Symmetry Principles and the Axiomatic Structure of Physical Theories

Theoretical Physics Research Unit, Birckbeck, University of London

Abstract

Symmetry plays an essential role in scientific thinking. When faced with a problem, the mind often tries symmetric solution. It also picks such solution when there are alternatives. Thus, symmetry plays a methodological role in acquiring scientific knowledge. In this article, we try to explain how symmetries inter naturally into the structure of our theories, and reveal its consequences.

Virtually all theorists agree that justified-true-belief is a necessary and sufficient condition for knowledge. Therefore, one can proceed to "define" science in general, and physics in particular, to be a mathematical representation of our knowledge about the universe. Thus, we can say that physics is based upon the assumption that nature can be understood mathematically. And implicit in that assumption is our belief that nature is not arbitrary, but that it evolves according to definite (mathematical) laws. What does mathematics has to do with it? Even though we do not expect a definite answer to this question, we may reason as follow: Man's position in nature determines his profound understanding of the universe. Our (scientific) conceptual framework seems to presuppose abstract entities (counting, for example, requires the set of positive integers). All relations between us and the world are in such a way that we have to destroy our earlier clear-cut (particular) picture about nature in order to grasp universal (general) concepts which often are of abstract nature (Newton introduced the point-mass concept in his laws of motion and Einstein had to introduce curved space-time into his general relativity). Now, since mathematics is about a realm of abstract objects, therefore its role in virtually every science is inevitable.

Having said all this, we still have to distinguish between the pure mathematical reasoning and the usually employed physical reasoning. Mathematical reasoning consists of definite logical rules on adopted (i.e. arbitrary) system of definitions and axioms. The process of mathematical deduction contains no further information other than those in the initial system of definitions and axioms. And mathematical assertions are valid for the abstract objects introduced by means of the definitions. However, if a correspondence has been established between these objects and real (physical) objects such that a mathematical structure, under certain conditions, gives a correct description of the behavior of physical objects, then we say that a physical realization of that mathematical structure has been found.

Since physics describes events occurring in space and time, the best description can be achieved by representing our space and time by an appropriate mathematical structure which involves as few assumptions as possible. The mathematical structure which has proved to be a good starting point is the so called differentiable manifold. In general, n-Manifold is a topological space which is locally homeomorphic to the n-dimensional real vector space . Local homeomorphism enables us to associate a set of n numbers, called the local coordinates, to each point of the manifold. In a local coordinate system, every point on the manifold must have unique coordinates and nearby points have nearby coordinates. So, we can think of the manifold as a collection of points, each of which will eventually correspond to a unique position in space and time and the entire collection represents the history of our universe. If the manifold is not homeomorphic to $\mathbb{R}^n$ globally, we have to introduce several overlapping coordinate systems, each covering a part of the manifold. Therefore, in the overlapping region, it is possible for a single point to have two or more coordinates. However, differentiability of functions must be preserved in all coordinate systems. Thus, we need a smooth transition (transformation) from one coordinate system to another. This allows us to describe physical events and quantities in terms of space and time changes of some differentiable functions (fields). Of course, we do not know for certain that space and time are as smooth as the manifold itself, but at least there is no evidence for any discreteness down to the smallest scale we are able to probe experimentally.

Now that we have modelled space and time by n-manifold, we realize that the class of mathematically permitted manifolds is too big for all physical purposes and that smoothness and other topological features do not help narrowing down our choices. So, one must present physical arguments for eliminating as many classes of “bad” manifolds as one possibly can:

1. One would perhaps not want to admit a space-time manifold with closed time-like curves as this would violate causality. If one chooses a time-oriented n-manifold, one can establish directionality to the time coordinate and foliate the whole space-time into stacks of space-like hypersurfaces (Cauchy surfaces). Events on each Cauchy surface cannot influence each other, and evolve independently along a global time coordinate. Information on a single Cauchy surface (i.e. suitable initial conditions) is sufficient to determine the subsequent state of the system in space-time. Thus, causality requires the space-time to be an oriented smooth n-manifold.

2. Topologically, this M^n is the product of a space-like (n -1)-manifold Σ^{n-1} and the time-like 1-manifold $\mathbb{R}$.

3. One would use the physical notion of "distinct events" to exclude (non-Hausdorff) manifolds in which there are two points which cannot be separated by disjoint neighborhoods.

4. Separate connected regions in space-time should be able to communicate. Therefore, one would want to exclude non-connected manifolds.

5. Our space-time manifold must admit a metric of Lorentz signature. Some compact connected manifold do possess such metric, but compact manifolds also have closed timelike curves mixing past and future. And, since paracompact manifold does admit a metric of Lorentz signature, one needs to exclude non-paracompact manifold (i.e. manifold in which some connected component cannot be covered by a countable collection of coordinate systems).

6. Since space-time has no edge (at least we have not seen an edge), one would exclude bounded manifold.

The conclusion from all these arguments, then, is that space and time can be modelled by a smooth oriented n-manifold with unbounded, connected, paracompact and Hausdroff included in the term "manifold".

Theoretical physics is characterized by the use of *mathematical models*, *measurable quantities* and "*fundamental principles*" to describe, understand and predict the behaviour of physical objects. The merit of a physical theory lies not just in its ability to explain observed phenomena, but also to predict a new one. To construct a physical theory by deductive manner, one must form a general *set of concepts* whose introduction is suggested by the observed phenomena; limit the range of application of these concepts by some sort of "*fundamental principles*" and show that the limited concepts, together with the mathematical relations between them, form a *self-consistent* scheme.

By observing nature, an ambiguous "interaction" takes place between man and the world outside. Mysteriously this process enables conscious observer to build a system of "qualitative" concepts whose nature is independent of the existence of such observer and his intervention during the act of observation. However, it is almost always possible to provide these concepts with quantitative features which bear the signs of different procedures of experimental study of physical objects. Thus, from the mathematical point of view, we must expect that our concepts are ill-defined.

The main function of any theory is to provide us with useful information about the world. This means that the theory must contain a number of elements which turn it into a meaningful language. In particular, there must be constant as well as variable elements, as a characteristic objects, in the theory. The simplest way of obtaining information can be achieved by means of experimental study which consists of frames of reference (or instruments) by means of which a well-defined measuring procedure is implemented. An experimental study will be possible (i.e. reproduces the results) provided that we choose a class of "equivalent" frames of reference to work with. In principle, an equivalence relation (between reference frames) can be established if, and only if, the observers: 1) are in well-defined state of motion, and 2) agree on a standard scale for length, time and mass. Regardless of how the equivalence of the frames is realized in practice, the equivalence relation has the structure of a group. Thus, any physical theory must contain, as an axiom, some symmetry principle defined by a group of transformations acting on the elements of the theory. In a class of equivalent frames with respect to a given group of transformations, the symmetry group can be used to translate observations made in one frame to any other frame reached by the transformations. This defines a principle of relativity which asserts that the laws of physics are the same for all equivalent observers, i.e. the equations of motion must be covariant (form-invariant) under the action of the symmetry group. This, in turn, determines the results which don't depend on the choice of the frame of reference. Therefore, the above-mentioned constant elements of the theory can be obtained in terms of the set of all invariant (Casmir's) operators of the given symmetry group. Thus, one expects that possible motion and interaction will be severely restricted by the values of these invariant operators. Indeed, all kinematical and dynamical properties of a given physical system are determined in terms of the set of all invariant operators associated with all possible symmetries of the system. That is to say that the collection of states of the system forms a manifold characterized by the values of the invariant operators. This manifold is also called the representation space of the symmetry group. For example, the manifold of (quantum) states for a particle can be labelled by the mass of the particle, the spin of the particle and certain other invariant parameters such as the electric and colour charge. In some sense, physics is very much similar to Klein's geometry where all properties of geometrical objects are determined by sets of invariants of the *global* symmetry group of the space.

With respect to the symmetry group of the theory, quantities with certain transformation laws can be identified with the variable elements of the physical theory. Obviously, they depend on the choice of the frame of reference. However, locally invariant combinations can always be constructed out of them. This in turn determines the locally measurable quantities. Thus, as advertised, the principles of symmetry and invariance limit the range of basic physical concepts and determine the structure of dynamical quantities in the physical theory.

Symmetry groups and their invariants provide, at the same time, good theoretical description of the experimental instruments. Therefore, there must be a room for the properties of the instruments in the axiomatic structure of our theory. Conversely, the choice of the instruments and procedures of an experiment predetermines the symmetry group of the theory describing the given experiment. Similar situation also exists in geometry. Indeed, physics and geometry share the same property regarding the relationship between theoretical and experimental ways of looking at the world.

The principles of special relativity, for example, deform and prevent the concept of a rigid body from being rigid. The deformation of rigid body leads us to the idea of describing extended objects by differentiable functions, i.e. fields over space-time. However, not every field represents an extended physical object. Again, this is because of the same relativity principles which limit the set of fields to those with definite transformation laws with respect to the Lorentz group. This way, the fields will describe the same physics in all frames of reference reached by the Lorentz transformations. Thus, physical field must carry a representation of the Lorentz group, i.e. it must be labelled by Lorentz invariant parameters (mass and spin).

Under the same relativity principles, to be *elementary,* the elementary particle must be structureless (mathematical) point with certain physical properties. However, this definition is not perfect because it brings about the problem of divergence and prevent the particle from having angular momentum of its own. The appearance of divergent quantities in a self-consistent theory (roughly) defines the scale at which the theory loses its predictive power and calls for a new theory. For example, the relativistic invariance of classical electrodynamics implies that the electron must be treated as a point-like charge in classical electrodynamics (indeed, down to distances of the order of 10^{-15}cm, scattering experiments on electrons show no evidence of structure or extension). However, point-like electron in classical electrodynamics leads to divergent electrostatic self-energy ($u = e^2/r \to \infty, \text{as } r \to 0$) and, therefore, an infinite electron mass ($m = u/c^2 \to \infty$). This absurd result means that classical electrodynamics breaks down at short distances and, therefore has little or no relevance to the real world of electrons and other charged elementary particles. But, how short is that "short distances"? An estimate can be obtained by comparing the electron rest energy, mc^2, with the electrostatic energy of an extended classical distribution of charge totaling the electronic charge, e^2/R. Thus, we conclude that classical electrodynamics breaks down at distances of the order of $R \sim e^2/mc^2 \sim 10^{-13}$cm. However, this unreal picture of the world (without point-like elementary particles) disappears at distances two order of magnitude higher. Indeed, we know that quantum electrodynamics, which is a unified description of particles and fields, become necessary at distances of the order of $\hbar/mc \sim 137R \sim 10^{-11}$cm.

We may now say that any consistent physical theory can be formulated in terms of the appropriate principles of symmetry and invariance. These are what we called earlier *"fundamental principles".* Using these principles, it is almost always possible to turn an abstract mathematical structure into useful apparatus. The main objective of any such apparatus is to describe the dynamics of physical systems, i.e. the differential equations of motion and the set of all conserved quantities. Self-consistency of the scheme is proved by showing that the symmetry is not violated on the dynamical level and this leads to the remarkable fact that the conserved quantities form a representation of the Lie algebra of the symmetry group. The equations of motion can be obtained by the following steps:

1. The dynamical variables of the system are represented by continuous (may

be complex) functions (fields) on the n-dimensional space-time manifold M^n, $\varphi_r(t, \vec{x}) \in \mathcal{C}^\infty(M^n, \mathbb{C}^r)$, and their derivatives to a finite order, $\partial_a\varphi_r$, $\partial_a\partial_b\varphi_r, \cdots$, we assume that the fields φ_r and their derivatives vanish sufficiently fast as $|\vec{x}| \to \infty$.

2. The invariance or covariance (i.e, the transformation laws) of these variables with respect to the symmetry group provides sufficient information to construct a real scalar function, $\mathcal{L} : \mathbb{C}^r \times \mathbb{C}^{nr} \times \cdots \to \mathbb{R}$, called the Lagrangian of the system, $\mathcal{L}(x) = \mathcal{L}(\varphi_r(x), \partial_a\varphi_r(x), \cdots)$. On the other hand, given $\mathcal{L}$, certain criteria for the symmetry can be defined (see below).

3. The whole dynamics then rests on the statement

$$\delta \int_{\Omega\subset\mathbb{R}^n} d^n x\, \mathcal{L}(x) = 0.$$

   This is an expression of general behaviour in nature. The vanishing variation of the action (or the principle of least action) is one of the most fundamental achievements of theoretical thought. Since the time of Hamilton, practically all observed phenomena have been described by equations shown to be the consequences of a similar principle. The equations of motion follow by assuming that the variations of the fields, $\delta\varphi_r$, vanish at the boundary, $\partial\Omega$, of some space-time domain Ω but arbitrary elsewhere. Various type of conserved quantities (depending on the symmetry group) can be then constructed when the dynamical variables satisfy the equations of motion.

If, by using only the transformation laws of the fields (i.e. without the use of the equations of motion), the Lagrangian changes according to

$$\delta\mathcal{L} = \partial_a\, \Lambda^a_\alpha(\varphi_r),$$

where $\Lambda^a_\alpha(\varphi_r)$ are some functions of the fields φ_a and α is a group index, then the action integral is unaffected (because the fields vanish on the boundary):

$$\delta\left(\int_\Omega d^n x\, \mathcal{L}\right) = \int_\Omega d^n x\, \partial_a \Lambda^a_\alpha = \int_{\partial\Omega} d\sigma_a\, \Lambda^a_\alpha(\varphi) = 0,$$

and the transformation is a *symmetry* transformation. Two situations are now distinguished: if $\Lambda^a_\alpha = 0$, we say that we are dealing with an *internal* symmetry; otherwise we have a *space-time* symmetry. When the fields satisfy the equations of motion, it is always true that

$$\delta\mathcal{L} = \partial_a \left(\frac{\partial \mathcal{L}}{\partial \left(\partial_a \varphi_r \right)} \, \delta_\alpha \varphi_r \right).$$

Thus, the symmetry implies that the object (Noether current)

$$J^a_\alpha \equiv \Lambda^a_\alpha - \frac{\partial \mathcal{L}}{\partial \left(\partial_a \varphi_r \right)} \, \delta_\alpha \varphi_r,$$

satisfies a continuity (conservation) equation, $\partial_a J^a = 0$, and allows us to define time-independent quantities, Q_α, (charges) by the integral

$$Q_\alpha = \int d^{n-1}x \, J^0_\alpha.$$

Then, our home work is to show that these charges have the following properties:

1. They define a set of Constants of motion,

$$\frac{d}{dt} \, Q_\alpha = 0.$$

2. They transform covariantly with respect to the symmetry group,

$$\delta_A Q_\alpha = \left(M_A \right)_\alpha{}^\beta \, Q_\beta,$$

   where M_A are finite-dimensional matrix representation of the generators of the symmetry group.

3. They generate the correct transformations on the fields,

$$\delta_\alpha \varphi_s(x) = [iQ_\alpha, \varphi_s(x)].$$

4. They satisfy the Lie algebra of the symmetry group,

$$[Q_\alpha, Q_\beta] = iC_{\alpha\beta}{}^\gamma Q_\gamma.$$

To describe the behaviour of a given system, the equations of motion need to be solved. To do this it is necessary to design a theoretical model of the system by setting up initial or boundary conditions, otherwise the solutions would have a great degree of generality. Fortunately, given a physical system, it is possible to identify a set of initial conditions, and given these same conditions, the resulting state (motion) of the system will be same and independent of where and when these conditions are realized. That is, the possible states of the system are independent of the space-time location of the observer, i.e. space-time is homogeneous. In the language of symmetry, this is equivalent to the statement that the equations of motion are covariant with respect to the group of translations $x^a \to \bar{x}^a = x^a + c^a$ by a constants c^a. Space is also isotropic: it is an experimental fact that the orientations of an event is another irrelevant initial condition. Thus, the equations of motion must be covariant with respect to spatial rotations. Experiments have also revealed the fact that light signals travel with same speed and pay no attention to the observer's state of motion. Maxwell's theory indicates that the motion of an observer, as long as it is uniform with constant velocity, is likewise an irrelevant initial condition. Putting all of the above together, we arrive at the principle of relativistic invariance which states that two observers moving with constant relative velocity will see the same physics. More precisely, the equations of motion (laws of nature) are covariant with respect to the Poincare group of transformations, $\bar{x}^a = \Lambda^a{}_b \, x^b + c^a$, where the transformation matrix $\Lambda^a{}_b$ depends on six parameters representing three rotation angles and three boosts. Two systems of coordinate related by Poincare transformation are said to be equivalent.

Conclusions

Finally, let us summaries our conclusions by saying that any consistent physical theory must satisfy the following conditions:

1. There should be a symmetry transformations, and all the frames which are connected by these transformations are equally good for the description of a given system: they define equivalent observers.

2. Equivalent observers should be able to communicate, i.e. there should be definite rules transform the dynamical variables of a given system from one reference frame to any equivalent frame. This is equivalent to the mathematical problem of finding all representations of the symmetry group.

3. Translating a physically possible situation should also be physically possible, i.e. a possible motion in one system should again appear possible in any equivalent coordinate system. That is to say that equivalent observers should make the same prediction regarding the outcome of an experiment carried out on a given system.

4. Condition (3) above should be identical for all equivalent observers, i.e. the equations of motion should have the same form in all equivalent coordinate systems.

For example, if we take the Lorentz transformation, $\bar{x}^a = \Lambda^a{}_b \, x^b$, to be the symmetry transformation, then condition (2) implies that the fields transform according to

$$\bar{\varphi}_r(\bar{x}) = D_r{}^s(\Lambda) \, \varphi_s(x),$$

with $D(\Lambda) = I$, if Λ is the identity transformation. Acting on the finite dimensional vector space of the dynamical variables $\varphi_r \, , r = 1, 2, \cdots, p$, $D(\Lambda)$ is a definite non-singular matrix representation of the Lorentz group $SO(1, n-1)$, i.e. $D(\Lambda)$ preserves the group multiplication law

$$D(\Lambda_1) \, D(\Lambda_2) = D(\Lambda_1 \Lambda_2).$$

Condition (3) then says that if $\varphi(x)$ describes a possible physical situation, then $D(\Lambda) \, \varphi(x)$, which is the possible situation as seen by the observer in $\bar{x}$ − system, is also a possible situation in the "original" x − system. And finally, condition (4) asserts that $D(\Lambda)$ depends only on the relation between the two coordinate systems and not on the intrinsic properties of either one.