Duality in field theory and statistical systems*

Robert Savit

Physics Department, University of Michigan, Ann Arbor, Michigan 48109

This paper presents a pedagogical review of duality (in the sense of Kramers and Wannier) and its application to a wide range of field theories and statistical systems. Most of the article discusses systems in arbitrary dimensions with discrete or continuous Abelian symmetry. Globally and locally symmetric interactions are treated on an equal footing. For convenience, most of the theories are formulated on a d-dimensional (Euclidean) lattice, although duality transformations in the continuum are briefly described. Among the familiar theories considered are the Ising model, the x-y model, the vector Potts model, and the Wilson lattice gauge theory with a Z_N or $U(1)$ symmetry, all in various dimensions. These theories are all members of a more general heirarchy of theories with interactions which are distinguished by their geometrical character. For all these Abelian theories it is shown that the duality transformation maps the high-temperature (or, for a field theory, large coupling constant) region of the theory into the low-temperature (small coupling constant) region of the dual theory, and vice versa. The interpretation of the dual variables as disorder parameters is discussed. The formulation of the theories in terms of their topological excitations is presented, and the role of these excitations in determining the phase structure of the theories is explained. Among the other topics discussed are duality for the Abelian Higgs model and related models, duality transformations applied to random systems (such as theories of a spin glass), duality transformations in the "lattice Hamiltonian" formalism, and a description of attempts to construct duality transformations for theories with a non-Abelian symmetry, both on the lattice and in the continuum.

CONTENTS

I. INTRODUCTION

A. Preview

This is a pedagogical review of duality in field theory and statistical systems. By duality we mean the natural extensions of the work of Kramers and Wannier (1941) on the two-dimensional Ising model, and related topics. In their work, Kramers and Wannier showed, using a certain transformation, that the $d=2$ Ising model could be exactly rewritten as another $d=2$ Ising model, but whose temperature was a monotonically decreasing function of the temperature of the original Ising model, as shown schematically in Fig. 3. Thus high-temperature regions of the original Ising model were transformed into low-temperature regions of its dual, and vice versa. This intriguing transformation led to many interesting insights into the structure of the model, as we shall discuss in Sec. II. It turns out that this transformation can be generalized and simply applied to almost any Abelian theory in any number of dimensions. The "any" of the last sentence includes theories with different kinds of local symmetry as well as globally symmetric theories, and also includes theories whose symmetry groups are Z_N, R, $U(1)$, or some combination of these, thereby covering systems with both discrete and continuous Abelian symmetries. The theory can be one which is formulated on a lattice or in the continuum, although we shall find it convenient to work almost exclusively on d-dimensional simple hypercubic lattices. The "almost" of three sentences back ap-

*Work supported by the National Science Foundation under grant No. PH78-08426.

pears because of the word "simply" in the same sentence. It is always possible to perform a duality transformation on an Abelian theory, but the result may be rather complicated and not manifestly useful. Unfortunately, it is much more difficult in general to construct a duality transformation for theories with a non-Abelian symmetry. We shall briefly discuss this problem in Sec. V.

In those Abelian cases when the duality transformation can be successfully applied we reap a number of benefits. First, the theory is expressed in terms of a new set of variables, the "disorder variables." These dual variables have small fluctuations at high temperatures when the original variables have large fluctuations. Thus the original theory is mapped into a dual theory whose temperature is low when the temperature of the original theory is high, and vice versa. Since the statistical systems we are discussing may also be considered as Euclidean field theories, the field theorist may replace the phrase high (low) temperature by large (small) coupling constant. In the language of field theory then, the duality transformation maps a theory with a large coupling constant into a theory with a small coupling constant, and vice versa. Clearly such a transformation can be of great help, especially in trying to understand the often difficult to handle strong coupling domains of field theories.

This feature of temperature or coupling constant inversion is similar to what Kramers and Wannier found for the $d=2$ Ising model. But unlike the $d=2$ Ising model which is self-dual, a dual theory in general does not have the same structure as the original theory to which it is dual. For example, it sometimes happens that the dual of globally symmetric theories are locally symmetric gauge theories, and vice versa.

Another important benefit derived from duality concerns the topological excitations of a theory. We will see that in some cases the dual form of a theory is an intermediate step to a third form in which the variables that appear in the partition function are the topological excitations of the original variables of the original form of the theory. This manifestation of the theory provides a good deal of insight into many qualitative features of the theory, especially its phase structure.

Our procedure in most of this review will be to examine various theories constructed on a d-dimensional hypercubic lattice. We shall not explicitly discuss duality for theories on other types of lattices, nor shall we discuss at any length duality for a class of theories in which one space-time direction is taken as continuous and the others are latticized (the so-called lattice Hamiltonian formulations.) Furthermore, we shall only briefly discuss theories defined in the continuum in Sec. V. Finally, we will not directly review the large body of literature concerned with monopoles in the usual sense, although this subject is very closely related to the subject of the present review and will be referred to from time to time. As a partial remedy for these omissions, Sec. I. B consists of a guide to the literature and includes references to review articles which discuss these and other briefly treated topics.

The rest of the review is organized as follows: In Sec. I. B we present a guide to the literature. Section II is devoted to a study of theories with a Z_2 symmetry. We begin in Sec. II. A with a fairly complete treatment of the dual properties of the simplest such theory, the two-dimensional Ising model. Included in this section is a brief allusion to spin glasses and the idea of frustration. Sec. II. B discusses the three-dimensional Ising model and its dual, the Z_2 lattice gauge theory, while Sec. II. C covers the four-dimensional Ising model. Appropriate correlation functions for the dual theories are introduced and discussed. Using the insights garnered from these theories we proceed in Sec. II. D to discuss the fundamental geometrical structure of the duality transformation. In so doing, we construct the general duality transformation for a large class of locally and globally Z_2 symmetric theories in arbitrary dimension. In Sec. III we turn to theories with a Z_N symmetry. After a short introduction (Sec. III. A), we discuss in Sec. III. B two-dimensional globally invariant Z_N theories. We touch on a variety of theories and treat in detail the vector Potts model and the "Z_N-Villain model." This last model has the interesting property of enforcing the periodicity associated with a Z_N symmetry by employing a set of auxiliary fields. The $U(1)$ invariant version of the model will turn out to be very useful for interpreting topological excitations. In Sec. III. C we generalize the results of Secs. II and III.B and construct the general duality transformation for a class of locally and globally Z_N invariant theories in arbitrary dimension. Section IV deals with theories with a $U(1)$ symmetry. The introduction, Sec. IV. A, is followed by a treatment in Sec. IV. B of the two-dimensional x-y model. We describe the duality transformation in some detail, and introduce the Coulomb gas representation of the model. We then use the "Villain approximation" to identify the charges of the Coulomb gas as the topological excitations of the x-y model spins. Section IV. B concludes with a qualitative discussion of the physics of the $d=2$ x-y model which relies heavily on its Coulomb gas representation. In Sec. IV. C the contents of Sec. IV. B are generalized to a large class of globally and locally $U(1)$ symmetric theories in d dimensions. The general duality transformation is constructed, and the representations of the theories in terms of their topological excitations are derived. At this point we see a very pretty pattern emerge which relates the dimension of the topological excitation to the nature of the $U(1)$ interaction (in particular, whether it manifests global invariance or local gauge invariance of the second, third, fourth, etc., kind) and the space-time dimension of the system. In Sec. IV. D we discuss three theories of physical interest, the $d=3$ x-y model and the $d=3$ and 4 $U(1)$ lattice gauge theories. We use their representations in terms of their topological excitations to give a qualitative description of their phase properties, and we argue that the onset of disorder in these theories (as in the d-2 x-y model) can be understood as being due to a condensation of topological excitations into something like a plasma phase. In the final section of the review, Sec. V, we briefly touch on several topics involving duality, but not covered in the previous chapters. These include the lattice Abelian Higgs model and related models, duality for Abelian random systems, topological excitations for Z_2 and Z_N symmetric theories, Abelian duality in the lattice Hamiltonian formalism, duality in the continuum, and a discussion of approaches to duality for non-Abelian theories.

B. Guide to the literature

For the convenience of the reader and to correct possible lapses in referencing in the body of the review, I present a list of references organized according to topic. The reader should be aware that since some papers overlap several categories, there is some arbitrariness in their classification. This also seems like a good time to extend my apologies to those authors to whose works, through my ignorance or oversight, I have failed to refer. Sorry folks.

Duality for Z_2 symmetric systems: Balian *et al.* (1975); Horn and Yankielowicz (1979); Kadanoff and Ceva (1971); Kramers and Wannier (1941); Wegner (1971)

Z_N symmetric theories: Cardy (1979); Casher (1978); Elitzur *et al.* (1979); Einhorn, Savit, and Rabinovici (1979); Horn *et al.* (1979); Korthes-Altes (1978); Savit (1980); Ukawa *et al.* (1979); Wegner (1973); Yoneya (1978)

Duality and the two-dimensional x-y model: Berezinskii (1970, 1972); Chui and Weeks (1976); Jose *et al.* (1977); Kosterlitz and Thouless (1973); Luther and Scalapino (1977); Savit (1978); Villain (1975)

Abelian Higgs model: Banks and Rabinovici (1979); Einhorn and Savit (1978, 1979); Fradkin and Shenker (1979); Israel and Nappi (1979); Jones, Kogut, and Sinclair (1979); Peskin (1978)

Other $U(1)$ invariant theories: Banks *et al.* (1977); Glimm and Jaffe (1977); Polyakov (1975, 1977); Savit (1977a, 1978); Stone and Thomas (1978); Sugamoto (1979)

Duality and Hamiltonian lattice formulations: Fradkin and Susskind (1978); Green (1978)

Duality and other Abelian theories: Drouffe (1978); Fradkin, Huberman, and Shenker (1978); Jose (1978); Kadanoff (1978); Mittag and Stephen (1971)

Duality for non-Abelian theories: Bellisard (1978); Drouffe *et al.* (1979); Englert and Windey (1978); Goddard *et al.* (1977); Halpern (1979); 't Hooft (1978, 1979); Kazama and Savit (1979); Mandlestam (1978); Montonen and Olive (1977); Seo *et al.* (1979)

In addition, four other reviews in the literature contain material related to our subject. First, the review of Syozi (1972) on exact Ising model transformations includes some discussion of duality transformations on lattices other than simple square or cubic. Next, the monograph by Gruber *et al.* (1977) contains an extensive discussion of duality, especially as applied to certain generalizations of the Ising model, such as the Potts models, the Askin-Teller model, and models on other than square lattices. Third, the paper by Goddard and Olive (1977) is an excellent pedagogical review of monopoles in the sense of Dirac, 't Hooft, Polyakov, and others. Finally, the recent review of Kogut (1979) on lattice spin and gauge theories includes some discussions of duality, particularly in the Hamiltonian formalism.

II. CASE OF THE Z_2 SYMMETRY

A. Two-dimensional Ising model

1. Duality transformation

The simplest nontrivial theory with which to illustrate duality transformations is the two-dimensional Ising model on a square lattice. The dual properties of this model were originally discussed by Kramers and Wannier (Kramers and Wannier, 1941). We will give a fairly detailed treatment of this model, since many of its features will find their counterparts in other theories we shall discuss.

Suppose we have a two-dimensional square lattice. On each site of the lattice, labeled by a pair of integers, $\vec{i}=(i_x, i_y)$, we place a variable (or spin) s_i which can take on the values ± 1. (For simplicity, we will usually drop the vector symbol on the site index $\vec{i}$.) The Hamiltonian of the system is

$$H=-J\sum_{\langle\,\rangle} s_i s_j\,, \tag{2.1}$$

where $\langle\,\rangle$ denotes a sum over all nearest neighbor pairs, and J is the coupling strength. J positive (negative) is a ferromagnetic (antiferromagnetic) coupling. This system possesses a global Z_2 symmetry: H is invariant under a change in the sign of all the s_i. The partition function of this system is

$$Z=\sum_{\{s\}} \exp\Big(\beta\sum_{\langle\,\rangle} s_i s_j\Big)\,, \tag{2.2}$$

where $\{s\}$ denotes a sum over all spin configurations, and $\beta=J/kT$, with k being Boltzmann's constant and T being the temperature.[1] Unless otherwise stated, we will consider the ferromagnetic case, $\beta\geq 0$. (For the moment we need not worry about the boundary conditions, but one may suppose that we have spherical boundary conditions.)

The first step in performing the duality transformation is to rewrite Eq. (2.2) in the form

$$Z=\sum_{\{s\}}\prod_{\langle\,\rangle} e^{\beta s_i s_j} \tag{2.3a}$$

$$=\sum_{\{s\}}\prod_{\langle\,\rangle}\sum_{k=0} C_k(\beta)(s_i s_j)^k\,, \tag{2.3b}$$

with

$$C_0(\beta)=\cosh\beta\,,\quad C_1(\beta)=\sinh\beta\,.$$

In Eq. (2.3a) we have just rewritten the exponential of the sum over pairs in Eq. (2.2) as a product over pairs of the exponential. In Eq. (2.3b) we introduce a new set of variables, $\{k\}$, one for each nearest neighbor pair to rewrite Eq. (2.3a) in the form indicated. Since there is one variable, k, associated with each link of the lattice, any given k can be labeled by a position and a direction $k_{\mu;i}$, where $\mu=1,2$ and i labels the position in the lattice. (By convention, a given link belongs to

[1]The expression (2.2) can also be regarded as a generating functional for a field theory in two Euclidean space-time dimensions. In that case, H/kT plays the role of the Lagrangian and the global Z_2 symmetry is a symmetry of the Lagrangian. We shall refer again to this correspondence below.

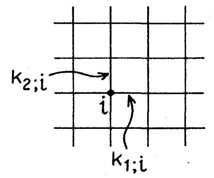

FIG. 1. Two-dimensional square lattice showing the location of the $k_{\mu;i}$ used in the duality transformation.

the site at its left or lower end.) Hence the k_μ's form a vector field over the lattice, as shown in Fig. 1.

We now rearrange the product in Eq. (2.3b) so that all factors of s_i associated with a given site are grouped together. We have

$$\begin{aligned} Z &= \sum_{\{s\}}\sum_{\{k\}}\prod_{l} C_{k_\mu}(\beta)\prod_i (s_i)^{\Sigma_i k_\mu} \\ &= \sum_{\{k\}}\prod_{l} C_{k_\mu}(\beta)\prod_i \sum_{s=\pm 1}(s_i)^{\Sigma_i k_\mu} \\ &= \sum_{\{k\}}\prod_{l} C_{k_\mu}(\beta)\prod_i 2\delta_2\left(\textstyle\sum_i k_\mu\right). \end{aligned} \tag{2.4}$$

The product over l (links) is a product of all the $C_{k_\mu}(\beta)$, there being one $C_{k_\mu}(\beta)$ associated with each link of the lattice. The product over i is a product over all sites of the lattice, and the notation $\sum_i k_\mu$ means the following: there are four links which impinge on each site of the lattice. $\sum_i k_\mu$ is the sum of the four k_μ's associated with those four links. Finally, $\delta_2(n)$ is a Kronecker delta function mod 2: it is zero if n is odd and one if n is even.

We now wish to find a representation for the $k_{\mu;i}$ which automatically satisfies the δ functions in Eq. (2.4). To do this, imagine constructing a new lattice from the original one by placing a vertex of the new lattice in the center of each square of the original lattice. Connecting nearest-neighbor pairs of the new lattice by links we see that we have constructed a new square lattice displaced from the originial lattice by half a lattice spacing in each direction. This is the dual lattice. (Note that if we repeat the process we get back the original lattice.) Now, we associate a variable σ_i taking on values ± 1 with each site of the dual lattice. Looking at Fig. 2, we see that to each link of the original lattice, we can uniquely associate a pair of σ_i's, namely, those that lie at the ends of the dual lattice link which crosses the given link of the original lattice. So, we can write each $k_{\mu;i}$ in the form

$$k_{\mu;i} = \tfrac{1}{2}(1-\sigma_i\sigma_{i-\hat{\nu}}), \quad \mu \neq \nu. \tag{2.5}$$

(Remember that the site label on k refers to a site of the original lattice, while the site label on σ refers to the corresponding site of the dual lattice.) Labeling the four dual sites which surround a given site of the original lattice by 1 through 4, we have

$$\textstyle\sum_i k_\mu = 2 - \tfrac{1}{2}(\sigma_1\sigma_2+\sigma_2\sigma_3+\sigma_3\sigma_4+\sigma_4\sigma_1), \tag{2.6}$$

which is even for any set of $\{\sigma_i = \pm 1\}$. We must also show that the representation (2.5) is necessary; i.e., that any set of the k_μ's satisfying the δ functions in Eq. (2.4) can be written in the form (2.5). This can be done by writing $k_\mu = \frac{1}{2}(1+e^{iL_\mu})$, with $L_\mu = 0, \pi$. Then the δ function can be regarded as a condition enforcing divergencelessness of the L_μ, and the representation (2.5) becomes the statement that L_μ may be written as a curl. [Actually, Eq. (2.5) is necessary only up to a sign; see below.] An analogous proof will be carried out more explicity when we discuss the Z_N and $U(1)$ symmetric theories.

Using Eq. (2.5) in Eq. (2.4), we can therefore drop the δ functions, and we have

$$Z = \tfrac{1}{2}2^N\sum_{\{\sigma\}}\prod_{l_d} C_{(1-\sigma_i\sigma_j)/2}(\beta), \tag{2.7}$$

where N is the number of lattice sites (we will eventually take $N\to\infty$), and the product over l_d means a product over links of the dual lattice, which is obviously the same as a product over links of the original lattice since they are in a one-to-one correspondence. Finally, the extra factor of $\frac{1}{2}$ appears in Eq. (2.7) since we now sum over $\{\sigma\}$ rather than $\{k\}$. From Eq. (2.5) it is clear that this counts each configuration of $\{k\}$ twice (i.e., $\{\sigma_i\}\to\{-\sigma_i\}$ gives the same $\{k\}$).

From Eq. (2.3) it is easy to see that since $k=0,1$, the $C_k(\beta)$ can be written in the form

$$\begin{aligned} C_k(\beta) &= \cosh\beta[1+k(\tanh\beta-1)] \\ &= \cosh\beta\exp[k\ln\tanh\beta] \\ &= (\cosh\beta\sinh\beta)^{1/2}\exp(-\sigma_i\sigma_j\tfrac{1}{2}\ln\tanh\beta), \end{aligned} \tag{2.8}$$

where in the last line we have used expression (2.5) for k. Inserting Eq. (2.8) in Eq. (2.7), we have

$$\begin{aligned} Z &= \tfrac{1}{2}(2\cosh\beta\sinh\beta)^N\sum_{\{\sigma\}}\exp\Big(-\tfrac{1}{2}\ln\tanh\beta\sum_{\langle\,\rangle}\sigma_i\sigma_j\Big) \\ &= \tfrac{1}{2}(\sinh 2\tilde{\beta})^{-N}\sum_{\{\sigma\}}\exp\Big(\tilde{\beta}\sum_{\langle\,\rangle}\sigma_i\sigma_j\Big), \end{aligned} \tag{2.9}$$

where $\tilde{\beta} = -\frac{1}{2}\ln\tanh\beta$ is the "dual inverse temperature," and the sum in the exponent is over nearest neighbor pairs on the dual lattice.

Now, except for an overall spin-independent factor, Eq. (2.9) is in the form of a partition function for a system of Ising spins but at an inverse temperature $\tilde{\beta}$. We note that $\tilde{\beta}$ is a monotonically decreasing function of β (shown schematically in Fig. 3) so that the high-temperature region of the original theory (2.2) is mapped into the low-temperature region of its dual rep-

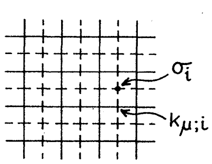

FIG. 2. Two-dimensional lattice (solid lines) and its dual (dashed lines). The dual variables σ_i sit on the sites of the dual lattice.

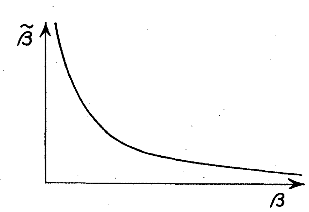

FIG. 3. Graph showing, schematically, the dual inverse temperature $\tilde{\beta}$ as a function of β.

resentation (2.9). This interchange of high and low temperatures (or, in the language of field theory, large and small coupling constants) as we go from the original theory to its dual is a general feature of the duality transformation, and is one of the properties that makes it so intriguing and so useful. Typically, one is able to perform some sort of perturbation theory when some parameter of the theory is small. In field theory it is the coupling constant. In the Ising model, we can do a high-temperature expansion when β is small. However, such a procedure may break down when the parameter gets too large. But since the duality transformation reverses the role of high and low temperatures, we can easily compute low-temperature properties of the Ising model by performing a high-temperature expansion in $\tilde{\beta}$ using Eq. (2.9).

Another interesting property of Eq. (2.9) is that is has the same functional form as Eq. (2.2); that is, the dual of the two-dimensional Ising model is also a two-dimensional Ising model. As we shall see, this property of self-duality is not a general feature of duality transformations, but occurs only for certain theories. In the present example, self-duality has the following simple consequence: Consider the free energy of the two-dimensional Ising model, $F_I(\beta)=\lim_{N\to\infty}(1/N)\ln Z(\beta)$, with Z defined in Eq. (2.2) for a lattice of N sites. Using Eq. (2.9), we have

$$F_I(\beta) = -\sinh 2\tilde{\beta} + F_I(\tilde{\beta})\,. \tag{2.10}$$

This is evidently a fairly strong restriction on the behavior of the free energy. For instance, let us suppose that $F_I(\beta)$ has only one point of nonanalyticity as a function of β. Then, since $\tilde{\beta}(\beta)$ is a monotonically decreasing function, the singularity must occur at the value

$$\beta_c = -\tfrac{1}{2}\ln\tanh\beta_c = \tilde{\beta}_c\,. \tag{2.11}$$

This is, in fact, the critical point of the two-dimensional Ising model. If the model had had more than one critical point, Eq. (2.10) would not have determined all their values, but would have determined relations between pairs of critical points.

2. Interpretation of the dual variables and disorder correlation functions

How are we to interpret the σ_i's that appear in Eq. (2.9)? We recall that the two-dimensional Ising model has two phases as a function of β. At low temperatures ($\beta>\beta_c$), the spins tend to point in the same direction and the order parameter defined by

$$\langle s_i\rangle = \lim_{h_j\to 0^+}\frac{d}{dh_i}\ln\sum_{\{s\}}\exp\Big(\beta\sum_{\langle\ \rangle}s_is_j+\sum_j h_j s_j\Big) \tag{2.12}$$

is nonzero. At high temperatures, on the other hand ($\beta<\beta_c$), $\langle s_i\rangle=0$. Because this model has only two phases and is self-dual, the situation for the σ_i's is precisely the reverse. Defining $\langle\sigma_i\rangle$ using the dual representation (2.9) in a way analogous to Eq. (2.12), we find that for $\beta>\beta_c$, when $\langle s_i\rangle\neq 0$, $\langle\sigma_i\rangle=0$, and for $\beta<\beta_c$ when $\langle s_i\rangle=0$, $\langle\sigma_i\rangle\neq 0$. $\langle\sigma_i\rangle$ is therefore a disorder parameter—at very high temperatures, when any configuration of the $\{s_i\}$ is almost equally likely, most configuration of the $\{\sigma_i\}$'s are relatively unlikely, and the σ_i's all tend to point in the same direction. It is important to stress that there is no simple one to one correspondence between configurations of the $\{s_i\}$'s and configurations of the $\{\sigma_i\}$'s. Rather, as in an ordinary Fourier transform, one must in general sum over all configurations of s's to produce a configuration of the σ's. Indeed, looking back to the derivation of Eq. (2.9), we see in Eq. (2.4) that it is the sum over all the s's which produces the δ functions which make the introduction of the σ_i's in Eq. (2.5) useful. Roughly speaking, one can say that when two nearest-neighbor s_i's are equally likely to point in the same or opposite directions (β small), the σ_i's which are connected by the dual link which crosses the link between the two s_i's will point in the same direction, whereas when the two s_i's are lined up, the corresponding σ_i's are with nearly equal probability aligned or misaligned. This statement is not precise because all the spins on the lattice are coupled, but it is a useful hueristic picture.

To gain some further insight into the meaning of the dual variables, it is instructive to calculate the disorder-disorder correlation function $\langle\sigma_l\sigma_m\rangle$, and express it in terms of the original spin variables by performing the duality transformation on the dual system defined by Eq. (2.9). Since $\tilde{\tilde{\beta}}=\beta$ (i.e., $-\frac{1}{2}\ln\tanh\tilde{\beta} = -\frac{1}{2}\ln\tanh(-\frac{1}{2}\ln\tanh\beta)=\beta$), it should be clear that if we perform the duality transformation (2.3)–(2.8) on the expression (2.9), we will just be led back to Eq. (2.2). But now we wish to perform the same operation on the object

$$\langle\sigma_l\sigma_m\rangle = \sum_{\{\sigma\}}\sigma_l\sigma_m\exp\Big(\tilde{\beta}\sum_{\langle\ \rangle}\sigma_i\sigma_j\Big)\Big/\sum_{\{\sigma\}}\exp\Big(\tilde{\beta}\sum_{\langle\ \rangle}\sigma_i\sigma_j\Big). \tag{2.13}$$

Since we know that the denominator just transforms into something proportional to Eq. (2.2), it is enough to apply the duality transformation to the numerator.

As in Eq. (2.3), we first rewrite the numerator

$$\begin{aligned} N_{l,m} &= \sum_{\{\sigma\}}\sigma_l\sigma_m\exp\Big(\beta\sum_{\langle\ \rangle}\sigma_i\sigma_j\Big)\\ &=\sum_{\{\sigma\}}\sigma_l\sigma_m\prod_{\langle\ \rangle}\exp(\beta\sigma_i\sigma_j)\\ &=\sum_{\{\sigma\}}\sigma_l\sigma_m\prod_{\langle\ \rangle}\sum_{k=0}^{1}C_k(\tilde{\beta})(\sigma_i\sigma_j)^k\,, \end{aligned}\tag{2.14}$$

where the C_k are given in Eq. (2.3) and we remind the reader that the sums and products are over pairs on the dual lattice, and l and m are two sites of the dual lattice. Collecting together all factors of a given σ_i, we rewrite Eq. (2.14) as

$$\begin{aligned} H_{l,m} &= \sum_{\{k\}}\sum_{l_d}C_k(\beta)\sum_{\{\sigma\}}(\sigma_l)^{1+\Sigma_l k}(\sigma_m)^{1+\Sigma_m k}\prod_i{}'\sigma_i\\ &=2^N\sum_{\{k\}}\prod_{l_d}C_k(\beta)\,\delta_2(1+\textstyle\sum_l k)\,\delta_2(1+\sum_m k)\\ &\quad\times\prod_i{}'\delta_2(\textstyle\sum_i k)\,, \end{aligned}\tag{2.15}$$

where the prime on the last product denotes a product over all dual sites except l and m.

Now, if we choose the represention analogous to Eq. (2.5) (with σ_i replaced by s_i) we will not satisfy all the

δ functions in Eq. (2.15), in particular the ones associated with sites l and m. To construct a satisfactory representation for the k's we draw an arbitrary line L along the dual lattice between l and m as in Fig. 4. We then choose the following representation for the k's:

$$k_{\mu;\,i}=\tfrac{1}{2}(1-s_i s_{i+\hat{\nu}})\,,\quad k_{\mu;\,i}\notin L \tag{2.16a}$$

and

$$k_{\mu;\,i}=\tfrac{1}{2}(1+s_i s_{i+\hat{\nu}})\,,\quad k_{\mu;\,i}\in L\,. \tag{2.16b}$$

That is, for each dual link not along the line connecting l and m we choose the representation (2.15a) which is analogous to Eq. (2.5), while for each $k_{\mu;\,i}$ which lies on L we choose Eq. (2.16b). As in Eq. (2.5), the s_i's are spins associated with the sites of the lattice dual to the one in question—in this case, the original lattice (2.2). It is easy to see that Eq. (2.16) satisfies all the δ functions of Eq. (2.15). Inserting Eq. (2.16) in Eq. (2.15) and recalling Eq. (2.8), we have

$$N_{l,m}=\tfrac{1}{2}(\sinh 2\beta)^{-N}\sum_{\{s\}}\exp\Big(\sum_{\langle\,\rangle}\beta_{ij}s_i s_j\Big)\,, \tag{2.17}$$

where the sum in the exponent is over nearest-neighbor pairs of the (original) lattice. $\beta_{ij}=+\beta$ for all lattice links except for those links which intersect the dual path connecting l and m (see Fig. 4). For those links, $\beta_{ij}=-\beta$. Thus $N_{l,m}$ is a partition function for an Ising system with a certain specific mixture of ferromagnetic and antiferromagnetic bonds, and $\langle\sigma_l\sigma_m\rangle$ is the ratio of that partition function to the partition function for the system with all ferromagnetic bonds. [This result was also obtained in Kadanoff and Ceva (1971).]

There are several comments to be made about this result. First we note that since the line chosen to connect l and m in Fig. 4 was arbitrary, there should be an infinite number of choices of sets of β_{ij} which will give the same result in Eq. (2.17). To see that this is true, we note that Eq. (2.17) is invariant under the combined operation of changing the definition of one spin, s_i, to $-s_i$, and simultaneously changing the sign of all the (four) β_{ij}'s associated with the four links which impinge on this spin. This corresponds to changing the path joining l and m as indicated in Fig. 5. Thus the exponent in $N_{l,m}$ has a restricted kind of local Z_2 gauge invariance analogous to the local gauge invariance of quantum electrodynamics. We shall have more to say about gauge invariance later.[2]

Next we observe from Figs. 4 and 5 that the product of the signs of the β_{ij} around any elementary square (or plaquette) of the lattice is +1, except those that surround the dual lattice sites l and m, where it is −1. Let us imagine that the system described by the partition function $N_{l,m}$, Eq. (2.17), is at a very low temperature (β large) and let us ask how the spins tend to point. Those connected by a ferromagnetic bond will

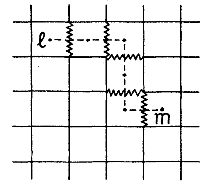

FIG. 4. Disorder–disorder correlation function for the $d=2$ Ising model. The bonds associated with the wiggly lattice links are antiferromagnetic in Eq. (2.17).

want to point in the same direction, while those connected by an antiferromagnetic bond will want to point in opposite directions. For the most part, the distribution of ferromagnetic and antiferromagnetic bonds is such that the spins easily lock into place when $\beta\gg 1$. Consider, however, the plaquette surrounding site l in Fig. 4. Start in the lower left-hand corner of this plaquette and let the spin on this site point up. Because of the ferromagnetic bond, the spin on its right will want to point up also, while the spin above that one will want to point down due to the antiferromagnetic bond. Continuing around the plaquette, counterclockwise, the spin in the upper left-hand corner will want to point down, and finally, we come back to our original spin which now wants to point down, contrary to its original assignment. This phenomenon occurs when there are an odd number of antiferromagnetic links surrounding a plaquette and is called frustration (Edwards and Anderson, 1975; Toulouse, 1977). Certain materials, notably spin glasses, can be modeled by spin systems in which a thermal average over spins is performed for a given distribution of antiferromagnetic bonds and the free energies thus obtained are then averaged over different distributions of antiferromagnetic bonds. Because of the gauge invariance discussed above, it is clear that different configurations of bonds give rise to different free energies only if they have different distributions of frustrated plaquettes. It is also clear that averages over all the n-point disorder correlation functions $\langle\sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_n}\rangle$ should be related in certain models to the spin glass average free energy. In the simplest, most primitive, version of the Edwards-Anderson model of a spin glass (Edwards and Anderson, 1975) we consider an Ising model with a random distribution of ferromagnetic and antiferromagnetic couplings. Suppose a given coupling can take on only two values and is equally likely to be ferromagnetic or antiferromagnetic, independent of the values of the other couplings. The free energy of a single sample of spin glass with a fixed distribution of antiferromagnetic bonds is

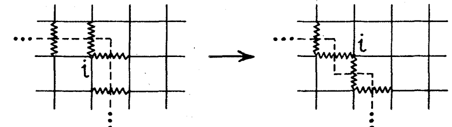

FIG. 5. Gauge transformation under which Eq. (2.17) is invariant. This figure shows how the gauge transformation is equivalent to a deformation of the path joining l and m in Fig. 4.

[2]Of course the ordinary Ising model ferromagnetic partition function can also be written such that certain configurations of the links are antiferromagnetic. The allowed configurations of antiferromagnetic links are those that correspond in the sense of Fig. 4 to closed paths on the dual lattice. Note that this includes the configuration with all links antiferromagnetic.

$$f(\beta)=\ln\sum_{\{s\}}\exp\left(\sum\beta_{ij}\,s_is_j\right),$$

where $\beta_{ij}=\pm\beta$. It is clear that there are many sets $\{\beta_{ij}\}$ which are related to each other by a redefinition of spins and links as discussed above. All these gauge related configurations will give rise to the same f. The property that gives rise to different values of f is different distributions of frustrated plaquettes. Since one associates a frustrated plaquette with the insertion of a disorder variable on the corresponding dual lattice site, it is clear that we can relate $f(\beta)$ to some n-point disorder correlation function. But because the two-dimensional Ising model is self-dual, these are related to the n-point spin correlation functions. In particular,

$$f(\beta)=F_I(\beta)+\ln\langle s_1\cdots s_n\rangle(\tilde{\beta})\,. \tag{2.18a}$$

$F_I(\beta)$ is the free energy of the usual Ising model with all ferromagnetic bonds, and the spins in the n-point correlation function are associated with the dual lattice sites which are located at the center of each frustrated plaquette. A quantity of particular interest in spin glass theory is the average of $f(\beta)$ over all possible configurations of ferromagnetic and antiferromagnetic bonds. Since there are $\sim 2^N$ configurations of different bonds that have the same distribution of frustrated plaquettes regardless of the distribution of these plaquettes, this average free energy becomes (in the simple case we are considering)

$$\begin{aligned}\bar{f}(\beta)&=\frac{1}{4^N}\sum_{\{\beta_{ij}\}}\ln\sum_{\{s\}}\exp\left(\sum\beta_{ij}s_is_j\right)\\&=F_I(\beta)+\frac{1}{4^N}\sum_n\sum_{i_1\cdots i_n}\ln\langle s_{i_1}\cdots s_{i_n}\rangle(\tilde{\beta})\,,\end{aligned} \tag{2.18b}$$

where the sum over the correlation functions is a sum first over all positions of the n spins $(i_k\neq i_l)$ and then a sum over all numbers of spins.

Finally, we note that the representation (2.17) is very useful for calculating $\langle\sigma_l\sigma_m\rangle$ at low temperatures. Suppose $|l-m|\equiv L\gg 1$ and let us imagine that we choose to have as few antiferromagnetic β_{ij}'s as possible. It is then easy to see (particularly if we imagine toroidal boundary conditions) that the most important contributions to Eq. (2.17) will occur when there are only of order L pairs of spins such that $\beta_{ij}s_is_j=-\beta$. Hence, we find

$$\langle\sigma_l\sigma_m\rangle=N_{l,m}/Z\sim e^{-2\beta L};\quad \beta\gg 1\,. \tag{2.19}$$

Thus $\langle\sigma_l\sigma_m\rangle$ falls exponentially to zero precisely because the original spin degrees of freedom are in a highly ordered state. From another point of view, we can say that $\langle\sigma_l\sigma_m\rangle$ goes to zero exponentially when β is very large because the σ degrees of freedom think they are at a high temperature (the dual temperature) and are therefore in a very disordered state.

B. Three-dimensional Ising model and the Z_2 gauge theory

1. Dual form of the three-dimensional Ising model

We now consider the Ising model defined by Eq. (2.2), but on a three-dimensional simple cubic lattice. As in two dimensions, the sum in the exponent runs over all nearest-neighbor pairs on the lattice. In this case the duality transformation will lead us to a theory which is much different than the one encountered in two dimensions.

Proceeding as in the last subsection, we first write

$$\begin{aligned}Z&=\sum_{\{s\}}\exp\left(\beta\sum_{\langle\,\rangle}s_is_j\right)=\sum_{\{s\}}\prod_{\langle\,\rangle}\exp(\beta s_is_j)\\&=\sum_{\{s\}}\prod_{\langle\,\rangle}\sum_{k=0}^{1}C_k(\beta)(s_is_j)^k\,,\end{aligned} \tag{2.20}$$

with the $C_k(\beta)$ defined below Eq. (2.3). Collecting together all the factors of each s_i, we have

$$\begin{aligned}Z&=\sum_{\{s\}}\sum_{\{k\}}\prod_l C_{k_\mu}(\beta)\prod_i(s_i)^{\Sigma_i k_\mu}\\&=\sum_{\{k\}}\prod_l C_{k_\mu}(\beta)\prod_i 2\delta_2\left(\textstyle\sum_i k_\mu\right).\end{aligned} \tag{2.21}$$

This is similar to expression (2.4), but since there is one k_μ for each link of the lattice, k_μ is a three vector in the present case, μ running over the three lattice directions. Furthermore, the sum $\sum_i k_\mu$ denotes here a sum over the six k's associated with the six links which impinge on the site i.

To satisfy the δ function in this case is slightly more complicated than in two dimensions. First we define the dual lattice by constructing another simple cubic lattice which is displaced from our original lattice by half a lattice spacing in each direction. Thus the vertices of the dual lattice lie in the centers of the elementary cubes of the original lattice and vice versa. In Fig. 6 we have drawn a piece of the dual lattice interleaved with the original one. As we see from this figure each link of the original lattice penetrates an elementary two-dimensional face or plaquette of the dual lattice. We associate with each link of the dual lattice a variable $A_{\mu;i}$ which takes on values ± 1. We now write

$$\begin{aligned}k_{\mu;i}&=\tfrac{1}{2}(1-A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat{\lambda}}A_{\lambda;i+\hat{\nu}})\\&=\tfrac{1}{2}\Bigl(1-\prod_{\substack{\text{dual}\\\text{plaquette}}}A_\nu\Bigr),\quad \mu\neq\nu,\lambda\,,\end{aligned} \tag{2.22}$$

where the A_ν's are those associated with the four dual links which border the dual plaquette through which the original link with which $k_{\mu;i}$ is associated passes. (Note that the site and direction labels on k refer to the original lattice while those on the A's refer to the dual lattice.) As in two dimensions, the representation (2.22) is (up to a sign) both necessary and sufficient to satisfy the δ functions. To see that this representation

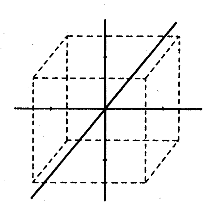

FIG. 6. Piece of the three-dimensional simple cubic lattice (solid lines) and its dual (dashed lines).

for the k's does automatically satisfy the δ functions in Eq. (2.21) is easy. Note first that the sum is zero when all the A_ν's are one, and second observe that since each A_ν appears in the expressions for two different k's in a given sum, changing the value of any A_ν changes two k's in the sum, each by one unit. Hence the total sum is always even if it is even for any configuration of A_ν's.

We now insert Eq. (2.22) in Eq. (2.21). Following the development from Eq. (2.7) to Eq. (2.9) we have

$$A=2^N \sum_{\{k(A)\}} \prod_{P_d} C_{1/2(1-A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu})}(\beta)$$
$$=2^N(\cosh\beta\sinh\beta)^{3N/2} \times \sum_{\{k(A)\}} \exp\Big(\tilde\beta \sum_{P_d} A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}\Big), \tag{2.23}$$

where the sum in the exponent is over all plaquettes on the dual lattice, and $\tilde\beta=-\frac{1}{2}\ln\tanh\beta$, as in Eq. (2.9). Now, the sum over $\{k(A)\}$ means that we are to sum over all distinct sets of k_μ which can be obtained from sets of A_ν via Eq. (2.22). Another way to say this is that we are to sum over sets of A_ν's such that each different set of A_ν's produces a distinct set of the k_μ's. As we shall now show, not all sets of A_ν's give distinct sets of k_μ's. Thus summing blindly over all sets of A_ν's could, in principle, produce an overcounting problem.

Consider some set $\{A_\nu\}$ which produces, using Eq. (2.22), a set $\{k_\mu\}$. Now pick a site of the dual lattice and change the sign of all the (six) A_ν's which lie on the six dual links attached to that site. Since each of the 12 dual plaquettes which have the chosen dual site as a corner have had two of the A_ν's which lie on their edge change sign, the k_μ's generated by Eq. (2.22) are unchanged. Since this operation can be done independently at each dual lattice site, there are of the order of 2^N different configurations of the A_ν which generate the same $\{k_\mu\}$. Moreover, since the exponent in Eq. (2.23) is only a function of $\{k_\mu\}$, we see that the Hamiltonian [or, if we think of Eq. (2.23) as describing a field theory, the Lagrangian] is also invariant under the local operation described in the last paragraph. Thus Eq. (2.23) is a theory with a local gauge invariance. In fact, the gauge invariance is very similar to the familiar gauge invariance of QED. To see this, we define $T_{\mu;i}$ by

$$A_{\mu;i}=e^{iT_{\mu;i}}, \tag{2.24}$$

so that $T_{\mu;i}=0,\pi$. The gauge symmetry described above can also be expressed by saying that the Lagrangian is invariant under the operation

$$T_{\mu;i}\to T'_{\mu;i}=T_{\mu;i}+\Delta_\mu\Lambda_i, \tag{2.25}$$

where the nearest-neighbor difference operator Δ_μ is defined by $\Delta_\mu c_i\equiv c_i-c_{i-\hat\mu}$, and Λ_i is an arbitrary scalar defined on the sites of the lattice in question (in this case the lattice dual to the original Ising model lattice) taking on values 0 or π on each site. If we identify $T_{\mu;i}$ with the vector potential of QED and Λ_i with the arbitrary gauge function, then Eq. (2.25) is the familiar form of the local gauge invariance of that theory. In ordinary QED the vector potential and the gauge function take on all real values, but in Eq. (2.25) these values are restricted to be 0 and π. Thus the group of the local gauge symmetry of Eq. (2.23) is Z_2 rather than R [or $U(1)$, in the presence of matter fields] as in ordinary QED.

Now there are two ways to deal with the overcounting problem in Eq. (2.23). The first is the usual procedure of choosing a gauge. We shall discuss gauge choices later in this section, and again in Sec. IV, so we shall say no more about them here. Our second option is to ignore the overcounting problem. This can certainly be done in principle since, as is clear, the overcounting is uniform; that is, by summing independently over all $A_{\mu;i}$ we reproduce each distinct configuration of k_μ's the same number of times (about 2^N), and so ignoring the gauge problem just gives us an extra (β-independent) factor in front of Eq. (2.23). Moreover, since Z_2 is compact we get no extra infinities (except one overall infinity when $N\to\infty$) when we ignore gauge fixing. Thus we expect that the procedure of not fixing a gauge will still allow us to calculate finite values of gauge invariant quantities. (We also note that the procedure of summing over all gauges gives zero when calculating any non-gauge-invariant quantity.) Thus we can write

$$Z=2^{-N/2}(\sinh 2\tilde\beta)^{-3N/2}\sum_{\{A\}}{}' \exp\Big(\tilde\beta\sum_{P_d} A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}\Big), \tag{2.26}$$

where the prime on the sum means that, in principle, we should choose a gauge, but it can usually be ignored.

2. Dual form of the three-dimensional Z_2 gauge theory

Before describing some of the physical consequences of Eq. (2.26), we want to demonstrate that, as in two dimensions, applying the duality transformation to Eq. (2.26) brings us back to the original theory (2.20). We proceed in analogy with the development from Eq. (2.20) to Eq. (2.23). We write Eq. (2.26) as

$$Z=2^{-N/2}(\sinh 2\tilde\beta)^{3N/2}\sum_{\{A\}}{}'\prod_{P_d}\exp(\tilde\beta A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu})$$
$$=2^{-N/2}(\sinh 2\tilde\beta)^{3N/2}\sum_{\{A\}}{}'\sum_{\{k\}=0}^{1}\prod_{P_d}C_{k_{\nu\lambda;i}}(\tilde\beta) \times(A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu})^{k_{\nu\lambda;i}}. \tag{2.27}$$

In the last line of Eq. (2.27), we have associated one integer $k_{\nu\lambda;i}$ with each plaquette of the dual lattice [i.e., the lattice on which the gauge theory (2.26) is defined]. To define a plaquette, we need two direction indices ν and λ for its orientation, as well as a site index i to tell us where in the lattice it is. The $C_k(\tilde\beta)$'s are again those that appear after Eq. (2.3b). Rearranging the factors in Eq. (2.27) we have

$$Z=2^{-N/2}(\sinh 2\tilde\beta)^{3N/2}2^{-N}\sum_{\{k\}}\prod_{P_d}C_k(\tilde\beta) \times\prod_{l_d}\sum_{A=0}^{1}(A_{\nu;i})^{\Sigma_{\nu;i}k}. \tag{2.28}$$

In Eq. (2.28) we have replaced $\sum'_{\{A\}}\to 2^{-N}\sum_{\{A\}}$, i.e., we do an unrestricted sum over the $A_{\nu;i}$, dividing out the redundancy as discussed above. (Note that this procedure explicitly does *not* involve choosing a gauge.) The symbol $\sum_{\nu;\,i} k$ means a sum over all those $k_{\mu\nu;\,i}$'s which are associated with plaquettes which have as one of their edges the link defined by lattice position i and direction ν. In three dimensions this sum involves four terms (see Fig. 7).

Summing over the $\{A\}$, Eq. (2.28) becomes

$$Z=(\tfrac{1}{2}\sinh\tilde{\beta})^{3N/2}\sum_{\{k\}}\prod_{P_d}C_k(\tilde{\beta})\prod_{l_d}2\delta_2\left(\textstyle\sum_{\nu;\,i}k\right). \tag{2.29}$$

As before we seek a representation for the $k_{\nu\lambda;\,i}$ which automatically satisfies the δ functions in Eq. (2.29). Let us denote the lattice on which the gauge theory is defined by D. The lattice on which the original Ising model (2.20) is defined we will call O. Now, we construct the lattice which is dual to D by shifting D by half a lattice spacing in each direction. This lattice, O', clearly coincides with O. We now note (Fig. 7) that each plaquette of D is penetrated by a link of O'. With each site of O', we associate a spin r_i, which can take on values ± 1. We now use the representation

$$k_{\nu\lambda;\,i}=\tfrac{1}{2}(1-r_i r_{i+\hat{\mu}}),\quad \mu\neq\nu,\lambda, \tag{2.30}$$

where the spins r are those that sit on the sites of O' joined by the link which penetrates the plaquette of D with which $k_{\nu\lambda;\,i}$ is associated. The δ functions in Eq. (2.29) are automatically satisfied by Eq. (2.30). This is because each r_i appears twice in the sum $\sum_{\nu;\,i}k$, so if this sum is even for any configuration of r's, it is always even (see Fig. 7). [As before, Eq. (2.30) is, up to the sign in front of the rr term, both necessary and sufficient to satisfy the δ functions.] Inserting Eq. (2.30) in Eq. (2.29), we find after a little algebra

$$Z=\sum_{\{r\}}\exp\left(\beta\sum_{\langle\,\rangle}r_i r_j\right), \tag{2.31}$$

where the sum in the exponent runs over nearest-neighbor pairs on O'. This is exactly Eq. (2.20), if we identify $r_i\equiv s_i$ and the lattice $O'\equiv O$.

It is instructive to carry out the transformation leading from Eq. (2.26) to Eq. (2.31) once again, but this time paying attention to the prime on the sum over $\{A\}$ in Eq. (2.26) and explicitly choosing a gauge. There are of course many ways to do this. Here we will perform the exercise choosing an axial gauge. Using the freedom implied by the gauge transformation (2.25), it

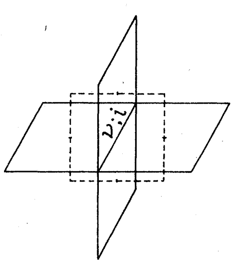

FIG. 7. Four plaquettes impinging on a lattice link in the $d=3$ Z_2 lattice gauge theory. The dashed lines are links of the lattice which is dual to *this* lattice. Ising spins are associated with the sites of the lattice to which the dashed links belong.

is easy to see that we can set all the $A_{\nu;\,i}$'s associated with all the links pointing in, say, the 3 direction equal to one. Now, let us suppose that our lattice has open edge boundary conditions, so that it is a cube with an $A_{\nu;\,i}$ associated with all the links, even those on the boundary. (These are the simplest boundary conditions with which to demonstrate axial gauge.) Then, going to the face of the lattice which lies at the negative end of the 3 axis, we can set all the $A_{\nu;\,i}$'s which point in the 2 direction and lie on this end face of the lattice equal to one. Finally, we can go the edge of the lattice which is the intersection of the faces at the negative ends of the 3 and 2 axes, and set all the $A_{\nu;\,i}$'s along this edge (all of which point in the 1 direction) to one. The freedom to fix the A_2's on a boundary face and the A_1's on a boundary edge is just a reflection of the usual residual gauge invariance which one encounters in axial gauge. In Eq. (2.25) it corresponds to the fact that after setting all A_3 equal to one, we can make additional gauge transformations with Λ_i which are independent of x_3.

Having completely specified a gauge in this manner, we must sum over all the remaining $A_{\mu;\,i}$'s. Now, for each gauge fixed $A_{\nu;\,i}$, we will be missing a δ function in Eq. (2.29). Let us concentrate on one site of the lattice D. Not summing over A_3 means that we have δ functions associated only with four of the six links which impinge on the site in question. There are four plaquettes associated with each link in the 3 direction. Label these by a, b, c, and d (for one link) and a', b', c', and d' for the other link. There remain another four plaquettes which have the chosen site as a corner. Label these α, β, γ, and δ. It is then easy to see that the four δ functions obtained by summing over the four unconstrained A_ν's imply

$$\begin{aligned}
&k_a+k_b+k_\alpha+k_\beta=\text{even}\,,\\
&k_{a'}+k_{b'}+k_\gamma+k_\delta=\text{even}\,,\\
&k_c+k_{c'}+k_\alpha+k_\gamma=\text{even}\,,\\
&k_d+k_{d'}+k_\beta+k_\delta=\text{even}\,.
\end{aligned} \tag{2.32}$$

Adding these equations together, we find

$$k_a+k_b+k_c+k_d+k_{a'}+k_{b'}+k_{c'}+k_{d'}=\text{even}\,. \tag{2.33}$$

Thus if the sum of k's about one x_3 link is even, so is the other.

Now, go to the face of the lattice which lies at the positive end of the 3 axis. Concentrate on one site on this face. Repeat the labeling procedure defined above for this site. Since there is only one link in the 3 direction coming into this site, the plaquettes labeled by a', b', c' and d' are absent. Hence, we find the analog of Eq. (2.33) for this site is

$$k_a+k_b+k_c+k_d=\text{even}\,,\quad \text{at the positive } x_3 \text{ face of the lattice}. \tag{2.34}$$

Using Eqs. (2.34) and (2.33) applied to a lattice site one layer in from the lattice edge, we find that $k_a+k_b+k_c+k_d$ is even for both links in the 3 direction which impinge on this site. Continuing in the same way we find that the sum of the k's around every lattice link in the 3 direction is even. Similar, but simpler arguments can be used to show that the gauge fixed links (in the 2 and 1 directions) on the lattice boundaries also have the sum of k's associated with the plaquettes which impinge on them equal to an even integer. Thus, all the information contained in the (overcomplete) set of δ functions in Eq. (2.29) is implicitly contained in the smaller set of δ functions which we get when we fix a gauge. Proceeding with this derivation, we find that the numerical factors (i.e., factors of 2^N) work out just right (as they must) to give the result identical to Eq. (2.31).

3. Correlation functions

The equivalence between three-dimensional spin and gauge systems is very intriguing. To help us understand this equivalence better, it is useful to write the expression for spin-spin correlation functions in terms of the gauge theory representation, and to write a gauge field correlation function in terms of the spin variables.

We begin by applying the duality transformation to the three-dimensional Ising model spin-spin correlation function:

$$\langle s_n s_m\rangle = \sum_{\{s\}} s_n s_m \exp\Big(\beta \sum_{\langle\ \rangle} s_i s_j\Big) \Big/ \sum_{\{s\}} \exp\Big(\beta \sum_{\langle\ \rangle} s_i s_j\Big). \tag{2.35}$$

The denominator of Eq. (2.35) just transforms into the partition function for the gauge theory, Eq. (2.26), so we only need consider the duality transformation applied to the numerator of Eq. (2.35). Rewriting this expression in the by now familiar way, we have

$$\begin{aligned}
N_{n,m} &= \sum_{\{s\}} s_n s_m \exp\Big(\beta \sum_{\langle\ \rangle} s_i s_j\Big) \\
&= \sum_{\{s\}} s_n s_m \prod_l \exp(\beta s_i s_j) \\
&= \sum_{\{s\}} s_n s_m \prod_l \sum_k C_k(\beta)(s_i s_j)^k \\
&= \sum_{\{k\}} \prod_l C_k(\beta) \sum_{\{s\}} s_n s_m \prod_i (s_i)^{\Sigma_i k} \\
&= 2^N \sum_{\{k\}} \prod_l C_k(\beta)\,\delta_2\big(1+\textstyle\sum_n k\big) \\
&\quad \times \delta_2\big(1+\textstyle\sum_m k\big) \displaystyle\prod_{i\neq n,m} \delta_2\big(\textstyle\sum_i k\big),
\end{aligned} \tag{2.36}$$

where there is one $k_{\mu;i}$ associated with each link of the lattice. We can satisfy the δ functions by choosing the following representations for the $k_{\mu;i}$: we draw an arbitrary line along the lattice connecting the two sites l and m. For each k associated with a link which lies on this line we write

$$k_{\mu;i} = \tfrac{1}{2}(1 + A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}), \quad k_{\mu;i} \in L, \tag{2.37a}$$

while for each k associated with a link which is not on this line, we write

$$k_{\mu;i} = \tfrac{1}{2}(1 - A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}), \quad k_{\mu;i} \notin L \tag{2.37b}$$

where the A_ν's are associated with the links of the dual lattice as described earlier in this section.

We can now use Eq. (2.37) in Eq. (2.36) and, following the steps leading from Eq. (2.21) to Eq. (2.23), we arrive at the expression

$$N_{n,m} = 2^{-N/2}(\sinh 2\tilde\beta)^{3N/2} \times \sum_{\{A\}}{}' \exp\Big(\tilde\beta \sum_{P_d} \eta_{\nu\lambda;i} A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}\Big), \tag{2.38}$$

where $\eta_{\nu\lambda;i}$ equals -1 for all those dual plaquettes penetrated by links which lie in L, Eq. (2.37a), and equals $+1$ otherwise. Thus

$$\langle s_n s_m\rangle = \Big[\sum_{\{A\}}{}' \exp\Big(\tilde\beta \sum_{P_d} \eta_{\nu\lambda;i} A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}\Big)\Big] \times \Big[\sum_{\{A\}}{}' \exp\Big(\tilde\beta \sum_{P_d} A_{\nu;i}A_{\lambda;i}A_{\nu;i+\hat\lambda}A_{\lambda;i+\hat\nu}\Big)\Big]^{-1} \tag{2.39}$$

and is the ratio of two gauge theory partition functions which differ only in that some of the complings of the partition function in the numerator are "antiferromagnetic."

This result is reminiscent of the form we obtained for the disorder-disorder correlation function in terms of the original degrees of freedom in the $d=2$ Ising model. Indeed, many of the comments we made in connection with expression (2.17) have their counterpart here. In particular, since the line, L, defined in Eq. (2.37) is arbitrary, the set of dual plaquettes for which $\eta_{\mu\nu;i} = -1$ is not unique, and can be changed by a redefinition of some of the A_ν's, analogous to the argument presented in the paragraphs following Eq. (2.17).

We now want to write a gauge field correlation function in terms of the Ising model spin variables. We note first that a simple correlation function such as $A_{\nu;l}A_{\mu;m}$ is not gauge invariant and so is not a meaningful quantity to calculate. We could compute $G_{\mu\nu;l}G_{\lambda\sigma;m}$, where $G_{\mu\nu;l} \equiv A_{\mu;l}A_{\nu;l}A_{\mu;l+\hat\nu}A_{\nu;l+\hat\mu}$, which is gauge invariant, but a more interesting object to examine is the "Wilson loop" (Wilson, 1974)

$$\begin{aligned}
\Gamma_c(\tilde\beta) &= \Big\langle \prod_c A_{\nu;i}\Big\rangle \\
&= \sum_{\{A\}}{}' \prod_c A_{\nu;i} \exp\Big(\tilde\beta \sum_{P_d} G_{\nu\lambda;i}\Big) \Big/ \sum_{\{A\}}{}' \exp\Big(\tilde\beta \sum_{P_d} G_{\nu\lambda;i}\Big),
\end{aligned} \tag{2.40}$$

where the product Π_c denotes a product around some closed curve c. That this is gauge invariant can easily be seen by observing that

$$\prod_c A_{\nu;i} = \prod_a G_{\mu\nu;i}, \tag{2.41}$$

where Π_a is a product over all plaquettes which lie on any two-dimensional surface bounded by the curve c. Since $G_{\mu\nu;i}$ is gauge invariant, so is the left-hand side of Eq. (2.41).

Now, the denominator of Eq. (2.40) is just proportional to Z, so under duality it will be transformed (up to overall spin-independent factors) into expression (2.20). To transform the numerator, we write

$$N_c = \sum_{\{A\}}{}' \prod_c A_{\nu;\,i} \exp\Big(\tilde\beta \sum_{P_d} G_{\mu\nu;\,i}\Big) = \sum_{\{A\}}{}' \prod_c A_{\nu;\,i} \prod_{P_d}\sum_k C_k(\tilde\beta)(G_{\nu\lambda;\,i})^k$$
$$= 2^{-N}\sum_{\{k\}}\prod_{P_d} C_k(\tilde\beta)\sum_{\{A\}}\prod_{l_d\notin c}(A_{\nu;\,i})^{\Sigma_{\nu;\,i}k}\prod_c (A_{\nu;\,i})^{1+\Sigma_{\nu;\,i}k}$$
$$= 2^{-N}\sum_{\{k\}}\prod_{P_d} C_k(\tilde\beta)\prod_{l_d\notin c} 2\delta_2\Big(\sum_{\nu;\,i}k\Big)\prod_c 2\delta_2\Big(1+\sum_{\nu;\,i}k\Big)\,. \tag{2.42}$$

We now seek a representation for the $k_{\mu\nu;\,i}$'s which will satisfy the δ functions in Eq. (2.42). To do this we first choose an arbitrary two-dimensional surface on the gauge theory lattice which is bounded by the curve c. Call this surface t. Then, for each $k_{\mu\nu}\notin t$, we write

$$k_{\mu\nu;\,i}=\tfrac{1}{2}(1-r_i r_{i+\hat\lambda});\quad \lambda\neq\mu,\nu\ \ (k\notin t)\,, \tag{2.43a}$$

while for $k_{\mu\nu}\in t$, we have

$$k_{\mu\nu;\,i}=\tfrac{1}{2}(1+r_i r_{i+\hat\lambda});\quad \lambda\neq\mu,\nu\ \ (k\in t)\,. \tag{2.43b}$$

Equation (2.43a) is the same as Eq. (2.30). The r_i's sit on the sites of the lattice O', which is dual to the gauge theory lattice D, and take on the values ± 1. That this representation satisfies all the δ functions in Eq. (2.42) can be seen by noting that in all of the sums $\sum_{\nu;\,i}k$ except those on c there are an even number (either zero or two) of the k's which have the representation (2.43b). Thus these sums are always even. On the other hand, each sum $\sum_{\nu;\,i}k$ on c has exactly one k represented by Eq. (2.43b), and these sums are therefore odd.

Using Eq. (2.43) in Eq. (2.42) and using expression (2.8) for $C_k(\tilde\beta)$ we find

$$N_c \propto \sum_{\{r\}}\exp\Big(\sum_{\langle\,\rangle}\beta_{ij}r_i r_j\Big) \tag{2.44}$$

and

$$\Gamma_c=\sum_{\{r\}}\exp\Big(\sum_{\langle\,\rangle}\beta_{ij}r_i r_j\Big)\Big/\sum_{\{r\}}\exp\Big(\beta\sum_{\langle\,\rangle}r_i r_j\Big)\,. \tag{2.45}$$

In Eq. (2.44) we have dropped over all r-independent factors (these factors cancel in the expression for Γ_c). $\beta_{ij}=+\beta$ for all links except those which penetrate the (dual) plaquettes which lie on t. For those links, $\beta_{ij}=-\beta$. As with other correlation functions we have studied this one too can be represented as a ratio of partition functions for systems with different distributions of ferromagnetic and antiferromagnetic bonds. In this case, the antiferromagnetic bonds of the numerator are associated with and normal to an open two-dimensional surface embedded in the three-dimensional lattice. Since the surface was arbitrary (except for its boundary) we should be able to change the distribution of antiferromagnetic bonds in Eq. (2.44) consistent with this arbitrariness. This can be done, as pictured in Fig. 8, by changing the sign of a spin on some site and simultaneously changing the sign of all six β_{ij}'s which are connected to this site. If we choose a site coupled to one of the links passing through t (as in Fig. 8) this operation just corresponds to producing a small dimplelike deformation of t.

4. Comments

There are several aspects of the equivalence between the three-dimensional Ising model and the three-dimensional Z_2 gauge theory that deserve to be stressed. First, remembering that the duality transformation maps high temperatures into low temperatures of the dual theory, we are led to the conclusion that the high-temperature region of the $d=3$ Ising model is most simply described in terms of the gauge degrees of freedom of the dual form. This is really quite remarkable. The high-temperature, disordered region of the $d=3$ Ising model behaves as if it were a very cool gauge theory. Of course the converse is also true: the high-temperature region of the Z_2 gauge theory behaves as if it were a very low-temperature Ising model. It is also important to stress another feature of this three-dimensional duality transformation; namely, we can transform the gauge theory into a theory with no gauge invariance *without choosing a gauge*. The set of variables of the Ising model is a complete set of variables which replaces the gauge potentials. One might hope that some such transformation could be effected for the gauge theory known as quantum chromodynamics (QCD). This theory is the leading contender for a theory of the strong interactions, and has interactions which are invariant under a gauge symmetry. On the other hand, one might suppose that the set of observed hadrons constitutes a complete set of variables for the description of the strong interactions. But there is apparently no gauge principle (at least no obvious gauge principle) that dictates *directly* the interactions between the hadrons. Is it possible that there is some transformation, not unrelated to the duality transformation which takes us from the set of variables of QCD to the set of observed hadrons? What we have in mind here is a transformation which would invert the gauge coupling constant g in the same way that duality transformations invert the temperature in statistical systems, to produce a representation for the QCD vacuum which is simple at large g. Because of the non-Abelian nature of QCD it is not easy to perform duality transformations in a simple way and so this question is a difficult one. We shall return briefly to this problem in Sec. V.

Finally, let us briefly describe the phase structure

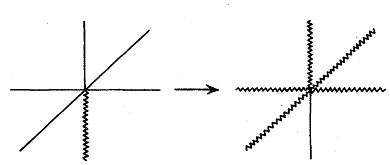

FIG. 8. Gauge transformation under which Eq. (2.44) is invariant. The wiggly lines indicate the bonds which are antiferromagnetic. This gauge transformation corresponds to a deformation of the surface subtended by the Wilson loop, Eq. (2.40), defined on the three-dimensional lattice dual to the one in this figure.

of the $d=3$ Ising model and the qualitative behavior of the spin and gauge correlation functions. We recall that the $d=3$ Ising model has two phases. In the low-temperature ordered phase $\langle s_i\rangle \neq 0$, and in the high-temperature disordered phase $\langle s_i\rangle = 0$. Moreover, we expect $\langle s_l s_m\rangle \to$ const ($\neq 0$) as $|l-m| = r \to \infty$ in the low-temperature phase, and $\langle s_l s_m\rangle \sim e^{-\mu r}$ as $r\to\infty$ in the high-temperature phase. (This latter behavior is very easy to deduce by simply expanding the expression for $\langle s_l s_m\rangle$ in powers of β.)

Now, since we can replace the set of Ising variables $\{s\}$ by the set of gauge variables $\{A_\nu\}$, we might expect that some correlation function of the gauge degrees of freedom should have qualitatively different large distance behavior in the two phases. Consider, in particular, Γ_c. Using Eq. (2.45), or Eq. (2.40), it is easy to see that when $\beta \gg 1$ (which implies that $\tilde\beta \ll 1$), $\Gamma_c \sim e^{-2\beta A}$, where A is the minimum area subtended by the curve c. This can be seen in Eq. (2.40) by expanding the exponents in powers of $\tilde\beta$ and noting that after summing over $\{A_\nu\}$, the lowest-order nonzero contribution to Γ_c requires that the entire surface subtended by c be filled with factors of $G_{\mu\nu;i}$. Each factor costs one power of $\tilde\beta$, so we have $\Gamma_c \sim (\tilde\beta)^A \sim e^{-2\beta A}$ ($\tilde\beta \ll 1$). In expression (2.45), we note that since $\beta \gg 1$, all the r_i connected by ferromagnetic bonds will want to be aligned, while those connected by antiferromagnetic bonds will want to be misaligned. Suppose we have periodic or fixed edge (i.e., all $r_i=+1$ on the edges of the lattice) boundary conditions. Then it is easy to see that the (of order A) antiferromagnetic bonds in the numerator of Eq. (2.45) will force a violation of the preferred spin alignment for a number of nearest neighbor pairs of order A. Hence for $\beta \gg 1$ the numerator is smaller than the denominator by a factor of order $e^{-2\beta A}$. So, in the low-temperature Ising phase, $\beta \gg 1$ (which is the high-temperature gauge theory phase $\tilde\beta \ll 1$), Γ_c falls like the exponential of the area enclosed by c. [This behavior is often taken as a signal for quark confinement in lattice gauge theories based on a non-Abelian group which are thought to represent QCD (Wilson, 1974). However, the criterion can only be taken literally in the absence of quark fields as dynamical variables in the functional integral. See the section on the Abelian Higgs model for a clarificatory discussion.]

The behavior of Γ_c in the high-temperature Ising phase ($\beta \ll 1$) could be obtained from the representation (2.40) or the representation (2.45) by expanding in powers of β. The expectation is (Wegner, 1971; Balian *et al.*, 1975) that $\Gamma_c \sim e^{-gP}$, where P is the perimeter of the curve c and g is a function of β. This result can be proven for sufficiently large β (Fontaine and Gruber, 1978; Gallavotti *et al.*, 1978). Furthermore, it is possible to show that for $\beta < \beta_{\text{critical}}$, Γ_c does not decrease as fast as an area law (i.e., the coefficient of the area in the exponent is zero) (Bricmont *et al.*, 1979). Thus the phase transition in the $d=3$ Z_2 lattice gauge theory should be reflected in the qualitatively different asymptotic behavior of Γ_c in the two phases.

C. Duality for the four-dimensional Ising model

We now want to briefly describe the duality transformation for the Ising model in four dimensions. This is a very useful exercise to do before discussing, in the next section, the general pattern which duality transformations follow.

Starting with the partition function of the $d=4$ Ising model

$$Z = \sum_{\{s\}} \exp\Big(\beta \sum_{\langle\,\rangle} s_i s_j\Big)\,, \tag{2.46}$$

we rewrite following the pattern developed in previous sections:

$$\begin{aligned} Z &= \sum_{\{s\}} \prod_l e^{\beta s_i s_j} = \sum_{\{s\}}\prod_l \sum_k C_k(\beta)(s_i s_j)^k \\ &= \sum_{\{k\}}\prod_l C_k(\beta) \sum_{\{s\}}\prod_i (s_i)^{\Sigma_i k} \\ &= \sum_{\{k\}}\prod_l C_k(\beta) \prod_i 2\delta_2\big(\textstyle\sum_i k\big)\,. \end{aligned} \tag{2.47}$$

As before, we seek a representation for the k_ν's which automatically satisfy the δ functions in Eq. (2.47). To do this we construct the dual lattice D in four dimensions, obtained from the original lattice O by shifting the lattice by half a lattice spacing in each direction. Applying a slight strain on the imagination, we can see that each link of the original lattice penetrates one elementary three-dimensional cube of the dual lattice. Moreover this elementary cube is oriented orthogonally to the link which penetrates it. (Recall, for clarity, the situation in two and three dimensions.) Let us associate a spin $L_{\mu\nu;i}$, which takes on values ± 1 with each plaquette of the dual lattice. Then we choose to represent the k_μ's in the form

$$\begin{aligned} k_{\mu;i} &= \tfrac{1}{2}(1 - L_{\nu\lambda;i} L_{\nu\sigma;i} L_{\lambda\sigma;i} L_{\nu\lambda;i+\hat\sigma} L_{\nu\sigma;i+\hat\lambda} L_{\lambda\sigma;i+\hat\nu}) \\ &\equiv \tfrac{1}{2}(1 - R_{\nu\lambda\sigma;i})\,, \quad \nu,\lambda,\sigma \neq \mu\,, \end{aligned} \tag{2.48}$$

where the six L's that appear in Eq. (2.48) are associated with the six faces of the elementary cube on D, penetrated by the link on O with which $k_{\mu;i}$ is associated. Recall also that the orientation and site indices of k refer to the lattice O, while those on L and R refer to the dual lattice D. It is now straightforward to use Eq. (2.48) in the expression $\sum_i k$ to see that the sum of these eight k's is indeed even. As in two and three dimensions, Eq. (2.48) is, up to the sign in front of the L's, both necessary and sufficient to satisfy the δ functions in Eq. (2.47). Using Eq. (2.48) in Eq. (2.47) we readily find

$$Z = 2^{-N}(\sinh 2\tilde\beta)^{-2N} \sum_{\{L\}}{}' \exp\Big(\tilde\beta \sum_{c_d} R_{\nu\lambda\sigma;i}(L)\Big)\,, \tag{2.49}$$

where $R(L)$ is defined in Eq. (2.48). As in three dimensions this dual theory has a local gauge symmetry, and the prime on the sum tells us that we must restrict oursleves to summing only over a subset of the possible configuration of L's such that each configuration summed over produces a distinct configuration of R's [or, equivalently, of the k's through Eq. (2.48)]. As in three dimensions, it will usually be possible to neglect this prime and do an unrestricted sum over all configurations of L's, so long as we remember that in so doing we have (uniformly) overcounted.

Let us now examine the gauge symmetry of this theory. From Eq. (2.48) we can see that $\{R\}$ is invariant under the following operation: choose a link of the dual lattice D. Attached to this link, there are, in four dimensions, six plaquettes. We now reverse the sign of each of the $L_{\mu\nu}$'s associated with these six plaquettes. By Eq. (2.48) we see that this operation changes the signs of exactly two of the L's in the expression for each R which has the chosen link as one of its edges. Thus the R's remain unchanged. This gauge symmetry is of a "higher" kind than the one present in the three-dimensional gauge theory of the previous subsection. If we write, in analogy with Eq. (2.24),

$$L_{\mu\nu;\,i} = e^{iQ_{\mu\nu;\,i}}, \tag{2.50}$$

with $Q_{\mu\nu;\,i} = 0, \pi$, then the gauge symmetry just discussed can be described by saying that the exponent of Eq. (2.49) (the statistical mechanics Hamiltonian or the field theory Lagrangian) is invariant under the operation[3]

$$Q_{\mu\nu;\,i} \to Q'_{\mu\nu;\,i} = Q_{\mu\nu;\,i} + \Delta_\mu \Lambda_{\nu;\,i}. \tag{2.51}$$

Thus the arbitrary gauge function $\Lambda_{\nu;\,i}$ is in this theory a vector field while in the gauge theory which is dual to the three-dimensional Ising model it is only a scalar [see Eq. (2.25)]. If, for comparison, we consider the gauge theory of the last subsection in four dimensions then we see the theory of Eq. (2.49) has many more (four times as many) gauge degrees of freedom as does Eq. (2.26).

We close this section with two comments. First, we note that the gauge-invariant correlation functions of the variables Q in Eq. (2.49) are defined on two-dimensional surfaces embedded in the four-dimensional space of the theory. In the absence of other fields coupled to Q, these must be closed surfaces. This is just the extension of what occurred in the gauge theory studied in the last subsection in which the gauge-invariant correlation functions (the Wilson loop integral) were defined on closed one-dimensional surfaces (closed loops). Finally, we remark that the theory defined in Eq. (2.49) has a gauge symmetry (2.51) similar to that of a theory studied in a different context by Kalb and Ramond (Kalb and Ramond, 1974). In the Kalb-Ramond theory the symmetry group is $U(1)$ rather than Z_2 in our case. In their study, Kalb and Ramond found that the spectrum of their theory was related to that of a theory with one scalar boson field. We can understand this result from our point of view by remembering that the Ising model can be regarded as a lattice analog of a relativistic field theory of a single scalar field. Since Eq. (2.49) and the Ising model in four dimensions are dual to each other [it should be clear that applying the duality transformation to Eq. (2.49) we get back the Ising model] their spectra are certainly related. In Sec. IV we will discuss theories with a $U(1)$ symmetry, one of which is the $U(1)$ analog of the Z_2 theory described here, and is very closely related to the Kalb-Ramond theory.

[3]A careful analysis of these theories indicates that the fields (or spins) should be regarded as antisymmetric in their direction indices. Thus the gauge transformation should also be antisymmetrized. We shall not explicitly concern ourselves with this point here.

D. Fundamental structure of the duality transformation

In the preceding sections we have begun to see a pattern emerge in the application of the duality transformation to Ising models in varying dimension. Here we will describe and generalize this pattern. Our discussion will be in the context of theories with a Z_2 symmetry, but the basic geometrical content of the result will persist when we describe theories with other symmetry groups.

To begin, we introduce the notion of a simplex. For our purposes, we take a simplex of dimension s to be an s-dimension element of our hypercubical lattice. Thus a simplex of dimension zero is a vertex of the lattice, a simplex of dimension one is a link joining two nearest neighbor vertices on the lattice, a simplex of dimension 2 is an elementary face, or plaquette of the lattice, a simplex of dimension 3 an elementary cube, and so forth. Obviously, in a d-dimensional lattice there are simplices of dimension $s \leq d$.

Let us next construct the lattice dual to our d-dimensional hypercubical lattice, which, we recall, can be done by shifting the lattice by half a lattice spacing in each direction. With a little thought, we can convince ourselves that each simplex of dimension s, of the original lattice, 0, is intersected by one simplex of dimension $\bar{s} = d - s$, of the dual lattice D. For example, in three dimensions, each link of O is bisected by a plaquette of D. Furthermore, each vertex of D ($\bar{s} = 0$) lies at the center of an elementary cube of O ($s = 3$), and vice versa.

Now, we can define a hierarchy of theories with a Z_2 symmetry, of which the Ising model is the simplest member. A theory in this heirarchy can be labeled by a simplex number s. The theory with simplex number s has a variable, or spin, which can take on values ± 1 associated with each simplex of dimension $s-1$ in the lattice. The variable, $Q_{\alpha_1 \cdots \alpha_{s-1};\,i}$, has $s-1$ direction labels since it takes that many directions to define the orientation of an $(s-1)$-dimensional simplex. Now each simplex of dimension s is bordered by $2s$ simplices of dimension $s-1$. The Hamiltonian (or Lagrangian, in field theory language) is defined by multiplying together the $2s$ variables Q, which border each simplex of dimension, s, and then summing over all s-dimensional simplices of the lattice. Thus the Ising model is the $s=1$ theory, the lattice gauge theory discussed in subsection B has simplex number $s=2$, the dual theory of subsection C has $s=3$, etc. The theories with $s \geq 2$ have a local Z_2 gauge symmetry as discussed for the dual theories in the last two subsections. When $s=1$ (Ising model) the Z_2 symmetry is global. As s increases the local gauge symmetry becomes increasingly directional in that the gauge functions have $s-2$ directional indices. Thus, for a given dimension d, the number of dynamically independent variables decreases as s increases. From simple geometrical considerations it is clear that in d dimensions we can only have theories with $s \leq d$; and, in fact, one can easily use the gauge symmetry to show that the theory with $s=d$ is trivial in that there are no dynamical degrees of freedom, and the partition function is just a simple product of numerical (β-dependent) factors. A simple example is the case $s=d=2$. The lack of dynamics in this case is just the analog in our Z_2 symmetric gauge theory of the well known result that there are no real photons in two-dimensional QED. Finally we observe that in the theory with simplex number s, the "natural" gauge invariant correlation function is defined on a closed surface of dimension $s-1$, as we found for the $s=2$ and 3 theories earlier.

From our treatment of the two-, and three- and four-dimensional Ising models the following result should be clear: the theory with simplex number s in d dimensions at an inverse temperature β is transformed by a duality transformation into a theory in d dimensions with simplex number $\tilde{s}=d-s$ at an inverse temperature $\tilde{\beta}=-\frac{1}{2}\ln\tanh\beta$. We will go through a proof of this result in Sec. IV in the context of $U(1)$ invariant theories. It will be clear at that time that the proof also applies to the special cases of Z_N symmetric theories. For the moment we just want to note a few of the implications of this result.

First, the theories with $s=\frac{1}{2}d$ are self-dual. The most important examples of these are the two-dimensional Ising model and the four-dimensional Z_2 gauge theory ($s=2$). As we discussed for the $d=2$ Ising model, if these theories have phase transitions the critical points must all map into themselves under the transformation $\beta\rightarrow\tilde{\beta}$. Next we note that the critical behavior of a theory must be reflected also in its dual. We discussed an example of this for the $d=3$ Ising model and its dual theory. In general, we might expect a phase transition in the simplex number s theory to be signaled by different asymptotic behavior of the gauge-invariant correlation function of that theory above and below the critical point (if there is only one). The kind of asymptotic behavior that is natural to expect is that $\Gamma\sim e^{-V}$ above the transition point and $\Gamma\sim e^{-B}$ below the transition temperature, where Γ is a correlation function of fields on simplices of dimension $s-1$, defined in analogy with Eq. (2.40) and the discussion following Eq. (2.51). V is the (minimal) s-dimensional volume enclosed by the fields defined in Γ, and B is the $(s-1)$-dimensional surface area enclosing V. (Note that the $s=1$ theory also obeys this pattern if we recall that the boundary B in that case consists of two points and does not grow as the separation between spins is increased. Thus $\langle s_i s_j\rangle$ need not approach zero asymptotically below T_c.) Finally, we refer to the comment made earlier that the theory with $s=d$ is trivial. This can be seen quite easily by performing the duality transformation on these theories. The result is

$$Z\sim\sum_{\{s_i=\pm1\}}\exp\Big(\tilde{\beta}\sum_i s_i\Big). \tag{2.52}$$

The s_i's are just spins on the sites of the dual lattice and Z is a partition function for a set of uncoupled Ising spins in an external magnetic field $\propto\tilde{\beta}$. Note that this theory has "interactions" defined only on the sites of the lattice, and is therefore consistent with the simplex number assignment $\tilde{s}=d-s=0$. Furthermore, there is no symmetry associated with Eq. (2.52) (for fixed $\tilde{\beta}$); not even a global Z_2 symmetry. This is what we expect, since decreasing simplex number means decreasing symmetry.

III. Z_N SYMMETRIC MODELS

A. Introduction

Here we will study duality for theories with a Z_N symmetry. There are several reasons to be interested in such theories. First, they are the simplest generalizations of the Z_2 symmetric theories discussed in the last section. Since they have a richer set of possible interactions, these theories provide us with some perspective on the structure of the Z_2 theories—in particular, a study of these theories will elucidate which of the properties of duality transformations described in the last section are peculiar to Z_2 symmetric theories, and which follow from the intrinsic geometrical structure associated with duality. Next, these theories are of great interest in the context of condensed matter physics. The interactions present in a regular lattice of atoms are usually symmetric though a finite rotation which brings the lattice back into itself. In two dimensions this is just a Z_N symmetry for some value of N. Among the interesting systems where such forces occur one may mention thin films of He^4 on graphite, and the general problem of melting in two dimensions [see, for example, Halperin and Nelson, (1978) and José *et al.* (1978]. Finally, gauge theories with a local Z_N symmetry are of interest in the problem of quark confinement. The reason is that the center of the group SU(N), which is Z_N, may be of particular importance in determining whether an SU(N) gauge theory is confining or not ('t Hooft, 1978, 1979).

Our plan in this section will be to examine in some detail the duality transformation for Z_N symmetric spin models in two dimensions. In Sec. III.B we will look first at the vector Potts model and a generalization thereof which includes other types of Z_N symmetric nearest neighbor interactions. Then we will study another model, the "Z_N-Villain model," which is self-dual. This model is in the same universality class as the vector Potts model (at least for $d\geq 2$) but is much easier to analyze. After completing these studies, our results can easily be extended to other Z_N models with global and local symmetries in all dimensions. This will be done in Sec. III.C.

B. Two-dimensional Z_N spin systems

1. Vector Potts model

Consider a two-dimensional square lattice with a complex phase, or spin $e^{i(2\pi/N)q}$ on each site of the lattice. q can take on integer values $0,1,2,\dots,N-1$. The vector Potts model is defined by the Hamiltonian (for a field theory, the Lagrangian)

$$\beta H=\beta\sum_{\langle\,\rangle}\cos\Big(\frac{2\pi}{N}\,\Delta_\mu q_i\Big), \tag{3.1}$$

where Δ_μ is a discrete difference operator

$$\Delta_\mu q_i\equiv q_i-q_{i-\hat{\mu}} \tag{3.2}$$

and the sum is over all nearest-neighbor pairs of the lattice. The partition function for this theory is a sum over all configurations of the system

$$Z=\sum_{\{q=0\}}^{N-1}\exp\left[\beta\sum_{\langle\,\rangle}\cos\left(\frac{2\pi}{N}\Delta_\mu q_i\right)\right]. \tag{3.3}$$

When $N=2$, Eq. (3.3) is just the partition function of the Ising model, Eq. (2.2), and when $N\to\infty$ Eq. (3.3) can be rewritten as the partition function of the classical x-y model (see Sec. IV). It is also interesting to note that when $N=4$ the partition function (3.3) simply reduces to the product of two mutually noninteracting Ising models. For most $N>4$, however, such wonderful simplifications do not generally occur.

It is clear that for $N\geq 4$ there are other Z_N symmetric nearest neighbor interactions which can be written in addition to the one appearing in Eq. (3.1). These are just of the form

$$H_p=\cos[(2\pi p/N)\Delta_\mu q_i], \tag{3.4}$$

where p is an integer $1\leq p\leq\frac{1}{2}(N-2)$. These are all the simple nearest-neighbor Z_N symmetric interactions one can write down without introducing additional fields. For $N=3$ there is only the one term with $p=1$ (because a rotation by $\frac{4}{3}\pi$ is equivalent to a rotation by $-\frac{2}{3}\pi$.)

There are several other models of common interest, besides Eq. (3.3) whose Hamiltonians can be expressed as a linear combination of the H_p. The most famous of these is the Potts model (Potts, 1952) (as distinct from the vector Potts model) in which the energy associated with nearest neighbor spins is, say, E_1 if the spins are aligned (i.e., if $\Delta_\mu q_i=0$), and E_2 if the spins are not aligned ($\Delta_\mu q_i\neq 0$).

Another model, whose Z_N symmetry is expressed by the introduction of a set of auxiliary fields in the "Z_N-Villain model" [Casher, 1978; Elitzur *et al.*, 1979; Cardy, 1978; Einhorn, Savit, and Rabinovici, 1979; Ukawa *et al.*, 1979). Its partition function has the form

$$Z=\sum_{\{q=0\}}^{N-1}\sum_{\{j_\mu=-\infty\}}^{\infty}\exp\left[\frac{-\beta}{2}\sum_{\langle\,\rangle}\left(\frac{2\pi}{N}\Delta_\mu q_i-2\pi j_{\mu;i}\right)^2\right], \tag{3.5}$$

where there is one integer $j_{\mu;i}$ associated with each link of the lattice, and the sum runs over all $j_{\mu;i}$ from $-\infty$ to ∞. Because this model involves an additional set of fields, its Hamiltonian can be expressed as a function of the H_p in Eq. (3.4) only after summing over the auxiliary fields. But then the coefficients of the H_p may be very complicated functions of β. We shall return to this model later.

Let us now turn our attention back to the vector Potts model (3.3), and derive its dual form. First we note that the argument of Eq. (3.3) can be expanded in a Fourier series for the group Z_N:

$$\exp\left[\beta\cos\left(\frac{2\pi}{N}\Delta_\mu q\right)\right]=\sum_{k=0}^{N-1}C_k(\beta)\exp\left(i\frac{2\pi}{N}k\Delta_\mu q\right), \tag{3.6}$$

where the $C_k(\beta)$ are determined by inverting Eq. (3.6) and using the completeness of the expansion, in the usual way for Fourier transforms. Thus the $C_k(\beta)$ have the form

$$C_k(\beta)\sim\sum_{L=0}^{N-1}\exp\left[\beta\cos\left(\frac{2\pi}{N}L\right)+i\frac{2\pi}{N}Lk\right], \tag{3.7}$$

which means that $\ln C_k(\beta)$ can be written

$$\ln C_k(\beta)=G_0(\beta)+G_1(\beta)\cos\frac{2\pi}{N}k+G_2(\beta)\cos^2\left(\frac{2\pi}{N}k\right)+\cdots+\begin{cases}G_{N/2}\cos^{N/2}[(2\pi/N)k], & N\text{ even},\\ G_{(N-1)/2}\cos^{(N-1)/2}[(2\pi/N)k], & N\text{ odd}.\end{cases} \tag{3.8}$$

We can now use these expressions in Eq. (3.3). We have

$$Z=\sum_{\{q\}}\exp\left[\beta\sum_{\langle\,\rangle}\cos\left(\frac{2\pi}{N}\Delta_\mu q_i\right)\right]$$

$$=\sum_{\{q\}}\prod_{\langle\,\rangle}\exp\left[\beta\cos\left(\frac{2\pi}{N}\Delta_\mu q_i\right)\right] \tag{3.9a}$$

$$=\sum_{\{q\}}\sum_{\{k_\mu\}}\prod_{\langle\,\rangle}C_{k_\mu}(\beta)\exp\left(i\frac{2\pi}{N}k_{\mu;i}\Delta_\mu q_i\right) \tag{3.9b}$$

$$=\sum_{\{k\}}\prod_{\langle\,\rangle}[C_{k_\mu}(\beta)]\sum_{\{q\}}\exp\left(-i\frac{2\pi}{N}\sum q_i\Delta_\mu k_{\mu;i}\right) \tag{3.9c}$$

$$=\sum_{\{k\}}\prod_{\langle\,\rangle}[C_{k_\mu}(\beta)]\prod_i N\delta_N(\Delta_\mu k_{\mu;i}). \tag{3.9d}$$

The form (3.9c) was obtained from Eq. (3.9b) by performing a summation by parts in the exponent. We assume spherical boundary conditions, so there is no surface term. In Eq. (3.9d), δ_N is a Kronecka δ function mod N; i.e., it restricts the argument to be N times an integer.

We now seek a representation for the $k_{\mu;i}$ which automatically satisfies the δ function in Eq. (3.9d). Since we do not insist that $\Delta_\mu k_{\mu;i}=0$, one might think that the required representations for the $k_{\mu;i}$ would be rather messy. However, we can define another set of variables $\bar{k}_{\mu;i}$, which are equal to the k_μ's mod N, and which satisfy $\Delta_\mu\bar{k}_{\mu;i}=0$. From Eq. (3.8), we see that changing $k_{\mu;i}$ by an integer multiple of N will leave the C_k's, and hence Z, invariant. Thus, we will be able to replace the k_μ's by the $\bar{k}_\mu$'s with impunity.

To define the $\bar{k}_\mu$'s we construct in the usual way the dual lattice D from an original lattice O (see Fig. 2). With each site of the dual lattice we associate a variable, ϕ_i, that can take on values $0,1,\ldots,N-1$. We then define

$$\bar{k}_{\mu;i}=\varepsilon_{\mu\nu}\Delta_\nu\phi_i. \tag{3.10}$$

The indices on $\bar{k}$ refer to lattice O, while those on the right-hand side of Eq. (3.10) refer to lattice D. Since Eq. (3.10) is of the form of a curl it follows (iff) that $\Delta_\mu\bar{k}_{\mu;i}=0$. Now, $-(N-1)\leq\bar{k}\leq N-1$. It is easy, however, to convince oneself that there is a one to one correspondence between sets of $\bar{k}$ satisfying $\Delta_\mu\bar{k}_{\mu;i}=0$ and sets of $k_{\mu;i}$ satisfying the δ functions in Eq. (3.9d). Consider, for instance, the example shown in Fig. 9 for $N=7$. In Fig. 9(a) we have a distribution of k_μ's which satisfy $\Delta_\mu k_\mu=-7$ at the depicted lattice site. In Fig. 9(b) we show the distribution of $\bar{k}_\mu$'s which satisfy $\Delta_\mu\bar{k}_\mu=0$ and which are equal mod 7 to the k's in Fig. 9(a). In Fig. 9(b) we also show the values of ϕ_i which generate the required $\bar{k}_\mu$'s via Eq. (3.10). (The ϕ_i's, of course, are defined only up to an overall constant.)

We can now substitute the $\bar{k}_\mu$'s defined through Eq.

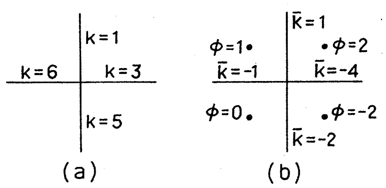

FIG. 9. Example of the change of variables $k_\mu \to \bar{k}_\mu$ for the case $N=7$. In (a) $\Delta_\mu k_\mu = -7$, while in (b) $\Delta_\mu \bar{k}_\mu = 0$. Shown also in (b) are the values of ϕ_i which generate the $\bar{k}_\mu$ according to Eq. (3.10). See the discussion following Eq. (3.9) for more details.

(3.10) for the k_μ's in Eq. (3.9d) and drop the δ functions. Equation (3.9d) then becomes, up to numerical constants,

$$\begin{aligned} Z &= \sum_{\{\phi\}} \prod_{\langle\ \rangle} \exp[\ln C_{\epsilon_{\mu\nu}\Delta_\nu \phi_i}(\beta)] \\ &= \sum_{\{\phi=0\}}^{N-1} \exp\Big[\sum_{l_d}\sum_{t=0}^{t_N} G_t(\beta)\cos^t\Big(\frac{2\pi}{N}\Delta_\mu\phi_i\Big)\Big] \\ &= \sum_{\{\phi\}} \exp\Big(\sum_{l_d}\sum_{t} G_t(\beta)(\mathbf{s}_i\cdot\mathbf{s}_{i+\hat{\mu}})^t\Big), \end{aligned} \tag{3.11}$$

where the first sum in the exponent is a sum over all nearest-neighbor pairs on D, and the sum over t just represents the expansion in Eq. (3.8). t_N is the largest integer $\frac{1}{2}N$ or $\frac{1}{2}(N-1)$. In the last form we have introduced the dual Z_N spin,

$$s_i = \Big[\cos\Big(\frac{2\pi}{N}\phi_i\Big), \sin\Big(\frac{2\pi}{N}\phi_i\Big)\Big].$$

There are several important features of this result. First we note that both the original model and its dual are Z_N symmetric, but the vector Potts model is not self-dual for general N. We can, however, define the more general theory whose Hamiltonian is a linear combination of the interactions of the form (3.4). It should be clear that some of the models so defined will be self-dual. At the very least, the model which includes all terms of the form (3.4) will, for arbitrary values of the coupling constants, generate a dual theory which also contains all terms of the form (3.4). Exceptions to these statements are the cases with $N=3$ and $N=4$ for which the vector Potts model is self-dual. As noted earlier, when $N=3$ there is only one independent Z_3 symmetric interaction, while when $N=4$, the vector Potts model reduces to a simple product of two Ising models.

Next, we note a very interesting feature of the behavior of the $G_t(\beta)$. From Eq. (3.7), we see that when $\beta \gg 1$, we can approximate the sum by

$$\sum_{L=0}^{N-1} \exp\Big(\beta\cos\frac{2\pi}{N}L + i\frac{2\pi}{N}LK\Big) \simeq \exp(\beta) + 2\exp\Big(\beta\cos\frac{2\pi}{N}\Big) \times \cos\frac{2\pi}{N}k + \cdots, \tag{3.12}$$

and so

$$\ln C_k(\beta) \underset{\beta\to\infty}{\simeq} \beta + 2\exp\Big\{\beta\Big[\cos\Big(\frac{2\pi}{N}\Big) - 1\Big]\Big\}\cos\frac{2\pi}{N}k + \cdots. \tag{3.13}$$

Hence, neglecting overall (β-dependent) factors

$$Z \underset{\beta\gg 1}{\simeq} \sum_{\{\phi\}} \exp\Big[G_1(\beta)\sum_{l_d}\cos\Big(\frac{2\pi}{N}\Delta_\mu\phi_i\Big)\Big] \tag{3.14}$$

with $G_1(\beta) \to 0$ as $\beta \to \infty$. Thus, in the very low-temperature limit the vector Potts model is approximately self-dual and, as expected, the dual temperature approaches infinity as the original temperature goes to zero. It is a straightforward and interesting exercise to generalize this observation to theories with other kinds of Z_N symmetric interactions of the form (3.4).

Finally, we note that despite the increased number of independent interactions, the underlying geometrical relation between the dual theory and the original theory still holds. That is, all the interactions in Eq. (3.11) are defined on a simplex of dimension one, and the Z_N symmetry is global. According to the discussion in Sec. II.D this is what we expect, since $\bar{s} = d - s = 1$ in this case. It should be clear that the entire geometric discussion of that section can be carried over to Z_N symmetric theories, as long as we remember the added complexity of several different kinds of interactions. In Sec. III.C we will make this explicit, and we will introduce theories with a local Z_N symmetry.

2. "Z_N-Villain model"

We turn now to the model described by the partition function (3.5). To see that this model has a global Z_N symmetry, we first observe that the Hamiltonian is invariant under the operation $q_i \to q_i + l$, where l is an integer, independent of i. However, to make the Hamiltonian truly Z_N invariant, we must also consider transformations of the form

$$q_i \to q_i + l + NL_i, \tag{3.15}$$

where L_i is an integer which may vary from site to site. This is just the statement that in the Hamiltonian, q_i is only defined mod$N$. [It is clear, for example, that the interactions (3.4) are invariant under this transformation, since the cosine is periodic.] In Eq. (3.5) the $j_{\mu;i}$'s are included to enforce this periodicity. The transformation (3.15) just amounts to a redefinition of the $j_{\mu;i}$'s,

$$j_{\mu;i} \to j_{\mu;i} - \Delta_\mu L_i, \tag{3.16}$$

but since we must sum over all integer $j_{\mu;i}$, the Hamiltonian, considered as a function of $\{q\}$ *after* summing over $j_{\mu;i}$ is obviously unchanged. This can be stated in another way: The exponent in Eq. (3.5) is invariant under the local gauge symmetry operation

$$q_i \to q_i + NL_i, \quad j_{\mu;i} \to j_{\mu;i} - \Delta_\mu L_i. \tag{3.17}$$

As stated above, once the $\{j_\mu\}$ is summed over we can consider the logarithm of the argument of the sum over $\{q\}$ as an (effective) Hamiltonian. This effective Hamiltonian has a true global Z_N symmetry, Eq. (3.15). It is the compactness of the group Z_N that is responsible for the additional formal freedom to add NL_i to any q_i without changing the Hamiltonian. On the other hand, it is precisely this compactness that is responsible for the local symmetry, Eq. (3.17), when Eq. (3.5) is considered as a function of both $\{q_i\}$ and $\{j_{\mu;i}\}$. Thus, a theory with a global compact symmetry may in some sense be regarded as a theory with a local gauge symmetry.[4,5]

There are several ways to approach the duality transformation for Eq. (3.5). We will follow a path which is most similar to the formalism we have encountered in previous theories. We begin by making use of the comment in footnote 5. Being cavalier about overall factors of infinity, we rewrite Eq. (3.5) as

$$Z=\sum_{\{q=-\infty\}}^{\infty}\sum_{\{j_\mu=-\infty\}}^{\infty}\exp\left[-\frac{\beta}{2}\sum_{\langle\,\rangle}\left(\frac{2\pi}{N}\Delta_\mu q_i-2\pi j_{\mu;i}\right)^2\right] \tag{3.18a}$$

$$=\sum_{\{p=-\infty\}}^{\infty}\sum_{\{j_\mu\}}\int_{-\infty}^{\infty}Dq\exp\left[\sum-\frac{\beta}{2}\left(\frac{2\pi}{N}\Delta_\mu q_i-2\pi j_{\mu;i}\right)^2+i2\pi p_i q_i\right] \tag{3.18b}$$

$$=(2\pi\beta)^{-V}\sum_{\{p\},\{j_\mu\}}\int_{-\infty}^{\infty}DqD\tau_\mu\exp\left[\sum-\frac{1}{2\beta}\tau_{\mu;i}^2+i\tau_{\mu;i}\left(\frac{2\pi}{N}\Delta_\mu q_i-2\pi j_{\mu;i}\right)+i2\pi p_i q_i\right]. \tag{3.18c}$$

In Eq. (3.18b) we have used the Poisson summation identity

$$\sum_{k=-\infty}^{\infty}\delta(x-k)=\sum_{p=-\infty}^{\infty}e^{i2\pi px}, \tag{3.19}$$

where x is a continuous variable which in general, ranges over $-\infty<x<\infty$. The form (3.19) is introduced for each q_i, thereby turning the sum over q_i into an integral over continuous q_i times a sum over p_i. In Eq. (3.18c) we have introduced a continuous Fourier conjugate variable $\tau_{\mu;i}$ for each link of the lattice, and, to avoid confusion, we have defined the number of lattice sites to be V. We now rearrange the terms in Eq. (3.18c) and we sum by parts so that $\tau_{\mu;i}\Delta_\mu q_i\to -q_i\Delta_\mu\tau_{\mu;i}$ (assuming spherical boundary conditions). We then have

$$Z=(2\pi\beta)^{-V}\sum_{\{p\}}\int D\tau_\mu\sum_{\{j_\mu\}}\int Dq\exp\left[\sum-\frac{1}{2\beta}\tau_{\mu;i}^2+iq_i\left(2\pi p_i-\frac{2\pi}{N}\Delta_\mu\tau_{\mu;i}\right)-i2\pi j_{\mu;i}\tau_{\mu;i}\right] \tag{3.20a}$$

$$=(2\pi\beta)^{-V}\sum_{\{p\}}\int D\tau_\mu\sum_{\{j_\mu\}}\exp\left(\sum-\frac{1}{2\beta}\tau_{\mu;i}^2-i2\pi\tau_{\mu;i}j_{\mu;i}\right)\prod_i\delta\left(2\pi p_i-\frac{2\pi}{N}\Delta_\mu\tau_{\mu;i}\right) \tag{3.20b}$$

$$=(2\pi\beta)^{-V}\sum_{\{p\}}\sum_{\{\tau_\mu=-\infty\}}^{\infty}\exp\left(\sum-\frac{1}{2\beta}\tau_{\mu;i}^2\right)\prod_i\delta\left(2\pi p_i-\frac{2\pi}{N}\Delta_\mu\tau_{\mu;i}\right). \tag{3.20c}$$

In Eq. (3.20b) we have performed the integral over $\{q_i\}$ to give us a product of Dirac delta functions, one for each site of the lattice. In Eq. (3.20c) we have used the identity (3.19) to observe that the sum over $\{j_{\mu;i}\}$ just reduces the integral over continuous τ_μ to a sum over integer τ_μ. The delta functions in Eq. (3.20c) may thus be regarded as Kronecker delta functions.

We now seek a representation for $\tau_{\mu;i}$ which automatically satisfies the constraints in Eq. (3.20c). The content of these delta functions is that $\tau_{\mu;i}$ must be a curl, mod$N$. This can be expressed by writing

$$\tau_{\mu;i}=\varepsilon_{\mu\nu}\Delta_\nu\phi_i-N\varepsilon_{\mu\nu}r_{\nu;i}. \tag{3.21a}$$

The ϕ_i's are integer-valued fields associated with the sites of the dual lattice, while the $r_{\nu;i}$'s are integer valued fields associated with the links of the dual lattice. Inserting Eq. (3.21a) in the argument of the δ function, we find that

$$p_i=-\varepsilon_{\mu\nu}\Delta_\mu r_{\nu;i}, \tag{3.21b}$$

so that the integer p_i is just the curl of the $r_{\nu;i}$'s which are associated with the links of the dual lattice which surround the original lattice site i. Using Eq. (3.21a) in Eq. (3.20c), we can replace the sums over p and τ_μ by summing over ϕ and r_ν, and we can drop the δ functions. The result is

$$Z=(2\pi\beta)^{-V}\sum_{\{\phi=-\infty\}}^{\infty}\sum_{\{r_\mu=-\infty\}}^{\infty}\times\exp\left[-\frac{N^2}{8\pi^2\beta}\sum_{l_d}\left(\frac{2\pi}{N}\Delta_\mu\phi_i-2\pi r_{\mu;i}\right)^2\right], \tag{3.22}$$

where the sum in the exponent is over all links of the dual lattice. As before, in going from Eq. (3.20c) to Eq. (3.22) we have neglected overall uniform factors of infinity. In particular, both Eqs. (3.18a) and (3.22) include redundancies, not only in the infinite sum over q_i (ϕ_i), but also in the unconstrained sum over $j_{\mu;i}$ ($r_{\mu;i}$). This can be seen easily from Eq. (3.21b). The sum over $r_{\mu;i}$ should not, in principle, include sums over configurations of $r_{\mu;i}$ that differ only by a gradient, since such configuration will produce the same configurations of the p_i. However, the overcounting is uniform, and so produces only a harmless multiplicative infinity. The interested reader can check the

[4]The local symmetry operation (3.17) resembles a local gauge transformation for a system consisting of an Abelian gauge field $J_{\mu;i}$ and a charge field whose "phase" is q_i. In Sec. IV we shall study the Abelian Higgs model after which study the reader will be able to show that Eq. (3.18a) corresponds to a discrete (Z_∞) Abelian Higgs model with a "scalar" field of charge N, in which the coefficient of the discrete gauge field kinetic energy term has been taken to zero.

[5]One might worry that adding NL_i to q_i will take q_i out of its range of definition. This is not really a problem though, since we can formally extend $-\infty<q_i<\infty$ and imagine dividing Z by the infinite number which counts the copies thereby produced.

veracity of these statements by redefining Eq. (3.18a) to keep track of the redundancy and repeating the steps from there to Eq. (3.22) with less abandon than we have done.

Up to an overall β-dependent factor, Eq. (3.22) has the same form as Eq. (3.18a) if we identify the dual temperature as $\frac{1}{2}\tilde{\beta}=N^2/8\pi^2\beta$. Thus the two-dimensional Z_N-Villain model is self-dual and possesses the usual property that high- (low-) temperature regions of the original theory get mapped monotonically into low- (high-) temperature regions of the dual theory. (It is also possible to construct a similar self-dual Z_N-symmetric model on a triangular lattice. See Savit, 1980, for details. For a very useful discussion of the self-duality of the Ising model on a triangular lattice see Wegner, 1973, and the reviews of Syozi, 1972, and Gruber, *et al.*, 1977.)

The model (3.5) has sometimes been regarded as an approximation to the vector Potts model, Eq. (3.3), especially at low temperatures. For large enough β one might expect that important configurations in Eq. (3.3) will be those for which all the $\Delta_\mu q_i=0,\pm 1$ and those in Eq. (3.5) for which all $\Delta_\mu q=0,\pm 1, \bmod N$. Thus, up to overall constants, Eqs. (3.5) and (3.3) could well have the same behavior for large enough β. The critical behavior of Eq. (3.5) has been studied (Elitzur *et al.*, 1979; Cardy, 1978; Einhorn, Savit, and Rabinovici, 1979; Ukawa *et al.*, 1979) and it has been conjectured that Eqs. (3.3) and (3.5) have a similar phase structure. In particular, for large enough N ($N \gtrsim 5$), Eq. (3.5) is expected to have three phases. Since the model is self-dual, the two critical points are related by the duality formula $\beta_1=N^2/4\pi^2\beta_2$. The reader should consult the references above for more details. In the next section we will study the $N\to\infty$ limits of these models which are the classical x-y model and its Villain approximation, respectively, whose critical properties are also thought to resemble each other.

Let us now ask whether the Z_N-Villain model in Eq. (3.5) can be generalized to include other quadratic terms which correspond to the H_p in Eq. (3.4) with $p>1$. Such a theory should have a Hamiltonian with the general form

$$\sum_{\langle\,\rangle}\sum_{p} V_p\left(\frac{2\pi p}{N}\Delta_\mu q_i+2\pi J^{(p)}_{\mu;i}\right)^2. \tag{3.23a}$$

The question is, what are the $J_\mu^{(p)}$'s which enforce the required periodicity? First we recall that all of the terms in Eq. (3.4) are invariant under the local transformation $q_i\to q_i+NL_i$, where L_i is an integer-valued scalar field defined on the lattice sites. Suppose we try to make Eq. (3.23a) invariant under this operation by setting $J_\mu^{(p)}=J_\mu$ for all p (remembering that we must sum over $\{J_\mu\}$ in the partition function). It is clear that this will not work. After transforming $q_i\to q_i+NL_i$, we can shift J_μ to cancel the extra factors of N in only one of the terms in the sum over p in Eq. (3.23a); the other quadratic terms will in general differ, and so Eq. (3.23a) will not have the required periodicity. However, if we define $J_\mu^{(p)}\equiv p j_\mu$, where j_μ is the same for every term in Eq. (3.23a) and takes on all integral values in the sum in the partition function, then it is clear that each term in Eq. (3.23a) will be invariant under the transformation $q_i\to q_i+NL_i$ if we simultaneously transform $j_{\mu;i}\to j_{\mu;i}-\Delta_\mu L_i$. Rewriting Eq. (3.23a), we have

$$\sum_{\langle\,\rangle}\sum_{p} p^2 V_p\left(\frac{2\pi}{N}\Delta_\mu q+2\pi j_\mu\right)^2. \tag{3.23b}$$

This is precisely the same form as the Hamiltonian in Eq. (3.5). Thus a straightforward attempt to generate addition quadratic terms with periodicity associated with a subgroup of Z_N degenerates immediately into the form (3.5). This is related to the fact that Eq. (3.5) [and Eq. (3.23)] are self-dual: The dual form of the vector Potts model involves different terms with periodicity N/p. But for the Z_N-Villain model these terms of would-be different periodicity are just proportional to the original interaction resulting in the model's self-duality.

C. Generalization to locally Z_N symmetric theories and to arbitrary dimension

In both of the globally Z_N symmetric models studied in the last section the dual formulation possessed a global Z_N symmetry with interactions defined on the links of the dual lattice. Thus these two-dimensional spin theories obey the geometrical pattern described in Sec. II.D for Z_2 symmetric theories, namely, that $\tilde{s}=d-s=2-1=1$, in this case. In the case of the vector Potts model, we saw that the duality transformation will induce of order $\frac{1}{2}N$ different (linearly independent) kinds of Z_N symmetric interactions, but all of them are defined on dual links. Here we will show that, as in the case of a Z_2 symmetry, one can define a heirarchy of locally Z_N invariant theories labeled by a simplex number s. These theories and their duals obey the general pattern described in Sec. II.D, namely, $\tilde{s}=d-s$. The only difference here is that for each value of s, there are of order $\frac{1}{2}N$ types of interactions; a theory with simplex number s, which contains some or all of these $\frac{1}{2}N$ interactions will, in general, have a dual form which may contain all of the $\frac{1}{2}N$ different interactions, although each one will have simplex number $d-s$. An exception to this are the Z_N-Villain models which, because of their quadratic nature, have dual forms which contain only quadratic Villain-type interactions, as in the theory discussed in Sec. III.B.2.

We turn now to a definition of the Z_N symmetric theories with $s>1$. In the next few paragraphs we shall also state some of the results of the duality transformation. The transformation itself is a straightforward generalization of what was done for the two-dimensional Z_N spin theories, following the pattern developed for the Z_2 symmetric theories discussed in the last section.[6]

We first discuss generalizations of theories with cosine interactions (e.g., the vector Potts model). To generate an interaction with simplex number s, we

[6]Any questions in regard to these matters can be addressed to the author. Please include a stamped self-addressed envelope and $2 cash, check, or money order to cover the cost of handling. Upon request responses will be sent in a plain brown wrapper. No C.O.D.'s.

associate with each simplex of dimension $s-1$ (in a d-dimensional lattice $d \geq s$) a Z_N field:

$$Q_{\mu_1,\mu_2\cdots\mu_{s-1};i} = \exp[i(2\pi/N)q_{\mu_1\cdots\mu_{s-1};i}], \qquad (3.24)$$

with q an integer, $0 \leq q \leq N-1$. As in the Z_2 case, we take a product of the $2s$ factors, Q, which lie on the $2s$ simplices of dimension $s-1$ which bound a simplex of dimension s. Since Eq. (3.24) is complex, we can choose some of the Q's in the product to be complex conjugated. Considering the boundary of the simplex to be an oriented surface one can give a geometric description of the rule dictating which Q's should be complex conjugated. It is much easier, however, to just write the general result. The interaction we seek is just the real part of this product of $2s$ Q's and is

$$I^{(1)}_{\mu_1\cdots\mu_s;i} = \cos\left(\frac{2\pi}{N(d-s)!}\varepsilon_{\mu_1\cdots\mu_s\alpha_1\cdots\alpha_{d-s}}\varepsilon_{\alpha_1\cdots\alpha_{d-s},\beta,\gamma_1\cdots\gamma_{s-1}}\Delta_\beta q_{\gamma_1\cdots\gamma_{s-1}}\right) \equiv \cos\left(\frac{2\pi}{N}\varepsilon\varepsilon\Delta q\right). \qquad (3.25)$$

As pointed out earlier, this is not the only Z_N symmetric interaction of simplex number s which we can write. The others, which are simple generalizations of the terms (3.4), are

$$I^{(p)} = \cos[(2\pi p/N)\varepsilon\varepsilon\Delta q],$$
$$p = 1, 2, \ldots, \tfrac{1}{2}N \text{ or } \tfrac{1}{2}(N-1). \qquad (3.26)$$

Now we can construct a locally (for $s>1$) Z_N symmetric theory by summing an arbitrary linear combination of the terms (3.26) over all s-dimensional simplices of the d-dimensional lattice, and then summing the exponential of this sum over different configurations of $\{q\}$. The partition function is

$$Z = \sum_{\{q=0\}}^{N-1}{}' \exp\left[\sum_s\sum_p A_p \cos\left(\frac{2\pi p}{N}\varepsilon\varepsilon\Delta q\right)\right]. \qquad (3.27)$$

Here the coefficients A_p are in general arbitrary.[7] But if Z describes a statistical system in the usual way, then the exponent in Eq. (3.27) should be $-\beta H$, and all the A_p should be proportional to β. In analogy to the Z_2 symmetric case, the exponent in Eq. (3.27) is invariant under a local gauge transformation of the form[3]

$$q_{\mu_1\cdots\mu_{s-1};i} \to q'_{\mu_1\cdots\mu_{s-1};i}$$
$$= q_{\mu_1\cdots\mu_{s-1};i} + \Delta_{\mu_1}L_{\mu_2\cdots\mu_{s-1};i}, \qquad (3.28)$$

where L is an integer-valued mod$N$ gauge function associated with the simplices of dimension $s-2$ of the lattice. As before, the prime on the sum over $\{q\}$ in Eq. (3.27) tells us that in principle we should sum only over those sets of q's such that each set $\{q\}$ gives a distinct configuration of $\varepsilon\varepsilon\Delta q$. That is, we should not sum over gauge copies. But, as before, we can usually ignore the prime so long as we remember that in so doing we uniformly overcount distinct configurations of $\varepsilon\varepsilon\Delta q$.

The duality transformation for Eq. (3.27) proceeds analogously to those we have already studied. One writes the sum over simplices in the exponent as a product over simplices of the exponentials. Performing a Fourier expansion for each simplex one expands the exponential of the interaction in a series of the form

$$\sum_{k=0}^{N-1} C_k \exp\left(ik\frac{2\pi}{N}\varepsilon\varepsilon\Delta q\right), \qquad (3.29)$$

where k has s direction indices. Summing over the q's, one obtains a set of constraints on the k's which are satisfied identically, if and only if the k's are parametrized in a specific form in terms of some set of integers ϕ, which are naturally associated with simplices of dimensions $d-s-1$ on the dual lattice. Inserting this parametrization in the partition function, one arrives at the dual form of Eq. (3.27) which is (up to overall field independent factors)

$$Z \simeq \sum_{\{\phi=0\}}^{N-1}{}' \exp\left[\sum_{\tilde{s}}\sum_p B_p \cos\left(\frac{2\pi p}{N}\varepsilon\varepsilon\Delta\phi\right)\right], \qquad (3.30)$$

where

$$\varepsilon\varepsilon\Delta\varphi \equiv \varepsilon_{\mu_1\cdots\mu_{d-s}\alpha_1\cdots\alpha_s}\varepsilon_{\alpha_1\cdots\alpha_s,\beta,\gamma_1\cdots\gamma_{d-s-1}} \times \Delta_\beta\phi_{\gamma_1\cdots\gamma_{d-s-1};i}. \qquad (3.31)$$

In Eq. (3.30), the sum over p is from 1 to $\frac{1}{2}N$ or $\frac{1}{2}(N-1)$ depending on whether N is even or odd. The sum over $\tilde{s}$ is a sum over all dual lattice simplices of dimension $\tilde{s} = d-s$. The prime on the $\{\phi\}$ sum indicates the necessity (in principle) to make a gauge choice (unless $\tilde{s}=0,1$) since the exponent in Eq. (3.30) (and the representation for the k's in terms of the ϕ's) is invariant under the local gauge transformation[3]

$$\phi_{\gamma_1\cdots\gamma_{d-s-1};i} \to \phi_{\gamma_1\cdots\gamma_{d-s-1};i} + \Delta_{\gamma_1}G_{\gamma_2\cdots\gamma_{d-s-1};i},$$

where G is an integer-valued mod$N$ field associated with the dual lattice simplices of dimension $d-s-2$. Notice that the geometrical relation between Eqs. (3.27) and (3.30) is precisely the same as that discovered for the Z_2 symmetric theories in the last section. The only difference is that one has several different kinds of interactions over which to sum.

The coefficients B_p in Eq. (3.30) are functions of the A_p in Eq. (3.27). Since we have come to expect duality transformation to map high-temperature theories into low-temperature theories and vice versa, it is interesting to ask how the B_p's vary when the A_p's are very large. Suppose that $A_p = \beta a_p$, and consider the limit $\beta \to \infty$. Examining the Fourier expansion of Eq. (3.27), one sees that in general in this limit, all the B_p's tend to zero like $e^{-c\beta}$, where c is some function of the a_p. This is the expected behavior for the dual theory.

We will describe now generalizations of the Z_N-Villain model discussed earlier. To define the (locally invariant) interaction on a simplex of dimension s, we again define fields on simplices of dimension $s-1$ as in Eq. (3.28). In addition, we define a field $J_{\mu_1\cdots\mu_s;i}$ associated with each simplex of dimension s, which can take on all integer values from $-\infty$ to $+\infty$. The Villain partition function is then

[7]Note, though, that if the only nonzero A_p's are those for which $\{N/p\}$ is a set of commensurate integers, the theory no longer evidences a Z_N symmetry, but only a Z_M symmetry where M is the largest integer N/p.

$$Z=\sum_{\{q=0\}}^{N-1}{}'\sum_{\{J=-\infty\}}^{\infty}{}'\exp\left[-\frac{\beta}{2}\sum_{s}\left(\frac{2\pi}{N}\varepsilon\varepsilon\Delta q+2\pi J\right)^{2}\right]. \quad (3.33)$$

As before the sum is over all simplices of dimension s. The prime on the sum over $\{q\}$ indicates the existence of the local gauge symmetry of the form (3.28), which leaves $\varepsilon\varepsilon\Delta q$ invariant. The prime on the sum over J indicates the (usually harmless) redundancy involved in summing over all J, as explained in the last section. This redundancy just reflects the local symmetry of the exponent in Eq. (3.33) with respect to the operation in which $\varepsilon\varepsilon\Delta q$ changes by N, and the corresponding J changes by one. Note that this symmetry is of a higher kind (i.e., more directional) than the one in Eq. (3.28), since the arbitrary gauge function has $s-1$ rather than $s-2$ direction indices. As we discussed for the globally Z_N symmetric models this higher gauge symmetry in Eq. (3.33) is necessary to ensure that Eq. (3.33), considered as a function of $\{q\}$, is Z_N symmetric with the local gauge symmetry of Eq. (3.28).

The duality transformation for Eq. (3.33) is just a straightforward generalization of what was done for Eq. (3.18a). The result is that, up to overall field-independent factors,

$$Z\simeq\sum_{\{\phi=0\}}^{N-1}{}'\sum_{\{r=-\infty\}}^{\infty}{}'\exp\left[\frac{-N^{2}}{8\pi^{2}\beta}\sum_{\tilde{s}}\left(\frac{2\pi}{N}\varepsilon\varepsilon\Delta\phi+2\pi r\right)^{2}\right]. \quad (3.34)$$

The sum in the exponent runs over all dual lattice simplices of dimension $\tilde{s}=d-s$. The symbol $\varepsilon\varepsilon\Delta\phi$ is that defined in Eq. (3.31) and $r=r_{\mu_1\cdots\mu_{d-s};i}$ is an integer-valued field associated with each simplex of dimension $d-s$ on the dual lattice. The primes on the ϕ and r sums indicate the existence of gauge symmetries in Eq. (3.34), analogous to those in Eq. (3.33), only here the symmetries are those associated with the simplex number $\tilde{s}=d-s$ interactions. We note that the dual form (3.34) of the Z_N-Villain theories (3.33) are again Z_N-Villain theories, and that in the special cases when $s=\frac{1}{2}d$ (e.g., the $d=2$ spin system or the $d=4$ gauge theory) the theories are self-dual. This is precisely the situation encountered for the Z_2 symmetric theories. The dual inverse temperature, however, is a different function of the original temperature, and depends on the group.

IV. $U(1)$ SYMMETRIC THEORIES

A. Introduction

Here we will discuss duality transformations for theories with an internal $U(1)$ symmetry. The simplest nontrivial model with such a symmetry is the two-dimensional x-y model. This is the (appropriately defined) $N\to\infty$ limit of the two-dimensional vector Potts model described in Sec. III. In Sec. III.B we will take up this model and show that it is dual in a certain approximation to the two-dimensional discrete Gaussian model. We will then show that the model can also be written as a two-dimensional Coulomb gas. With the help of the Villain approximation introduced in the last section, we will demonstrate that the charges in this Coulomb gas can be understood as topological excitations, or vortex penetrations of the original x-y model spins. Thus we will demonstrate that for this model, the topological excitations (plus spin waves) form a complete set of variables for a description of the theory. This remarkable property will obtain in all the theories discussed in this section.

In Sec. III.C we will construct the duality transformation for the general $U(1)$ symmetric lattice theory with simplex number s. We will show that such a theory is dual to a theory with simplex number $\tilde{s}=d-s$ and possessing an internal symmetry Z_∞, the additive group of integers. Moreover, we will show how the theory can be expressed directly in terms of its spin waves and topological excitations, and we will show that those excitations have dimension $d-s-1$. Finally, in Sec. III.D we will describe in more detail three $U(1)$ theories of physical interest: the three-dimensional x-y model, and the three- and four-dimensional $U(1)$ lattice gauge theories ($s=2$). In particular we will show how the formulation of these theories in terms of their topological excitations can lead to a simple and illuminating picture of their phase properties.

B. Two-dimensional x-y model

1. Duality transformation

This model is the $N\to\infty$ limit of the model described in Eq. (3.3). Consider a square two-dimensional lattice, and on each site of the lattice place a spin, $s_j=e^{i\theta_j}$, $-\pi<\theta\leq\pi$. The spins interact through nearest-neighbor couplings as in Eq. (3.3). The partition function of this model can be written

$$Z=\int_{-\pi}^{\pi}D\theta\exp\left(\beta\sum_{\langle\,\rangle}\cos(\Delta_\mu\theta)\right), \quad (4.1)$$

where the functional integration is defined as

$$\int_{-\pi}^{\pi}D\theta\equiv\prod_j\int_{-\pi}^{\pi}d\theta_j, \quad (4.2)$$

the product over j running over all sites of the lattice. We note also that there is a whole sequence of other simple $U(1)$ invariant interactions of the form $\cos(p\Delta_\mu\theta)$, where p is any integer. We will not discuss these interactions further in the present context, but we refer the reader to Sec. III.B where similar terms are treated in the context of Z_N symmetric theories. (See also, José, *et al.*, 1977).

To write the dual form of Eq. (4.1), we follow the recipe used before. First, we rewrite Eq. (4.1) using the Fourier expansion

$$e^{\beta\cos(\tau)}=\sum_{n=-\infty}^{\infty}I_n(\beta)\,e^{in\tau}. \quad (4.3)$$

The $I_n(\beta)$'s are the modified Bessel functions. We have

$$\begin{aligned}Z&=\int_{-\pi}^{\pi}D\theta\prod_{\langle\,\rangle}e^{\beta\cos(\Delta_\mu\theta)}\\&=\sum_{\{k_\mu\}}\int D\theta\prod_{\langle\,\rangle}I_{k_{\mu;j}}(\beta)\,e^{ik_{\mu;j}\Delta_\mu\theta_j}\\&=\sum_{\{k_\mu\}}\prod_{\langle\,\rangle}I_{k_{\mu;j}}\,\beta)\int D\theta\exp\left(i\sum_{\langle\,\rangle}k_{\mu;j}\Delta_\mu\theta_j\right),\end{aligned} \quad (4.4)$$

where we have introduced one $k_{\mu;j}$ which takes on all integer values for each link of the lattice. Assuming periodic boundary conditions, we may sum by parts the exponent of the last factor in Eq. (4.4). Doing the θ integral we find

$$Z=\sum_{\{k\}}\prod_{\langle\ \rangle} e^{\ln[I_{k_{\mu;j}}(\beta)]}\prod_i\int_{-\pi}^{\pi}d\theta_i\, e^{i\theta_i\Delta_\mu k_{\mu;i}}$$

$$=\sum_{\{k\}}\exp\Big(\sum_{\langle\ \rangle}\ln[I_{k_{\mu;j}}(\beta)]\Big)\prod_i\delta(\Delta_\mu k_{\mu;i})\,. \tag{4.5}$$

The delta functions in Eq. (4.5) are Kronecker delta functions, and enforce the condition that the $k_{\mu\ i}$ are divergenceless at each lattice site. This condition is satisfied if and only if we are able to write $k_{\mu\ i}$ as a curl. To this end, we introduce on each dual lattice site an integer-valued field ϕ and identify

$$k_{\mu;i}=\varepsilon_{\mu\nu}\Delta_\nu\phi_i\,, \tag{4.6}$$

where, as before, the indices on the right-hand side of Eq. (4.6) refer to the dual lattice, and those on the left-hand side to the original lattice. Using Eq. (4.6) in Eq. (4.5), we have, up to an overall uniform infinite factor,

$$Z=\sum_{\{\phi=-\infty\}}^{\infty}\exp\Big(\sum_{l_d}\ln[I_{\epsilon_{\mu\nu}\Delta_\nu\phi_j}(\beta)]\Big). \tag{4.7}$$

The field ϕ is the disorder parameter for this system. As such we expect that when $\beta\ll1$ (high temperatures) and the original variables $e^{i\theta}$ are disordered, large fluctuations in ϕ will be suppressed and the major contribution to Z will come from $\{\phi\}$'s which do not vary too much over the dual lattice. To see this we invert Eq. (4.3) and examine the integral representation for $I_n(\beta)$:

$$I_n(\beta)=\frac{1}{\pi}\int_0^{\pi}dx\, e^{\beta\cos x}\cos(nx)\,. \tag{4.8}$$

For $\beta\ll1$ we can expand Eq. (4.8) in powers of β. We find

$$I_n(\beta)=\delta(n)+\tfrac12\beta[\delta(n+1)+\delta(n-1)]$$
$$+\tfrac14\beta^2[\delta(n+2)+\delta(n-2)+2\delta(n)]+\cdots, \tag{4.9}$$

where the delta functions are Kronecker delta functions. Using Eq. (4.7), we have

$$Z=\sum_{\{\phi\}}\prod_{l_d}\Big(1+\frac{\beta^2}{2}\Big)\delta[(\Delta_\mu\phi_j)^2]+\frac{\beta}{2}\delta[(\Delta_\mu\phi_j)^2-1]$$
$$+\frac{\beta^2}{4}\delta[(\Delta_\mu\phi_j)^2-4]+O(\beta^3)+\cdots\,. \tag{4.10}$$

Thus, at infinite temperature, the only contribution to Z comes when all the ϕ_j's have the same value. As β is increased from zero the system becomes more and more disordered in ϕ. First nearest neighbor jumps by one unit become likely and as β is increased still further, higher-energy configurations of the disorder parameter become more likely with $|\Delta_\mu\phi_j|$ taking on larger and larger values.

Let us now discuss Eq. (4.7) in the low-temperature limit, $\beta\gg1$. To do this, we observe that Eq. (4.8) can be expanded in powers of n^2. Taking the logarithm of this series and inserting it in Eq. (4.7) we obtain

$$Z=\sum_{\{\phi\}}\exp\Big(\sum_{l_d}\sum_{p=1}^{\infty}\frac{D_p(\beta)}{p!}(\Delta_\mu\phi_j)^{2p}\Big). \tag{4.11}$$

We can now use the Poisson summation formula (3.19) to write Eq. (4.11) in the form

$$Z=\sum_{\{m_j=-\infty\}}^{\infty}\int_{-\infty}^{\infty}D\phi\,\exp\Big(\sum_{l_d}\sum_{p=1}^{\infty}\frac{D_p(\beta)}{p!}(\Delta_\mu\phi_j)^{2p}+i2\pi m_j\phi_j\Big). \tag{4.12}$$

ϕ is now a continuous valued field, and there is one integer m_j for each site of the dual lattice.

Now, the coefficients $D_p(\beta)$ can be analyzed in various limits (Savit, 1978). It is enough for our purposes, however, to note that for $\beta\gg1$, the leading behavior of the $D_p(\beta)\sim\beta^{1-2p}$, and that the term with $p=1$ is a good low-temperature approximation to Eqs. (4.11) or (4.12). [*A priori* this might not have been true since the $D_p(\beta)$'s alternate in sign and so there could have been cancellations among large terms when the field variables become large. But this does not happen (Savit, 1978).] Thus, for large β, we can keep the quadratic terms in Eq. (4.11) or (4.12), and treat the higher-order terms as a perturbation.

In most of the rest of this section we will explicitly keep only the $p=1$ term in Eqs. (4.11) and (4.12). We do this primarily for simplicity since, as we shall see, much of the qualitative physics is already contained in this term. It is also important to realize that the higher-order terms have the same symmetry properties as the $p=1$ term, so by keeping only the $p=1$ term we do not include any extraneous symmetries. Furthermore, the important qualitative feature that ϕ_j becomes ordered as $\beta\to0$ is true even if we drop terms with $p>1$ in Eq. (4.11). [The rate at which the system orders in ϕ for small β does change, however. This can be seen by computing $D_1(\beta)$ for small β and comparing the resulting expression with Eq. (4.10).] On the other hand, one should bear in mind that this is only a low-temperature approximation to Eq. (4.1) and that the other terms can by systematically included.

In this spirit we now truncate Eq. (4.12), retaining only the quadratic terms. The integral over φ is a Gaussian integral which we can carry out. The result is

$$Z=Z_0\sum_{\{m=-\infty\}}^{\infty}{}'\exp\Big(\frac{\pi}{2D_1(\beta)}\sum_{i,j}m_iV_{ij}m_j\Big) \tag{4.13a}$$

$$\simeq Z_0\sum_{\{m\}}{}'\exp\Big(\pi\beta\sum_{i\neq j}m_i\ln|i-j|\,m_j-\beta c\sum_j m_j^2\Big). \tag{4.13b}$$

In Eq. (4.13b) we have used the low-temperature approximation to $D_1(\beta)\approx1/2\beta$, and we have used a very good approximation to the lattice Green's function V_{ij}. c is a constant the value of which is irrelevant for our purposes. Z_0 is just the Gaussian integral with all $m_j=0$. In calculating Eq. (4.13) we have assumed spherical boundary conditions. With this choice of boundary conditions, there is a restriction (at least at low temperatures—see below) on the configurations of m_j's which are allowed in the functional sum. The only allowed configurations in the limit of an infinite lattice are those for which $\sum_j m_j=0$: There is actually an extra term in the exponent of Eq. (4.13) of the form $-\beta E\sum_j m_j$, but $E\sim\ln s$, where s is the total volume of the system, so that in the limit $s\to\infty$ all configurations for which $\sum_j m_j \neq 0$ are energetically suppressed.

The partition function in Eq. (4.13) has the form of a partition function for a (neutral) two-dimensional Coulomb gas in which the charges can have any integer value, multiplied by a partition function for a system of noninteracting harmonic excitations. We want to ask two questions about this expression: First, how are we to interpret these Coulomb charges and harmonic excitations, and second, what can Eq. (4.13) tell us, at least qualitatively, about the physics of the x-y model? We will address the former question first.

2. Interpretation of the Coulomb gas and the Villain approximation

To understand the fields m and ϕ in Eqs. (4.12) and (4.13), it is very convenient to introduce an approximation to Eq. (4.1) which is accurate at low temperatures. This is the so-called Villain approximation (Berezinski, 1970, 1972; Villain, 1975) and is just the $U(1)$ generalization of the expression (3.5). Let us consider Eq. (4.1) in the limit that $\beta \gg 1$. In that case, the only important configurations in the partition function will be those for which $\cos(\Delta_\mu \theta_j)$ is close to one. Thus we can expand the cosine keeping only the quadratic term. However, this is not the whole story, since if $\Delta_\mu \theta_j = 2\pi n$ and n is an integer, $\cos(\Delta_\mu\theta_j)=1$. To take this periodicity into account we introduce an auxiliary integer-valued vector $J_{\mu;j}$ associated with the lattice links, and we write

$$Z \simeq e^{2N\beta} \sum_{\{J_\mu=-\infty\}}^{\infty} \int_{-L}^{L} D\theta \exp\Big(-\frac{\beta}{2}\sum_l (\Delta_\mu\theta_j - 2\pi J_{\mu\,j})^2\Big). \tag{4.14}$$

Here N is the total number of lattice sites. Now, there are at least two ways to deal with Eq. (4.14), and these have to do with different ways of defining the limits on the sum over J_μ and the integral over θ. First we note that the most natural definition for the range of θ in Eq. (4.14) is $-\pi<\theta\le\pi$, since this is how θ is defined in Eq. (4.1). With that definition of L, we evidently need to sum independently over the $2N$ J_μ's. On the other hand, we can imagine formally extending the range of θ from $-\infty$ to ∞. It is clear that, at worst, this will produce a uniform overcounting of states in Eq. (4.14). But if we do let $L=\infty$, then we clearly do not need to sum over all J_μ's independently: shifting $J_{\mu;j}\to J_{\mu;j}+\Delta_\mu\rho_j$, where ρ_j is an integer-valued scalar field, just amounts to a redefinition of θ_j and counts again configurations which have already been counted. It is instructive to compute Eq. (4.14) with both these definitions of the range of θ and J_μ and show that formally they give the same result, although a different interpretation of the fields J_μ emerge.

First let us take $L=\pi$, and imagine doing an unrestricted sum over J_μ. We introduce a new continuous variable $k_{\mu;j}$, associated with the links of the lattice, and write Eq. (4.14) as

$$\begin{aligned} Z &= (2\beta e^{2\beta})^N \sum_{\{J\}} \int_{-\pi}^{\pi} D\theta \int_{-\infty}^{\infty} Dk_\mu \exp\Big(\sum -\frac{1}{2\beta}k_{\mu;j}^2 + ik_{\mu;j}(\Delta_\mu\theta_j - 2\pi J_{\mu;j})\Big) \\ &= (2\beta e^{2\beta})^N \int_{-\pi}^{\pi} D\theta \sum_{\{k_\mu=-\infty\}}^{\infty} \exp\Big(\sum -\frac{1}{2\beta}k_{\mu;j}^2 - i\theta_j\Delta_\mu k_{\mu;j}\Big), \end{aligned} \tag{4.15}$$

where we have performed the sum over J_μ and used Eq. (3.19) to write the integral over k_μ as a sum. Since the k_μ's are now integers, the integral over the θ_j will just produce a flock of Kronecker delta functions:

$$Z = (4\pi\beta e^{2\beta})^N \sum_{\{k\}} \exp\Big(-\frac{1}{2\beta}\sum_l k_{\mu,j}^2\Big)\prod_i \delta(\Delta_\mu k_{\mu;i}). \tag{4.16}$$

In the usual way, we associate integer-valued fields ϕ, with the sites of the dual lattice and use expression (4.6) to satisfy the delta functions. We can then use Eq. (3.19) again to turn the sum over integer-valued ϕ's into a integral over continuous ϕ's times a sum over a new integer-valued variable, m. We have

$$Z = (4\pi\beta e^{2\beta})^N \int D\theta \sum_{\{m\}} \exp\Big(\sum -\frac{1}{2\beta}(\Delta_\mu\phi_j)^2 + i2\pi m_j\phi_j\Big). \tag{4.17}$$

which agrees with Eq. (4.12) when $\beta\gg 1$ [up to overall factors which were neglected in Eq. (4.12)]. Note that the J_μ's in Eq. (4.15) do not have an obvious interpretation in Eq. (4.17).

Now let us take $L=\infty$ in Eq. (4.14) and do a restricted sum over $\{J_\mu\}$, avoiding summing over sets of J_μ which differ from each other only by a gradient. Once again we introduce a continuous Fourier conjugate variable k_μ, and write Eq. (4.14) as

$$\begin{aligned} Z &= (2\beta e^{2\beta})^N {\sum_{\{J\}}}' \int_{-\infty}^{\infty} D\theta \int_{-\infty}^{\infty} Dk_\mu \\ &\quad\times \exp\sum -\frac{1}{2\beta}k_\mu^2 + ik_{\mu;j}(\Delta_\mu\theta_j - 2\pi J_{\mu;j}) \\ &= (4\pi\beta e^{2\beta})^N {\sum_{\{J\}}}' \int_{-\infty}^{\infty} Dk_\mu \\ &\quad\times \exp\Big(\sum -\frac{1}{2\beta}k_\mu^2 - i2\pi k_{\mu;j}J_{\mu;j}\Big)\prod_i \delta(\Delta_\mu k_{\mu;i}), \end{aligned} \tag{4.18}$$

where we have done the integral over the θ_j to produce Dirac delta functions. To satisfy these delta functions, we associate a continuous variable ϕ_j with the sites of the dual lattice. We can now use Eq. (4.6) to satisfy the delta functions, remembering that $k_{\mu;j}$ and ϕ_j are now continuous valued functions. Substituting into Eq. (4.18) and doing an integration by parts, we have

$$\begin{aligned} Z = (4\pi\beta e^{2\beta})^N {\sum_{\{J\}}}' \int_{-\infty}^{\infty} D\phi \exp\Big(&\sum -\frac{1}{2\beta}(\Delta_\mu\phi_j)^2 \\ &+ i2\pi\phi_j\varepsilon_{\mu\nu}\Delta_\mu J_{\nu;j}\Big). \end{aligned} \tag{4.19}$$

If we identify

$$m_j \equiv \varepsilon_{\mu\nu}\Delta_\mu J_{\nu;j}\,, \tag{4.20}$$

then this is identical to Eq. (4.17), since we recall that the prime on the sum over J_μ in Eq. (4.18) means precisely a sum over distinct configurations of $\varepsilon_{\mu\nu}\Delta_\mu J_\nu$ (i.e., we do not sum over $\{J_\mu\}$'s that differ from each other only by a gradient).

Equation (4.20) tells us how to interpret the fields m_j in Eqs. (4.17) and (4.12). From Eq. (4.14), we see that $J_{\mu;j}$ represents an "integer-valued piece" of the angle difference $\Delta_\mu\theta_j$. Furthermore, interpreted this way, we see that we can define an angle Θ_j, that is not necessarily single-valued, by writing

$$\Delta_\mu\Theta_j = \Delta_\mu\theta_j - 2\pi J_{\mu;j}\,. \tag{4.21}$$

It is a simple matter to see that if we compute a phase sum around a closed loop c of $\Delta_\mu\Theta_j$, the result is

$$\oint\!\!\!\!\sum \Delta_\mu\Theta_j = \sum_s m_j\,, \tag{4.22}$$

where m_j is given in Eq. (4.20), and the sum over s is over all dual lattice sites contained in c. Thus the m_j's are just the vortices of the angles Θ_j. On the other hand, we can effectively identify the Θ_j with the angles of the original x-y model, Eq. (4.1). This requires some discussion: First, examining Eq. (4.1) we see that a state of the x-y model is defined by assigning a unique value to the phase $e^{i\theta}$ at each site of the lattice. In Eq. (4.1), we have defined $-\pi < \theta_j \leq \pi$. We can, however, formally extend the range of θ_j to $-\pi\lambda < \theta_j \leq \lambda\pi$, where λ is some integer. Then we see that θ_j need not be single valued. That is, we could go around some closed loop on the lattice in which the spins were varying (even slowly) from site to site, and have the quantity $\oint\!\!\!\!\sum \Delta_\mu\theta_j$ be nonzero. (Consider, for example, a simple circular configuration in which the head of spin chases its neighbor's tail in a clockwise direction.) The possibility of allowing θ_j in Eq. (4.1) to be non-single-valued exists precisely because the action is periodic, which is also the reason we had to include the J_μ's in the model (4.14) to make it a good low-temperature approximation to Eq. (4.1). Now a good large β approximation to Eq. (4.1) is just to retain the quadratic term in $\Delta_\mu\theta_j$, if we remember that we must not lose the periodicity of the original action. Using Eq. (4.21) we see that the action in Eq. (4.14) is approximately $(\Delta_\mu\Theta_j)^2$, and properly takes account of the periodicity, as explained before. Thus we may effectively identify Θ_j with the angles appearing in Eq. (4.1), and so the m_j can be interpreted as representing the vortices of the original x-y model degrees of freedom. Note that this interpretation requires use of the Villain form (4.14) and is completely obscure in the argument leading to Eq. (4.12). In fact, in that presentation, the m_j's (and the ϕ_j's) appear as a complete set of fields replacing the original x-y model spins, whereas in the derivation leading to Eq. (4.19), the m_j's appear as a piece of the effective x-y model angles Θ_j.

The existence of vortexlike excitations in this model follows from very general homotopy arguments. The vortices are topologically stable and exist because the symmetry of the model (4.1) is $U(1)$. [Equation (4.14) is effectively $U(1)$.] Following the philosophy of the duality transformation leading to Eq. (4.12), we are led to the interpretation that the x-y model can be completely and exactly expressed in terms of the topological excitations (the m_j) plus a set of spin waves (the continuous-valued ϕ_j). This same result, that the theory can be written exactly in terms of topological excitations, will obtain in the other $U(1)$ theories we shall study in this section. Further interpretive comments on this topic may be found in Kostelitz and Thouless (1973), Jose *et al.* (1977), and Savit (1978).

3. Qualitative physics of the two-dimensional x-y model

The vortex field m_j plays an important role in determining the behavior of the x-y model at various temperatures. For the purposes of this discussion, it is enough to limit ourselves to the quadratic approximation. The quantitative effects of higher-order terms are discussed in Jose *et al.* (1977). [See, however, Luther and Scalapino (1977) for a divergent treatment. It should be pointed out that the conclusions of this paper are in contradiction with those of Jose *et al.* (1977), and are generally believed to be incorrect. Nevertheless, the approach is very interesting.]

From Eqs. (4.12) and (4.13), we see that the model can be described by the vortex field and another, continuous field ϕ_j. Using the Villain approximation, one easily sees that this continuous Gaussian field is just the Fourier transform field (in the sense of the duality transformation) of the Gaussian spin waves which appear when we approximate $\cos(\Delta\theta) \sim 1 - \frac{1}{2}(\Delta\theta)^2$ and ignore the periodicity. From Eq. (4.13), we see that at any nonzero temperature we can excite the noninteracting, massless Gaussian spin waves (whose partition function is Z_0), while for large enough β there is suppression of certain vortex excitations. In particular, because of the "chemical potential" term in Eq. (4.13) the density of vortices will be small at low temperatures, and due to the logarithmic potential the vortices will tend to be bound in vortex-antivortex pairs. In this region the spin-spin correlation function, $\langle e^{i\theta_i}e^{i\theta_j}\rangle$, will fall to zero like a power of the spin separation with a temperature dependent exponent. [Consistent with the Mermin-Wagner theorem (Mermin and Wagner, 1966), $\langle e^{i\theta_i}\rangle = 0$ at all nonzero temperatures in this (infinite) two-dimensional system.]

Above a certain temperature T_c, the situation is dramatically altered. To understand this let us focus on the behavior of a vortex-antivortex pair. It is easy to see that the entropy of a vortex-antivortex pair separated by a distance r is proportional to $\ln r$. There is therefore a competition in the free energy for such a pair between the entropy and energy terms. For $T > T_c$ the entropy term dominates and the free energy for a single vortex-antivortex pair has its minimum at $r = \infty$. Thus for $T > T_c$ the vortex-antivortex pairs become unbound, creating a plasma of vortices with arbitrary separation. This plasma "screens" the spin-spin correlation function, causing it to behave like $e^{-\mu r}$, where r is the spin separation. The behavior of the theory at T_c is nonanalytic, and this phase transition is due precisely to the unbinding of the vortex-antivortex pairs and the concurrent condensation of the plasma of the "free" topological excitations (vortices) of the system. It is also interesting to note that for $T>T_c$ the prime on the sum over $\{m\}$ in Eq. (4.13) can be removed. This can be seen by placing the system on a finite lattice of linear dimension s. The energy of a single vortex $\propto \ln s$. However, the entropy of a single vortex is also proportional to $\ln s$. Thus in the free energy of a single vortex there is also a competition between energy and entropy. For $T>T_c$ the entropy dominates, and taking $s\to\infty$ we see that we no longer have a suppression factor for configurations with $\sum_j m_j \neq 0$.

The picture painted in this section is only impressionistic and the arguments crude and qualitative, but, nonetheless, substantially correct. It will be useful to keep in mind the idea of a phase transition induced by a condensation of topological excitations as we study other theories with a $U(1)$ symmetry.

C. Dual forms of general $U(1)$ invariant theories

Here we will generalize the results of Sec. IV.B and derive the duality transformation for a $U(1)$ invariant theory with simplex number s in d dimensions (Savit, 1977). As in the discussion of the two-dimensional x-y model, we will be able to write these theories directly in terms of their topological excitations. As we shall see, a very pretty pattern for these topological excitations will emerge. The derivation we present will be somewhat loose in that we shall not pay careful attention to problems of gauge redundancy. As in previously discussed theories this redundancy gives rise to overall infinite factors and while harmless, should in principle be removed. We will post appropriate warning signs at slippery places in the derivation. For a careful treatment of gauge problems in one $U(1)$ invariant theory, the $d=3$ x-y model, see Savit (1978).

We begin by introducing the $U(1)$ invariant theory with simplex number s. It is defined in precise analogy with the Z_N symmetric theories of Sec. III.C. Consider a d-dimensional lattice. On each simplex of dimension $s-1$ place a complex phase

$$Q_{\mu_1\cdots\mu_{s-1};i}=\exp(i\,\theta_{\mu_1\cdots\mu_{s-1};i}), \tag{4.23}$$

with $-\pi<\theta\leq\pi$. These phases interact according to the form

$$I^{(1)}_{\mu_1\cdots\mu_s;i}=\cos\left(\frac{1}{(d-s)!}\,\varepsilon_{\mu_1\cdots\mu_s,\alpha_1\cdots\alpha_{d-s}}\,\varepsilon_{\alpha_1\cdots\alpha_{d-s},\beta,\gamma_1\cdots\gamma_{s-1}}\Delta_\beta\theta_{\gamma_1\cdots\gamma_{s-1};i}\right)\equiv\cos(\varepsilon\varepsilon\Delta\theta), \tag{4.24}$$

as in Eq. (3.25). The partition function of the simplex s theory is

$$Z=\int_{-\pi}^{\pi\,\prime} D\theta\exp\left(\beta\sum_s\cos(\varepsilon\varepsilon\Delta\theta)\right), \tag{4.25}$$

where the sum in the exponent runs over all simplices of dimension s on the d dimensional lattice. The prime on the integral indicates that there is (except for the case $s=1$) a harmless, uniform, and usually ignorable gauge redundancy in the functional integral over θ. This redundancy follows from the existence of a local gauge invariance in Eq. (4.25). The Hamiltonian therein is invariant under the operation[3]

$$\begin{aligned}\theta_{\mu_1\cdots\mu_{s-1};i}\to\theta'_{\mu_1\cdots\mu_{s-1}}\\ =\theta_{\mu_1\cdots\mu_{s-1};i}+\Delta_{\mu_1}\Lambda_{\mu_2\cdots\mu_{s-1};i},\end{aligned} \tag{4.26}$$

where Λ is an angle valued field associated with the simplices of dimension $s-2$ [see, by analogy, Eq. (3.28)]. As in the case of the two-dimensional x-y model, we can define a whole sequence of $U(1)$ invariant simplex number s interactions according to the formula

$$I^{(p)}_{\mu_1\cdots\mu_s;i}=\cos(p\,\varepsilon\varepsilon\Delta\theta), \tag{4.27}$$

where p is any integer. We shall have no more to say about these interactions here. See Secs. III.B and III.C for a discussion of similar interactions in the context of Z_N symmetric theories.

We now rewrite Eq. (4.25) by Fourier expanding each factor in the partition function. We have

$$Z=\int_{-\pi}^{\pi\,\prime}D\theta\prod_s\sum_{k=-\infty}^{\infty}I_k(\beta)\,e^{ik\cdot\varepsilon\varepsilon\Delta\theta} \tag{4.28a}$$

$$=\sum_{\{k\}}\prod_s I_k(\beta)\int' D\theta\exp\left(i\sum_s k\cdot\varepsilon\varepsilon\Delta\theta\right) \tag{4.28b}$$

$$=\sum_{\{k\}}\prod_s I_k(\beta)\int' D\theta\exp\left(-i\sum_t\theta\cdot\varepsilon\varepsilon\Delta k\right). \tag{4.28c}$$

We have introduced a set of integer-valued Fourier conjugate variables $k_{\mu_1\cdots\mu_s;i}$ associated with the s-dimensional simplices of the lattice. In Eq. (4.28c) we have summed by parts in the exponent to produce

$$\sum_t\theta_{\mu_1\cdots\mu_{s-1};i}\varepsilon_{\mu_1\cdots\mu_{s-1},\beta,\alpha_1\cdots\alpha_{d-s}}\times\varepsilon_{\alpha_1\cdots\alpha_{d-s},\gamma_1\cdots\gamma_s}\Delta_\beta k_{\gamma_1\cdots\gamma_s;i}, \tag{4.29}$$

where the sum t runs over all simplices of dimension $s-1$.

Next, we want to integrate over θ, and so we must decide what to do with the prime on the functional integral. The simplest procedure is to ignore it. This will just produce extra factors of 2π (an infinite number in an infinite lattice). The situation here is exactly analogous to that encountered in our treatment of the $d=3$ Z_2 gauge theory (with simplex number 2) discussed in Sec. II.B.2. In that case we found that failing to choose a gauge just produced extra factors of the volume of the gauge group (for Z_2 extra factors of 2) as well as introducing redundant Kronecker delta functions in Z. It is a straightforward matter to show that the same thing happens here. Thus we will formally ignore the prime on the integral over θ secure in the understanding that in so doing we just change Z by a constant factor.

Proceeding with the θ integrals and dropping overall factors of 2π, we find for Eq. (4.28),

$$Z=\sum_{\{k\}}\exp\Big(\sum_s \ln[I_k(\beta)]\Big)\prod_t \delta(\varepsilon\varepsilon\Delta k). \tag{4.30}$$

To satisfy the Kronecker delta functions, it is necessary and sufficient to write k in the form

$$k_{\mu_1\cdots\mu_s;i}=\frac{1}{(d-s-1)!}\varepsilon_{\mu_1\cdots\mu_s,\beta,\alpha_1\cdots\alpha_{d-s-1}}\times\Delta_\beta\phi_{\alpha_1\cdots\alpha_{d-s-1};i}\equiv\epsilon\Delta\phi, \tag{4.31}$$

where the φ's are a set of integer-valued fields associated with the simplices of dimension $d-s-1$ on the dual lattice. {The factor $[(d-s-1)!]^{-1}$ is for normalization; the k's and ϕ's may be taken to be antisymmetric in their direction indices.}

For simplicity, we now make the same low-temperature approximation to $\ln I_k(\beta)$ that was made before Eq. (4.13) [i.e., keeping only the quadratic term in the exponent of Eq. (4.11) and approximating $D_1(\beta)\approx 1/2\beta$ for $\beta\gg 1$]. Using Eq. (4.31) in Eq. (4.30), we can write

$$Z\approx\exp(N_s\beta)\sum_{\{\epsilon\Delta\phi\}}\exp\Big(-\frac{1}{2\beta}\sum_s(\epsilon\epsilon\Delta\phi)^2\Big),\quad \beta\gg 1, \tag{4.32}$$

where N_s is the number of s-dimensional simplices on the lattice. In the exponent we have written $k^2=(\epsilon\Delta\phi)^2=(\epsilon\epsilon\Delta\phi)^2$ [with an extra normalization factor included in $\epsilon\epsilon\Delta\phi$, in analogy to Eq. (3.25)]. Writing k^2 this way makes it more apparent that the sum in the exponent can be considered to be over all $\tilde{s}=d-s$ dimensional simplices on the dual lattice since $\epsilon\epsilon\Delta\phi$ has $\tilde{s}$ direction indices, naturally associated with the dual lattice. Finally, the sum over states in Eq. (4.32) is understood to be a sum over the largest set of configurations of ϕ's such that each configuration gives a distinct configuration of $k=\epsilon\Delta\phi$.

Equation (4.32) possesses the expected characteristics for the theory dual to Eq. (4.25). First, it has dual lattice interactions defined on simplices of dimension $\tilde{s}=d-s$, which are similar in form to the interactions in Eq. (4.25) (compare $\epsilon\epsilon\Delta\theta$ with $\epsilon\epsilon\Delta\phi$). Second, high-temperature regions of Eq. (4.25) are mapped into low-temperature regions of Eq. (4.32) and vice versa. Finally, the symmetry group of Eq. (4.32) is Z_∞, the additive group of integers, which is the natural group to associate with the indices which label the Fourier expansion on the group $U(1)$. All of these characteristics are independent of the low-temperature approximation made in Eq. (4.32) and hold for the full theory as well.

Next we shall introduce the topological excitations. This is done by using the Poisson summation formula (3.19) to replace the sum over $\{\epsilon\Delta\phi\}$ in Eq. (4.32) by an integral over $\epsilon\Delta\phi$ times a sum over another integer-valued variable. We have

$$Z\approx\exp(N_s\beta)\int_{-\infty}^{\infty}D(\epsilon\Delta\phi)\times\sum_{\{L=-\infty\}}^{\infty}\exp\Big(\sum-\frac{1}{2\beta}(\epsilon\epsilon\Delta\phi)^2+i2\pi L\epsilon\Delta\phi\Big), \tag{4.33}$$

where

$$L\epsilon\Delta\phi\equiv\frac{1}{s!\,(d-s-1)!}\times L_{\mu_1\cdots\mu_s;i}\epsilon_{\mu_1\cdots\mu_s,\beta,\alpha_1\cdots\alpha_{d-s-1}}\times\Delta_\beta\phi_{\alpha_1\cdots\alpha_{d-s-1};i}. \tag{4.34}$$

(Note that this could also have been written $\tilde{L}\epsilon\epsilon\Delta\phi$, in which case $\tilde{L}$ would have $d-s$ indices and be defined on the dual lattice.) Now we rearrange the last term in the exponent of Eq. (4.33) by summing by parts. Recalling that we have spherical boundary conditions there is no surface term, and the last term becomes

$$\begin{aligned}2\pi i\sum(L\epsilon\Delta\phi)&=-2\pi i\sum\frac{1}{s!\,(d-s-1)!}\phi_{\alpha_1\cdots\alpha_{d-s-1};i}\times\epsilon_{\mu_1\cdots\mu_s,\beta,\alpha_1\cdots\alpha_{d-s-1}}\Delta_\beta L_{\mu_1\cdots\mu_s;i}\\&\equiv-2\pi i\sum\frac{1}{(d-s-1)!}\phi_{\alpha_1\cdots\alpha_{d-s-1};i}\times J_{\alpha_1\cdots\alpha_{d-s-1};i}\equiv-2\pi i\sum(\phi J),\end{aligned} \tag{4.35}$$

with

$$J_{\alpha_1\cdots\alpha_{d-s-1};i}=\frac{1}{s!}\epsilon_{\mu_1\cdots\mu_s,\beta,\alpha_1\cdots\alpha_{d-s-1}}\times\Delta_\beta L_{\mu_1\cdots\mu_s;i}. \tag{4.36}$$

J is an integer-valued current associated with the simplices of dimension $d-s-1$ on the dual lattice. From Eq. (4.36) we easily see that J satisfies the property that

$$\Delta_{\alpha_r}J_{\alpha_1\cdots\alpha_r\cdots\alpha_{d-s-1};i}=0 \tag{4.37}$$

for any r between 1 and $d-s-1$. So finally we may write Eq. (4.33) as

$$Z\simeq\exp(N_s\beta)\int_{-\infty}^{\infty\,\prime}D\phi\sum_{\{J\}}{}'\exp\Big(\sum\frac{-1}{2\beta}(\epsilon\epsilon\Delta\phi)^2+i2\pi\phi J\Big) \tag{4.38a}$$

$$=\exp(N_s\beta)Z_0\sum_{\{J\}}{}'\exp\Big(\beta\sum_{i,j}J_iV_{ij}J_j\Big). \tag{4.38b}$$

The prime on the integral over ϕ reminds us that we should integrate only over ϕ's that produce distinct configurations of $\epsilon\Delta\phi$, i.e., we should choose a gauge. The prime on the sum over J means that we should sum over integer J's from $-\infty$ to $+\infty$, but only over those J's satisfying Eq. (4.36), or, equivalently, Eq. (4.37).

The J's are the topological excitations of the model and are just generalizations of the m's in Eq. (4.12), which represented the vortices of the two dimensional x-y model. In this context the geometrical significance of Eq. (4.37) is clear: The J's are associated with $d-s-1$ dimensional dual simplices. Our interpretation of the J's tells us that the allowed topological excitations of the theory have dimension $d-s-1$, and Eq. (4.37) tells us that these excitations are defined on closed manifolds. Consider, for example, the $d=3$ x-y model. In that case the J's are just vortex string bits, and Eq. (4.37) tells us that we can have only closed vortex strings, since $\Delta_\mu J_\mu \neq 0$ at the end of a string. Similarly, the $d=4$ x-y model has topological excitations which are two-dimensional surfaces embedded in the four-dimensional dual lattice. If we try to lay down a configuration of J's which has boundaries, we will find that $\Delta_\alpha J_{\alpha\beta} \neq 0$ at the boundaries. Thus we are allowed to have only closed two-dimensional surfaces.

These topological excitations interact with each other and with themselves. The interaction is expressed in the exponent of Eq. (4.38b), wherein we have suppressed the direction indices on the J's and on the matrix V_{ij}. V_{ij} has a diagonal piece which is just the self-energy of the J's. For large separations, $V_{ij} \sim |i-j|^{-d+2}$ for $d \neq 2$, and $V_{ij} \sim \ln|i-j|$ for $d=2$. The factor Z_0 just represents the *dual* spin waves of the system, which are described by antisymmetric $d-s-1$ forms, in precise analogy to the scalar spin waves of the two-dimensional x-y model.

We have stated that the J's represent the topological excitations of the original degrees of freedom, but that interpretation is certainly not obvious. To make it clear, one should go back and construct the analog of the Villain approximation for the general case considered here and follow the steps leading to Eq. (4.19). The calculation is quite straightforward and we shall not perform it. Those interested will have little difficulty in confirming our interpretation.

Before continuing, we wish to make an important technical comment which concerns the gauge invariance of a given configuration of J's and the derivation of Eq. (4.38). To perform the Gaussian integral over ϕ in Eq. (4.38a) requires a gauge choice (for $s>1$). Suppose, for instance, that we choose an axial-like gauge. This would mean setting some of the ϕ's equal to zero and integrating over the rest. The argument of Eq. (4.38b) will therefore not contain some factors of J, in particular, these which are coupled in Eq. (4.38a) to the ϕ's which have been set equal to zero. One might then be concerned that the energy of some of the allowed configurations of J's would have been changed by the gauge choice. However, this is not the case. Recall that the only allowed configuration of J's is that satisfying Eq. (4.37). Suppose we consider open edge boundary conditions, and let us choose a complete axial gauge, eliminating all residual gauge invariance. One can show that even with such a complete gauge choice, any configuration satisfying Eq. (4.37) will still have some of its J's appearing in the exponent of Eq. (4.38b). (That is, a complete axial gauge fixing does not involve eliminating all degrees of freedom on a closed manifold.) Moreover, the energy associated with each allowed configuration of J's will be independent of the gauge choice. This will be enforced in Eq. (4.38b) by the appearance in V_{ij} of long-range Coulomb potentials in special directions, depending on the gauge choice. Thus each allowed divergenceless configuration of J's has a gauge invariant meaning, as we formally expect from Eq. (4.38a), and as it must if our interpretation of the J's as topological excitations is to be correct. A detailed discussion of this question for the three-dimensional x-y model can be found in Savit (1978).

It may be helpful to conclude this section with a brief summary: We have studied the large class of $U(1)$ invariant lattice theories in d dimensions with simplex number s. We have derived the dual form for these theories. The dual theory is a Z_∞ invariant theory, also in d dimensions, with simplex number $\tilde{s}=d-s$. It has the usual property of a dual theory, namely, that the high- (low-) temperature region of the original theory is mapped into the low- (high-) temperature region of the dual theory. Furthermore, we have found that the dual theory can be easily converted into a third form. In this form the physical degrees of freedom that appear are the spin waves [actually, their dual (Fourier) conjugates] and the topological excitations of the original $U(1)$ invariant spins. These topological excitations exist on closed manifolds of dimension $d-s-1$.

In the next subsection we will discuss some of the physics of these $U(1)$ invariant theories. In Sec. V we will briefly describe the Abelian Higgs model which is a kind of hybrid of the theories treated in this section, and we will also briefly address a question which may have occurred to the reader: What are the topological excitations of the Z_2 and Z_N symmetric theories?

D. Some physical comments: Phase transitions and topological excitations

We have already remarked that dual formulations of a theory are useful for obtaining a simple qualitative picture of the behavior of a theory at high temperatures, when fluctuations in the disorder parameter are small. Furthermore, one can perform a "high-temperature" expansion in $\tilde{\beta}$ which corresponds to an expansion about $T=0$, in terms of the original variables of the theory. In addition to these benefits, the formulation of the theory in terms of its topological excitations (4.38) can provide very important qualitative (and quantitative) understanding of other aspects of the behavior of the theory, in particular its phase structure. We have already seen this in our discussion of the two-dimensional x-y model. Here we shall present variations on that theme for three theories of physical interest: the three-dimensional x-y model and the three- and four-dimensional $U(1)$ gauge theory ($s=2$) (Banks *et al.*, 1977; Savit, 1978; Peskin, 1978; Einhorn and Savit; 1979).

Consider first the $d=3$ x-y model. Among other reasons, this theory is of physical interest because its critical behavior is thought to be related to that of superfluid ^{4}He. The theory has topological excitations which are closed vortex strings, or vortex strings which terminate on the boundary of the system. In its application to ^{4}He, these strings represent the vortices in bulk superfluid. It is interesting to note that if the model really does have something to do with ^{4}He, then we have an example of an extended system with macroscopic excitations which interact through a gauge principle. The reason is that when $s=1, d=3$, Eq. (4.38a) has the form of the generating functional for three-dimensional QED without electrons, but coupled to conserved (integer-valued) currents.

For this theory the exponent in Eq. (4.38b) takes the approximate form

$$-\beta \sum_{i\neq j} aJ_{\mu;i}\frac{D_{\mu\nu}}{|i-j|}J_{\nu;j}+bJ^2_{\mu;i}\,, \tag{4.39}$$

where a and b are positive constants and $D_{\mu\nu}=\delta_{\mu\nu}$ +gauge terms which do not contribute so long as $\Delta_\mu J_\mu=0$. From this we see that vortex currents with the same sense (both positive or both negative) repel each other, while those with opposite sense attract. This is the opposite of what happens with electric currents in real QED and can be traced to the factor of i in front of the J term in Eq. (4.38a).

Now, suppose β is very large (low temperatures). Because of the term bJ^2 in Eq. (4.39) we expect to have only a low density of small vortex loops, in addition to the spin waves represented by the Z_0 factor in Eq. (4.38b). As β decreases, we pay a decreasing penalty in probability for creating vortex strings, and we expect their size and density to grow. We now ask whether we expect this size and population increase to occur smoothly or nonanalytically. To answer this question, consider the free energy of a closed vortex loop of total length L and unit flux. Neglect for a moment the r^{-1} interaction term in Eq. (4.39). In that case the energy of the loop is just proportional to L. The entropy of the loop is (up to logarithmic corrections) also proportional to L. To see this just notice that the number of configurations of the loop of length L are those of a modified nonrepeating nonoverlapping random walk which returns to the origin. This number is of the form $e^{pL}f(L)$ where $F(L)$ is a slowly varying function (powers and logarithims), and p is a numerical constant. The log of this is proportional to the entropy, and so the free energy has the approximate form

$$\beta F=\beta bL-pL+O(\ln L)\,. \tag{4.40}$$

For large β, this has a minimum at $L=0$, while for $\beta<p/b\equiv\beta_c$, the minimum is at $L=\infty$. If we believe this argument, then we expect a phase transition at $\beta\simeq\beta_c$. For low temperatures we have a soup (alphabet) of small vortex rings with some density. As we raise the temperature the density and size of the rings increases, until suddenly at $\beta=\beta_c$ we find that the rings grow dramatically, and we leave the alphabet soup phase and enter the spaghetti phase. This high-temperature phase can therefore be described as a condensation of topological excitations. We expect that this transition is the usual order-disorder transition of the $d=3$ x-y model, but looked at in a rather different way. In the spaghetti phase, therefore, the x-y model spin-spin correlation function should fall exponentially to zero for large separations, while in the alphabet soup phase it should approach a nonzero constant. From this point of view we can say that the phase transition and accompanying disordering is caused by the condensation of the topological excitations.

This picture is clearly quite similar to that which we presented for the $d=2$ x-y model, and is very attractive and reasonable. However, there is an important difference in its derivation: In the $d=2$ case, the vortex-vortex interaction was crucial in determining the critical temperature. Indeed, it was the competition between this term and the entropy which indicated the existence of a phase transition. In Eq. (4.40) we have ignored the current-current interaction term, keeping only the self-energy piece. If the interaction energy fell off very rapidly with distance [like an exponential, say, as in the Abelian Higgs model (see Sec. V)] this would probably be a very good approximation. However, the interaction in Eq. (4.39) is long range, and this could affect the veracity of our picture. Although one can generate arguments on both sides, to firmly resolve the issue requires a careful renormalization group treatment. Nevertheless, it is the considered opinion of the author that the picture presented here is essentially correct, and that while including the interaction term in Eq. (4.39) will certainly have a (perhaps strong) quantiative effect, it will not change things qualitatively.

In a very similar way, one can discuss the $d=4$ $U(1)$ lattice gauge theory ($s=2$). The dual form of this theory has $\tilde{s}=2$, just as in the $d=3$ x-y model. Here again, the topological excitations are closed vortex loops,[8] and in this case the exponent of Eq. (4.38b) has a form very similar to Eq. (4.39), viz.,

$$-\beta\sum_{i\neq j} aJ_{\mu;i}\frac{D_{\mu\nu}}{|i-j|^2}J_{\mu;j}+bJ^2_{\mu;i}\;. \tag{4.41}$$

If we neglect the interaction term in Eq. (4.41), we can clearly repeat the arguments leading to Eq. (4.40) for this theory. Thus we are led to expect a phase transition here also.

Now, a useful gauge invariant correlation function for the $s=2$ theory is the Wilson loop integral (of the original $U(1)$ spins),

$$\Gamma_c=\left\langle \exp\left(i\sum_c \theta_{\mu;i}\right)\right\rangle, \tag{4.42}$$

where c is some closed contour on the original lattice. In the low-temperature (alphabet soup) phase, (which should have properties very much like those of real QED) we expect that $\Gamma_c\sim e^{-P}$, where P is the perimeter of the contour, while in the spaghetti phase $\Gamma_c\sim e^{-A}$, where A is a (minimal) area subtended by c.[9] The different behavior of Γ_c signals the analog of quark confinement in the spaghetti phase, and the absence of quark confinement in the alphabet soup phase; see Banks *et al.* (1977), Einhorn and Savit (1979), Peskin (1978), and Banks and Rabinovici (1979) for further discussion. As with the $d=3$ x-y model, one must remember that our argument for the phase structure of this system neglects the r^{-2} interaction term in Eq. (4.41) and therefore must be handled gingerly.

Finally, we briefly describe the physics of the $d=3$ $U(1)$ lattice gauge theory ($s=2$). The dual form of this

[8] When factors of the lattice spacing are appropriately included, and a naive continuum limit is taken, only quadratic terms survive and the theory becomes just a theory of free photons [the Z_0 factor (4.38)]. In this limit the $U(1)$ symmetry group becomes the group R, and the topological excitations disappear.

[9] Since the dual theory of the $d=4$ $s=2$ theory is also a gauge theory, we can define an analogous "Wilson loop" also for the dual variables $\tilde{\Gamma}_{\tilde{c}}$. We expect that when $\Gamma_c\sim e^{-P}$, $\tilde{\Gamma}_{\tilde{c}}\sim e^{-A}$, and vice versa.

theory has $\tilde{s}=1$, and pointlike topological excitations (a kind of monopole). The exponent in Eq. (4.38b) has the form

$$-\beta\sum_{i\neq j} am_i\frac{1}{|i-j|}m_j+bm_i^2\,. \qquad (4.43)$$

The behavior of this theory has been analyzed by Polyakov (1975, 1977). We will not repeat his analysis here, but will just state the results. The $d=3$ $U(1)$ lattice gauge theory has, for any nonzero temperature, only one phase, a "quark confining" phase characterized by $\Gamma_c\sim e^{-A}$. In this phase, the topological excitations in Eq. (4.43) are condensed into a plasma of unbound monopoles which causes the original $U(1)$ degrees of freedom to be sufficiently disordered to produce the area law fall off for Γ_c. The situation here should be compared with that of the two-dimensional x-y model. In that case, the logarithmic monopole-monopole potential was strong enough to bind monopole-antimonopole pairs at sufficiently low temperature. In Eq. (4.43), the r^{-1} intermonopole potential is not sufficiently strong to do this: the entropy of a monopole-antimonopole pair separated by a distance r is proportional to $\ln r$, and that term dominates over the energy term in the free energy for such a pair at any nonzero temperature, causing the free energy to have its minimum at $r=\infty$.

V. OTHER TOPICS

A. Abelian Higgs model and related models

A model of broad physical interest is the Abelian Higgs model. In the continuum (Euclidean d-dimensional space) this model is described by the Lagrangian

$$\mathcal{L}=\int d^dx\,\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}+|(\partial_\mu-iqA_\mu)\phi|^2+V(\phi)\,, \qquad (5.1)$$

where $F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu$, A_μ is the electromagnetic vector potential, and ϕ is a complex scalar field. q is the charge of the scalar field in units of the fundamental charge e. We take for the potential

$$V(\phi)=\lambda(|\phi|^2-f^2)^2\,, \qquad (5.2)$$

with $\lambda\geq 0$.

If f^2 is negative, V has its minimum when $|\phi|=0$, and Eq. (5.1) describes massive scalar QED in the symmetric phase with a $|\phi|^4$ interaction for the scalar field. If f^2 is positive, V is minimized when $|\phi|=f$. In this case Eq. (5.1) is forced into a broken symmetry phase, and by the usual Higgs mechanism the massless photon field and part of the scalar field combine to form a massive vector field. In the broken symmetry phase, Eq. (5.1) takes on the form of the Ginzburg-Landau theory of superconductivity.

Now in the theory of superconductivity, strings of magnetic flux play a central role. These magnetic flux lines can be understood as topological excitations of the degrees of freedom in Eq. (5.1). These remarks suggest that duality transformations may be useful in studying this theory. To this end it is convenient to formulate the theory on a lattice. There is no unique way to do this, but one simple and elegant procedure is to parallel the philosophy of Wilson for pure gauge theories (Wilson, 1974) and start in the broken symmetry phase, freezing the radial degree of freedom of the ϕ field. This roughly corresponds to taking $\lambda\to\infty$. The precise presciption for doing this can be found in Einhorn and Savit (1978). The result is

$$L=\sum\beta\cos(\Delta_\mu\theta_{\nu;j}-\Delta_\nu\theta_{\mu;j})+\kappa\cos(\Delta_\mu\tau_j-q\theta_{\mu;j})\,, \qquad (5.3a)$$

with the partition function

$$Z=\int_{-\pi}^{\pi}D\tau D\theta\, e^L\,. \qquad (5.3b)$$

β and κ are coupling constants related to the coupling constants in Eq. (5.1) and the lattice spacing a. q in Eq. (5.3a) is an integer which is the charge of the Higgs field in Eq. (5.1). The lattice theory has two kinds of fields or spins. On each link of the d-dimensional hypercubic lattice is a phase $e^{i\theta_{\mu;j}}$. $\theta_{\mu;j}$ is related to the original gauge field A_μ by $\theta_\mu\sim aeA_\mu$. Thus, as $a\to 0$, $-1/a\leq A_\mu\leq 1/a$, while $-\pi<\theta_\mu\leq\pi$. On each site of the lattice is another spin $e^{i\tau_j}$. This is just the phase of the original Higgs field ϕ and so we have $-\pi<\tau_j\leq\pi$ independent of a. The sum in Eq. (5.3a) runs over all plaquettes for the first term and all links for the second term. This model is therefore a kind of hybrid of the simplex number one and two models discussed in the last section, since the interaction proportional to β is an $s=2$ interaction, while the one proportional to κ has $s=1$. Note that because of the form of the κ term (5.3a) does have a local gauge invariance of the form

$$\tau_j\to\tau_j+\Lambda_j\,,\quad \theta_{\mu;j}\to\theta_{\mu;j}+(1/q)\Delta_\mu\Lambda_j\,, \qquad (5.4)$$

where Λ_j is an angle-valued gauge function.

The duality transformation for Eq. (5.3) is a relatively straightforward generalization of the technique described in the last section (Einhorn and Savit, 1978, 1979; Peskin, 1978). The result is that the dual of Eq. (5.3) in d dimensions is another hybrid theory which is Z_∞ invariant and has two interaction terms. One term, defined on a simplex of dimension $d-2$, is generated by the plaquette interaction in Eq. (5.3) and has a coupling constant which for large β is proportional to $1/\beta$. The other term associated with the link interaction of Eq. (5.3) sits on dual simplices of dimension $d-1$ and for large κ has a coupling constant proportional to $1/\kappa$. As in Eq. (5.3a), these two dual interactions are not independent of each other, but are connected by a local gauge invariance appropriate to the simplex number of the dual interaction. Note that for $d=3$, the dual of Eq. (5.3) is another (Z_∞) Abelian Higgs model in which the plaquette interaction has been generated by the link interaction of Eq. (5.3) and vice versa.

As with the pure theories of Sec. IV, the model (5.3) can be written in terms of its topological excitations. It is found that these are of two types: one type lives on closed manifolds of dimension $d-2$, and the other type lives on open manifolds of dimension $d-2$, including the boundary, which is of dimension $d-3$. For example, in three dimensions we have closed vortex strings (of dimension $d-2=1$) and open vortex strings with monopoles on the ends (the boundary in this case having dimension $d-3=0$). Roughly speaking, the strings exist because of the link interaction in Eq. (5.3) and the monopoles exist because of the plaquette interaction. Note, however, that because the terms in Eq. (5.3) are related by gauge invariance, there are no free monopoles.[10] As for the theories discussed in Sec. IV.D, the formulation of the Abelian Higgs model in terms of its topological excitations is very useful for describing the phase properties of the model. For $d>2$ and $q>1$, there is apparently, in general, a phase transition as a function of, say, β for fixed β/κ. Below the transition (β large) a Wilson loop correlation function of a "test charge" λ, with λ an integer less than q, behaves like e^{-P}, where P is the perimeter of the loop, while above the transition (β small), it behaves like e^{-A}, where A is the minimum area subtended by the loop. This is the analog of quark confinement for this theory for a quark of charge λ/q. The transition can be understood as being caused by a condensation of the topological excitations of the theory. The case of two dimensions in special, there being only one phase. Furthermore, when $q=1$, the phase structure of the model (for $d>2$) is problematical. In this case, the smallest possible test charge is $\lambda=q$, and the Wilson loop behaves like e^{-P} for all β. In addition, from a result of Osterwalder and Seiler (Osterwalder and Seiler, 1978; Fradkin and Shenker, 1979) one can show that it is possible to go analytically between the low-β and high-β regions of the theory. However, it is still possible that a phase boundary between these two regions exists, extending over some finite region of the coupling constant plane, and terminating in a critical point as in the usual PT diagram for a liquid-gas transition. More work is needed to resolve this issue.

In closing this section we note that one may construct other kinds of hybrid theories. It is possible, for example, to cross-pollinate interactions with simplex number s and $s-1$ (and so forth), and to construct such theories based on the Z_N as well as the $U(1)$ symmetry group. An especially interesting case is the Abelian Higgs model based on the Z_2 group (Balian *et al.*, 1975; Horn and Yankielowicz, 1979). For $d=3$ this model is self-dual.

B. Random systems

Duality transformations have yielded some insight in the study of random systems displaying the property of frustration. The simplest way of imposing frustration on a ferromagnetic system is to make some of the bonds antiferromagnetic. (This was described for the $d=2$ Ising model in Sec. II.A, in connection with disorder correlation functions.) Duality has been applied to such theories with a $U(1)$ and Z_2 symmetry (random x-y and Ising models) by Fradkin, Huberman, and Shenker (1978) and with a $U(1)$ symmetry by Jose (1978).

[10]It is also possible to put Eq. (5.1) on a lattice in such a way that the gauge field interaction [proportional to β in Eq. (5.3)] is not periodic. This has the consequence that there are no topological excitations on open manifolds with boundaries. For example, when $d=3$ there would be no open vortex strings with monopoles on the end—only closed vortex loops.

A rather nice picture emerges for the $U(1)$ case in two dimensions. One can show that frustrated plaquettes give rise to topological excitations of half-integer winding number. This is very easy to understand when we remember that as we circumscribe a frustrated plaquette at low temperatures, we find a mismatch in the spin orientation when we return to our starting point (see Sec. II.A). In the $U(1)$ case a configuration of spins which rotates through π as we traverse a closed contour has this property, and can also be thought of as a vortex with half-integer winding number. Like their integer-valued counterparts, these half-vortices are located on the sites of the dual lattice. They appear in pairs so that the total winding number is still integral, and at low temperatures are bound in pairs by a logarithmic potential. In three dimensions, the picture is a little more complicated, but is a natural extension of the two-dimensional case. Frustration gives rise to magnetic flux tubes of fractional flux, in addition to the integer-valued flux tubes which exist in the ordinary ferromagnetic case.

The random Ising model in two and three dimensions has properties not too dissimilar from the random x-y model. In two dimensions (as we saw in Sec. II.A) frustrations occur in pairs. At low temperatures, these are bound by a linear potential. When $d=3$, frustrations in the Ising model give rise to simple closed tubes of "frustration flux" on the dual lattice. These flux lines penetrate the frustrated plaquettes of the original random Ising model lattice.

The dual forms of these theories are helpful in analyzing the effects of frustration on various quantities of physical interest. The reader should consult Fradkin, Huberman, and Shenker (1978) and Jose (1978) for details of these calculations.

C. Topological excitations in Z_2 and Z_N symmetric theories

In our treatment of $U(1)$ symmetric theories (Sec. IV) we found that we could write the theories directly in terms of their topological excitations. Here we want to describe the topological excitations one expects in theories with a Z_N symmetry and their effect on the physics. The Z_2 case is real straightforward, but the Z_N case with $N>2$ is somewhat more complex.

In theories with a Z_2 symmetry the topological excitations are "kinks" and generalizations thereof. Consider, for instance, the $d=2$ Ising model. The model can be described in terms of domains of aligned spins. A single up spin with all its neighbors pointing down is just a very small domain. Suppose we now consider the dual lattice and draw a line along each dual lattice link which crosses an original lattice link joining two oppositely aligned spins. It is easy to see that for any configuration of spins, we will have drawn a set of closed loops. These are the domain boundaries of the Ising model. They are also the linelike topological excitations. If we think of one direction of our lattice as time, these domain boundaries can be thought of as the world-lines of kinks, or solitons (actually, in the case of a closed domain boundary, soliton-antisoliton pairs) in a one-space-dimensional world.

In three dimensions, the Ising model has domain boundaries which are closed two-dimensional surfaces (again associated with the dual lattice). Consider now the Z_2 lattice gauge theory ($s=2$) in two dimensions. This theory has pointlike topological excitations associated with the sites of the dual lattice. These appear whenever the product of the four Z_2 gauge potentials (associated with an original lattice plaquette) which surrounds the given dual lattice site is -1. Note that this is the natural extension of the Ising model. In that case, we placed a topological excitation on a dual lattice simplex associated with an original lattice link whenever the product of the spins bounding that link was -1. In the $d=3$ Z_2 gauge theory, the topological excitations become closed strings, in four dimensions closed two-dimensional surfaces, etc. From these considerations it is clear that the Z_2 invariant theory with simplex number s in d dimensions has topological excitations which are of dimension $d-s$ and live on closed manifolds on the dual lattice. Notice that the topological excitations are of one higher dimension than those that appear for the corresponding theory with a $U(1)$ symmetry. This can be understood by remembering that the spins in the Z_2 theory take their values on the zero-dimensional surface of a one-dimensional sphere (i.e., two points) while the $U(1)$ spins take their value on the one-dimensional surface of a two-dimensional sphere (a circle). A general homotopy analysis indicates that these are the first two members of a family of spins defined on the surface of an n-dimensional sphere. Each time we add one component to the spin variable, the dimension of the topological excitation decreased by one for fixed d and s.

As with the $U(1)$ invariant theories, the phase transitions in the Z_2 invariant theories can be understood as being due to a condensation of the topological excitations. This is most easily seen in the $d=2$ Ising model. In the low-temperature ordered phase, most spins are aligned and domains of misaligned spins are small and scarce, resulting in a low density of small topological loops. At the critical point, they condense into a spaghetti phase consisting of a relatively high density of arbitrarily large loops. This is the disordered phase in which nearest neighbor spins can easily be misaligned. Qualitatively similar arguments apply to other Z_2 invariant theories.

We turn now to theories with a Z_N symmetry for $N>2$. For definiteness, let us focus on the $d=2$ vector Potts model discussed in Sec. III.B. Since N is finite, it is clear that we can define domain boundaries between regions of differently oriented spins. If the angle difference between two nearest-neighbor spins is $(2\pi/N)k$, we can associate a piece of a domain boundary of strength (or flux) k with the dual lattice link which crosses the link joining the two misaligned spins. If we think of the boundary of strength k as being made up of a superposition of k boundaries of strength one, then one might suppose that these boundaries of strength one just form closed loops on the dual lattice, in analogy with the Ising model. But this is not quite correct as we shall see in a moment.

Suppose for the moment that N is very large, but finite. As N increases the vector Potts model looks more and more like the x-y model. But we know that the x-y model in two dimensions has topological excitations which are vortex points. Can we make such an excitation in the vector Potts model? The best we can do is shown schematically in Fig. 10 for the case $N=7$. The dashed circle has a diameter of order N lattice spacings. Inside that circle the Z_N spins rotate (as smoothly as possible, to try to minimize the energy) through 2π as we traverse a closed loop surrounding the dual lattice site that sits (roughly) at the center of the circle. If we look only at the spin configurations inside the circle, we will not be able to tell whether we are looking at a $U(1)$ vortex or at a configuration of Z_N spins. The reason is that the minimum energy configuration for the spins making up a vortex in the x-y model is obtained when the rotation through 2π is shared equally among the spins as we traverse a circle of any radius surrounding the vortex. Thus, for a distance from the x-y model vortex center less than N lattice spacings, nearest-neighbor spins will on the average not have angle differences less than $2\pi/N$, so the configurations will look roughly like the minimum energy configurations of the Z_N model close to the Z_N vortex center. For a distance greater than about N lattice spacings away from the vortex center the situation is different. In this region the Z_N spins cannot rotate through a small enough angle to share equally a rotation of 2π as we pass around a circle surrounding the vortex. Instead, for the Z_N model, the minimum energy configuration is effected by forming approximately wedge shaped domains of spins all of which point in the same direction. The spin orientation changes by $2\pi/N$ as we pass from one domain to its neighbor as shown in Fig. 10. In general (for example, in the vector Potts model), the domain boundaries radiating from a vortex center have a finite energy per unit length, and the energy of a single Z_N vortex increases linearly with the size of the system. However, one can produce finite energy configurations of a vortex-antivortex pair bound by N strings.

Thus the $d=2$ globally symmetric Z_N theories in general have two types of topological excitations: strings and vortices. Considering Fig. 10 it is clear that the Z_N strings only form closed loops in the absense of vortices; N units of string flux (e.g., N strings of unit flux) can coterminate in a small region around a vortex center. The extension to higher dimensions and to Z_N theories with other simplex numbers is clear: The Z_N symmetric theory in d dimensions with simplex number s has topological excitations defined on closed manifolds of dimension $d-s$ and, in addition, topological excitations defined on N (or fewer) open manifolds

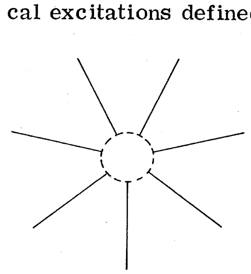

FIG. 10. Vortex in a two-dimensional globally Z_N symmetric theory such as the vector Potts model. See Sec. V.C for an explanation.

of dimension $d-s$ which coterminate on a closed manifold of dimension $d-s-1$. (These are similar, in some ways to the topological excitations of the Abelian Higgs model; see Sec. V.A.)

As with the other theories we have studied, these topological excitations have a strong influence on the phase properties of Z_N theories. For example, one can understand the two phase transitions that occur (for large enough N) in the $d=2$ vector Potts model as sequential condensations of these two types of excitations. This is discussed in more detail elsewhere (Einhorn, Savit, and Rabinovici, 1979).

D. Duality in the lattice Hamiltonian formalism

In Secs. II–IV, we have discussed duality in the context of d-dimension lattice theories. These theories can be regarded either as classical statistical mechanics systems in d space dimensions, or as Euclidean cutoff quantum field theories in d space-time dimensions. In the former case, the argument of the partition function should be thought of as the exponential of the Hamiltonian of the classical statistical system, while in the latter case the partition function's argument is the exponential of the Lagrangian of the quantum field theory. If we adopt the latter interpretation, we can rework the lattice theory so that it is expressed in terms of the Hamiltonian of the quantum field theory. (This is *not* the same as the Hamiltonian of the classical statistical system described above.) This formalism was developed by Kogut and Susskind (1975) as an alternative formulation of Wilson's lattice gauge theory (Wilson, 1974). Some aspects of duality have been studied using this method. We will first briefly describe how one develops the Hamiltonian formalism, and then remark on its use in conjunction with duality.

In the Hamiltonian formulation of latticized quantum field theory, one space-time direction is identified as time, and the lattice spacing in that direction is sent to zero. Thus the d-dimensional lattice becomes a $(d-1)$-dimensional lattice plus one continuous time axis. In order to have a chance of retaining the same (large distance) physics when the time direction lattice spacing is sent to zero, it is necessary to allow lattice coupling constants in the time direction to vary with the lattice spacing so that the effective interaction over some fixed distance is unchanged. Details may be found in Kogut and Susskind (1975) and Kogut (1979). Since time is now continuous, it is relatively straightforward to define momenta which are canonically conjugate to the coordinates (the fields) of the theory. The Hamiltonian can then be written, in the usual way, in terms of these coordinates and momenta. Thus, unlike the Lagrangian or the classical statistical mechanics Hamiltonian, this field-theoretic Hamiltonian contains noncommuting operators. As an example, let us look at the Hamiltonian derived for the $d=2$ Ising model by this method. It can be written

$$H=-\sum_i \lambda\sigma_i^{(3)}\sigma_{i+1}^{(3)}+\sigma_i^{(1)}, \tag{5.5}$$

where the Pauli matrices are

$$\sigma^{(3)}=\begin{pmatrix}1&0\\0&-1\end{pmatrix},\quad \sigma^{(1)}=\begin{pmatrix}0&1\\1&0\end{pmatrix}. \tag{5.6}$$

The matrices act on states in which a spin up is represented by $\binom{1}{0}$ and a spin down by $\binom{0}{1}$. The sum over i is a sum over points on a one-dimensional lattice and matrices acting on different sites commute. λ is a coupling parameter which is related to the original coupling constants (and temperature) of the $d=2$ Ising model. The first term in Eq. (5.5) is just an operator representation for the original nearest neighbor spin-spin interactions in the x (space) direction, while the $\sigma_i^{(1)}$ term results from the process of taking the y (time) direction lattice spacing to zero.

The Hamiltonian (5.5) can be rewritten in a new set of variables $\mu_i^{(j)}$, defined as

$$\mu_i^{(1)}=\sigma_i^{(3)}\sigma_{i+1}^{(3)}, \tag{5.7a}$$

$$\mu_i^{(3)}=\prod_{j<i}\sigma_j^{(1)}. \tag{5.7b}$$

The $\mu_i^{(j)}$'s sit on the sites of the dual one-dimensional lattice, i.e., the bonds of the original lattice. The product in Eq. (5.7b) runs from the left-hand side of the one-dimensional lattice up to the site i. [The cognoscenti will recognize this as a kind of Jordan-Wigner transformation (Jordan and Wigner, 1928).] Using Eq. (5.7), Eq. (5.5) becomes

$$H=-\lambda\sum_i \lambda^{-1}\mu_i^{(3)}\mu_{i+1}^{(3)}+\mu_i^{(1)}, \tag{5.8}$$

which, up to an overall factor of λ is the same as Eq. (5.5) with $\lambda\leftrightarrow\lambda^{-1}$. It is easy to check that the commutation relations among the μ's is the same as those among the σ's, and so this formulation of the $d=2$ Ising model is self-dual. Writing λ in terms of the original two-dimensional lattice formulation of the theory, one sees that the point $\lambda=1$ corresponds to the correct dual temperature (critical point) of the $d=2$ Ising model. (Since the Hamiltonian procedure requires from the beginning the introduction of an anisotropic two-dimensional lattice, one finds that the condition $\lambda=1$ actually generates a line of critical points in the space of the horizontal and vertical nearest-neighbor couplings.) A little reflection reveals that the μ's really can be regarded as disorder operators. For instance, an eigenstate of $\sigma_i^{(3)}$ is a state in which the spin at site i has a definite orientation, while an eigenstate of $\mu_i^{(3)}$ is a superposition of states in which all spins to the left of link (dual site) i have opposite orientation.

Similar duality transformations have been constructed in the context of the lattice Hamiltonian formalism for the Z_2 lattice gauge theory in four dimensions (Fradkin and Susskind, 1978) and the Z_N lattice gauge theory in three and four dimensions (Horn *et al.*, 1979; Green, 1978). As expected, the $d=3$ Z_N lattice gauge theories are dual to spin systems with a global symmetry, while the $d=4$ gauge theories are dual to gauge theories with a local symmetry.

E. Abelian duality in the continuum

There are two related but distinct questions to be asked about duality transformations and Abelian theories defined on a continuous manifold. First, if we start with an Abelian theory on a lattice and proceed to the continuum limit via the renormalization group, what happens to the dual representation of the theory as the limit is taken? The second question is, can we sensibly formulate the duality transformation directly in the continuum? These questions are related because they essentially involve the fate of integer-valued fields on a continuous manifold.

As for the first question, it is likely that the following scenario will be generally true: Suppose we consider some theory with more than one phase, for instance the $d=3$ x-y model. We now imagine starting our renormalization group calculation with some value of the temperature T_0. If $T_0 < T_c$, the critical temperature of the model, then we expect that as we scale to larger and larger distances (or smaller and smaller lattice spacings) the large distance physics will be that which we expect to find in the low-temperature phase of the lattice model. Alternatively, if $T_0 > T_c$, we expect that the renormalization group will lead us to a theory whose large distance physics is that of the high-temperature phase of the lattice theory. Now, the large distance structures of the two phases of the $d=3$ x-y model differ because for $T > T_c$, the closed vortex strings are condensed and can with high probability be arbitrarily large, while for $T < T_c$, only small closed vortex strings exist, and these will not strongly affect the large distance structure of the theory. Hence, we might expect that for $T_0 < T_c$, the renormalization group will lead us to a continuum theory without topological excitations (for the $d=3$ x-y model, just a theory of a massless scalar field), while for $T_0 > T_c$ the renormalization group will lead us to a continuum theory with vortex strings. Thus in this sense, the integer-valued fields should have a sensible continuum definition in this phase. Renormalization group studies of the $d=2$ x-y model have been done by José *et al.* (1977). The reader is referred to that paper and references therein for more insight.

We turn now to the second question, whether a dual transformation can be performed directly for a theory defined in the continuum. Whatever problems arise here are not related to any intrinsic difficulty of defining duality in the continuum, but, as before, are related to the problem of handling integer-valued fields in the continuum. To see this graphically, we note that one can easily construct the dual form of a continuum free field theory. For instance, the generating functional for free photons in three Euclidean dimensions can be rewritten as the generating functional for a massless scalar field in three dimensions, following the by now familiar routine of Fourier transformation. On the other hand, one quickly runs into trouble trying to work with a theory with integer-valued fields, for example, some sort of analog of the Villain approximation (say, for a scalar theory) defined in the continuum. Although one may formally proceed by simply replacing discrete differences by derivatives, one does not really know what one is doing: for instance, if $n(x)$ is an integer-valued field, what does $\partial_x n(x)$ mean? To have a chance of controlling these objects some sort of short distance cutoff is evidently necessary. Of course, such a cutoff is provided by the lattice formulation.

These comments notwithstanding, it is sometimes possible to use duality arguments profitably in continuum models with vortexlike structures. An example is the paper by Sugamoto (1979) in which he studies the Abelian Higgs model in four dimensions in the continuum. His duality transformation is related to, but is not identical with the one discussed by Einhorn and Savit (1978, 1979) and by Peskin (1978) for the latticized Abelian Higgs model. Sugamoto introduces a Fourier variable conjugate to $F_{\mu\nu}$ and then integrates over the vector potential A_μ. Since the Lagrangian still contains terms quadratic in A_μ (coupled to the Higgs scalar field) integrating over A_μ does not produce the familiar delta functions; instead we are left with a kind of magnetic field strength formalism. Sugamoto describes a classical solution for this formulation of the model and discusses its relation to the Nielsen-Olesen vortex string. However, questions of renormalization and the ultimate fate of these classical solutions are not treated. See his paper for more details.

F. Approaches to duality for non-Abelian theories

Many important theories in condensed matter and high-energy physics have non-Abelian symmetries, and so it is natural to try to generalize duality transformations which have been so successful for Abelian theories to non-Abelian theories. Unfortunately, one encounters tremendous difficulties in the program. Here we will briefly describe some of the work that has been done on these problems.

Let us first discuss the most straightforward generalization of Abelian duality. To illustrate the procedure, consider the two-dimensional 0(3) Heisenberg ferromagnet. We place on each site of a square lattice a "spin" which takes its value on the surface of a sphere. The Hamiltonian of the system is

$$H = -J \sum_{\langle\,\rangle} \cos(\Delta_\mu \Omega_i)\,, \tag{5.9}$$

where the sum runs over all nearest-neighbor pairs on the lattice, and $\Delta_\mu \Omega_i$ is the angle between the orientations of the nearest-neighbor spins. The partition function is

$$Z = \int D\Omega \exp\Big(\beta \sum_{\langle\,\rangle} \cos(\Delta_\mu \Omega_i)\Big)\,, \tag{5.10}$$

with $\beta = J/kT$. Z may also be considered to be the generating functional (on a square lattice) for the 0(3) nonlinear sigma model in two Euclidean dimensions with βH playing the role of the Lagrangian.

For Abelian theories we recall that the first step in constructing the duality transformation was to Fourier expand the exponential of the interaction. For the theory (5.10), the natural set of Fourier expansion variables is the set of spherical harmonics $Y_{l,m}(\Delta\Omega)$. Writing Eq. (5.10) as a product over all lattice links of exponentials of the interaction in the usual way one can carry out such an expansion. Next, one can use the addition formula for spherical harmonics to write $Y_{l,m}(\Delta_\mu \Omega_i)$ as a sum of products of spherical harmonics that depend separately on Ω_i and $\Omega_{i-\hat{\mu}}$. Rearranging the factors, Eq. (5.10), can be written in the form

$$Z=\sum_{\{l,m\}}F(\beta;\{l,m\})\prod_i\int D\Omega_i\prod_{j=1}^{4}Y_{l_j,m_j}(\Omega_i). \qquad (5.11)$$

F is essentially a product over the lattice links of the coefficients of the Fourier expansion in $Y_{l,m}$'s, times some factors coming from the use of the addition formula. The last factor in Eq. (5.11) is a product over all lattice sites of integrals over the spin orientation. For each integral there are four factors of $Y_{l,m}$ because there are four lattice links that impinge on each lattice site. Equation (5.11) should be compared with, say, Eq. (4.5) for the x-y model. The point is that the integral over the θ_i in Eq. (4.5) gives a set of delta functions which allow us to use in Eq. (4.5) the representation (4.6) for the integers k. In Eq. (5.11), on the other hand, no such simple result emerges: the integral over Ω gives constraints on the set of four l's and m's, but nothing as simple as a delta function (except, of course, for the m's). It is thus difficult to find a representation for the integers l, which will automatically satisfy the constraints imposed by the integrals over Ω_i.

Despite the failure of this approach to produce an elegant dual theory in the usual sense, one does generate a representation for Z in terms of the l's and m's which is relatively simple at high temperatures. The key to understanding this representation is the observation that since the spins are almost completely disordered at very high temperatures, they are all predominantly in a relative s state, and so the dominant contribution to Z will be from values of l near zero. Hence it is possible to do a kind of configuration expansion in l and m about $\{l=0, m=0\}$ (Savit, 1977b). This expansion is related to (but is not the same as) the high-temperature expansion of statistical mechanics.

The complication described for the 0(3) case persists generally for non-Abelian theories regardless of dimension or whether the theories have a global or a local symmetry. However, in the special case of solvable non-Abelian groups, Drouffe, Itzykson, and Zuber (Drouffe *et al.*, 1979) have shown that for a globally symmetric theory of spins with nearest neighbor interactions it is possible to construct a dual theory. The procedure is to first show that the theory with a solvable non-Abelian symmetry is equivalent to a theory with an Abelian symmetry, and then to apply a duality transformation to the Abelian theory. Bellisard (1978) has also demonstrated this transformation for the special case of the group S_3. Unfortunately, the procedure is apparently not useful for (lattice) gauge theories.

For theories with a non-Abelian symmetry group which is not solvable, a number of workers have tried to construct some sort of dual representation using a variety of approaches. Most of the work we shall describe has been done for gauge theories. A major motivation has usually been to find a simple calculable representation for the theory in the strong coupling regime. However, some authors have suggested that a *simple* dual representation may not exist.

The first paper we shall mention is one to which other papers (particularly those of 't Hooft, Mandelstam, and Englert and Windey) may be related. This is the work of Goddard, Nuyts, and Olive (Goddard *et al.*, 1977) who have noted that for any of the classical Lie groups there exists a dual group whose generating algebra is associated with an inversion of the root diagram of the algebra of the original group. They suggest that in a gauge theory this dual group may be the symmetry group appropriate to a description of the theory in terms of its magnetic degrees of freedom. In the specific case of the Georgi-Glashow model, Montonen and Olive (1977) have presented circumstantial evidence to support the conjecture that (at least in some limit) the monopole of the theory may be the gauge boson of the theory when expressed in terms of its dual (magnetic) degrees of freedom. Unfortunately, one has not yet been able to complete the program and express the theory directly in terms of its dual degrees of freedom, but the evidence gathered so far in support of the conjecture is tantalizing.

Halpern's approach to duality for non-Abelian gauge theories involves two steps (Halpern, 1979). First, he argues that with a completely fixed axial gauge the vector potential A is a unique function of the field strength G. He then shows that with a completely fixed axial gauge the generating functional for a gauge theory can be written in the form

$$Z=\int DG\,\delta(I(G))\exp\left(-\frac{1}{4}\int G^2d^dx\right) \qquad (5.12a)$$

$$=\int DG\,DB\exp\left(-\frac{1}{4}\int G^2+iBI(G)\right). \qquad (5.12b)$$

$I(G)$ is the Bianchi identity for G with the vector potential A regarded as a function $A(G)$. Equation (5.12a) follows from Eq. (5.12b) by integrating over B. Finally, Eq. (5.12b) can be integrated over G to obtain Z purely in terms of B. It is the field B in Eq. (5.12b) that Halpern identifies as the dual potential. He argues that B can be interpreted as a potential for a kind of disorder field, as one should expect for a dual variable. Many more properties of the representation (5.12) are contained in his paper which the reader should consult for more details.

Another approach to non-Abelian duality which relies heavily on first-order formalism is the paper by Seo, Masonori, and Sugamoto (Seo *et al.*, 1979). This paper is essentially a generalization of the paper by Sugamoto (1979) described in Sec. V.E. In it two non-Abelian Higgs models, one with a single Higgs doublet, the other with two Higgs triplets, are studied. The dual transformation in this paper is defined by introducing a variable conjugate to $F^a_{\mu\nu}$, and writing

$$\exp\left(-\frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}d^dx\right)$$
$$\propto\int DW^a_{\mu\nu}\exp\left(-\int W^a_{\mu\nu}W^a_{\mu\nu}+iF^a_{\mu\nu}\tilde{W}^a_{\mu\nu}d^dx\right). \qquad (5.13)$$

The functional integral of the Higgs model will now be quadratic in the vector potential A^a_μ, and so the integration of A^a_μ can be done. What results is a functional integral over the exponential of a rather complicated Lagrangian which is a function of the fields $W^a_{\mu\nu}$, the Higgs fields and their derivatives. This formulation is of interest because there is some hope that one could define a sensible strong coupling expansion using it.

Another approach to a dual theory has been studied by Kazama and Savit (1979). They start from the observation that, in the dual transformation for an Abelian theory, integrating over the original variable produces a delta function which enforces a kind of "Bianchi identity." For example, the delta functions in Eq. (4.5) tell us that the $k_{\mu;i}$ must be expressible as a curl. To mimic this in a non-Abelian gauge theory one first uses the identity (5.13) so that the functional integral is quadratic in A_μ^a. Next, one isolates the terms quadratic in A_μ^a and Fourier transforms these again with respect to A_μ^a. Now the theory contains two Fourier variables, but the functional integral is linear in A_μ^a. Performing the integral over A_μ^a produces functional delta functions which have the form of a kind of Bianchi identity. This procedure results in a form with some very intriguing properties one of which is the possibility of defining a strong coupling expansion. Details will be reported elsewhere.

Finally, we briefly mention the work of 't Hooft (1978, 1979) and Mandelstam (1978). Work along similar lines has also been done by Englert and Windey [Englert and Windey (1978) and references therein]. These authors study non-Abelian gauge theories and try to construct magnetic operators which are dual to gauge invariant operators of the electric degrees of freedom. The latter may be taken to be the Wilson loop operators (Wilson, 1974). In four dimensions the dual (magnetic) operators are certain kinds of Nielsen-Olesen vortex loops associated with the non-Abelian theory. By examining the commutation relations between the magnetic flux loops and the Wilson loops, 't Hooft and Mandelstam, in their related, but not identical approaches, delineate possible phases in which the non-Abelian gauge theory could find itself. One of these phases is the long-sought quark confinement phase. Whether this phase is in fact realized depends on the dynamics of the theory. Unfortunately, both 't Hooft and Mandelstam conclude that the non-Abelian gauge theory expressed in terms of their magnetic vortex string operators (and the corresponding dual vector potentials) is terribly complicated. Thus, while one may gain much insight from these studies, the magnetic variable may not be computationally useful, even for strong coupling.

ACKNOWLEDGMENTS

This review is based in part on lectures given at the University of Michigan during the winter of 1979. I thank my colleagues there and elsewhere for many interesting comments.

REFERENCES

Anderson, P. W., and S. F. Edwards, 1975, J. Phys. F **5**, 965.

Balian, R., J. M. Drouffe, and C. Itzykson, 1975, Phys. Rev. D **11**, 2098.

Banks, T., R. Meyerson, and J. Kogut, 1977, Nucl. Phys. B **129**, 493.

Banks, T., and E. Rabinovici, 1979, IAS preprint.

Bellisard, J., 1978, CNRS Marseille preprint No. 78/p. 1024.

Berezinskii, V. L., 1970 Zh. Eksp. Teor. Fiz. **59**, 907 [Sov. Phys.-JETP **32**, 493 (1971)].

Berezinskii, V. L., 1972 Zh. Eksp. Teor. Fiz. **61**, 1144 [Sov. Phys.-JETP **34**, 610 (1972)].

Bricmont, J., J. Lebowitz, and C. Pfister, 1979, Rutgers University preprint, to appear in Ann. N. Y. Acad. Sci.

Cardy, J., 1978, UCSB preprint No. TH-30.

Casher, A., 1978, DESY preprint 78/43.

Chui, S. T., and J. D. Weeks, 1976, Phys. Rev. B **14**, 4978.

Drouffe, J. M., 1978, Phys. Rev. D **18**, 1174.

Drouffe, J. M., C. Itzykson, and J. B. Zuber, 1979, Nucl. Phys. B **147**, 132.

Einhorn, M., and R. Savit, 1978, Phys. Rev. D **17**, 2583.

Einhorn, M., and R. Savit, 1979, Phys. Rev. D **19**, 1198.

Einhorn, M., R. Savit, and E. Rabinovici, 1979, University of Michigan preprint no. 79-25.

Elitzur, S., R. B. Pearson, and J. Shigemitsu, 1979, IAS preprint.

Englert, F., and P. Windey, 1978, Université Libre de Bruxelles preprint.

Fontaine, J., and C. Gruber, 1978, "Surface Tension and Phase Transition for Lattice Systems," preprint to appear in Commun. Math. Phys.

Fradkin, E., B. A. Huberman, and S. Shenker, 1978, SLAC-PUB-2112.

Fradkin, E., and S. Shenker, 1979, Phys. Rev. D **19**, 3682.

Fradkin, E., and L. Susskind, 1978, Phys. Rev. D **17**, 2637.

Gallavotti, A., F. Guerra, and S. Miracle-Sole, 1978, in Vol. 80 of "Lecture Notes in Physics" (Springer, Berlin).

Glimm, J., and A. Jaffe, 1977, Phys. Lett. B **66**, 67.

Goddard, P. Nuyts, and D. Olive, 1977, Nucl. Phys. B **125**, 1.

Goddard, P., and D. I. Olive 1978, Rep. Prog. Phys. **41**, 1357.

Green, M. B., 1978 Nucl. Phys. B **144**, 473.

Gruber, C., A. Hintermann, and D. Merlini, 1977, *Group Analysis of Classical Lattice Systems*, Vol. 60 of "Lecture Notes in Physics" (Springer, Berlin).

Halperin, B. I., and D. R. Nelson, 1978, Phys. Rev. Lett. **42**, 121.

Halpern, M. B., 1979, Phys. Rev. D **19**, 517.

't Hooft, G., 1978, Nucl. Phys. B **138**, 1.

't Hooft, G., 1979, University of Utrecht preprint.

Horn, D., M. Weinstein, and S. Yankielowicz, 1979, Tel-Aviv University preprint no. TAUP 723-79.

Horn, D., and S. Yankielowicz, 1979, Tel-Aviv University preprint no. TAUP 746-79.

Israel, R., and C. Nappi, 1979, Commun. Math. Phys. **64**, 177.

Jones, D., J. Kogut, and D. Sinclair, 1979, Phys. Rev. D **19**, 1882.

Jordan, P., and E. Wigner, 1928, Z. Phys. **47**, 631.

Jose, J., L. P. Kadanoff, S. Kirkpatrick, and D. R. Nelson, 1977, Phys. Rev. B **16**, 1217.

Jose, J., 1978, University of Chicago preprint.

Kadanoff, L. P., and H. Ceva, 1971, Phys. Rev. B **3**, 3918.

Kadanoff, L. P., 1978, J. Phys. A **11**, 1399.

Kalb, M., and P. Ramond, 1974, Phys. Rev. D **9**, 2273.

Kazama, Y., and R. Savit, 1979, University of Michigan preprint no. 79-39.

Kogut, J. B., and L. Susskind, 1975, Phys. Rev. D **11**, 395.

Kogut, J. B., 1979, Rev. Mod. Phys. **51**, 659.

Korthes-Altes, C. P., 1978, CNRS Marseille preprint no. 79/p. 1003.

Kosterlitz, J. M., and D. J. Thouless, 1973, J. Phys. C **6**, 1181.

Kramers, H. A., and G. H. Wannier, 1941, Phys. Rev. **60**, 252.

Luther, A., and D. Scalapino, 1977, Phys. Rev. B **16**, 1153.

Mandelstam, S., 1978, Berkeley preprint No. UCB-PTH-78/6.

Mermin, N. D., and H. Wagner, 1966, Phys. Rev. Lett. **17**, 1133.

Mittag, L., and M. J. Stephen, 1971 J. Math. Phys. **12**, 441.
Montonen, C., and D. Olive, 1977 Phys. Lett. B **72**, 117.
Osterwalder, K., and E. Seiler, 1978, Ann. Phys. (NY) **110**, 440.
Peskin, M., 1978, Ann. Phys. (NY) **113**, 122.
Polyakov, A. M., 1975, Phys. Lett. B **59**, 82.
Polyakov, A. M., 1977 Nucl. Phys. B **120**, 429.
Potts, R. B., 1952, Proc. Cambridge Philos. Soc. **48**, 106.
Savit, R., 1977a, Phys. Rev. Lett. **39**, 55.
Savit, R., 1977b, unpublished.
Savit, R., 1978, Phys. Rev. B **17**, 1340.
Savit, R., 1980, University of Michigan preprint No. UM-HE 80-3.
Seo, K., O. Masonori, and S. Sugamoto, 1979, University of Tokyo preprint No. INS-REP-329.
Stone, M., and P. R. Thomas, 1978, Phys. Rev. Lett. **41**, 351.
Sugamoto, A., 1979, University of Tokyo preprint No. INS-REP-319, Phys. Rev. D, to be published.
Syozi, I., 1972, "Transformation of Ising Models," in *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic, New York), Vol. 1, p. 270.
Ukawa, A., P. Windey, and A. Guth, 1979, Princeton University preprint.
Villain, J., 1975, J. Phys. C **36**, 581.
Wegner, F., 1971, J. Math. Phys. **12**, 2259.
Wegner, F., 1973, Physica **68**, 570.
Wilson, K. G., 1974, Phys. Rev. D **10**, 2445.
Yoneya, T., 1978, CCNY preprint.