10

Renormalization of Gauge Theories

GERARD 'T HOOFT

Born Den Helder, The Netherlands, 1946; Ph.D., 1972 (physics), University of Utrecht; Professor of Physics at the Institute for Theoretical Physics, University of Utrecht; high-energy physics (theory).

Like most other presentations by scientists in this Symposium, my account of the most important developments that led toward our present view of the fundamental interactions among elementary particles is a personal one, recounting discoveries I was just about to make when someone else beat me to it. But there is also something else I wish to emphasize. This is the dominant position reoccupied during the last two decades by theory, in its relation to experiment. In particular quantum field theory not only fully regained respectability but has become absolutely essential for understanding those basic facts now commonly known as the "Standard Model." So much happened here, so many discoveries were made, that the space allotted to theory in this volume runs far too short to cover it all. Therefore, I will limit myself only to the nicest goodies among the many interesting developments in theory, and of those I'll only pick the ones that were of direct importance to me.

Renormalization

Before the seventies there was only one renormalizable quantum field theory that seemed to give a reasonable and useful description of (parts of) the real world: quantum electrodynamics. Its remarkable successes in explaining, among others, the Lamb shift and the anomalous magnetic moment of the electron did not go unnoticed.[1] Yet the idea that other interactions should also be described in the context of renormalizable field theories became less and less popular. Indeed, the notion of renormalizability was quite controversial, and to some it still is.

The reason for this controversy is quite understandable: there are many misconceptions concerning the real meaning of renormalization in quantum field theories, and these are – partly – due to inaccurate pre-

sentations of the notion of renormalization, in particular the "infinite" renormalization apparently required in these constructs. A correct presentation would have to explain elaborately *why* theories are constructed the way they are, in a logically coherent way. Instead of that, however, it is often much more convenient to explain how renormalization works *in practice.*

In the latter case one is tempted to short-circuit the original delicate physical arguments. And then one gets some useful mathematical prescriptions, roughly to be summarized as follows:

> Start with the "naive" unrenormalized theory. You will see that it contains "infinities." Renormalization simply amounts to "subtracting" or "removing" the infinite terms.

Now this sounds like: "You hit upon difficulties; just ignore them, cover them up!" As if by miracle, the resulting prescriptions are now claimed to be completely unique and self-consistent. But of course the explanations as to why they work are then lacking, and many textbooks that contain only this version of the argument have added to the widespread mistrust and contempt for such an obviously shaky procedure, in spite of its experimental success, which, according to some, had to be accidental.[2]

Quite a few investigators tried to launch "infinity subtraction" as a first principle in renormalization. That renormalization turned theories with infinities into finite – hence useful – theories was used as a commercial that, in my opinion, did not betray deep insight concerning the real underlying physics.

To resolve this confusion one must realize that all known quantum field theories (in $3+1$ space–time dimensions) must be viewed as *models.* They do not pretend to describe any possible system of interacting particles with *infinite* accuracy, although some models allow us to make far more accurate predictions than others. The reason for this is that nothing in the model can be calculated with infinite precision.

All one has is some power expansion. An amplitude Γ will always be represented in terms of a series such as

$$\Gamma = a_o + a_1 g^2 + a_2 g^4 + \dots , \tag{10.1}$$

where g is some coupling constant. In all known realistic theories this series will be an asymptotic series at best, which means that there is no value for g small enough for the series to converge completely, apart from $g = 0$ (in which case the particles do not interact at all).

In practice the convergence question is often of little importance, that is, when we have g so small that the first few terms suffice. But if it comes to mathematical rigor, we have to state that mathematically these models are well-defined only if the coupling strength(s) g is (are) *infinitesimally* small.

Since in reality the coupling strengths are non-zero, we must admit that our quantum field theories must be viewed upon as *effective* field theories having a very accurate, but not infinitely accurate, predictive power. One must terminate the series when the next term becomes bigger than the previous. The value of that next term then roughly represents the error bar. This is mathematically acceptable if we simply replace the *field* of (real or complex) *numbers* by the field of *asymptotic series expansions.*

In our effective field theory we must assume that at a very tiny length scale $a = 1/\Lambda$ the basic interactions are not understood, but can be approximated by a simple model with a cutoff, for instance defined by a lattice, or by assuming the presence of unphysical particles as described by Wolfgang Pauli and F. Villars.[3] Now at this point we must replace all numbers by power-series expansions in terms of some expansion parameter z (for instance the coupling strength g^2), which tends to zero when all interactions vanish.

As a next step we express all quantities that can be directly observed in an experiment at low energy, hence large distance scale, in terms of z, and then we also replace z itself by an expansion parameter that can be observed at large distances (such as the physically observed electric charge of an electron). For instance, we will not consider the "bare" (i.e., original) mass, but only the physically observed mass (i.e., "renormalized mass") of a particle. One then discovers that for a certain class of models the limit $a \Rightarrow 0 \quad \Lambda \Rightarrow \infty$ exists, in the sense that all expansion coefficients of the asymptotic series remain finite. All artifacts due to the (lattice or Pauli–Villars) cutoff disappear in the limit. This class of models is called renormalizable.

The expansion coefficients for the *bare* mass and charge do not exist in the limit but may diverge, logarithmically in most cases. This means that for finite z one should not allow a to become much smaller than some exponential function like $\exp(-1/z)$, but in practice this is of little concern because it is far beyond the region where we expect the model to be physically reliable anyway. Thus the answer to many critical objections against the renormalization procedure is that the limits $z \Rightarrow 0$ and $a \Rightarrow 0$ *must* be taken in this order: z first, a last.

It is only when we streamline and short-circuit this long series of arguments in order to obtain a convenient manual for calculating the coefficients $a_o, a_1, \ldots$, that we find as a prescription that "infinities must be subtracted."

Actually one may consider five categories of sophistication for quantum field theories:

1. Nonrenormalizable field theories. If z is the expansion parameter representing the coupling strength, these theories allow us only to consider the lowest expansion term, for instance:

$$\Gamma \;=\; a_1 z \;+\; \mathcal{O}(z^2)\;. \tag{10.2}$$

   Examples are the old (but still quite useful) Fermi theory for the weak interactions,[4] and quantum gravity with quantized matter fields.
2. One-loop renormalizable field theories. In some theories such as Yang–Mills theory with mass term,[5] and pure quantum gravity without matter,[6] the existing symmetry allows us to compute unambiguously the next term but not more:

$$\Gamma \;=\; a_1 z \;+\; \;a_2 z^2 \;+\; \mathcal{O}(z^3)\;, \tag{10.3}$$

   where both a_1 and a_2 are unique and calculable.
3. Renormalizable theories. For these all expansion coefficients are uniquely defined and calculable, but the series are only asymptotic. Hence one has typically

$$\Gamma \;=\; \Sigma a_n z^n \;+\; \mathcal{O}(e^{-1/z}); \quad a_n \;=\; \mathcal{O}(n!). \tag{10.4}$$

4. Asymptotically free theories. These theories are also renormalizable, but have as an additional bonus that if we scale to very small distances the expansion parameter z approaches to zero, so that there the expansion (10.4) becomes extremely accurate. Consequently these theories are very accurately defined even if at large distances z is large. However, we still do not know whether these theories allow for *infinitely* precise calculations, although this is generally conjectured. Examples are pure non-Abelian gauge theories coupled to a limited number of fermion species only, such as quantum chromodynamics.
5. Borel summable theories.[7] These are theories that allow a rigorous definition of all amplitudes, typically obtaining

$$\Gamma(z) \;=\; \int_o^\infty B(u)\; e^{-u/z}\, du, \tag{10.5}$$

   where the power expansion of B in terms of u not only has a finite

radius of convergence but also allows for an analytic extension toward the entire real axis. Theories of this sort are not known in $3+1$ dimensions, apart from some special limiting cases.[8]

The early days of Yang–Mills theory

Just a few classical papers in the older literature stand out as real jewels, and they were inspiring examples of theoretical reasoning to all of us for many years. First let me mention the marvelous paper by Chen Ning Yang and Robert Mills.[9] They pointed out that the only interparticle force that was well understood at that time, QED, can be seen as a construction built upon a fundamental principle: local gauge invariance. And this principle can be generalized if we have more than one type of fermionic fields $\psi(x,t)$, which we can arrange as isovectors:

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}. \tag{10.6}$$

Consider transformations of the type

$$\psi \Rightarrow \Omega(x,t)\ \psi\ , \tag{10.7}$$

where Ω is a 2×2 (or possibly larger) matrix. One can then construct the *covariant derivative* $D_\mu\psi$ as follows:

$$D_\mu\psi = \partial_\mu\psi + g b^a_\mu T^a\psi, \tag{10.8}$$

which transforms just as (10.7) if the new fields b^a_μ transform in a very special way. Here g is just some coupling constant, and the matrices T^a are the generators of infinitesimal rotations. One can formulate dynamical equations of motion for the new fields b^a_μ by first defining the covariant fields

$$F^a_{\mu\nu} = \partial_\mu b^a_\nu - \partial_\nu b^a_\mu + g f^{abc}\, b^b_\mu\, b^c_\nu, \tag{10.9}$$

where f^{abc} are the structure constants of the Lie group of matrices Ω. The field equations are generated by the Lagrangian

$$\mathcal{L}^{inv} = -\frac{1}{4} F^a_{\mu\nu}F^a_{\mu\nu} - \bar\psi(\gamma_\mu D_\mu + m)\psi. \tag{10.10}$$

It is invariant under local gauge transformations and as such a direct generalization of QED.

Since the rigid, space–time independent analog of the transformation group (henceforth called the *global* group) was known as isospin invariance for the strong interactions, Yang and Mills viewed their theory as

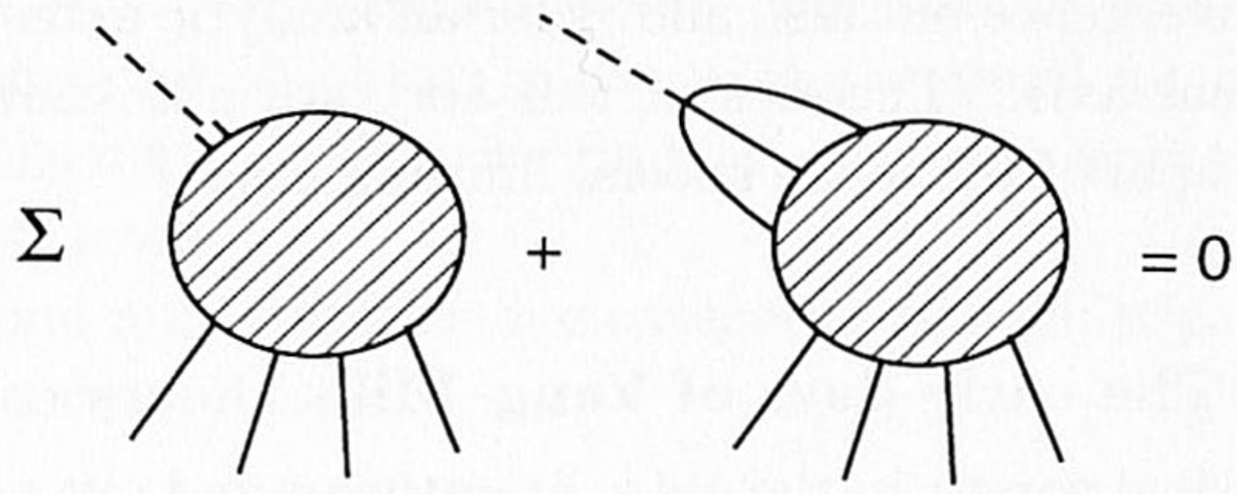

Fig. 10.1. Veltman–Ward identity among diagrams.

a scheme to turn isospin into a *local* symmetry, but they immediately recognized that then there was a problem: the Lagrangian describes a *massless* vector particle with three (or more) components, in general electrically charged as well as neutral ones. In spite of its beauty, this theory was therefore considered to be unrealistic. Besides, since these massless particles interact with each other, the theory showed horrible infrared divergences.

Proposals to cure this "disease" were made several times. Richard Feynman, who looked upon this model as a toy model for quantum gravity, proposed simply to add a small mass term just to avoid the infrared problem:[10]

$$\mathcal{L} \;=\; \mathcal{L}^{inv} \;-\; \frac{1}{2}\, M^2\, (b^a_\mu)^2. \qquad (10.11)$$

Sheldon Glashow and Martinus Veltman proposed to use the same Lagrangian as a model for the weak intermediate vector boson.[11] It was hoped that the mass term would not spoil the apparent renormalizability of the Lagrangian. Probably the philosophy here was that the mass term is only a mild symmetry-breaking correction of a kind we see more often in Nature: isospin invariance itself is also softly broken.

Indeed Veltman initially reported progress here: the theory (10.11) is renormalizable at the one-loop level.[12] He made use of field transformations that look like gauge transformations, even though the mass term in (10.11) is not gauge invariant:

$$b^{a'}_\mu = b^a_\mu \;+\; g f^{abc}\, \Lambda^b b^c_\mu \;-\; \partial_\mu \Lambda^a \quad ; \qquad \psi' \;=\; \psi \;+\; g \Lambda^a T^a \psi. \quad (10.12)$$

Here Λ may be any function of some arbitrarily chosen field variable. Veltman called this a "Bell–Treiman trasformation." The identities among amplitudes corresponding to different Feynman diagrams obtained in this way (see Fig. 10.1) should have been called *Veltman–Ward* identities.

To me it came as a surprise that Veltman managed to renormalize his theory up to one loop with this method. The mass term renders the longitudinal part of the gauge field observable, in spite of the fact that the Lagrangian carries no kinetic term for it. This theory should self-destruct. This it does, as Veltman found out, but only if you try to renormalize diagrams with two or more loops. To render the "massive Yang–Mills theory" renormalizable, a better theory was needed.

The Gell-Mann–Lévy sigma model

It was one of those caprices of fate that brought me, as a young student of Veltman's, to the 1970 Cargèse Summer Institute. (I had first applied to Les Houches, where my application was turned down.) The champions of renormalization were gathered there to discuss the Gell-Mann–Lévy sigma model, which had been proposed by Murray Gell-Mann and Maurice Lévy in 1960 in another classic jewel.[13] In order to explain the existence of a partially conserved axial-vector current, they added a fourth component to the three pion fields, the sigma field, transforming together as a 2×2 representation of chiral SU(2)⊗SU(2). The Lagrangian was

$$\begin{aligned}\mathcal{L}\left(\vec{\pi},\,\sigma,\,\psi,\,\bar{\psi}\right) \;=\; & \\ & -\frac{1}{2}\left[\partial_\mu\vec{\pi}^2+\partial_\mu\sigma^2\right]-\frac{1}{2}\mu_o^2\left[\vec{\pi}^2+\sigma^2\right]-\frac{1}{4}\lambda_o^2\left[\vec{\pi}^2+\sigma^2\right]^2 \\ & -\bar{\psi}\left[\gamma_\mu\partial_\mu+g_o\left(\sigma+i\gamma_5\vec{\pi}\cdot\vec{\tau}\right)\right]\psi \;+\; c\sigma. \qquad (10.13)\end{aligned}$$

If we take μ_o^2 here to be negative, then the potential for the scalar fields has the by now familiar dumbbell shape. The sigma field gets a vacuum expectation value,

$$\langle\sigma\rangle \;=\; F \;=\; |\mu_o|/\lambda_o, \qquad (10.14)$$

so in a perturbative expansion we write $\sigma = F + s$, and expand in s. The nucleon fields ψ get a mass g_oF, the pions have a tiny mass-squared proportional to the small constant c, whereas the sigma field s becomes a heavy resonance.

Jean-Loup Gervais, Benjamin Lee, and Kurt Symanzik explained in their Cargèse lectures how this model could be renormalized, and that its beautiful features would not be seriously affected by renormalization.[14] It was clear to me at that time that one can produce mass terms for Yang–Mills fields in a way very similar to this sigma model. I did not ask many questions in this school, but I did ask one question to Lee and

to Symanzik: "Do your methods also apply to the Yang–Mills case?" They both gave me the same answer: "If you are Veltman's student, you should ask him; I am not an expert in Yang–Mills theory."

Massless Yang–Mills

This I did, as soon as I was back in Utrecht. But Veltman replied that he found it difficult to believe in such a spontaneous symmetry breakdown in particle theory. His opinion was that if that happens the vacuum would have a tremendously large energy density, which would give the physical vacuum an enormously large cosmological constant.

But we know it happens in the sigma model, which describes strong interactions pretty nicely. And if it is not symmetry breaking, then at least all other vacuum fluctuation effects also contribute to the cosmological constant, not as much as in a weak interaction theory with Higgs mechanism, but still far more than the experimental upper bound. The cosmological constant problem should be postponed until we solve quantum gravity; we should not let it affect our theories at the GeV or TeV scale.

It was then that we decided what my research program would be. First I would try to really understand all details of the massless, unbroken Yang–Mills system, and then I would add the mass, by a "spontaneous local symmetry-breaking mechanism."[15]

The status of pure Yang–Mills theory was somewhat vague. Strong *formal* arguments existed that this theory had to be renormalizable. But there were competing and conflicting ideas as to what its Feynman rules were. One paper on this subject was my third classical gem: a short *Physics Letters* paper by Ludwig Faddeev and Victor Popov.[16] It was all I needed to understand what was going on. Faddeev and Popov argued that a gauge-invariant functional integral expression for the amplitudes had to have the form

$$\Gamma \;=\; \int e^{i\int \mathcal{L}^{\rm inv}(B)d^4x} \prod_x dB(x), \qquad (10.15)$$

where $B(x)$ stands for all field components of the gauge and matter system. However, since the integrand is invariant under gauge transformations, one only needs to integrate over the inequivalent field configurations, each being constrained by some gauge condition. As a gauge

condition one typically takes

$$\partial_\mu b^a_\mu \;=\; 0. \tag{10.16}$$

If we impose this constraint on the integrand, however, we need a Jacobian factor. So if we keep track of the measure, this turns the integral into

$$\Gamma \;=\; C\int e^{i\int \mathcal{L}^{inv}(B)d^4x}\prod_x \left(dB(x)\;\delta(\partial_\mu b^a_\mu)\right)\det\left(\frac{\delta\;\partial_\mu b^a_\mu}{\delta\Lambda}\right). \tag{10.17}$$

The theory produces a transverse propagator:

$$\frac{\delta_{\mu\nu}\;-\;\frac{k_\mu k_\nu}{k^2-i\epsilon}}{k^2-i\epsilon}. \tag{10.18}$$

Other theories led to a Feynman gauge propagator,

$$\frac{\delta_{\mu\nu}}{k^2-i\epsilon}, \tag{10.19}$$

and how this could be related to a functional integral was not clear.[17] More important, I thought, was that none of the existing papers provided for a precise prescription as to how the infinities should be subtracted. The *formal* arguments were there, but how does it work in practice?

This became the subject of my first publication.[18] Several things had to be done. First, the formalism to obtain the Feynman rules from the functional integrals could be simplified. The existing procedure to deduce the ghost Feynman rules from the determinant was not satisfactory. I observed that one can write

$$(\det\mathcal{M})^{-N} \;=\; C\int \mathcal{D}\vec{\phi}\mathcal{D}\vec{\phi}^*\;e^{-\vec{\phi}^*\mathcal{M}\vec{\phi}}, \tag{10.20}$$

where $\vec{\phi}$ is a complex Lorentz-scalar field with N components. One now reads off directly the Feynman rules for closed loops of ϕ fields. A factor N goes with each closed loop. Since we want N to be -1, our closed loops will usually go with a factor -1, just like the rules for fermions. Indeed, one can also write

$$\det\mathcal{M} \;=\; C\int \mathcal{D}\eta\mathcal{D}\bar{\eta}\;e^{-\bar{\eta}\mathcal{M}\eta}, \tag{10.21}$$

where η is an anticommuting (Grassmann) variable.

Next, I could also see how Faddeev and Popov's trick could produce the Feynman gauge. Just take an auxiliary field variable F and impose the gauge

$$\partial_\mu b^a_\mu \;= F^a. \tag{10.22}$$

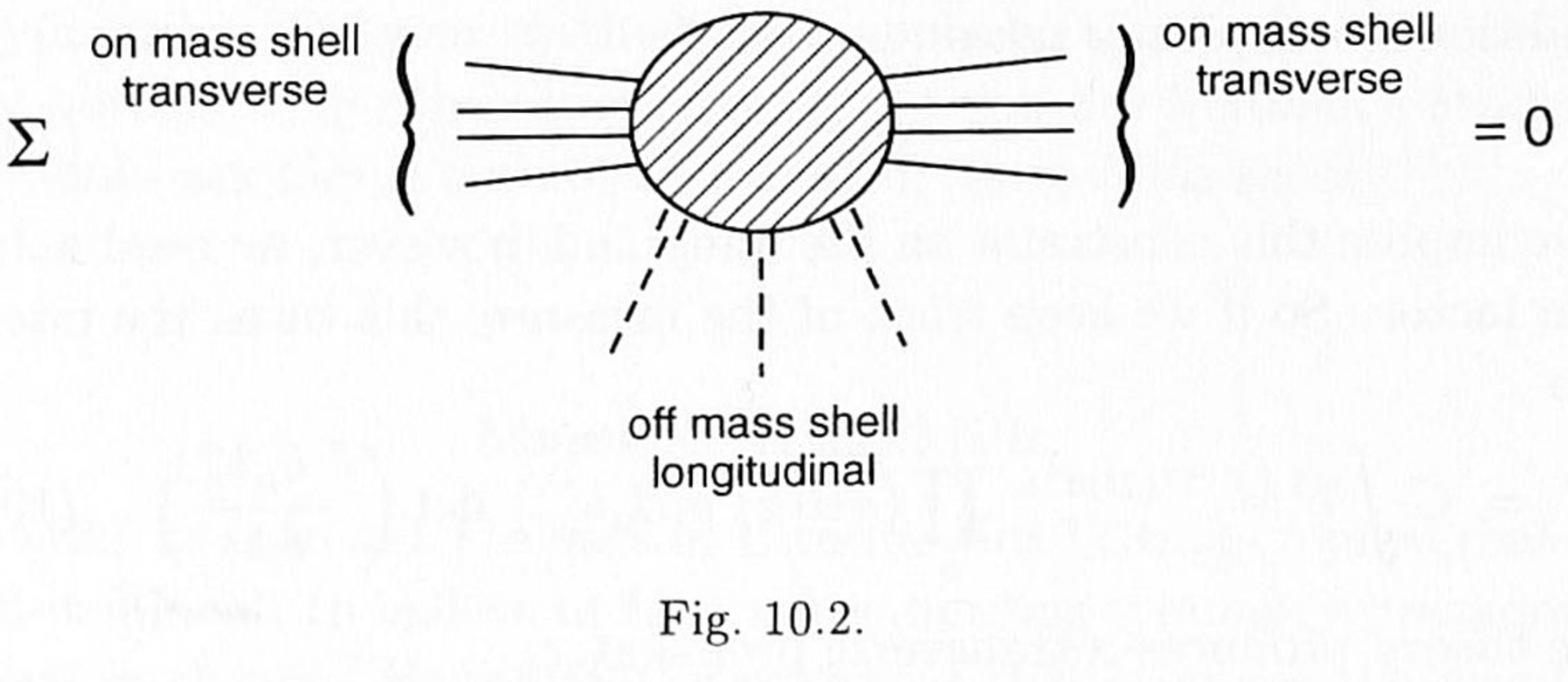

Fig. 10.2.

One then sees that

$$e^{-\frac{1}{2}\left(\partial_\mu b^a_\mu\right)^2} = \int \mathcal{D}F\; e^{-\frac{1}{2}F^2}\,\delta\left(\partial_\mu b^a_\mu - F^a\right). \tag{10.23}$$

To see that the renormalization counterterms do not spoil gauge invariance, we needed Ward identities. It turned out to be sufficient to prove identities of the form of Fig. 10.2.

Since reducible and irreducible diagrams must all be added together, these identities are sufficient to restrict all counterterms completely up to gauge-invariant ones. This point was often not realized by later investigators. The proof of these Ward identities was much more complicated than the Veltman–Ward identities mentioned before, because we had to disentangle carefully the contributions of various ghost lines. See Fig. 10.3, which was an intermediate step.

I was annoyed that I could not use a simple symmetry argument for the proof as Veltman had done for his case. Only much later it was discovered how to do this. Becchi, Rouet, and Stora found that the underlying symmetry for this identity is an *anticommuting* one.[19] Their marvelous discovery was this. Take as an invariant Lagrangian, for instance

$$\mathcal{L}^{inv} = -\frac{1}{4}F^a_{\mu\nu}F^a_{\mu\nu} - D_\mu\phi^* D_\mu\phi - V\left(\phi,\phi^*\right) - \bar{\psi}\left(\gamma D + m\right)\psi + \ldots \tag{10.24}$$

and add as a gauge-fixing term

$$\mathcal{L}^{gauge} = i\frac{1}{2}\left(\ell^a\right)^2, \tag{10.25}$$

where ℓ^a is anything like $\partial_\mu b^a_\mu$, b^a_4, and so on. Introduce the ghost fields η^a and $\bar{\eta}^a$, which must be anticommuting. Consider then the

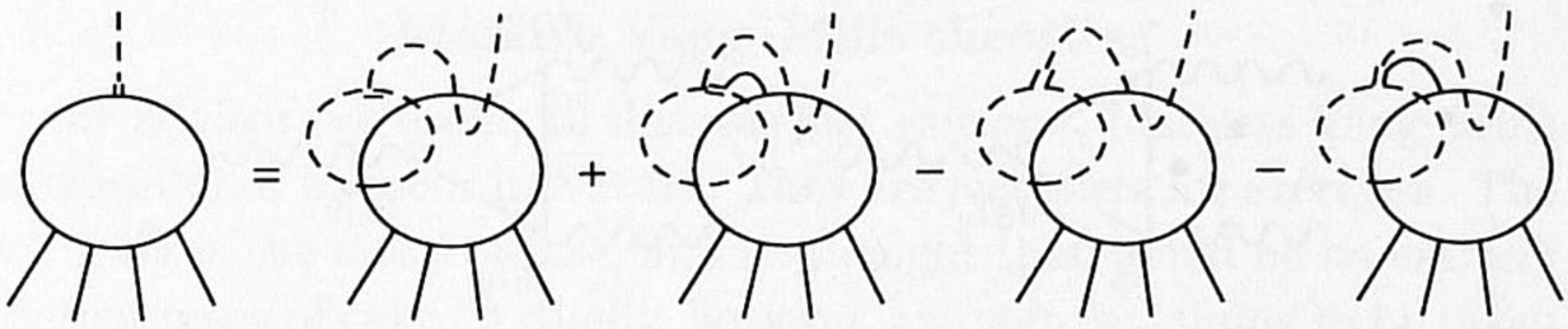

Fig. 10.3.

anticommuting variations:

$$\begin{aligned} \delta b^a_\mu &= D_\mu \eta^a; \\ \delta \phi &= -igT^a \eta^a \phi; \\ \delta \eta^a &= \tfrac{1}{2} g f^{abc} \eta^b \eta^c; \\ \delta \bar{\eta}^a &= \ell^a(b, \phi, \ldots). \end{aligned} \tag{10.26}$$

Here the first two equations are just gauge transformations. Then the total Lagrangian of the theory when taken to be

$$\mathcal{L} = \mathcal{L}^{inv} + \mathcal{L}^{gauge} + \mathcal{L}^{ghost}, \tag{10.27}$$

with

$$\mathcal{L}^{ghost} = -\bar{\eta}^a \, \delta \ell^a(b, \phi, \ldots, \eta), \tag{10.28}$$

is invariant under this *global* transformation. The above identities are nothing but an expression of this invariance, now called BRS invariance.

I had to convince myself that the rules obtained produced a *unitary* theory. The new identities were sufficient to guarantee this. Just one problem remained: the identities *overdetermined* the renormalization counterterms. Would there never be a conflict? There was a well-known example of just such a conflict in the literature: the Adler–Bell–Jackiw anomaly. Steve Adler, and independently from him John Bell and Roman Jackiw, had discovered that diagrams of the kind depicted in Fig. 10.4 cannot be renormalized in such a way that both the vector current and the axial-vector current are conserved.[20] If something like this would happen in a gauge theory, there would be deep trouble. I could prove that if no gauge fields are coupled to the axial charge, clashes of this sort will not destroy renormalizability in diagrams with up to one loop. The trick was to use a fifth dimension for the internal lines inside the loop.

What if you have more than one loop? I tried to use six, seven, or more dimensions but this does not work. (Recently a book appeared in which a "proof" of renormalizability along these lines appeared. The

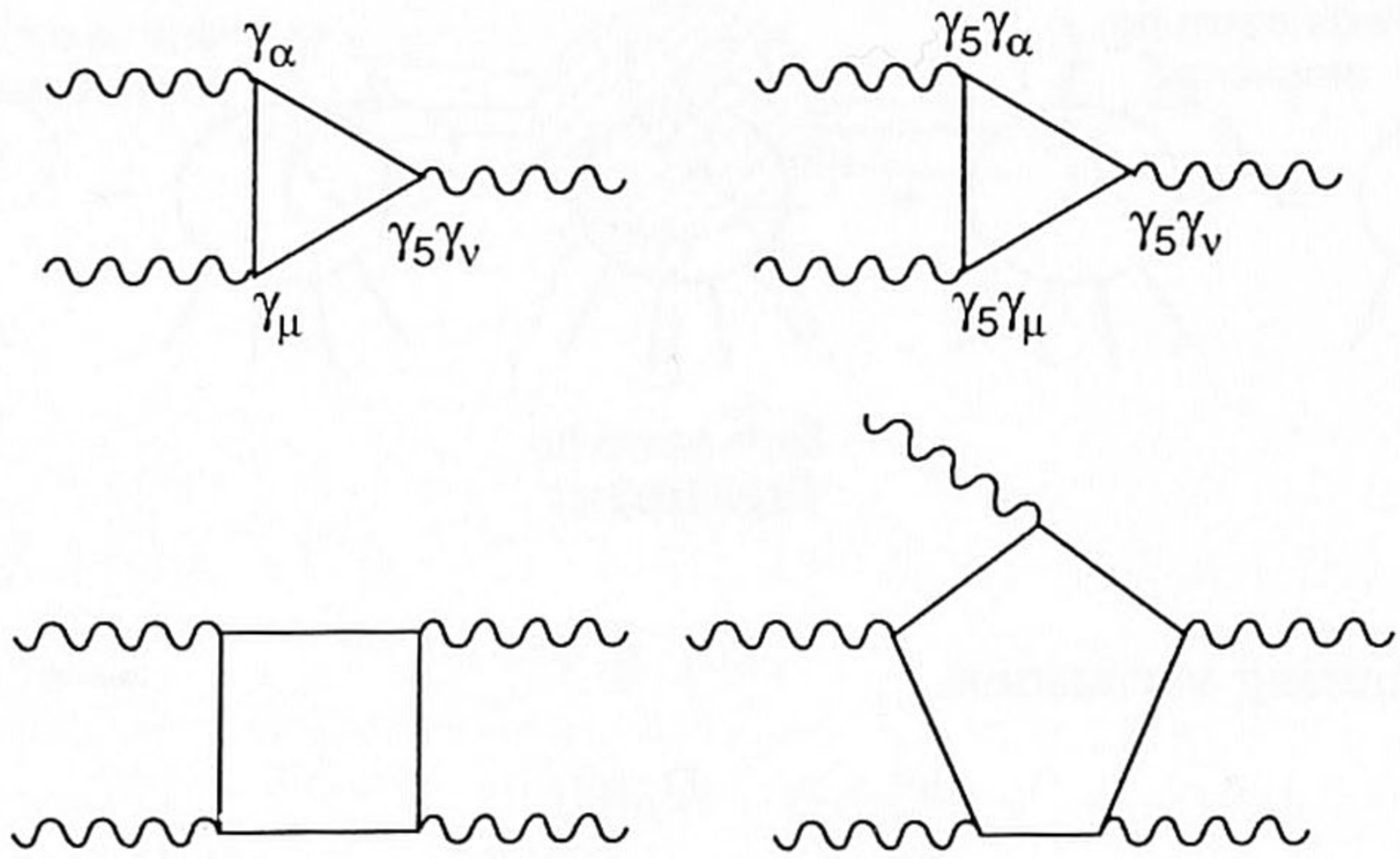

Fig. 10.4. Diagrams that contribute to the Adler–Bell–Jackiw anomaly.

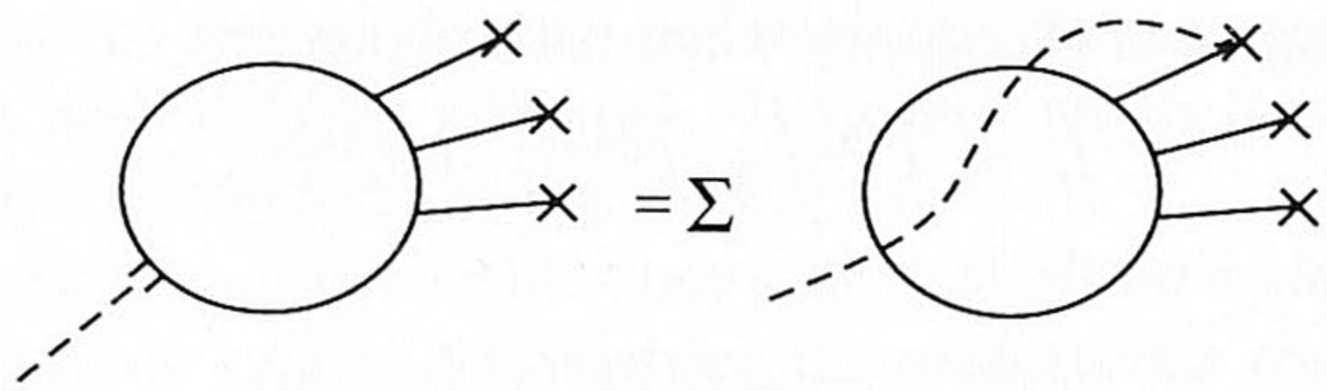

Fig. 10.5.

proof is incorrect.) I was confident the problem could be solved, but was unable to do it then.

Soon after my paper had come out, two other papers appeared, one by Andrei Slavnov and one by John Taylor.[21] Both observed that the identities I had written down could be generalized. If some of the external lines are neither longitudinal nor on mass shell, one gets extra contributions where the ghost line ends up at one of these lines. See Fig. 10.5.

The derivation went the same way as that for my own identities. The only reason why I had not written these identities in this new form before was that I thought the extra pieces would be cumbersome, requiring new renormalization counterterms of their own, and, furthermore, I didn't need them. It is clear now that these newer identities are more complete. And so it happened that they were to become known as the Slavnov–Taylor identities.

Massive Yang–Mills theories

For my advisor, Veltman, all this was just *spielerei*. Massless Yang–Mills fields seem not to occur in Nature. They are just there for exercises. The real thing is the massive case, and he thought that would be an entirely different piece of cake. Actually, however, the step remaining to be taken was a small one.[22] As I knew from Cargèse, the actual nature of the vacuum state has little effect upon renormalization counterterms. All that needed to be done was to add to the gauge-invariant Lagrangian the by-now familiar Higgs terms,

$$\mathcal{L}^{Higgs} \;=\; -\frac{1}{2}\left(D_\mu\phi\right)^2 \;-\; V(\phi), \tag{10.29}$$

where $V(\phi)$ has the familiar dumbbell shape, just as in the Gell-Mann–Lévy sigma model (for simplicity I take the ϕ field here to be a real multiplet). Writing $\phi = F + \varphi$ we get a gauge-invariant Lagrangian for the b and φ fields, such that now the b fields get the required mass. To appease Veltman, I wrote the self-interaction as

$$V = \;\frac{1}{8}\lambda\left(2F\varphi + \varphi^2\right)^2, \tag{10.30}$$

so that at least at lowest order the vacuum energy density vanishes. In terms of the φ fields, the gauge-transformation laws for these and the b field look very similar:

$$\begin{aligned} b_\mu^{a'} &= b_\mu^a + f^{abc}\Lambda^b b_\mu^c - \tfrac{1}{g}\,\partial_\mu\Lambda^a; \\ \varphi' &= \varphi + T^a\Lambda^a\varphi + T^a\Lambda^a F. \end{aligned} \tag{10.31}$$

Everything else went exactly as in the previous paper. Because the local gauge invariance is still exact, we again have Slavnov–Taylor identities and BRS invariance, and from them one can prove unitarity and equivalence of the various gauge choices. A judicious gauge choice was found such that the propagators for the massive gauge fields and the other fields became as simple as possible.

The problem of regularizing and renormalizing diagrams with two or more loops was still there. Veltman and I discussed a lot about this problem and eventually agreed that the best strategy was continuous variation of the number of space–time dimensions.[23] As if it were a seed from outer space, this idea germinated simultaneously in various places as an answer to different problems. Kenneth Wilson and Michael Fisher were writing a paper proposing to calculate critical phenomena in statistical physics in $4 - \epsilon$ dimensions as an expansion in ϵ.[24] And

independently of us C. Bollini and J. Giambiagi, and J. Ashmore, also suggested to use analyticity in space–time dimensions as a regulator.[25]

One may notice that by now I entirely address the problem of renormalization as a procedure for infinity subtraction. As explained at the beginning, this is not at all what renormalization really is from a physical point of view. It is preferable to talk about *regularization* first, and then *renormalization* afterwards. Regularization is the replacement of a theory by a slightly mutilated theory, using a cutoff. We must show that the effects of the cutoff become negligible at large distance scales, and then make the transition toward renormalized observables, after which it must be demonstrated that the limit where the cutoff goes away exists and is perturbatively finite. It does not matter much how crazy the mutilation was in the beginning, as long as the limit is well behaved. Going to $4 - \epsilon$ dimensions is just such a crazy regularization scheme. It turns out to be extremely elegant technically. Anyway, the important thing was that this method works fine at all orders of perturbation expansion and not just up to one loop, like the five-dimensional procedure found earlier.

We now had a general scheme for producing theories with interacting massive vector particles.[26] At first I was thinking about applying it to ρ mesons, as a nice generalization of the Gell-Mann–Lévy sigma model. But of course Veltman could convince me that the weak interactions were a much more promising application. I had practically reproduced Weinberg's model before I saw his 1967 paper. I also reproduced an error (the neutral cross section calculated by Weinberg was much too large because of a sign error in the Fierz transformation), but managed to correct it before my paper was published.[27] My own paper said in a footnote that the anomalies do not render the theory nonrenormalizable. Of course, this should be interpreted as saying that renormalizability can be restored by adding an appropriate amount of various kinds of fermions (quarks), but I admit that I also thought that perhaps this was not even necessary.[28] Now we know it certainly is necessary to have the anomalies cancel (see below).

More important to my mind was that we now had a large class of renormalizable theories with massive and massless vector mesons. A crucial argument was added to this by Chris Llewellyn-Smith and by John Cornwall, David Levin, and George Tiktopoulos; they showed that requiring unitarity implies that the *only* such theories are gauge theories.[29] So not only do we have a large class of new models, we have the *complete* class of renormalizable vector theories.

Asymptotic freedom

While searching for a decent regularization method for gauge theories, I had also studied scaling behavior; in 1971 I already knew that when you scale all momenta by a common factor upwards, then the gauge coupling constant decreases, whereas for QED the coupling strength increases. (I still had an error in the coefficient, now known as the β function. I found the right coefficient in 1972.)

I did dream about the possibility of pure gauge theories for quarks, but since I thought that strong interactions were infinitely complicated, I did not dare to ruin my reputation by launching such crackpot ideas. After the work with Veltman in 1972 in which we had carefully exhibited all detailed properties of the renormalization counterterms, I knew how to do the scaling calculation precisely.[30] Veltman convinced me, however, that our work on the counterterms for quantum gravity was much more important.[31]

In a Marseille conference in 1972 I met Symanzik again.[32] He explained his attempts to construct a theory with a negative β coefficient in order to explain Bjorken scaling.[33] I was delighted to announce after Symanzik's talk that what he was looking for was a non-Abelian gauge theory, and wrote its β function, in modern notation:

$$\beta\left(g^2\right) = \frac{1}{16\pi^2}\left(-\frac{11}{3}C_1 + \frac{1}{6}C_2\, N_{scalar} + \frac{2}{3}C_3 N_{fermion}\right), \quad (10.32)$$

so that, if you take SU(2), 11 fermion species would be needed to cancel the vector-boson contribution ($C_1 = 2, C_3 = 1$). Symanzik said to me that he believed I had made a sign error, but if it was correct I should publish it, because it was important, and if I did not, somebody else would.

I knew I had made no sign error (the origin of the sign differences was evident in the calculations). But there was still much work to do on quantum gravity, and also I would have to explain in detail my calculational procedure for which much time was needed. And so, my remarks remained largely unnoticed, by all except Symanzik. He first remained quiet (reportedly to allow me to correct my "mistake"), but then mentioned my result to Giorgio Parisi, who came to CERN and discussed the topic with me. Then, when the news about "asymptotic freedom" came, from the United States, Symanzik was the first to point out to everyone involved that the discovery was first made in Europe, and that an announcement made at a conference counts when matters of priority are concerned.[34]

Topological aspects of gauge theories

In weak interaction theories all phenomena related to the nonlinearity of the field equations are rare and weak. The perturbative expansion converges rapidly. And so, in the early days, it was natural to think that topologically nontrivial field configurations would never play a significant role whatsoever.

Yet the first interesting idea about topologically nontrivial structures in gauge theories was launched by Holger-Bech Nielsen and Poul Olesen, and independently by Bruno Zumino.[35] They considered stable magnetic flux vortices in the Abelian Higgs model, suspecting that these might have something to do with the dual string theory for mesons. Then, when you try to generalize this to the non-Abelian case, you hit upon a paradox, and this led to the discovery of the magnetic monopole.[36] This monopole was also found along a different route. Alexander Polyakov at the Landau Institute studied three-dimensional "hedgehog solutions" and found that these topologically stable objects have finite energy when they are coupled to a non-Abelian gauge theory.[37] According to a footnote in his publication, it was Lev Okun who remarked that his hedgehog must carry magnetic charge, which is why I believe that perhaps also Okun's name should be attached to the monopole.

The flux vortex is stable in two dimensions, the monopole in three; is there anything that is topologically stable in four dimensions? Sure there is! Alexander Belavin, Polyakov, Albert Schwarz, and Yuri Tyupkin were the first to point out that pure gauge theories allow for such a structure.[38] But what are the physical consequences of this idea? Naturally, the fields are localized not only in three-space, but also in time. Hence they describe an event. This is why we called this an "instanton."[39]

And instantons may be important events, in particular if the gauge theory is coupled to fermions. The solutions of the coupled Dirac equation near an instanton are so special that several kinds of fermionic conservation laws may be broken there. One of those laws was a chiral symmetry law that would prevent the η meson from having a mass. We now know that the η mass is entirely due to QCD instantons.[40] The electroweak instanton, on the other hand, gives rise to more drastic violations of conservation laws, but it is very rare. Three quarks of each generation and one lepton of each generation could be absorbed by one

instanton:

$$u + u + d + c + c + s + t + t + b \Rightarrow e^+ + \mu^+ + \tau^+ \qquad (10.33)$$

Now this will probably never be seen. (There are claims that this process becomes detectable at energies in the 10–20 TeV region.[41] It is, however, practically certain that even at high accelerator energies, it remains exponentially damped.) However, one experimental consequence of the electroweak instanton is immediate. We see that (10.33) would be at odds with electric charge conservation if the number of baryonic and leptonic generations were not equal. These numbers *must* be equal for the theory to be self-consistent. We observe that experiment agrees with us here.

Further developments

It is a characteristic of successful theories that they provide further understanding in many different areas of the field, in elegant and unsuspected ways. As for the Standard Model, we now know that the roles of asymptotic freedom, monopoles, and instantons are crucial in our present picture of quark confinement, the hadron spectrum, the scaling phenomena, and jet physics. The renormalized theory allows us to reproduce the observed data on the Z and W bosons with unprecedented precision. The Standard Model, as a gauge theory with fermions and at most only one scalar, is indeed tremendously successful.

Of course, after two decades have passed, the deficiencies in our theory are also standing out clearly. A theory that explains why the local symmetry is as it is, where the fermion spectrum comes from, and how the values of some 20 constants of Nature are determined is still being sought, but it is difficult to believe that such a giant leap in particle theory as occurred in the 1970s will be repeated in the near future.

Notes

1 B. E. Lautrup, A. Peterman, and E. de Rafael, "Recent developments in the comparison between theory and experiment in QED," *Phys. Rep. 3C* (1972), pp. 196–259.

2 A nice account of these views is presented in T. Y. Cao and S. S. Schweber, "The Conceptual Foundations and the Philosophical Aspects of Renormalization Theory," *Synthese 97* (1993), pp. 33–108; S. Schweber, "A Historical Perspective on the Rise of the Standard Model, Traditions, Men," Chapter 38, this volume.

3 W. Pauli and F. Villars, "On Regularization in Quantum Electrodynamics," *Rev. Mod. Phys. 21* (1949), pp. 434–41.
4 R. P. Feynman and M. Gell-Mann, "Theory of the Fermi Interactions," *Phys. Rev. 109* (1958), pp. 193–8; E. C. G. Sudarshan and R. E. Marshak, "Charality Invariance and the Universal Fermi Interactions," *Phys. Rev. 109* (1958), pp. 1860–1.
5 M. Veltman, "Perturbation Theory of Massive Yang–Mills Fields," *Nucl. Phys. B7* (1968), pp. 637–50; J. Reiff and M. Veltman, "Massive Yang–Mills Fields," *Nucl. Phys. B13* (1969), pp. 545–64; M. Veltman, "Generalized Ward Identities and Yang–Mills Fields," *Nucl. Phys. B21* (1970), pp. 288–302.
6 G. 't Hooft and M. Veltman, "One-loop divergencies in the theory of gravitation," *Annales de l' Institut Henri Poincaré 20* (1974), pp. 69–94.
7 G. 't Hooft, in "The Whys of Subnuclear Physics," ed., A. Zichichi (New York, London: Plenum), p. 943; G. 't Hooft, "Borel Summability of a Four-Dimensional Field Theory," *Phys. Lett. 119B* (1982), pp. 369–71; G. Parisi, "On Infrared Divergences," *Nucl. Phys. B150* (1979), pp. 163–72.
8 G. 't Hooft, "On the Convergence of Planar Diagram Expressions," *Comm. Math. Phys. 86* (1982), pp. 449–63; G. 't Hooft, "Rigorous Construction of Planar-Diagram Field Theories in Four-Dimensional Euclidian Space," *Comm. Math. Phys. 88* (1983), pp. 1–25.
9 C. N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance," *Phys. Rev. 96* (1954), pp. 191–5; see also R. Shaw, Cambridge University Ph.D. thesis (unpublished).
10 R. P. Feynman, "Quantum Theory of Gravitation," *Acta Phys. Pol. 24* (1963), pp. 697–722.
11 S. L. Glashow, "Partial Symmetries of Weak Interactions," *Nucl. Phys. 22* (1961), pp. 579–88. M. Veltman, "Perturbation Theory of Massive Yang–Mills Fields"; J. Reiff and M. Veltman, "Massive Yang–Mills Fields"; M. Veltman, "Generalized Ward Identities."
12 M. Veltman, "Perturbation Theory of Massive Yang–Mills Fields"; J. Reiff and M. Veltman, "Massive Yang–Mills Fields"; M. Veltman, "Generalized Ward Identities."
13 M. Gell-Mann and M. Lévy, "The Axial Vector Current in Beta Decay," *Nuovo Cimento 16* (1960), pp. 705–26.
14 B. W. Lee, "Renormalization of the $\sigma-$ Model," *Nucl. Phys. B9* (1969), pp. 649–72; J.-L. Gervais and B. W. Lee, "Renormalization of the $\sigma-$ model," *Nucl. Phys. B12* (1969), pp. 627–46; B. W. Lee, "Chiral Dynamics," in *Cargèse Lectures in Physics*, Vol. 5 (New York: Gordon and Breach 1972); K. Symanzik, "Renormalizable Models with Simple Symmetry Breaking," *Lett. Nuovo Cimento 2* (1969), p. 10 and *Comm. Math. Phys. 16* (1970), pp. 48–80.
15 Strictly speaking a local gauge symmetry is never spontaneously broken, since the vacuum is always completely symmetric. The Higgs mechanism is really something else: a rearrangement of the spectrum of states. But because of its analogy with spontaneous breakdown of a global symmetry one often uses this incorrect phrase. Hence the quotation marks.
16 L. D. Faddeev and V. N. Popov, "Feynman Diagrams for the Yang–Mills Field," *Phys. Lett. 25B* (1967), pp. 29–30; see also L. D. Faddeev, "The Feynman Integral for Singular Lagrangians," *Theoretical and*

Mathematical Physics 1 (1969), pp. 3–18 (in Russian), pp. 1–13 (Engl. transl).

17 S. Mandelstam, "Feynman Rules for Electromagnetic and Yang–Mills Fields From the Gauge-Independent Field-Theoretic Formalism," *Phys. Rev. 175* (1968), pp. 1580–1603.

18 G. 't Hooft, "Renormalization of Massless Yang–Mills Fields," *Nucl. Phys. B33* (1971), pp. 173–99.

19 C. Becchi, A. Rouet, and R. Stora, "Renormalization of Gauge Theories," *Ann. Phys. (N.Y.) 98* (1976), pp. 287–321; I. V. Tyutin, "Cargèse Lectures," *Lebedev Report No. FIAN39* (1975), unpublished; R. Stora, "Continuum Gauge Theories," in M. Lévy and P. Mitter, eds., *New Developments in Quantum Field Theory and Statistical Mechanics* (New York: Plenum Press, 1977), pp. 201–24; J. Thierry-Mieg, "Geometrical Reinterpretation of Faddeev–Popov Ghost Particles and BRS Transformations," *J. Math. Phys. 21* (1980), pp. 2834–8.

20 S. L. Adler, "Axial Vector Vortex in Spin Electrodynamics," *Phys. Rev. 177* (1969), pp. 2426–38; J. S. Bell and R. Jackiw, "A PCAC Puzzle $\pi^0 \to \gamma\gamma$ in the $\sigma-$ Model," *Nuovo Cimento A60* (1969), pp. 47–60.

21 A. Slavnov, "Ward Identities in Gauge Theories," *Theoretical and Mathematical Physics 10* (1972) (English translation), pp. 99-104; J. C. Taylor, "Ward Identities and Charge Renormalization of the Yang–Mills Fields," *Nucl. Phys. B33* (1971), pp. 436–44.

22 G. 't Hooft, "Renormalizable Lagrangians for Massive Yang–Mills Fields," *Nucl. Phys. B35* (1971), pp. 167–88.

23 G. 't Hooft and M. Veltman, "Regularization and Renormalization of Gauge Fields," *Nucl. Phys. B44* (1972), pp. 189–213.

24 Kenneth G. Wilson, "Renormlization Group and Strong Interactions," *Phys. Rev. D3* (1971), pp. 1818–46; Kenneth G. Wilson and Michael E. Fisher, "Critical Exponents in 3.99 Dimensions," *Phys. Rev. Lett. 28* (1972), pp. 240–3.

25 C. Bollini and J. Giambiagi, "Dimensional Renormalization," *Nuovo Cim. 12B* (1972), pp. 20–6; J. Ashmore, "A Method of Gauge-Invariant Regularization," *Lett. Nuovo Cim. 4* (1972), pp. 289–90.

26 G. 't Hooft and M. Veltman, "Combinatorics of Gauge Fields," *Nucl. Phys. B50* (1972), pp. 318–53.

27 G. 't Hooft, "Prediction for Neutrino–Electron Cross-Sections in Weinberg's Model of Weak Interactions," *Phys. Lett. 37B* (1971), pp. 195–6.

28 Stephen L. Adler and William A. Bardeen, "Absence of Higher-Order Corrections in the Anomalous Axial-Vector Divergence Equation," *Phys. Rev. 182* (1969), pp. 1517–36; William A. Bardeen, "Anomalous Ward Identities in Spinor Field Theories," *Phys. Rev. 184* (1969), pp. 1848–59; David G. Boulware, "Quantum Field Theory in Schwarzschild and Rindler Spaces," *Phys. Rev. D11* (1975), pp. 1404–23; David G. Boulware, "Hawking Radiation and Thin Shells," *Phys. Rev. D13* (1976), pp. 2169–87.

29 C. Llewellyn-Smith, "High Energy Behaviour and Gauge Symmetry," *Phys. Lett. B46* (1973), pp. 233–6. John M. Cornwall, David N. Levin, and George Tiktopoulos, "Uniqueness of Spontaneously Broken Gauge Theories," *Phys. Rev. Lett. 30*, (1973), pp. 1268–70.

30 G. 't Hooft and M. Veltman, "Combinatorics of Gauge Fields."

31 M. Veltman, "Perturbation Theory of Massive Yang–Mills Fields"; J. Reiff and M. Veltman, "Massive Yang–Mills Fields"; M. Veltman, "Generalized Ward Identities."
32 C. P. Korthals-Altes, ed., *Renormalization of Yang–Mills Fields and Applications to Particle Physics*, Marseille, 19–23 June 1972 (Marseille: Centre de Physique Théorique, 1972).
33 K. Symanzik, "On Theories with Massless Particles," in C. P. Korthals-Altes, ed., *Renormalization of Yang–Mills Fields*; K. Symanzik, "A Field Theory with Computable Large-Momenta Behaviour," *Lett. Nuovo Cimento 6* (1973), pp. 77–80.
34 David J. Gross and Frank Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories," *Phys. Rev. Lett. 30* (1973), pp. 1343–6; H. David Politzer, "Reliable Perturbative Results for Strong Interactions," *Phys. Rev. Lett. 30* (1973), pp. 1346–9; H. David Politzer, "Asymptotic Freedom: An Approach to Strong Interactions," *Phys. Rep. 14C* (1974), pp. 129–80.
35 H. B. Nielsen and P. Olesen, "Vortex-Line Models for Dual Strings," *Nucl. Phys. B61* (1973), pp. 45–61. Bruno Zumino, "Relativistic Strings and Supergauges," pp. 367–81, and "Application of Gauge Theories to Weak and Electromagnetic Interactions," pp. 383–98, both in Eduardo R. Caianiello, ed., *Renormalization and Invariance in Quantum Field Theory*, Capri Summer Meeting, July 1973 (New York: Plenum Press, 1974).
36 G. 't Hooft, "Magnetic Monopoles in Unified Gauge Theories," *Nucl. Phys. B79* (1974), pp. 276–84.
37 A. M. Polyakov, "Particle Spectrum in Quantum Field Theory," *JETP Lett. 20* (1974), pp. 194–5.
38 A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Yu. S. Tyupkin, "Pseudoparticle Solutions of the Yang–Mills Equations," *Phys. Lett. 59B* (1975), pp. 85–7.
39 G. 't Hooft, "Symmetry Breaking Through Bell–Jackiw Anomalies," *Phys. Rev. Lett. 37* (1976), pp. 8–11; "Computation of the Quantum Effects due to a Four-Dimensional Pseudoparticle," *Phys. Rev. D14* (1976), pp. 3432–50; R. Jackiw and C. Rebbi, "Vacuum Periodicity in a Yang–Mills Quantum Theory," *Phys. Rev. Lett. 37* (1976), pp. 172–5; C. G. Callan, Jr., R. F. Dashen, and D. J. Gross, "The Structure of the Gauge Theory Vacuum," *Phys. Lett. 63B* (1976), pp. 334–40; Curtis G. Callen, Jr., Roger Dashen, and David J. Gross, "Toward a Theory of the Strong Interactions," *Phys. Rev. D17* (1978), pp. 2717–63.
40 G. 't Hooft, "How Instantons Solve the U(1) Problem," *Phys. Rep. 142* (1986), pp. 357–87.
41 A. De Rujula, H. Georgi, S. L. Glashow, and H. R. Quinn, "Fact and Fancy in Neutrino Physics," *Rev. Mod. Phys. 46* (1974), pp. 391–407.