BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 40, Number 1, Pages 21–29
S 0273-0979(02)00969-2
Article electronically published on October 9, 2002

PHYSICAL LAW AND THE QUEST FOR MATHEMATICAL UNDERSTANDING

EDWARD WITTEN

ABSTRACT. The theoretical physics of the first quarter of the twentieth century —centering around relativity theory and nonrelativistic quantum mechanics— has had a broad influence mathematically. The main achievement of theoretical physics in the following half-century was the development of quantum field theory or QFT. Yet the mathematical influence of QFT still belongs largely to the 21^{st} century, because its mathematical foundations are still not well-understood.

PHYSICS OF THE TWENTIETH CENTURY

I will begin by reviewing some of the discoveries in theoretical physics in the twentieth century.

In doing so, I will for brevity omit the experimental discoveries that in most instances set the stage. Moreover, I will concentrate on theoretical work that aims to discover the laws of nature, as opposed to work that aims to solve the equations in different situations and to understand the resulting phenomena.

I consider first what are widely seen as the three most central discoveries made up to 1925:

(1) Special Relativity—the theory that describes the behavior of objects moving at speeds not necessarily small compared to the speed of light, in situations in which gravitation can be neglected. Spacetime is described as a Minkowski space with the flat Lorentz signature metric $ds^2 = c^2\,dt^2 - d\vec{x}^2$.

(2) General Relativity—Einstein's extension of Special Relativity according to which gravitation is described by (pseudo)-Riemannian geometry with the Einstein equations that (in the absence of additional fields) read $R_{\mu\nu} = 0$.

(3) Quantum Mechanics—in which the concept of the trajectory of a particle is blurred, as the position x and momentum p become noncommuting operators obeying $[p, x] = -i\hbar$. Atomic and molecular physics and chemistry are described via the Schrödinger equation

$$i\hbar\frac{\partial\Psi}{\partial t} = H\Psi.$$

One thing that stands out about the discoveries on this list is that they are relatively familiar to mathematicians, and their mathematical interest has certainly

Received by the editors August 3, 2000, and, in revised form, February 21, 2002.
2000 *Mathematics Subject Classification.* Primary 51P05, 81T30.
This work was supported in part by NSF Grant PHY-9513835 and the Caltech Discovery Fund.

been felt. For example, the development of Riemannian geometry and of functional analysis was greatly intensified because of the relevance of these fields to the description of nature.

Now let us make a similar list of theoretical highlights of the half century after 1925:

(1′) The quantization, in the late 1920's, of Bose fields (and of the electromagnetic field in particular) and the resulting explanation of why light is made of quanta or "photons".

(2′) The quantization, also in the late 1920's, of Fermi fields, and the prediction, using the Dirac equation, of "antimatter".

(3′) Renormalization theory and the emergence (by 1950) of Quantum Electrodynamics as an ultra-precise theory for computing quantum relativistic corrections to the behavior of charged particles.

(4′) The explanation (in the period 1967-72) of the weak interactions via quantum non-abelian gauge theories (or Yang-Mills theories) with spontaneous symmetry breaking.

(5′) The description (completed in 1973-4) of the nuclear force via nonabelian gauge theory with "asymptotic freedom".

These developments plus others that I have omitted added up by 1975 to "the standard model of particle physics", in which the strong, weak, and electromagnetic interactions (but not gravity) are described in a common framework. All this extended our knowledge of nature and opened new horizons, at least as much as did the discoveries of the first quarter of the twentieth century. But I hope it is obvious from this list that the theoretical physics of the half-century after 1925 is much less familiar mathematically than that of the first quarter of the century.

I don't think that the main reason for this is that mathematicians have pursued abstract interests unrelated to physics. It is obvious, in fact, from the range of lectures at this meeting that mathematicians have maintained a lively interest in the natural sciences.

Rather, the difference is that the mathematical foundations of the theoretical discoveries in our second list are much more obscure. The theories of the first quarter of the century—Relativity and Quantum Mechanics—involved equations that were more or less rigorously defined from the start, though their physical application was perhaps surprising. By contrast (and with the aid of some hindsight), in the period 1925-75, the main theme was the development of "Quantum Field Theory" (QFT), in which the quantum concepts are applied to fields, like the electromagnetic field, and not just to particles, as in the more familiar case of the Schrödinger equation.

In quantum mechanics of particles, the position x and momentum p become noncommuting operators, and the quantum state can be represented by a wave-function $\Psi(x)$, bringing us into the realm of functional analysis. What happens instead if we quantize a field? In quantizing, say, the electromagnetic field, the components $\vec{E}(x)$ and $\vec{B}(x)$ of the electric and magnetic fields become noncommuting operators, so we have to consider an infinite-dimensional noncommutative algebra. This algebra can be represented on a Hilbert space consisting of wavefunctions $\Psi(\vec{B})$, so now we are one step "higher up"— the quantum state is a function on a function space (the function space being in this example the space of possible $\vec{B}$'s). Having to do functional analysis in a space with infinitely many variables brings a whole new level of analytical difficulty. To make sense of all this, at the level of precision used

in theoretical physics, the key concepts are renormalization theory and asymptotic freedom.

Being a much more difficult theory than Relativity or Quantum Mechanics, Quantum Field Theory has taken much longer to develop, and it is much harder to establish the mathematical foundations. Moreover the important constructions are subtle to describe and at first sight may tend to look rather specialized to many mathematicians. Rigorous models of QFT are hard to come by, and when available (as in Sinai's lecture at this conference, and in extensive developments in constructive field theory) they are far from what is needed to get in touch with elementary particle physics. Since the asymptotic freedom of quantum non-abelian gauge theory was discovered in 1973, we have known what conjectures one should aim to prove to give a proper mathematical foundation to the standard model of particle physics. I will say more about this later. But proofs have yet to appear.

Motivations for understanding quantum field theory mathematically

The gap that therefore still exists is, I think, the main reason that post-1925 theoretical physics is not better known mathematically. I can think of at least three reasons that mathematicians may wish to remedy this:

(A) Understanding natural science has been, historically, an important source of mathematical inspiration. So it is frustrating that, at the outset of the new century, the main framework used by physicists for describing the laws of nature is not accessible mathematically. The same point has been made for decades, since the start of axiomatic and constructive quantum field theory.

(B) Although QFT is not yet widely recognized as a mathematical subject, in the last 20 years or so, the QFT's that are important in physics have proved to have many geometrical applications that may be of interest even if one is not principally motivated by physics. Examples include applications to

* Donaldson theory of four-manifolds,
* Jones polynomial of knots and related three-manifold invariants,
* mirror symmetry,
* cohomology of moduli spaces,
* elliptic cohomology,
* $SL(2, \mathbf{Z})$ symmetry of characters of Kac-Moody algebras.

As the QFT's from which various geometrical deductions are made generally do not yet have a rigorous mathematical basis, the role of QFT has until now tended to be to motivate conjectures that are then proved, piecemeal, by independent methods, without direct reference to the "underlying" QFT's. This is the best that one can do now, but it clearly has a major drawback: one misses the insights that come from the QFT methods, and in particular one fails to see the unifying QFT origin of results in seemingly different areas of geometry.

To give just one example, in the last six or seven years, physicists have extensively studied nonlinear "dualities" of four-dimensional supersymmetric quantum gauge theories. These dualities generalize the Montonen-Olive duality between four-dimensional $N = 4$ supersymmetric Yang-Mills theory with gauge group G and the same theory with the Langlands dual group G'. To physicists, these dualities seem to be somewhat similar in spirit to mirror symmetry, but perhaps broader in scope.

Apart from exciting physical results such as new approaches to quark confinement, consequences of these dualities include the relation of Donaldson and Seiberg-Witten invariants of four-manifolds, predictions about L^2 cohomology of monopole

moduli spaces, and modular symmetry of the generating function of Euler characteristics of instanton moduli spaces. Probably one day the nonlinear dualities of the quantum gauge theories will be considered an interesting chapter in mathematics, but for now each spinoff is studied by separate methods.

(C) Finally, life—and theoretical physics—did not end when the standard model was completed by the mid-1970's. Physicists have continued to seek to understand QFT more deeply and to search for an even broader framework in which it would be possible to describe gravitation, as well as the other forces, in the light of relativistic quantum mechanics. String theory has emerged as a remarkably rich and exciting candidate, with many physical and mathematical hints that it is on the right track.

For many reasons, a knowledge of QFT is the basic prerequisite for learning about string theory. For one thing, the motion of a string is governed by a two-dimensional QFT. At perhaps an even more basic level, the aim of string theory is to generalize QFT (while incorporating gravity); to appreciate string theory, one must understand what it is that is being generalized.

The quest to understand string theory may well prove to be a central theme in physics of the twenty-first century. To understand this quest in mathematical terms and reap the full fruits, it will be necessary to develop QFT as a mathematical subject.

Four-dimensional quantum gauge theory

What mathematical problem best embodies the challenge of understanding quantum field theory?

We want a problem that is

* central in physics,
* important mathematically, and
* representative of the difficulties of QFT.

To me, the outstanding problem with these features is this one: *Prove the existence and mass gap of quantum Yang-Mills theory on* $\mathbf{R}^4$*, with gauge group a compact simple non-abelian Lie group* G.

The existence part would essentially mean making sense of the standard model of particle physics ("essentially" because the standard model includes fields other than the gauge fields, leading to some additional issues).

The mass gap, on the other hand, is a fundamental—though still poorly understood—statement about nature. It means that strongly interacting particles, such as the proton and pion, have positive masses and so travel at velocities less than the speed of light, even though, classically, Yang-Mills theory describes waves that propagate at the speed of light. (These waves obey a nonlinear Yang-Mills equation, somewhat analogous to the Einstein equations for gravitational waves and in contrast to the linear Maxwell equations for electromagnetic waves.) When nonabelian gauge theory was formulated by Yang and Mills in the 1950's, there certainly was interest in applying it to the strong interactions, but this seemed impossible because classical waves that travel at the speed of light would seemingly correspond in the quantum theory to massless particles, in contrast to what we see in the world of the strong interactions.

Ultimately physicists learned—largely from experiment, partly from computer simulations, partly from crude theoretical calculations making use of asymptotic freedom—that when four-dimensional Yang-Mills theory is quantized, there is a mass gap: the quantum particles have positive masses even though the classical

waves travel at the speed of light. The masses are very tiny; they vanish exponentially fast as Planck's constant $\hbar$ (or the dimensionless gauge coupling g) is taken to zero. This makes the mass gap difficult to understand theoretically.

What does the mass gap of four-dimensional quantum gauge theory mean from a mathematical point of view? Mathematically, it is natural to try to define quantum Yang-Mills theory on a general four-manifold; the mass gap is specifically a statement about what happens on $\mathbf{R}^4$. The mass gap, in fact, implies a principle of exponential decay of correlations at long distances that makes it possible to deduce global results about four-manifolds from a knowledge of how the theory behaves on $\mathbf{R}^4$. For example (in the supersymmetric case, with minimal or $N = 1$ supersymmetry), the mass gap is closely related to the behavior of the Donaldson invariants on algebraic surfaces.

A mathematical proof that quantum Yang-Mills theory exists in four dimensions would be a milestone in coming to grips mathematically with twentieth century theoretical physics. The reaction of physicists, however, would be that with the renormalization group and asymptotic freedom, one already understands why this theory exists, and that mathematicians would have merely succeeded in supplying the ϵ's and δ's. For the mass gap, it is different: a proof of the mass gap, should it appear now, would shed light on a fundamental aspect of nature that physicists still do not properly understand. This last statement, however, is not guaranteed for all time! An eventual rigorous proof of the mass gap might follow (or depend upon) much better heuristic explanations than are currently available. If such explanations emerge, that will certainly change how physicists think of the problem.

The existence and mass gap problem is natural for any compact non-abelian simple Lie group G. The question arises of whether the problem is easier for some class of G's. In fact, there are reasons to think that the simplest case is the case that $G = SU(n)$ (or $SO(n)$, or $Sp(n)$) for sufficiently large n. It is suspected that four-dimensional quantum gauge theory is equivalent to a string theory with $1/n$ as the string coupling constant. This equivalence would be a nonlinear "duality" somewhat analogous to those that I discussed earlier. If valid, it might give an effective way to demonstrate the mass gap (and other properties that I have not explained, such as quark confinement and chiral symmetry breaking) for sufficiently large n. Such a relation of gauge theory to string theory is an old idea (dating back to a suggestion by 't Hooft around 1973), and there has been dramatic progress in the last few years (centering around a conjecture by Maldacena for a case with maximal supersymmetry). This seems like much the most plausible known approach to the problem, but an answer along these lines is not yet in sight, even at a heuristic level.

Comments

To define quantum Yang-Mills theory, we are supposed to start with the Yang-Mills Lagrangian

$$L = \frac{1}{4g^2}\int_{M_4} \mathrm{Tr}\ F \wedge *F,$$

with M_4 a four-manifold, g a real number called the coupling constant, $F = dA + A \wedge A$ the curvature two-form of a connection A, and $*$ the Hodge duality operator. Then, one considers the Feynman path integral over the space $\mathcal{A}$ of connections.

The integral is formally

$$Z = \frac{1}{vol(\widehat{G})}\int_{\mathcal{A}} DA\, \exp(-L).$$

Here as $\mathcal{A}$ is an affine space, it formally has a translation-invariant measure DA (unique up to a multiplicative constant that will cancel out when we define correlation functions). The Feynman integral for gauge theories is usually formulated as an integral over $\mathcal{A}/\widehat{G}$, where $\widehat{G}$ is the group of gauge transformations; instead I have divided by the volume of $\widehat{G}$ in defining Z.

Next we pick points $x_i \in M_4$, and "local operators" $\mathcal{O}_i(x_i)$ that are gauge-invariant polynomials in the curvature F and its covariant derivatives at the point x_i. We set

$$Z_{\mathcal{O}} = \frac{1}{vol(\widehat{G})}\int_{\mathcal{A}} DA\, \exp(-L)\, \prod_{i=1}^{t}\mathcal{O}_i(x_i).$$

Finally, we define the "expectation values" or "correlation functions"

$$\langle\prod_i \mathcal{O}_i(x_i)\rangle = Z_{\mathcal{O}}/Z.$$

To prove existence of quantum Yang-Mills theory, one must make sense of these correlation functions (perhaps by making sense of the path integrals in the above heuristic definition) and show that they obey certain axioms that are related, among other things, to the fact that the $\mathcal{O}_i(x_i)$ can be interpreted as operator-valued distributions acting on a Hilbert space.

One approach (used in numerical simulations) to making sense of the path integral is to make a k-dimensional approximation to the space $\mathcal{A}$ of connections, for some integer k, and then take the limit as $k \to \infty$. For example, this can be done by triangulating M_4, making a discrete approximation to the gauge theory using the triangulation, and then taking a limit in which the triangulation is refined. Asymptotic freedom gives a precise recipe, supported by computer simulations, for how this limit should be taken. If the triangulation is determined by a cubic lattice with lattice spacing a, then as one takes $a \to 0$, one must adjust

$$g^2 \sim \frac{b_0}{|\ln a|},$$

where b_0 is a known constant that depends on G. Proving the existence of this limit would essentially establish the standard model as part of mathematics, as I have already noted.

Precisely in four dimensions, the classical Yang-Mills Lagrangian

$$\frac{1}{4g^2}\int \mathrm{Tr}\ F\wedge *F$$

is conformally invariant. Indeed, F is a Lie algebra valued two-form, the Hodge $*$ operator is conformally invariant precisely in the middle dimension, and a two-form is in the middle dimension precisely in dimension four. Conformal invariance is ruined by the lattice regularization and is not expected to return for $a \to 0$. So if we formulate the theory on a four-manifold M_4, say compact, with a metric γ, the results will not be invariant under a conformal transformation of the metric

$$\gamma \to e^{2\phi}\gamma,$$

where ϕ is a function on M_4. For $\phi \to -\infty$, or in other words as M_4 becomes small, asymptotic freedom gives asymptotic formulas for the correlation functions $\langle \prod_{i=1}^s \mathcal{O}_i(x_i)\rangle$. These formulas show that in the limit of a small four-manifold, the correlation functions differ only logarithmically from semiclassical formulas. By contrast, the behavior for $\phi \to +\infty$, or in other words as M_4 becomes large, is determined by whether the theory has a mass gap. The mass gap means that the limit of the correlation functions for $\phi \to +\infty$ is independent of the x_i as well as of γ. (The correlation functions in this limit are hence four-manifold invariants, but they are believed to be trivial; in three dimensions, the story is different, as I explain later.) As I have stressed, there are good reasons, based on experiment and computer simulation, to believe that the mass gap exists, but there is no satisfactory theoretical demonstration of it, even at a heuristic level.

The interpolation from $\phi = -\infty$ to $\phi = +\infty$ has many analogs in mathematics and physics. An example arises whenever one is trying to solve a well-defined time evolution equation (such as a PDE) for any system. Having a well-defined equation means that, given the initial data, one knows the behavior for short times. If one also knows the behavior for long times, one says that one knows how to solve the equations. Short time and long time correspond, respectively, to $\phi \to -\infty$ and $\phi \to +\infty$. Another example is the heat kernel proof of the Atiyah-Singer index theorem. Here one considers a trace $\mathrm{Tr}\,(-1)^F \exp(-\beta H)$, where β corresponds to e^ϕ. For β near zero (ϕ near $-\infty$) one gets a cohomological formula for the trace, while for β large (ϕ near $+\infty$), one gets an interpretation of this trace as the index of the Dirac operator. Finally, another example of this sort of interpolation arises in the study of topologically invariant correlation functions in twisted $N = 2$ supersymmetric Yang-Mills theory on a four-manifold. In that example, for $\phi \to -\infty$ one gets the Donaldson invariants, and for $\phi \to +\infty$ one gets the corresponding Seiberg-Witten invariants.

Practice problems

"Existence and mass gap" of four-dimensional quantum gauge theory is a worthy 21^{st} century challenge, but it is too hard for now.

Easier problems are needed, and given past experience, they should be problems chosen to minimize the (inevitably huge) analytical difficulties and maximize the geometrical interest. Successes in applying simplified quantum field theory models to problems that are of interest outside of QFT may be the key to attracting wider mathematical attention and bringing new energies and new methods to bear on the subject.

There are a variety of places that one can look for such practice problems. The examples that I will describe come from quantum gauge theory below four dimensions. One dimension is too easy, because quantum gauge theory in one dimension (which amounts to the problem of constructing the invariant subspace of a given representation of the gauge group G) is part of standard mathematics. This leaves us with two cases to consider, namely dimension two or three.

$D = 2$. Yang-Mills theory without additional fields is "trivial" in two dimensions. That is lucky, since if it is trivial, perhaps it will not be too hard!

Classically, the "triviality" can be seen from the Euler-Lagrange equation $d_A * F = 0$. As F is a two-form, $*F$ is a zero-form, and the equation simply states that the zero-form $*F$ is covariantly constant. This condition is so simple that, even

on an arbitrary compact two-manifold Σ, one can explicitly describe its general solution.

However, as Atiyah and Bott showed twenty years ago, this "trivial" classical theory is of mathematical significance. Indeed, they showed that the classical action

$$L = \frac{1}{4g^2}\int_\Sigma \mathrm{Tr}\, F\wedge *F$$

is a "perfect equivariant Morse function" for the action of the gauge group on the space of connections. They used this to determine the Betti numbers of the moduli space $\mathcal{M}_g$ of stable G-bundles on a Riemann surface of genus g. The perfectness of L as a Morse function is certainly linked to the fact that the classical theory is trivial and does not describe propagating nonlinear waves, as it would above two dimensions.

There is a quantum analog of this. The quantum theory is also "trivial". The path integral was essentially computed by A. Migdal in 1974. In a quantum version of the Atiyah-Bott result, this trivial path integral has been used to compute the intersection ring of $\mathcal{M}_g$.

Can this computation be justified mathematically? Such a result would surely have a considerable impact mathematically, even though by now the formulas describing the intersection ring have been proved by Jeffrey and Kirwan without using QFT.

$D = 3$. In three dimensions, on an oriented manifold M_3, we can consider a more general Lagrangian:

$$L = \frac{1}{4g^2}\int_{M_3} \mathrm{Tr}\ F\wedge *F - 2\pi i k I_{CS}(A),$$

where $I_{CS}(A)$ is the Chern-Simons invariant of the connection A:

$$I_{CS}(A) = \frac{1}{4\pi^2}\int_{M_3}\mathrm{Tr}\ \left(A\wedge dA + \frac{2}{3}A\wedge A\wedge A\right).$$

I_{CS} is well-defined as a map to $\mathbf{R}/\mathbf{Z}$, so for k an integer, $\exp(-L)$ is a well-defined complex number and we can attempt to define the quantum path integral.

For $k = 0$, on a flat three-torus or on $\mathbf{R}^3$ there are significant results in this direction (notably in work by Balaban). This should be completed and generalized to all k and to any compact three-manifold M_3.

A mass gap is expected for all k. I think that for the time being, this result will be out of reach, but a good goal for the relatively near future is to prove that the mass gap exists for sufficiently large $|k|$. This is much easier because the mass gap exists classically for $k\neq 0$, and the classical description is approximately valid for large k; hence any existence proof for the theory will be likely to entail a proof of the mass gap for sufficiently large k.

If this is accomplished, I recommend the following (undoubtedly hard) problem. Consider the theory on a compact M_3. (An important technical subtlety is that in defining the quantum theory, it is expected that one needs a choice of framing of M_3.) Let γ be a fixed metric on M_3 and ϕ a function; consider the theory with metric $e^{2\phi}\gamma$. Show that for $\phi\to\infty$, the path integral Z is *independent of* γ.

If the path integral is independent of γ, it is a (framed) three-manifold invariant. Indeed, Z should be the "quantum three-manifold invariant" of M_3 associated with

the group G, at

$$q = \exp(2\pi i/(k+h)).$$

Here h is the dual Coxeter number of G.

Many other examples could be cited in two or three dimensions of simplified QFT models of geometrical interest and relatively modest (but still very large) analytic complexity. For example, supersymmetric gauge theories in two dimensions are a very fertile area related to many questions involving mirror symmetry and cohomology of moduli spaces. Whatever examples are considered first, successes in deducing interesting geometrical theorems from simplified QFT models would hopefully attract wider mathematical interest and help turn QFT into a mathematical subject.

Department of Physics, Cal Tech, Pasadena, California 91125; and CIT-USC Center for Theoretical Physics, University of Southern California, Los Angeles, California 90089

Current address: School of Natural Sciences, Institute for Advanced Study, Olden Lane, Princeton, New Jersey 08540

E-mail address: witten@ias.edu