descent approximation,

$$\text{im } J = \text{Im} \int_1^{1+i\infty} dz(2\pi\hbar)^{-\frac{1}{2}} e^{-S(1)/\hbar} e^{-\frac{1}{2}S''(1)(z-1)^2/\hbar}$$
$$= \tfrac{1}{2} e^{-S(1)/\hbar} |S''(1)|^{-\frac{1}{2}}. \tag{2.52}$$

Note the factor of $\frac{1}{2}$; this arises because the integration is over only half of the Gaussian peak.

(If we had passed from one potential to the other in the conjugate manner, the contour would have been distorted into the lower half plane, and we would have obtained the opposite sign for the imaginary part. This is just a reflection of the well-known fact that what sign you get for the imaginary part of the energy of an unstable state depends on how you do your analytic continuation.)

Now, we have studied a one-dimensional integral, but we can always reduce our functional integral to a one-dimensional integral simply by integrating (in the Gaussian approximation) over all the variables orthogonal to our path. These directions involve only positive or zero eigenvalues near the stationary point and give us no trouble. In this manner we obtain Sagredo's answer, Eq. (2.49), except that the negative eigenvalue carries a factor of $\frac{1}{2}$ with it; that is to say, we obtain Eq. (2.50).

3 The vacuum structure of gauge field theories[9]

3.1 *Old stuff*

This subsection is a telegraphic compendium of formulae from gauge field theories. Its purpose is to establish notational conventions and possibly to jog your memory. If you do not already know the fundamentals of gauge field theory, you will not learn them here.[10]

Lie algebras. A representation of Lie algebra is a set of N anti-Hermitian matrices, T^a, $a=1\ldots N$, obeying the equations

$$[T^a, T^b] = c^{abc} T^c, \tag{3.1}$$

where the cs are the structure constants of some compact Lie group, G. It is always possible to choose the Ts such that $\text{Tr}(T^a T^b)$ is proportional to δ^{ab}, although the constant of proportionality may depend on the representation. The Cartan inner product is defined by

$$(T^a, T^b) = \delta^{ab}. \tag{3.2}$$

Thus this is proportional to the trace of the product of the matrices.

So far I have not stated a convention that gives a scale to the structure constants and thus to the Ts. For SU(2), the case I will spend most time

discussing, I will choose c^{abc} to be equal to ε^{abc}. Thus, for the isospinor representation,

$$T^a = -i\sigma^a/2, \tag{3.3}$$

where the σs are the Pauli spin matrices. In this case,

$$(T^a, T^b) = -2\,\text{Tr}(T^a T^b). \tag{3.4}$$

Occasionally I will discuss SU(n), in particular SU(3). In this case I will choose the structure constants to agree with the preceding convention for the SU(2) subgroup composed of unitary unimodular transformations on two variables only. Thus, for SU(3), T^a is $-i\lambda^a/2$, where the λs are Gell-Mann's matrices.

Gauge fields. The gauge potentials are a set of vector fields, $A^a_\mu(x)$. It is convenient to define a matrix-valued vector field, $A_\mu(x)$, by

$$A_\mu = g A^a_\mu T^a, \tag{3.5}$$

where g is a constant called the gauge coupling constant. The field-strength tensor, $F_{\mu\nu}(x)$, is defined by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]. \tag{3.6}$$

Pure gauge field theory is defined by the Euclidean action,

$$S = \frac{1}{4g^2} \int d^4x (F_{\mu\nu}, F_{\mu\nu}). \tag{3.7}$$

Sometimes I will write this in a shorthand form,

$$S = \frac{1}{4g^2} \int (F^2). \tag{3.8}$$

Gauge transformations. A gauge transformation is a function, $g(x)$, from Euclidean space into the gauge group, G. In equations,

$$g(x) = \exp \lambda^a(x) T^a, \tag{3.9}$$

where the λs are arbitrary functions. (Please do not confuse $g(x)$ with the coupling constant, g.) Under such a transformation,

$$A_\mu \to g A_\mu g^{-1} + g \partial_\mu g^{-1}, \tag{3.10}$$

and

$$F_{\mu\nu} \to g F_{\mu\nu} g^{-1}. \tag{3.11}$$

Thus, S is gauge-invariant. If $F_{\mu\nu}$ vanishes, then A_μ is a gauge-transform of zero; that is to say,

$$A_\mu = g \partial_\mu g^{-1}, \tag{3.12}$$

for some $g(x)$.

Covariant derivatives. The covariant derivative of the field strength tensor is defined by

$$D_\lambda F_{\mu\nu} = \partial_\lambda F_{\mu\nu} + [A_\lambda, F_{\mu\nu}]. \tag{3.13}$$

Equation (3.7) leads to the Euclidean equations of motion

$$D_\mu F_{\mu\nu} = 0. \tag{3.14}$$

Given a field ψ that gauge-transforms according to

$$\psi \to g(x)\psi, \tag{3.15}$$

then the covariant derivative of ψ,

$$D_\mu \psi = \partial_\mu \psi + A_\mu \psi, \tag{3.16}$$

transforms in the same way.

3.2 *The winding number*

I propose to study Euclidean gauge field configurations of finite action (not necessarily solutions of the equations of motion).

Why?

The naive answer, sometimes given in the literature,[11] is that configurations of infinite action are unimportant in the functional integral, since, for such configurations, $e^{-S/\hbar}$ is zero. *This is wrong.* In fact, it is configurations of finite action that are unimportant; to be precise, they form a set of measure zero in function space. This has nothing to do with the divergences of quantum field theory; it is true even for the ordinary harmonic oscillator. (For a proof, see Appendix 3.) The only reason we are interested in configurations of finite action is that we are interested in doing semiclassical approximations, and a configuration of infinite action does indeed give zero if it is used as the center point of a Gaussian integral.

The convergence of the action integral is controlled by the behavior of A_μ for large r, where r is the radial variable in Euclidean four-space. To keep my arguments as simple as possible, I will assume that, for large r, A_μ can be expanded in an asymptotic series in inverse powers of r. (This assumption can be relaxed considerably without altering the conclusions.)[12] Thus, for the action to be finite, $F_{\mu\nu}$ must fall off faster than $1/r^2$ as r goes to infinity; that is to say, $F_{\mu\nu}$ must be $O(1/r^3)$. One's first thought is that this implies that A_μ is $O(1/r^2)$, but this is wrong: vanishing $F_{\mu\nu}$ does not imply vanishing A_μ, but merely that A_μ is a gauge transform of zero. Thus A_μ can be of the form

$$A_\mu = g\partial_\mu g^{-1} + O(1/r^2), \tag{3.17}$$

where g is a function from four-space to G of order one, that is to say, a function of angular variables only.

Thus, with every finite-action field configuration there is associated a group-element-valued function of angular variables, that is to say, a mapping of a three-dimensional hypersphere, S^3, into the gauge group, G. Of course, this assignment is not gauge-invariant. Under a gauge transformation, $h(x)$

$$A_\mu \to hA_\mu h^{-1} + h\partial_\mu h^{-1}. \tag{3.18}$$

Thus,

$$g \to hg + O(1/r^2). \tag{3.19}$$

If one could choose h to equal g^{-1} at infinity, one could transform g to one and eliminate it from Eq. (3.17). In general, though, this is not possible. The reason is that h must be a continuous function not just on the hypersphere at infinity, but throughout all four-space, that is to say, on a nested family of hyperspheres going all the way from r equals zero to r equals infinity. In particular, at the origin, h must be a constant, independent of angles. Thus, h at infinity can not be a general function on S^3, but must be one that can be obtained by continuous deformation from a constant function. Since any constant gauge transformation can trivially be obtained by continuous deformation from the identity transformation (all gauge groups are connected), we might as well say that h at infinity must be obtainable from $h=1$ by a continuous deformation.

Given two mappings of one topological space into another, such that one mapping is continuously deformable into another, mathematicians say the two functions are 'homotopic' or 'in the same homotopy class'. What we have shown is that by a gauge transformation we can transform $g(x)$ into any mapping homotopic to $g(x)$, but we can not transform it into a function in another homotopy class. Thus, the gauge-invariant quantity associated with a finite-action field configuration is not a mapping of S^3 to G but a homotopy class of such mappings. Our task is to find these homotopy classes for physically interesting Gs.

To warm up for this task, let me consider a baby version of the problem for which the geometry is somewhat easier to visualize. I will work with the simplest of all gauge groups, U(1), the group of complex numbers of unit modulus. Thus the gauge field theory is ordinary electromagnetism. (However, I will still keep to the notational conventions established in Sect. 3.1; in particular, A_μ will be an imaginary quantity, i times the usual vector potential.) Also, I will work not in Euclidean four-space but in Euclidean two-space. I will still study fields obeying Eq. (3.17), although, of course, in two-space this condition is not a consequence of finiteness of the action. Because we are working in two-space, we have, instead of a hypersphere, S^3, an ordinary circle, S^1.

Now to work:

(1) G is the unit circle in the complex plane; thus, topologically, G is also S^1, and we have to study homotopy classes of mappings of S^1 into S^1. We will label the circle in space, the domain of our functions, in the standard way, by an angle θ ranging from 0 to 2π.

(2) It will be useful to define some standard mappings from S^1 to S^1. One is the trivial mapping,

$$g^{(0)}(\theta)=1. \tag{3.20a}$$

Another is the identity mapping,

$$g^{(1)}(\theta)=e^{i\theta}. \tag{3.20b}$$

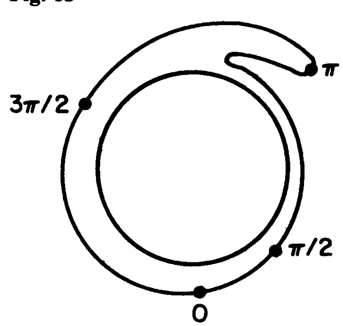
These are both part of a family of mappings,

$$g^{(\nu)}(\theta)=[g^{(1)}(\theta)]^{\nu}=e^{i\nu\theta}, \tag{3.20c}$$

where ν is an integer (positive, negative, or zero). ν is called the 'winding number', because it is the number of times we wind around G when we go once around the circle at infinity in two-space. (By convention, winding around minus once means winding around once in the negative direction.)

(3) Every mapping from S^1 to S^1 is homotopic to one of the mappings (3.20c). We do not have the mathematical machinery to prove this rigorously, but I hope I can make it plausible. Imagine taking a rubber band and marking on it in ink a sequence of values of θ running from 0 to 2π. We then wrap the band about a circle representing G, such that each value of θ lies above the point into which it is mapped. (Fig. 13 shows such a construction.) We can continuously deform the band, first to eliminate any folds, like the one on the top of the figure, and second to stretch the band so it lies uniformly on the circle. In this way we obtain some $g^{(\nu)}(\theta)$. (In the case shown, we obtain $g^{(1)}$.) Thus we can associate a winding number with every mapping. (Note that I have not yet shown that this number is uniquely defined.)

Fig. 13



(4) I will now show that the winding number defined above is given by the integral formula

$$\nu=\frac{i}{2\pi}\int_0^{2\pi} d\theta g\, dg^{-1}/d\theta. \tag{3.21}$$

Firstly, by direct calculation, this gives the right answer for the standard mappings, Eq. (3.20c). Secondly, this quantity is invariant under continuous deformations. To prove this assertion it suffices to demonstrate invariance under infinitesimal deformations. A general infinitesimal deformation is of the form

$$\delta g=i(\delta\lambda)g, \tag{3.22}$$

where $\delta\lambda$ is some infinitesimal real function on the circle. Thus

$$\delta(g dg^{-1}/d\theta)=-id(\delta\lambda)/d\theta, \tag{3.23}$$

and the change in ν vanishes upon integration. (We now know that all of our standard mappings are in different homotopy classes and that the winding number is uniquely defined.)

(5) If

$$g(\theta)=g_1(\theta)g_2(\theta), \tag{3.24a}$$

then

$$\nu=\nu_1+\nu_2. \tag{3.24b}$$

The proof is simple. The winding number is unchanged by continuous deformations. We can deform g_1 such that it is equal to one on the upper half of the circle ($0\leqslant\theta\leqslant\pi$) and g_2 such that it is equal to one on the lower half of the circle ($\pi\leqslant\theta\leqslant 2\pi$). The integrand in Eq. (3.21) is then the sum of a part due to g_1 (vanishing on the upper semicircle) and a part due to g_2 (vanishing on the lower semicircle).

(6) Let us define

$$G_\mu=\frac{i}{2\pi}\varepsilon_{\mu\nu}A_\nu.\text{[13]} \tag{3.25}$$

By Eqs. (3.17) and (3.21),

$$\nu=\lim_{r\to\infty}\int_0^{2\pi} r d\theta \hat{r}_\mu G_\mu, \tag{3.26}$$

where $\hat{r}_\mu$ is the radial unit vector. Thus, by Gauss's theorem,

$$\nu=\int d^2x\partial_\mu G_\mu. \tag{3.27}$$

Hence,

$$\nu=\frac{i}{4\pi}\int d^2x\varepsilon_{\mu\nu}F_{\mu\nu}. \tag{3.28}$$

I will now return to four-space, and take G to be SU(2). As we shall see, every argument will be a (mild) generalization of the arguments I have given for the baby problem.

(1) SU(2) is the group of unitary unimodular two-by-two matrices. It is well known that any such matrix can be uniquely written in the form

$$g = a + i\mathbf{b}\cdot\sigma, \tag{3.29}$$

where $a^2 + |b|^2 = 1$. Thus, topologically, SU(2) is S^3, and we have to study homotopy classes of mappings from S^3 to S^3.

(2) It will be useful to define some standard mappings from S^3 to S^3. One is the trivial mapping,

$$g^{(0)}(x) = 1. \tag{3.30a}$$

Another is the identity mapping,

$$g^{(1)}(x) = (x_4 + i\mathbf{x}\cdot\sigma)/r. \tag{3.30b}$$

These are both part of a family of mappings,

$$g^{(\nu)}(x) = [g^{(1)}(x)]^{\nu}. \tag{3.30c}$$

where ν is an integer, called the winding number. (It is also sometimes called the Pontryagin index.) It measures the number of times the hypersphere at infinity is wrapped around G. (By convention, we say the hypersphere is wrapped around G in a negative sense if a right-handed triad of tangent vectors is mapped into a left-handed triad.)

(3) Every mapping from S^3 to S^3 is homotopic to one of our standard mappings (3.30c). We do not have the mathematical machinery to prove this assertion rigorously, but a plausibility argument can be constructed just as in the baby problem, with hyperspheres replacing circles. (If you have problems envisioning hyperspheres wrapped around hyperspheres, just accept the assertion on faith.) In this way we can associate a winding number with every mapping. (Note that I have not yet shown that this number is uniquely defined.)

(4) Let us define

$$\nu = \frac{1}{48\pi^2}\int d\theta_1\, d\theta_2\, d\theta_3\, \varepsilon^{ijk}(g\partial_i g^{-1}, g\partial_j g^{-1}\, g\partial_k g^{-1}). \tag{3.31}$$

where θ_1, θ_2 and θ_3 are three angles that parametrize S^3. How these angles are chosen is irrelevant to Eq. (3.31); the Jacobian determinant that comes from changing the angles is canceled by the Jacobian determinant from the ε-symbol. Equation (3.31) is written using the Cartan inner product, that is to say, in a representation-independent way. Of course, for any particular representation of SU(2), we can rewrite Eq. (3.31) in terms of traces; for example, for the two-dimensional representation, by

Eq. (3.4),

$$\nu = -\frac{1}{24\pi^2}\int d\theta_1\, d\theta_2\, d\theta_3\, \mathrm{Tr}\,\varepsilon^{ijk} g\partial_i g^{-1}\, g\partial_j g^{-1}\, g\partial_k g^{-1}. \tag{3.32}$$

I will show that this quantity is, firstly, a homotopy invariant, and secondly, agrees with the winding number as defined for our standard mappings. As before, a corollary of this proof will be that all of our standard mappings are in different homotopy classes and that the winding number is uniquely defined.

To show invariance under continuous deformations it suffices to show invariance under infinitesimal deformations. For any Lie group, a general infinitesimal transformation can be written as an infinitesimal right multiplication:

$$\delta g = g\delta\lambda^a(x)T^a \equiv g\delta T. \tag{3.33}$$

Under this transformation,

$$\delta(g\partial_k g^{-1}) = -g(\partial_k \delta T)g^{-1}. \tag{3.34}$$

The three derivatives in Eq. (3.32) make equal contributions to $\delta\nu$; thus,

$$\delta\nu \propto \int d\theta_1\, d\theta_2\, d\theta_3\, \varepsilon^{ijk}\, \mathrm{Tr}\, g\partial_i g^{-1}\, g\partial_j g^{-1}\, g(\partial_k\delta T)g^{-1}. \tag{3.35}$$

If we use the identity,

$$0 = \partial_i(gg^{-1}) = g\partial_i g^{-1} + (\partial_i g)g^{-1}, \tag{3.36}$$

this becomes

$$\delta\nu \propto \int d\theta_1\, d\theta_2\, d\theta_3\, \varepsilon^{ijk}\, \mathrm{Tr}\, \partial_i g^{-1}\, \partial_j g \partial_k \delta T, \tag{3.37}$$

which vanishes upon integration by parts, because of the antisymmetry of the ε-symbol. This completes the proof of invariance under continuous deformations.

(5) Now to evaluate Eq. (3.32) for our standard mappings. The task is easiest for $g^{(1)}$, for the integrand is here obviously a constant, and we need evaluate it only at the north pole of the unit hypersphere, $x_4 = 1$, $x_i = 0$. At this point we might as well choose θ_i to equal x_i. Thus, from Eq. (3.30b),

$$g\partial_i g^{-1} = -i\sigma_i, \tag{3.38}$$

and

$$\mathrm{Tr}\,\varepsilon^{ijk} g\partial_i g^{-1}\, g\partial_j g^{-1}\, g\partial_k g^{-1} = -12. \tag{3.39}$$

Since the area of a unit hypersphere is $2\pi^2$, we obtain the desired result, $\nu = 1$.

For the other standard mappings, the simplest way to proceed is to observe that if

$$g=g_1g_2, \tag{3.40a}$$

then

$$\nu=\nu_1+\nu_2. \tag{3.40b}$$

The argument is the same as for the baby problem, with semihyperspheres replacing semicircles.

(6) Let us define

$$G_\mu=2\varepsilon_{\mu\nu\lambda\sigma}(A_\nu, \partial_\lambda A_\sigma+\tfrac{2}{3}A_\lambda A_\sigma). \tag{3.41}$$

A straightforward computation shows that

$$\partial_\mu G_\mu=\tfrac{1}{2}\varepsilon_{\mu\nu\lambda\sigma}(F_{\mu\nu}, F_{\lambda\sigma}). \tag{3.42}$$

The dual of an antisymmetric tensor (denoted by a tilde) is conventionally defined by

$$\tilde{F}_{\mu\nu}\equiv\tfrac{1}{2}\varepsilon_{\mu\nu\lambda\sigma}F_{\lambda\sigma}. \tag{3.43}$$

(The factor of $\frac{1}{2}$ is inserted in the definition so that $\tilde{\tilde{F}}=F$.) Equation (3.42) can thus be rewritten as

$$\partial_\mu G_\mu=(F_{\mu\nu}, \tilde{F}_{\mu\nu})\equiv(F, \tilde{F}). \tag{3.44}$$

From the definition of $F_{\mu\nu}$,

$$G_\mu=\varepsilon_{\mu\nu\lambda\sigma}(A_\nu, F_{\lambda\sigma}-\tfrac{2}{3}A_\lambda A_\sigma). \tag{3.45}$$

This expression is useful in evaluating

$$\int d^4x(F, \tilde{F})=\int d^3S\hat{r}_\mu G_\mu, \tag{3.46}$$

where d^3S is the element of area on a large hypersphere. The first term in Eq. (3.45) is $O(1/r^4)$ and makes no contribution to the integral; the second term simply gives (up to a multiplicative constant) the integral formula for the winding number, Eq. (3.31). Thus we obtain

$$\int d^4x(F, \tilde{F})=32\pi^2\nu. \tag{3.47}$$

Summary and generalizations. This has been a long analysis, and you may have lost track of what we were doing, so let me summarize the main results of this subsection. For a gauge field theory based on the group SU(2), every field configuration of finite action in four-dimensional Euclidean space has an integer associated with it, the Pontryagin index or winding number, ν. It is not possible to continuously deform a configuration of one winding number into one of a different winding number while maintaining the finiteness of the action. We have two integral formulae for the winding number, one in terms of a surface integral over a large sphere, Eq. (3.31), and one in terms of a volume integral over all four-space, Eq. (3.47).

How much of this depends on the gauge group being SU(2)? Firstly, if the gauge group is U(1), it is easy to see that every mapping of S^3 into U(1) is continuously deformable into the trivial mapping (all of S^3 mapped into a single point). *Thus, for an Abelian gauge field theory, there is no analog of the winding number.* Secondly, for a general simple Lie group, G, there is a remarkable theorem due to Raoul Bott[14] that states that any continuous mapping of S^3 into G can be continuously deformed into a mapping into an SU(2) subgroup of G. *Thus, everything we have discovered for* SU(2) *is true for an arbitrary simple Lie group; in particular, it is true for* SU(n). I stress that 'everything' means *everything*. In particular, not a single numerical factor in the integral formulas for the winding number needs alteration, so long as we choose the normalization of the Cartan inner product appropriately (as we have). Finally, since a general compact Lie group is locally the direct product of an Abelian group and a string of simple groups, *for a general gauge field theory, there is an independent winding number for every simple factor group.*

3.3 *Many vacua*

We have learned a lot about classical gauge field theories; now it is time to confront the quantum theory. In principle, the Euclidean functional integral tells how to go from the classical theory to the quantum theory. As I explained in Sect. 2, we can use the functional integral to study the energy eigenstates of the theory; also, by adding appropriate source terms to the Hamiltonian (equivalently, to the Euclidean action) and then differentiating with respect to the sources at the end of the computation, we can study the expectation values of strings of operators, Euclidean Green's functions. However, for gauge field theories, there is a famous complication: to make the functional integral well-defined, we must impose a gauge-fixing condition.[15]

I will choose to work in axial gauge, $A_3=0$. I have several reasons for this choice. (1) It is possible to show[16] that every non-singular gauge field configuration can be put in axial gauge by a non-singular gauge transformation. It is by no means clear whether this is true for covariant gauges, for example. (2) In axial gauge the functional integral is directly equivalent to a canonical formulation of the theory;[17] there is no need of the ghost terms that occur in covariant gauges, or of the subsidiary conditions on the space of states that are needed in such gauges as $A_0=0$. (3) Most of the treatment in the literature of the phenomena we are about to discuss is in the gauge $A_0=0$. It is nice to show explicitly that the answers do not depend on this gauge choice. (4) Although axial gauge is terribly awkward for specific computations, once we have obtained functional-integral expressions for quantities of interest, we can use the standard Faddeev–Popov methods to transform these into some more convenient gauge.

In field theory, we normally plunge directly into infinite space. However, I will here study gauge field theory in a finite box of three-volume V, with definite boundary conditions, which I shall specify shortly. Just as in Sect. 2, I will also restrict the theory to a finite range of Euclidean time, T, with appropriate boundary conditions at initial and final times. Thus we are integrating over a box in Euclidean four-space, with boundary conditions on the (three-dimensional) walls of the box. Of course, I will eventually send both V and T to infinity. I again have reasons for this choice. (1) Certainly nothing is lost by beginning in a finite box; if the transition to infinite space goes smoothly, at worst we will have wasted a little time. (2) In some theories, we can gain information about the structure of the theory by seeing how things depend on the boundary conditions imposed on the walls of the box. For example, in a scalar field theory with spontaneous symmetry breakdown, the expectation value of the scalar field in the center of the box depends on the boundary conditions on the walls, no matter how large the box; this is one of the easiest ways to see that the theory has many vacua. (3) In the canonical quantization of the theory, it is necessary to eliminate A_0 from the action. To do this, it is necessary to find A_0 from $\partial_3^2 A_0$. In infinite space, this problem has many solutions; this ambiguity is usually resolved by applying *ad hoc* conditions on the behavior of A_0 at infinity. In a box with appropriate boundary conditions, this problem always has a unique solution.

There are many possible types of boundary conditions we could impose: we could fix some components of A_μ, some components of $F_{\mu\nu}$, some combinations of these, etc. A clue to a wise choice of boundary conditions is given by the surface term in the expression for the variation of the action. For example, for a free scalar field theory,

$$\delta S = \int d^3 S n^\mu \partial_\mu \phi \delta\phi + \cdots . \tag{3.48}$$

Here, d^3S is the element of surface area, n^μ is the normal vector to the surface, and the triple dots denote the usual volume integral of the Euler–Lagrange equations. From this expression we see that one way to make the surface terms vanish is to fix the value of ϕ on the walls of the box. Likewise, for a gauge field theory,

$$\delta S = \frac{1}{g^2} \int d^3 S n^\mu F_{\mu\nu} \delta A^\nu + \cdots . \tag{3.49}$$

From this expression we see that one way to make the surface term vanish is to fix the tangential components of A_μ on the surface. Note that there is no need to fix the normal component of A_μ; because $F_{\mu\nu}$ is antisymmetric, this makes no contribution to the surface integral.

We are not totally free to choose the tangential components of A_μ arbitrarily. Firstly, they must be chosen consistent with our gauge condition, $A_3=0$. Secondly, because we want to do semiclassical computations, we must choose our boundary conditions to be consistent with finiteness of the action, as the box goes to infinity. Equivalently, the boundary conditions must be consistent with the box being filled with a field configuration of a definite winding number. Furthermore, for fixed boundary conditions, this winding number is fixed, for only the tangential components of A_μ are needed to compute the normal component of G_μ. (See Eq. (3.41).)

Thus at least one relic of our boundary conditions remains no matter how large the box: we can not put an arbitrary finite-action field configuration in the box, but only one of a definite winding number. It turns out that the winding number is the *only* relic of the boundary conditions that survives as the box goes to infinity. The hand-waving argument for this is that the winding number is the *only* gauge-invariant quantity associated with the large-distance behavior of the fields. If you do not find this argument convincing, you will find a more careful one in Appendix 4.

Thus, for large boxes, we can forget about the boundary conditions in the functional integral and simply integrate over all configurations where the winding number, ν, has some definite value, n. I will denote the result of such an integration by $F(V, T, n)$. In equations,

$$F(V, T, n) = N \int [dA] e^{-S} \delta_{\nu n}. \tag{3.50}$$

where $[dA]$ denotes $[dA_1][dA_2][dA_4]$. Also, I have set $\hbar$ to one; we can always keep track of the powers of $\hbar$ by keeping track of the powers of g, as explained in Sect. 1.

$F(V, T, n)$ is a transition matrix element from some initial state to some final state (determined by our boundary conditions). What these states are will not be important to us. What is important is that for large times, T_1 and T_2,

$$F(V, T_1+T_2, n) = \sum_{n_1+n_2=n} F(V, T_1, n_1) F(V, T_2, n_2). \tag{3.51}$$

This follows from Eq. (3.47), the expression for the winding number as the integral of a local density; this tells us that the way to put total winding number n in a large box is to put winding number n_1 in one part of the box and winding number n_2 in the remainder of the box, with $n=n_1+n_2$. (Of course, such counting misses field configurations with significant action density on the boundary between the two sub-boxes, for there is no reason for the winding-number integral for each sub-box to be an integer for such configurations. However, we expect this to be a negligible surface effect for sufficiently large boxes.)

Pretty as it is, Eq. (3.51) is not what we would expect from a transition-matrix element that has a contribution from only a single energy eigenstate. Such an object would be a simple exponential, and would obey a multiplicative composition law for large times, not the convolutive composition law of Eq. (3.51). However, it is easy enough to turn convolutions into multiplications. The technique is called Fourier transformation:

$$F(V,T,\theta)\equiv\sum_n e^{in\theta}F(V,T,n) = N\int [dA]e^{-S}e^{i\nu\theta}. \tag{3.52}$$

From Eq. (3.51),

$$F(V,T_1+T_2,\theta)=F(V,T_1,\theta)F(V,T_2,\theta). \tag{3.53}$$

This is the correct composition law for a simple exponential. Thus we identify $F(V,T,\theta)$ as being (up to a normalization constant) the expectation value of e^{-HT} in an energy eigenstate, which we denote by $|\theta\rangle$ and call the θ vacuum.

$$F(V,T,\theta)\propto\langle\theta|e^{-HT}|\theta\rangle = N'\int [dA]e^{-S}e^{i\nu\theta}. \tag{3.54}$$

where N' is a new normalization constant.

Our analysis has been simple and straightforward (I hope), but we have been led to a very unintuitive conclusion. Our original gauge field theory seems to have split up into a family of disconnected sectors, labeled by the angle θ, each with its own vacuum. Furthermore, in each of these sectors, the computational rules are the same as those we would have naively written down if we had not gone through any of this analysis, except that an extra term, proportional to $(F,\tilde{F})$, has been added to the Lagrangian density. Probably half the people who have played with gauge field theories have thought, at one time or another, of adding such a term, and they have discarded the possibility, because the added term is a total divergence (see Eq. (3.44)) and thus has no effect on the equations of motion and therefore 'obviously' has no effect on the physics of the

theory. Of course, at this stage in our investigation, it is still possible that we have been fooling ourselves, that the extra term indeed has no effect on the physics, and that all the θ vacua we think we have discovered are simply duplicates of the same state. We shall eliminate this possibility immediately.

(I should remark that what we have done here closely parallels the treatment of a periodic potential in Sect. 2.3, except the arguments are somewhat more abstract and in a different order. The winding number is something like the total change in x (the difference between the number of instantons and the number of anti-instantons) in Sect. 2.3, and the θ vacua are something like the $|\theta\rangle$ eigenstates. The two big differences are that we found the analogs of the $|\theta\rangle$ states without pausing to talk about the analogs of the $|j\rangle$ states, and that we did the Fourier transform that untangled the energy spectrum before we saturated the functional integral with instantons. The first difference is unimportant; if I had wanted to, I could have added two extra paragraphs when I was talking about $F(V,T,n)$ and discussed the analogs of the $|j\rangle$ states. (They are called n vacua.) As for the instantons, they are the subject of the next subsection.)

3.4 Instantons: generalities

In the next subsection I shall explicitly construct instantons, finite-action solutions of the Euclidean gauge-field equations with $\nu=1$. Most of the qualitative consequences of these solutions are independent of their detailed structure and follow merely from the fact of their existence. Therefore, in this subsection, I will simply assume that instantons exist and draw some conclusions from this assumption.

I will denote the action of an instanton by S_0. Because S_0 is finite, the instanton can not be invariant under spatial translations. Thus there exists at least a four-parameter family of instanton solutions; I will call these parameters 'the location of the center of the instanton'. The winding number is parity-odd. Thus there must also exist at least a four-parameter family of solutions with $\nu=-1$, the parity transforms of the instanton solutions, which I will call anti-instantons. Just as in Sect. 2, we can build approximate solutions consisting of n instantons and $\bar{n}$ anti-instantons, with their centers at arbitrary widely separated locations. These approximate solutions have $\nu=n-\bar{n}$.

Again as in Sect. 2, we approximate Eq. (3.54) by summing over all these configurations. Thus we obtain

$$\langle\theta|e^{-HT}|\theta\rangle\propto\sum_{n,\bar{n}}(Ke^{-S_0})^{n+\bar{n}}(VT)^{n+\bar{n}}e^{i(n-\bar{n})\theta}/(n!\,\bar{n}!) = \exp(2KVTe^{-S_0}\cos\theta), \tag{3.55}$$

where K is a determinantal factor, defined as in Sect. 2. Thus, the energy of a θ vacuum is given by

$$E(\theta)/V = -2K \cos\theta\, e^{-S_0}. \tag{3.56}$$

Note that, as should be the case in a field theory, the different vacua are distinguished not by different energies, but by different energy densities. (Also note the similarity with the energy spectrum of a periodic potential, Eq. (2.45).)

We can go on and compute the expectation values of various operators. A particularly easy (and particularly instructive) computation is that of the expectation value of $(F, \tilde{F})$. By translational invariance,

$$\langle\theta|(F(x), \tilde{F}(x))|\theta\rangle = \frac{1}{VT}\int d^4x\, \langle\theta|(F, \tilde{F})|\theta\rangle. \tag{3.57}$$

Thus, by Eq. (3.47),

$$\begin{aligned}\langle\theta|(F, \tilde{F})|\theta\rangle &= \frac{32\pi^2 \int [dA]\nu e^{-S} e^{i\nu\theta}}{VT \int [dA] e^{-S} e^{i\nu\theta}} \\ &= -\frac{32\pi^2 i}{VT}\frac{d}{d\theta}\ln\left(\int [dA] e^{-S} e^{i\nu\theta}\right).\end{aligned} \tag{3.58}$$

Hence there is no need to do a fresh summation over a dilute instanton–anti-instanton gas, since we have just evaluated the quantity in parentheses in Eq. (3.55). Thus in our approximation,

$$\langle\theta|(F, \tilde{F})|\theta\rangle = -64\pi^2 i K e^{-S_0} \sin\theta. \tag{3.59}$$

Some comments:

(1) The expectation value is independent of V and T, as it should be.

(2) The expectation value is an imaginary number, again as it should be. The reason is that

$$(F, \tilde{F}) = (F_{12}, F_{34}) + \text{permutations}. \tag{3.60}$$

When we continue from Euclidean space to Minkowski space, F_{12} remains F_{12}, but, just as x_4 becomes ix_0, so does F_{34} become iF_{30}. Thus, if we had obtained a real answer, we would have found that in Minkowski space (the real world) a Hermitian operator would have had an imaginary vacuum expectation value, a disaster.

(3) Both the vacuum energy density and the vacuum expectation value depend non-trivially on θ. Thus the θ-vacua are indeed all different from each other.

3.5 Instantons: particulars

$$\begin{aligned}\int d^4x(F, F) &= \left[\int d^4x(F, F)\int d^4x(\tilde{F}, \tilde{F})\right]^{\frac{1}{2}} \\ &\geq \left|\int d^4x(F, \tilde{F})\right|,\end{aligned} \tag{3.61}$$

by the Schwartz inequality. Thus, for any winding number, we have an absolute lower bound on the action,

$$S \geq \frac{8\pi^2}{g^2}|\nu|. \tag{3.62}$$

Furthermore, equality is attained if and only if

$$F = \pm\tilde{F}, \tag{3.63}$$

where the positive (negative) sign holds for positive (negative) ν.

This inequality was first derived by Belavin, Polyakov, Schwartz, and Tyupkin,[9] who used it to search for instantons. Their idea was to look for solutions of Eq. (3.63). If such solutions exist, they are minima of the action for fixed winding number, and thus stationary points of the action under local variations, that is to say, solutions of the field equations. Furthermore, since they have lower action than any other solutions of the same winding number (if other solutions exist), they dominate the functional integral, and, for our purposes, are the only solutions we need worry about. Finally, as a bonus, Eq. (3.63) is a first-order differential equation and considerably more tractable than the second-order field equations.

Let us begin the search with $\nu = 1$. We know that any field configuration with $\nu = 1$ can be gauge-transformed such that

$$A_\mu = g^{(1)}\partial_\mu[g^{(1)}]^{-1} + O(1/r^2), \tag{3.64}$$

where

$$g^{(1)} = \frac{x_4 + i\mathbf{x}\cdot\sigma}{r}. \tag{3.65}$$

Equation (3.64) is rotationally invariant, in the sense that the effect of any four-dimensional rotation can be undone by an appropriate gauge transformation. This is a consequence of the statement that a rotation is a continuous deformation and thus does not change the winding number. There is also a short direct proof: Under a general rotation

$$g^{(1)} \to g g^{(1)} h^{-1}, \tag{3.66}$$

where g and h are elements of SU(2) determined by the rotation. (This is a standard formula; it is the usual way of demonstrating the isomorphism between SO(4) and SU(2)⊗SU(2).) Thus,

$$A_\mu \to g A_\mu g^{-1} + O(1/r^2). \tag{3.67}$$

This, as promised, can be undone by a gauge transformation, indeed, by a gauge transformation of the first kind, a constant gauge transformation.

This suggests that we search for a solution of Eq. (3.63) that is rotationally invariant in the same sense. That is to say, we make the Ansatz,

$$A_\mu = f(r^2) g^{(1)} \partial_\mu [g^{(1)}]^{-1}, \tag{3.68}$$

where, to avoid a singularity, f must vanish at the origin. From here on it is straightforward plug-in-and-crank, which I will spare you. It turns out that we do indeed obtain a solution in this way, if

$$f = \frac{r^2}{r^2 + \rho^2}, \tag{3.69}$$

where ρ is an arbitrary constant, called 'the size of the instanton'. The existence of solutions of arbitrary sizes is a necessary consequence of the scale invariance of the classical field theory. (This fact will occasion some embarrassment shortly.)

Once we have a solution to any field theory, we can obtain new solutions by applying the invariances of the theory. In the case at hand, these are generated by (1) scale transformations, (2) rotations, (3) the four-parameter group of spatial translations, (4) the four-parameter group of special conformal transformations, and (5) gauge transformations. Scale transformations simply change the size of the instanton; thus they just shift around the members of our one-parameter family of solutions but generate no new solutions. Rotations, as I have shown, can always be undone by gauge transformations. Spatial translations generate genuinely new solutions, and give us four more parameters, the 'location of the center of the instanton'. Although I do not have time to demonstrate it here, it turns out[18] that special conformal transformations can be undone by gauge transformations and translations.

Gauge transformations, as usual, require special consideration. It is easy to see that any non-trivial gauge transformation changes (3.68). Because $g^{(1)}$ is a function of angles only, the radial component of A_μ, A_r, vanishes. Thus, under a general non-singular gauge transformation, $g(x)$,

$$A_r \to g A_r g^{-1} + g \partial_r g^{-1} = g \partial_r g^{-1}. \tag{3.70}$$

Hence, if the gauge transformation is not to change A_μ, g must be independent of r. That is to say, its value everywhere must be its value at the origin; g must be a constant gauge transformation. But the only constant

gauge transformation that leaves A_μ unchanged is the identity. (Remember, the effect of a constant gauge transformation is the same as that of a rotation.)

You might think that this discussion of gauge transformations is irrelevant. After all, when we do the quantum theory, we must work in a fixed gauge, such as axial gauge, and it is commonly said that once we have fixed the gauge we have no freedom to make gauge transformations. However, although commonly said, this is not strictly true; all standard gauges still allow constant gauge transformations.[19] This is as it should be. Constant gauge transformations act like ordinary symmetries; they put particles into multiplets (if there is no spontaneous symmetry breakdown), impose selection rules on scattering processes, etc. Thus, in a sensible formulation of the theory, they should remain as manifest symmetries of the Hamiltonian. Whether you accept this philosophy or not, the fact remains that constant gauge transformation applied to an instanton solution (transformed to obey the gauge conditions) will generate a different solution still obeying the gauge conditions. Thus we have found an eight-parameter family of solutions, one parameter from scale transformations, four from translations, and three from constant gauge transformations.

Are there other solutions with unit winding number? Atiyah and Ward[20] state that there are none. I can not give their proof here because I do not understand it. Nevertheless, mathematicians I trust say that their argument is not only legitimate but brilliant, so let us assume they are right and continue.

Solutions of higher winding number (if they exist) are of no interest to us. We have used approximate solutions consisting of n widely separated objects (instantons or anti-instantons) to evaluate the functional integral. These approximate solutions depend on $8n$ parameters, 8 for each object. Now suppose there are exact solutions that can be interpreted as n objects; that is to say, they depend on $8n$ (or fewer) parameters and become our approximate solutions when some of the parameters (the separations between the objects) become large. In this case, all we learn by knowing these exact solutions exist is that the dilute-gas approximation is better than we think it is – but we already know that it is good enough for our purposes. There might also be exact solutions that can not be interpreted in this way. To have a definite example, let me suppose there were a 'binstanton', a brand-new solution of winding number two. Then in evaluating the functional integral, we would have to sum over a dilute gas of instantons, anti-instantons, binstantons, and anti-binstantons. Thus, Eq. (3.56) would be replaced by

$$E(\theta)/V = -2K\cos\theta\,\mathrm{e}^{-S_0} - 2K'\cos 2\theta\,\mathrm{e}^{-S_0'}, \qquad (3.71)$$

where the primed quantities are the action and determinantal factor for a binstanton. But S_0' is twice S_0, so the new term is exponentially small compared to the old one and should be neglected.[21]

3.6 *The evaluation of the determinant and an infrared embarrassment*

We now know enough to go a long way towards explicitly evaluating the right-hand side of Eq. (3.56).

(1) S_0 is $8\pi^2/g^2$.

(2) We have an eight-parameter family of solutions and thus eight eigenmodes of eigenvalue zero in the small-vibration problem. Thus K contains a factor of $(1/\hbar^{\frac{1}{2}})^8$, or, equivalently $1/g^8$. Everything else in K is independent of $\hbar$, and thus independent of g.

(3) We have already done the integral over instanton location. The integral over constant gauge transformations is an integral over a compact group and thus gives only a constant numerical factor, the volume of SU(2). The integral over instanton sizes is potentially troublesome, since ρ can be anywhere between zero and infinity, so we will, for the moment, keep it as an explicit integral.

(4) Thus we obtain

$$E(\theta)/V = -\cos\theta\,\mathrm{e}^{-8\pi^2/g^2}g^{-8}\int_0^\infty \frac{\mathrm{d}\rho}{\rho^5}\,f(\rho M), \qquad (3.72)$$

where f is an unknown function and M is the arbitrary mass (more properly, arbitrary inverse wavelength) that is needed to define the renormalization prescription in a massless field theory. (I have avoided mentioning renormalization until now, but renormalization is essential in any computation that involves an infinite number of eigenmodes, as does this one. In Sect. 5 I will give a more detailed discussion of the ultraviolet divergences in determinantal factors and their removal by the usual one-loop renormalization counterterms.) The form of the integral is determined by dimensional analysis; an energy density has dimensions of $1/(\text{length})^4$.

(5) However, M and g are not independent parameters. Renormalization-group analysis[22] tells us that they must enter expressions for observable quantities only in the combination

$$\frac{1}{g^2} - \beta_1 \ln M + O(g^2), \qquad (3.73)$$

where β_1 is a coefficient which can be computed from one-loop perturbation theory. In the case at hand, β_1 is $11/12\pi^2$.

(6) This fixes the form of f. Thus,

$$E(\theta)/V = -A\cos\theta\,\mathrm{e}^{-8\pi^2/g^2}g^{-8}\int_0^\infty \frac{\mathrm{d}\rho}{\rho^5}(\rho M)^{8\pi^2\beta_1}[1+O(g^2)], \qquad (3.74)$$

where A is a constant independent of g, ρ, and M.

(7) To determine A requires a lot of hard work,[23] so I shall stop the calculation here. Even though we have not been able to carry things out to the end, it is remarkable how far we have been able to go with so little effort.

No doubt you have noticed that the integral we have derived is infrared-divergent. The origin of the divergence is clear from the derivation of the integral: the effective coupling constant (in the sense of the renormalization group) becomes large for large instantons, and this makes the integrand blow up. Thus the divergence is an embarrassment but not a catastrophe. It would be a catastrophe if we obtained a divergent answer in a regime in which we trusted our approximations. This is not the situation here; the divergence arises in the regime of large effective coupling constant, where all small-coupling approximations are certainly wrong. Phrased another way, the fact that the integrand has the wrong behavior for large ρ is overshadowed by the fact that it is the wrong integrand. Thus we are free to hope that strong-coupling effects (which we can not at the moment compute) introduce some sort of effective infrared cutoff in the integrand. This hope might be wrong, but it is not ruled out by anything we have done so far.

I admit that this argument is blatant hand-waving. However, it is not some new hand-waving special to instanton calculations, but the same old hand-waving that accompanies any discussion of the large-scale behavior of non-Abelian gauge field theories. For example, there is evidence that the observed hadrons are made of weakly coupled quarks. But if the quarks are weakly coupled, why can we not knock them out of the hadron? Well, in a gauge field theory the effective coupling constant grows at large distances, etc., much hand-waving, infrared slavery and quark confinement.

Everything that we have done for SU(2) can be extended straightforwardly to SU(3). To begin with, an SU(2) instanton solution can trivially be made into an SU(3) instanton solution; all that needs to be done is to say that three of the gauge fields, those associated with an

SU(2) subgroup, are of the form given, while the other five vanish. It is believed that these exhaust the set of solutions of Eq. (3.63) with unit winding number, although, unlike the SU(2) case, there is, to my knowledge, no rigorous proof of this statement. If this is indeed the case, there are only two minor differences between the SU(3) computation and the SU(2) one. (1) Instead of three parameters associated with constant gauge transformations, we have seven. (One of the eight SU(3) generators commutes with the SU(2) subgroup and does not change the solution.) Thus the factor of g^{-8} in Eq. (3.74) is replaced by one of g^{-12}. (2) β_1 has the proper value for an SU(3) gauge theory, $11/8\pi^2$.

4 The Abelian Higgs model in 1+1 dimensions[24]

In this section I will discuss a field theory in which instanton effects drastically change the particle spectrum, the Abelian Higgs model in two-dimensional space-time.

In any number of dimensions, this is the theory of a complex scalar field with quartic self-interactions, minimally coupled to an Abelian gauge field with gauge coupling constant e, called the electric charge. In our notation, the theory is defined by the Euclidean Lagrangian density,

$$\mathscr{L}=\frac{1}{4e^2}(F,F)+D_\mu\psi^*D_\mu\psi+\frac{\lambda}{4}(\psi^*\psi)^2+\frac{\mu^2}{2}\psi^*\psi, \tag{4.1}$$

where λ is a positive number and μ^2 may be either positive or negative. To this must be added renormalization counterterms; however, renormalization will play no part in our computations, and, to keep things as simple as possible, I will not distinguish between bare and renormalized parameters.

Perturbation theory tells us that for weak coupling the qualitative properties of the theory depend critically on the sign of μ^2:

(1) If μ^2 is positive, the theory is simply the electrodynamics of a charged scalar meson. The mass spectrum consists of the charged meson, its antiparticle, and a massless vector meson, the photon. The force between widely separated external charges is the ordinary Coulomb force. These statements require some modification in two dimensions. Firstly, because there are no transverse directions, there is no photon. Secondly, because the Coulomb force is independent of distance, it is impossible to separate a meson and an antimeson; in contemporary argot, the charged particles are confined. The spectrum of the theory consists of a sequence of meson–antimeson bound states, rather like the spectrum of positronium, except that these states are all stable, since they can not decay through the emission of (nonexistent) photons.

(2) If μ^2 is negative, the Higgs phenomenon takes place. In the ground state of the theory,

$$|\langle\psi\rangle|^2=-\mu^2/\lambda\equiv a^2. \tag{4.2}$$

The particle spectrum consists of a massive neutral scalar meson and a massive neutral vector meson. The force between widely separated external charges falls off exponentially rapidly. These statements require no modification in two dimensions.

In the remainder of this section, I will argue that the preceding sentence is a lie; contrary to the predictions of perturbation theory, the qualitative properties of the model for negative μ^2 are the same as those for positive μ^2; the two-dimensional Abelian Higgs model does not display the Higgs phenomenon. To be precise, I will show that, for negative μ^2, the theory admits instantons, and, when the effects of these instantons are taken into account, the long-range force between external charges is independent of their separations. Also, I will be able to argue, from the behavior of the long-range force, that the theory contains (confined) charged particles. There is a quantitative difference between positive and negative μ^2, though: for positive μ^2, the strength of the long-range force is independent of $\hbar$; for negative μ^2, the strength of the long-range force is exponentially small in $\hbar$, the mark of an instanton effect.

Just as in Sect. 3, we must begin the analysis by classifying classical field configurations of finite action. Of course, before doing this, we must add a constant to the Lagrangian density so the minimum of the action is zero. Thus we write

$$\mathscr{L}=\frac{1}{4e^2}(F,F)+|D_\mu\psi|^2+\frac{\lambda}{4}(|\psi|^2-a^2)^2. \tag{4.3}$$

This is the sum of three positive terms. In order that the third term not make a divergent contribution to the action, it is necessary that $|\psi|$ approach a as r goes to infinity. However, there is no restriction on the phase of ψ. In equations,

$$\lim_{r\to\infty}\psi(r,\theta)=g(\theta)a, \tag{4.4}$$

where g is a complex number of unit modulus, an element of U(1). In order that the second term not make a divergent contribution to the action, it is necessary that

$$A_\mu=g\partial_\mu g^{-1}+O(1/r^2). \tag{4.5}$$