Asymptotic freedom

The force between quarks varies with the distance between them: The dynamics of the vacuum enhance the force at large distances, while at short distances the interaction grows weaker.

David J. Gross

Nuclei are very strongly bound aggregates of protons and neutrons. The nucleons in turn, we now believe, are extremely strongly bound aggregates of quarks: so strongly bound that we have never seen a free quark. The nuclear force is not a constant, however, but varies with the distance between the quarks. The dynamics of the vacuum enhance the force at large distances, while at short distances the interaction grows weaker. The notion that the force between quarks becomes vanishingly small as the quarks come close together, or, equivalently, that the quarks become free particles at very large energies, is called asymptotic freedom. I was very fortunate to be able to contribute to the discovery of asymptotic freedom, so I shall start with a few historical remarks, describing my own personal road to this discovery, and then discuss the current status and significance of asymptotic freedom.

As I was composing this review I was struck by an intense feeling of nostalgia for elementary-particle physics as it functioned in the 1960s, with close connections between experiment and theory. Today, we theorists long for experimental discoveries that would be as exciting, surprising and consequential as the discovery of scaling in deep inelastic scattering. However, it appears that the standard theory (usually called the "standard model," but by now it surely has earned the right to be called a theory) is consistent with all contemporary experimental data. Thus we are forced, as in the current attempts at grand unification, to adopt a new style of guesswork that is both riskier and less enjoyable.

David J. Gross is professor of physics at Princeton University. This article is adapted from a talk he gave at the April 1986 meeting of The American Physical Society in Washington, DC, on the occasion of his receiving the J. J. Sakurai Prize.

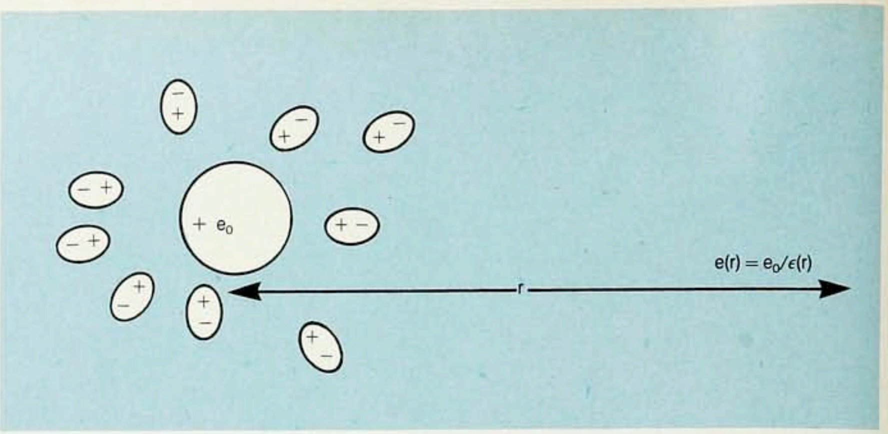
History

After graduating from Berkeley, in the fall of 1966 I went to Harvard as a junior fellow. This was the heyday of current algebra, and the air was buzzing with marvelous results. I was very impressed by the fact that one could assume a certain structure of current commutators and derive measurable results. Clearly the properties of these currents placed strong restrictions on hadronic dynamics. By this time most of the easy stuff had been done, and the implications of global current algebra were well understood as consequences of spontaneously broken chiral symmetry. Like others, I therefore studied the less understood properties of the algebra of local current densites. These were model dependent—but that was fine, because they might therefore contain dynamical information that went beyond statements of global symmetry. Furthermore, it was soon realized that one could check assumptions about the structure of local current algebra by deriving sum rules that could be tested in deep inelastic lepton–hadron scattering experiments.

In 1967 Curtis Callan and I proposed a sum rule to test the then popular "Sugawara model," a dynamical model of local currents. James D. Bjorken then noted that this sum rule, as well as dimensional arguments, would imply the scaling of deep inelastic scattering cross sections, that is, that they were homogeneous functions of only a single variable involving energy and momentum transfer. This prediction was shortly confirmed by experiments at the newly operational Stanford linear accelerator, which were to play an important role in elucidating the structure of hadrons. Soon thereafter Callan and I discovered that by measuring the ratio σ_L/σ_T (where σ_L and σ_T are the cross sections for the scattering of longitudinally and transversely polarized virtual photons) one could determine whether the constituents of hadrons had spin zero ($\sigma_T = 0$) or spin $1/2$ ($\sigma_L = 0$). The experiments quickly showed that σ_L is pretty nearly zero.

These experiments had a profound impact: They clearly showed that the proton behaved, when observed over short times, as if it was made out of pointlike objects (Bjorken and Richard

0031-9228 / 87 / 0100 39- 06 / \$01.00



Screening in QED. The virtual electron–positron pairs that surround a bare charge e_0 are polarized by the charge. As a result, the charge that is seen at a distance r from the charge is reduced from e_0 to $e_0/\epsilon(r)$. Correspondingly, at short distances there is less screening, and the effective charge increases. Figure 1

Feynman called them "partons"). Furthermore, these pointlike constituents had spin $1/2$ and (as later neutrino–proton scattering indicated) baryon number $1/3$—in other words they looked like quarks. From then on I was convinced of the reality of quarks, not just as the mnemonic devices for summarizing hadronic symmetries that they were then universally regarded to be, but as physical pointlike constituents of the nucleon. But how could that be? Surely there must exist some strong interaction among the quarks that would smear out their pointlike behavior. It soon became clear that, in a field-theoretic context, only a free, noninteracting theory could produce exact scaling. Once interactions were introduced, scaling, as well as my beloved sum rules, went down the tube. Yet the experiments indicated that scaling was in fine shape. But one could hardly turn off the interactions between the quarks or make them weak, as one would then expect hadrons to break up easily into their quark constituents. Why then had no one ever observed free quarks? This paradox and the search for an explanation of scaling were to preoccupy me for the following four years.

I remember well the 1970 Kiev conference on high-energy physics. There I met A. M. Polyakov, who was an uninvited but already impressive participant. Polyakov, A. A. Migdal and I had long discussions about deep inelastic scattering. Polyakov knew about the renormalization group and explained to me that naive scaling can't be right, because in field theory one expects to get anomalous dimensions. The reason is as follows: Renormalization changes the physics at different distance scales, which breaks scale invariance and changes the dimensions of operators with the scale of the physics being probed. Thus, dimensionless couplings change with scale, approaching at small distances fixed point values that are generically those of a strongly coupled theory. Such a theory might have scaling behavior, but it would be anomalous scaling behavior, quite different from the scaling behavior seen in the naive theory for pointlike particles. In fact, one would expect the cross sections to fall off with energy much more rapidly than naive dimensional arguments would suggest. I retorted that the experiments showed otherwise. He responded that this contradicts field theory. We departed; he convinced, as many were, that higher energies would change the experimental picture, I that the theory would have to be changed.

Renormalization

By the end of 1972 I had learned enough field theory, especially renormalization-group methods from Kenneth Wilson, to tackle the problem head on. I decided, quite deliberately, to prove that local field theory could not explain the experimental fact of scaling and thus was not an appropriate framework for the description of the strong interactions. The plan of attack was twofold. First I would prove that "ultraviolet stability," the vanishing of the effective coupling at short distances later called asymptotic freedom, was necessary to explain scaling; and second, that there existed no asymptotically free field theories. This was to be expected. After all, the paradigm of quantum field theory, quantum electrodynamics, was "infrared stable"—in other words, the effective charge grew larger at short distances—and no one had ever constructed a theory in which the opposite occurred.

Infrared stability, one aspect of charge renormalization, was understood by the early 1950s with the development of renormalization theory. Renormalization theory was always regarded with much suspicion as a way of sweeping infinities under the rug: Its role in eliminating unphysical divergences overshadowed its other features. Today we more fully recognize the physical reality of the variation with energy or distance scale of the strength and nature of fundamental interactions. Charge renormalization is nothing more (certainly in the case of QED) than vacuum polarization. The vacuum, or ground state of the universe, should be thought of as a medium containing virtual electron–positron pairs. If a charge e_0 is put into it, the medium becomes polarized. The resulting virtual electric dipoles screen the charge, so that the actual charge e observed at large distances differs from e_0; it is, in fact, e_0/ϵ, where ϵ is the dielectric constant (see figure 1). The dielectric constant depends on frequency (or energy, or distance), and thus one can introduce the notion of an effective coupling $e(r)$ that governs the strength of electrodynamic interactions at a distance r. As r increases there is more medium that screens, so $e(r)$ decreases with increasing r and, of course, increases with decreasing r. The function

$$\beta(r) = -\frac{\mathrm{d}\ln e(r)}{\mathrm{d}\ln r}$$

is therefore positive.

If the effective coupling were, contrary to QED, to decrease at short distances, one might explain how the strong interactions turn off at high energies and produce scaling. In fact one might suspect that this is the only way to get pointlike behavior at short distances. Indeed, by the spring of 1973 Callan and I had completed a proof of this argument (extending an idea of Giorgio Parisi's), and Sidney Coleman and I were close to a proof of the argument that all field theories behaved like QED. There was one hole in the line of argument, non-Abelian gauge theories, sometimes also called Yang–Mills theories. These, for technical reasons, could not be dealt with by the same methods. With Frank Wilczek, who had started his graduate work with me that year, we tried to close that last hole.

Our discovery that non-Abelian gauge theories—alone among four-dimensional field theories—were asymptotically free came in the spring of 1973. The discovery was made simultaneously by David Politzer, who was working with Coleman on a thesis problem that required knowledge of the β function. To me it was a total surprise. I didn't expect to find an asymptotically free theory—I was trying to prove that there were none. The discovery that one existed was almost sufficient, all by itself, to convince me that it had to provide the basis for the theory of the strong interactions. In addition, Wilczek and I realized immediately that color gauge theories of quarks could easily explain all of the then observed features of deep inelastic scattering (albeit with logarithmic corrections to scaling, which turned out to be a bonus because they provided an experimental test of the theory) and that the infrared growth of the coupling ("infrared slavery") might provide a mechanism for quark confinement. Rather than killing field theory, we had discovered a cornerstone of a unique field theory of the strong interactions, later dubbed quantum chromodynamics.

Why are non-Abelian gauge theories asymptotically free? The easiest way to understand this is by considering the magnetic-screening properties of the vacuum in electrodynamics. In a relativistic theory one can calculate the dielectric constant ϵ in terms of the magnetic permeability μ because their product is the speed of light. In classical physics all media are diamagnetic because classically all magnets arise from electric currents and the response of a system to an applied magnetic field is to set up currents that act to decrease the field (Lenz's law). Thus μ is less than 1 (in units in which the speed of light is 1), so that ϵ must be larger than 1, corresponding to electric screening. However, in quantum systems, paramagnetism is possible; in that case μ may be larger than 1, and ϵ less, resulting in an antiscreening effect. This is the case in non-Abelian gauge theories where the gluons are charged particles of spin 1. They behave as permanent color-magnetic dipoles, which align themselves parallel to an applied external field, thus increasing its magnitude and making μ larger than 1 (see figure 2). The antiscreening of the Yang–Mills vacuum can therefore be regarded as paramagnetism!

QCD is asymptotically free because the antiscreening of the gluons overcomes the screening due to the quarks. The arithmetic works as follows: The contribution to μ (in some units) of a particle of charge q is $-q^2/3$, corresponding to ordinary dielectric (or diamagnetic) screening. If the particle has spin s (and thus a permanent dipole moment proportional to γs, where γ is the usual gyromagnetic g factor) it contributes $\gamma^2 s(s+1)$ to μ. Thus a spin-1 gluon (with $\gamma = 2$, as in the Yang–Mills theory) gives a contribution to μ of

$$\begin{aligned}\delta\mu &= (-\tfrac{1}{3} + 2^2\times 1\times 2)q^2\\ &= \tfrac{22}{3}\, q^2\end{aligned}$$

whereas a spin-$\frac{1}{2}$ quark contributes

$$\begin{aligned}\delta\mu &= -(-\tfrac{1}{3} + 2^2\times\tfrac{1}{2}\times\tfrac{3}{2})q^2\\ &= -\tfrac{8}{3}\, q^2\end{aligned}$$

(the extra minus arises because quarks are fermions). Today this calculation is regarded as quite simple and even assigned as a homework problem in courses on quantum field theory. At the time it was not so easy. This change in attitude is the analog in theoretical physics of the familar phenomenon in experimental physics whereby yesterday's great discovery becomes today's background. In any case, the upshot is that as long as there aren't too many quarks the antiscreening due to the gluons wins out over the screening due to the quarks—an SU(3) gauge group such as QCD can accommodate as many as 16 triplets of quarks before their screening overcomes the gluon antiscreening.

Applications

The first and most important application of asymptotic freedom was the final stage of development of the color gauge theory of the strong interactions. Two important ingredients of this theory, the quarks and the suggestion of the color quantum number, had appeared years before. With the discovery of asymptotic freedom the dynamical role of color became evident and the argument for a particular Lagrangian became compelling. One could, as we did, argue for QCD from the deep inelastic experiments alone. They indicated that the flavor-charged constituents of hadrons were quarks (now called up, down, strange and so forth) and the glue that held them together was flavor neutral. Scaling meant that the glue must consist solely of Yang–Mills gluons.

With the discovery of asymptotic freedom the attitude of particle physicists toward the strong interactions changed, almost overnight. Before, there was a feeling that the strong interactions were an extremely hard and messy problem, where no early progress was likely—to quote Freeman Dyson in 1959, "The correct theory will not be found in 100 years." Suddenly there was a compelling theory that said that at short distances the coupling gets weaker, so that one could do, and trust, perturbative calculations. There is nothing that physicists like better than to calculate measurable quantities, so people started immediately to apply QCD to those problems where asymptotic freedom allows one to use perturbation theory.

Originally the applications were to deep inelastic scattering and to electron–positron annihilation experiments. In subsequent years the range of applicability of perturbative QCD has widened, and it continues to widen. The theory is now applied, and with apparent success, to a large number of high-energy experiments including the Drell–Yan process, large-transverse-momentum hadron scattering, the structure of heavy-quark bound states such as charmonium, jet production and many other processes. The scaling deviations predicted by the theory have been observed and the theory has passed many quantitative tests (although, given the tendency to logarithmic energy dependence, high-precision experimental tests are still lacking). A measure of the confidence in the validity of QCD is its widespread use by experimental physicists, a notoriously skeptical lot, as a tool for calculating the rate of background events. Recent experiments have shown no deviations from the background as calculated from perturbative QCD. This lack of new discoveries—although disappointing to experimenters—has provided some of the most striking confirmation of QCD.

Understanding the short-distance behavior of QCD has proved to be very important in other contexts. For example, it clears up an important point with respect to the weak interactions, namely why the strength of nonleptonic decays, in comparison with that of leptonic or semileptonic interactions, is not modified by the strong interactions. The answer, provided by asymptotic freedom, is that the strong interactions are actually weak at the distance scale of the weak interactions.

Asymptotic freedom also allows one to probe the behavior of a strongly interacting system at very high densities or temperatures. In this regime the mean energies are large, and the behavior of a hadronic medium can be described in terms of an effective coupling that turns off as the temperature, say, increases. This allows one to argue that QCD undergoes a phase transition at temperatures of order 200 MeV (10^{10} K), or at densities of order three to four times the nuclear density, to an unconfined phase—a quark–gluon plasma. Such temperatures or densities are of course unavailable in terrestrial laboratories, but might exist in the cores of neutron stars or in the early universe.

The flip side of asymptotic freedom, the decrease of the effective coupling at short distances, is the increase of the effective coupling at large distances, sometimes called "infrared slavery." It was this phenomenon that we originally invoked as providing a mechanism for quark confinement—the permanent binding of quarks in colorless hadronic bound states. This argument is rather naive, but it has turned out to be qualitatively correct. One can picture the QCD vacuum as a perfect paramagnet, with an infinite effective coupling, or magnetic permeability, at very large distances. Such a medium (the opposite of a perfect screening medium or a superconductor, which cannot tolerate magnetic fields) cannot admit color-electric fields and thus doesn't allow isolated quarks. One reasonably successful model of hadronic structure—the MIT bag model—is based on this idea. Real progress toward establishing confinement came, however, with Wilson's formulation of lattice gauge theory, which provided a qualitative picture of confinement (on a lattice) in the strong-coupling regime. To establish confinement for real quarks, one then had to show that this behavior persisted as the lattice spacing was shrunk to zero and, correspondingly, the coupling vanished according to asymptotic freedom. This approach has proved very fruitful. It has provided strong evidence for confinement and, in recent years, with the advent of new, powerful computers, the beginnings of quantitative calculation of the hadronic spectrum from first principles.

Consistency of field theory

Doubts as to the consistency of quantum field theory arose at the first signs of trouble in the quantization of electrodynamics—the ubiquitous ultraviolet divergences. Faced with this problem, the inventors of quantum field theory were quite willing to contemplate radical revisions of its basic principles to eliminate these divergences. Even the development of renormalization theory, although successful in expressing the predictions of QED in terms of finite physical parameters, appeared to be sweeping the problem under the rug, whence it would surely emerge at short distances. It is only with the discovery of asymptotic freedom that these concerns have been laid to rest and one is assured of the physical soundness of renormalization. Asymptotically free theories have eliminated most doubts as to the consistency of quantum field theory, have, in some cases, been rigorously proven to exist and have provided us with theories of the sort nature seems to prefer: theories that require no cutoffs, contain no infinities and do not break down at arbitrarily short distances.

The first problem that might concern us is that renormalization is a program that is formulated and carried out in perturbation theory. However, the true Hamiltonian of a local relativistic theory involves a coupling that describes the interaction at arbitrarily short distances—the so-called bare coupling. In theories that are not asymptotically free, the closer one gets to seeing the bare coupling, the larger the coupling appears to be, so a perturbative renormalization is suspect. The converse holds for asymptotically free theories, where the bare coupling vanishes and perturbation theory becomes better and better at short distances. This means that an asymptotically free theory does not really contain any meaningful divergences at all! The only way infinities appear is when one attempts to express a physical coupling, measured at some finite distance, in terms of the (infinitesimal) bare coupling, measured at infinitely small distances.

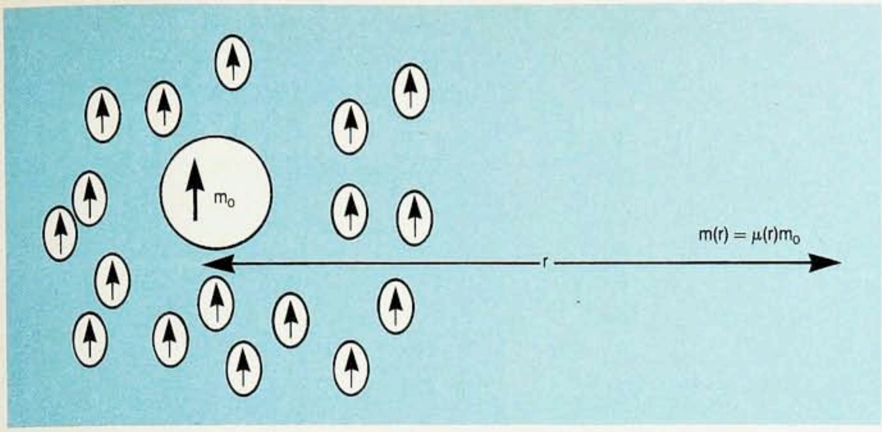
A more sophisticated criticism of quantum field theory was presented by Lev Landau in the early 1960s. He noted that because of screening the effective coupling e^2 of QED varies with momentum according to

$$e^2(p) = \frac{e^2(\Lambda)}{1 + e^2(\Lambda)\ln(\Lambda/p)}$$

where Λ is an ultraviolet cutoff, defined as the momentum at which the coupling equals the bare coupling:

$$e_{\text{bare}} = e(\Lambda)$$

Although the formula for the momentum variation is suspect (it only takes into account the coupling renormalization to lowest order), the concern it raises has proven to be qualitatively correct. The point is that the effective



Quantum chromodynamic screening. In QCD a magnetic moment m_0 causes virtual gluons (which also have magnetic moments) in the vacuum to align parallel to it. This increases the magnetic field observed at a distance r to $\mu(r)m_0$. Because the product of the dielectric constant ϵ and the permeability μ must equal the speed of light (1 in the units used here), an electric charge must be antiscreened—that is, the effective charge decreases as one approaches the charge. Figure 2

coupling, as given by the formula, blows up at some value of the momentum, namely

$$p = \Lambda \exp[-1/e^2(\Lambda)]$$

This pole in the effective coupling, known as the "Landau pole," is disastrous, producing unphysical singularities in measurable quantities. Another way of stating the problem is to note that if one takes $e_{\rm bare}$ to be finite (which would appear reasonable) then as Λ becomes infinite the physical coupling $e^2(p)$ must vanish for any finite p. This is the so-called zero-charge problem. It led Landau to conclude that "weak-coupling electrodynamics is a theory which is logically incomplete." In the case of electrodynamics the concern is only academic, because the effective coupling becomes strong only for very high energies, but if one were to try to construct non-asymptotically-free theories of the strong interactions the problem would show up for energies of a few GeV. Landau's critique had the effect of discouraging theorists in the Soviet Union from working on field theory (in particular as applied to the strong interactions) from the early 1960s until the discovery of asymptotic freedom.

Asymptotically free theories eliminate Landau's concern. Charge is now antiscreened, so that

$$e^2(p) = \frac{e^2(\Lambda)}{1 - e^2(\Lambda)\ln(\Lambda/p)}$$

This simple change of sign has dramatic effects.

▶ The effective coupling now blows up in the infrared regime, where it can be tamed by a variety of physical effects. For example, the Higgs mechanism, which, by breaking the gauge symmetry and giving the gauge bosons masses, can introduce an infrared cutoff. Another possibility is that the infinite growth of the effective coupling at large distances is real and leads to confinement—this, in fact, was the basis for our idea that infrared slavery confines quarks.

▶ The problem of zero physical charge is absent: The bare coupling is now zero and the physical coupling finite.

▶ Finally, there are some rigorous results to report. A few asymptotically free theories (simpler than QCD, of course) have been proved to exist and to be consistent; an essential feature of the proofs is the use of asymptotic freedom to control the ultraviolet behavior. Conversely, and in agreement with Landau's argument, a few non-asymptotically-free theories have been shown to be inconsistent for nonvanishing couplings.

Dimensional transmutation

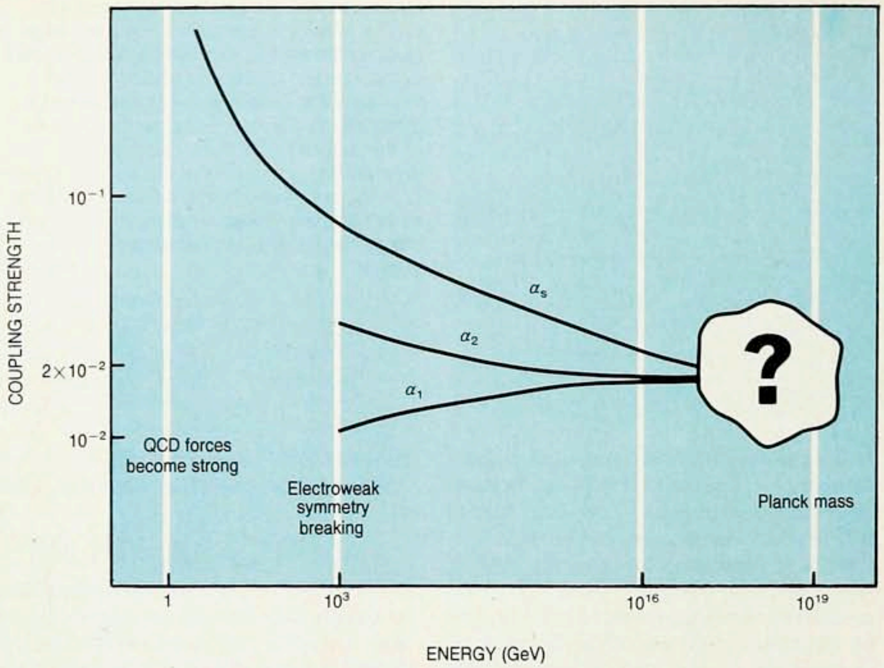
One of the remarkable and most appealing features of QCD is its high degree of uniqueness. If we take as given the SU(3) structure of the gauge group and the number and masses of the quarks, then the theory contains no arbitrary parameters. Arbitrary adjustable parameters in a fundamental theory are embarrassing. Many of us believe, as Einstein did, that "nature is so constituted that it is possible logically to lay down such strongly determined laws that within these laws only rationally, completely determined constants occur, not ones whose numerical value could be changed without destroying the theory." Unfortunately most theories have many free parameters. Of these, three parameters are always arbitrary, and they can be chosen to fix the units of length, time and mass. These are usually chosen to be the velocity of light c, Planck's constant h and some unit of mass or length. All other parameters can be expressed in terms of these and dimensionless constants. Thus QED is characterized by various mass ratios and by the fine-structure constant α, whose value is about $1/137$ and which plays the role of a dimensionless coupling constant. QCD is the first example of a physical theory in which the dimensionless coupling constant is not a free parameter. This possibility is realized because of asymptotic freedom and dimensional transmutation.

The Lagrangian that describes QCD is

$$L = \tfrac{1}{4}\,\mathrm{Tr}[F_{\mu,\nu}F^{\mu,\nu}] + \bar{\Psi}(i\partial\!\!\!/ + gA\!\!\!/)\Psi$$

The number of parameters is restricted by gauge and chiral symmetry and by the demand of renormalizability. (Note that we have set the masses of the quarks to zero. This is actually a good approximation for the up, down and strange quarks and is a consequence of chiral symmetry. The origin of the quark masses, in fact, has nothing to do with the strong interactions; rather it is a consequence of the spontaneous symmetry breaking of the electroweak interactions.) It would appear that the dimensionless gauge coupling g is a free adjustable parameter, just as the value of α (or e^2) is a free parameter in QED. This is an illusion. As I have noted, the strength of an interaction in a quantum field theory depends on the distance. But QCD contains no parameter with the dimensions of length or mass (in units where the velocity of light and Planck's constant are equal to unity, a mass m is equivalent to an inverse length, h/mc). Were it not for the dependence of the coupling on distance, QCD would be a scale-invariant theory. In a scale-invariant theory there is no way a physical length parameter could arise, because there is nothing it could depend on. However, the physics of antiscreening breaks the exact scaling and gives us a way of measuring distances, or producing physical masses. That is, because g depends on distance, we can trade this dimensionless parameter for a unit of length or mass—which we can then use to express physical quantities with these dimensions. This procedure is called "dimensional transmutation."

To be more precise, imagine introducing into the theory a length scale Λ that defines the coupling (much as one defines the electric charge in terms of the force between two identical charges at a given distance). That is, we define the coupling $g(\Lambda)$ as a function of this



Coupling strengths. The effective coupling constant α_s for the strong interactions decreases with increasing energy as $\ln(1/E)$. Extrapolation from low energies suggests that at an energy on the order of 10^{16}–10^{19} GeV (corresponding to distances of order 10^{-29}–10^{-32} cm) α_s becomes equal to the electroweak coupling strengths α_1 and α_2. At these energies, perhaps, all gauge interactions are unified. Figure 3

length scale. Now if the theory were actually scale invariant $g(\Lambda)$ would depend on Λ in a trivial fashion (it would be independent of Λ), but because renormalization breaks scale invariance, g depends nontrivially on Λ. However, Λ is totally arbitrary, so that any measurable, physical entity can depend only on a combination of $g(\Lambda)$ and Λ that is invariant under a change in Λ. Such a parameter with dimensions of mass is known as an invariant mass. It is given by

$$M(g,\Lambda) = \Lambda \exp\left[-\int^{g(\Lambda)} \mathrm{d}x/\beta(x)\right]$$

where $\beta(x)$ is the function that expresses the variation of $g(\Lambda)$ with Λ:

$$\beta(x) = \frac{\mathrm{d}\ln g(x)}{\mathrm{d}\ln x}$$

Any physical parameter P with dimension D_P (in units of mass) must be of the form

$$N[M(g,\Lambda)]^{D_P}$$

where N is a pure number that cannot depend on g (because the only dependence on g must be through M). We can then trade $M(g,\Lambda)$ for the mass M_p of the proton, and it follows that all physical parameters can be calculated in terms of h, c and M_p. Note that the feasibility of this program depends on the theory's being asymptotically free because the behavior of $M(g,\Lambda)$ depends, for small g, on the behavior of the β function. In fact, for g near zero,

$$M(g,\Lambda) \approx \exp(\pm 1/g^2)$$

depending on whether the theory is asymptotically free $(-)$ or not $(+)$. The behavior with the plus sign is absurd—it implies that a dynamically produced mass blows up as the interaction that produces it is turned off. It is no surprise that there are many examples of dimensional transmutation in asymptotically free theories, but none for infrared-stable theories.

Grand unification

Traditionally, when "fundamental" theories of nature have broken down at short distances, it has been a signal that there was new physics to be discovered once there were experimental instruments of high enough resolution (or energy) to explore at smaller scales or higher energies. In an asymptotically free theory this is not necessarily the case—the decrease of the effective coupling for large energies means that no new physics need arise at short distances, and indeed, were it not for the electroweak interactions and gravity, we might be satisfied with QCD as it stands.

The standard theory, which encompasses both the strong and the electroweak interactions, cries out for unification. Asymptotic freedom has played an important role in the search for a unified theory of all interactions. It explains how these separate interactions, whose strengths are so different at ordinary energies, can be thought of as emerging from a single unified theory at very high energies. The strong coupling decreases, according to asymptotic freedom, so as to equal the value of the electroweak couplings at the unification scale, and the apparent disparity between the strong and weak interactions disappears (see figure 3).

The variation of coupling with energy is logarithmic. Thus very high energies are needed to achieve unification—energies of 10^{16} GeV or higher. Such energies are remarkably close to the Planck mass M_P (about 10^{19} GeV), at which gravity becomes strong, which suggests gravity may play a role in the unified theory, as it does in the recent unified string theories. If the Planck mass is indeed the unification scale, then the logarithmic variation of the gauge couplings can explain a great mystery of nature, namely, why are we so big? The only truly fundamental length scale we know of is the ridiculously small Planck length

$$L_P \approx hG_N/c$$
$$\approx 10^{-33}\ \mathrm{cm}$$

formed from Newton's constant G_N, Planck's constant h and the velocity of light c. Most contemporary attempts to explain this hierarchy of scales, that is, the enormous ratio of the proton mass to the Planck mass,

$$M_p/M_P \approx 10^{-19}$$

are based on this logarithmic variation of the couplings. Perhaps we are so large compared with the fundamental length L_P because the effective coupling strength g that determines our size—by breaking the electroweak gauge symmetry and creating hadronic bound states—becomes significant only at energies of order $\exp(-1/g^2)$. □