BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 44, Number 3, July 2007, Pages 361–391
S 0273-0979(07)01167-6
Article electronically published on April 23, 2007

FROM SUPERCONDUCTORS AND FOUR-MANIFOLDS TO WEAK INTERACTIONS

EDWARD WITTEN

Abstract. The goal of this article is to describe the concept of "gauge symmetry breaking" and its applications to superconductors, four-manifold theory, and particle physics.

1. Introduction

What is now understood as gauge symmetry breaking first appeared in 1950 in the Landau-Ginzburg model of superconductivity [7]. From a modern point of view, that model is largely correct as a macroscopic description of a superconductor, but it took a long time for this to be clear. Schwinger [16] in 1962 was the first to propose what we now think of as gauge symmetry breaking in the relativistic case, albeit from a somewhat abstract point of view. Anderson [1] the following year linked the two problems, showing that superconductivity gives a real-world illustration of Schwinger's ideas. In 1964, Higgs [12], as well as Englert and Brout [6], described concrete and straightforward relativistic models of electroweak symmetry breaking, analogous to the Landau-Ginzburg model of superconductivity. These models had much impact because of their simplicity.

As for what they might be good for, early authors assumed that the applications would be to the problem of understanding the strong interactions (which, among other things, bind protons and neutrons to form atomic nuclei). This turned out to be not quite right; the strong interactions are indeed described by a quantum gauge theory, as became clear in the 1970's, but the key ideas are different ones, more difficult to describe in a classical language and beyond the scope of the present article. Instead it turned out via the work of Weinberg [22] and Salam [15] (which improved on previous constructions by Glashow [9] and Salam and Ward [14] with partial symmetry) that gauge symmetry breaking is a key idea in understanding the weak interactions. The weak interactions, which are responsible among other things for the radioactive decay of certain atoms, are, roughly speaking, one of the three main forces in the atomic world, along with electromagnetism and the strong interactions.

Related ideas also appear in mathematics, an assertion that I will illustrate with one important example. Taubes [20, 21] a decade ago obtained spectacular results about symplectic four-manifolds by using a slightly perturbed version of

Received by the editors February 26, 2007.
2000 *Mathematics Subject Classification.* Primary 81Q99, 53D45.
Key words and phrases. Superconductors, weak interactions, four-manifolds.
Supported in part by NSF Grant PHY-0503584.

the Seiberg-Witten equations. Though the equations studied by Taubes do not precisely coincide with those of Landau and Ginzburg, they are close enough so that a familiarity with the theory of superconductivity is excellent background for understanding the work of Taubes.

This article[1] will aim at an introduction to gauge symmetry breaking in superconductivity, four-manifold theory, and weak interactions. We discuss superconductors in section 2, four-manifolds in section 3, and the weak interactions in section 4. We are able to give brief introductions to superconductivity and weak interactions because we concentrate on aspects that do not depend deeply on quantum mechanics. As Anderson wrote in [1], "the quantum nature of the gauge field is irrelevant" for the basic idea of symmetry breaking.

As our brief historical sketch has probably indicated, the ideas that will be surveyed here are not new, but they are still fresh. That is so in all three areas and is certainly true in particle physics. Although most aspects of the Standard Model of known elementary particle phenomena are well-tested experimentally, the mechanism of gauge symmetry breaking is a conspicuous exception. In fact, learning how this symmetry breaking comes about is one of the main goals of the Large Hadron Collider (LHC), the new accelerator that will operate at CERN, the European Laboratory for Particle Physics, starting at the end of 2007.

2. Superconductivity

2.1. Classical Electromagnetism.

Classically, electromagnetism is described in terms of the electric and magnetic fields $\vec{E}$ and $\vec{B}$, which are one-forms or vector fields on space, that is on $\mathbb{R}^3$. ($\mathbb{R}^3$ is endowed with the Euclidean metric, so one-forms can naturally be associated with vector fields.) Relativistically, it is convenient to combine $\vec{E}$ and $\vec{B}$ to a two-form F on spacetime.

In special relativity, spacetime is Minkowski spacetime $\mathbb{R}^{3,1}$, endowed with a flat pseudo-Riemannian metric of the indicated signature. Picking a time coordinate t and spatial coordinates $\vec{x} = (x^1, x^2, x^3)$, we write $ds^2 = -dt^2 + (d\vec{x})^2$ for the metric. Then $\vec{E}$ and $\vec{B}$ combine naturally to the two-form

$$F = \vec{E} \cdot d\vec{x} \wedge dt + \frac{1}{2}\vec{B} \cdot d\vec{x} \times d\vec{x}. \tag{2.1}$$

Another way to say this is that $\vec{E}$, regarded as a one-form, is $-\iota_{\partial/\partial t} F$, while $\vec{B}$ is obtained by restricting F to a slice of fixed time and then applying the three-dimensional Hodge star operator.

Maxwell's equations read

$$\begin{aligned} d \star F &= J \\ dF &= 0, \end{aligned} \tag{2.2}$$

where J is the electromagnetic current and $\star$ is the Hodge star operator. In the nineteenth century, it was found convenient, as a step toward solving these equations, to introduce a "vector potential". By the 1920's, with the advent of quantum

[1] Earlier versions were presented in lectures at the Fields Institute, the 90th Birthday Celebration of I. M. Gelfand at Rutgers University, and the Women and Mathematics program at the IAS. All audiences made very helpful comments.

mechanics, a formulation using the vector potential was not just convenient but necessary; without the vector potential, one cannot write a Schrödinger equation for an electron in a magnetic field.

From a modern point of view, the vector potential is a connection, which we will call A, on a unitary complex line bundle $\mathcal{L}$. F is interpreted as the curvature of the connection; thus if A is represented locally as a one-form, we have $F = dA$. The second Maxwell equation, $dF = 0$, is the Bianchi identity. The first Maxwell equation, $d \star F = J$, is regarded as the dynamical equation. It determines, for given initial data and up to a gauge transformation, the time dependence of the connection. This equation can be interpreted as the Euler-Lagrange equations for the following action functional:

$$I = \int_{\mathbb{R}^{3,1}} \left(-\frac{1}{2} F \wedge \star F + A \wedge J \right). \tag{2.3}$$

In other words, the condition for a stationary point of I is the Euler-Lagrange equation $d \star F = J$.

We define a gauge-covariant extension d_A of the exterior derivative d by saying that if w is a section of $\mathcal{L}$, then[2] $d_A w = (d - ieA)w$. Here e is a real constant, the electric charge of the electron (customarily considered to be negative); this factor could be absorbed in the definition of A by setting $\widehat{A} = eA$, but for the moment we will not do so. Here, I should note that in conventional units, the definition of d_A reads

$$d_A w = \left(d - \frac{ieA}{\hbar c} \right) w, \tag{2.4}$$

where $\hbar$ and c are Planck's constant and the speed of light. We work primarily in units with $\hbar = c = 1$.

2.2. **The Landau-Ginzburg Model.** Now many materials, including familiar metals such as lead, tin, and aluminum, become "superconducting" at low temperatures (typically a few degrees Kelvin).[3] As will become clear, the key property of a superconductor is the following. To describe the macroscopic state of a superconductor, in addition to the obvious macroscopic variables such as temperature, pressure, magnetic field, and the like, we need also a section s of the line bundle $\widehat{\mathcal{L}} = \mathcal{L}^2$. Moreover, superconductivity occurs when it is energetically favored to have $|s| = a$ and $d_A s = 0$, where here a is a positive constant that depends on the material as well as on the temperature and other variables.

It probably seems rather mysterious *why* such an abstract-sounding entity as s would appear along with the more obvious variables in a macroscopic description of a piece of lead. The answer to this question emerged from the microscopic theory of superconductors, for which Bardeen, Cooper, and Schrieffer [2] won the 1972 Nobel Prize. Our purpose here, however, is not to explain why s is needed in the macroscopic description of a superconductor, but to explain the peculiar properties that follow from this.

[2] Physicists customarily think of A as a real one-form, and then unitarity of the connection requires the factor of $i = \sqrt{-1}$.

[3] For more on some of the following and a treatment of topics that are omitted here, including the Josephson effect and the microscopic BCS theory of superconductivity, the reader may wish to consult section 21.6 of [23].

The fact that s is a section of the square of $\mathcal{L}$, rather than of $\mathcal{L}$ itself, is important in the microscopic theory (it reflects the fact that a "Cooper pair" is a bound state of two electrons). But it will not be important for us. For our purposes, we can just think of $\widehat{\mathcal{L}}$ as the fundamental line bundle of electromagnetism. However, we will retain the factor of 2 in the key formulas, writing $2A$ for the connection on $\widehat{\mathcal{L}}$ and writing the covariant differential as $d_A s = (d - 2ieA)s$.

Under what conditions will it be true that the energy is minimized for $|s| = a$ and $d_A s = 0$? The most obvious possibility is that the energy depends on s via the Landau-Ginzburg functional:

$$I(s) = \int_{\mathbb{R}^3} d^3x \left(\frac{1}{2} |d_A s|^2 + \frac{\gamma}{2} \left(|s|^2 - a^2 \right)^2 \right). \tag{2.5}$$

This certainly has its minimum for $|s| = a$ and $d_A s = 0$, assuming that $\gamma > 0$. We have scaled s by a positive real factor so that the coefficient of $|d_A s|^2$ is precisely $1/2$; the coefficient of the $(|s|^2 - a^2)^2$ term then involves the new constant γ. There is no claim here that the energy function $I(s)$ gives a particularly good description of a superconductor; it is simply the most obvious way to make a model in which the energy is minimized for $|s| = a$, $d_A s = 0$. All such models lead to superconductivity.

2.3. **The Meissner Effect.** To have $s \neq 0$ with $d_A s = 0$ means that s gives a covariantly constant trivialization of the line bundle $\widehat{\mathcal{L}}$, and in particular it implies that $\widehat{\mathcal{L}}$ is flat when restricted to the interior of the superconducting material. Explicitly, $d_A s = 0$ implies that $0 = d_A^2 s = -2ieFs$, so that the curvature F vanishes wherever s is non-zero. This in particular gives us the Meissner effect: the magnetic field vanishes deep inside a superconductor. (Near the surface of the superconductor there are edge effects that we will discuss more carefully in subsection 2.9. The electric field also vanishes deep inside a superconductor, but this is less dramatic as it occurs for ordinary conductors.) If a lead sphere is placed in a constant magnetic field, then at high temperatures the magnetic field is essentially unaffected by the lead, but below a temperature of 7.2 degrees Kelvin, the lead sphere becomes superconducting and the magnetic field is expelled, as sketched in fig. 1.

This explanation of the Meissner effect is reasonable only if the magnetic field is not too strong. Let $\mathcal{V}$ be the volume of the sphere in fig. 1. If we set $s = 0$

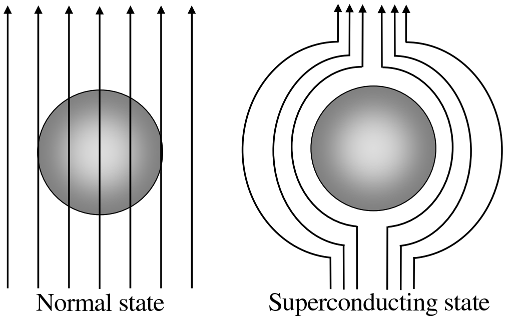

FIGURE 1. A lead sphere in a magnetic field above the superconducting temperature (left) and below (right). On the right, the magnetic field, indicated by the arrows, is expelled from the sphere.

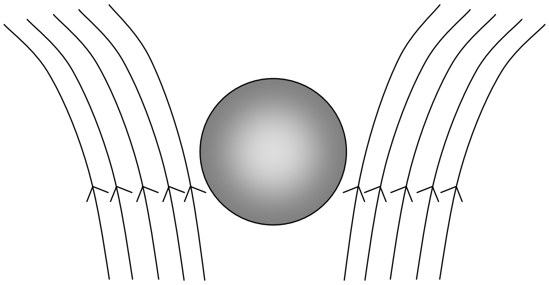

FIGURE 2. A superconducting material is repelled from a region in which the magnetic field is strong. This can lead to magnetic levitation.

throughout the sphere, then the magnetic field can freely penetrate. According to (2.5), this has an energetic cost $\gamma a^4\mathcal{V}/2$. On the other hand, if we take $|s| = a$ and $d_A s = 0$, the magnetic field must be expelled from the sample; the energetic cost of expelling a magnetic field B from a volume $\mathcal{V}$ is of order $|B|^2\mathcal{V}/2$. The Meissner effect will occur if the energetic cost of expelling the field is less than the cost of setting s to zero, or in other words if B is less than the critical magnetic field $B_{\text{crit}} = \gamma^{1/2}a^2$ (which typically is of order 10,000 times the magnetic field of the Earth and relatively close to the strongest available magnetic fields). What happens when B exceeds B_{crit} is a little more complicated than one might expect from this explanation; we return to this later.

Because the energy required to expel the magnetic field from a superconductor is proportional to $|B|^2$, this energy grows when B is increased. As a result, a superconductor is repelled from a region in which the magnetic field is large. This leads to the phenomenon of magnetic levitation (fig. 2), which is commonly seen in science demonstrations and is used in maglev trains, such as the trains running to and from the Shanghai airport.

Now let us understand why a material with the properties that we have assumed is a superconductor. We actually can see this directly from the Meissner effect. On the left-hand side of fig. 1, we have an ordinary conductor in a magnetic field that is created by unspecified external currents. On the right-hand side of fig. 1, the external currents are the same, but the configuration of the magnetic field is different. This results from electric currents that flow in the superconductor in such a way that inside the superconductor the combined magnetic field, due to the external currents and those in the superconductor, will vanish. The electric currents that flow in the superconductor will flow indefinitely, since it is energetically favored to maintain the Meissner effect. The fact that electric currents flow indefinitely with no energy required to maintain them is what we call superconductivity.

Maxwell's equations, which in a time-independent situation read $\vec{\nabla} \times \vec{B} = \vec{J}$, show that as $\vec{B}$ vanishes deep inside the superconductor, $\vec{J}$ must also. Hence the superconducting currents are carried on the surface of the material. We will be more precise about this later.

2.4. **A Superconducting Ring.** However, it is instructive to also consider another geometry, one in which superconducting current flows in a homotopically

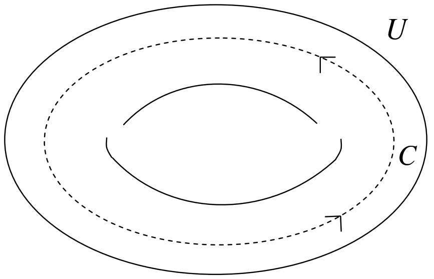

FIGURE 3. A ring U filled with superconducting material. The circle $C \subset U$ (indicated by the dotted line) is homotopically non-trivial in U but is the boundary of a disc $D \subset \mathbb{R}^3$ (not drawn here).

non-trivial path. In addition to being closer to most applications of superconductors, this is conceptually simpler, with no need to consider an externally applied field.

In fig. 3, we sketch a ring U of superconducting material. A circle running around the ring, deep inside the superconductor, has been labeled C. C represents a non-trivial homology class in U, but its homology class in $\mathbb{R}^3$ is, of course, zero. So C is the boundary of a disc D in $\mathbb{R}^3$, though we cannot take D to lie in U.

The line bundle $\widehat{\mathcal{L}}$ restricted to D is topologically trivial because D is contractible. However, the section s of $\widehat{\mathcal{L}}$ is non-zero inside U and hence gives a trivialization of $\widehat{\mathcal{L}}$ over C, that is, over the boundary of D. The pair $(D, \partial D)$, that is, D with its boundary collapsed to a point, is equivalent to S^2. A complex line bundle over S^2 has a first Chern class. So $\widehat{\mathcal{L}}|_D$ (that is, $\widehat{\mathcal{L}}$ restricted to D) has a "relative first Chern class" $c_1(\widehat{\mathcal{L}}, s)$ defined relative to the trivialization s on the boundary of D. We set $n = c_1(\widehat{\mathcal{L}}, s)$. n can be measured by integrating the magnetic flux through D. In conventional units,

$$n = 2 \int_D \frac{eB}{2\pi\hbar c}, \tag{2.6}$$

where the factor of 2 in front reflects the fact that $\widehat{\mathcal{L}} = \mathcal{L}^2$.

Since n is an integer, it cannot change in time as long as it is true that s is everywhere non-zero deep inside the superconductor. If an isolated superconducting ring, far from any relevant charges and currents, is in a state with $n \neq 0$, it will soon settle down to its state of lowest energy with the given value of n. In view of eqn. (2.6), a state with $n \neq 0$ necessarily has a non-zero magnetic field. Maxwell's equations tell us that to have a non-zero magnetic field, there must be non-zero electric currents. In the present case, these currents will be contained in the superconductor itself (and in fact on its surface), since we assume it to be isolated. To get a non-zero flux through the ring, the currents must flow on a homotopically non-trivial path around the ring U (fig. 4). The currents are proportional to n, with the constant of proportionality (and the details of where on the surface of U the currents flow) depending on the detailed geometry.

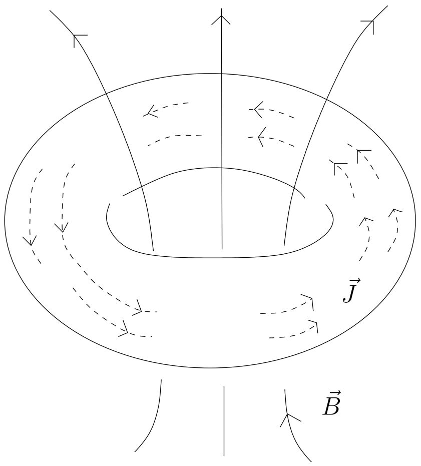

FIGURE 4. A magnetic flux through the ring arises, via Maxwell's equations, from a flow of current around the ring. The current, labeled $\vec{J}$, runs on the surface of the ring.

To the extent that n is truly constant, the electrical currents that sustain the magnetic field will persist forever, with no external input of energy. This again is superconductivity.

2.5. **Flux Lines.** The reason that n eventually will change is that although there is a cost in energy to have $s = 0$, this will sometimes happen by thermal or quantum fluctuations. So let us consider a superconductor in which $s = 0$ somewhere. As s is a section of a complex line bundle, by transversality, we expect to have $s = 0$ in real codimension two, as sketched in fig. 5. Thus, we expect s to vanish on a line or one-manifold in space. Lines on which $s = 0$ are of great scientific and even technological importance and are known as Abrikosov-Gorkov flux lines.

A flux line L is characterized by an integer according to essentially the same reasoning that we used to analyze the superconducting ring. As sketched on the

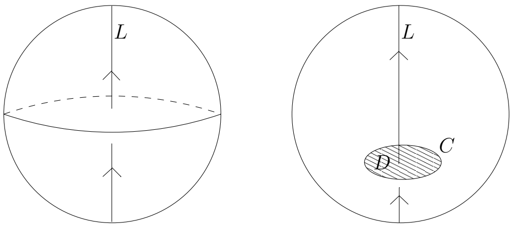

FIGURE 5. Shown on the left is a superconducting sphere with a line L – known as a flux line – along which $s = 0$. On the right we sketch a circle C that encloses the flux line. It is the boundary of a disc D that intersects L transversely at a single point.

right of fig. 5, we consider a circle C that is deep inside the superconducting material, and disjoint from L, so that $s \neq 0$ everywhere on C. We further choose C to "link" L, so that a disc D with boundary C intersects L transversely at a single point. The line bundle $\widehat{\mathcal{L}}$, when restricted to D, has a relative first Chern class $n = c_1(\widehat{\mathcal{L}}, s)$, which characterizes the flux line. It can be evaluated by the integral in eqn. (2.6).

We describe the flux line mathematically as a local minimum of the appropriate energy function. In the Landau-Ginzburg model, the energy function is taken to be the sum of the energy function for s, which was given in eqn. (2.5), and the magnetic energy, which is the integral of $|B|^2/2$. Thus the function that must be minimized is

$$V(s,B) = \int_{\mathbb{R}^3} d^3x \left(\frac{|B|^2}{2} + \frac{1}{2}|d_A s|^2 + \frac{\gamma}{2}\left(|s|^2 - a^2\right)^2 \right). \tag{2.7}$$

Making this expression stationary gives the Euler-Lagrange equations of Landau-Ginzburg theory:

$$\begin{aligned} d_A^* d_A s + 2\gamma s(|s|^2 - a^2) &= 0 \\ d\star F + ie\star(\bar{s} d_A s - (d_A \bar{s})s) &= 0. \end{aligned} \tag{2.8}$$

(Here $d_A^* = \star d_A \star$ is the adjoint of d_A. Also, we have written the equations in terms of the curvature two-form F rather than the magnetic field $B = \star F$.) To describe a vortex line, we consider an idealized problem in which the superconductor fills all of $\mathbb{R}^3$, and we look for a solution that is invariant under translation in one direction and is "pulled back" from $\mathbb{R}^2$. The Landau-Ginzburg equations thus reduce to partial differential equations with the same form as those in eqn. (2.8), but now on $\mathbb{R}^2$. They can be further reduced to ordinary differential equations by assuming rotational symmetry on $\mathbb{R}^2$. Solutions of these equations for which $|s| \to a$ and $d_A s \to 0$ at infinity are labeled, as usual, by an integer n, the relative first Chern class. The basic Abrikosov-Gorkov vortex line has $n = 1$. Such a solution exists but cannot be described in closed form.

2.6. **Decay of the Current.** Now we can describe how the current in a superconducting ring relaxes. This happens when a flux line is "nucleated" on one side of the ring (because of thermal or quantum fluctuations), migrates across the ring and then disappears, as sketched in fig. 6. In this process, the relative first Chern class n changes by ± 1, depending on which way the flux line crosses the ring. The case that is energetically feasible at very low temperatures is the case in which $|n|$ is reduced, so that the energy in the superconducting current becomes smaller.

If we include time in this description, then the superconductor fills out a four-manifold $\widehat{U} = U \times \mathbb{R}$ in spacetime (where U is the spatial volume occupied by the superconductor and $\mathbb{R}$ parametrizes the time), and the flux line, as it crosses the ring, fills out a two-manifold $\widehat{D} \subset \widehat{U}$. We can think of $\widehat{D}$ as the submanifold of $\widehat{U}$ characterized by $s = 0$. Topologically, for the process in which the current relaxes, $\widehat{D}$ is a closed two-dimensional disc. If we forget the time, $\widehat{D}$ projects to the disc $D \subset U$ sketched in fig. 6. The quantum or thermal tunneling process by which the flux line crosses the ring is very rare, but it is most likely to occur if $\widehat{D}$ is a

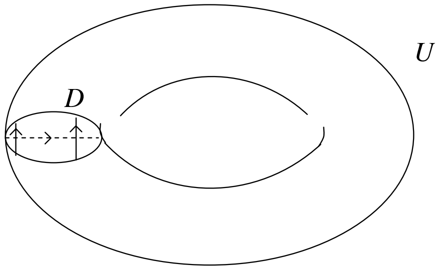

FIGURE 6. A flux line is nucleated on the outside of a superconducting ring and "migrates" across to the inside, where it disappears. In the process, it cuts across the disc labeled D. The two vertical lines in D show the position of the flux line at two successive times.

two-manifold of minimal area in $\widehat{U}$. Thus, to compute the rate at which the current decays, one must find a minimal area two-manifold inside a four-manifold.[4]

In the real world, for a superconducting ring in Minkowski spacetime $\mathbb{R}^{3,1}$, the four-manifold and two-manifold that appear here are very simple. If, however, it were practical to study superconductors in a more general four-dimensional spacetime, an analysis of the current decay would lead to a much more general problem of finding a minimal area two-manifold in four dimensions.

Since the current in a superconducting ring does eventually decay (at least in theory) by the process just described, one may ask precisely what it means to speak of superconductivity. Is a superconductor simply a very good conductor? In fact, like many really interesting concepts in statistical mechanics and condensed matter physics, superconductivity acquires a precise meaning in the limit that the size of the sample becomes large. Let w be the width of our superconducting ring and r its radius. In a ring of given w and r, the current will eventually decay, regardless of what material the ring is made from. In a normal material, the rate at which the current decays is of order r/w^2; the coefficient of r/w^2, in the limit that r and w are both large, is known as the resistivity. In a superconductor, the rate of current decay is exponentially small for w large compared to atomic dimensions, and the resistivity, defined in terms of the coefficient of the power law, vanishes. In practice, the difference between power law decay of the current and exponentially slow decay is that in an ordinary conducting loop, the decay of the current is obvious, while a superconducting loop can carry current for months or years with no measurable change.

2.7. **Type I and Type II Superconductors.** Now let us re-examine the Landau-Ginzburg equations (2.8). As written, the equations depend on parameters e, γ, and a (and units have already been chosen to eliminate $\hbar$ and c). However, by elementary rescalings, one can eliminate all parameters except for a single dimensionless

[4] The tunneling computation is most usefully done in Euclidean signature – with a Riemannian rather than pseudo-Riemannian metric on $\widehat{U}$ – and then $\widehat{D}$ is simply the product of $D \subset U$ with a point in $\mathbb{R}$.

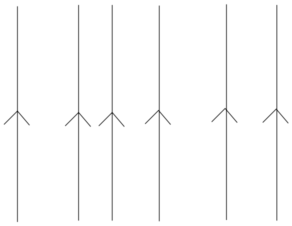

Figure 7. The vertical lines represent parallel flux lines in a superconductor.

ratio $\psi = \gamma/e^2$. We simply introduce a new variable $\widehat{s} = s/a$, new spatial coordinates $\widehat{x} = eax$, and a new connection $\widehat{A} = eA$ and curvature $\widehat{F} = eF$. After these rescalings (but omitting the "hats" to keep the equations readable), the Landau-Ginzburg equations become

$$\begin{aligned} -d_A{}^* d_A s + 2\psi s(|s|^2 - 1) &= 0 \\ d \star F + i \star (\bar{s} d_A s - (d_A \bar{s}) s) &= 0, \end{aligned} \tag{2.9}$$

showing explicitly that ψ is the only relevant parameter.

A sort of phase transition occurs at $\psi = 1$. Consider a space-filling superconductor with many parallel flux lines (fig. 7). For $\psi < 1$, the flux lines attract each other and combine together to form a single flux line of large n. Such a material is called a Type I superconductor; examples include most pure metals except niobium. For $\psi > 1$, the flux lines repel. A material with this property is called a Type II superconductor; examples include niobium and many alloys. The borderline case $\psi = 1$ is discussed later.

The difference between Type I and Type II superconductivity is of substantial practical importance. To understand why, let us return to the Meissner effect (fig. 1). What happens when the external magnetic field is increased to the critical value B_{crit}? Magnetic flux begins to penetrate the material, initially in the form of flux lines. In the Type I case, the flux lines attract, forming a large region of normal metal, and superconductivity is soon lost altogether. In the case of a Type II superconductor, when one reaches the critical magnetic field, flux lines appear inside the superconductor. But since they repel each other, they can form a stable arrangement, a "lattice" of parallel flux lines, as sketched in fig. 8. As a result,

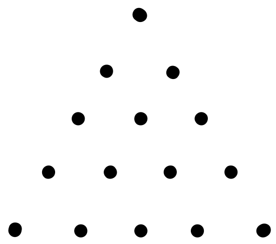

Figure 8. Part of a lattice of flux lines in a Type II superconductor, seen here from above.

a Type II superconductor for $B > B_{\mathrm{crit}}$ can reach a stable arrangement in which part of an externally applied magnetic field is expelled to the outside world, while part penetrates the superconductor in the form of the flux lattice. The material remains superconducting in such a state, and as a result, Type II superconductors can support considerably higher magnetic fields and currents, making them more useful for many applications.

Superconductivity is finally lost at a higher value of the magnetic field called the upper critical field. For more on this, see section 21.6 of [23].

2.8. **The Borderline Case.** Now as in [3] and [19], we are going to use the Landau-Ginzburg theory to discuss the borderline case $\psi = 1$. As we will see momentarily, this case has very special and beautiful properties. But if we are interested in applications to real superconductors, there are a few things to bear in mind.

First of all, a real superconductor will not ordinarily have $\psi = 1$; it will have some other, more generic, value of ψ. In principle, one could try to manufacture a material with $\psi = 1$ by adjusting the chemical composition of an alloy. Even if one does construct such a material, the Landau-Ginzburg model may not be an adequate model for answering subtle questions about its behavior. The Landau-Ginzburg model is not a particularly good model of a superconductor; it is just, in a sense, the simplest possible model. To answer a delicate question about superconductors, such as a question about the boundary between Type I and Type II behavior, one should expect that the Landau-Ginzburg model may be inadequate. The microscopic BCS model might be a much better starting point.

Nevertheless, we will describe the predictions of Landau-Ginzburg theory for flux lines at $\psi = 1$, both because the topic is pretty and because it is good preparation for our study of four-manifolds in section 3. We write x^1, x^2 for Euclidean coordinates on $\mathbb{R}^2$ and introduce the complex variable $z = x^1 + ix^2$. Also, write $d_A = \bar\partial_A + \partial_A$, where $\bar\partial_A$ is of type $(0,1)$ and ∂_A of type $(1,0)$. Precisely at[5] $\psi = 1$, after reduction to two dimensions, the energy function of eqn. (2.7) can be written

$$V = \int_{\mathbb{R}^2} d^2x\big(|\bar\partial_A s|^2 + \frac{1}{2}(\star F - 1 + |s|^2)^2\big) + \pi n, \tag{2.10}$$

where n is the usual topological invariant

$$n = \int_{\mathbb{R}^2} \frac{F}{\pi}. \tag{2.11}$$

One can also write a second formula that differs from the first by a reversal of orientation:

$$V = \int_{\mathbb{R}^2} d^2x\big(|\partial_A s|^2 + \frac{1}{2}(- \star F - 1 + |s|^2)^2\big) - \pi n. \tag{2.12}$$

From eqn. (2.10) it is clear that if the equations

$$\begin{aligned} \bar\partial_A s &= 0 \\ F &= 1 - |s|^2 \end{aligned} \tag{2.13}$$

can be solved, we will get the absolute minimum of V for given n. Moreover, the value of V at its minimum will be πn. Of course, the absolute minimum will

[5] Setting $\psi = 1$ after the preliminary scaling that eliminated e, γ, and a in favor of ψ is equivalent to simply setting $e = \gamma = a = 1$ in eqn. (2.7).

automatically give a solution of the Euler-Lagrange equations. Similarly, from eqn. (2.12), we see that if we can solve

$$\begin{aligned} \partial_A s &= 0 \\ F &= -(1-|s|^2), \end{aligned} \tag{2.14}$$

then we will get the minimum of V, and the value at the minimum will be $-\pi n$. Since the original formula (2.7) made clear that V is non-negative, we can only hope to solve eqn. (2.13) if $n > 0$ or eqn. (2.14) if $n < 0$. The two cases are related simply by a reversal of orientation of $\mathbb{R}^2$, so we will consider just the case that $n > 0$.

The equations (2.13) can be given a simple interpretation. In complex dimension 1, the operator $\bar\partial_A$ automatically obeys $\bar\partial_A{}^2 = 0$, simply because there is no room for a $(0,2)$-form on $\mathbb{R}^2 \cong \mathbb{C}$. So whatever A may be, the $\bar\partial_A$ operator endows the line bundle $\widehat{\mathcal{L}} \to \mathbb{C}$ with a holomorphic structure. The integer n is the degree of this line bundle, relative to its trivialization at infinity (where we assume that $|s|$ approaches 1 and $d_A s$ vanishes). Since a holomorphic line bundle $\widehat{\mathcal{L}} \to \mathbb{C}$ is holomorphically trivial, we can trivialize it, whereupon s becomes an ordinary holomorphic function of z. The integer n becomes the number of zeroes of s, which thus takes the form

$$s = \alpha \prod_{i=1}^{n} (z - z_i), \tag{2.15}$$

where $(z_1, \dots, z_n)$ are an n-plet of unordered and not necessarily distinct points in $\mathbb{C}$ and α is a constant that can be set to 1 by suitably choosing the trivialization of $\widehat{\mathcal{L}}$.

Thus, up to the appropriate equivalence, the solutions of the first equation in (2.13) are in one-to-one correspondence with such n-plets. One then shows [19] by a symplectic or moment map argument that for each n-plet, the second equation in (2.13) has a unique solution, up to a gauge transformation.

So there is a unique multi-vortex solution for each n-plet $(z_1, \dots, z_n)$. We can think of the z_i as the positions of the individual vortex lines. Such a family of solutions does not exist for $\psi \neq 1$. If ψ is not 1, the vortex lines attract or repel one another, and one cannot find a critical point of the energy with vortices at prescribed positions.

2.9. **The Boundary Layer.** In its minimum energy state, a superconductor has $|s| = a$, $d_A s = 0$, and $F = 0$. Since F vanishes, we can pick a gauge (that is, a trivialization of the line bundle $\widehat{\mathcal{L}}$) relative to which the connection form is simply $A = 0$. In this gauge, s is a constant, which by a further choice of gauge we can take to be $s = a$.

Now we would like to understand solutions of the Landau-Ginzburg equations that are close to this minimum energy solution. For this, we write $s = a + h$ and expand the Landau-Ginzburg equations to first order in h and A, around the solution with $h = A = 0$.

In doing this, we can make a gauge choice such that h is real. The gauge group acts by $s \to \exp(if)s$, for an arbitrary real-valued function f. Given that a is real and positive, such a gauge transformation can clearly be chosen to make h real.

Once h is real, the linearized Landau-Ginzburg equations simplify and give separate equations for h and for A. The equation for h becomes

$$\left(\Delta + 4a^2\gamma\right) h = 0, \tag{2.16}$$

where $\Delta = d^*d$ is the Laplacian. The solutions of this linear equation are not difficult to analyze. For an important example, consider a superconductor that occupies the half-space $x \geq 0$, where x is one of the Euclidean coordinates of $\mathbb{R}^3$. A solution depending only on x is a linear combination of

$$h_\pm = \exp(\pm x/\xi), \tag{2.17}$$

where $\xi = 1/2a\sqrt{\gamma}$ is known as the correlation length. To ensure that $|s| = a$ deep inside the superconductor, we must pick the solution h_-, and then we see that $|s|$ departs significantly from a only within a distance of order ξ from the boundary. This conclusion is unchanged if we allow dependence on the other variables.

The story is similar for A. The Landau-Ginzburg equations give

$$\begin{aligned} \left(\Delta + 4e^2a^2\right) A &= 0 \\ d \star A &= 0. \end{aligned} \tag{2.18}$$

Now the solutions that depend only on one coordinate x are proportional to $\exp(\pm x/\lambda)$, where $\lambda = 1/2ea$ is called the penetration depth. Again, for a superconductor that fills a half-space, we must pick the exponentially decaying solution. So A vanishes exponentially for $x >> \lambda$, as does the curvature $F = dA$. Thus we get a more accurate statement of the Meissner effect: the magnetic field in a superconductor is confined to a boundary layer with a thickness of order λ. In view of Maxwell's equation $\vec{\nabla} \times \vec{B} = \vec{J}$ (or as one can see in more detail from the Landau-Ginzburg equations), the electrical currents in a superconductor are confined to the same boundary layer.

Typical values of ξ and λ are of order 10^{-5} or 10^{-6} centimeters – in other words, very small by ordinary standards but hundreds of times the interatomic distances. The dimensionless parameter ψ that distinguishes Type I and Type II superconductors is $\psi = (\lambda/\xi)^2$. For $\xi > \lambda$, one has a Type I superconductor, while $\lambda > \xi$ corresponds to Type II.

Though we have considered a half-plane, the structure of the boundary layer is the same for any superconducting sample whose dimensions are much greater than ξ and λ. (The surface of such a sample looks locally like the boundary of a half-space.) In general, ξ and λ determine the rate of approach to the "ground state" of a superconductor with $|s| = a$ and $d_As = 0$. For example, although this statement requires a little more study of the Landau-Ginzburg equations, the "thickness" of an Abrikosov-Gorkov vortex line is the larger of ξ and λ.

In this article, we have formulated the Landau-Ginzburg model only in the time-independent case. It is not difficult to generalize it to allow time-dependence and to analyze the time-dependent fluctuations around the ground state. In this case, at low frequencies, the results are qualititatively similar to what we have just described: perturbations vanish exponentially outside a boundary layer near the surface of the superconductor. But as explained in [1] and as we will describe in section 4 in the relativistic case, above certain critical frequencies (which differ for h and A), waves can propagate in the superconductor.

3. Four-manifolds

3.1. **Spin^c Structures.** Some of what we have explained for superconductors has a rough analog in four-manifold theory, via the work of Taubes [20, 21] on a perturbed version of the Seiberg-Witten equations. We will have to first explain a number of preliminaries (all of which are presented in a more leisurely fashion in [17], for example), though many of the details are ultimately not important for our main point.

Let X be a smooth four-manifold endowed with a Riemannian metric g. The structure group of the tangent bundle of X is $SO(4)$. A spin structure on X is a lifting of the principal $SO(4)$-bundle of trivializations of the tangent bundle to a principal $\mathrm{Spin}(4)$-bundle over X. Here $\mathrm{Spin}(4)$ is the simply-connected double cover of $SO(4)$. Since $\mathrm{Spin}(4) \cong SU(2) \times SU(2)$, it has two irreducible two-dimensional representations. Associated to a spin structure are therefore a pair of rank two complex vector bundles S_+ and S_-, known as the positive and negative spin bundles. Conversely, the spin structure is determined by the bundle S_+ (or S_-) with $SU(2)$ structure group and a map σ that is defined in subsection 3.2.

In general, the four-manifold X may not admit a spin structure, but it always admits a Spin^c structure. A Spin^c structure is a lifting of the principal $SO(4)$-bundle associated to the tangent bundle to a principal $\mathrm{Spin}^c(4)$ bundle, where $\mathrm{Spin}^c(4) = (\mathrm{Spin}(4) \times U(1))/\mathbb{Z}_2 = (SU(2) \times SU(2) \times U(1))/\mathbb{Z}_2$. The group $\mathrm{Spin}^c(4)$ has two maps to $(SU(2) \times U(1))/\mathbb{Z}_2 = U(2)$ by forgetting one of the $SU(2)$ factors. Using the natural two-dimensional representation of $U(2)$, we get a pair of two-dimensional representations of $\mathrm{Spin}^c(4)$. So a Spin^c structure on X determines a pair of rank two complex vector bundles V_+ and V_- over X. Again, the Spin^c structure is completely determined by V_+ (or V_-) and the map σ.

Locally, a spin structure does exist and is unique up to isomorphism, and one can describe $V_\pm$ as $S_\pm \otimes \mathcal{R}$ where $\mathcal{R}$ is a complex line bundle. Globally, $S_\pm$ and $\mathcal{R}$ may not quite exist (and may not be unique if they exist), but the tensor product does. A unitary connection A on $\mathcal{R}$, together with the Riemannian or Levi-Civita connection on $S_\pm$, determines a connection on $V_\pm$. By a Spin^c connection on $V_\pm$, we mean a connection that is everywhere locally of this type for one (and therefore any) local decomposition of $V_\pm$ as $S_\pm \otimes \mathcal{R}$. The curvature of a Spin^c connection is a two-form valued in $\mathfrak{u}(2)$, the Lie algebra of $U(2)$. As $\mathfrak{u}(2) = \mathfrak{su}(2) \oplus \mathfrak{u}(1)$, the curvature can be projected to $\mathfrak{su}(2)$ or to $\mathfrak{u}(1)$. The projection to $\mathfrak{su}(2)$ gives part of the Riemann tensor of X (because of the definition of a Spin^c connection), while the projection to $\mathfrak{u}(1)$ is a closed two-form F. Locally, F can be regarded as the curvature of the connection A on the line bundle $\mathcal{R}$.

A Spin^c structure determines a line bundle $\det V_+$. Hence it determines an integral cohomology class $c_1(\det V_+)$. If $V_+ = S_+ \otimes \mathcal{R}$, then $\det V_+ = \mathcal{R}^2$ (the determinant of S_+ is trivial, since S_+ has structure group $SU(2)$), so the natural connection on $\det V_+$ has curvature $2F$, and $c_1(\det\, V_+)$ is represented in de Rham cohomology by $2 \cdot F/2\pi$.

3.2. **Dirac Operator and Seiberg-Witten Equations.** Now, let $\Gamma(X, V_\pm)$ be the space of sections of $V_\pm$. There is a natural Dirac operator $\mathcal{D} : \Gamma(X, V_\pm) \to \Gamma(X, V_\mp)$. It is an elliptic, first-order differential operator. In local coordinates, $\mathcal{D} = \sum_{i=1}^4 \gamma^i\, D/Dx^i$, where γ^i (sometimes called the Dirac matrices) generate a Clifford algebra. The Dirac operator was originally introduced by Dirac [4] in

Quantum Electrodynamics. (This original framework for the Dirac operator is closely related to our subject, since the Schrödinger equation for electrons, which would be the starting point in a more precise treatment of superconductivity than we gave in section 2, is a non-relativistic approximation to the Dirac equation.) The Dirac operator is also a central example in the Atiyah-Singer index theorem.

The Seiberg-Witten equations are equations for a pair consisting of a section M of V_+ and a Spin^c connection. Before writing the equations, we need one more preliminary. If X is spin, then $S_+ \otimes S_+ \cong \Omega^0 \oplus \Omega^{2,+}$, where Ω^0 is the bundle of zero-forms and $\Omega^{2,+}$ is the bundle of self-dual two-forms (that is, two-forms that are invariant under the action of the Hodge $\star$ operator, which in four dimensions maps two-forms to two-forms). This statement follows from the representation theory of Spin(4). So there is a natural bilinear map $\sigma : S_+ \otimes S_+ \to \Omega^{2,+}$. If we write $\overline{E}$ for the complex conjugate of a complex vector bundle E, then on a four-dimensional spin manifold, $\overline{S}_+$ is naturally isomorphic to S_+ (this follows from the fact that S_+ is defined using a representation of $SU(2)$ that is equivalent to its complex conjugate). In the Spin^c case, V_+ and $\overline{V}_+$ are not isomorphic, but (since V_+ is locally isomorphic to $S_+ \otimes \mathcal{R}$, where $\mathcal{R}$ is a complex line bundle with a unitary structure, that is a trivialization of $\mathcal{R} \otimes \overline{\mathcal{R}}$), the tensor product $V_+ \otimes \overline{V}_+$ is isomorphic to $S_+ \otimes S_+$. So there is a natural map $\sigma : V_+ \otimes \overline{V}_+ \to \Omega^{2,+}$. We write $\sigma(M)$ for $\sigma(M \otimes \overline{M})$, where $\overline{M}$ is the complex conjugate of M, a section of $\overline{V}_+$.

With this understood, the Seiberg-Witten equations are

$$\begin{aligned} \mathcal{D}M &= 0 \\ F^+ &= \sigma(M). \end{aligned} \tag{3.1}$$

Here $F^+ = \frac{1}{2}(1 + \star)F$ is the self-dual projection of F, a section of $\Omega^{2,+}$.

The Seiberg-Witten equations are elliptic differential equations. They can be used to define invariants of smooth four-manifolds in close analogy with Donaldson theory. Donaldson [5] defined invariants of a smooth four-manifold X by, roughly speaking, counting the instanton solutions in $SU(2)$ gauge theory for a given choice of an $SU(2)$-bundle $E \to X$. Similarly one defines the Seiberg-Witten invariants by counting the solutions of the Seiberg-Witten equations on X for a given choice of the Spin^c structure. The physics of supersymmetric gauge theories [18] leads one to expect that the four-manifold information contained in the Seiberg-Witten invariants is the same as that contained in the Donaldson invariants. This has been explained most fully in [13].

3.3. **Perturbed Equations.** The Seiberg-Witten equations can usefully be perturbed using a closed two-form ω. An important case considered by Taubes [20, 21] is that X is a symplectic four-manifold, with symplectic form ω. Moreover, we orient X using the four-form $\omega \wedge \omega$, in which case it is possible to pick a metric on X such that ω is self-dual, that is $\star\omega = \omega$, and the Riemannian volume form coincides with $\omega \wedge \omega$. (The space of such metrics is contractible, so a choice of one contains no topological information.) Having done so, we perturb the Seiberg-Witten equations to

$$\begin{aligned} \mathcal{D}M &= 0 \\ F^+ &= \sigma(M) + \frac{b}{2}\,\omega \end{aligned} \tag{3.2}$$

with real b. Since the equations are first-order elliptic equations and the perturbation is of zeroth order, general results about elliptic operators (together with compactness properties of the Seiberg-Witten equations) ensure that the counting of solutions is independent of b, and hence the Seiberg-Witten invariants can be computed by counting the solutions of (3.2) for any b. At $b = 0$, we get the original definition. The goal will be to get an alternative description by taking b large.

Let us first consider some important examples of solutions of the perturbed equations. We take X to be $\mathbb{R}^4$ with the Euclidean metric and endowed with a standard symplectic form $\omega = dx^1 \wedge dx^2 + dx^3 \wedge dx^4$. We look for a solution of the equations on $\mathbb{R}^4$ that is invariant under translations in the x^3 and x^4 directions and is "pulled-back" from the $x^1 - x^2$ plane. $\mathbb{R}^4$ is a spin manifold; the spin bundles $S_\pm$ are trivial and endowed with a flat connection. So V_+ is simply the tensor product of a line bundle $\mathcal{R}$ with a two-dimensional vector space $S_{0,+}$. The rotations of the $x^1 - x^2$ plane act in a natural way on $S_{0,+}$, decomposing it as a sum of one-dimensional eigenspaces. Relative to this decomposition, M splits as the direct sum of a pair of sections s and u of $\mathcal{R}$.

As in subsection 2.8, we set $z = x^1 + ix^2$, and we write the covariant differential as $d_A = \bar{\partial}_A + \partial_A$, where $\bar{\partial}_A$ and ∂_A are of types $(0,1)$ and $(1,0)$, respectively. After suitably normalizing and labeling s and u, the equations (3.2) take the form

$$\begin{aligned} \bar{\partial}_A s &= 0 \\ \partial_A u &= 0 \\ \star F &= b - |s|^2 + |u|^2. \end{aligned} \tag{3.3}$$

The value of b is inessential as long as it is not zero, since a dilation or homothety of $\mathbb{R}^2$ multiplies b by a positive real number, and a reversal of orientation of $\mathbb{R}^2$ reverses the sign of b (while exchanging s and u and reversing the sign of the $\star$ operator). For definiteness, we take b positive. If we now set[6] $u = 0$, the equations reduce apart from obvious scalings to the familiar equations (2.13) of Landau-Ginzburg theory at the critical value $\psi = 1$.

For $u = 0$, there is a trivial solution with $A = F = 0$ and $s = \sqrt{b}$. There are also non-trivial solutions. The most important of these, still with $u = 0$, is the solution that in a superconductor is interpreted as the Abrikosov-Gorkov vortex line. In this particular solution, s has a simple zero at, say, the origin, and $|s| \to \sqrt{b}$ at infinity. The correlation length, in the language of subsection 2.9, is $1/\sqrt{b}$, so the difference of $|s|$ from its asymptotic value vanishes for large z like $\exp(-\sqrt{b}|z|)$.

In its embedding as a solution of the Seiberg-Witten equations, the vortex line has thus been scaled or dilated so that its "width" is of order $1/\sqrt{b}$. This result just reflects the dilation or homothety by which we could set $b = 1$.

3.4. **Symplectic and Almost Complex Structures.** On a symplectic four-manifold X, with g any metric such that ω is self-dual and the Riemannian measure agrees with $\omega \wedge \omega$, the expression $J = g^{-1}\omega$ defines an almost complex structure for which ω is of type $(1,1)$. A special case of this is that X may be a Kähler

[6] If b is negative, we instead set $s = 0$, and compare to (2.14) after exchanging s and u.

Actually, solutions of the equations (3.3) with $b > 0$ and suitable behavior at infinity all have $u = 0$. The appropriate conditions are that $|s|$ and $|u|$ are constant at infinity and not both zero, and that $d_A s = d_A u = 0$ at infinity. Vanishing of u for such solutions is proved using the argument sketched in subsection 3.6 below, which more generally shows vanishing of u on a Kähler manifold.

manifold with complex structure J and Kähler form ω. However, in general, J is not integrable, and g is a hermitian metric but not necessarily Kähler.

Without assuming integrability of J, one can decompose the space of complex-valued n-forms as the direct sum of spaces of forms of type (p, q) with $p + q = n$. In particular, an almost complex four-manifold X has a canonical line bundle K whose sections are forms of type $(2, 0)$. X is spin if and only if K admits a square root; indeed, a square root of K determines a spin bundle on X, namely $S_+ = K^{1/2} \oplus K^{-1/2}$. The analogous decomposition on a Spin^c manifold is $V_+ = K^{1/2} \otimes \mathcal{R} \oplus K^{-1/2} \otimes \mathcal{R}$, where $K^{\pm 1/2}$ and $\mathcal{R}$ do not exist globally, but the line bundles $\mathcal{T}_\pm = K^{\pm 1/2} \otimes \mathcal{R}$ do. We have $\mathcal{T}_+^2 = K \otimes \det V_+$, so $K \otimes \det V_+$ always admits a natural square root.

Even though J may not be integrable, there is a natural notion [10] of a pseudo-holomorphic curve: a two-dimensional submanifold $D \subset X$ whose tangent space is J-invariant. If X is Kähler, this coincides with the usual definition of a holomorphic curve. Like holomorphic curves, pseudo-holomorphic curves are curves of minimal area in their homology class.

The equation defining a pseudoholomorphic curve is elliptic, and counting of such curves gives powerful invariants of symplectic manifolds. As we will describe momentarily, Taubes [20, 21] shows that in the four-dimensional case, these are closely related to the Seiberg-Witten invariants.

3.5. **Analogy with Superconductors.** Relative to the decomposition $V_+ \cong K^{1/2} \otimes \mathcal{R} \oplus K^{-1/2} \otimes \mathcal{R}$, M decomposes as $s \oplus u$, where s and u are respectively sections of $K^{1/2} \otimes \mathcal{R}$ and $K^{-1/2} \otimes \mathcal{R}$.

Taubes analyzes solutions of the perturbed Seiberg-Witten equations (3.2) with a parameter $b >> 1$. The following picture emerges. In such a solution, u is uniformly small for large b; indeed, it is everywhere of order $1/b$. For topological reasons, being a section of $\mathcal{T}_+ = K^{1/2} \otimes \mathcal{R}$, s must vanish on a two-dimensional cycle D that is Poincaré dual to $c_1(K^{1/2} \otimes \mathcal{R}) = \frac{1}{2}(c_1(K) + c_1(\det V))$. In the limit of large b, D turns out to be a pseudoholomorphic curve. The structure in the normal bundle to D is everywhere given by the fundamental Abrikosov-Gorkov vortex solution, embedded in Seiberg-Witten theory as we have just described. In other words, the structure in each normal plane to D looks like this solution, scaled to size $1/\sqrt{b}$. Except in a small layer near D with a width of order $1/\sqrt{b}$, $|s|$ is very close to $\sqrt{b}$. The picture is sketched in fig. 9.

Conversely, given a pseudoholomorphic curve in this homology class (with a certain restriction on $\det V_+$), Taubes constructs a solution of the Seiberg-Witten equations. This construction gives much information about the Seiberg-Witten invariants of symplectic four-manifolds, along with a powerful existence theorem for pseudo-holomorphic curves in them. It has had many consequences for the study of symplectic four-manifolds, which are important examples in the theory of smooth four-manifolds.

The large b behavior that we have just summarized is strikingly similar to the description of how the current relaxes in a superconductor (fig. 6). Each problem, that is the superconductor or the symplectic four-manifold, involves an abelian gauge field A coupled to an additional field, which we have called s in each case. s is a section of a complex line bundle that we have called $\widehat{\mathcal{L}}$ or $\mathcal{T}_+$, as the case may be. (The equations involving the symplectic four-manifold also contain another variable u that ultimately is small and does not play an important role.) The equations

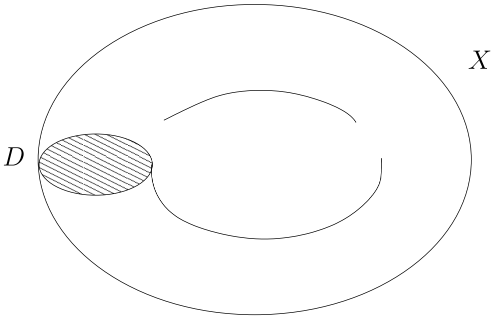

Figure 9. A schematic depiction of a pseudo-holomorphic curve D in a symplectic four-manifold X.

favor a certain value for $|s|$ (which we have called a or $\sqrt{b}$ in the two cases). One can look for a solution in which s is covariantly constant and has the appropriate magnitude. However, such a solution exists only if the topology is right. In the case of the superconductor, if one wants a process in which the superconducting current changes, s must vanish along a two-manifold D. The same happens for the symplectic four-manifold if $c_1(\mathcal{T}_+) \neq 0$. In either case, the important solution looks like the vortex solution near D and like the trivial solution elsewhere.

This simple picture arises for the superconductor if the size of the sample is much larger than ξ and λ. Otherwise, we need to use a more microscopic description. For the symplectic four-manifold, the analogous simplification occurs if $|b| >> 1$. In either case, the simplification occurs when the vortex line can be considered thin, compared to the superconducting sample or the four-manifold.

3.6. **The Weitzenbock Formula.** We perhaps should conclude by giving some idea of how one shows that u is uniformly small for large b. This enables the symplectic four-manifold to be studied using the same variables A, s as the superconductor. The starting point is a Weitzenbock formula. Clearly, the quantity

$$I = \int_X d^4x\sqrt{g}\left(\left(F^+ - \sigma(M) - \frac{b}{2}\omega\right)^2 + |\mathcal{D}M|^2\right) \tag{3.4}$$

is non-negative and vanishes precisely for a solution of the perturbed Seiberg-Witten equations. On the other hand, after integrating by parts, we have[7]

$$I = \int_X d^4x\sqrt{g}\left(|F^+|^2 + |DM|^2 + |\sigma(M) + b\omega/2|^2 + \frac{1}{4}R|M|^2\right). \tag{3.5}$$

R is the scalar curvature of X, and DM is the covariant derivative of M, a section of $T^*X \otimes V_+$.

[7] The functional I is a certain limit of the bosonic part of the effective action of $\mathcal{N} = 2$ super Yang-Mills theory [18]. (To physicists, "Seiberg-Witten theory" is the deduction of this action from an underlying $SU(2)$ gauge theory.) The representation of I as a sum of squares in several different ways is part of the supersymmetric structure of the theory.

In terms of $M = s \oplus u$, we have

$$|\sigma(M) + b\omega/2|^2 = 4|s|^2|u|^2 + (|s|^2 - |u|^2 - b)^2 = |u|^4 + |u|^2(2b + 2|s|^2) + \dots \tag{3.6}$$

where the ellipsis represents terms that are independent of u. To explain the next step in a simplified fashion, suppose that one wishes to minimize as a functional of u an expression

$$J = \int_X d^4x\sqrt{g}\left(|u|^4 + |u|^2(2b + 2|s|^2) + P|u|^2 + \mathrm{Re}(Qu)\right), \tag{3.7}$$

where P and Q are bounded functions on X and independent of b. We write $||Q||$ for the maximum value of the function $|Q|$. The minimum of J is at a value of u that is bounded above, for large b, by $||Q||/b$. In fact P is unimportant for large b; the $|u|^4$ and $|s|^2|u|^2$ terms only make $|u|$ smaller at the minimum of J; and if we simply drop the terms $|u|^4$, $|s|^2|u|^2$, and $P|u|^2$, the minimum of J is at $u = -\overline{Q}/4b$, so $|u|$ is bounded above in that case by $||Q||/4b$.

Now consider minimizing the functional I as a function of u for fixed A and s (as a step toward finding an absolute minimum of I). The problem is similar to the problem considered in the last paragraph, and $|u|$ is bounded above by $1/b$ for similar reasons. The analog of $\int d^4x\sqrt{g}P|u|^2$ is $\int d^4x\sqrt{g}\left(|Du|^2 + R|u|^2/4\right)$. The second term is of the form $P|u|^2$ from the last paragraph, and though the first term (since it involves derivatives of u) is not quite of this form, it is positive semi-definite and hence has the effect of making $|u|$ smaller at the minimum. Finally, a term of the form $\int_X d^4x\sqrt{g}\,\mathrm{Re}(Qu)$ comes from the $|DM|^2$ term in I, which gives rise to cross terms[8] between s and u. After integrating by parts to remove derivatives from u, these terms can be put in the form $\int_X d^4x\,\sqrt{g}\mathrm{Re}(Qu)$ for some Q.

The way that physicists would describe this is to say that because of the $b|u|^2$ term in I, u is for $b >> 1$ a "massive field" or (in the language of the renormalization group) an "irrelevant variable" that can be "integrated out". After doing so, the perturbed Seiberg-Witten equations for symplectic four-manifolds can be described in terms of A and s, the same variables that enter the Landau-Ginzburg theory of superconductivity.

This process of integrating out inessential variables is ubiquitous in physics. For example, a really precise description of a superconductor would involve the Schrödinger equation for electrons and nuclei (or an even more complete theory incorporating additional elementary particle forces). To arrive at the Landau-Ginzburg theory of a superconductor, one has "integrated out" a vast assortment of inessential variables. The four-manifold problem is simpler than superconductivity (and many other realistic problems in physics) in that the process of integrating out irrelevant variables can be carried out in a much more precise way.

We have here discussed only the minimization of I as a function of u, with s and A held fixed. The aim has been to explain how u disappears from the problem for large b, which greatly improves the analogy with superconductivity. However, we will conclude by giving a hint of how [20, 21] to solve the equations for s and A. If $X = \mathbb{R}^4$, we know of two types of solutions. One is the trivial solution with $A = 0$ and $s = \sqrt{b}$, and the second is the Abrikosov-Gorkov vortex. If X is compact and not necessarily flat but $\mathcal{T}_+$ is trivial, then we can find a solution that is everywhere close to the trivial solution. If $\mathcal{T}_+$ is non-trivial, we cannot do this

[8] The cross terms vanish if X is Kähler, so in that case u simply vanishes at the minimum of I.

everywhere, but we can do it away from a two-manifold D, near which we use the vortex solution. This gives an approximate solution everywhere. To get an exact solution, we must minimize I with respect to the choice of D, and this leads to D being a pseudoholomorphic curve.

4. The weak interactions

4.1. **Particle Physics and Supeconductivity.** There is a close analog of superconductivity in the world of elementary particle physics. (Historical references were presented in the introduction.) The forces that are important for ordinary interactions of atoms and subatomic particles are electromagnetism, the weak interactions (which are responsible, for example, for certain forms of radioactivity), and the strong interactions (which, for example, bind protons and neutrons into atomic nuclei). In the Standard Model of particle physics, which was put in its modern form in the 1970's, all of these forces are described via gauge theory. We will focus here on the weak and electromagnetic interactions, as a discussion of the strong interactions would take us too far afield. Moreover, we will omit the quarks and leptons, so our discussion will be highly oversimplified.[9]

Electromagnetism can be described by $U(1)$ gauge theory, as undergraduates are taught, though not always in precisely that language. According to the Standard Model, a more complete description of nature, including the weak interactions as well as electromagnetism, can be obtained by starting with a gauge group that is $U(2)$ (or $SU(2) \times U(1)$); the global structure will not be very important in our discussion and depends on considerations that we will omit, such as the existence of quarks).

If $U(2)$ gauge theory is a better approximation to nature than $U(1)$ gauge theory, why is it that undergraduates learn about $U(1)$? The answer to this is gauge symmetry breaking. For largely unknown reasons, it is energetically favored for a phenomenon to occur that effectively reduces the structure group from $U(2)$ to $U(1)$. Physically, this means that most observations at distances and times that are large compared to the nuclear scale can be adequately described by $U(1)$ gauge theory.

The simplest way to achieve symmetry breaking is quite analogous to the Landau-Ginzburg theory of superconductivity. Thus, some of the key ideas have already been described in section 2. However, before trying to understand the Standard Model of weak and electromagnetic interactions, we will practice with some simpler examples of field theory.

4.2. **Brief Review of Field Theory.** The most elementary example of all is a point particle in mechanics. If t is the time, m is the particle mass, x is the position, and $\dot{x}$ is the velocity, then the kinetic energy is $T = \frac{1}{2}m\dot{x}^2$. If the particle has also a potential energy V, then the action is $I = \int dt(T - V)$ with an important minus sign. The minus sign in the action is needed so that the Euler-Lagrange equations come out to be Newton's laws $m\ddot{x} = -V'(x)$, or equivalently so that the conserved energy is $T + V$ with a plus sign.

[9] A much more complete account of the Standard Model can be found, for example, in [23]. When we speak of "ordinary interactions" of particles, we are omitting gravity, which becomes important for large assemblies of particles but is negligible for a few particles at accessible energies.

Now let us practice with an elementary example of field theory. We consider a real massless scalar field ϕ. We write x^μ, $\mu = 0, \dots, 3$ for coordinates on $\mathbb{R}^{3,1}$, such that the metric is $ds^2 = \sum_{\mu,\nu} \eta_{\mu\nu} dx^\mu \, dx^\nu$, with

$$\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1). \tag{4.1}$$

If a and b are tangent vectors to $\mathbb{R}^{3,1}$, we write $a \cdot b$ for their Lorentz invariant inner product $\sum_{\mu\nu} \eta_{\mu\nu} a^\mu b^\nu$, and we write a^2 (or occasionally $|a|^2$) for $a \cdot a$. To compare to non-relativistic experience, we also separate the time and space coordinates $t = x^0$ and $\vec{x} = (x^1, x^2, x^3)$. Now an action for a massless scalar field can be written

$$I = -\frac{1}{2} \int_{\mathbb{R}^{3,1}} d^4x \sum_{\mu,\nu} \eta^{\mu\nu} \partial_\mu \phi \, \partial_\nu \phi = -\frac{1}{2} \int d\phi \wedge \star d\phi. \tag{4.2}$$

The second version is elegant, but the first version will help in understanding in what sense field theory generalizes elementary mechanics. With our explicit choice (4.1) for the metric, we can write

$$I = \int_{\mathbb{R}^{3,1}} d^4x \left(\frac{1}{2} \left(\frac{\partial \phi}{\partial t} \right)^2 - \frac{1}{2} \sum_{i=1}^{3} \left(\frac{\partial \phi}{\partial x^i} \right)^2 \right). \tag{4.3}$$

We see that if we are going to interpret this as $I = \int dt(T - V)$, then T should be $\frac{1}{2} \int d^3x (\partial\phi/\partial t)^2$, and V should be $\frac{1}{2} \int d^3x \sum_{i=1}^{3} (\partial\phi/\partial x^i)^2$. Thus, T is the integral over all space (as opposed to spacetime) of half the square of the time derivative of the field, while as in mechanics, V contains the terms that lack time derivatives.

A "mass" term for ϕ is a term in the action proportional to ϕ^2. Since this contains no time derivatives, it will be a contribution to V and hence must appear in the action with a minus sign (if we want the energy to be positive). So our action should be

$$I = \int_{\mathbb{R}^{3,1}} d^4x \left(-\frac{1}{2} \sum_{\mu,\nu=0}^{3} \eta^{\mu\nu} \frac{\partial \phi}{\partial x^\mu} \frac{\partial \phi}{\partial x^\nu} - \frac{1}{2} m^2 \phi^2 \right), \tag{4.4}$$

where m^2 should be positive.

To learn more, let us study the Euler-Lagrange equations. These take the form

$$\left(\Box + m^2 \right) \phi = 0, \tag{4.5}$$

where

$$\Box = d^* d = -\sum_{\mu,\nu} \eta^{\mu\nu} \frac{\partial^2}{\partial x^\mu \partial x^\nu} \tag{4.6}$$

is the d'Alembertian (the analog of the Laplacian for Lorentz signature). As these are linear equations with constant coefficients, we can solve them by Fourier analysis, beginning by finding the "plane wave" solutions of the form $\phi = \exp(ik \cdot x)$ with real k. We find that

$$-k^2 = m^2. \tag{4.7}$$

In non-relativistic terminology, one would write $\phi = \exp(-i\omega t) \exp(i\vec{k} \cdot \vec{x})$, where ω is called the angular frequency of the wave and $\vec{k}$ is the wave-vector. So $\omega = -k_0$ and (4.7) reads

$$\omega^2 = \vec{k}^2 + m^2. \tag{4.8}$$

For real $\vec{k}$, there is a minimum possible value of ω, namely $|\omega| = m$. If ω is smaller than this, we have to let $\vec{k}$ acquire an imaginary part, and then the solution grows exponentially in some direction in $\mathbb{R}^{3,1}$. This is not physically sensible for a relativistic field in Minkowski spacetime. However, for a field that is defined only in part of space, solutions with imaginary $\vec{k}$ can be physically sensible. In subsection 2.9, we used solutions to the Landau-Ginzburg equations with imaginary $\vec{k}$ to describe the boundary layer of a superconductor. (In that discussion, we did not include the time-dependence, so in effect we considered only the case $\omega = 0$.)

Quantum mechanically, though an explanation of this would take us too far afield, the waves that we have just described are re-interpreted in terms of particles with energy $\varepsilon = \hbar\omega$ and momentum $\vec{p} = \hbar\vec{k}$. So from (4.8), the energy becomes $\varepsilon = \sqrt{p^2 + m^2}$; one must take the positive sign of the square root, as one learns upon carrying out the quantization procedure. We have here set the speed of light to 1, but if we restore it, the formula becomes $\varepsilon = \sqrt{(pc)^2 + (mc^2)^2}$, which may be familiar from Special Relativity. The minimum frequency corresponds to a minimum energy $\varepsilon_{\min} = mc^2$, usually called the rest energy. The parameter m is called the mass of the particle.

The particular action that we just considered was rather special in that it leads to linear Euler-Lagrange equations. What additional terms could one add to the action? The answer is highly constrained. We want to maintain Poincaré invariance, that is, invariance under the Poincaré group of isometries of $\mathbb{R}^{3,1}$. If we count dimensions so that ϕ has dimension 1 and a derivative also has dimension 1, then it turns out that the quantum theory, in 4 spacetime dimensions, behaves badly if one adds to the action a term of dimension greater than 4. (In general, in n spacetime dimensions, with a slightly different definition of the dimension, one wants to add to the action only terms with dimension at most n.) We are also not interested in terms that can be eliminated by redefining ϕ by $\phi \to a\phi + b$, $a, b \in \mathbb{R}$, or by integration by parts. Nor are we interested in a possible additive constant (in the density whose integral is the action), as this affects neither the classical theory nor its quantization.

Given all this, there are actually only two more parameters by which (4.4) can be usefully modified. The important one for us is that one can add a term $-(\gamma/4)\int d^4x\,\phi^4$ for some real γ. The action thus becomes

$$I = \int_{\mathbb{R}^{3,1}} d^4x \left(-\frac{1}{2} \sum_{\mu,\nu=0}^{3} \eta^{\mu\nu} \frac{\partial\phi}{\partial x^\mu} \frac{\partial\phi}{\partial x^\nu} - \frac{1}{2} m^2\phi^2 - \frac{\gamma}{4}\phi^4 \right). \tag{4.9}$$

This gives a much-studied field theory model known as ϕ^4 theory. The potential energy of ϕ^4 theory is minus the spatial integral of the part of the action that does not involve time derivatives:

$$V = \int d^3x \left(\frac{1}{2} \sum_{i=1}^{3} \left(\frac{\partial\phi}{\partial x^i} \right)^2 + \frac{1}{2} m^2\phi^2 + \frac{\gamma}{4}\phi^4 \right). \tag{4.10}$$

For V to be bounded below, we need $\gamma \geq 0$, since the $\gamma\phi^4$ term is the dominant one for large ϕ. However, if $\gamma > 0$, either sign of the parameter m^2 is physically sensible. (If $\gamma = 0$, positivity of the energy requires, as above, that $m^2 \geq 0$.)

It is useful to reparametrize the action (4.9) as follows:

$$I = \int_{\mathbb{R}^{3,1}} d^4x \left(-\frac{1}{2} \sum_{\mu,\nu=0}^{3} \eta^{\mu\nu} \frac{\partial\phi}{\partial x^\mu} \frac{\partial\phi}{\partial x^\nu} - \frac{\gamma}{4}(\phi^2 - a^2)^2 \right). \tag{4.11}$$

(An irrelevant constant has been added to the action density.) We assume that γ is positive. The model makes sense for either sign of the parameter a^2, but the interesting case for us is that $a^2 > 0$; we thus consider a to be real and positive. The potential energy is now

$$V = \int d^3x \left(\sum_{i=1}^{3} \left(\frac{\partial\phi}{\partial x^i} \right)^2 + \frac{\gamma}{4}(\phi^2 - a^2)^2 \right) \tag{4.12}$$

and is minimized if $\phi = a$.

By setting $\phi = a$ to minimize the classical energy, we get a classical approximation to the quantum vacuum state. To learn what it looks like to live in such a world, we write $\phi = a + h$, where we assume that h is small. In expanding the action near $h = 0$, there is no linear term in h, because $h = 0$ is a solution of the Euler-Lagrange equations. There is, however, a quadratic term:

$$I_2(h) = \int_{\mathbb{R}^{3,1}} d^4x \left(-\frac{1}{2}|dh|^2 - \gamma a^2 h^2 \right). \tag{4.13}$$

This is equivalent to (4.4), but now with $m^2 = 2\gamma a^2$. The mass of the particle is thus (in this approximation) $m_h = a\sqrt{2\gamma}$. What in subsection 2.9 was called the correlation length in a superconductor is the analog of $1/m_h$.

Of course, (4.13) is only an approximation to the action for h. There actually are non-linear corrections to the action, as a result of which the particles obtained by quantizing this Lagrangian scatter each other. The reader who wishes to understand this, however, will have to study Quantum Field Theory.

The action (4.11) has the $\mathbb{Z}_2$ symmetry $\phi \to -\phi$. This symmetry is "spontaneously broken", since it does not leave invariant the minimum energy state at $\phi = a$. We could just as well minimize the energy by taking $\phi = -a$. Of course, in expanding around $\phi = -a$, we would have found the same value for m_h.

The second term we could have added in (4.9) is a term linear in ϕ. This spoils the symmetry and leaves a unique state of least energy. In the weak interaction problem that we come to later, no such term is possible.

4.3. **More on Electromagnetism.**

Now we will even more briefly consider the analog of some of this for electromagnetism. In the electromagnetic case, with connection A and curvature $F = dA$, the action should be a Poincaré-invariant integral of a density that is polynomial in F and its derivatives. We define dimensions so that F has dimension 2 and a derivative has dimension 1, and (in 4 spacetime dimensions) we construct the action from terms of dimension 4 or less. With these rules, the only possible terms in the action are $\int F \wedge \star F$ and $\int F \wedge F$. The second term is a topological invariant (a multiple of the square of the first Chern class) and as such will not be important for our purposes, though it does play a role in the theory. We will likewise omit terms involving the Chern classes when we get to the Standard Model.

Consequently, as long as the only field considered is the connection, the general form of the action is

$$I_A = -\frac{1}{2e^2}\int_{\mathbb{R}^{3,1}} F \wedge \star F, \tag{4.14}$$

where e is a "coupling constant", interpreted in a more complete version of the theory as minus the electric charge of the electron. Our notation differs slightly from subsection 2.1, where we omitted the factor of $1/e^2$ in the action (2.3) but included a factor of e in the definition of the covariant derivative in (2.4). This gives the formulas that one is most likely to see in an undergraduate course, but ultimately it is more convenient to place the coupling parameters in the action rather than in the definition of the covariant derivatives. The two descriptions differ by $A \to \widehat{A} = eA$, which is a change of variables that we actually made, for related reasons, in subsection 2.7.

Now let us describe the plane wave solutions of electromagnetism. Maxwell's equations $d \star F = 0$ are equivalent to $d \star dA = 0$. We can supplement them with a gauge condition $d \star A = 0$. Maxwell's equations plus the gauge condition imply $\Box A = 0$, where $\Box = d^*d + dd^*$ is the d'Alembertian acting on differential forms. For a plane wave of the form $A = \epsilon \cdot dx \exp(ik \cdot x)$ (with constants ϵ and k), we have $\Box A = k^2 A$, so a plane wave solution of Maxwell's equations must have

$$k^2 = 0. \tag{4.15}$$

This is the $m^2 = 0$ case of (4.7), so there is no minimum frequency for electromagnetic waves, and the particles (known as photons) that are obtained by quantizing them have zero mass.

That is actually an essential part of why electromagnetism is obvious in everyday life. The electromagnetic waves that we detect with our eyes have a very long wavelength by atomic standards. We experience electromagnetic effects that have a human scale and greater: lightning, the magnetic field of the Earth, etc. By contrast, an elementary particle of mass m produces significant effects only at a frequency above $mc^2/\hbar$ (the minimum frequency of a plane wave, as we found above) or at a distance scale below $\hbar/mc$ (the analog of the penetration depth in a superconductor, as described in subsection 2.9). The frequency is normally far too high and the distance too small to be accessible without modern experimental equipment.

4.4. **The Standard Model.**

Now we come to the Standard Model of weak and electromagnetic interactions. It is based on $U(2)$ gauge theory, which means to begin with that we are given a rank two complex vector bundle $E \to \mathbb{R}^{3,1}$, with a hermitian metric $(\ , \ )$ and thus structure group $U(2)$. The basic field of $U(2)$ gauge theory is a connection that we will call $\mathcal{C}$. In terms of Lie algebras, $\mathfrak{u}(2) = \mathfrak{su}(2) \oplus \mathfrak{u}(1)$, leading to a decomposition $\mathcal{C} = C \oplus B$, where C and B are respectively $\mathfrak{su}(2)$-valued and $\mathfrak{u}(1)$-valued. Following the usual physics convention, we interpret C and B as, respectively, a traceless hermitian 2×2 matrix of one-forms and a real one-form. Moreover, we normalize these fields so that the gauge-covariant exterior derivative $d_{\mathcal{C}}$, acting on a section H of E, is

$$d_{\mathcal{C}} H = \left(d - iC - \frac{1}{2} iB\right) H. \tag{4.16}$$

The factor of 1/2 multiplying B is for our purposes just a convenient normalization.[10] The factors of $i = \sqrt{-1}$ make the connection unitary.

We write F_C and F_B for the curvatures of C and B, respectively. By the same sort of reasoning that leads to (4.14) as the action for a $U(1)$ connection, we deduce that the part of the action that depends only on $\mathcal{C}$ should be the natural analog of that formula:

$$I_{\mathcal{C}} = -\int_{\mathbb{R}^{3,1}} d^4x \left(\frac{1}{g^2} \operatorname{Tr} F_C \wedge \star F_C + \frac{1}{2(g')^2} F_B \wedge \star F_B \right). \tag{4.17}$$

Here g and g' are real parameters known as coupling constants.

To achieve symmetry breaking, we introduce a field H which is a section of E. Its role will be analogous to that of the field that we called s in the theory of superconductivity. The energy function will be chosen so as to be minimized if $H \neq 0$, reducing the structure group from $U(2)$ to $U(1)$. The same reasoning that led to (4.11) as the general action for a real scalar field gives us the general possible form of the part of the action that depends on H, namely

$$I_H = \int_{\mathbb{R}^{3,1}} d^4x \left(\frac{1}{2} |d_{\mathcal{C}} H|^2 - \frac{\gamma}{4} \left(|H|^2 - a^2 \right)^2 \right) \tag{4.18}$$

with real parameters γ and a^2. Here $|H|^2 = (H, H)$ is defined using the hermitian metric on E, and $|d_{\mathcal{C}} H|^2$ is defined using this metric and the Lorentz metric of $\mathbb{R}^{3,1}$. Just as in our discussion of (4.11), the model is only physically sensible if $\gamma \geq 0$. If γ is positive, as we will assume, the model makes sense for either sign of the parameter a^2. However, the case we want is $a^2 > 0$, so we will take a to be real and positive. The total action of the system is

$$I = I_{\mathcal{C}} + I_H. \tag{4.19}$$

Rather as in our discussion of superconductivity, the action has been chosen so that the energy is minimized if the connection is trivial and H is a non-zero constant of appropriate magnitude. Up to a gauge transformation, the minimum energy is for $B = C = 0$ and

$$H = \begin{pmatrix} 0 \\ a \end{pmatrix}. \tag{4.20}$$

Mathematically, one would say that the choice of the everywhere non-zero section H of the rank two complex vector bundle E reduces the structure group of E from $U(2)$ to the subgroup $U(1)$ consisting of gauge transformations of the form

$$\begin{pmatrix} * & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.21}$$

Physically, one says the same thing, except one says that this reduction occurs at low energy. It is because of this low energy reduction of the structure group that undergraduates learn about $U(1)$ gauge theory, not $U(2)$ gauge theory.

As in the practice problems of subsection 4.2, as well as the discussion of the boundary layer of a superconductor in subsection 2.9, once we have found the state of minimum energy, the next step is to look at the small fluctuations around this

[10] One might interpret it to mean that the global form of the gauge group is not $SU(2) \times U(1)$ but $(SU(2) \times U(1))/\mathbb{Z}_2 = U(2)$. However, this interpretation does not work nicely when quarks and leptons are included, and as noted earlier, the global form of the gauge group will not be important in this article.

state. Furthermore, as in the superconducting case, a judicious gauge choice greatly simplifies the analysis. By a gauge transformation, we can put H in the form

$$H = \begin{pmatrix} 0 \\ a+h \end{pmatrix} \tag{4.22}$$

with real h. This is not a complete choice of gauge; it is invariant precisely under $U(1)$-valued gauge transformations of the form (4.21). So we will do some further gauge-fixing later.

As in the case of a superconductor, the advantage of this gauge condition is that it partially diagonalizes the equations for small fluctuations: h can be treated separately from B and C. For example, we can find the mass of h by inserting the ansatz (4.22) for H in the action I_H and finding the terms that are quadratic in h. One quickly finds that the analysis is exactly equivalent to the analysis that we have already made of (4.11). The quadratic action for h is precisely the one already written in (4.13), and the mass of the h particle is

$$m_h = a\sqrt{2\gamma}. \tag{4.23}$$

To make the equivalent analysis for the gauge fields, we need to work a little harder. We explicitly write C as a matrix of one-forms:

$$C = \frac{1}{2}\begin{pmatrix} C_3 & C_1 + iC_2 \\ C_1 - iC_2 & -C_3 \end{pmatrix}. \tag{4.24}$$

A priori, the quadratic part of the action might be a general bilinear expression in C_1, C_2, C_3, and B. However, we can partially diagonalize the problem using the unbroken $U(1)$ symmetry, that is the symmetry (4.21) that leaves H invariant. In the Lie algebra $\mathfrak{u}(2)$, this $U(1)$ symmetry has two one-dimensional eigenspaces corresponding to the complex conjugate one-forms

$$W^\pm = \frac{C_1 \pm iC_2}{\sqrt{2}} \tag{4.25}$$

and a two-dimensional eigenspace spanned by the real one-forms C_3 and B. To diagonalize the quadratic part of the action in that two-dimensional space, one introduces the linear combinations

$$\begin{aligned} A &= \frac{(g')^2 C_3 + g^2 B}{g^2 + (g')^2} \\ Z &= C_3 - B. \end{aligned} \tag{4.26}$$

It will also be useful momentarily to parametrize g and g' by

$$g = \frac{e}{\sin\theta_W},\ g' = \frac{e}{\cos\theta_W}, \tag{4.27}$$

where (in a more complete form of the model) $-e$ turns out to be the electric charge of the electron and θ_W is called the weak mixing angle.

The point of introducing the linear combinations A, Z, and $W^\pm$ is that in the quadratic approximation, the action for the gauge fields is the sum of an action I_A for A, an action I_Z for Z, and an action I_W for $W^\pm$. As an example of this diagonalization, a short computation shows that in the quadratic approximation,

$$|d_{\mathcal{C}}H|^2 = |dh|^2 + \frac{a^2|W|^2}{2} + \frac{a^2 Z^2}{4}. \tag{4.28}$$

Here $Z^2 = \sum_{\mu,\nu} \eta^{\mu\nu} Z_\mu Z_\nu = Z \cdot Z$, and likewise $|W|^2 = W^- \cdot W^+$. (This simple and "diagonal" result depends on the gauge condition (4.22).) In the quadratic approximation, the rest of the action for the gauge fields has a similar diagonal structure, and we find:

$$\begin{aligned} I_A =& -\frac{1}{2e^2}\int_{\mathbb{R}^{3,1}} F_A \wedge \star F_A \\ I_Z =& \frac{1}{2(g^2+(g')^2)}\int_{\mathbb{R}^{3,1}} \left(-F_Z \wedge \star F_Z - \frac{a^2}{8} Z \wedge \star Z\right) \\ I_W =& \frac{1}{g^2}\int_{\mathbb{R}^{3,1}} \left(-F_{W^-} \wedge \star F_{W^+} - \frac{a^2}{4} W^- \wedge \star W^+\right). \end{aligned} \tag{4.29}$$

Here for L equal to A, Z, or $W^\pm$, we write F_L for dL.

Now let us discuss the consequences of this. Since I_A is the usual action of Maxwell theory, the analysis of electromagnetism is unchanged (in this quadratic approximation) by the embedding in a larger theory. One uses the residual gauge symmetry (4.21) to impose a gauge condition

$$d \star A = 0. \tag{4.30}$$

The Euler-Lagrange equations are simply Maxwell's equations

$$0 = d \star F_A = d \star dA. \tag{4.31}$$

Just as in the derivation of (4.15), these imply that $\Box A = 0$, where $\Box$ is the d'Alembertian. A plane wave solution $A = \epsilon \cdot dx \exp(ik \cdot x)$ must thus have

$$k^2 = 0, \tag{4.32}$$

describing the propagation of electromagnetic waves of arbitrarily small frequency.

Next we consider the field Z. The Euler-Lagrange equations derived from I_Z read

$$d \star dZ + \frac{a^2(g^2+(g')^2)}{4} \star Z = 0. \tag{4.33}$$

By acting with d, we can deduce from this equation that

$$d \star Z = 0. \tag{4.34}$$

This is precisely analogous to (4.30), though it was deduced as an equation of motion rather than as a gauge condition.[11] Once we know that $d \star Z = 0$, we can write $d \star dZ = \star \Box Z$ in terms of the d'Alembertian $\Box = d^*d + dd^* = \star d \star d + d \star d \star$. So the Euler-Lagrange equations imply that $(\Box + a^2(g^2+(g')^2)/4)Z = 0$, and hence in a plane wave solution $Z = \epsilon \cdot dx \exp(ik \cdot x)$, we get

$$-k^2 = \frac{a^2(g^2+(g')^2)}{4}. \tag{4.35}$$

Accordingly, the particles obtained by quantizing these plane waves have mass

$$m_Z = \frac{a\sqrt{g^2+(g')^2}}{2}. \tag{4.36}$$

[11] A better explanation of (4.34) and the analogous condition for $W^\pm$ is as follows. After making a gauge transformation to put H in the form of eqn. (4.22), we should not merely forget about the remaining part of H. We must impose Euler-Lagrange equations for the fields that have been set to zero. This leads to (4.34) and the analog for $W^\pm$.

For $W^{\pm}$, the story is precisely analogous. The Euler-Lagrange equations are

$$d\star dW + \frac{a^2g^2}{4}\star W^{\pm} = 0. \tag{4.37}$$

From this, we deduce that $d\star W^{\pm} = 0$ and that $(\Box + a^2g^2/4)W^{\pm} = 0$. So the particles obtained by quantizing the W field have mass

$$m_W = \frac{ag}{2}. \tag{4.38}$$

Comparing these formulas, we get

$$\frac{m_W}{m_Z} = \frac{g}{\sqrt{g^2 + (g')^2}} = \cos\theta_W. \tag{4.39}$$

All these results, which we have obtained by classical reasoning, are subject to small quantum corrections.

A qualitative explanation of what we have learned is that from a low energy point of view, the relevant gauge group is not the underlying gauge group $U(2)$, but only the subgroup $U(1)$ that leaves H invariant. Associated with the "unbroken" $U(1)$ is the massless gauge field A. The other gauge fields W and Z receive mass from symmetry breaking, or in other words from the non-zero value of H.

4.5. **Experimental Values.** In a more complete version of the model including quarks and leptons, the parameter e is the electric charge of the proton (or minus the electric charge of the electron). The experimental value is

$$\frac{e^2}{4\pi\hbar c} = \frac{1}{137.0359991(46)}. \tag{4.40}$$

The best measurement of the weak mixing angle is

$$\sin^2\theta_W = .23120 \pm .00015. \tag{4.41}$$

The W and Z particles were discovered at the CERN accelerator in the early 1980's. Experimental data give

$$\begin{aligned} m_W &= 80.403 \pm .029 \text{ GeV}/c^2 \\ m_Z &= 91.1876 \pm .0021 \text{ GeV}/c^2. \end{aligned} \tag{4.42}$$

Here GeV/c^2 is a unit of mass, perhaps best described by saying that the proton mass is a little less than 1 GeV/c^2:

$$m_p = .938272029(80) \text{ GeV}/c^2. \tag{4.43}$$

Thus, the W and Z masses are nearly 100 times the proton mass. The unbroken $U(1)$ group (4.21) is the group of electromagnetic gauge transformations, so Z is electrically neutral and $W^{\pm}$ have charges $\pm e$.

The W and Z masses make a dramatic difference in the real world. Classical W and Z fields have a minimum frequency of oscillation $mc^2/\hbar \sim 10^{26}\,\text{sec}^{-1}$, so high that we do not see classical W and Z fields (as we see electromagnetic fields) but only individual quanta or particles. Moreover, it takes sophisticated modern technology to produce or detect these particles. The W and Z masses also cause weak interaction processes (which include decays of various unstable atomic nuclei and elementary particles) to be extremely slow or rare, compared to other subatomic processes.

Thus, according to the Standard Model, electromagnetism and weak interactions have a common origin in gauge theory. But symmetry breaking – the reduction in the low energy gauge group occasioned by the non-zero value of H – causes them to be completely different in the real world.

In most respects, the Standard Model is extremely well-tested experimentally; for an up-to-date survey, see the website [11]. Regrettably, the topic is too vast to be summarized here. To do so, we would need to complete the Standard Model (including the quarks and leptons) and to explain much more about quantization of fields, the resulting particle interactions, and particle physics in general.

4.6. **The Symmetry-Breaking Mechanism.** The exception to the statement that the Standard Model is generally well-tested is that we have almost no experimental information about how the gauge symmetry is broken. The particle that we have labeled as h – generally known as the Higgs particle – has not yet been discovered, and so its mass m_h has not been measured. m_h depends on the parameter we called γ, which does not enter any of our other formulas, so from a theoretical point of view, we do not get any immediate way to predict m_h in terms of the data summarized above.

From a general point of view, should we expect that the Standard Model of gauge symmetry breaking is correct? In the case of a superconductor, we do not expect the Landau-Ginzburg model that we explored in section 2 to be a precise description. Rather, a superconductor can be described more precisely via the microscopic BCS theory [2], or even more precisely (but much less usefully) via the Schrödinger equation for electrons and nuclei. The role of the Landau-Ginzburg model is not that it is a very accurate model of a superconductor, but that it is, in some sense, the simplest conceivable model of gauge symmetry breaking in a superconductor, and hence it accounts for the many important phenomena that depend only on gauge symmetry breaking.

The analogy suggest that the simple Standard Model of $U(2)$ symmetry breaking, via the field H, might be only an approximation to something much more elaborate. (For a brief introduction to such ideas, see section 21.4 of [23].) Most Standard Model tests are mainly sensitive to properties of W and Z particles (and photons, quarks, and leptons) rather than the Higgs particle, and so we have only limited experimental information about the question.

However, what experimental evidence we have seems to hint that for weak interactions, unlike superconductivity, the simple model with the H-field may be a good quantitative description of how symmetry breaking comes about. For one thing, the relation (4.39) is well-obeyed (especially once one evaluates some small quantum corrections to it). This relation would not work in a generic competing model of gauge symmetry breaking. More generally, detailed studies of the weak interactions fit the Standard Model well [11] and have been troublesome for proposed alternatives. But it is always possible that the right alternative has not yet been proposed.

At any rate, the question should be resolved within a few years. If the Standard Model description of gauge symmetry breaking is correct, then the mass of the Higgs particle is at least 115 Gev/c^2 (or it would have already been found) and no more than 200 GeV/c^2 (or the detailed fits to experimental data [11] do not work). The unknown range from 115 to 200 GeV/c^2 will be accessible to the Large Hadron Collider (LHC), which is to begin operating at CERN at the end of 2007.

Indeed, the LHC can reach far beyond this range, so in case the Standard Model description of electroweak symmetry breaking is incorrect or incomplete, we should learn quite a lot about what happens instead.

4.7. **Further Unification?** Having at least partly unified the weak and electromagnetic interactions in a $U(2)$ or $SU(2)\times U(1)$ gauge theory, one naturally wonders if it is possible to do better. There are two immediate directions in which one might hope for more. It would be desirable to also include the strong interactions, which in the Standard Model are described by $SU(3)$ gauge theory (though to explain how this works would take us far deeper into the wilds of quantum theory than we have required for the present article). Also, technically $U(2)$ and $SU(2)\times U(1)$ are not simple Lie groups, as a result of which the action of the Standard Model has independent gauge couplings g and g'. It would be nicer to use a simple Lie group.

The most obvious simple Lie group that contains $SU(3)\times SU(2)\times U(1)$ – which describes the strong, weak, and electromagnetic interactions – is $SU(5)$. We simply consider $SU(5)$ matrices of the form:

$$\begin{pmatrix} * & * & * & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & 0 & 0 \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}. \tag{4.44}$$

A simple model [8] based on $SU(5)$ gauge theory is remarkably successful at accounting for the quantum numbers of quarks and leptons. In a supersymmetric extension, it also gives a very successful prediction for the weak mixing angle θ_W (whose experimental value was recorded in eqn. (4.41)). It may well be that this model is part of a more complete description of nature.

In the $SU(5)$ model, gauge symmetry breaking is carried out in two stages, each of them much like what we have described in the present article. At high energies, the structure group is reduced from $SU(5)$ to $SU(3)\times SU(2)\times U(1)$, and then at much lower energies $SU(2)\times U(1)$ is reduced to $U(1)$ by the field that we have called H. What happens to the $SU(3)$ factor cannot, unfortunately, be usefully described in classical terms.

Beyond unifying the usual elementary particle forces, one would like to also include gravity. Much is achieved in this direction in string theory, but elucidation of this would again take us too far afield.

About the author

Edward Witten is a professor at the Institute for Advanced Study in Princeton. He is a recipient of the Nemmers Prize and the Fields Medal.

References

1. P. W. Anderson, "Plasmons, Gauge Invariance, and Mass," Phys. Rev. **130** (1962) 439-442. MR0153388 (27:3355)
2. J. Bardeen, L. N. Cooper, and J. R. Schrieffer, "Theory of Superconductivity," Phys. Rev. **108** (1957) 1175-1204. MR0095694 (20:2196)
3. E. B. Bogomolny, "Stability of Classical Solutions," Sov. J. Nucl. Phys. **24** (1976) 449. MR0443719 (56:2082)
4. P. A. M. Dirac, "The Quantum Theory of the Electron," Proc. Roy. Soc. (London) **A117** (1928) 610.

5. S. Donaldson, "Polynomial Invariants for Smooth Four-Manifolds," Topology **29** (1990) 257-315. MR1066174 (92a:57035)
6. F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons," Phys. Rev. Lett. **13** (1964) 321-3. MR0174314 (30:4520)
7. V. Ginzburg and L. Landau, "On the Theory of Superconductivity," JETP **20** (1950) 1064-1082.
8. H. Georgi and S. L. Glashow, "Unity of All Elementary Particle Forces," Phys. Rev. Lett. **32** (1974) 438-441.
9. S. L. Glashow, "Partial Symmetries of Weak Interactions," Nucl. Phys. **22** (1961) 579-588.
10. M. Gromov, "Pseudoholomorphic Curves in Symplectic Manifolds," Invent. Math. **82** (1985) 307-347. MR809718 (87j:53053)
11. LEP Electroweak Working Group, http://lepewwg.web.cern.ch/LEPEWWG/.
12. P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," Phys. Rev. Lett. **13** (1964) 508-9. MR0175554 (30:5738)
13. G. Moore and E. Witten, "Integration over the u-Plane in Donaldson Theory," Adv. Theor. Math. Phys. **1** (1998) 298-387. MR1605636 (99k:57070)
14. A. Salam and J. Ward, "Electromagnetic and Weak Interactions," Phys. Lett. **13** (1964) 168-171. MR0192825 (33:1050)
15. A. Salam, "Weak and Electromagnetic Interactions," in N. Svartholm, ed., *Elementary Particle Physics* (Almqvist and Wiksells, Stockholm, 1968), 367-377.
16. J. Schwinger,"Gauge Invariance and Mass," Phys. Rev. **125** (1962) 397-8. MR0154597 (27:4543)
17. A. Scorpan, *The Wild World of 4-Manifolds* (American Mathematical Society, 2005). MR2136212 (2006h:57018)
18. N. Seiberg and E. Witten, "Electric-Magnetic Duality, Monopole Condensation, and Confinement in $\mathcal{N}=2$ Supersymmetric Yang-Mills Theory," Nucl. Phys. **B426** (1994) 19-52. MR1293681 (95m:81202a)
19. C. Taubes, "Arbitrary N-Vortex Solutions to the First-Order Ginzburg-Landau Equations," Commun. Math. Phys. **72** (1980) 277-292. MR573986 (83c:81124)
20. C. Taubes, "The Seiberg-Witten and Gromov Invariants," Math. Res. Lett. **2** (1995) 221-238. MR1324704 (96a:57076)
21. C. Taubes, *Seiberg-Witten and Gromov Invariants for Symplectic 4-Manifolds* (International Press, 2000). MR1798809 (2002j:53115)
22. S. Weinberg, "A Model of Leptons," Phys. Rev. Lett. **19** (1967) 1264.
23. S. Weinberg, *The Quantum Theory of Fields*, volumes 1 and 2 (Cambridge University Press, 1996).

School of Natural Sciences, Institute for Advanced Study, Princeton, New Jersey 08540

E-mail address: witten@ias.edu