The reciprocal interaction between mathematics and natural law

Sunil Mukhi

Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India

In this article it is argued that the relationship between mathematics and physics has undergone a qualitative change in the last two decades. We have entered a new era in which research in theoretical physics is providing a stream of new mathematical ideas and relationships, some of which are yet to be understood using conventional mathematical tools. Some examples of this new paradigm are provided.

IN ancient times the distinction among mathematics, physics and other natural sciences was not clearly articulated, nor was there necessarily a distinction among the practitioners of these sciences. Many mathematicians were also physicists and vice versa, Newton being only the most famous example. But over the last century or so, it is harder to find examples of individuals straddling both fields, and the modern university typically has distinct mathematics and physics departments in pursuit of distinct goals.

In the era of 'modern physics', starting from the early twentieth century, physicists have seen mathematics primarily as a language in which their observations can be codified. In his famous article 'The unreasonable effectiveness of mathematics in the natural sciences', Wigner put it as follows: 'The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift....' Many new developments have taken place in the four-and-a-half decades since Wigner made this observation. The miracle to which he was referring remains alive and well today, but there are also indications that major changes are on the way.

As Wigner indicated, the new physical notions that came up during the twentieth century readily found a setting in established areas of mathematics. One such example is the theory of continuous groups, known as 'Lie groups', and their associated algebras. It was realized that rotations in space form a group, the orthogonal group in three dimensions, O(3), whose representations play an especially important role in quantum mechanics. In special relativity, the rotation group gets subsumed into the Lorentz group, O(3, 1), which contains besides rotations, the 'boosts' between space and time. In the 1960s it became necessary to invoke a new class of groups, the unitary groups, to classify fundamental particles. The experimentally observed particles conveniently fell into representations, or multiplets, of these unitary groups.

e-mail: mukhi@theory.tifr.res.in

It was just as well that orthogonal and unitary groups had already been discovered and classified by the time they were needed in physics. Otherwise, for example, the discoverers of particle spectra would have been at a loss to deal with the regularities they were observing, which could have found no explanation without an understanding of unitary symmetry.

Once the concept of Lie groups entered physics there was considerable work done by physicists to further develop the mathematics. However, it is fair to say that from the mathematicians' point of view, this work did not add significant conceptual material to what was already present in the mathematics literature. To be sure, physicists did discover new and important facts about Lie groups, but they were typically not the sort of facts, and did not have the degree of beauty and depth, that would appeal to mathematicians. Hence it is not surprising that the enthusiasm of physicists to learn mathematics has usually been greater than that of mathematicians to learn physics, the latter being less confident that they would have something to gain.

Many other twentieth-century developments, such as the introduction of differential geometry in the study of gravity, of infinite-dimensional vector spaces in quantum mechanics, and of fibre bundles in gauge theory, exemplify the same story. The role of mathematics in physics in all these cases was limited in two ways. First, the core mathematical results needed had already been obtained, and second, despite considerable mathematical work by physicists, mathematicians did not learn much of significance to themselves from the physicists.

Of course this is a value judgement, and heavily depends on what we mean by 'of significance'. Wigner himself wrote a textbook entitled *Group theory*. But then, the subtitle of this book was *and its Application to the Quantum Mechanics of Atomic Spectra* and the primary audience for the book is physicists. Taking other similar cases into account, this value judgement does not seem far off course.

However, during the 1980s, something about this situation changed quite drastically. The change is hard to understand at present, because we are still in the middle of it, or maybe close to the beginning. It will be the focus of this article. My main observation will be that the interface between mathematics and physics has turned into a two-way relationship of a unique and unprecedented nature. Not only does mathematics provide physics with a formulation it can use, but physics has started to provide mathematics with methods and results that it needs, desires and appreciates on its

own terms. It would be tempting to describe this new situation in terms of 'The unreasonable effectiveness of physics in mathematics', which would have made an appropriate title for this article. But an internet search reveals that this very title has been used in articles and colloquia by distinguished mathematicians and physicists like Michael Atiyah, Robbert Dijkgraaf and Arthur Jaffe, just over the last two years. So there appears to be a growing consensus that the role reversal which will be the subject of our discussion is really taking place.

Consistency and experiment

Before turning to the new paradigms thrown up in the 1980s, I would like to briefly touch upon an extra ingredient in the mathematics–physics relationship – the remarkable role played by consistency. In the twentieth century it happened, more than once, that mathematics did not merely act as a codifying tool, but appeared to have predictive power over nature. Merely by requiring mathematical consistency, one was sometimes forced to believe in new, as yet undiscovered, physical phenomena. The phenomena in question were then later confirmed by experiment.

The classic example was Dirac's equation for the electron, which he postulated in 1928. It was soon realized that while the equation worked well for electrons, it also predicted the existence of another particle with the same mass as an electron but opposite electric charge. This prediction arose because for a charged particle of spin one-half, a Lorentz-invariant equation necessarily has at least two kinds of solutions. So it was mathematical consistency that required the new particle and no way could be found to get rid of it from the equation. Instead, in 1932 the new particle, the 'positron', was experimentally detected.

As is widely known, Dirac later commented: 'it is more important to have beauty in one's equations that to have them fit experiment', a statement that comes across as bold, controversial, even downright unscientific. What would Dirac, or anyone else, have done if the positron had not been discovered? Or was he saying that the positron had no choice but to be discovered?

The rest of Dirac's comment, which is less widely publicized, gives a better picture of what he had in mind. 'If there is no complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may well be due to minor features that are not properly taken into account and that will get cleared up with further development of the theory.' He was only saying that mathematics is a more reliable guide to nature than one might expect, but not that it is an ultimate arbiter of natural law.

In 1931, Dirac proposed the existence of magnetic monopoles from similar considerations of mathematical consistency. He argued that monopoles were consistent with all known facts about quantum mechanics, therefore they ought to exist. Fifty years later, during which period no monopole had been discovered, he retracted his proposal in the most abjectly empiricist language. He wrote to Abdus Salam in 1981: 'I am inclined now to believe that monopoles do not exist. So many years have gone by without any encouragement from the experimental side'.

Taken as a composite, the quotations above give a more accurate picture of the situation. Dirac's views about the role of mathematics were far from extremist. He simply believed that beauty in the equations was a more reliable guide than most people had hitherto believed. I take this as a slight strengthening of Wigner's observations on the effectiveness of mathematics in the natural sciences, but no more than that.

Mathematics and quantum field theory

The major influence in creating a different paradigm for the mathematics–physics relationship is undoubtedly Edward Witten at Princeton, though others, notably Witten's former student Cumrun Vafa at Harvard, have been extremely active at the forefront.

Despite being a physicist, Witten received the highest honour in mathematics, the Fields Medal, in 1990 for his contributions to mathematics using tools of physics. In the citation for the award, Ludwig Faddeev said that 'Physics was always a source of stimulus and inspiration for Mathematics... In classical time(s) its connection with mathematics was mostly via Analysis, in particular through Partial Differential Equations. However, (the) quantum era gradually brought a new life. Now Algebra, Geometry and Topology, Complex Analysis and Algebraic Geometry enter naturally into Mathematical Physics and get new insights from it.'

Putting it as simply as possible, Witten's strategy was to rephrase mathematical concepts in the language of Quantum Field Theory, a framework developed for the description of fundamental particles and their interactions. He would then use insights and tools native to quantum field theory, such as symmetries and invariances, and the (notoriously hard to define) Feynman–Kac path integral. This would lead to a 'natural setting' for the mathematics problem and often provide not only a solution of it, but an enormous variety of generalizations. The solutions that Witten provided were not rigorous, but in many cases were rendered rigorous after some additional work. In an article that was intended to be Witten's Fields citation but which he was unable to actually deliver at the award ceremony, Michael Atiyah notes 11 papers of Witten that justify his being awarded this prestigious prize. To give the reader a flavour of what was done, I will briefly touch upon a few of them.

Knot theory

The study of knots in three-dimensional space was initiated in the 19th century by the physicist Lord Kelvin. He had a theory of 'vortex atoms', according to which the differ-

ent atoms that occur in nature are made of the same 'substance' but knotted in different ways. According to this theory, the classification of atoms could be reduced to the classification of knots, and Kelvin accordingly embarked on the latter problem.

Over the ensuing decades, considerable progress on classifying knots was made by mathematicians. Two knots are distinct if they cannot be continuously deformed into each other. But for complicated configurations, it is hard to establish whether or not this is possible just by looking at pictures of the knots. Two pictures can look quite different from each other, but there might be some 'moves' that smoothly deform one of them to the other – then they would describe the same knot. The idea then emerged of associating an 'invariant' to a knot. This would be a number that is calculable starting from any picture of the knot, and that remains invariant under smooth deformations of the knot. Now if from two pictures we get two different invariants, then we can be sure the knots are distinct. The converse is not in general true, but a powerful knot invariant will be able to distinguish more and more subtly differing knots from each other.

Knot invariants were extensively developed and the most powerful ones (some of which were found by the mathematician Vaughn Jones) turned out to be polynomials in one or more parameters. Witten sought a relationship between these polynomials and a certain gauge field theory in three spatial dimensions called 'Chern
In this theory, the natural observable is a 'Wilson line' (a line of generalized electric flux). The flux can run along an arbitrary closed contour, the same thing as a knot. Being an observable, the Wilson line may have a quantum mechanical 'expectation value' in the vacuum of the theory. When computed using techniques of quantum field theory, this expectation value turned out to be a polynomial in some parameters related to the coupling constant of the theory and the representation (the type of charges) flowing around the Wilson line. In fact, the expectation value was none other than the Jones polynomial.

For knot theory, this was a dazzling illumination: it gave an interpretation to the polynomials, it explained (via a relation to conformal field theory) some of their properties and relations to the theory of quantum groups, it related the parameters in the knot polynomials to an expansion parameter in a perturbations series, and it provided an easy way to generalize the polynomial, by changing the gauge group of the gauge theory from SU(2) to an arbitrary Lie group. Moreover, as Atiyah points out, Witten's field-theory interpretation 'is the only intrinsically three-dimensional interpretation of the Jones invariants; all previous definitions employ a presentation of a knot by a plane diagram or by a braid.'

Now the Chern–Simons gauge theory at the root of this brilliant discovery is a cousin of the Yang–Mills theory, that describes three of the four fundamental interactions in nature. Yang–Mills theory, in turn, was a generalization of Maxwell's theory of electromagnetism, a relativistic field theory with a gauge symmetry. So without the whole circle of ideas and experiments relating to relativity and electromagnetism at the start of the twentieth century, we (more precisely, Witten) would probably not have found a new way to describe knot invariants, and a major mathematical illumination could never have taken place.

Morse theory

Morse theory studies the topology of differentiable manifolds. It relates two different quantities: on the one hand the critical points of some suitable function defined on them, and on the other hand their homology, or the kinds of non-trivial 'cycles' that can be embedded in them. One can gain insight into the topology of a manifold by knowing about its cycles and, via Morse theory, one gets this by studying critical points. Witten provided a proof of certain inequalities, the 'Morse inequalities' by setting up an equivalence between the homology of the manifold and a toy physical system called supersymmetric quantum mechanics.

For the lay person, this is harder to appreciate as compared to knot polynomials. But here too, one can see where the physics input came from. Starting around 1974, theoretical physicists made a breakthrough in their search for a symmetry that unifies bosons and fermions, the two different kinds of particles that occur in nature. The new symmetry was dubbed 'supersymmetry'. There are compelling reasons to believe that, although it is hidden from view in everyday life, nature makes use of it in some subtle way. So compelling are these reasons, that the Large Hadron Collider (LHC) in Geneva will start looking for evidence of supersymmetry from 2007 and the chances of finding it are considered rather high. Witten's supersymmetric quantum mechanics is a highly simplified version of the physical theories whose experimental confirmation is being sought at LHC, but uses the same basic structure. By mapping it onto the problem of Morse theory, Witten was able to use some properties of the supersymmetric path integral to make headway into pure mathematics. Again, without the experimental fact of bosons and fermions and the physicist's desire to unify them, this mathematical insight into Morse inequalities would not have been possible.

The Moduli spaces of Riemann surfaces

To a lay person, a Riemann surface is just a two dimensional surface, examples being a sphere, a torus (like a bicycle tyre) and multi-handled generalizations. They are more naturally described in the language of complex analysis. Mathematicians would like to know how these surfaces can be smoothly varied and in what way their mathematical properties change as we vary them. The 'moduli space' of such a surface is the space of parameters that label this variation. Sophisticated results due to Mumford and Morita

and others, had given considerable insight into the global structure (topology) of these moduli spaces.

Witten was inspired to address this by considering a problem in gravitation. However, his theory of gravity was not the one discovered by Einstein, but a simpler version of it where spacetime is two-dimensional. From a physical point of view the theory has very little dynamics in this setup, but Witten analysed the operators that made up the observables of this theory, and asked what were their correlation functions – which measure how quantum fluctuations are interlinked in a field theory. What he discovered was that these correlation functions were identical to the Mumford–Morita classes on the moduli space of Riemann surfaces. As usual, this approach also gave rise to important generalizations of the known mathematical results.

This time the physical inspiration came from gravity, and its theoretical structure as elucidated by Einstein in his theory of General Relativity. Though Witten's gravity theory was very different, it is clear that if General Relativity had remained undiscovered, or if there were simply no gravitational force in nature (and assuming we still existed), this insightful approach to moduli spaces would never have been found. And therefore some profound results about the abstract manifolds called Riemann surfaces would not have been known to mathematicians.

Atiyah once made a comment about Witten's contribution to mathematics on the following lines. Usually, physics provides an intuitive notion, or a way of thinking about things, and then mathematicians prove a corresponding rigorous result. But in the case of Witten's work, rigorous results in mathematics already existed and Witten then provided the intuitive explanation for them. In this sense, the mathematicians were ahead of the game, though to say this in no way undermines the contribution of Witten's work. As already indicated, his physical re-interpretations led to generalizations of the mathematics which would have been almost impossible to guess if one only knew the rigorous methods. And by providing a new setting for old ideas, they provided a number of pointers towards new mathematical directions. We will look at one of those directions now.

Mathematics and string theory

The subject of this section will be an area of research in which not only is physics in the process of influencing mathematics, but, for the moment at least, the physics approach seems to be ahead of the game.

We will not need to know much string theory for the purpose of this discussion. It is enough to mention here that string theory is the most serious candidate known for a consistent quantum theory of gravity. It has not received direct experimental confirmation, but it is built on many experimental facts: the number of dimensions we live in, the presence in nature of both fermions and bosons, and the existence of gauge symmetry as in Yang–Mills theory and of general coordinate invariance as in Einstein's theory of General Relativity.

Einstein taught us that gravity is the geometry of spacetime. But he was talking about classical gravity, not the quantum version, for the simple reason that in his lifetime, no quantum version was known. The classical field theory of gravity used a great deal of mathematics from the realm of Riemannian geometry. Spacetime is a (pseudo)-Riemannian manifold. A quantum theory should be a theory of fluctuating manifolds, consistent with the basic principles of quantum mechanics but such a theory proved very elusive in Einstein's own lifetime.

Even in Einstein's time, there were some theories of gravity that required not only a physically observable spacetime, but also an extra hidden space that was too small to observe experimentally. This hidden space affected our world only indirectly. The notion of such hidden spaces and their promising role in the unification of forces was proposed by Kaluza and Klein early in the twentieth century. The hidden space was to be a Riemannian manifold and the possibilities for what it could be were governed by standard geometric considerations, well-known to mathematicians.

Now in discovering string theory, it appears that we have finally discovered a quantum theory of gravity. Like the older Kaluza–Klein theories, string theory too requires a hidden space in addition to the observable spacetime. But since the new theory is a quantum theory, the observable and hidden spaces can no longer be thought of as purely classical objects. They need not correspond to conventional Riemannian geometry except in a limit where quantum effects are negligible. To what sort of geometry do they correspond in general?

The answer is not completely known, but there is by now considerable evidence that there is a vast geometrical structure, which we will call 'stringy quantum geometry' for lack of a better word, that generalizes the usual mathematical notions of geometry in an exceedingly strange way.

Here I will just highlight two (related) properties of the quantum geometry associated to string theory. One is called target–space duality and the other is called mirror symmetry.

Target–space duality

In conventional Riemannian geometry, one has the notion of a metric which defines the distance between two points. This is a very appealing notion to a physicist. It plays a key role in general relativity where the metric of spacetime is itself the fundamental dynamical variable. In conventional geometry, a distance can take any value from zero to infinity.

Now let us consider an internal spatial direction that is compactified into a circle. Classically this circle has a radius, given by the minimum distance one traverses in making a complete tour of the circle and returning to the starting point. In the limit of large radius, we would call the direction 'non-compact' and roughly that is what the three familiar spatial dimensions in the real world look like.

Suppose now that we consider a finite value of the radius, and then vary it so that it becomes smaller and smaller. In classical geometry this process never ends, and the circle simply continues to shrink. If we place a quantum mechanical particle on this space, it will have difficulty exciting itself into a mode that can propagate on the small circle. The reason is that a well-defined wave on a tiny circle needs to be rapidly varying and therefore must have a high wave number, or momentum. Therefore to probe this circle, the particle must have a high energy. This is precisely what we mean by saying that a compact internal dimension is physically unobservable when its radius is small.

Now replace the particle with a string. Physically it is clear that at low energies a string behaves very much like a point particle, and it will have the same difficulty entering the small internal space unless we give it a large energy. However, the string can do something else that also probes presence of this space. It can wind itself around this circular direction. This is a classical configuration and it is easy to see that it carries an energy proportional to the length of the direction. So the energy required to excite this 'stringy' mode actually decreases as the circle shrinks. Therefore the string has a spectrum of heavy momentum modes as well as light winding modes.

Now suppose we replace the circle of a given radius by one of the inverse radius (in appropriate units). In this way a small-radius circle gets mapped to a large-radius one. In conventional geometry this is a major change. But a string on this space again has a spectrum of light modes (but now they are the modes of a string propagating on the large circle) and a spectrum of heavy modes (in which the strings winds over the large circle). In other words, the spectrum of a string remains invariant on replacing a tiny direction with a huge one. Going beyond the intuitive picture presented here, inversion of the radius can be shown to be an exact symmetry in string theory. This symmetry goes by the name of target–space duality or 'T-duality'.

What is the consequence of this for mathematics? Geometry, the study of shapes of objects, clearly originated with something or someone being able to actually probe an object. That probe is the macroscopic-sized human being, or perhaps one of the point particles out of which we are made. But with the introduction of a new probe, the string, the observed geometry is very different. In this new geometry, there would be a minimum length scale and the geometry would be invariant under inversion of a compact direction.

It is too early to say what are the implications of this result within mathematics. Perhaps one day, schoolchildren will be taught that a circle of a given radius is the same as a circle of the inverse radius, even though pointlike objects would not be the appropriate probes to demonstrate this fact. Perhaps it will enter the lore in a different way. Whatever the case, stringy quantum geometry as illustrated by this example promises to be as new and as strange as Riemannian geometry was when compared to the older Euclidean geometry of flat space.

Mirror symmetry

In string theory we believe that the internal space hidden from our view is a 6-dimensional geometrical manifold. The number 6 arises because strings like to propagate in 10 dimensions, while we live in only 4. This means the remaining 6 dimensions must be invisible, so they are assumed to be compact and small. String theory requires them to be special manifolds called 'Calabi Yau' spaces, with a definite set of properties.

Now in the 1980s some groups of researchers discovered that such 6-dimensional manifolds come in pairs, with each member of the pair having little or no resemblance to the other one in the conventional geometric sense. However, strings propagate in precisely the same way on both members of the pair, and this is again known to be an exact symmetry of string theory. In other words, in the domain of stringy quantum geometry these manifolds would be completely indistinguishable from each other. They are called 'mirror

Here is a regularity that was totally unexpected at the outset. On encountering it, the first thing a physicist would do is to enquire if mathematicians understand this regularity. And here is the remarkable surprise: they do not.

Just like a small circle and a large circle, two Calabi–Yau spaces that form a mirror pair appear to be completely different from each other, but string theory recognizes them as being the same. In fact, the underlying feature of string theory responsible for this symmetry is the same in both cases, target–space duality and mirror symmetry. The latter is a more complicated manifestation of the same phenomenon.

In summary, we believe there is a new branch of mathematics, 'stringy quantum geometry', within which the symmetries discussed above are manifest. But today all that we know about this new branch of mathematics is what string theory tells us. As mentioned earlier, in this case the physics technique is ahead and the mathematicians have yet to catch up, though many are working hard on it at present.

I would guess that the newly two-sided relationship between mathematics and physics is likely to occupy a part of centre-stage in both fields for the next few decades.