatter will interpret the results of the experiment equally well. We must somehow ascertain before-hand whether they are made of matter or antimatter. It turns out that this is in fact possible. There are subnuclear reactions that violate time-reversal symmetry and eventually provide us with a method of ascertaining the material versus antimaterial nature of the distant world. The details are much too complicated, but the happy ending is that the Ozma problem is solved in principle.

1.4 Gauge Symmetry

We have been talking mostly about the geometric symmetries of space-time. These are the general framework symmetries without which the physical world will hardly be comprehensible. They seem so natural, almost *a priori*, that we take them for granted. Thus, the failure of symmetry under space reflection, even though a discrete and non-performable one, came as a great shock. Now we are approaching

a symmetry of an entirely different kind — the *gauge symmetry*. It is special, it is abstract and it appeals only to a preoccupied mind. Here we are requesting invariance of the law that there be, with respect to transformations that are simply outrageous. And yet the experience of the last five decades points to these gauge symmetries as the basic dynamical principles on which the fundamental interactions (forces) of Nature are designed. The familiar *electromagnetic* interaction that controls much of the low-energy physics and all of chemistry, the *strong* interactions that hold neutrons and protons together in the nucleus, the *weak* interactions responsible for the radioactive decay of unstable nuclei, and possibly even the universal *gravitation* that holds the planets and the stars together, all seem to fit in with this general scheme as gauge fields.

A proper understanding of gauge symmetry in physics requires a background knowledge of the framework theory, *quantum mechanics*, which is frankly outside the scope of this discussion (see, however, Appendix B). It is possible to get acquainted with the basic idea of gauge symmetry from an example that we know from our first year in college — the example of a simple harmonic oscillator (SHO). It will be a caricature, but real enough for our purpose. Let us get down to it without further apology.

Consider a particle performing a simple harmonic motion in a plane, *i.e.*, a two-dimensional SHO. What it means is that both its x and y coordinates oscillate sinusoidally with the same frequency. Thus the particle will in general describe an elliptical trajectory in the x, y -plane. It is convenient to combine the two motions along the x and the y axes into the motion of a single complex variable $z = x + iy$, where $i = \sqrt{-1}$ is the imaginary unity that keeps the real and the imaginary parts of z from getting scrambled up. The position of the particle in the x, y -plane is now labeled by a single complex variable z . This is the familiar Argand diagram, or the Gauss plane, for complex numbers (Fig. 1.12).

The magnitude of z is $r = \sqrt{x^2 + y^2}$, which is the distance of the particle from the origin O , and the polar angle θ is its angular position, where $\tan \theta = y/x$. The SHO is described by the equation $d^2z/dt^2 + \omega^2z = 0$. Here $2\pi/\omega$ is the time

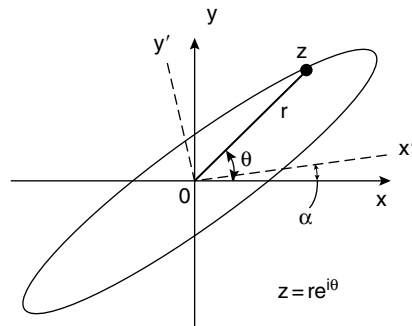


Figure 1.12: Elliptical trajectory of a two-dimensional harmonic oscillator in the complex plane representation.

period of oscillation. (We can even make ω time dependent and have parametric oscillations). As remarked before, the real and the imaginary parts of this equation do indeed describe the simple harmonic motions along the x and y axes.

Now comes the crucial observation. We have reckoned the angle θ from the x -axis. But this is just a matter of convenience. The absolute origin of the angle is irrelevant. And with this irrelevancy comes the freedom of choice. We could, for example, rotate our x, y axes anticlockwise by an angle α to new axes x', y' and reckon θ from the new x' -axis. This trivially amounts to replacing θ by $\theta - \alpha$. We say that we have re-gauged θ . All we have to do is to multiply our equation by $e^{-i\alpha}$ and absorb this phase factor by redefining $z' = ze^{-i\alpha}$, and our equation reads the same in terms of z' . Nothing really has changed. We could do this, of course, because α was a constant, *i.e.*, time independent. This irrelevance of absolute θ and the associated invariance of our equation is what we call the *global gauge freedom* and invariance. Global because it was an overall shift of θ , fixed and the same for all time. Encouraged by this, we now become more demanding. We demand freedom of choosing α differently at different times. That is to say we demand invariance under time-dependent shift $\alpha(t)$. This is the *local gauge invariance*, *i.e.*, local in time. But with $\alpha(t)$ varying with time, the factor $e^{i\alpha}$ can no longer be absorbed by the re-definition of z because of the time-derivative occurring in our equation. It will generate additional terms involving time-derivatives of $\alpha(t)$. Our earlier invariance of the equation is obviously lost. The question is if we can regain it with as little and as reasonable, or natural, a modification as possible of our original equation. In other words, can we introduce something that will *compensate* for these additional terms? It comes as a pleasant surprise that the answer is *yes*. All we have to do is to replace the time-derivative d/dt occurring in our equation by $d/dt - iA(t)$ with the proviso that re-gauging θ locally as $\theta - \alpha(t)$ should be accompanied by a re-gauging of $A(t)$ as $A(t) - d\alpha/dt$. Here $A(t)$ is the compensatory, or the 'gauge' field. That is all! But what have we gained after all this, you may well ask. Let us see. The time-dependent shift $\alpha(t)$ amounts to rotating our reference frame with an angular velocity $d\alpha/dt$. Now we may recall from our high-school mechanics that such a rotation gives rise to 'fictitious forces,' namely the centrifugal force and the Coriolis force acting on our particle. The centrifugal force is the radially outward directed force you feel while riding a merry-go-round. This is the force that makes the rotating earth bulge out at the equator. The Coriolis force is the force that makes you swerve sideways when you try to walk on a rotating platform. This is the force that deflects the winds and the ocean currents to the right (left) in the Northern (Southern) hemisphere due to Earth's rotation. After a little calculus our equation will show that the 'gauge field' $A(t)$ generates precisely these forces automatically. Thus, the requirement of local gauge invariance has created the right kind of forces acting on the particle in accord with experience. Is this not wonderful? This is the essence of local gauge symmetry.

It is now believed that all the fundamental forces of nature, the *electromagnetic*, the *weak*, the *strong*, and even the *gravitational*, are generated just this way. One

has to simply identify the correct global symmetry (the irrelevancy) that is to be gauged locally. This is where all the ingenuity and the insight of the theorist lie. We have spoken of the irrelevance of the absolute origin of space and time, and the irrelevance of the absolute orientation in the Minkowski space-time. When these global symmetries are gauged locally, we get Einstein's *general theory of relativity* that replaces the old-fashioned Newtonian gravitation acting in the old-fashioned Euclidean space. Thus gravitation appears as a gauge field. The idea of local gauge invariance really comes into its own only when it is combined with the framework of quantum mechanics, with all its built-in redundancies, irrelevancies and unobservables. For instance, as we have remarked earlier, the absolute phase of the wave function ψ of an electron is irrelevant. It can be changed globally by an arbitrary constant. But when we gauge it locally, the compensating force turns out to be just the electromagnetic force that we know so well from our experience. It couples to (acts on) the charges and the currents as it should. In point of fact, should we replace the single independent variable t in our oscillator equation by the three space co-ordinates (x, y, z) , generalize the gradients appropriately, and let $z(t)$ become $\psi(x, y, z)$, our equation will become the Schrödinger equation for a charged particle moving in a magnetic field represented by the gauge field $A(x, y, z)$, the so-called 'vector potential.'

When this gauge principle is applied to relativity-plus-quantum mechanics, it becomes the formidable gauge-field theory of physics today. The principle of local gauge invariance has become the guiding principle in our quest of fundamental understanding in the domain of the very small as well as the very large. Let us hasten to add that the same general principle appears again and again in our world of middle dimensions — the physics of *condensed matter*. So, next time you hear of gauge invariance, it may well be the gauge theory of ordinary glass, or its magnetic cousin, the 'spin glass.'

1.5 Spontaneous Symmetry Breaking (SSB)

Finally, we come to discussing an idea which is as deep as the idea of symmetry itself, or perhaps even deeper. Its time came much later. But now it is seen as a physical principle that holds the key to unifying all the fundamental forces of Nature, the *electromagnetic*, the *weak*, the *strong* and possibly even the *gravitational*. This has been in one form or another, the all-time dream of physicists. It is already partially realized now, and some say that the end is in sight. But, first, what is *spontaneous symmetry breaking*? Let us define it. A symmetry is said to be broken spontaneously if the symmetry of the state of the system is lower than (is a subgroup of) the symmetry of the force law governing the system. Mark you, we do not break the symmetry of the law itself. We have already hinted at such a possibility — remember the elliptical orbit of the earth around the sun in spite of the spherical

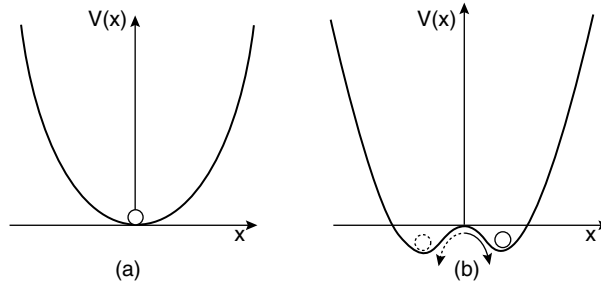


Figure 1.13: Particle in a symmetrical potential well: (a) symmetric state; (b) spontaneously-broken-symmetry state.

symmetry of the gravitational force of the sun! To fix the idea, let us consider two examples, the first a trivial one, taken from mechanics, and the second a highly non-trivial one taken from statistical mechanics where it all began.

Take a piece of wire. Bend it in a U-like shape and hold it vertically. Now slip a bead on the wire and let it slide freely on it. It is common knowledge that the bead will oscillate for a while and eventually settle down (due to friction) at the bottom of the U-wire (Fig. 1.13), this being the state of lowest potential energy (equilibrium).

The gravitational potential energy measured from the bottom is proportional to the height. Thus, the U-wire is really a ‘potential well’ with a single potential minimum at the bottom. Notice that the potential is symmetrical about the vertical through this minimum. Thus the state of the system has the same symmetry as the potential (the force law). Now, let us flatten the bottom part of our U-wire and finally make it convex upwards. We will now have two local minima of the potential located symmetrically about the midpoint which now becomes a local maximum. What should we expect now? The potential is still symmetrical about the vertical through the midpoint, but this is now a state of unstable equilibrium. A disturbance, however small, will tilt the balance in favor of one or the other of the two minima and the bead will roll down accordingly. Let it roll down to the right-side minimum. Now, the symmetry of this lopsided state is definitely lower than the symmetry of the potential which is still symmetrical about the vertical axis. This is *spontaneous symmetry breaking*. Broken symmetry agreed, but what is spontaneous about it, you may ask. After all we did need some disturbance to break it. Well, the point is this. The disturbance needed to break the symmetry of the state can be made arbitrarily small. Even the tiny thermal jiggling of molecules in the wire will do. The effect produced, namely the rolling down to one of the minima, is totally out of proportion to this tiny disturbance which could in principle be made almost zero — we have here a *critically poised atom*! This is why it is called spontaneous. (One is reminded of *Buridan’s ass*. The hapless ass was placed symmetrically between two identical bales of hay. The ass was hungry but the very symmetry (equidistance) of the two options forbade him from making up his mind and, as the parable goes,

he starved to death. But, of course, we know that the ass will eat — the slightest bias, even if an autosuggestion, or merely thinking about it, may make him turn to one or the other of the two stacks of hay!).

Now, we turn to the physically interesting example of a system of many interacting particles in thermal equilibrium. The branch of physics that deals with such systems is called *statistical mechanics* (see Appendix C). Most inanimate systems are of this kind. A good example is that of a ferromagnet. Take a piece of iron. For our purpose, we may regard the atoms of iron as tiny magnets, compass needles if you like. The origin of these tiny magnets, or the magnetic moments as they are called, lies in the spinning electrons. But this detail is not relevant for our discussion. These tiny magnets, shown as arrows in Fig. 1.14, interact with each other. The interaction is due to the quantum-mechanical ‘exchange’ of electrons because of their indistinguishability. But this is again a detail not important for our discussion. What is really important is that the interaction energy depends on the *angle between* these magnetic moments. Thus if we turn all the atomic magnetic moments around by the same angle about the same axis, the energy of the system will remain the same. We say that the law governing the system is spherically symmetric. Furthermore, for a ferromagnet the energy is minimum when the magnetic moments are all parallel to each other. It is clear, therefore, that these atomic moments will tend to align parallel to each other. At high temperatures, however, the thermal agitation will make these moments point in different directions at random so that there is no net magnetization. The state of the system will be spherically symmetric — it has the same symmetry as the law of interaction. We call this disordered, high-temperature symmetric phase the paramagnetic phase. As the sample is sufficiently cooled, however, the interaction energy favoring parallel alignment of the magnetic moments wins over the disrupting tendency of the thermal agitation. When this happens the magnetic moments align parallel to each other on average and thus the

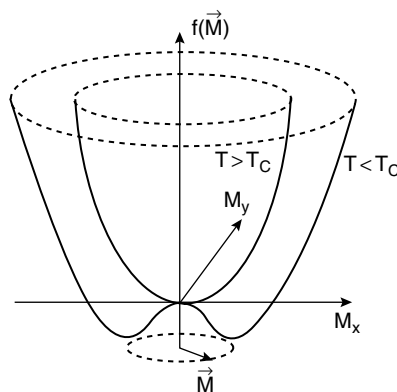


Figure 1.14: Spontaneous symmetry breaking in a ferromagnet at T_c . $f(\mathbf{M})$ denotes thermodynamic potential.

system develops a net magnetization \mathbf{M} , which grows in magnitude with decreasing temperature. This low-temperature ordered state is called the ferromagnetic phase. The temperature T_c at which the system makes the continuous transition from the high-temperature disordered phase to the low-temperature ordered phase is called the critical temperature, or the Curie temperature. The magnetization which is a measure of order is referred to as the *order parameter*. (The physics of continuous phase transition, often called the second-order phase transition, at and about the critical temperature has been an extremely active area of research of our times. It is determined almost entirely by the symmetry and the dimensionality of the order parameter and is quite independent of the microscopic details of chemical composition, etc. This ‘*universality*’ of the ‘critical behavior’ is most fascinating but we must let it pass). The question now is what should be the direction of this net magnetization \mathbf{M} . Inasmuch as the energy depends only on the relative orientation of the atomic moments, all directions of \mathbf{M} are equally probable statistically. And yet in a given realization some direction of \mathbf{M} must get selected. This state is then symmetric only for rotations about this direction of \mathbf{M} — it has only an axial symmetry which is a subgroup of the full spherical symmetry of the interaction law. The symmetry is thus *spontaneously broken!* For a large, in principle infinite, system an arbitrarily small magnetic field or anisotropy will fix the direction of \mathbf{M} . The connection with our mechanical example should be obvious. The order parameter (magnetization \mathbf{M}) plays the role of displacement x . Instead of mechanical potential energy $V(x)$ which was to be minimized, here we have a thermodynamic potential (free energy) $f(\mathbf{M})$, which is to be minimized. The only detail that differs is that whereas the position x in our mechanical case was a scalar (one-dimensional), the order parameter \mathbf{M} is a vector. Thus, in the mechanical case the two equivalent minima were separated by a potential barrier in the broken symmetry phase, while in the ferromagnetic case all the equivalent minima (differing only by the direction of \mathbf{M}) are degenerate (*i.e.*, have the same free energy) and \mathbf{M} can in principle *freely* gyrate among them. Perhaps a better mechanical analogue would have been the marble in a punted wine bottle. (Incidentally, this freedom leads to the possibility of certain waves propagating in the broken symmetry phase whose frequency tends to zero as the wavelength tends to infinity (*i.e.*, they are massless). We call these *Goldstone modes*. For the (antiferro-)magnetic case, these are the spinwaves). All this plays an important role in the physics of phase transition in *condensed matter*. We emphasize that this is a highly cooperative phenomenon resulting from interaction among large number, infinite in principle, of particles, their spins in this case.

Now, how can all this possibly bring about unification of the fundamental forces of Nature? This is a magnificent and highly technical obsession of contemporary physics. We will try to give just the flavor of it in plain words. Any symmetry can be broken spontaneously. In particular and most importantly, it can be the *gauge symmetry*. It is the combination of gauge symmetry and *spontaneous symmetry breaking* that is central to unification. Consider the simplest case when the

matter consists of charged particles (electrons). As we have already seen, the global symmetry (namely, the irrelevance of absolute phase) when gauged locally, generates the electromagnetic field automatically, which in the quantum version is the photon. (It is inherent in this mechanism that the photon — the gauge field — be massless). Now in quantum theory, the interactions between particles are mediated by the exchange of quanta of some field (much the same way as the exchange of a handball between two players will exert an effective force of repulsion between them. For attraction, let them exchange boomerangs!). Thus, the photon mediates interaction between charges. It also follows from general quantum principles that the range of interaction be inversely proportional to the rest mass of the quanta exchanged. This is why the range of the electromagnetic interaction is infinite. This intimate ‘genetic’ connection between matter (electrons) and the gauge field (photons) leads us to expect an induced change in the character of the gauge field when the matter undergoes a phase transition in which the very global symmetry, whose local gauging generated the gauge field, undergoes spontaneous breakdown. In point of fact, it would be very surprising if it were otherwise. We already see this effect in a superconductor in the laboratory (see Chapter 3 on Superconductivity). Here electrons undergo a transition in which the phase of this collective (macroscopic) wave function takes on a definite value, breaking thus the global symmetry spontaneously. This induces the photon to acquire a non-zero mass, making it impossible for it to propagate very far into the superconductor. This explains the famous Meissner effect, namely that the magnetic field is expelled from the bulk of a superconductor.

This was the simplest, but a most striking demonstration, of the change of character of a gauge field induced by the SSB of the matter (field). All we have to do now is to generalize to more complicated internal symmetries that can be imagined and indeed have been postulated. Thus, there may be several gauge fields generated by local gauging. They may be all symmetry related and thus of the same character. Now, if the matter undergoes SSB, the group of these gauge fields may be split into subgroups, and different subgroups may acquire different masses. Successive phase transitions (and the associated SSB’s) may generate thus a gamut of fields with different characters. This generation of different gauge fields (fundamental forces) by the descent of symmetry due to SSB, from the single most symmetric initial entity is the dream of the *grand unified theory* (GUT). It has already been partially realized in the *unification* of the *electromagnetic* and the *weak* interaction by Glashow, Salam and Weinberg for which they won the 1979 Nobel Prize in Physics.

One has a plausible scenario in mind that the universe began totally symmetric with a Big Bang some 15 billion years ago. As it expanded it cooled and underwent successive phase transitions. The associated SSB’s led to the diversity of fields that survive at the present epoch. And what a diversity — if the strong interaction measures unity on a certain scale, the electromagnetic interaction will measure 10^{-2} , the weak interaction 10^{-5} and the gravitational interaction 10^{-34} !

The strong interaction acts on hadrons (protons, neutrons, pions, etc., or their postulated building blocks, the quarks) but not on leptons (electrons, neutrinos, etc.) and is short-ranged. The weak interaction involves neutrinos and has a still shorter-range. The electromagnetic interaction acts on all charged particles and has infinite range. The gravitational interaction is the weakest of all, but acts universally on everything and has infinite range. And yet all may have a common origin. This reminds one of the concept of the Nirguna Brahma of the ancient Hindus — formless, featureless, totally symmetric pure existence, from which all diversity originated, shall we say, by spontaneous symmetry breaking!

This brings us to the end of our exploration of symmetry. We have seen its power. Obviously, symmetry cannot answer all the *why's* and *how's*, but it does reduce them to fewer *why's* and *how's*. To the philosophical question of why Nature is so symmetric, we can perhaps answer thus. Symmetry is, in the ultimate analysis, absence of bias. It is an expression of justice. There is a principle of insufficient reason against asymmetry. A sphere is admitted. But a deviation from sphericity must bide our question.

Galileo had spoken of the great Book of Nature. We should perhaps add that the first and the last Chapters of this Book are on *symmetry* and its *spontaneous breakdown*, respectively.

1.6 Summary

Symmetry means invariance of an object with respect to a set of operations called symmetry operations to be performed on it. The object may be the geometrical form of body such as a crystal of common salt, and the set of operations may be the geometrical operations of translation along a direction, rotation about an axis, or reflection in a plane. The symmetry operations may be continuous or discrete, physically performable or non-performable. More importantly, the object may be a law of nature itself expressed mathematically by a certain equation. Symmetry then means the invariance, or rather covariance, of the form of the equation under the mathematical transformations corresponding to the symmetry operations, that may not be geometrical in nature. Symmetry is a powerful physical principle that helps us not only simplify calculations and classify and unify diverse objects, it also restricts the possibilities in the absence of complete knowledge of the physical world. It creates new physics when a symmetry is requested on intuitive grounds. There is a branch of mathematics called *group theory* that provides the proper and powerful language for dealing with symmetry. Symmetry has played a fundamental role in quantum physics, particularly in the domain of high-energy physics, where its predictive power has been fully vindicated. The rather abstract idea of *gauge symmetry* is one of the profoundest concepts produced by the human mind. Any symmetry, however compelling aesthetically it may be, must be established experimentally.

Thus, *parity* signifying the left-right symmetry between an object or a process and its mirror image turned out to be false in certain fundamental processes involving neutrinos. Symmetry can also be broken spontaneously when there is a phase transition. The idea of spontaneous symmetry breaking has played a decisive role in our understanding of phase transitions in general, and in the context of grand unified theories and the early universe in particular. The search for deeper, hidden symmetries of Nature continues.

1.7 Further Reading

Books

- M. Gardner, *The New Ambidextrous Universe*, 3rd Ed. (W. H. Freeman and Company, New York, 1995).
- H. Weyl, *Symmetry* (Princeton University Press, Princeton, 1952).
- I. Hargittai, *Symmetry 2: Unifying Human Understanding* (Pergamon Press, Oxford, 1989).
- A. Zee, *Fearful Symmetry: The Search for Beauty in Modern Physics* (Macmillan Publishing Company, New York, 1986).