6

Classical lumps and their quantum descendants

(1975)

1 Introduction

A stone thrown into a still body of water makes ripples that spread out and eventually die away. The stone disturbs the water, gives it energy, but, even if we ignore friction, this energy tends in the course of time to spread out over the water. If we imagine the water to be infinite in extent then, if we wait long enough, at no point is the water appreciably different from its state before the stone was cast. The disturbance dissipates.

This concept of dissipation can be generalized beyond the hydrodynamic example I have given. Consider a classical field theory, with energy density Θ_{00}, such that Θ_{00} is always greater than or equal to zero and is everywhere zero for the ground state(s) of the theory. Then we will say that a solution of the classical equations of motion is dissipative if

$$\lim_{t\to\infty} \max_{\mathbf{x}} \Theta_{00}(\mathbf{x}, t) = 0. \tag{1.1}$$

Most of the simple field theories with which we are familiar have the property that all of their non-singular solutions of finite total energy are dissipative. This is the case for Maxwell's equations, the Klein–Gordon equation, etc. However, there are some perfectly sensible classical field theories that possess non-singular non-dissipative solutions of finite energy. Among these are some (though not all) of the spontaneously broken gauge theories that have risen to such prominence in recent years.

In the most striking and simplest case, and the one that will occupy most of our attention, these solutions may be time-independent, lumps of energy holding themselves together by their own self-interaction. There are also more complicated kinds of non-dissipative solutions, solutions that are periodic in time or have even more exotic time dependences, oscillating or quivering lumps. Also, if our theory is Lorentz-invariant (or Galileo-invariant), if we can construct stationary lumps, we can also

construct steadily moving lumps, and then it is reasonable to imagine that we can construct solutions that correspond to several widely separated lumps, all moving away from each other.

(I should tell you that the concept of 'lump' which I have vaguely defined here is called 'soliton' in much of the recent literature. I avoid the word 'soliton' here because it has a very precise and narrow definition in applied mathematics, and most of the lumps we will talk about are not solitons in this narrow sense.)[1]

Once one knows that non-dissipative solutions exist, many obvious questions arise. Which classical theories possess non-dissipative solutions? Time-independent non-dissipative solutions? What are the properties of these solutions? What about the quantum theories obtained from these classical theories by canonical quantization? Do the classical lumps become particles in the quantum theory? If so, how do we compute the properties of these particles?

No one knows the complete answer to any of these questions. However, much is known, and much of what we know, especially about the quantum problem, has been learned within the last year. These lectures are an attempt to survey this knowledge.

Section 2 deals with a set of classical field theories so simple that we can find lumps by quadrature. These are theories of a single scalar field in a two-dimensional space-time. Here we shall meet such famous lumps as the kink of ϕ^4 theory and the soliton of the sine–Gordon equation.

Unfortunately, direct integration of the equations of motion is not a very attractive method for more than a single field or more than two space-time dimensions. Section 3 discusses the powerful method of topological conservation laws, which enables us to prove the existence of non-dissipative solutions without getting tangled up in the detailed structure of the equations of motion. In the course of this section, we shall meet some more famous lumps: the flux lines of superconductivity (lumps in two spatial dimensions), and the lumps in three spatial dimensions discovered independently last year by Polyakov and 't Hooft, and called hedgehogs by the first and monopoles by the second.

(A warning: when I delivered these lectures at Erice, many of the students found the material in Sect. 3 the most difficult part of the lectures. If you have trouble with Sect. 3, *skip it*; nothing in the remainder of these notes depends on it.)

In Sect. 4, we turn to the quantum problem. I discuss in some detail two approximation methods for treating time-independent classical solutions, a systematic weak-coupling expansion and a variational

method. I also discuss much more briefly a method for treating periodic classical solutions, based on WKB ideas.

Section 5 deals with a very special system, the quantum sine–Gordon equation, for which it is possible to gain some information without recourse to approximations. Everything we shall learn about this system that falls within the expected domain of validity of the approximation methods of the previous section is consistent with them. However, we shall also discover some surprises.

Much of what I know about this subject has been learned from Ludwig Faddeev, Sheldon Glashow, Jeffrey Goldstone, Roman Jackiw, T. D. Lee, André Neveu, and Gian-Carlo Wick, all of whom I gratefully acknowledge.

2 Simple examples and their properties[2]

2.1 *Some time-independent lumps in one space dimension*

(Notation: The signature of the metric tensor is such that a time-like vector has positive norm. x always denotes a space-time point, with time and space components x^0 and $\mathbf{x}$. An exception is two-dimensional space-time, where x is used indiscriminately for a space-time point and its space component; which is meant should always be clear from the context. Greek indices run over all space-time coordinates, latin indices from the middle of the alphabet over space coordinates only, latin indices from the beginning of the alphabet over internal-symmetry variables. Summation over repeated indices is always implied unless otherwise stated. $\hbar = c = 1$.)

Let us consider the simplest family of classical relativistic field theories: theories of a single scalar field in one space and one time dimension, with non-derivative interactions. The dynamics of such a theory are described by the Lagrange density,

$$\mathscr{L} = \tfrac{1}{2}\partial_\mu \phi \, \partial^\mu \phi - U(\phi), \tag{2.1}$$

where μ equals 0 and 1, and U is some arbitrary function of ϕ, but not of its derivatives. The energy of any field configuration is given by

$$E = \int \mathrm{d}x [\tfrac{1}{2}(\partial_0 \phi)^2 + \tfrac{1}{2}(\partial_1 \phi)^2 + U]. \tag{2.2}$$

We will sometimes have occasion to write this as the sum of kinetic and potential energy,

$$E = T + V, \tag{2.3}$$

where T, as usual, is the term quadratic in time derivatives (the first term in Eq. (2.2)), and V, also as usual, is the term involving no time derivatives (the sum of the second two terms).

If the energy is to be bounded below, U must be bounded below. In this case, we can always add a constant to U such that the minimum value of U is zero; I will assume in what follows that this has always been done. It is evident from Eq. (2.2) that a state of minimum energy (a ground state) is one where ϕ is independent of space and time and equal to one of the zeros of U. Of course, if U has several zeros, the theory has several (degenerate) ground states. The energy of a ground state is zero.

There are two theories which I will use throughout these lectures as examples. One is familiar ϕ^4 theory, with the sign of the quadratic term in U such that there are two ground states. (This is the situation that, in the quantum theory, is called 'spontaneous breakdown of symmetry'.) In equations,

$$U=\frac{\lambda}{2}\phi^4-\mu^2\phi^2+\frac{\mu^4}{2\lambda}, \tag{2.4a}$$

with λ and μ^2 positive parameters. This can also be written as

$$U=\frac{\lambda}{2}(\phi^2-a^2)^2,$$

where a^2 is μ^2/λ. The two ground states are $\phi=\pm a$.

The other example is somewhat less familiar. It is the so-called[3] sine–Gordon equation:

$$U=\frac{\alpha}{\beta^2}(1-\cos\beta\phi), \tag{2.4b}$$

where α and β are real parameters. Since we are free to redefine the sign of ϕ, we can with no loss of generality take β to be positive. Likewise, since we are free to shift ϕ by π/β (and add a constant to U), we can also take α to be positive. The ground states are those for which ϕ is an integral multiple of $2\pi/\beta$. If we study small oscillations about one of these ground states, say $\phi=0$, we expand the cosine in power series:

$$U=\frac{\alpha}{2}\phi^2-\frac{\alpha\beta^2}{4!}\phi^4+\cdots.$$

Thus we see that, for small oscillations, α is something like the parameter μ^2 in the previous example, while $\alpha\beta^2$ is similar to λ.

Now let us turn to the problem of finding possible time-independent solutions of the equations of motion. For time-independent solutions, Hamilton's principle reduces to

$$\delta V=\delta\int \mathrm{d}x[\tfrac{1}{2}(\partial_1\phi)^2+U(\phi)]=0. \tag{2.5}$$

This is mathematically identical to a familiar problem in particle mechan-

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

ics, the motion of a particle of unit mass in a potential equal to *minus* $U(x)$,

$$\delta \int dtL = \delta \int dt[\tfrac{1}{2}(dx/dt)^2 + U(x)] = 0. \tag{2.6}$$

Fig. 1 shows the potential for this equivalent mechanical problem in the case of ϕ^4 theory.

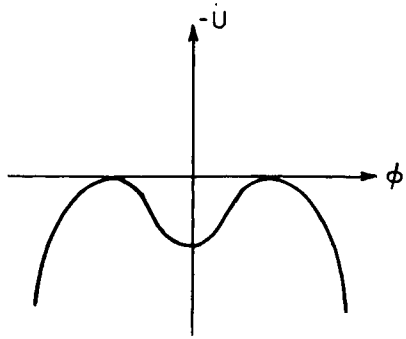
Every motion of the particle in the potential corresponds to a time-independent solution of the field equations. However, most of these solutions are not of finite energy. For the energy integral (2.2) to converge, it is necessary for ϕ to go to a zero of U as x goes to either plus or minus infinity. In terms of the particle problem of Fig. 1, this means that the particle must go to the zeros of the potential as t goes to plus or minus infinity. Of course, the ground states (motions where the particle stays forever at the top of one of the two potential hills) satisfy this criterion, but there are also less trivial motions that will do the job, motions where the particle starts on top of one hill and moves to the top of the other. These are motions of zero particle energy. If the particle energy were negative, the particle would never reach the top of the hill; if it were positive, the particle would overshoot and travel on forever. This analysis is in no way special to the ϕ^4 potential sketched in Fig. 1. It only depends on the fact that U has two adjacent zeros between which the particle can move; the detailed form of U between the zeros is irrelevant, as is the behavior of U outside the region between the zeros.

Thus we find:

(1) If U has only one zero – that is to say, if the ground state of the theory is unique – there are no non-trivial time-independent solutions of finite energy.

(2) On the other hand, if U has more than one zero – that is to say, if the theory has more than one ground state – there always exist non-

Fig. 1



https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

trivial time-independent solutions of finite energy. For such solutions, ϕ moves monotonically from one zero of U at x equals minus infinity to an adjacent zero of U at x equals plus infinity. By convention, we will call those solutions for which ϕ is monotone increasing 'lumps' and those for which it is monotone decreasing 'antilumps'.

(3) It is easy to reduce the problem of finding the explicit form of these solutions to quadratures. The particle motion is one of zero particle energy; hence, since the particle moves in the potential $-U(x)$,

$$\frac{1}{2}\left(\frac{dx}{dt}\right)^2 - U(x) = 0. \tag{2.7}$$

Translating this back into field language, we find

$$\tfrac{1}{2}(\partial_1\phi)^2 = U(\phi). \tag{2.8}$$

This has the solution

$$x = \pm \int_{\phi_0}^{\phi} d\phi\, [2U(\phi')]^{-\frac{1}{2}}. \tag{2.9}$$

Here the plus sign is for lumps, the minus sign for antilumps and the arbitrary parameter ϕ_0 is the value of ϕ at $x=0$, and can be any number between the two adjacent zeros.

Of course, the presence of the arbitrary parameter ϕ_0 is just an expression of spatial translation invariance. Another way of saying the same thing is to say that, if $f(x)$ is any solution of (2.9) with some fixed ϕ_0, then the general solution with arbitrary ϕ_0 is

$$\phi = f(x-b), \tag{2.10}$$

where b is arbitrary. In other words, the center of a lump can be anywhere.

(4) These formulae also enable us to simplify somewhat the expression for the energy of a lump. In particular, it follows from Eq. (2.8) that

$$\begin{aligned} E &= \int dx[\tfrac{1}{2}(\partial_1\phi)^2 + U] \\ &= \int dx(\partial_1\phi)^2 \\ &= \int d\phi[2U(\phi)]^{\frac{1}{2}}. \end{aligned} \tag{2.11}$$

For both of our specific examples, these integrals are elementary. Just to have some concrete examples in front of us, I will write out the answers. For ϕ^4 theory, we find for the form of the lump

$$\phi = a \tanh(\mu x). \tag{2.12a}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

The energy of the lump is given by

$$E = 4\mu^3/3\lambda. \tag{2.13a}$$

The lumps of ϕ^4 theory are frequently called 'kinks' in the literature.

For the sine–Gordon theory, we find

$$\phi = \frac{4}{\beta}\tan^{-1}\exp(\alpha^{\frac{1}{2}}x), \tag{2.12b}$$

where which branch of the inverse tangent we choose determines which two zeros we move between. The energy of the lump is

$$E = \frac{8\alpha^{\frac{1}{2}}}{\beta^2}. \tag{2.13b}$$

The lumps of sine–Gordon theory are frequently called 'solitons' in the literature. In both cases the antilumps are obtained simply by multiplying by minus one.

2.2 *Small oscillations and stability*

We have shown that any theory of a single scalar field in one space dimension with more than one ground state admits time-independent solutions of finite energy, but we have not yet shown that these solutions are stable under small perturbations. I now turn to this problem.

The equation of motion for our system is

$$\Box^2\phi + U'(\phi) = 0, \tag{2.14}$$

where the prime denotes differentiation with respect to ϕ. Let us consider a solution of the form

$$\phi(x, t) = f(x) + \delta(x, t), \tag{2.15}$$

where $f(x)$ is our time-independent solution and δ is the small perturbation. Inserting this expression in the equation of motion and only retaining terms of first order in the perturbation, we find

$$\Box^2\delta + U''(f)\delta = 0. \tag{2.16}$$

This equation is invariant under time translations, so we can express a general small perturbation as a superposition of normal modes. That is to say, the general solution is of the form

$$\delta(x, t) = \mathrm{Re}\sum_n a_n e^{i\omega_n t}\psi_n(x), \tag{2.17}$$

where the as are arbitrary complex coefficients, and the ψs and ωs obey the equation

$$-\frac{d^2\psi_n}{dx^2} + U''(f)\psi_n = \omega_n^2\psi_n. \tag{2.18}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

Note that this is a one-dimensional Schrödinger equation, with potential $U''(f)$. Our solution is stable under small perturbations if and only if none of the energy eigenvalues of this Schrödinger equation are negative. I will now show that this is always the case.

Spatial translation invariance tells us that if $f(x)$ is a solution of the equation of motion, so is $f(x+a)$. Thus we already know one energy eigenfunction of our Schrödinger equation,

$$\psi_0(x) = \mathrm{d}f/\mathrm{d}x. \tag{2.19}$$

We also know the associated eigenvalue; it is zero. In Sect. 2.1 I showed that f was always a monotone function of x; therefore ψ_0 has no nodes. It is a well-known theorem that for a one-dimensional Schrödinger equation with arbitrary potential the eigenfunction with no nodes is the eigenfunction of lowest energy.[4] Q.E.D.

This takes care of the question of stability, but I should like to make some additional remarks about Eq. (2.18). If we could explicitly solve this equation, we could explicitly obtain (in the linearized approximation) a rich set of solutions to our equations of motion. These solutions are especially interesting if Eq. (2.18) possesses a discrete eigenvalue (other than the trivial zero eigenvalue). In this case, the associated eigenfunction yields a solution to the (linearized) equation of motion which is of finite energy and periodic in time. If we are daring enough to use quantum language before we get to the quantum theory, we can think of this situation as a meson bound to a lump. On the other hand, from continuum eigenfunctions, we can only form a solution of finite energy by forming a wave packet. In the same spirit, we can think of this wave packet as representing a meson scattering off a lump. (When we do get to the quantum theory, we shall see that these are not bad descriptions.)

2.3 *Lumps are like particles (almost)*

Although we are working with classical continuum theories, the objects we have found are much like classical particles (or, better yet, since the energy density is not concentrated at a point, classical extended bodies). They have a definite finite rest mass and a definite location, the location of the center-of-mass. Furthermore, from the stationary solutions we have found, we may construct, by Lorentz transformation, solutions corresponding to lumps moving with any velocity less than that of light. By Lorentz invariance, these moving solutions obey the traditional energy–momentum relation,

$$E = (P^2 + M^2)^{\frac{1}{2}}, \tag{2.20}$$

where M is the rest mass.

These are all properties which we would expect to hold for more traditional extended bodies (e.g. billiard balls) in relativistic classical mechanics. However, there is one important sense in which lumps differ from classical particles.

One classical particle at rest is always a solution of classical mechanics. From this solution one can build an approximate many-particle solution that consists of many particles at rest, all widely separated from each other. I say 'approximate' because there may be interactions between the particles (e.g. Yukawa forces); the solution becomes exact only as the separations go to infinity. Can we construct the same sort of asymptotic many-particle solutions for lumps?

To phrase the problem more precisely, let a_i be a set of points on the line ($i = 1 \ldots n$), arranged in increasing order,

$$a_1 < a_2 \ldots < a_n. \tag{2.21}$$

Let $f_i(x)$ be a set of time-independent solutions of the field equation. Can we find an approximate solution of the field equation, such that, in the neighborhood of a_i,

$$\phi \approx f_i(x - a_i), \tag{2.22}$$

and such that the solution becomes exact as the distance between successive as goes to infinity?

The answer is obviously *no*, unless

$$f_i(\infty) = f_{i+1}(-\infty). \tag{2.23}$$

This simple condition for patching adjacent solutions together has far-reaching consequences for any attempt to interpret lumps as particles.

To show this, let me begin with the sine–Gordon equation. For f_1, we may choose any solution, soliton or antisoliton. For f_2, we may again choose either soliton or antisoliton, but which soliton or antisoliton we choose (i.e. which branch of the tangent) is determined by our previous choice, and so on down the line. This strongly suggests that properly we should not say that there are many solitons and many antisolitons, corresponding to the many branches of the tangent, for this picture would lead to a gross overcounting of the many-soliton states. We get the right counting if we say instead that there is only one soliton and only one antisoliton. If we adopt this position, we must also say that the only observable functions of ϕ are those that are unchanged by the transformation

$$\phi \to \phi + 2\pi/\beta. \tag{2.24}$$

In ϕ^4 theory, the situation is even more peculiar. If we begin with a kink, we do not have the option to follow it with either a kink or an anti-

kink; we must follow it with an antikink. Likewise, the next lump along the line must be a kink, etc. Just as above, this strongly suggests that we should not consider the kink and antikink as independent objects, but as the same object. (This observation is due to Goldstone and Jackiw.)[2] If we adopt this position, we must also say that the only observable functions of ϕ are those that are unchanged by the transformation

$$\phi \to -\phi. \tag{2.25}$$

I should emphasize that these are just questions of interpretation, of what words we use to describe the mathematically unambiguous properties of well-defined classical field theories. We could just as well say that the kink and the antikink are different, and that there are somewhat peculiar forces between two kinks that do not fall off with distance, but which are shielded by an intervening antikink. This is not so strange in in one spatial dimension as it would be in three; after all, in one dimension, even the familiar Coulomb force is independent of distance.

In Sect. 3 we will return to the question of patching together distant lumps, for gauge theories in two and three spatial dimensions. We will find the situation here more closely resembles the assembly of distant particles, but still there will be some novelties.

2.4 *More dimensions and a discouraging theorem*

The obvious next step is to attempt to find time-independent solutions for scalar field theories in more than one spatial dimension, and perhaps also with more than one scalar field. Unfortunately, this step leads us into a dead end, for such theories possess no such solutions.

Derrick's theorem. Let ϕ be a set of scalar fields (assembled into a big vector) in one time dimension and D space dimensions. Let the dynamics of these fields be defined by

$$\mathscr{L} = \tfrac{1}{2}\partial_\mu \phi \cdot \partial^\mu \phi - U(\phi), \tag{2.26}$$

and let U be non-negative and equal to zero for the ground state(s) of the theory. Then, for $D \geqslant 2$, the only non-singular time-independent solutions of finite energy are the ground states.[5]

Proof. Define

$$V_1 = \tfrac{1}{2} \int \mathrm{d}^D\mathbf{x} (\nabla\phi)^2, \tag{2.27}$$

$$V_2 = \int \mathrm{d}^D\mathbf{x}\, U(\phi). \tag{2.28}$$

V_1 and V_2 are both non-negative and are simultaneously equal to zero only for the ground states. Let $\phi(\mathbf{x})$ denote a time-independent solution.

Consider the one-parameter family of field configurations defined by

$$\phi(\mathbf{x};\lambda)\equiv\phi(\lambda\mathbf{x}), \tag{2.29}$$

where λ is a positive number. For this family, the energy is given by

$$V(\lambda)=\lambda^{(2-D)}V_1+\lambda^{-D}V_2. \tag{2.30}$$

By Hamilton's principle, this must be stationary at $\lambda=1$. Thus,

$$(D-2)V_1+DV_2=0. \tag{2.31}$$

For $D>2$, this immediately implies that both V_1 and V_2 vanish, and the proof is complete. For $D=2$, however, Eq. (2.31) only implies the vanishing of V_2, and a small amount of further argument is required. If V_2 vanishes it is stationary, since zero is its minimum value. Thus we may apply Hamilton's principle to V_1 alone, from which it trivially follows that V_1 also vanishes. Q.E.D.

I should emphasize that Derrick's theorem denies the existence of time-independent solutions *only*; it says nothing about time-dependent (but non-dissipative) solutions. Indeed, T. D. Lee[6] has constructed models involving scalar fields only which possess such solutions; I report on his construction in Appendix 1.

Nevertheless, the theorem is sufficiently discouraging to make us investigate theories of more than just scalar fields. The obvious next step is to look at theories of scalar fields and gauge fields.[7] This is not only the next step, it is almost as far as we can go, at least in three dimensions. Anything much more complicated would be non-renormalizable; this is no problem in classical physics, but it is a considerable problem if we want to go on to the quantum theory eventually.

Of course, we could consider fermions. However, I remind you that the classical limit of a Fermi field is not an ordinary c-number but an anticommuting c-number. Thus, if we were to consider fermions, the values of our classical fields would not be numbers but elements of a Grassman algebra, a situation I would just as soon avoid in these lectures.

There is, of course, another reason for looking at gauge field theories: there is reason to believe that they have something to do with the real world. At any rate, it is gauge theories that will occupy us in the next section.

3 Topological conservation laws[8]

3.1 *The basic idea and the main results*

In the previous section we demonstrated the existence of non-dissipative solutions for a family of simple field theories by explicitly displaying time-independent solutions. I would now like to establish

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

exactly the same result by a more indirect line of argument. This would be a bootless exercise were it not that the indirect argument is generalizable to cases where explicit solutions are hard to come by, in particular to theories of gauge fields and scalar fields in three dimensions.

Instead of focusing on non-singular solutions of finite energy, let us focus on non-singular initial-value data (ϕ and $\partial_0\phi$ at some fixed time) of finite energy. Of course, these two sets of objects are in one-to-one correspondence; for every solution, there is initial-value data, and for every set of initial-value data, there is a solution. (For the class of theories under consideration, non-singular initial-value data can not develop a singularity in the course of time, for this would violate the conservation of energy.[9] Mathematical purists can take 'non-singular' to mean 'continuously differentiable'.)

For such initial-value data, just as for the time-independent solutions of Sect. 2.1, finiteness of the energy implies that

$$\lim_{x\to\pm\infty} \phi(x, t)\equiv\phi(\pm\infty, t), \tag{3.1}$$

must exist and must be zeros of U. Now, non-singular solutions are, in particular, continuous in time. Since, for our examples, the zeros of U form a discrete set, this implies that

$$\partial_0\phi(\pm\infty, t)=0. \tag{3.2}$$

If U has more than one zero, this equation is non-trivial, and can be used to prove the existence of non-dissipative solutions.

To show how this works, let me take ϕ^4 theory as an example. I will show that any solution for which

$$\phi(\infty, t)=-\phi(-\infty, t), \tag{3.3}$$

is non-dissipative. Proof: By continuity in x, for any t, there must be some x for which $\phi=0$. At this point,

$$\Theta_{00}(x, t)\geqslant U(0)=\lambda a^4/2. \tag{3.4}$$

Hence,

$$\max_x \Theta_{00}(x, t)\geqslant\lambda a^4/2, \tag{3.5}$$

which contradicts Eq. (1.1).

A slightly more abstract way of expressing Eq. (3.2) is to say that we have divided the space of non-singular finite-energy solutions at a fixed time into subspaces, labeled by $\phi(\pm\infty, t)$. (Thus, for ϕ^4 theory, we have four such subspaces.) These subspaces are disconnected components of the whole space, in the normal topological sense; it is not possible to continuously change a solution in one component into a solution in another component. Since time evolution is continuous, this implies that if a

solution is in one component at any one time, it is in the same component at any other time.

This way of putting things emphasizes the difference between Eq. (3.2) and familiar conservation laws. Eq. (3.2) is a conservation law; just like the conservation of energy, it states that something is independent of time. However, unlike the conservation of energy, it is not a consequence of a symmetry of the theory. It is instead a consequence of a topological property of the space of non-singular finite-energy solutions, the fact that it is not connected. For this reason we shall call Eq. (3.2), and similar results we shall establish for more complicated theories, 'topological conservation laws'.

The remainder of this section will be devoted to the study of such topological conservation laws for spontaneously broken gauge field theories, in both two and three spatial dimensions. Since the analysis is lengthy, I will try to whet your appetite by stating some of the main results for three dimensions now.

(1) If the theory has no spontaneous symmetry breakdown, the space of non-singular finite-energy solutions has only one component, and there are no non-trivial topological conservation laws.

(2) The same disgusting situation obtains if the symmetry breakdown is total, that is to say, if no massless gauge mesons survive symmetry breakdown.

(3) We now come to the interesting case in which only one massless gauge meson (by convention, called the photon) survives symmetry breakdown. In this case there are two possibilities:

(3a) The gauge group when written as a product of simple Lie groups contains a U(1) factor, and the generator of this U(1) factor enters into the expression for the electric charge. (This is the case, for example, in the Weinberg–Salam model.) In this case, there are again no non-trivial topological conservation laws.

(3b) In all other cases (for example, in the model of Glashow and Georgi), the space of finite-energy non-singular solutions has an infinite number of components; there are non-trivial topological conservation laws. Just as above, these can be used to prove the existence of non-dissipative solutions. The components can be labeled by a quantity that can be interpreted physically as magnetic flux emerging from the solution. (This despite the fact that the theory contains no magnetic monopoles in the sense of Dirac, just ordinary charged and neutral scalar and vector fields!) This magnetic flux is quantized, always comes in integral multiples of a basic unit. (This despite the fact that these are classical, not quantum, theories!)

(4) Similar results hold if many massless gauge mesons survive symmetry breakdown, but I can not explain the results in detail at this early stage, because they involve the generalization of the concept of magnetic flux to non-Abelian theories. (This concept will be developed in full as our analysis proceeds, and the appropriate results stated precisely.)

Along with these appetizers, I should give you two warnings:

(1) Topological conservation laws enable us to establish the existence of non-dissipative solutions, *not* of time-independent solutions. For example, in ϕ^4 theory every set of initial conditions obeying Eq. (3.3) is non-dissipative, but, as we have seen in Sect. 2.1, only a small subset of these are time-independent. Another example: for the sine–Gordon equation it is possible to show, by arguments identical to those given above, that any solution for which

$$\phi(\infty, t)-\phi(-\infty, t)=4\pi/\beta, \tag{3.6}$$

is non-dissipative, but we know that none of these are time-independent. (For all time-independent solutions the field changes by $2\pi/\beta$, half this amount.)

(2) Topological conservation laws give sufficient conditions for non-dissipative solutions, *not* necessary ones. It is quite possible for there to be non-dissipative solutions even when there are no non-trivial topological conservation laws. (See Appendix 1 for an example.)

These warnings emphasize that topological conservation laws do not tell us everything we want to know. Nevertheless, they tell us quite a lot, and at a very low cost in analytic effort, and thus are very much worth pursuing.

3.2 *Gauge field theories revisited*

Before proceeding to our main task, I would like to briefly[10] review gauge field theories, and to develop some of their properties which will be important to us later.

The theories we shall consider will involve a set of n scalar fields (not necessarily real), which we assemble into an n-vector, ϕ. We also have a compact connected Lie group, G, called the gauge group, and an n-dimensional unitary representation of G, $D(g)$. Given any function from space-time to G, $g(x)$, we define a transformation of the scalar fields, called a gauge transformation, by

$$g(x)\colon \phi(x)\to D(g(x))\phi(x). \tag{3.7}$$

To avoid long equations with nested parentheses, we will identify the abstract group G with the representation $D(g)$, and write this as

$$g(x)\colon \phi(x)\to g(x)\phi(x). \tag{3.8}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

The generators of $D(g)$, a set of $n \times n$ orthonormal Hermitian matrices, are denoted by T^a. These obey the commutation relations

$$[T^a, T^b] = ic^{abc}T^c, \tag{3.9}$$

where the cs are the structure constants of G, and the sum over repeated indices is implied. With these generators we associate a set of real vector fields, A^a_μ, called the gauge fields, with gauge transformation properties given by

$$A^a_\mu(x)T^a \to g(x)A^a_\mu(x)T^a g(x)^{-1} + ie^{-1}\,\partial_\mu g(x) g(x)^{-1}, \tag{3.10}$$

where e is a real constant, called the gauge coupling constant. For an infinitesimal gauge transformation,

$$g(x) = 1 - iT^b\omega^b(x), \tag{3.11}$$

this becomes

$$\delta A^a_\mu = c^{abc}A^b_\mu \delta\omega^c + e^{-1}\,\partial_\mu \delta\omega^a. \tag{3.12}$$

Equation (3.12) is more common in the literature than Eq. (3.10), but they are equivalent, and (3.10) is better suited for our purposes. (If G is not simple, there may be several gauge coupling constants, but for notational simplicity I will ignore this possibility here. None of our conclusions will be affected by the presence of several coupling constants.)

From these objects, we define the covariant derivative of ϕ,

$$\mathrm{D}_\mu\phi = \partial_\mu\phi + ieA^a_\mu T^a\phi, \tag{3.13}$$

and the field strengths,

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - ec^{abc}A^b_\mu A^c_\nu. \tag{3.14}$$

These have simple gauge-transformation properties:

$$\mathrm{D}_\mu\phi(x) \to g(x)\mathrm{D}_\mu\phi(x), \tag{3.15}$$

and

$$F^a_{\mu\nu}(x)T^a \to g(x)F^a_{\mu\nu}(x)T^a g(x)^{-1}. \tag{3.16}$$

From these objects, we construct the gauge-invariant Lagrange density that defines our theory,

$$\mathscr{L} = -\tfrac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} + \mathrm{D}_\mu\phi^\dagger \cdot \mathrm{D}^\mu\phi - U(\phi), \tag{3.17}$$

where U is some invariant function of the scalar fields.[11] Just as before, we assume U is always greater than or equal to zero and equals zero for the ground state(s) of the theory. The simplest theory of this kind is the electrodynamics of charged scalar mesons (possibly with self-interactions). In this case, ϕ consists of a single complex field, and G is U(1), the group of complex numbers of unit modulus.

What about spontaneous symmetry breakdown? The ground states of the theory can easily be seen to all have energy zero; up to a gauge trans-

formation, they are states where the gauge fields vanish and ϕ is a constant and a zero of U. For such a ground state, defined by $\phi = \phi_0$, we define H, a subgroup of G, called the unbroken group, by saying that h is an element of H if and only if $h\phi_0 = \phi_0$. If we study small oscillations about this ground state, we find that the gauge fields associated with the generators of H remain massless, while the other gauge fields acquire masses.

Because U is G-invariant, if ϕ_0 is a zero of U, so is $g\phi_0$, for any g in G. I will now make an important simplifying assumption: I will assume that *all* the zeros of U are of the form $g\phi_0$ for some g in G. This assumption excludes both the case of accidental degeneracy, in which U has zeros that are not forced on it by any symmetry, and the case in which the theory has an ordinary (non-gauge) internal symmetry group in addition to the gauge group, G. Accidental degeneracy is of no real interest to us, because we are interested in classical theories only as limits of quantum theories, and, in general, quantum corrections remove accidental degeneracies. The case of an internal symmetry group is more interesting, and I exclude it only to simplify the discussion. The generalization of all of our results to this case is sketched in Appendix 3.

The simplifying assumption is equivalent to the statement that the set of zeros of U can be identified with the coset space G/H, and therefore I will denote this set by G/H in the sequel. (If you do not remember what a coset space is, do not worry; just accept 'G/H' as my eccentric shorthand for 'the set of zeros of U' and you will have no difficulties.)

This completes the lightning review of standard lore. I now turn to less familiar material. In particular, I shall describe a method of using the gauge fields to associate an element of G with every path in space-time; these path-dependent group elements will be very useful later.

Let P be a path in space-time, going from some initial point x_0 to some final point x_1. Let us consider a field ϕ such that its covariant derivative vanishes along P:

$$\frac{dx^\mu}{ds} D_\mu \phi = 0 \tag{3.18}$$

where s is some parameter along the path, chosen such that s goes from 0 to 1 as x goes from x_0 to x_1. Eq. (3.18) can be rewritten as

$$\frac{d\phi}{ds} = -ie \frac{dx^\mu}{ds} A_\mu^a T^a \phi. \tag{3.19}$$

Aside from the fact that this involves finite-dimensional matrices rather than infinite-dimensional ones, this is structurally identical to the familiar time-evolution equation of interaction-picture perturbation theory, and can be solved in the same way, by Dyson's formula. In particular, ϕ at

the end of the path is given in terms of ϕ at the beginning of the path by

$$\phi(x_1)=g(P)\phi(x_0), \tag{3.20}$$

where

$$g(P)=S\exp\left(-ie\int_0^1 \mathrm{d}s\,\frac{\mathrm{d}x^\mu}{\mathrm{d}s}A_\mu^a T^a\right), \tag{3.21}$$

and S indicates s-ordering (defined in the same way as T-ordering in Dyson's formula.)

The $g(P)$s are very useful objects. If we know $g(P)$ for every path, we know the gauge fields, by differentiation. Also, they have simple gauge-transformation properties,

$$g(P)\to g(x_1)g(P)g(x_0)^{-1}. \tag{3.22}$$

I shall now use the $g(P)$'s to prove the following:

Theorem. It is always possible to make a gauge transformation (or, as we say, choose a gauge) such that

$$A_0^a=0. \tag{3.23}$$

Proof. For any space-time point x, define P_x to be the straight-line path from (x, 0) to x. The desired gauge transformation is defined by

$$g(x)=g(P_x)^{-1}, \tag{3.24}$$

for, under this transformation,

$$g(P_x)\to g(P_x)^{-1}g(P_x)g(P_0)=1. \tag{3.25}$$

from which Eq. (3.23) follows by differentiation.

Note that if the original fields are non-singular, so are the transformed fields. In the future, we shall always assume that we have chosen our gauge such that Eq. (3.23) holds. (This still leaves us the freedom to make time-independent gauge transformations.) The advantage of working in such a gauge is that the time-derivative terms in the Lagrangian (3.17) simplify enormously:

$$\mathrm{D}_0\phi=\partial_0\phi, \tag{3.26a}$$

and

$$F_{0i}^a=\partial_0 A_i^a. \tag{3.26b}$$

Thus, the structure of the initial-value problem is the same as that for a theory of scalar fields only; a complete and independent set of initial-value data is given by the fields and their first time derivatives at any fixed time. Also, the expression for the energy simplifies:

$$E=T+V, \tag{3.27}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

where

$$T = \int \mathrm{d}^D\mathbf{x}[\tfrac{1}{2}(\partial_0 A_i^a)^2 + \partial_0\phi^\dagger\cdot\partial_0\phi], \tag{3.28}$$

$$V = \int \mathrm{d}^D\mathbf{x}[\tfrac{1}{4}(F_{ij}^a)^2 + \mathbf{D}\phi^\dagger\cdot\mathbf{D}\phi + U(\phi)], \tag{3.29}$$

and D is the number of spatial dimensions. Finally, we have nothing to lose by working in such a gauge, since the definition of a dissipative solution, Eq. (1.1), is gauge-invariant.

3.3 *Topological conservation laws, or homotopy classes*

We are now ready to tackle the problem of finding topological conservation laws, that is to say, to divide the space of initial value data (fields and their first time derivatives at some fixed time) into disconnected components. Since we are always working at a fixed time, I will drop the explicit t-dependence from our formulae, to simplify notation. I will also use our freedom to make time-independent gauge transformations to insist that

$$A_r^a(\mathbf{x}) = 0,\ r \geqslant 1, \tag{3.30}$$

where the subscript r indicates the radial component. (I have excluded a sphere around the origin because the radial component is ill-defined at the origin.) The existence of an appropriate gauge transformation follows from an argument identical with that at the end of Sect. 3.2, with t replaced by r.

I will carry out the necessary arguments in detail for two spatial dimensions; the generalization to three dimensions will be straightforward.

From the formula for the energy,

$$E \geqslant \int_1^\infty r\mathrm{d}r \int_0^{2\pi} \mathrm{d}\theta[\partial_r\phi^\dagger\cdot\partial_r\phi + U(\phi)]. \tag{3.31}$$

Thus, if the energy is to be finite, by the same reasoning as in one dimension,

$$\lim_{r\to\infty} \phi(r, \theta) = \phi(\infty, \theta) \tag{3.32}$$

must exist and must be a zero of U, that is to say, an element of G/H, for every θ.

Now we come to the key point: can $\phi(\infty, \theta)$ depend non-trivially on θ? At first glance, one would think not, for, if it did,

$$\mathbf{e}_\theta\cdot\nabla\phi \xrightarrow[r\to\infty]{} \frac{1}{r}\frac{\mathrm{d}\phi(\infty, \theta)}{\mathrm{d}\theta}, \tag{3.33}$$

and this would make the energy integral diverge. However, the quantity

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

that enters the energy is not the gradient, but the covariant gradient,

$$\mathbf{e}_\theta \cdot \mathbf{D}\phi = \frac{1}{r}\frac{\mathrm{d}\phi}{\mathrm{d}\theta} + \mathrm{i}eA_0^a T^a \phi, \tag{3.44}$$

and for any given non-singular $\phi(\infty, \theta)$, we can choose the gauge fields so this goes to zero for large r as fast as we please. Note that our simplifying assumption – that the set of zeros of U is the coset space G/H – is critical for this argument, for it is only this assumption that allows us to assert that we can cancel an infinitesimal change in ϕ (the first term in the covariant derivative) with an infinitesimal group transformation (the second term). Observe that the finiteness of the energy links the large-r behavior of the gauge fields with that of the scalar fields:

$$\lim_{r\to\infty} rA_\theta^a T^a \phi(\infty, \theta) = -\mathrm{i}e^{-1}\frac{\mathrm{d}\phi(\infty, \theta)}{\mathrm{d}\theta} \tag{3.35}$$

In particular, this implies that the gauge fields must go like $1/r$ at large r. This is perfectly consistent with the finiteness of the energy, for the field strengths then go like $1/r^2$, and their squares like $1/r^4$, which causes no large-r blowup in the energy integral.

The energy also has a contribution from the time derivatives of the fields, Eq. (3.28), but the time derivatives are independent initial-value data and can be independently adjusted to make only a finite contribution.

All of this may be trivially extended to three space dimensions; the only difference is that the angle θ is replaced by the two polar angles θ and ϕ. However, the process of extension stops at three dimensions; in four or more dimensions, a term in the energy density proportional to $1/r^4$ at large r does make the total energy blow up. (Fortunately, physical interest also stops at three dimensions.)

Let us summarize what we have found. In two dimensions, with every non-singular finite-energy set of initial value data we have associated the function $\phi(\infty, \theta)$. This function is a mapping of a circle, or, as mathematicians call it, S^1 (i.e. a one-dimensional sphere), into G/H. Likewise, in three dimensions, we have a mapping of S^2 (i.e. an ordinary two-dimensional sphere) into G/H.

If two sets of initial-value data have the same mapping associated with them, it is obvious that we can continuously turn one into the other without making the energy blow up; once we have tacked things down at spatial infinity, nothing continuous we do in the finite part of space can cause any trouble. On the other hand, if the two sets of initial-value data have different mappings associated with them, they can be continuously deformed into each other if and only if the associated mappings can be continuously deformed into each other.

Thus, we have greatly simplified our search for topological conservation laws. To determine whether two sets of initial-value data lie in the same connected component of the space of non-singular finite energy solutions, we need only determine whether the associated mappings from S^1 (in two space dimensions) or S^2 (in three) into G/H can be continuously deformed into each other.

Now, mathematicians[12] have a special word for 'continuously deformable'; they call it homotopic. For the benefit of purists, I will give the definition: let X and Y be two topological spaces, and let $f_0(x)$ and $f_1(x)$ be two continuous functions from X to Y. Let I denote the unit interval on the real line, $1 \geqslant t \geqslant 0$. We say the two functions f_0 and f_1 are homotopic if and only if there is a continuous function from the Cartesian product of X and I to Y, $F(x, t)$, such that $F(x, 0) = f_0(x)$ and $F(x, 1) = f_1(x)$. The function F is called a homotopy. As promised, this is just a fancy way of saying that f_0 is continuously deformable into f_1.

Homotopy is obviously an equivalence relation, and thus we can divide all functions from X to Y into homotopy classes, such that two functions are in the same class if they are homotopic and in different classes if they are not. Thus, with no loss in content and with a considerable elevation in tone, we can rephrase the result of two paragraphs back as the following:

Theorem. The connected components of the space of nonsingular finite-energy solutions are in one-to-one correspondence with the homotopy classes of mappings from S^{D-1} to G/H, where D is the number of space dimensions.

As a trivial corollary, if U has only one zero, so G/H consists of a single point, there is only one homotopy class (indeed, only one mapping), so there are no non-trivial topological conservation laws. This is result (1) of Sect. 3.1. On the other hand, if there is more than one homotopy class, there are non-trivial topological conservation laws.

You might worry that even if we do find several homotopy classes, the topological conservation laws might be physically uninteresting, because the solutions in different components might be gauge transforms of each other. This worry is eliminated by the following:

Theorem. Let $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ be two non-singular finite-energy field configurations such that

$$\phi_2(\mathbf{x}) = g(\mathbf{x})\phi_1(\mathbf{x}). \tag{3.36}$$

where $g(\mathbf{x})$ is non-singular. Then the associated mappings of S^{D-1} into G/H are homotopic.

Proof. For notational simplicity, I will give the proof only for $D=2$. The extension to $D=3$ is trivial. Since G, like all gauge groups, is connected, we can continuously distort $g(\mathbf{x})$ so that $g(\mathbf{0})=1$. This obviously defines a homotopy, so, with no loss of generality, we can restrict ourselves to the case $g(\mathbf{0})=1$. We now rewrite Eq. (3.36) in polar coordinates:

$$\phi_2(r,\theta)=g(r,\theta)\phi_1(r,\theta), \tag{3.37}$$

where, by assumption, $g(0,\theta)=1$. We now define a homotopy:

$$F(\theta,t)=g\left(\frac{t}{1-t},\theta\right)\phi_1(\infty,\theta). \tag{3.38}$$

This obviously equals $\phi_1(\infty,\theta)$ when $t=0$ and $\phi_2(\infty,\theta)$ when $t=1$. Q.E.D.

3.4 *Three examples in two spatial dimensions*

I shall now try to put some flesh on these bare and abstract bones by giving some specific examples. I begin with three examples in two spatial dimensions. In each case, the existence of non-trivial topological conservation laws can be used to prove the existence of non-dissipative solutions, along the lines of the arguments in Sect. 3.1; I will not bother to give these arguments explicitly, though, and will restrict myself to finding the topological conservation laws.

Example 1. G is U(1), the multiplicative group of complex numbers of unit modulus, and the set of scalar fields consists of a single complex field, ϕ, on which G acts by multiplication. The self-interactions of ϕ are given by

$$U=\frac{\lambda}{2}(\phi^*\phi-a^2)^2, \tag{3.39}$$

where λ and a are positive numbers, just as in the ϕ^4 theory of Sect. 2.1. The set of zeros of U, G/H, consists of all ϕs of the form

$$\phi=ae^{i\sigma}, \tag{3.40}$$

where σ is a real number. That is to say, G/H is a circle, S^1. Thus we must find the homotopy classes of mappings from S^1 (in ordinary space) to S^1 (in field space).

This is a famous problem with a well-known and obvious solution. The homotopy classes are characterized by an integer, the 'winding number', which tells how many times you wind around the circle in field space when you go once around the circle in ordinary space. In equations,

$$n=\int_0^{2\pi}\frac{d\theta}{2\pi}\frac{d\sigma}{d\theta}. \tag{3.41}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

Thus, the space of finite-energy non-singular solutions decomposes into an infinite number of disconnected components, each labeled by the winding number, n. An easy argument (construct it yourself) shows that any solution with $n \neq 0$ is non-dissipative. In fact, although topological conservation laws are not strong enough to show it, there are actually time-independent solutions with $n = \pm 1$. Solutions with higher values of n tend to decompose into widely separated solutions with $n = \pm 1$.

This theory, at least for time-independent solutions, is mathematically identical with the Landau–Ginzburg theory of Type II superconductors, and the lumps we have found are called 'flux lines' in superconductor theory.[13] They are called lines because superconductor theorists think of the two-space in which we are working as the x–y plane in a three-dimensional superconductor in which everything is independent of the z coordinate. The 'flux' requires more explanation:

By Stokes's theorem, the magnetic flux passing through our plane is given by

$$\Phi = \lim_{r \to \infty} r \int_0^{2\pi} d\theta A_\theta(r, \theta). \tag{3.42}$$

By Eqs. (3.35) and (3.40),

$$\Phi = \int_0^{2\pi} d\theta e^{-1} \frac{d\sigma}{d\theta} = \frac{2\pi n}{e}, \tag{3.43}$$

whence 'flux lines'.

It is a bit surprising that we have obtained a quantization condition, the quantization of flux, from a purely classical theory. (You should find it especially surprising if you have encountered more traditional treatments of flux quantization, which emphasize (quite properly) the essentially quantum-mechanical nature of the effect.) The reason for the apparent contradiction is that our e is the coupling constant for a field, ϕ. In classical physics, this quantity has no connection with the electric charge of a particle; the two quantities even have different dimensions. It is only the wave–particle duality (essentially quantum-mechanical) that enables us to connect the two, through

$$\hbar e_{\text{field}} = e_{\text{particle}}. \tag{3.44}$$

Thus, when expressed in terms of the electron charge (or, more properly, the Cooper-pair charge), Eq. (3.43) does have Planck's constant in it, and is quantum-mechanical.

Example 2. G is SO(3), and the scalar fields form an isovector, ϕ. U is as before, with ϕ^2 replacing $\phi^*\phi$. (This is the Georgi–Glashow model, without fermions, and in two space dimensions.) G/H is now a two-dimensional sphere, S^2.

Thus we must find the homotopy classes of mappings from S^1 to S^2. It is easy to show that every such mapping is homotopic to the trivial mapping (that which maps all of S^1 into a single point in S^2); this is the famous theorem that it is impossible to lasso a basketball. A more formal proof goes like this: S^1 is one-dimensional and S^2 is two-dimensional. Thus there is at least one point of S^2 into which no point in S^1 is mapped. Remove this point from S^2. S^2 with one point removed is topologically identical with an open disc. Define a homotopy by shrinking the disc to its central point. You cannot lasso a basketball.

Thus there is only one homotopy class, and no non-trivial topological conservation laws.

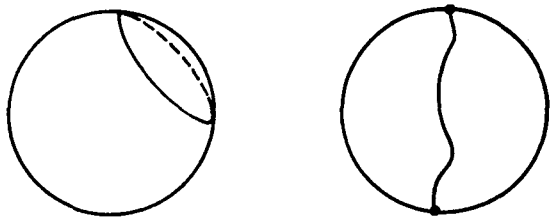
Example 3. G is SO(3), as before. ϕ is an isotensor, which we represent, as usual, as a traceless real symmetric 3×3 matrix. I will assume that U is such that a general zero of U is of the form

$$\phi = 2aee^{\mathrm{T}} - a(I - ee^{\mathrm{T}}). \tag{3.45}$$

where a is some parameter, determined by the detailed form of U, and e is a real unit vector. (This is easy to arrange with a U that is a sum of terms proportional to $\mathrm{Tr}\phi^2$, $\mathrm{Tr}\phi^3$, and $\mathrm{Tr}\phi^4$.) Thus, there is one eigenvector of ϕ, e, with eigenvalue $2a$, and two orthogonal eigenvectors, both with eigenvalue $-a$. The infinitesimal structure of symmetry breakdown here is the same as in Example 2; there is one massless gauge meson corresponding to infinitesimal rotations about the axis e. However, G/H is very different in structure. For although the vectors e do indeed lie on a sphere (S^2), $-e$ defines the same ϕ as e. Thus G/H is a sphere with antipodal points identified.

Let us represent mappings from S^1 into G/H by drawing the closed path in G/H traced out as θ goes around S^1. Fig. 2 shows two such drawings. (Fig. 2(b) may not look like a closed path to you, but remember, antipodal points are identified.) Clearly Fig. 2 shows the only two possible closed paths, up to homotopies. Either a path is manifestly closed when drawn on a sphere (Fig. 2(a)) or it goes between antipodal points (Fig. 2(b)), and there is no way to continuously distort a path of the first type into one of the second.

Fig. 2



Thus there are two homotopy classes, and the space of non-singular finite-energy solutions has two connected components. I emphasize the contrast between this example and Example 1. There, the components were labeled by integers; in its algebraic structure, the topological conservation law was much like the conservation of electric charge in quantum field theory. Here, the topologically conserved quantity can only take on two values; the situation is much more like parity, for example.

3.5 *Three examples in three dimensions*

Example 4. The field theory is the same as in Example 2 (the Georgi–Glashow model), except that we are now in three space dimensions. Thus, we must study the homotopy classes of mappings from S^2 into S^2, rather than of S^1 into S^2.

Mappings of spheres into spheres are harder to visualize than mappings of circles into spheres, but I will try and convince you (by a very non-rigorous argument) that in this case there is at least one[14] non-trivial homotopy class. Argument: you cannot peel an orange without breaking the skin. The skin of the orange (idealized as being infinitely thin and infinitely flexible) is a sphere, and so is the surface of the flesh of the orange upon which it lies. When the skin lies on the surface it defines a mapping of a sphere (the skin) onto a sphere (the surface of the flesh). If this mapping were homotopic to the trivial mapping, we could distort the skin on the surface, without breaking it, until the whole skin lies over a single point. We could then lift the skin off the orange. This is obviously impossible.

Thus there is at least one mapping of a sphere into a sphere, to wit, the identity mapping, that is not in the same homotopy class as the trivial mapping. By arguments which should by now be standard, this implies that any set of non-singular finite-energy initial value data such that

$$\lim_{r\to\infty} \phi^b(\mathbf{x}) = ax^b/r, \tag{3.46}$$

where $b = 1, 2, 3$, cannot dissipate.

In fact, there are not only non-dissipative solutions obeying the boundary condition (3.46) but actual time-independent solutions, found independently by Polyakov[15] and 't Hooft.[16] (It is not so hard to find such solutions as one might think; see Appendix 4.) Polyakov called the solution a 'hedgehog', for obvious reasons; 't Hooft called it a 'magnetic monopole', for reasons that will become clear in Sect. 3.7.

Example 5. The group is SU(2), and the scalar fields transform like a complex isospinor, K. The zeros of U are all fields such that $\bar{K}K$ is some positive number, a^2. (This is the Weinberg–Salam model, without leptons and without the U(1) gauge field.)

If we write the two-component complex K field in terms of four real fields, we see that G/H is a hypersphere, S^3. All mappings from S^2 to S^3 are homotopic to the trivial mapping, by the argument given for Example 2, with 'disc' replaced by 'ball'. Thus, there are no non-trivial topological conservation laws.

Example 6. G is any compact connected Lie group, H is any discrete subgroup.

At this point, for the first and last time in this discussion, I will have to appeal to authority: it is a trivial corollary of a famous theorem of Elie Cartan[17] that, if H is discrete, all mappings of a sphere into G/H are homotopic to the trivial mapping.

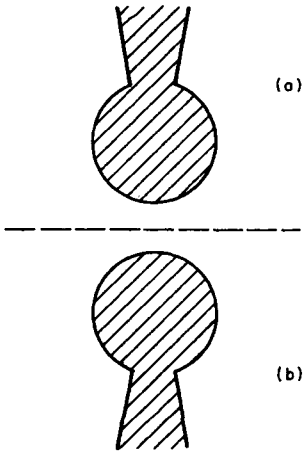
Thus, there are no non-trivial topological conservation laws.[18] This is a result (2) of Sect. 3.1. (Note that the previous example is a special case of this one.)

3.6 *Patching together distant solutions, or, homotopy groups*

I now turn to the question of patching together distant lumps, the question we analyzed for our simple models in Sect. 2.3. In one sense, the problem is much simpler for gauge theories than for the simple scalar models: we need not worry about patching together the gauge fields, for they vanish at large distances, and we need not worry about patching together the scalar fields, for we can always force them to match by an appropriate gauge transformation.

Fig. 3 explains this statement in more detail for the case of two spatial dimensions. Fig. 3(a) shows the plane for some fixed time scalar field, $\phi_1(r, \theta)$. In order to fit the drawing onto a finite page, I have assumed that ϕ_1 attains its asymptotic value, $\phi_1(\infty, \theta)$, and $\mathbf{D}\phi_1$ vanishes, at some finite radius, everywhere outside the shaded disc in the drawing. (The wedge protruding from the disc will be explained shortly.) Now, starting from some fixed angle, say $\theta = 0$, we can gauge transform $\phi_1(\infty, \theta)$ to some

Fig. 3



standard value, say $\phi_1 = \phi_0$, and continue gauge-transforming ϕ_1 to ϕ_0 as we move around the circle. Of course, we can not go all the way around the circle, for we would surely reach a discontinuity when we return to our starting point, if the solution is not in the trivial homotopy class. Thus we gauge transform ϕ_1 to ϕ_0 only in the unshaded region. Fig. 3(b) shows the plane for a second solution, ϕ_2. Here also we have gauge-transformed to make $\phi_2 = \phi_0$ in the entire unshaded portion of the plane. We can now path the two halves of the figure together, along the dashed line, with no discontinuity.

Fig. 3 also serves to illustrate the situation in three spatial dimensions, if we think of it as a cross-section of a surface of revolution. The discs are then balls and the wedges cones, and the argument is unchanged.

Thus, in both two and three dimensions, the assembly of distant lumps looks much more like the assembly of distant particles than it did in one dimension. Nevertheless, there is one tricky point: I said that we had to make a gauge transformation such that ϕ_1 became ϕ_0 everywhere in the unshaded region of Fig. 3(a). However, such a gauge transformation is not uniquely defined; once we have set ϕ_1 equal to ϕ_0, we do not change the situation if we make a gauge transformation, $g(\mathbf{x})$, such that g is an element of H everywhere in the unshaded region. If we consider two such gauge-equivalent versions of ϕ_1, and patch them on to the same ϕ_2, it is not clear that the two compound solutions obtained in this way will be gauge-equivalent. (Of course they are equivalent under a discontinuous (across the dashed line) gauge transformation, but this is not allowed; a discontinuous gauge transformation introduces terrible singularities in the gauge fields.)

It is easy to see that no ambiguity of this kind can arise in our patching prescription if H is connected, as it is for Examples 1 and 4, for in this case we can continuously distort $g(\mathbf{x})$ to one as we approach the dashed line, without changing ϕ_1 at all. Even if H is disconnected, as it is for Example 3, the same argument shows that we need check for ambiguity only in the case when $g(\mathbf{x})$ is a constant in the unshaded region; indeed, we need only check for some one constant g from each component of H. But check we must, for there is no general theorem that assures us there is no ambiguity in our patching prescription. (As we shall see shortly, there is no ambiguity even for Example 3. However, it is possible to construct somewhat more complicated models for which the patching prescription is ambiguous.)[19]

I have given a prescription for patching together two widely separated solutions of our field equations. Now, each of our two original solutions is associated with a homotopy class, as is the compound solution, and our patching prescription is obviously homotopically invariant. Thus the (possibly ambiguous) physical operation of patching together two solutions to make a third defines a (possibly ambiguous) mathematical law for composing two homotopy classes to make a third. That is to say, it gives a group structure to the set of homotopy classes. The groups we obtain in this way are objects well-known to mathematicians; they are called homotopy groups.[20]

I will now give the standard mathematical definitions of homotopy groups,[21] and show that they correspond to our patching prescription. As usual, I will begin in two dimensions, and later go on to three.

Consider a subset of the set of continuous mappings from S^1 into G/H, mappings such that some fixed point in S^1 (say $\theta=0$) is mapped into some fixed point in G/H (say ϕ_0). Any such mapping can be thought of as a continuous function $\phi(\theta)$ from the interval $[0, 2\pi]$ into G/H, such that

$$\phi(0)=\phi(2\pi)=\phi_0. \tag{3.47}$$

Given two such mappings ϕ_1 and ϕ_2, we define the compound mapping $\phi_1\cdot\phi_2$ simply by putting the two intervals together end to end (with the first on the left and the second on the right). This defines a continuous function on the interval $[0, 4\pi]$. We now rescale the coordinate by two to turn this into $[0, 2\pi]$, thus producing a new function of the standard form. In equations,

$$\begin{aligned}\phi_1\cdot\phi_2(\theta)&=\phi_1(2\theta), && 0\leqslant\theta\leqslant\pi\\ &=\phi_2(2\theta-2\pi), && \pi\leqslant\theta\leqslant 2\pi.\end{aligned} \tag{3.48}$$

Another way of thinking of this is to think of ϕ_1 and ϕ_2 as describing two closed paths in G/H, each of which begins and ends at ϕ_0. $\phi_1\cdot\phi_2$ then

describes the path obtained by first going around ϕ_1 and then going around ϕ_2. This way of describing things makes it especially clear why we must impose Eq. (3.47) to get a definition of multiplication; there is no sense in the prescription to go first around one path and then around the other unless the second path begins where the first ends.

We define two such mappings to be homotopic just as in Sect. 3.3, except that we restrict the continuous deformation, $F(\theta, t)$ to obey Eq. (3.48) for each t. More loosely, we are allowed to continuously distort our closed paths as we will, except that the starting point is nailed down. Our definition of multiplication is obviously invariant under (restricted) homotopies, and thus defines a multiplication law among (restricted) homotopy classes. This multiplication obeys all the group axioms: associativity is trivial, the identity is the trivial homotopy class, and the inverse is obtained by going around the path in the opposite direction.

The group obtained in this way is called the first homotopy group, and is denoted by $\pi_1(G/H)$. We do not bother to specify ϕ_0 in this expression, because the group structure does not depend on ϕ_0. (Proof: gauge invariance.) From now on, we will reserve the term 'homotopy classes' for our old homotopy classes, and refer to our restricted homotopy classes (with ϕ_0 nailed down) as 'elements of the homotopy group'.

Given a general mapping from S^1 into G/H, we can always turn it into a mapping of the form (3.47) by multiplying it by an appropriate element of G. Thus we can always assign to every homotopy class an element of the homotopy group. It is easy to see that the trivial homotopy class is always assigned to the unit element of the homotopy group;[22] however, for other homotopy classes, there may be ambiguity in this prescription, because the 'appropriate element' of G is not unambiguously defined. Given one such element, we can always obtain another by left-multiplying by an element of H. If H is connected, there is no problem, because we can then continuously distort the element of H to 1, and this defines a homotopy. However, if H is disconnected, we have to check to see if there is or is not ambiguity.

You may have the feeling that you have read the preceding paragraph before, because it is exactly like the description of possible ambiguities in the patching procedure. This is as it should be, because the composition of homotopy group elements *is* our patching procedure. This can be seen from Fig. 3: if we draw a big circle around the figure, and traverse it in a positive sense, starting at $\theta = 0$, we begin at ϕ_0 in G/H, then circle the path associated with ϕ_1 (as we pass through the top shaded wedge), then that associated with ϕ_2 (the bottom wedge) and then return to ϕ_0.

It is an easy exercise to check that, for Example 1, π_1 is the additive

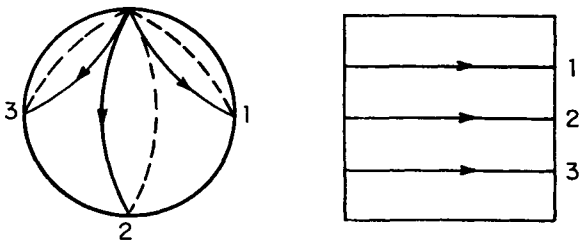
group of the integers; this is the law of addition of winding number (or of magnetic flux). For Example 3, π_1 is the additive group of the integers modulo 2 (i.e. $1+1=0$). If, in Fig. 2, we take a path that goes from the north pole to the south pole, this is the same as a path that goes from the south pole to the north pole, since antipodal points are identified. Thus if we go around the path twice, we obtain a path that goes from the north pole to the north pole, that is to say, one that is homotopic to the trivial path. $(1+1=0.)$

The extension to three dimensions, that is to say, from S^1 to S^2, is straightforward. We consider mappings from S^2 into G/H such that some fixed point in S^2 (say the north pole) is mapped into some fixed point in G/H (say ϕ_0). Just as we can represent the circle as a line interval with boundary points identified, so we can represent the sphere as a square with boundary points identified. Fig. 4 shows such a representation; to make things clear, I have drawn three paths on S^2 and their images in the square. Note that the entire perimeter of the square represents a single point on the sphere, the north pole. (If, instead of a square, I had drawn a (topologically equivalent) closed disc, this representation would be that used in polar projection maps of the earth, like the United Nations flag, where the entire boundary of the disc represents the south pole.)

We can now define the multiplication of two mappings by putting two squares side by side and rescaling the horizontal coordinate to produce a new square. The remainder of the construction is step-by-step the same as in the two-dimensional case. The group we arrive at in this way is called the second homotopy group, and is denoted by $\pi_2(G/H)$.

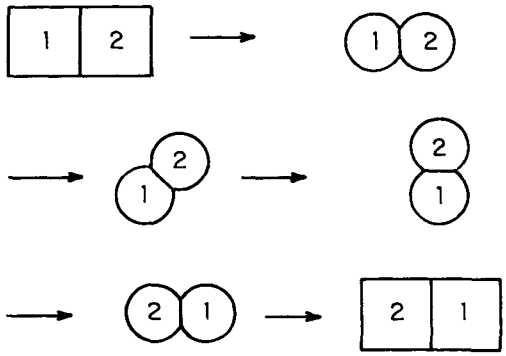
There is one important new feature in three dimensions: the second homotopy group is always Abelian. (Although the first homotopy group is Abelian in our examples, this is not always the case.) Fig. 5 is a sequence of frames from a motion picture of the homotopy which proves this assertion. We first map the two squares with their sides in common into two discs with a portion of their perimeters in common; we then con-

Fig. 4



https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

Fig. 5



tinuously change the common portion of the perimeter until it has moved by π; we then map the two discs back into two squares. The essential reason we can perform this construction is that the boundary of a square is connected; we can not perform a similar construction for π_1 because the boundary of a line interval is not connected.

Of course, we can also see the same thing in terms of patching together solutions, as in Fig. 3. In two dimensions, we can not, by a continuous deformation, exchange the top and the bottom of Fig. 3, without either moving a shaded wedge through the ray $\theta=0$ (destroying Eq. (3.47)) or moving the wedges through each other (destroying our ability to make sense of the drawing). In contrast, in three dimensions, the shaded wedges become cones, and there is plenty of room in three-space to move one cone around another.

This takes care of all of the physics and all of the mathematics of patching together distant solutions, but I would like to waste a few minutes worrying about problems of interpretation, that is to say, about the best words to use to describe the situation we have uncovered.

Let us suppose, for simplicity, that instead of patching together general solutions, we are patching together time-independent solutions, like the lumps of Sect. 2. In Sect. 2.3 we found that we could not always patch together widely separated lumps. This meant that if we were to make the assembly of distant lumps look like the assembly of distant particles, we had to say that several different solutions represented the same object. (In ϕ^4 theory, we had to identify the kink and the antikink.)

We now have exactly the opposite situation: we can always patch together distant solutions, but there may be several ways of doing the patching, if there is ambiguity in identifying homotopy classes with homotopy group elements. Thus, if we are to make the assembly of distant

lumps look like the assembly of distant particles, we must adopt the opposite stratagem, and say that a single solution represents several different objects, one for each way of assigning a homotopy group element to the solution. (As Rudolf Haag has pointed out,[23] this is closely parallel to the situation in quantum parastatistics. Here, if you restrict yourself to the one-particle subspace of Fock space, there is no way of telling that there is more than a single species of particle. It is only by doing experiments in the many-particle subspace that you can detect that there are in fact several species.)

Even this stratagem does not work in the special case of two-dimensional theories for which $\pi_1(G/H)$ is non-Abelian. When you assemble distant particles in two dimensions, the two-particle state you obtain does not depend on the angular order of imaginary wedges radiating from the particles. In this special case, I know of no way of interpreting things such that the assembly of distant lumps is like the assembly of distant particles.

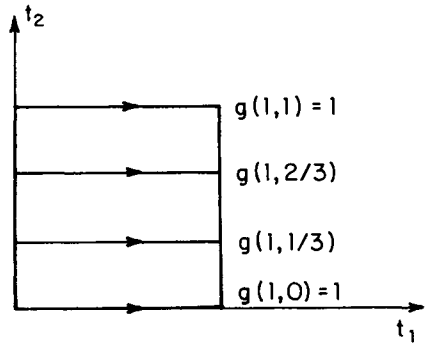
3.7 *Abelian and non-Abelian magnetic monopoles, or, $\pi_2(G/H)$ as a subgroup of $\pi_1(H)$*

From now on we will restrict ourselves to three spatial dimensions. Although some of the phenomena we shall encounter have two-dimensional analogues, these are essentially trivial; the heavy artillery is needed only in three dimensions.

We shall use heavily the representation of a sphere by a square of Fig. 4. In particular, we shall introduce Cartesian coordinates on the square, t_1 and t_2, $1 \geqslant t_{1,2} \geqslant 0$, and represent the asymptotic values of our scalar fields by a function $\phi(t_1, t_2)$, which equals ϕ_0 on the boundary of the square (the north pole of the sphere).

Although we have not been paying much attention to them lately, we of course have gauge fields as well as scalar fields in our theory. These go to zero like $1/r$ for large r; thus we can associate a group element with every path on the sphere (equivalently, on the square), defined by Eq. (3.21). For every point in the square, we will choose a standard path, defined to be a horizontal line starting from the t_2 axis and terminating at the given point. We thus define a function from the square into G, $g(t_1, t_2)$. Note that $g(t_1, t_2)$ does *not* define a continuous function on the sphere. It is equal to 1 on the left-hand side of the square, because this is the starting point of all our paths, and it is equal to 1 at the top and bottom of the square, because these are trivial paths that never leave the north pole, but it is not necessarily equal to 1 on the right-hand side of the square. This situation is shown in Fig. 6.

Fig. 6



$\phi(t_1, t_2)$ and $g(t_1, t_2)$ are not independent functions; they are connected by the three-dimensional analogue of Eq. (3.35), the statement that $\mathbf{D}\phi$ goes to zero for large r more rapidly than $1/r$ (indeed, in three dimensions, more rapidly than $1/r^{\frac{3}{2}}$). Thus,

$$\phi(t_1, t_2) = g(t_1, t_2)\phi_0, \tag{3.49}$$

which implies that

$$\phi_0 = g(1, t_2)\phi_0. \tag{3.50}$$

Thus, $g(1, t_2)$ defines a path in H, which starts at 1 and ends at 1. A continuous distortion of our field configuration (restricted so ϕ at the north pole stays fixed at ϕ_0) produces a continuous distortion of the path in H. That is to say, Fig. 6 defines a mapping from $\pi_2(G/H)$ into $\pi_1(H)$.

From Fig. 6, we can read off some interesting properties of this mapping:

(1) The arguments associated with Fig. 5 show that we can define group multiplication in π_2 by putting our two squares together any which way. If we put them together one on top of the other, it is clear that the product of two elements in $\pi_2(G/H)$ is mapped into the product of the corresponding two elements in $\pi_1(H)$. That is to say, our mapping is a homomorphism of one group into the other.

(2) Every line of fixed t_1 on the square defines a path in G (not necessarily in H), which begins at 1 and ends at 1. As t_1 goes from 0 to 1, this path changes continuously from the trivial path to $g(1, t_2)$. Thus, $g(1, t_2)$ can not be an arbitrary path in H, but must be one that is homotopic to the trivial path when H is embedded in G.

This result can be rephrased in group-theoretical language: every closed path in H is also a closed path in G; thus there is a natural homomorphism of $\pi_1(H)$ into $\pi_1(G)$. What we have found is that our homomorphism maps $\pi_2(G/H)$ into the kernel of this natural homomorphism. (I remind you that

the kernel of a homomorphism is the set of all elements that are mapped into 1.)

(3) Conversely, given a closed path in H that is homotopic to the trivial closed path when H is embedded in G, we can run through our construction backwards and find a $\phi(t_1, t_2)$ consistent with this path being $g(1, t_2)$. That is to say, we can use the homotopy to define a function on the square, $g(t_1, t_2)$, and then use Eq. (3.49) to define ϕ. In group-theoretical language, our homomorphism maps $\pi_2(G/H)$ *onto* the kernel of this natural homomorphism.

Thus we have established that our homomorphism is onto. I shall now show that it is also one-to-one, and thus we have an invertible homomorphism, i.e. an isomorphism.

Theorem. $\pi_2(G/H)$ is isomorphic to the kernel of the natural homomorphism of $\pi_1(H)$ into $\pi_1(G)$.

Proof. As stated, we have to establish the one-to-one nature of our homomorphism. Since we are dealing with a homomorphism, a mapping that preserves group multiplication, all we need to show is that only the identity element of $\pi_2(G/H)$ is mapped into the identity element of $\pi_1(H)$. In other words, given that $g(1, t_2)$ is homotopic to the trivial path $[g(t)=1]$ in H, we wish to show that $\phi(t_1, t_2)$ is continuously deformable to the constant function, $\phi=\phi_0$. We can obviously continuously deform g (and therefore ϕ) so that it runs through all its changes in the left-hand half of the square, and is independent of t_1 in the right-hand half, $t_1 \geqslant \frac{1}{2}$. Now we can replace g for $t_1 \geqslant \frac{1}{2}$ by the homotopy that turns $g(1, t_2)$ into 1. This is an abrupt change, not a continuous deformation, but it does not affect ϕ at all, since the homotopy stays in H. We now have arranged matters such that $g(t_1, t_2)=1$ on the entire boundary of the square. That is to say, $g(t_1, t_2)$ defines a continuous mapping from S^2 into G. By the theorem of Cartan, quoted in Example 6, such a mapping can always be continuously deformed into the trivial mapping. Through Eq. (3.49), this defines a continuous deformation of $\phi(t_1, t_2)$ into the trivial mapping. Q.E.D.[24]

As a specific instance, let us consider the field theory of Example 4. Here G is SO(3) and H is U(1). Thus $\pi_1(H)$ is the additive group of the integers. If we embed U(1) in SO(3), only paths with even winding numbers, rotations by even multiples of 2π, are deformable to the trivial path. (This is why SO(3) has double-valued representations.) Thus, the kernel of the natural homomorphism is the subgroup of the even integers.

We have finally found the structure of the topological conservation laws for Example 4; each connected component of the space of nonsingular finite-energy solutions is labeled by an even integer. This is

reminiscent of Example 1, where the components were labeled by integers, and where these integers were identified with quantized magnetic flux. I shall now show that the same identification holds here.

To do this, I will first have to find the definition of magnetic flux. We have three field-strength tensors, and thus three candidates for the magnetic field vector:

$$H_i^b = \varepsilon_{ijk} F^{bjk}. \tag{3.51}$$

What linear combination of these is the magnetic field depends on what linear combination of our three gauge fields is massless, that is to say, is the photon field. Thus, for a ground state, a constant ϕ which is a zero of U, the answer is unambiguous:

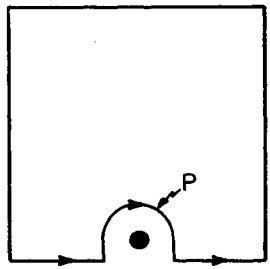
$$H_i = \phi^b H_i^b / a, \tag{3.52}$$

where a is the magnitude of ϕ at a zero of U. This is certainly the quantity that would be measured by a magnetometer in this case. It is also the quantity that would be measured by a magnetometer if $U(\phi)$ and $\mathbf{D}\phi$ were approximately zero over the region of space occupied by the magnetometer. ('Approximately zero' means small compared to the characteristic length and energy scales of the theory.) However, if this condition is not satisfied, then you need the detailed theory of the mechanism of the magnetometer to know what the magnetometer measures. The mathematical reflection of this is that we can add all sorts of terms to Eq. (3.52) (proportional to $\mathbf{D}\phi$, etc.) that vanish for a ground state. Thus we have an infinite variety of possible definitions of the magnetic field, and any one of them is as good as any other; they just describe different models of idealized magnetometers. Fortunately, we do not wish to compute the magnetic field point by point, but only to compute the net magnetic flux emerging from our solution. For this we only need the field for infinite r, where all definitions reduce to Eq. (3.52). Thus,

$$\Phi = \lim_{r \to \infty} r^2 \int \mathrm{d}\Omega H_r. \tag{3.53}$$

To evaluate this quantity, let us choose ϕ_0 to point in the 3-direction in isospin space, and let us gauge transform ϕ so that it equals ϕ_0 over the whole of the sphere at infinity except for a small region near the north pole. (It is easy to show that such a gauge transformation always exists: the sphere with the small region excised is topologically equivalent to a disc with the north pole at its center. On such a disc, we can define the gauge transformation at every point by transforming continuously outward from the center along radial lines.) Fig. 7 shows this situation using the by-now-familiar representation of the sphere as a square; the excised

Fig. 7



region is indicated by the black dot. Since $\mathbf{D}\phi$ vanishes on the sphere at infinity, outside the black dot the only non-vanishing component of A_i^b is A_i^3; thus,

$$\mathbf{H} = \nabla \times \mathbf{A}^3, \tag{3.54}$$

outside the black dot. Furthermore, since H is a continuous gauge-invariant function, if we make the black dot sufficiently small, we can evaluate Eq. (3.53) with negligible error by integrating only over the portion of the sphere that lies above the path P in Fig. 7. Thus

$$\Phi = \oint_P \mathbf{A}^3 \cdot d\mathbf{x}, \tag{3.55}$$

by Stokes's theorem.

Now, in the same gauge, let us evaluate the element of $\pi_1(H)$ associated with this field configuration. For any path across the square, P', that avoids the black dot, the formula for $g(P')$, Eq. (3.21), is trivial to evaluate, because only $\mathbf{A}^3$ is non-vanishing:

$$g(P') = \exp\left[-ieT_3 \oint_{P'} \mathbf{A}^3 \cdot d\mathbf{x}\right]. \tag{3.56}$$

Just as before, by gauge invariance and continuity, we can evaluate the winding number with negligible error by just going down to the path P. Thus,

$$2\pi n = e \oint_P \mathbf{A}^3 \cdot d\mathbf{x}. \tag{3.57}$$

(Of course, I am assuming that T^3 has been normalized in the standard way, such that the gauge fields carry charges e, $-e$, and 0.) I remind you that we showed above that the winding number must be an even integer.

Thus we find

$$\Phi = 2\pi n/e, \tag{3.58}$$

where n is an even integer. This is a remarkable result: all solutions to this

field theory that do not belong to the trivial homotopy class are magnetic monopoles, carrying quantized magnetic charge, and all our topological conservation laws are equivalent to the conservation of magnetic charge.

Nearly all of this analysis extends trivially to a general field theory for which the connected part of H is isomorphic to U(1). (Only the connected part of H is relevant to $\pi_1(H)$.) Here also all topological conservation laws are equivalent to the conservation of magnetic charge. The only part of the analysis that depends on the details of the theory, on how H is embedded in G, is the condition that determines the allowed values of magnetic charge, the generalization of Eq. (3.58). (We shall shortly develop a simple algorithm for finding this generalization in any given theory.) The extension to general Abelian H is also trivial. In this case we may have several massless vector mesons, several photons, and for each photon we have a magnetic charge. On the other hand, if H is not Abelian, there is no way of extending our analysis. Of course, we can always state flatly that the non-Abelian generalization of magnetic charge is an element of $\pi_1(H)$, but there is no general way of representing such a quantity as the integral of a numerical valued field over a large sphere. For example, $\pi_1(\mathrm{SO}(3))$ is the additive group of the integers modulo 2, and there is simply no way of obtaining this structure by the addition of ordinary numbers.

I should emphasize that the objects we have found are very different from the old magnetic monopoles of Dirac. The field of a Dirac monopole is simply the magnetostatic analog of the usual electrostatic Coulomb field. Like the Coulomb field, the Dirac monopole field is a singular solution of free electrodynamics carrying infinite electromagnetic energy. To make the Coulomb field part of a dynamical theory, it is necessary to introduce non-electromagnetic dynamical degrees of freedom, charged particles that constitute the source of the Coulomb field; the masses and spins of these charged particles are free parameters. Likewise, to make the Dirac monopole field part of a dynamical theory, it is necessary to introduce non-electromagnetic dynamical degrees of freedom, the magnetic monopoles; the masses and spins of these monopoles are free parameters.

Our monopoles could not be more different. They have nothing to do with ordinary electrodynamics, and only appear in theories of non-Abelian gauge fields interacting with scalar fields. They are totally non-singular and carry finite energy. They necessitate no new dynamical degrees of freedom, and all of their properties are determined in terms of the same parameters that govern the dynamics of the ordinary scalar and vector mesons.

There is one point of similarity: Dirac monopoles also carry quantized magnetic charge. We can see how this arises using our homotopy methods.

Outside the Dirac monopole, all that exists is electromagnetic field: $G=H=\mathrm{U}(1)$. In addition to the gauge-independent singularity at the origin, the gauge field A has a gauge-dependent line of singularities ('the string') going from the origin to infinity. The string can be moved about by gauge transformations, but can never be gauged away altogether.[25] (The relation between the string and the monopole is rather like that between a branch line (movable) and a branch point (immovable) in the theory of a multi-valued analytic function.) We can now reinterpret Fig. 7 as representing a sphere about the monopole (not necessarily of infinite radius), with the black dot representing the location of the string. Just as before, we can associate with this field configuration an element of $\pi_1(H)=\pi_1(G)$, and identify the winding number with the magnetic flux. Since there is no way of getting rid of the string (the black dot), there is no constraint on this element; thus we obtain Eq. (3.58) for arbitrary (integer) n. We can also see that if n is not zero, there must be a real singularity, not removable by a gauge transformation, somewhere within the sphere. For if there is no real singularity as we shrink the sphere to zero, n must change continuously. Since n is an integer, this means it must be independent of the radius of the sphere. But unless there is a singularity, n must vanish for a sphere of zero radius. We can also see the generalization of a Dirac monopole field to a gauge theory with a non-Abelian gauge group, G. It is a field configuration, necessarily singular, associated with an element of $\pi_1(G)$ other than the identity.[26]

I would now like to pick up a question I mentioned a few paragraphs back, and find the allowed values of the magnetic charge for a general theory in which the connected part of H is isomorphic to U(1). Every connected Lie group, G, has a simply-connected covering group, $\bar{G}$. All multiple-valued representations of G are ordinary single-valued representations of $\bar{G}$. (For example, the covering group of SO(3) is SU(2), and the covering group of U(1) is the additive group of the real numbers.) Let us change our conventions and declare that the gauge group of our theory is not G but $\bar{G}$. Of course, our scalar fields may not form a faithful representation of $\bar{G}$ – several elements of $\bar{G}$ may effect the same transformation of the fields – but faithfulness was never used in our arguments. Also of course, when we enlarge G we must also enlarge H, but let it be so enlarged. Once we have done this enlargement, the kernel condition of our previous theorem is trivial, for $\bar{G}$ is simply connected, and every closed path in $\bar{G}$ can be shrunk to the trivial path. Thus the only thing we have to check is that when we run down the right-hand side of Fig. 6, we really get a closed path in the enlargement of H, that is to say a closed path in $\bar{G}$. To check this, it is sufficient to check that the path is closed in every

representation of $\bar{G}$. In terms of magnetic flux, this means that for every non-zero electric charge Q occurring in any representation of $\bar{G}$, the magnetic flux must be an integral multiple of $2\pi/Q$.

There are now two possibilities: (1) the representations of $\bar{G}$ may contain arbitrarily small non-zero charges. In this case, only zero magnetic flux is possible, and there are no non-trivial topological conservation laws; (2) there is a minimum positive charge, $Q_{\min}$. Now, let Q be some other charge. By forming the direct product of the representation that contains Q with a string of representations that contain $Q_{\min}$ (or their complex conjugates), we can form a representation that contains an object of charge $Q+nQ_{\min}$, for any positive (or negative) integer n. Thus Q must be an integral multiple of $Q_{\min}$, if we are not to obtain a contradiction. Hence the only constraint on the magnetic flux is that it is an integral multiple of $2\pi/Q_{\min}$.

We summarize all this in the following:

Theorem. For a general theory for which the connected part of H is isomorphic to U(1), all topological conservation laws are equivalent to the conservation of magnetic flux. Magnetic flux comes in integral multiples of a minimum unit, given by

$$\Phi_{\min}=2\pi/Q_{\min}, \tag{3.59}$$

where $Q_{\min}$ is the smallest positive electric charge occurring in any single-valued or multiple-valued representation of G. If there is no smallest positive charge, there are no non-trivial topological conservation laws.

Corollary. If G contains a U(1) factor, and if the generator for this U(1) factor enters into the expression for the electric charge, there are no non-trivial topological conservation laws.

Proof. For, in this case, the U(1) factor has infinitely-multiple-valued representations, and thus we can make the electric charge as small as we please.

In all other cases, the standard representation theory of Lie groups tells us that there is a smallest non-zero charge, and thus there are non-trivial topological conservation laws.

This theorem and its corollary constitute result (3) of Sect. 3.1; result (4) is distributed throughout this subsection.

It might be helpful to see these principles at work in some examples:

Example 7. G is SU(3). H is U(2), embedded in SU(3) in the same way the isospin–hypercharge U(2) group is embedded in strong-interaction SU(3). Although this is not strong-interaction SU(3), I will save on definitions by using strong-interaction language and referring to the generators

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

of U(2) as 'isospin' and 'hypercharge'. For orientation, let me begin by considering the case in which the symmetry breaks down all the way to hypercharge U(1). This is topologically the same as electric charge, treated above. Thus, in this case, all topological conservation laws are equivalent to the conservation of 'hypermagnetic flux' and the minimum value of this flux is 2π divided by the minimum hypercharge, that of the non-strange quarks. If U(2) were isomorphic to SU(2)⊗U(1), the additional factor of SU(2) would have no effect; SU(2) is simply connected and makes no contribution to the homotopy structure. However, the mapping of SU(2)⊗U(1) into U(2) is two-to-one. (If we represent an element of the first group by (U, λ), where U is an SU(2) matrix and λ is a complex number of unit modulus, then (U, λ) is mapped into the same element of U(2), $U\lambda$, as $(-U, -\lambda)$.) As a consequence of this, paths that go only half-way around SU(2)⊗U(1), from (1, 1) to $(-1, -1)$, are mapped into closed paths in U(2), and the minimum unit of magnetic flux is half what it would be if H were U(1).

Example 8. G is SU(3) and H is SO(3), the subgroup of real unimodular unitary matrices. Since SU(3) is simply connected, the kernel condition is trivial. Thus, $\pi_2(G/H)$ is $\pi_1(H)$ is the additive group of the integers, modulo two. This is the same algebraic structure we found in two spatial dimensions back in Example 3.

Example 9. This is the same field theory as in Example 3, but in three spatial dimensions. G is SO(3), and the connected part of H is U(1), the subgroup of z-rotations. Thus, as far as the second homotopy group goes, this is the same as Example 4, and the same conclusions hold: the magnetic flux is given by Eq. (3.58), etc. However, there is one novelty; H has two components, and the component that does not contain the identity contains a rotation by π about the x-axis. This changes the sign of a z-rotation, and thus changes the sign of electric charge and magnetic flux. Thus we finally have a concrete example of the patching ambiguity we worried so much about in Sect. 3.6. In this theory, a monopole and an antimonopole are represented by gauge-equivalent solutions of the field equations; it is only when we try and put them together to make many-monopole states that we find that they are in fact different objects.

4 Quantum lumps

4.1 *The nature of the classical limit*

For example, consider the sine–Gordon equation:

$$\mathscr{L} = \tfrac{1}{2}(\partial_\mu \phi)^2 + \frac{\alpha}{\beta^2}(\cos \beta\phi - 1). \tag{4.1}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

If we define

$$\phi' = \beta\phi, \tag{4.2}$$

then

$$\mathscr{L} = \frac{1}{\beta^2}\left[\tfrac{1}{2}(\partial_\mu\phi')^2 + \alpha(\cos\phi' - 1)\right]. \tag{4.3}$$

We see that, in classical physics, β is an irrelevant parameter; if we can solve the sine–Gordon equation for any non-zero β, we can solve it for any other β. The only effect of changing β is the trivial one of changing the energy and momentum assigned to a given solution of the equation.

This is not true in quantum physics, because the relevant object for quantum physics is not $\mathscr{L}$ but

$$\frac{\mathscr{L}}{\hbar} = \frac{1}{\beta^2\hbar}\left[\tfrac{1}{2}(\partial_\mu\phi')^2 + \alpha(\cos\phi' - 1)\right]. \tag{4.4}$$

Another way of saying the same thing is to say that in quantum physics we have one more dimensional constant of nature, Planck's constant, than in classical physics. Thus the quantum version of the sine–Gordon theory involves one more dimensionless parameter, $\beta^2\hbar$, than the classical theory.

Because the only way either β or $\hbar$ enters Eq. (4.4) is through the combination $\beta^2\hbar$, the classical limit, vanishing $\hbar$, is exactly the same as the small-coupling limit, vanishing β. Once we have this knowledge, there is no need to keep cluttering up our equations with $\hbar$s, so from now on I will adopt standard quantum units, and set $\hbar$ equal to one.

Nothing in this analysis is special to the sine–Gordon equation; every step goes through in exactly the same way for ϕ^4 theory, with λ replacing β^2. Here again the classical limit is the small-coupling limit. Likewise, for two-coupling-constant theories, like the gauge theory defined by Eq. (3.39), the classical limit is vanishing e, with λ/e^2 fixed.

These manipulations are trivial, but they teach us something important: if there are particles in the quantum theory that correspond to classical lumps, they are most likely to resemble their classical ancestors for weakly coupled theories. (Conversely, there is no more reason to trust classical analysis for strongly coupled theories than there is to trust the Born approximation.)[27] This suggests that the most direct way to construct quantum lumps is by an expansion in powers of the coupling constant. The leading term in such an expansion should give the classical results, appropriately reinterpreted in quantum language, and the higher terms should give quantum corrections.

I will now derive the first few terms in such an expansion for the case of a time-independent lump.[28]

4.2 *Time-independent lumps: power-series expansion*

For simplicity, I will do the analysis for a single scalar field in one spatial dimension, the sort of theory we discussed in Sect. 2. The generalization to more complicated theories, like those of Sect. 3, will be straightforward.

I will describe the theory in terms of a rescaled field, as in Eq. (4.2), and, for uniformity of notation, I will in all cases denote the coupling constant by β^2. Thus,

$$\beta^2 \mathscr{L} = \tfrac{1}{2}(\partial_\mu \phi')^2 - U(\phi'). \tag{4.5}$$

The canonical momentum density is given by

$$\pi' = \frac{\partial \mathscr{L}}{\partial(\partial_0 \phi')} = \frac{\partial_0 \phi'}{\beta^2}, \tag{4.6}$$

and the Hamiltonian by

$$\begin{aligned} \beta^2 H &= \int [\beta^4 \pi'^2/2 + (\nabla \phi')^2/2 + U(\phi')]\, \mathrm{d}x \\ &= \int (\beta^4 \pi'^2/2)\, \mathrm{d}x + V[\phi']. \end{aligned} \tag{4.7}$$

I will assume that U has at least two zeros, so there exist finite-energy non-singular stable time-independent classical solutions. I denote these by

$$\phi'(x) = f(x-b). \tag{4.8}$$

I also define

$$E_0 \equiv V[f]. \tag{4.9}$$

I emphasize that these are rescaled quantities, and thus independent of β^2.

Equation (4.7) defines a rather peculiar Hamiltonian from the viewpoint of ordinary perturbation theory. Firstly, there is an explicit β^2 on the left-hand side of the equation. Of course, this is a trivial peculiarity; if we can find a power-series expansion for the energy eigenfunctions and eigenvalues of $\beta^2 H$, we can find one for those of H. Secondly, the small parameter multiplies the kinetic energy rather than the potential energy. This is very strange indeed; have we ever encountered such a system before?

Yes, we have. For this is precisely the situation for a diatomic molecule:

$$H = \frac{\mathbf{P}^2}{2M} + V(r), \tag{4.10}$$

where M is the nuclear reduced mass. The standard expansion procedure in the study of the spectra of diatomic molecules uses as the small parameter $1/M$, the coefficient of the kinetic energy. Of course, our system is not

exactly a diatomic molecule. What it is exactly is a polyatomic molecule,

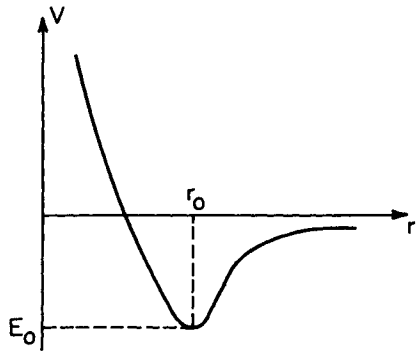
$$\sum_{i=1}^{N} \frac{\mathbf{P}^{(i)2}}{M_i} + V[\mathbf{x}^{(1)} \ldots \mathbf{x}^{(N)}]. \tag{4.11}$$

To be precise, it is an infinitely polyatomic molecule, where all the nuclei have mass $1/\beta^4$. Thus the problem of constructing quantum lumps is one that was solved completely more than forty years ago.

I will now explain this solution, first by reminding you of the familiar results for a diatomic molecule,[29] then by telling you the trivial extension to a polyatomic molecule, and, finally, by making the even more trivial transcription of this extension into the language of field theory.

For the diatomic molecule, we assume the interatomic potential is as shown in Fig. 8. The minimum of V is at $r=r_0$, and $V(r_0)=E_0$. The first

Fig. 8



three approximations to the low-lying energy eigenstates and their eigenvalues are shown in Table 1.

Table 1. *The diatomic molecule*

Order of approximation	Energy eigenstate	Energy eigenvalue
0	$\|r_0, \Omega\rangle$	E_0
1	$\|n, \Omega\rangle$	$+(n+\frac{1}{2})[V''(r_0)/M]^{\frac{1}{2}}$
2	$\|n, l, m\rangle$	$+\dfrac{l(l+1)}{2Mr_0^2}+\ldots$

As we see from the table, the proper expansion parameter for energy eigenvalues is $1/M^{\frac{1}{2}}$; this will become β^2 in the field-theory problem.

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

(The right-hand column of the table is additive; that is to say, the energy in first order is the sum of the first two entries, etc.) I will now explain the origin of the table.

In zeroth order, we neglect the kinetic energy altogether. The particle sits at the bottom of the potential, in an eigenstate of the position operator, **x**. The magnitude of the particle position is fixed at r_0, but its angular position is arbitrary. (As usual, Ω is shorthand for the pair of spherical coordinates θ and ϕ.) This is not much like the real spectrum of low-lying states; in particular, there is a totally spurious degeneracy in Ω, which, as we shall see, is removed only in second order.

To first order, we begin to see the effects of the vibration of the particle about its equilibrium position. Since M is very large, the particle does not vibrate very far, and, to first order, we can replace the potential near equilibrium by a harmonic potential

$$V(r) = E_0 + \tfrac{1}{2}V''(r_0)(r-r_0)^2. \tag{4.12}$$

The energy eigenfunctions are now harmonic-oscillator wave functions in r, but still delta-functions in Ω. They are labeled by the usual harmonic-oscillator excitation number, n, and have the usual harmonic-oscillator energies. (These are the famous vibrational levels.) They are no longer eigenstates of the position operator, but it is still easy to compute the expectation value of any component of **x**. For example, if we normalize the eigenstates such that

$$\langle n', \Omega' | n, \Omega \rangle = \delta_{nn'}\delta(\Omega - \Omega'), \tag{4.13}$$

then, for the z-component of the position operator, in the ground state,

$$\langle 0, \Omega' | z | 0, \Omega \rangle = r_0 \cos\theta\, \delta(\Omega - \Omega'). \tag{4.14}$$

It is only in second order that we begin to see the effects of rotation; this is because the zeroth-order moment of inertia of the molecule is Mr_0^2. The degeneracy in Ω is removed; angle eigenstates are replaced by angular-momentum eigenstates

$$|n, l, m\rangle = \int d\Omega\, Y_{lm}(\Omega) |n, \Omega\rangle, \tag{4.15}$$

and a rigid-rotator term is added to the energy. (These are the famous rotational levels.) Note that the rotational structure is determined purely by group theory; it involves no properties of V that have not entered earlier approximations. In addition, we begin to see the effects of departures from the harmonic approximation, Eq. (4.12). I have indicated these terms (vibrational–vibrational coupling) in the table by triple dots. They depend on the detailed form of V (in particular, on $V'''(r_0)$ and $V^{iv}(r_0)$).

Unlike the rotational term, they do not affect the qualitative features of the problem, nor do the higher terms in the expansion.

The extension of all this to a polyatomic molecule is trivial. Unless the equilibrium configuration of the molecule is one in which all the nuclei are aligned, the equilibrium configurations are labeled, like the positions of a rigid body, by three Euler angles rather than two polar angles. As a consequence of this, the rotational spectrum, once it appears in second order, will be that of a rigid body, rather than that of a rigid rotator. Also, there are many ways to vibrate about equilibrium, and the single integer n is replaced by a string of integers n_i, one for each normal mode. Likewise, in the first-order eigenvalue,

$$\left(\frac{V''}{M}\right)^{\frac{1}{2}}(n+\tfrac{1}{2}) \to \sum_i \left(\frac{K_i}{M}\right)^{\frac{1}{2}}(n_i+\tfrac{1}{2}), \tag{4.16}$$

where the Ks are the spring constants for the normal modes. (There is a possible error here into which one is less likely to fall in the diatomic case. The count of normal modes should not include the three zero-frequency normal modes that correspond to infinitesimal rotations; these degrees of freedom are already taken care of by the Euler angles, and it would be double counting to include them with the ns.)

Now, as promised, to treat the Hamiltonian (4.7) requires hardly more than a change of notation. This produces Table 2, which I will now explain.

To zeroth order, all we have to do is replace $1/M^{\frac{1}{2}}$ by β^2, and divide the β^2 out of the left-hand side of Eq. (4.7). Instead of a degenerate family of eigenstates of the position operator, connected by rotations, we have a degenerate family of eigenstates of the field operator, connected by space translations. I have labeled these by b, the position of the center of the lump.

Table 2. *The quantum lump*

Order of approximation	Energy eigenstate	Energy eigenvalue
	$\lvert b\rangle$	E_0/β^2
1	$\lvert n_1, n_2 \ldots ; b\rangle$	$+\sum_i (n_i+\frac{1}{2})\omega_i$
2	$\lvert n_1, n_2 \ldots ; p\rangle$	$+\dfrac{\beta^2 p^2}{2E_0}+\ldots$

To first order, we merely have to replace the normal-mode spring constants of Eq. (4.16) by the eigenfrequencies of Eq. (2.18). Just as for the

molecule, the count of normal modes does not include the zero-frequency mode, Eq. (2.19). (Of course, the eigenfrequencies form a continuous set, so the sum should be an integral. For the moment, let me imagine we have put the system in a box, so the eigenfrequencies are discrete; I will take care of the continuum problem shortly.) As usual when passing from particle mechanics to field theory, we reinterpret harmonic-oscillator excitation numbers as meson normal-mode occupation numbers. This justifies the last paragraph of Sect. 2.2; the small perturbations about the classical solution do indeed describe mesons interacting with a lump. The state where all the ns vanish, which we denote by $|0; b\rangle$, is an isolated lump; states with non-vanishing n correspond to one or more mesons bound to or passing by the lump. To this order, there is no sign of meson–meson interactions in the energy because meson–meson interactions are of order β^2; their effects are analogous to the vibrational–vibrational coupling in the molecule, and, like it, they are lurking behind the three dots in the second-order energy. In analogy to Eqs. (4.13) and (4.14), if we normalize the first-order ground states, for example, such that

$$\langle 0; b'|0; b\rangle = \delta(b'-b), \tag{4.17}$$

then

$$\langle 0; b'|\phi'(x)|0; b\rangle = \delta(b'-b)f(x-b). \tag{4.18}$$

In second order, the spurious degeneracy is lifted, and the translation eigenstates are replaced by momentum eigenstates. Just as for the molecule, the proper eigenstates are determined purely by group theory, by the symmetries of the problem,

$$|n_1 \ldots; p\rangle = \int \frac{db}{(2\pi)^{\frac{1}{2}}} e^{ipb}|n_1 \ldots; b\rangle, \tag{4.19}$$

as is the degeneracy-lifting part of the energy,

$$\begin{aligned}(p^2+M^2)^{\frac{1}{2}} &= M + p^2/2M + O(M^{-3}) \\ &= \frac{E_0}{\beta^2} + \frac{\beta^2 p^2}{2E_0} + O(\beta^6).\end{aligned} \tag{4.20}$$

As a consequence of Eqs. (4.18) and (4.19), we can give an interpretation of the classical solution, $f(x)$, in terms of the conventional concepts of particle physics:

$$\langle 0; p'|\phi'(0)|0; p\rangle = \int \frac{dx}{(2\pi)} e^{-i(p-p')x} f(x). \tag{4.21}$$

In words, $f(x)$ is the leading approximation to the Fourier transform of the ϕ' form factor of the quantum lump.

Now let me return to the problems associated with the fact that the sum

over normal modes is actually an integral. Obviously, we can take care of everything if we can take care of the isolated lump, so let us concentrate on this.

I remind you that Eq. (2.18) tells us that the *squares* of the eigenfrequencies are the eigenvalues of the positive operator

$$K = -\frac{d^2}{dx^2} + U''(f). \tag{4.22}$$

If we put the system in a box, so the eigenfrequencies are discrete, the sum of the eigenfrequencies is the trace of the operator $K^{\frac{1}{2}}$, the positive square root of K. Of course, what we are really interested in is not the energy of the lump, but the difference between the energy of the vacuum and that of the lump. Thus we define

$$K_0 = -\frac{d^2}{dx^2} + U''(\phi_0), \tag{4.23}$$

where ϕ_0 is a zero of U. The desired energy difference is

$$\tfrac{1}{2}\left(\sum_i \omega_i|_{\text{lump}} - \sum_i \omega_i|_{\text{vacuum}}\right) = \tfrac{1}{2}\text{Tr}(K^{\frac{1}{2}} - K_0^{\frac{1}{2}}). \tag{4.24}$$

As we send the box to infinity, the left-hand side of this equation becomes difficult to interpret, but the right-hand side remains unambiguous; it is what should properly appear in Table 2.

Although I will not attempt to explicitly compute Eq. (4.24) here for any of our models, I should tell you that the trace is ultraviolet divergent, and also that this is as it should be. There are no counterterms in Eq. (4.7), so the parameters occurring in it (like α for the sine–Gordon equation, or μ^2 for ϕ^4 theory) should be interpreted as bare parameters. As we shall see explicitly in Sect. 4.3, these bare parameters are finite quantities plus divergent terms proportional to β^2 (and sometimes also to higher powers of β). Thus the zeroth order expression for the energy, E_0/β^2, generates a divergence of order β^0, which must cancel against a corresponding divergence in Eq. (4.24) if everything is to remain finite to this order.

I do not have time in these lectures to show explicitly these divergence cancellations, and must refer you to the literature if you want a demonstration. However, I should stress that it would be a genuine shock if the divergences did not cancel. After all, for all renormalizable theories, our experience has been that the usual counterterms are sufficient to get rid of all the divergences in all observable processes – not just scattering amplitudes but also bound-state energies, etc. A quantum lump is just another bound state (I will expand on this remark later); its energy is an observable quantity, and we would expect it to be purged of divergences by renormalization just like every other observable.

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

The extension to three spatial dimensions is straightforward. Of course, since three-dimensional theories with time-independent solutions necessarily involve much more than just a single scalar field, the actual computations are considerably more onerous. Indeed, so far as I know, no one has even made an approximate computation of the first-order correction to the lump energy, Eq. (4.24), for any three-dimensional theory.

The only qualitative difference occurs in the second-order degeneracy-removing term. This is because there is a larger group of geometrical symmetries in three dimensions (the six-parameter Euclidean group), and thus the manifold of minima of V may have as many as six dimensions. There are three cases. (1) The classical lump is spherically symmetric. In this case, the classical solutions are labeled by a three-vector, $\mathbf{a}$, the position of the center of the lump, and the degeneracy-removing term is simply $\beta^2\mathbf{p}^2/2E_0$, just as in one dimension. (2) The classical lump is not spherically symmetric, but it has an axis of symmetry. In this case, the classical solutions are labeled by $\mathbf{a}$ and by the pair of angles θ and ϕ, that give the orientation of the symmetry axis. We then obtain an additional degeneracy-removing term, $\beta^2 s(s+1)/2I_0$, where s is the intrinsic angular momentum of the lump and I_0 is its classical moment of inertia. This is exactly like the situation for the diatomic molecule; the quantum lump has rotational excitations. (3) The classical lump has no axis of symmetry. In this case, in addition to $\mathbf{a}$, we need three Euler angles. This is exactly like the polyatomic molecule; the quantum lump has rotational excitations, and their spectrum is given by the quantum rigid-body formula.

Of course, in any number of dimensions, we can also have degeneracy if our classical lump is not invariant under some unbroken internal symmetry group of the theory. In this case the degeneracy would also be removed in second order, leading to what we might call 'internal rotational excitations'.[30]

We have learned a lot about the structure of those quantum lumps that correspond to time-independent classical solutions. They are really quite like more familiar composite systems, ordinary bound states. There is one main difference. Ordinary bound states are, for weak coupling, approximate eigenstates of particle number. (Thus, in lowest order, positronium is made up of one electron and one positron, but, as we go to higher orders, we find contributions from two electrons and two positrons, etc.) As a result of this, as the coupling vanishes, the mass of an ordinary bound state goes to some multi-particle threshold. In contrast, for weak coupling, the quantum lump is an approximate eigenstate of the field operator, ϕ. This does not commute with particle number, and thus there is no reason for the mass of the lump to go to a threshold as the coupling vanishes. Indeed, as we have seen, it goes to infinity.

4.3 *Time-independent lumps: coherent-state variational method*[31]

In the previous subsection, we tackled the quantum lump by a method from non-relativistic quantum mechanics. This is not the only non-relativistic method that can be used on this problem; the Rayleigh–Ritz variational method can also be used. Of course, to say 'the variational method' is to say nothing until one specifies the trial states; for a large enough set of trial states the variational method is guaranteed to be arbitrarily accurate and arbitrarily impractical.

I will restrict myself to theories of scalar fields in one spatial dimension with non-derivative interactions, like the theories of Sect. 2. This time, the restriction is not made for reasons of simplicity, but because I believe that these are the only theories for which I have a chance of making a reasonable guess for the trial states. The reason is that these theories are locally Fock; this means that if we restrict the theories to a box, their exact energy eigenstates are in ordinary Fock space. This is connected to the mild ultraviolet divergences of these theories; it is definitely not true for theories in more spatial dimensions, where the divergences are worse. Since all trial states that have been proposed in the literature are in Fock space (when restricted to a box), they are guaranteed to be grotesquely bad guesses for the true energy eigenstates in more than one spatial dimension. This does not mean that the variational method is *a priori* inapplicable; it just means that better trial states than any yet proposed are needed.

(In fairness, I should tell you that most workers in this field disagree completely with the preceding paragraph and think that I am a fuss-budget.)

Another advantage of sticking to one spatial dimension is that the divergence structures of the theories are so simple that we can write the Hamiltonian in terms of finite parameters in closed form. Thus we preserve the great advantage of the variational method, that we always obtain an upper bound on the energy. Of course, even in three spatial dimensions we would obtain an upper bound, but it would be useless because it would be expressed in terms of the bare parameters of the theory, known in terms of finite parameters only as infinite power series. There is no advantage in obtaining exact upper bounds on the eigenvalues of a Hamiltonian that is only known approximately.

Thus, before I describe variational computations, I will discuss how to sum up the divergences which arise in all orders of perturbation theory for the theory defined by

$$\mathscr{H} = \tfrac{1}{2}\pi^2 + \tfrac{1}{2}(\partial_1\phi)^2 + U(\phi). \tag{4.25}$$

(Note that we are back with canonical fields, not rescaled fields as in Sect. 4.2.)

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

For this theory, the only ultraviolet divergences that occur in any order of perturbation theory come from graphs that contain a closed loop consisting of a single internal line, that is to say, graphs in which two fields at the same vertex are contracted with each other. For example, the graph of Fig. 9(a) is ultraviolet divergent, while that of Fig. 9(b) is not. Thus, all ultraviolet divergences can be removed by normal-ordering the interaction Hamiltonian in the interaction picture.

However, variational computations are not carried out in the interaction picture, but in the Schrödinger picture, where the Hamiltonian is given as a function of fixed-time coordinates and momenta, as in Eq. (4.25). It is easy to see what interaction-picture normal-ordering becomes in the Schrödinger picture: we simply define the 'creation' and 'annihilation' parts of ϕ and π as if they were free fields.

$$\phi^{(\pm)}(x, m)=\frac{1}{2}\left[\phi(x)\mp\frac{\mathrm{i}}{(-\nabla^2+m^2)^{\frac{1}{2}}}\pi(x)\right], \tag{4.26a}$$

and

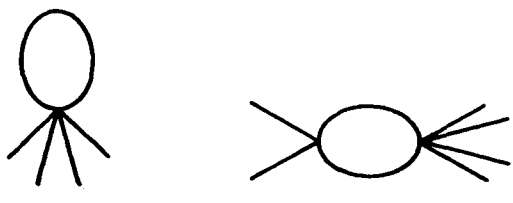
$$\pi^{(\pm)}(x, m)=\mp\mathrm{i}(-\nabla^2+m^2)^{\frac{1}{2}}\phi^{(\pm)}(x, m), \tag{4.26b}$$

where $+$ indicates creation and $-$ annihilation. We then define the normal-ordered version of any function of these operators as the function rearranged with all the creation operators on the left and all the annihilation operators on the right.

Note that this prescription is ambiguous, because m is not specified. Of course, if we were really doing interaction picture perturbation theory, it would be senseless to choose m to be other than that mass which occurs in the free Hamiltonian. However, we are doing variational computations, and we should avoid premature decisions about the best way to divide H into a free and an interacting part. Thus, for the moment, I will keep m free, and denote by N_m the normal-ordering operation defined by the mass m.

(This is off the main line of argument, but it is amusing to consider what happens in interaction-picture perturbation theory, if m is chosen

Fig. 9



https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

to be different from μ, the mass in the free Hamiltonian. In this case, it is easy to see that the divergent loop integral of Fig. 9(a) is replaced by a convergent integral according to the prescription:

$$\int \frac{d^2k}{k^2-\mu^2} \to \int d^2k \left(\frac{1}{k^2-\mu^2} - \frac{1}{k^2-m^2} \right). \tag{4.27}$$

Note that if $m=\mu$ (the conventional rule) the graph is canceled completely (the conventional result). However, even if $m \neq \mu$, the divergent integral is converted into a convergent one.)

Whatever m we choose, we can normal-order the Hamiltonian by Wick's theorem. We begin by computing the contraction of two fields,

$$\phi^2(x) = N_m \phi^2(x) + \int \frac{dk}{4\pi} \frac{1}{(k^2+m^2)^{\frac{1}{2}}}. \tag{4.28}$$

The integral is divergent; as usual, we introduce a high-momentum cutoff and find

$$\int_{-\Lambda}^{\Lambda} \frac{dk}{4\pi} \frac{1}{(k^2+m^2)^{\frac{1}{2}}} = \frac{1}{4\pi} \ln \frac{4\Lambda^2}{m^2}, \tag{4.29}$$

for $\Lambda \gg m$. Whence, for any function of ϕ,

$$U(\phi) = N_m \left\{ \exp \left[\frac{1}{8\pi} \left(\ln \frac{4\Lambda^2}{m^2} \right) \frac{d^2}{d\phi^2} \right] U(\phi) \right\}. \tag{4.30}$$

This is just Wick's theorem; the successive terms in the expansion of the exponential give us the terms with no contractions, one contraction, etc.

We can also use Wick's theorem to pass from one normal-ordering prescription to another. Since, from Eq. (4.28),

$$N_m \phi^2 = N_\mu \phi^2 + \frac{1}{4\pi} \ln \frac{m^2}{\mu^2}, \tag{4.31}$$

it follows that

$$N_m U(\phi) = N_\mu \left\{ \exp \left[\frac{1}{8\pi} \left(\ln \frac{m^2}{\mu^2} \right) \frac{d^2}{d\phi^2} \right] U(\phi) \right\}. \tag{4.32}$$

Note that this is free of all reference to the cutoff, as it should be.

Equations (4.30) and (4.32) take care of the last term in the Hamiltonian density, Eq. (4.25). The first two terms can be treated similarly. For brevity of notation, I define

$$\mathscr{H}_0 = \tfrac{1}{2}\pi^2 + \tfrac{1}{2}(\partial_1 \phi)^2. \tag{4.33}$$

Normal-ordering this is trivial, since it is a mere quadratic function of ϕ and π; thus

$$\mathscr{H}_0 = N_m \mathscr{H}_0 + \text{divergent constant}. \tag{4.34}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

An easy integration reveals that

$$N_m \mathscr{H}_0 = N_\mu \mathscr{H} + \frac{1}{8\pi}(\mu^2 - m^2). \tag{4.35}$$

To have a definite example, let me apply these formulae to the sine–Gordon equation,

$$\mathscr{H} = \mathscr{H}_0 - \frac{\alpha_0}{\beta^2}\cos\beta\phi + \gamma_0. \tag{4.36}$$

Here I have put a nought on α, in anticipation of its impending renormalization. I have also added a constant to the Hamiltonian density, to acknowledge the fact that we do not know the ground-state energy of the quantum theory. γ_0 is not a new free parameter; it is to be chosen at the end of our computations so the ground-state energy is zero. As we shall see immediately, there is no need to renormalize β.

Applying Eqs. (4.30) and (4.34), we see that, for any m,

$$\mathscr{H} = N_m\left[\mathscr{H}_0 - \frac{\alpha_0}{\beta^2}\left(\frac{m^2}{4\Lambda^2}\right)^{\beta^2/8\pi}\cos\beta\phi + \gamma_0 + \text{divergent constant}\right]. \tag{4.37}$$

Thus we can absorb all divergences into a multiplicative renormalization of α and an additive renormalization of γ. That is to say, we define

$$\alpha = \alpha_0\left(\frac{m^2}{4\Lambda^2}\right)^{\beta^2/8\pi}, \tag{4.38}$$

and

$$\gamma = \gamma_0 + \text{divergent constant}, \tag{4.39}$$

and obtain

$$\mathscr{H} = N_m\left[\mathscr{H}_0 - \frac{\alpha}{\beta^2}\cos\beta\phi + \gamma\right]. \tag{4.40}$$

We can now let the cutoff go to infinity, and we shall. From now on we will work only with the finite form of $\mathscr{H}$, Eq. (4.40). Of course, if at some future time we find we are unhappy with our original choice of m, we can change it in a moment, using Eqs. (4.32) and (4.36),

$$\mathscr{H} = N_\mu\left[\mathscr{H}_0 - \frac{\alpha}{\beta^2}\left(\frac{\mu^2}{m^2}\right)^{\beta^2/8\pi} + \gamma + \frac{1}{8\pi}(\mu^2 - m^2)\right]. \tag{4.41}$$

Note that the result of changing m can be absorbed in a redefinition of α and γ. In this sense the arbitrary normal-ordering mass m is like the arbitrary normalization mass that appears in the theory of the renormalization group.

We are now in a position to do a simple sample variational computation. Let us consider

$$\mathscr{H} = N_m[\mathscr{H}_0 + \tfrac{1}{2}M^2\phi^2]. \tag{4.42}$$

This is a free-field Hamiltonian density, perversely normal-ordered; that is to say, m is not necessarily equal to M. Of course, any fool can solve this problem exactly, and there is no real need to use the variational method. Nevertheless, the computation is instructive.

As our trial states, let us choose the vacuum states appropriate to a free field of mass μ. These are defined by

$$\phi^{(-)}(x, \mu)|0, \mu\rangle = \pi^{(-)}(x, \mu)|0, \mu\rangle = 0. \tag{4.43}$$

Since these states are translationally invariant, we need only compute the expectation value of the Hamiltonian density. This is made trivial by our re-normal-ordering equations,

$$\mathscr{H} = N_\mu\left[\mathscr{H}_0 + \tfrac{1}{2}M^2\phi^2 + \frac{1}{8\pi}\left(\mu^2 - m^2 - M^2 \ln\frac{\mu^2}{m^2}\right)\right]. \tag{4.44}$$

Whence,

$$\langle 0, \mu|\mathscr{H}|0, \mu\rangle = \frac{1}{8\pi}\left(\mu^2 - m^2 - M^2 \ln\frac{\mu^2}{m^2}\right). \tag{4.45}$$

This completes the first step of the Rayleigh–Ritz method, the computation of the expectation value of the Hamiltonian in the trial state. The next step is to differentiate this with respect to the variational parameter, μ,

$$1 - \frac{M^2}{\mu^2} = 0. \tag{4.46}$$

Hence the best trial state is $|0, M\rangle$. Of course, this is also the exact ground state. This is reassuring.

Now let us try exactly the same procedure, with exactly the same trial states, for the sine–Gordon equation. From Eq. (4.41),

$$\langle 0, \mu|\mathscr{H}|0, \mu\rangle = -\frac{\alpha}{\beta^2}\left(\frac{\mu^2}{m^2}\right)^{\beta^2/8\pi} + \gamma + \frac{1}{8\pi}(\mu^2 - m^2). \tag{4.47}$$

Now comes the surprise: this expression is unbounded below, as μ goes to infinity, if β^2 exceeds 8π. The variational method always gives a rigorous upper bound on the ground-state energy; in this case the upper bound is minus infinity. We have rigorously proved that the energy of the sine–Gordon theory is unbounded below if β^2 exceeds 8π; the theory has no ground state, even if we put it in a box, and is physically nonsensical. If β^2 is less than 8π, Eq. (4.47) does have a minimum; of course, this does not prove that the theory is physically sensible in this case.

Although we have gone surprisingly far with the states $|0, \mu\rangle$, they are obviously not a rich enough set to be good trial states for the vacuum in the general case, let alone for a lump. For one thing, the expectation value of ϕ vanishes in any such state. I will now describe a family of states that do not have this deficiency; they are called coherent states.

Coherent states are best explained for a very simple system, a harmonic oscillator,

$$H = \tfrac{1}{2}(p^2 + q^2). \tag{4.48}$$

As usual, we define the raising and lowering operators,

$$a = (q + ip)/\sqrt{2}, \tag{4.49a}$$

$$a^\dagger = (q - ip)/\sqrt{2}. \tag{4.49b}$$

A coherent state is labeled by an arbitrary complex number, z, and is defined by

$$|z\rangle = e^{-\frac{1}{2}z^*z}\, e^{za^\dagger}|0\rangle, \tag{4.50}$$

where $|0\rangle$ is the oscillator ground state. It is easy to check that

$$\langle z|z\rangle = 1, \tag{4.51}$$

and

$$a|z\rangle = z|z\rangle, \tag{4.52a}$$

or, equivalently,

$$\langle z|a^\dagger = \langle z|z^*. \tag{4.52b}$$

From these equations, we see that the complex number z can be given in terms of the expectation values of p and q,

$$z = \frac{1}{\sqrt{2}}(\langle q\rangle + i\langle p\rangle). \tag{4.53}$$

For any normal-ordered function of p and q,

$$\langle :F(p, q):\rangle = F(\langle p\rangle, \langle q\rangle). \tag{4.54}$$

This is because normal ordering puts all the lowering operators on the right, where we can use Eq. (4.52a), and all the raising operators on the left, where we can use Eq. (4.52b).

All of this extends trivially to an infinite assembly of harmonic oscillators, a free field of mass μ. Here, starting from the state $|0, \mu\rangle$, we can construct a family of states labeled by two arbitrary functions of space $\langle\phi(x)\rangle$ and $\langle\pi(x)\rangle$. We will use these as trial states to find the energy eigenstates of the theory defined by

$$\mathscr{H} = N_m[\mathscr{H}_0 + U(\phi)]. \tag{4.55}$$

Just as before, the computation is simplified by re-normal-ordering,

$$\mathscr{H} = N_\mu[\mathscr{H}_0 + U(\phi, \mu)], \tag{4.56}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

where, by our re-normal-ordering equations,

$$U(\phi, \mu) = \exp\left[\frac{1}{8\pi}\left(\ln\frac{m^2}{\mu^2}\right)\frac{d^2}{d\phi}\right] U(\phi) + \frac{1}{8\pi}(\mu^2 - m^2). \tag{4.57}$$

Thus,

$$\langle \mathscr{H}(x)\rangle = \tfrac{1}{2}\langle \pi(x)\rangle^2 + \tfrac{1}{2}(\partial_1\langle \phi(x)\rangle)^2 + U(\langle \phi(x)\rangle, \mu). \tag{4.58}$$

The easiest thing to compute is the best trial state(s) for the ground state(s) of the theory. The first two terms in (4.58) are obviously minimized by choosing $\langle \pi(x)\rangle$ to be zero and $\langle \phi(x)\rangle$ to be a constant. The ground state(s) are then determined by the absolute minimum (minima) of $U(\langle \phi\rangle, \mu)$. This is almost as easy as the corresponding computation in the classical case; the only difference is that we have to search for minima in a two-parameter space rather than a one-parameter space. Once the best values of $\langle \phi\rangle$ and μ have been found, it is convenient to add a constant to Eq. (4.58) so the ground-state energy is zero.

Let us denote the value of μ found from the ground-state computation by μ^*. Any of our coherent states that do not have μ equal to μ^* will have an infinite energy expectation value, for the space integral of the last term in Eq. (4.58) will diverge. Thus the problem of finding other variational energy eigenstates is *identical* to the problem of Sect. 2.1, the problem of finding classical time-independent lumps, with the classical U equal to $U(\phi, \mu^*)$. Hence if the ground-state variational computation yields more than one ground state, we will always obtain a variational lump, the energy of the lump will be given by the classical formula, etc.

In atomic physics, variational computations cut across the perturbation expansion, picking up a part of every order of perturbation theory. This is also the case here. The weak-coupling expansion of Sect. 4.2 encountered infinities, which had to be removed by renormalization. For the sine–Gordon equation, as we can see from Eq. (4.38), these infinities occur in all orders of β^2. The variational computation is phrased in terms of finite parameters; it is 'already renormalized', and therefore contains at least a piece of all orders of the β^2 expansion. On the other hand, it does not contain all of each order. In particular, it does not contain the most significant feature of the second-order computation, the removal of the classical translational degeneracy. Our coherent states are not momentum eigenstates (except for the ground-state trial states), and we have an infinite degenerate family of variational energy eigenstates, labeled by the position of the center of the lump.

There is another way to look at this. Again, let me use the sine–Gordon equation as an example. The method of Sect. 4.2 gives us a power series in β^2, with α_0 held fixed. We can turn this into a renormalized expansion by

expressing α_0 as a finite parameter, α, times a power series in β^2, with divergent coefficients, and regrouping terms in the usual way. There are, of course, many such renormalized expansions, because there are many ways of defining α, many possible renormalization conventions. If we choose α to be the coefficient which occurs in $U(\phi, \mu^*)$, then, for this definition (and for no other), the variational computation is identical with the zeroth order of the renormalized series expansion. In this way of looking at things, the variational computation of the ground state serves to determine the 'best choice' of renormalized parameters for the series expansion.

Of course, it is always possible to improve a variational computation by enlarging the set of trial states. For example, it is possible to enforce translational invariance by building momentum eigenstates out of superpositions of coherent states. Also, it is possible to enlarge the set of trial states by generalizing the concept of normal-ordering, replacing $(-\nabla^2+m^2)$ in Eq. (4.26) by a more general positive operator. Such enlargements have been considered in the literature, but I will not discuss them here. The calculations become considerably more difficult, and the qualitative features of the output are unchanged.

As I said at the beginning, I do not believe there are any plausible variational computations using coherent states or simple generalizations of them in more than one spatial dimension. It would be nice to have reasonable variational states in this case. A good place to begin exploring would be a super-renormalizable theory in two spatial dimensions. Here, although all divergences are not removed by normal ordering, it is still possible to sum them up in closed form. Thus we have at least a minimal criterion for a reasonable trial state: the expectation value of the Hamiltonian density should not be infinite. Coherent states do not meet this test.

4.4 *Periodic lumps: the old quantum theory and the DHN formula*

Time-independent solutions are not the only non-dissipative solutions of classical field theories, although they are certainly the easiest to study. Among others, there are periodic solutions.

Once again, the sine–Gordon equation provides a useful example, for an exact periodic solution is known:

$$\phi(x, t)=\frac{4}{\beta}\tan^{-1}\left[\frac{\eta \sin \omega t}{\cosh \eta\omega x}\right], \tag{4.59}$$

where

$$\eta=(\alpha-\omega^2)^{\frac{1}{2}}/\omega, \tag{4.60}$$

and ω ranges from 0 to α. This solution has a simple physical interpreta-

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

tion, which can most easily be seen if $\omega \ll \alpha$. In this case,

$$\phi(x, t) \approx \frac{4}{\beta} \tan^{-1}\left[\frac{\alpha^{\frac{1}{2}} \sin \omega t}{\omega \cosh \alpha^{\frac{1}{2}} x}\right]. \tag{4.61}$$

This simplifies further if we restrict ourselves to portions of the period such that

$$\omega \ll \alpha^{\frac{1}{2}} |\sin \omega t|. \tag{4.62}$$

Let us also assume $\sin \omega t$ is positive. In this case, the argument of the tangent is very large near $x = 0$, and ϕ is approximately $2\pi/\beta$. As we go to large negative x, $\cosh \alpha^{\frac{1}{2}} x$ eventually grows large and ϕ departs from $2\pi/\beta$. In this region, we may approximate $\cosh \alpha^{\frac{1}{2}} x$ by an exponential, and obtain,

$$\phi(x, t) \approx \frac{4}{\beta} \tan^{-1} \exp[\alpha^{\frac{1}{2}} x + \ln(2\alpha^{\frac{1}{2}} \sin \omega t/\omega)]. \tag{4.63}$$

That is to say, there is a soliton far to the left. Likewise, there is an antisoliton far to the right. As $\sin \omega t$ increases to one, the soliton and antisoliton move farther apart from each other. When $\sin \omega t$ passes through one, they turn around and begin to approach each other. As $\sin \omega t$ comes down to zero, Eq. (4.62) no longer applies, and we can no longer describe ϕ in terms of a widely separated soliton–antisoliton pair. This is reasonable, because the soliton and antisoliton are on top of each other. However, we can pick up the description again when $\sin \omega t$ becomes negative, and the soliton and antisoliton have passed each other.

Thus, Eq. (4.59) can be thought of as a soliton and an antisoliton oscillating about their common center-of-mass. For this reason, it is called 'the doublet solution'.

For a later computation, we will need the energy of the doublet. The calculation is done most quickly at $t = 0$:

$$\begin{aligned} E &= \tfrac{1}{2} \int \mathrm{d}x (\partial_0 \phi)^2 \\ &= \frac{8\eta^2\omega^2}{\beta^2} \int \mathrm{d}x \frac{1}{(\cosh \eta\omega x)^2} \\ &= \frac{16\eta\omega}{\beta^2} \\ &= 2M\left(1 - \frac{\omega^2}{\alpha}\right)^{\frac{1}{2}}, \end{aligned} \tag{4.64}$$

where

$$M = 8\alpha^{\frac{1}{2}}/\beta^2, \tag{4.65}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

is the soliton mass. Note that the mass of the doublet is always less than twice the soliton mass, as we would expect for bound motion of a soliton–antisoliton pair.

We would expect such bound classical motions to become quantum bound states, energy eigenstates, in the quantum version of the theory. Our problem is to compute, in some plausible approximation, the energy eigenvalues. Fortunately for us, a crude but serviceable approximation is at hand, for this is the fundamental problem of the old quantum theory, solved by Bohr and Sommerfeld. Their quantization formula says that if we have a one-parameter family of periodic motions, labeled by the period, T, then an energy eigenstate occurs whenever

$$\int_0^T \mathrm{d}t p\dot{q} = 2\pi n, \tag{4.66}$$

where n is an integer. Note that this is *not* the WKB formula, which has $n+\frac{1}{2}$ rather than n on the right. The WKB formula is a refinement of Eq. (4.66), but it is a refinement only for special circumstances, a particle moving in a potential. (For example, there is no $\frac{1}{2}$ in the corresponding quantization condition for angular motion.) Eq. (4.66) is cruder than the WKB formula, but it is much more general; it is always the leading approximation for any dynamical system.

For the benefit of those of you who have forgotten the stratagems of the old quantum theory, I will briefly recapitulate the arguments that lead to Eq. (4.66). For any classical system, the energy is given by

$$E = p\dot{q} - L. \tag{4.67}$$

Integrating over a period, we find

$$ET = 2\pi n - \int_0^T \mathrm{d}t L \tag{4.68}$$

where n is defined by Eq. (4.66). Differentiating this with respect to T, with $q(0)$ fixed, we find

$$E\mathrm{d}T + T\mathrm{d}E = 2\pi \mathrm{d}n - L(T)\mathrm{d}T + p(T)\dot{q}(T)\mathrm{d}T. \tag{4.69}$$

(The last term comes from differentiating the integrand, integrating by parts, and using the Euler–Lagrange equations.) Thus,

$$\mathrm{d}E = \omega \mathrm{d}n. \tag{4.70}$$

Thus, if the motion is quantized at integer ns, the frequency of the quantum emitted when the system makes a transition from one quantum state to the next is the same as the frequency of the classical motion. This is Bohr's correspondence principle. (Of course, the WKB formula also obeys Eq.

(4.70). In this way of looking at things, the only new thing produced by WKB analysis is the integration constant for Eq. (4.70).)

There is nothing in this argument that makes any reference to the details of the dynamics, and it was only notational laziness that made me carry it through only for a system with a single degree of freedom. It is just as valid for a system with an infinite number of degrees of freedom, that is to say, a field theory:

$$\int_0^T \mathrm{d}t \int \mathrm{d}x \pi(x, t) \partial_0 \phi(x, t) = 2\pi n. \tag{4.71}$$

I will now apply this formula to the sine–Gordon doublet. The fastest way to do the integral is to integrate Eq. (4.70),

$$\frac{\mathrm{d}n}{\mathrm{d}E} = \omega^{-1} = \alpha^{-\frac{1}{2}} \left(1 - \frac{E^2}{4M^2}\right)^{-\frac{1}{2}}, \tag{4.72}$$

and use as boundary condition the obvious fact that the left-hand side of Eq. (4.71) vanishes when $E = 0$. Thus we find a sequence of quantum states,

$$E_n = 2M \sin(\beta^2 n/16) \tag{4.73}$$

where $n = 0, 1, 2, \ldots < 8\pi/\beta^2$.

We can check this formula for small β. In this case,

$$E_n = n\alpha^{\frac{1}{2}} + O(\alpha^{\frac{1}{2}}\beta^2). \tag{4.74}$$

Now, $\alpha^{\frac{1}{2}}$ is the mass of the fundamental meson, for small β. It is not surprising that the fundamental meson should emerge as the lowest quantum state of the doublet, other than the vacuum. For small E, the soliton and antisoliton never separate much, and Eq. (4.59) looks much more like a small perturbation of the ground state than an oscillating pair. The other quantum states also have a simple interpretation. For small coupling, the fundamental mesons have only weak interactions among themselves; if we expand the interaction in powers of β, the leading term is an attractive quartic interaction proportional to β^2. A weak attractive interaction makes only weakly bound states, for which non-relativistic reasoning is sufficient. In the non-relativistic limit, the quartic interaction becomes an attractive two-body delta-function potential. It is known that for a system of identical bosons in one dimension, such an interaction produces exactly one two-body bound state, exactly one three-body bound state, etc. These are the other quantum states in Eq. (4.74). (You might wonder if all of these states are really stable, since there is no obvious conservation law that keeps a four-meson bound state from decaying into two mesons. It turns out that, in tree approximation, all of the bound states *are* stable. The secret reason for this is that the classical sine–Gordon equation possesses an infinite number of non-obvious (but also non-trivial) con-

servation laws. Whether the higher states are stable when loops are taken into account is at the moment unknown.)

It is worth pointing out that the spacing of energy levels predicted by Eq. (4.74) is independent of β for small β. This is a general feature of Bohr–Sommerfeld quantization applied to any classical field theory with periodic solutions; it is simply a consequence of Eq. (4.70) and the fact that ω is not affected when we scale out β, as in Sect. 4.1. On the other hand, the lowest energy eigenvalue is independent of β for small β only because the energies of our classical periodic motions go all the way down to zero. If this were not the case, the lowest eigenvalue would be proportional to β^{-2}.

In the theory of a particle in a potential, there is a more accurate formula than the Bohr–Sommerfeld quantization formula, the WKB formula. Likewise, there is a more accurate formula for periodic lumps, the analog of the WKB formula. This formula was derived and applied to several cases in a brilliant sequence of papers by Dashen, Hasslacher, and Neveu;[32] I will call it the DHN formula. The derivation of the DHN formula is extremely lengthy, and I do not have time even to sketch it in these lectures. However, the final formula is simple enough to state, once some preliminary definitions are made. As usual, for notational simplicity, I will restrict myself to the case of a single scalar field in one space dimension; the generalization to more complicated theories is trivial.

Let us consider a one-parameter family of periodic solutions of our theory, parametrized by the period T. I denote these solutions by $\phi_T(x, t)$. Let us study small perturbations about one of these solutions,

$$\phi(x, t) = \phi_T(x, t) + \delta(x, t). \tag{4.75}$$

Inserting this in the equation of motion, and retaining only terms linear in δ, we obtain

$$\Box^2\delta + U''(\phi_T)\delta = 0. \tag{4.76}$$

This equation is invariant under time translations by T. Thus, following closely the procedures of Sect. 2.2 for a time-independent lump, we look for solutions that are eigenfunctions of such a time translation, that is to say, solutions obeying

$$\delta_i(x, t+T) = e^{-i\nu_i}\delta_i(x, t).$$

The νs are called stability angles. The reason for this terminology is that ϕ_T is infinitesimally stable if and only if all the νs are real. From now on I will assume that this is the case. Because of space-time translational invariance, there are two eigenfunctions, $\partial_\mu\phi_T$, with zero stability angles. Because Eq. (4.76) is real, the eigenfunctions with non-zero stability angles

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

occur in complex-conjugate pairs; the stability angles are of the same magnitude but opposite sign.

The interpretation of the δs is much the same as in Sect. 2.2. Discrete eigenfunctions can be thought of as mesons bound to the oscillating lump, continuum eigenfunctions as mesons scattering off the lump. For notational simplicity I will treat all eigenfunctions as discrete. (We can always put the system in a box, and let the box go to infinity at the end of our computations.)

I am now in a position to state the DHN formula.[33] Let n_i be a sequence of non-negative integers, one for each positive stability angle. For each such sequence, we select a set of periodic classical motions by the quantization condition

$$\int_0^T dt \int dx \pi_T(x,t)\partial_0 \Phi_T(x,t) + \sum_{\nu_i>0} (n_i+\tfrac{1}{2})\left(T\frac{d\nu_i}{dT} - \nu_i\right) = 2n\pi, \tag{4.77}$$

where n is an integer. Let us denote the energy of one of these selected classical motions by E_{cl}. Then there is a quantum energy eigenstate with eigenvalue

$$E = E_{cl} + \sum_{\nu_i>0} (n_i+\tfrac{1}{2})\frac{d\nu_i}{dT}. \tag{4.78}$$

Some remarks:

(1) There is an obvious similarity between Eq. (4.78) and the first-order term in the weak-coupling expansion of Sect. 4.2. Actually, the weak-coupling expansion can be thought of as a degenerate case of the DHN formula. A time-independent solution is *ipso facto* a periodic solution, of arbitrary period T, and thus can be thought of as constituting a one-parameter family of periodic solutions all by itself. In the notation of Sect. 4.2, E_{cl} is E_0/β^2, and ν_i is $\omega_i T$. Eq. (4.77) is trivially satisfied for $n=0$, and Eq. (4.78) becomes the first-order approximation of Sect. 4.2.

(2) Just like the sum of frequencies in Sect. 4.2, the sum over stability angles in the DHN formula is divergent. Just as in Sect. 4.2, the divergence is canceled once the bare parameters in the classical theory are re-expressed in terms of renormalized parameters.

(3) The DHN formula obeys the correspondence principle. The differential of Eq. (4.77), with fixed n_i, is

$$T dE_{cl} + \sum (n_i+\tfrac{1}{2}) T \frac{d^2\nu_i}{dT^2} dT = 2\pi dn. \tag{4.79}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

The differential of Eq. (4.78) is

$$dE = dE_{cl} + \sum (n_i + \tfrac{1}{2}) \frac{d^2 \nu_i}{dT^2} dT. \tag{4.80}$$

Thus,

$$T dE = 2\pi dn. \tag{4.81}$$

This is Eq. (4.70).

(4) Equation (4.77) is ambiguous. A stability angle is defined only modulo 2π, but if we add 2π to one of the stability angles in Eq. (4.77), we change the equation. On the other hand, if we add 4π to a stability angle (or 2π to each of two angles), we make no change; the result can be absorbed in n. Thus we are missing one binary bit of information. It is possible to give rules for eliminating this ambiguity, but I will not do so here. In the example at hand, the sine–Gordon doublet, the stability angles are unambiguously known for $T = 2\pi/\alpha^{\frac{1}{2}}$, when the doublet becomes the ground state, and are known for other T by continuity.

In a spectacular analysis, Dashen, Hasslacher, and Neveu[34] computed the stability angles for the sine–Gordon doublet and evaluated the sum over stability angles. Their results are: (1) There are no discrete stability angles. Thus there is only a single series of bound states, labeled by the integer n. (2) The energies of these bound states are given by

$$E_n = 2M \sin(\beta'^2 n/16), \tag{4.82}$$

where $n = 0, 1, 2 \ldots < 8\pi/\beta'^2$,

$$\beta'^2 = \frac{\beta^2}{1 - \beta^2/8\pi}, \tag{4.83}$$

and M is the soliton mass. M is not given by Eq. (4.65), but is the soliton mass corrected by the DHN formula, or, equivalently, by the first-order weak coupling expansion. (See remark (1) above.) I have written the equation in this form, rather than giving separate expressions for E_n and M, to eliminate α, and thus avoid worries about renormalization conventions. Note that the DHN formula is identical to the Bohr–Sommerfeld formula, except that β is replaced by β'.

Dashen *et al.* speculate that Eq. (4.82) is exact. They have three reasons for this. (1) Equation (4.83) blows up at 8π. This is precisely the place where we found the bottom dropping out of the energy spectrum in Sect. 4.3. (2) Since the first quantum doublet state is just the ordinary meson, and the second is just a two-meson bound state, the ratio E_2/E_1 can be computed in ordinary perturbation theory through the Bethe–Salpeter equation. Dashen *et al.* have done this computation up to and including

terms of order β^8. The result agrees with Eq. (4.82) to this order. (3) We shall get to the third reason in Sect. 5.

I emphasize that even if the DHN formula does turn out to be exact for the sine–Gordon doublet, this is pure fortuity, like the exactness of the WKB formula for the hydrogen atom. If a ϕ^6 term is added to the sine–Gordon Lagrangian, the DHN formula and the perturbation expansion no longer agree. From our viewpoint, the possible exactness of the DHN formula for this special system is unfortunate. As we shall see in Sect. 5, the sine–Gordon equation is a marvelous theoretical laboratory for checking various approximation schemes, and the exactness of the DHN formula makes the sine–Gordon laboratory as useless for studying the errors in the DHN approximation as the hydrogen atom would be for studying the errors in the WKB approximation.

5 A very special system

5.1 *A curious equivalence*[35]

Several times in these lectures I have alluded to a peculiar property of the sine–Gordon equation. It is now time to reveal the secret: the sine–Gordon equation is equivalent, in a sense I shall make precise, to the massive Thirring model. This is a surprise, because the massive Thirring model is a canonical field theory whose Hamiltonian is expressed in terms of fundamental Fermi fields only. Even more surprising, when $\beta^2 = 4\pi$, the sine–Gordon equation is equivalent to a free massive Dirac theory, in one spatial dimension.

To establish these results, I will have to devote a little time to defining the massive Thirring model.

The (massless) Thirring model[36] is a theory of a single Dirac field in one spatial dimension, with

$$\mathscr{L} = \bar{\psi} i \gamma_\mu \partial^\mu \psi - \tfrac{1}{2} g j_\mu j^\mu, \tag{5.1}$$

where

$$j^\mu = \bar{\psi} \gamma^\mu \psi, \tag{5.2}$$

and g is a free parameter, the coupling constant. The formal definition of j_μ as a product of two fields at the same point is plagued with ambiguities, just as in higher-dimensional theories, but, again just as in higher-dimensional theories, these ambiguities can be removed by demanding that

$$[j^0(x, t), \psi(y, t)] = -\delta(x - y)\psi(y, t). \tag{5.3}$$

Once this has been done, the Hamiltonian derived from Eq. (5.1) requires no further renormalizations, aside from a zero-point energy subtraction. The model is exactly soluble and physically sensible for g greater than minus π.

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

Within the massless model, it is possible to define renormalized chiral eigendensities,

$$\sigma_{\pm} = Z\bar{\psi}(1 \pm \gamma_5)\psi, \tag{5.4}$$

where Z is a cutoff-dependent constant. (Note that Z is determined only up to a finite multiplicative constant.) The massive Thirring model is formally defined by adding a mass term to the Hamiltonian density of the massless model:

$$\mathscr{H} \to \mathscr{H} + \mathscr{M} \tag{5.5}$$

where

$$\mathscr{M} = \tfrac{1}{2}m'(\sigma_+ + \sigma_-). \tag{5.6}$$

Here m' is merely a real parameter; it is not to be identified with the mass of any presumed one-particle state. The massive model is not exactly soluble; indeed, it is a difficult and still not totally solved problem even to show that these equations define a physically sensible theory.[37] However, it is certainly true that the matrix elements of $\mathscr{M}$ between any two states of the massless model is a well-defined object, free of ultraviolet divergences. If the massive model does exist, these matrix elements should be sufficient to define it.

In fact, I will not study matrix elements between general states, but only states obtained by applying space-time smeared polynomials in $\sigma_{\pm}$ to the vacuum state. This family of states is invariant under both $\mathscr{M}$ and the massless Hamiltonian; therefore, a state in this subspace should stay in this subspace under the time evolution defined by the massive Hamiltonian. For this subspace, the massive model is defined by the vacuum expectation value of a string of $\sigma_{\pm}$s, computed in the massless model. Some of these will be smeared to make the state on the right, some smeared to make the state on the left, and one left unsmeared to make $\mathscr{M}$.

These vacuum expectation values can be found in the literature on the massless model.[38] I will write down the explicit form only for the case when all separations are space-like; the answers for other configurations are uniquely determined from this expression by analytic continuation.

$$\left\langle 0 \middle| \prod_{i=1}^{n} \sigma_+(x_n)\sigma_-(y_n) \middle| 0 \right\rangle = A^{2n} \frac{\prod_{i>j} [(x_i - x_j)^2 (y_i - y_j)^2]^{\delta}}{\prod_{i,j} [(x_i - y_j)^2]^{\delta}}, \tag{5.7}$$

where

$$\delta = \frac{1}{1 + g/\pi}, \tag{5.8}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

and A is an arbitrary constant, undetermined until we fix the renormalization convention that determines the finite part of Z. For this section only, I have defined

$$x^2 = -x_\mu x^\mu, \tag{5.9}$$

so the square of a spacelike vector is positive. Because of the chiral invariance of the massless model, expectation values of strings with unequal numbers of σ_+s and σ_-s vanish.

We now have all the information about the Thirring model we need to prove the equivalence.

Now let us consider the theory of a free massless scalar field

$$\mathscr{H} = \tfrac{1}{2} N_m[\pi^2 + (\nabla\phi)^2] = N_m \mathscr{H}_0, \tag{5.10}$$

where m is an arbitrary normal-ordering mass. In this theory, let us define

$$S_\pm = N_m e^{\pm i\beta\phi}, \tag{5.11}$$

where β is a real number.

I will shortly demonstrate that the vacuum expectation of a string of Ss is given by an expression of exactly the form of Eq. (5.7) with

$$\delta = \beta^2/4\pi. \tag{5.12}$$

Thus, we can identify the space of states made by applying strings of Ss to the vacuum in the scalar theory with the set of states made by applying strings of σs to the vacuum in the Thirring model. Furthermore, we can identify the mass term in the Thirring model with the sine–Gordon interaction,

$$\mathscr{M} = -\frac{\alpha}{\beta^2} N_m \cos\beta\phi. \tag{5.13}$$

Of course, in order to do this consistently, we must match Eqs. (5.8) and (5.12), that is, we must say

$$\frac{\beta^2}{4\pi} = \frac{1}{1+g/\pi}. \tag{5.14}$$

Equations (5.13) and (5.14) define the correspondence between the massive Thirring model and the sine–Gordon equation.[39] Note that if $\beta^2 = 4\pi$, $g = 0$, and the sine–Gordon equation is the theory of a free massive Dirac field.

Now for the computation of the expectation value of a string of Ss. The computation is complicated by the extreme infrared divergences of a massless scalar field in two dimensions. For example, for a free scalar field of mass μ,

$$\triangle_+(x;\mu) = -\frac{1}{4\pi}\ln(c\mu^2x^2) + O(\mu^2x^2), \tag{5.15}$$

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

where c is a numerical constant, related to Euler's constant. This blows up badly when μ goes to zero. I will circumvent this problem by adding a small mass term to Eq. (5.10),

$$N_m \mathscr{H}_0 \to N_m[\mathscr{H}_0 + \tfrac{1}{2}\mu^2\phi^2] = N_\mu[\mathscr{H}_0 + \tfrac{1}{2}\mu^2\phi^2], \tag{5.16}$$

plus an irrelevant constant. At the end of the computation I will let μ go to zero.

The computation is simplified by re-normal-ordering Eq. (5.11),

$$N_m e^{\pm i\beta\phi} = \left(\frac{\mu^2}{m^2}\right)^{\beta^2/8\pi} N_\mu e^{\pm i\beta\phi}. \tag{5.17}$$

Thus, what we wish to compute is

$$\left(\frac{\mu^2}{m^2}\right)^{\Sigma\beta_i^2/8\pi} \left\langle 0, \mu \left| \prod_i N_\mu e^{i\beta_i\phi(x_i)} \right| 0, \mu \right\rangle, \tag{5.18}$$

where the β_is are constants, equal in the case at hand to either plus or minus β. The computation is a trivial application of Wick's theorem. Because the operators are normal-ordered with respect to the free-field mass, μ, there are no contractions of an operator with itself, and thus we obtain

$$\left(\frac{\mu^2}{m^2}\right)^{\Sigma\beta_i^2/8\pi} \prod_{i>j} e^{-\beta_i\beta_j\Delta_+(x_i - x_j;\,\mu)}. \tag{5.19}$$

We now let μ go to zero, and use Eq. (5.15). We find, in the small μ limit,

$$\left(\frac{\mu^2}{m^2}\right)^{\Sigma\beta_i^2/8\pi} \prod_{i>j} [c\mu^2(x_i - x_j)^2]^{\beta_i\beta_j/4\pi}. \tag{5.20}$$

The most important thing about this expression is that it is proportional to

$$(\mu^2)^{(\Sigma\beta_i)^2/8\pi}. \tag{5.21}$$

Thus, if the βs do not sum to zero, that is to say, if there are not equal numbers of S_+s and S_-s, the limit is zero. This is the recovery in the scalar theory of the selection rule we obtained from chirality conservation in the Thirring model. On the other hand, if the βs do sum to zero, (5.20) is independent of μ. Indeed it is identical to Eq. (5.7), aside from terms that can be absorbed in the arbitrary constant A.

This completes the argument.

Although I will not show it explicitly here, very similar arguments can be used to express the Thirring model current in sine–Gordon language.[38] The result is

$$j^\mu = \frac{\beta}{2\pi} \varepsilon^{\mu\nu}\partial_\nu\phi, \tag{5.22}$$

where $\varepsilon^{01} = -\varepsilon^{10} = 1$.

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

5.2 *The secret of the soliton*

One of the most intriguing features of the equivalence we have found is the relation between the coupling constants, Eq. (5.14). Large g is small β. Thus we have a rare animal indeed, a non-trivial relativistic field theory for which it is possible to make sensible strong-coupling approximations. The theory is the massive Thirring model, and the approximations are the small-β approximations for the sine–Gordon equation discussed in Sect. 4.

The central topic of Sect. 4 was the appearance of a coherent bound state, the quantum soliton. At first glance, one might think that our equivalence was proved only for states made from the vacuum by the application of strings of local field operators, the *S*s. Any such state has the property that the expectation value of ϕ has the same value at plus infinity as at minus infinity. Thus we can not investigate a one-soliton state, where the expectation value of ϕ changes by $2\pi/\beta$.

However, this objection does not apply to a state consisting of a widely separated soliton–antisoliton pair. For such a state, for small β,

$$\langle\phi(x)\rangle = f(x-b) - f(x+b), \tag{5.23}$$

where f is the classical soliton solution, and the separation is $2b$. If we apply Eq. (5.22) to the widely separated pair, we find

$$\begin{aligned}\langle j_0\rangle &= \frac{\beta}{2\pi}\langle\partial_1\phi\rangle \\ &= \frac{\beta}{2\pi}\frac{\mathrm{d}}{\mathrm{d}x}[f(x-b) - f(x+b)].\end{aligned} \tag{5.24}$$

The right-hand side of this equation is peaked about the points $-b$ and b, the locations of the antisoliton and soliton. The integrated charge over the soliton peak is $+1$; that over the antisoliton peak is -1. Now, the charge in question is fermion number, normalized such that the ψ field carries charge -1, as we see from Eq. (5.3). Thus the soliton is a particle of fermion number 1, and the antisoliton is a particle of fermion number -1. Also, unless we have missed something in Sect. 4, the soliton and the antisoliton are the lightest particles in the theory with non-zero fermion number. Thus the soliton of the sine–Gordon equation is nothing other than the fundamental fermion of the massive Thirring model.

It is a bit surprising to see a fermion appearing as a coherent state of a Bose field. Certainly this could not happen in three dimensions, where it would be forbidden by the spin-statistics theorem. However, there is no spin-statistics theorem in one dimension, for the excellent reason that there is no spin.

In fact, the appearance of fermions occurs even in a much simpler one-dimensional theory, the theory of a free massless scalar boson. The two-particle subspace of this theory contains states where each particle is in a normalizable state with spatial momentum support restricted to the positive axis. Even though this is a normalizable state, it is still an eigenstate of $P_\mu P^\mu$, with eigenvalue zero; all the two-momenta in the individual wave functions are aligned null vectors. Thus, if we adopt the usual definition of a particle, a normalizable eigenstate of P^2, we must say that the two-particle sector of Fock space contains single massless particles, as does the three-particle sector, the four-particle sector, and the coherent-state sector. As Streater and Wilde[40] discovered some time ago, some of these coherent states are fermions; they can be created from the vacuum by fields which anticommute with each other for space-like separations. (These fields are non-local functions of ϕ, but that is neither here nor there.)

Thus, the theory of a massless scalar field is chock-full of massless particles. (This is one way of understanding its terrible infrared divergences.) Typically, when we add an interaction, most of these particles dissolve into the continuum; which particles survive depends on the detailed form of the interaction. For example, if we add an interaction proportional to ϕ^2, only the particle in the one-particle sector of Fock space survives, and we have the theory of a free massive boson. On the other hand, as we have seen, if we add an interaction proportional to $\cos(2\pi^{\frac{1}{2}}\phi)$, only a fermion survives, and we have a theory of a free massive fermion.

So much for the soliton. What is the fundamental meson of sine–Gordon theory in Thirring-model language? In the Thirring model, positive g corresponds to an attractive force between fermions and antifermions. In the non-relativistic limit, this is a delta-function potential, and produces a bound state for arbitrarily small g. As we increase g, we would expect the bound state mass to decrease, perhaps becoming as small as possible (zero) as g goes to infinity. That is to say, we expect the fermion/bound-state mass ratio to go to infinity for vanishing β. But this is just where the soliton/meson mass ratio goes to infinity. Thus the lowest fermion–antifermion bound state of the massive Thirring model is an obvious candidate for the fundamental meson of sine–Gordon theory.

A check on this picture: The bound state has a symmetric wavefunction. Since the particles being bound are fermion and antifermion, this means the bound state is odd under both parity and charge conjugation. Equation (5.22) tells us how to transport the definitions of parity and charge conjugation from the Thirring model to the sine–Gordon equation. We see that the ϕ field is pseudoscalar and charge-conjugation odd.

This is consistent with the computations of Sect. 4.4, where the fundamental meson emerged as the lowest bound state of a soliton–antisoliton doublet. At this point I can state the postponed third reason for believing the DHN formula is exact: equation (4.82) predicts that all the doublet bound states disappear when β^2 exceeds 4π. This is precisely the point where the Thirring model interaction switches from attractive to repulsive.

The picture we have developed for this special system has some of the features of nuclear democracy in the sense of Chew. There is no way of deciding whether the fermion is fundamental and the boson a bound state or the boson is fundamental and the fermion a quantum lump. One is the natural way of putting things if one is describing the massive Thirring model and the other is the natural way of putting things if one is describing the sine–Gordon equation, but these two theories define identical physics. Which you choose to use is purely a matter of taste.

5.3 *Qualitative and quantitative knowledge*

We can learn both qualitative and quantitative information from the sine–Gordon equation.

Qualitatively, the sine–Gordon equation demonstrates explicitly that 'a lump' is a classical concept, like 'a wave' or 'a particle'. To ask whether the sine–Gordon soliton is a lump or a fundamental particle is as silly as asking whether light is a particle or a wave. Light acts very much like a wave in a certain approximation and very much like a particle in a different approximation. Whether it is better to think of it as a classical wave with quantum corrections or a classical particle with quantum corrections depends upon which approximation is appropriate. Likewise, for small β, it is certainly useful to think of the quantum soliton as a classical lump with quantum corrections. However, when β^2 increases to 4π, the quantum soliton becomes the most fundamental particle imaginable, the quantum of a free field theory.

Quantitatively, we can use the sine–Gordon equation as a theoretical laboratory in which we can test the approximation methods of Sect. 4. In Table 3 I have computed the predictions of these methods for the ratio of the soliton mass to the meson mass for three values of $\beta^2 : 4\pi$ (where the qualitative picture of the soliton as a lump totally breaks down), 2π, and π. At 4π we know the exact answer, because this is the point where the meson (in its guise as a fermion–antifermion bound state) is just on the verge of binding. (I happen to know the exact answer for 2π, so I have included this in the table.)[41]

Some features of the table could have been anticipated: The variational method gives the same answer as zeroth-order weak-coupling, the DHN

Table 3. *Soliton/meson mass ratio in sine–Gordon theory*

Method	$\beta^2=\pi$	$\beta^2=2\pi$	$\beta^2=4\pi$
Zeroth-order weak-coupling expansion[a]	2.55	1.27	0.64
Coherent-state variation[b]	2.55	1.27	0.64
First-order weak-coupling expansion[c]	2.23	0.95	0.32
Bohr–Sommerfeld quantization[d]	2.56	1.31	0.71
DHN formula[e]	2.25	1.00	0.50
Exact	?	1.00	0.50

[a]Equation (2.13b).
[b]This always gives the same result as the previous method.
[c]The explicit formula is not given in the text; it is $(8/\beta^2)-(1/\pi)$.
[d]Equation (4.64).
[e]Equation (4.82).

formula is exact, wherever it can be checked, and the stronger the coupling, the more the various approximation methods differ from each other and from the exact result.

However, the smallness of the errors, even for large coupling, is surprising. $\beta^2=4\pi$ is very strong coupling; the fundamental meson decays into a soliton–antisoliton pair, and the whole classical picture of the system is nonsense. Nevertheless, all of our approximations are still good to within 40% or so, not at all bad for a strong-coupling problem. $\beta^2=2\pi$ is still fairly strong coupling: the soliton and meson masses are identical; this is certainly very far from the small β limit, in which the soliton is very much heavier than the meson, and we would expect the weak-coupling expansion to be very bad. Nevertheless, the first-order weak-coupling approximation is good to within 5%.

Before I computed this table, I did not anticipate that the picture of the soliton as a lump would give such accuracy for such strong coupling. Unfortunately, since I know of no system other than the sine–Gordon equation soluble in the strong-coupling regime, I have no idea of whether this accuracy is a general feature of these approximations or one peculiar to this very special system.

5.4 *Some opinions*

This has been a long series of physics lectures with no reference whatsoever to experiment. This is embarrassing.

Certainly objects much like the lumps we have discussed play an important role in many fields of physics. There are lumps in superconductivity theory, in plasma physics, in solid-state physics. However, this is a school on subnuclear physics, and all of my lectures have been oriented

to relativistic quantum field theory. Is there any chance that the lump will be more than a theoretical toy in our field?

I can think of two possibilities.

One is that there will appear a theory of strong-interaction dynamics in which hadrons are thought of as lumps, or, more probably, as systems of quarks bound to lumps. (The SLAC bag[42] involves something like this idea, as, in a different way, does the Nielsen–Olesen string.[43]) At the moment, I am pessimistic about the success of such a theory. Classical ideas seem to me to be linked to weak couplings, as we saw in Sect. 4.1, and therefore unlikely to be useful in the study of the strong interactions. However, this intuition can be wrong, as we saw in Sect. 5.3, and I stand ready to be converted in a moment by a convincing computation.

The other possibility is that a lump will appear in a realistic theory with weak coupling constants, a theory of weak and electromagnetic interactions. If such a lump were to be kept from dissipating by the topological conservation laws of Sect. 3, the theory would have to imbed the $U(1)\otimes SU(2)$ group of the Weinberg–Salam model in a larger group without U(1) factors, but there is no shortage of such models. In this case, the lump, once discovered, would be unmistakable. It would be much heavier than the other particles in the theory (including the intermediate vector bosons) and it would be a magnetic monopole. Such monopoles are theoretically much more attractive than Dirac monopoles; for example, the renormalization difficulties that plague a field theory of Dirac monopoles are totally absent for lumps. Thus, if a monopole is found, it will be a very attractive possibility that it is a lump, and the properties of the monopole will tell us a great deal about the structure of the weak and electromagnetic interactions. Unfortunately, all monopole searches conducted to date have been failures.

Appendix 1: A three-dimensional scalar theory with non-dissipative solutions

Consider, in three spatial dimensions, the theory of a real scalar field ϕ and a complex scalar field ψ, with

$$U=(a\phi-m)^2\psi^*\psi+b\phi^2+c\phi^4, \tag{A.1}$$

where a, b, c, and m are positive numbers. T. D. Lee showed that this model has non-dissipative solutions.[6]

The unique ground state is $\psi=\phi=0$. For any dissipative solution, the fields become arbitrarily small for sufficiently large times, and thus we can compute the energy with arbitrary accuracy by only retaining the quadratic terms in U. In equations,

$$E=\lim_{t\to\infty}\left[\int d^3\mathbf{x}(\partial_0\psi^*\partial_0\psi+\nabla\psi^*\cdot\nabla\psi+\tfrac{1}{2}\partial_0\phi\partial_0\phi+\tfrac{1}{2}\nabla\phi\cdot\phi+m^2\psi^*\psi+b\phi^2)\right], \tag{A.2}$$

for a dissipative solution.

The theory is invariant under ψ phase transformations and thus possesses a conserved charge,

$$Q=\mathrm{Im}\int d^3\mathbf{x}\psi^*\partial_0\psi. \tag{A.3}$$

For any dissipative solution,

$$\begin{aligned} Q^2 &\leqslant \left|\int d^3\mathbf{x}\psi^*\partial_0\psi\right|^2 \\ &\leqslant \left[\int d^3\mathbf{x}\psi^*\psi\right]\left[\int d^3\mathbf{x}\partial_0\psi^*\partial_0\psi\right] \\ &\leqslant E^2/m^2. \end{aligned} \tag{A.4}$$

Here I have used the Schwarz inequality on the second line, and Eq. (A.2) on the third. I will now show that there exist non-dissipative solutions by displaying initial-value data that violate this inequality.

The initial-value data are:

$$\begin{aligned} \phi &= m/a, && r\leqslant L, \\ &= me^{-(r-L)}/a, && r\geqslant L, \end{aligned} \tag{A.5}$$

$$\partial_0\phi=0, \tag{A.6}$$

$$\begin{aligned} \psi &= \frac{A}{r}\sin(\pi r/L), && r\leqslant L, \\ &= 0, && r\geqslant L, \end{aligned} \tag{A.7}$$

and

$$\partial_0\psi=i\pi\psi/L, \tag{A.8}$$

where A and L are positive numbers.

ϕ and ψ have been chosen to make the first term in U vanish everywhere. Also, ψ has been chosen such that

$$\int d^3\mathbf{x}\nabla\psi^*\cdot\nabla\psi=\pi^2L^{-2}\int d^3\mathbf{x}\psi^*\psi. \tag{A.9}$$

Thus it is trivial to compute

$$\lim_{A\to\infty}(Q^2/E^2)=L^2/4\pi^2. \tag{A.10}$$

This violates the inequality for $L>2\pi/m$.

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

Appendix 2: A theorem on gauge fields

Theorem. In the standard theory of gauge fields only,

$$\mathscr{L} = -\tfrac{1}{4}F^a_{\mu\nu}F^{a\mu\nu}, \tag{A.11}$$

in D spatial dimensions, the only non-singular time-independent finite-energy solutions are gauge transforms of $A^a_\mu = 0$, for $D \neq 4$.

Note that this theorem does not depend on any special choice of gauge. If it did, it would be uninteresting, for a solution that is time-independent in one gauge can be horribly time-dependent in another gauge. The theorem asserts that there are no time-independent solutions in any gauge.

Proof. Evaluating the Lagrangian for a time-independent solution, we find

$$L = L_1 - L_2, \tag{A.12}$$

where

$$L_1 = \tfrac{1}{2}\int \mathrm{d}^D\mathbf{x}(F^a_{0i})^2 = \tfrac{1}{2}\int \mathrm{d}^D\mathbf{x}(\partial_i A^a_0 + ec^{abc}A^b_0 A^c_i)^2, \tag{A.13}$$

and

$$L_2 = \tfrac{1}{4}\int \mathrm{d}^D\mathbf{x}(F^a_{ij})^2. \tag{A.14}$$

Note that this is not the same as the total energy. The energy is the sum of these two terms, not the difference. However, finiteness of the energy does imply that each of the two terms must be finite.

Hamilton's principle for a time-independent solution implies that L must be stationary under time-independent variations. Consider the two-parameter family of field configurations defined by

$$A^a_0(\mathbf{x};\sigma,\lambda) = \sigma\lambda A^a_0(\lambda\mathbf{x}), \tag{A.15}$$

$$A^a_i(\mathbf{x};\sigma,\lambda) = \lambda A^a_i(\lambda\mathbf{x}). \tag{A.16}$$

An easy computation shows that

$$L(\sigma,\lambda) = \sigma^2\lambda^{(4-D)}L_1 - \lambda^{(4-D)}L_2. \tag{A.17}$$

This must be stationary at $\sigma = \lambda = 1$. Thus,

$$L_1 = L_2 = 0, \tag{A.18}$$

for $D \neq 4$. This, in turn, implies

$$F^a_{\mu\nu} = 0, \tag{A.19}$$

from which the stated result follows by standard arguments. (If you are not familiar with the standard arguments, there is a proof at the end of Appendix 5.)

Appendix 3: A trivial extension

In this appendix I sketch the extension of the analysis of Sect. 3 to the case where our theory of gauge fields and scalar fields possesses ordinary internal symmetries as well as gauge symmetries.

Let the full symmetry group of the theory be some compact Lie group G', not necessarily connected. The gauge group, G, is a closed connected invariant subgroup of G'. Let the full group of unbroken symmetries (for some choice of ϕ_0) be H'. H, the group of unbroken gauge symmetries, is the intersection of G and H'.

So far, this is just definitions. I will now make the assumption announced in Sect. 3.2. I will assume there is no accidental degeneracy; thus the set of zeros of U can be identified with the coset space G'/H'. This is much weaker than the simplifying assumption of Sect. 3.2; G'/H' can be much bigger than G/H.

G'/H' is the union of the orbits of G, the minimal sets invariant under the action of G. One of these orbits is G/H. All the orbits are topologically identical to G/H, since any orbit can be obtained from any other by the action of some element of G'.

The scalar fields at infinity must lie on a single orbit, for otherwise we could not cancel their space derivatives with gauge fields. (See the discussion after Eq. (3.34).) Thus, at least at first glance, it seems as if we have a richer classification of solutions than in Sect. 3. Our solutions are labeled not just by an element of $\pi_{D-1}(G/H)$, but also by an orbit of G in G'/H'.

However, this richness is illusory. Solutions associated with different orbits can not exist in the same world. For, if we try to patch together two distant solutions associated with different orbits, as in Fig. 3, we can not make a gauge transformation such that the solutions are identical in the unshaded region. (If you find this argument unconvincing, consider a theory with spontaneous symmetry breakdown but no gauge fields, and instead of a lump, consider an ordinary meson, like the sigma meson in the sigma model. Does the fact that there are many vacuum states mean that an experimenter exploring the universe of the sigma model would find many kinds of sigma mesons?)

Thus, the existence of the larger group G' is totally irrelevant to the phenomena discussed in Sect. 3. The only relevant objects are the gauge group G and its unbroken subgroup H.

Appendix 4: Looking for solutions

When studying a linear partial differential equation, it is frequently useful to write the field variables as sums over a set of functions

that transform according to the irreducible representations of the symmetry group of the theory. In this way, for example, the Schrödinger equation for a particle in a central potential is converted into a system of uncoupled ordinary differential equations, one for each angular-momentum eigenstate. Of course, this stratagem does not work if we deal with nonlinear equations, as we must if we look for a time-independent lump. However, there is a stratagem based on the same ideas which yields less far-reaching, but still useful, results.

Let G^* be some subgroup of the full symmetry group of the theory (now to be thought of as including space-time symmetries). Every field configuration can be written as the sum of two terms, a symmetric part, invariant under G^*, and an asymmetric part, a sum of terms each of which transforms according to some non-trivial irreducible representation of G^*. If we plug this decomposition into the expression for the energy, we find that the energy contains no terms linear in the asymmetric part, by Schur's Lemma. Thus the energy, evaluated at a symmetric field configuration, is automatically stationary under asymmetric variations. If we wish to search for symmetric time-independent solutions, we need only insert the most general symmetric field configuration into the energy, and stationarize this restricted problem. There is no need to check for stationarity under asymmetric variations. Similar remarks apply to time-dependent solutions, if 'energy' is replaced by 'action'.

I will give two examples.

(1) The field theory of Example 4, the 't Hooft–Polyakov model. The boundary conditions for the simplest kind of non-dissipative solution, Eq. (3.46), are invariant under simultaneous space–isospin rotations. They are also invariant under parity, if we define the ϕ fields to be pseudoscalar. We choose G^* to be this group of combined rotations plus parity. The most general field configuration invariant under G^* is of the form

$$\begin{aligned} \phi^b &= r^b f(r), \\ A^{ib} &= \varepsilon^{ibc} r^c g(r), \end{aligned} \tag{A.20}$$

where f and g are arbitrary functions of r. Thus the problem has been reduced from one that involves six functions of three variables to one that involves only two functions of one variable. The reduced problem is not beyond the reach of numerical analysis, even if one is armed only with a desk calculator.

(2) The model of Appendix 1. Inspired by the non-dissipative initial-value data, we look for solutions that are invariant under (*a*) rotations, (*b*) simultaneous time translations by T and phase transformations by ωT, where ω is some real number, and (*c*) the transformation

$$\Psi(\mathbf{x}, t) \to \psi^*(\mathbf{x}, -t)$$
$$\phi(\mathbf{x}, t) \to \phi(\mathbf{x}, -t). \tag{A.21}$$

The most general set of fields symmetric under this choice of G^* is

$$\psi = e^{i\omega t} f(r),$$
$$\phi = g(r) \tag{A.22}$$

where f and g are arbitrary real functions of r. Once again the problem has been reduced to one that involves only two unknown functions of r.

Some remarks:

(1) This method does not help in finding general solutions, only in finding symmetric solutions. Still, knowing only a few symmetric solutions is better than knowing nothing at all.

(2) Schur's Lemma guarantees stationarity, but not stability, for stability is a question of second variations. After a symmetric solution is found, stability under asymmetric variations has to be checked. To the best of my knowledge, this has not yet been done even for the solution of Example 4 obeying Eq. (A.20) found numerically by 't Hooft.

Appendix 5: Singular and non-singular gauge fields

This is an appendix to the discussion of Dirac monopoles in Sect. 3.7. A Dirac monopole has a genuine gauge-invariant singularity at the origin of coordinates, where the magnetic field blows up; it also has a gauge-variant singularity, the string. The string can be moved around by gauge transformations, but not eliminated altogether. We want to understand why the singularity at the origin necessitates the string. As a by-product, we shall gain a deeper understanding of non-singular gauge fields, and prove the theorem mentioned at the end of Appendix 2. The arguments have nothing to do with the metric structure of Minkowski space, and I will therefore work in a general real n-space, with points $x = (x^1 \ldots x^n)$.

Definition. Let R be a region in n-space. We say that there is a locally non-singular field defined in R if R can be covered with a family of open sets such that (1) in each of these open sets there is defined a non-singular gauge field, $A_i^a(x)$, (2) whenever two open sets overlap, there is a non-singular gauge transformation defined in the overlap region that turns the gauge field in one set into the gauge field in the other.

The first condition is the statement that for every point in R there is some neighborhood of the point that contains no gauge-invariant singularities; the second condition is the statement that when two open sets overlap, all gauge-invariant quantities are uniquely defined in the region of overlap.

An ordinary non-singular gauge field is *a fortiori* locally non-singular. The Dirac monopole is locally non-singular for all of three-space with the origin excised. Our fundamental problem is, given a locally non-singular gauge field, when can we make a gauge transformation in each of the open sets such that the end result is a non-singular gauge field defined in all of R? Phrased more briefly, when are locally non-singular gauge fields gauge-equivalent to non-singular gauge fields?

Theorem. If R is a closed n-cube of side L, then any locally non-singular gauge field is gauge-equivalent to a non-singular gauge field.

Proof. Choose the origin of coordinates to lie within R and construct a regular cubic lattice of points based at the origin with lattice spacing ε. Let each point of this lattice be the center of an open cube with side 2ε. We thus obtain a covering of R by a regular lattice of open cubes. Consider the family of such coverings with $\varepsilon=(\frac{1}{2})^m$, m an integer. Every point of R is contained in some open cube from one of these coverings that is, in turn, contained in one of the open sets in the definition of local non-singularity. Thus we can replace the covering of the definition by a covering by these open cubes. R is compact, and thus we can replace this covering, in turn, by a finite subcovering. A finite covering must contain a smallest cube. Thus we can replace the covering in the definition by a covering by a regular lattice of open cubes, that which contains this smallest cube. We are now ready to construct the required gauge transformation.

The construction is exactly the same as that which we used to construct the gauge $A_0^a=0$ in Sect. 3.2. Starting from the hyperplane $x^1=0$, we construct the group elements $g(P)$ associated with paths parallel to the x^1 axis. There is no problem in changing gauge as we pass from cube to cube, because the gauge-transformation properties of $g(P)$ depend only on the value of the gauge transformation at the end point of the path. (See Eq. (3.22).) We then use $g(P)$ to gauge transform the fields in each cube such that $A_1^a=0$ everywhere. This implies that the gauge transformation connecting two overlapping cubes is independent of x^1. (See Eq. (3.10).) Thus, we can move along a column of cubes parallel to the x^1 axis, and make this gauge transformation in each successive cube. We have thus transformed the open cubes into open columns parallel to the x^1 axis, such that in each column a single-valued non-singular gauge field is defined. We may still have to gauge transform between adjacent columns, but the gauge transformation is independent of x^1.

We now iterate this procedure. On the hyperplane $x^1=0$, we gauge transforms such that $A_2^a=0$, and apply the same (x^1-independent) gauge

https://doi.org/10.1017/CBO9780511565045.007 Published online by Cambridge University Press

transformation throughout the covering. We thus turn the columns into slabs parallel to the x^1–x^2 plane. After $n-1$ such iterations, we have gauge transformed such that

$$A_r^a = 0,\ x^1 = x^2 = \cdots = x^{r-1} = 0, \tag{A.23}$$

and have arrived at a single-valued non-singular gauge field defined in the entire cube. Q.E.D.

Theorem. If R is all n-space, any locally non-singular gauge field is gauge-equivalent to a non-singular gauge field.

Proof. Consider two n-cubes, both centered at the origin, with one larger than the other. If we do the construction of the previous theorem in the larger cube, and then restrict the answer to the smaller cube, it is easy to see that we get the same result we would have obtained if we had originally done the construction only in the smaller cube. Thus we can fill n-space with an ever-increasing set of cubes, such that when we increase the size of the cube, we do not have to make further gauge transformations in already conquered territory. Thus, this procedure trivially has a limit, which defines the required gauge transformation in all of n-space.

Theorem. If R is all of n-space, excluding a single point, then any locally non-singular gauge field is gauge-equivalent to a non-singular gauge field defined in all of R except for a line going from the excluded point to infinity, parallel to the x^1 axis. (This is the Dirac string.)

Proof. Choose coordinates such that the excluded point is (1, 0, 0...). Exclude from R a small open sphere about this point. The intersection of R with a small n-cube is now a compact set, and we can do the construction of the first theorem. However, as we merrily gauge-transform along, we can not reach that portion of the columns of the covering that lies above the excluded sphere; this remains unknown territory. As the excluded sphere becomes arbitrarily small, this unknown territory shrinks down to the string.

Comments. (1) Since we have never used the metric structure of n-space, we can do all of our constructions in any non-singular coordinate system. This means that the string can be any non-self-intersecting non-singular curve going from the excluded point to infinity. (2) The generalization to any finite set of excluded points is trivial. There is one string for each point.

Theorem. Any non-singular gauge field in all n-space such that $F_{ij}^a = 0$ everywhere is gauge equivalent to a vanishing gauge field. (This is the theorem referred to at the end of Appendix 2.)

Proof. We go to the gauge of Eq. (A.23). Since A_1^a vanishes everywhere,

$$F_{1i}^a = \partial_1 A_i^a = 0. \tag{A.24}$$

Thus A_i^a is independent of x^1. On the surface $x^1 = 0$, A_2^a vanishes. But it is independent of x^1; therefore it vanishes everywhere. Thus,

$$F_{2i}^a = \partial_2 A_i^a = 0. \tag{A.25}$$

Etc.

Notes and references

1. For the narrow definition, and also for much information about a system we shall encounter shortly, the sine–Gordon equation (and about many similar systems), see A. Scott, F. Chu, and D. McLaughlin, *Proc. IEEE* **61**, 1443 (1973).
2. Much of the material in this section is plagiarized from two long papers on the quantum problem: J. Goldstone and R. Jackiw, *Phys. Rev.* **D11**, 1486 (1975); N. Christ and T. D. Lee, *Phys. Rev.* **D12**, 1606 (1975).
3. This equation has an interesting history. It was studied extensively by differential geometers in the last quarter of the nineteenth century, because of its role in the theory of surfaces of constant negative curvature. Bianchi called it, 'l'equazione fondamentale di tutta la teoria delle superficie pseudosferiche'. (*Lezioni di geometria differenziale*, 3rd ed., I, 658.) The equation enters geometry in the following way. On any two-dimensional Riemannian manifold, it is possible to choose coordinates in some neighborhood of any point such that

   $$ds^2 = du^2 + dv^2 + 2du\, dv \cos\theta(u, v).$$

   In terms of such coordinates, a simple computation shows that the statement that the manifold has constant negative curvature is equivalent to

   $$\partial^2\theta/\partial u\,\partial v = \alpha \sin\theta,$$

   where α is a constant related to the magnitude of the curvature. This is the sine–Gordon equation, in light-cone coordinates.
   The equation entered particle physics through the work of Skyrme (*Proc. Roy. Soc.* **A247**, 260 (1958); **A262**, 237 (1961)) who studied it as a simple model of a nonlinear field theory. Skyrme's work prefigured many of the results we shall obtain in Sect. 5.
   The silly name is of recent origin. From a letter from David Finkelstein: 'I am sorry that I ever called it the sine–Gordon equation. It was a private joke between me and Julio Rubinstein, and I never used it in print. By the time he used it as the title of a paper he had earned his Ph.D. and was beyond the reach of justice.'
4. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, (McGraw-Hill, New York, 1953), I, 721.
5. G. H. Derrick, *J. Math. Phys.* **5**, 1252 (1964).
6. R. Friedberg, T. D. Lee and A. Sirlin, *Phys. Rev.* **D13**, 2739 (1976).
7. An extension of Derrick's theorem can be proved for theories of gauge fields alone. See Appendix 2 (but read Sec. 3.2 first).
8. I have reorganized the material in this section so much that it is difficult to give specific credits at appropriate places in the text, but I would be astonished if any of its contents comes as a surprise to the authors of the following papers: D. Finkelstein, *J. Math. Phys.* **7**, 1218 (1966); J. Arafune, P. Freund, and C.

Goebel, *J. Math. Phys.* **16**, 433 (1975); Yu. S. Tyupkin, V. A. Fateev, and A. S. Shvarts, *JETP Lett.* **21**, 42 (1975); M. I. Monastyrskiĭ and A. M. Perelemov, *ibid.*, 43 (1975); A. Patrascioiu, *Phys. Rev.* **D11**, 523 (1975). The contents of several long conversations with J. Goldstone and L. Faddeev are also distributed uniformly throughout this section.

9. The initial-value theorem established by C. Parenti, F. Strocchi, and G. Velo, (*Phys. Lett.* **59B**, 157 (1975), *Comm. Math. Phys.* **53**, 65 (1977)) should be sufficient to demonstrate this rigorously for all of the models we shall consider.
10. For a lengthier discussion, see Chapter 5 in this volume.
11. If the scalar fields are real, a factor of $\frac{1}{2}$ is conventionally inserted in front of the second term in Eq. (3.17).
12. Two standard mathematical expositions of homotopy theory are: P. J. Hilton, *An Introduction to Homotopy Theory*, No. 43, Cambridge Tracts in Mathematics and Mathematical Physics; N. E. Steenrod, *The Topology of Fibre Bundles*, No. 14, Princeton Mathematics Series.
13. See, for example, A. Fetter and P. Hohenberg, Chapter 14 of *Superconductivity*, ed. by R. Parks.
14. We shall find them all, by and by.
15. A. M. Polyakov, *JETP Lett.* **20**, 194 (1974).
16. G. 't Hooft, *Nucl. Phys.* **B79**, 276 (1974).
17. E. Cartan, *Œuvres complètes*, I, 2, 1307.
18. This is a result of Tyupkin *et al.*[8] It was also found independently, and by a very different method, by Patrascioiu.[8]
19. We shall construct one at the end of Sec. 3.7.
20. To my knowledge, the first use of homotopy groups to derive topological conservation laws was by D. Finkelstein and C. Misner, *Ann. Phys.* **6**, 230 (1959).
21. For slightly more details and much more rigor, see the texts of Note 12.
22. If it is not obvious to you, here is the proof. We want to show that given $\phi(\theta)$ such that $\phi(0)=\phi_0$, and given a homotopy $\phi(\theta, t)$ such that $\phi(\theta, 0)=\phi(\theta)$ and $\phi(\theta, 1)=$ a constant, then there exists a second homotopy, $\phi'(\theta, t)$, obeying all the previous conditions and in addition obeying the condition $\phi'(0, t)=\phi_0$ for all t. Proof. Since $\phi(0, t)$ is continuous, there exists continuous $g(t)$ such that $g(t)\phi(0, t)=\phi_0$. Define $\phi'(\theta, t)=g(t)\phi(\theta, t)$.
23. I thank Professor Haag for clarifying this situation for me, and also for paying for the coffee.
24. Homotopy experts will recognize the preceding eight paragraphs as the world's most ham-handed exposition of the exact homotopy sequence,

    $$\pi_2(G)\to\pi_2(G/H)\to\pi_1(H)\to\pi_1(G).$$

25. For more on strings, and singular gauge fields in general, see Appendix 5.
26. This is a discovery of E. Lubkin, *Ann. Phys.* **23**, 233 (1963). Lubkin may have been the first to realize the utility of homotopy theory in describing the global properties of gauge fields.
27. This casts doubt on the utility of the analysis of Sec. 2 when H is non-Abelian. The effective coupling constant (in the sense of the renormalization group) grows with distance for non-Abelian gauge fields, so there is not much point in using classical reasoning to study the fields at large distances. The same thing phrased another way: if quarks are indeed confined because they carry non-Abelian charges, then the same mechanism should confine non-Abelian monopoles.
28. This expansion has been derived by many authors using many different methods. In addition to the papers of Note 2, there are: J.-L. Gervais and B. Sakita, *Phys. Rev.* **D11**, 2943 (1975); **D12**, 1038 (1975); V. E. Korepin,

P. P. Kulish, and L. Faddeev, *JETP Lett.* **21**, 139 (1975); C. Callan and D. Gross, *Nucl. Phys.* **B93**, 29 (1975); E. Tomboulis, *Phys. Rev.* **D12**, 1678 (1975); A. Klein and F. Krejs, *Phys. Rev.* **D13**, 3295 (1981).

29. For more details, see any graduate-level text on quantum mechanics.
30. R. Rajaraman and E. J. Weinberg, *Phys. Rev.* **D11**, 2950 (1975).
31. Two early papers on lumps as coherent states are: P. Vinciarelli, *Lett. Nuovo Cimento* 2, **4**, 905 (1972); K. Cahill, *Phys. Letters* **53B**, 174 (1974).
32. R. C. Dashen, B. Hasslacher, and A. Neveu, *Phys. Rev.* **D10**, 4114; 4130; 4138 (1974). A pedagogical review of these methods has been written by R. Rajaraman *(Phys. Reports* **21**, 227 (1975)).
33. The form given is essentially that of Rajaraman's review.[32]
34. R. Dashen, B. Hasslacher, and A. Neveu, *Phys. Rev.* **D11**, 3424 (1975). A lot of the early material in this section is plagiarized from this paper.
35. Most of this section is based on S. Coleman, *Phys. Rev.* **D11**, 2088 (1975), although some of the arguments have been rearranged.
36. The Thirring model was proposed by W. Thirring, *Ann. Phys.* (N.Y.) **3**, 91 (1958). A formal operator solution was found by V. Glaser, *Nuovo Cimento* **9**, 990 (1958), and a rigorous solution by K. Johnson, *Nuovo Cimento* **20**, 773 (1961). More references are in Note 35.
37. J. Fröhlich has been able to show that the model exists and obeys the usual axioms for a range of coupling constants. His work is summarized in his lectures at the 1975 Erice School on mathematical physics. See J. Fröhlich, *Comm. Math. Phys.* **47**, 233 (1976), J. Fröhlich and E. Seiler, *Helv. Phys. Acta* **49**, 889 (1976).
38. See Note 35 for more details.
39. S. Mandelstam has gone further than this and actually constructed the Fermi fields as (non-local) functions of the Bose field (*Phys. Rev.* **D11**, 3026 (1975)).
40. R. Streater and I. Wilde, *Nucl. Phys.* **B24**, 561 (1970).
41. S. Coleman, *Ann. Phys.* **101**, 239 (1976).
42. See, for example, the lectures of S. Drell at this school.
43. H. Nielsen and P. Olesen, *Nucl. Phys.* **B61**, 45 (1973).