2
Physical Motivation

2.1 General Framework for Classical Gauge Theories

Our objective now is to use the machinery assembled in the previous chapter, and that to be developed in subsequent chapters, to study a number of rather specific "classical gauge theories" arising in modern physics. The discussion is heuristic and informal and its intention is to indicate how the topology and geometry that are our real concern here arise naturally in meaningful physics. In order to have a context within which to place these examples we will devote this rather brief section to an explicit enumeration of the basic mathematical ingredients required to describe, at the classical level, the interaction of a particle with a gauge field.

1. A smooth, oriented, (semi-) Riemannian manifold X.

Generally, this will be space ($\mathbb{R}^3$), a spacetime (e.g., $\mathbb{R}^{1,3}$; see Section 2.2), a Euclidean ("Wick rotated") version of a spacetime (e.g., $\mathbb{R}^4$), a compactification of one of these (e.g., $S^4 = \mathbb{R}^4 \cup \{\infty\}$), or an open submanifold of one of these. The particles "live" in X.

2. A finite dimensional vector space $\mathcal{V}$.

The particles have wavefunctions that take values in $\mathcal{V}$. The choice of $\mathcal{V}$ is dictated by the internal structure of the particle (e.g., phase, isospin, spin, etc.) and so $\mathcal{V}$ is called the **internal space**. Typical examples are $\mathbb{C}$, $\mathbb{C}^2$, $\mathbb{C}^4$, or the Lie algebra of some Lie group, e.g., $u(1)$, or $su(2)$. $\mathcal{V}$ is equipped with an inner product $\langle\ ,\rangle$ by which one computes squared norms $\|\psi\|^2$ and thereby the probabilities with which quantum mechanics deals.

3. A matrix Lie group G and a representation

$$\rho : G \longrightarrow GL(\mathcal{V})$$

of G on $\mathcal{V}$ that is orthogonal with respect to the inner product on $\mathcal{V}$:

$$\langle\, \rho(g)(v), \rho(g)(w)\,\rangle = \langle\, v, w\,\rangle.$$

This will generally be one of the classical groups (e.g., $U(1)$, $SU(2)$) or a product of these and plays a dual role.

G.L. Naber, *Topology, Geometry and Gauge fields: Interactions*,
Applied Mathematical Sciences 141, DOI 10.1007/978-1-4419-7895-0_2,

(i) The inner product on $\mathcal{V}$ determines a class of orthonormal bases, or "frames" (e.g., isospin axes) and these are related by the elements of G, i.e., $g \in G$ "acts" on a frame p to give a new frame $p \cdot g$. By fixing some frame at the outset (the elements of which will correspond to certain "states" of the particle) one can identify the elements of G with the frames.

(ii) G also acts on $\mathcal{V}$ via the representation $\rho(v \longrightarrow \rho(g)(v) = g \cdot v)$ and so acts on the wavefunction ψ at each point. If $\psi(p)$ is a value of the wavefunction, described relative to the frame p, then

$$\psi\,(p \cdot g) = g^{-1} \cdot \psi\,(p)$$

is its description relative to the new frame $p \cdot g$.

4. A smooth principal G-bundle over X:

$$G \hookrightarrow P \xrightarrow{\mathcal{P}} X.$$

Typical examples are trivial bundles (e.g., $SU(2) \hookrightarrow \mathbb{R}^4 \times SU(2) \longrightarrow \mathbb{R}^4$) and Hopf bundles (e.g., $U(1) \hookrightarrow S^3 \longrightarrow S^2$, $SU(2) \hookrightarrow S^7 \longrightarrow S^4$). At each $x \in X$ the fiber $\mathcal{P}^{-1}(x)$ is a copy of G thought of as the set of all frames in the internal space $\mathcal{V}$ at x. A local cross-section $s : V \longrightarrow P$ is a smooth selection of a frame at each point in some open subset V of X (a local "gauge") relative to which wavefunctions can be described on V.

5. A connection $\boldsymbol{\omega}$ on $G \hookrightarrow P \xrightarrow{\mathcal{P}} X$ with curvature $\boldsymbol{\Omega}$.

If $s : V \longrightarrow P$ is a local cross-section (local gauge), then the pullback $\boldsymbol{\mathcal{A}} = s^*\boldsymbol{\omega}$ is the **local gauge potential** and $\boldsymbol{\mathcal{F}} = s^*\boldsymbol{\Omega}$ is the **local field strength**. Generally, these exist only locally since nontrivial principal bundles do not admit global cross-sections and pullbacks by different local cross-sections usually do not agree on the intersection of their domains. Particles coupled to (i.e., experiencing the effects of) the field determined by $\boldsymbol{\omega}$ have locally defined wavefunctions ψ taking values in $\mathcal{V}$ that are obtained by solving equations of motion (see #8 below) involving the local potentials $\boldsymbol{\mathcal{A}}$. A change of gauge $(s \longrightarrow s \cdot g)$ changes the wavefunction by the representation ρ $(\psi \longrightarrow g^{-1} \cdot \psi)$. These local wavefunctions piece together into a globally defined object called a **matter field** which can be described in two equivalent ways:

6. A global cross-section of the vector bundle $P \times_\rho \mathcal{V}$ associated to $G \hookrightarrow P \xrightarrow{\mathcal{P}} X$ by ρ (equivalently, a $\mathcal{V}$-valued map $\phi : P \longrightarrow \mathcal{V}$ on P that is equivariant: $\phi(p \cdot g) = g^{-1} \cdot \phi(p)$).

Physically, such a matter field has potential energy which we describe with

7. A non-negative, smooth, real-valued function

$$U : \mathcal{V} \longrightarrow \mathbb{R}$$

that is invariant under the action of G on $\mathcal{V}$:

$$U(g \cdot v) = U(v).$$

U is to be regarded as a potential function with $U \circ \phi$ describing the self-interaction energy of the matter field ϕ. Typically, this will depend only on $\| \phi \|^2$, e.g., $\frac{1}{2} m \| \phi \|^2$, or $\frac{\lambda}{8}(\| \phi \|^2 - 1)^2$, where m and λ are non-negative constants.

8. An action (energy) functional $A(\boldsymbol{\omega}, \phi)$, the stationary points of which are the physically significant field configurations $(\boldsymbol{\omega}, \phi)$.

Typically, this functional is of the following general form:

$$A(\boldsymbol{\omega}, \phi) = c \int_X \Big[\| F_{\boldsymbol{\omega}} \|^2 + c_1 \| d^{\boldsymbol{\omega}} \phi \|^2 + c_2 U \circ \phi \Big].$$

We will spell out in detail what each of these terms means in the concrete examples to follow. Briefly, c is a normalizing constant, c_1 and c_2 are "coupling constants," $F_{\boldsymbol{\omega}}$ is a global 2-form on X with values in the ad-joint bundle $\operatorname{ad} P$ which locally pulls back to the gauge field strengths $\boldsymbol{\mathcal{F}}$, $d^{\boldsymbol{\omega}} \phi$ is the covariant exterior derivative of the matter field ϕ (thought of as a cross-section of the associated vector bundle) and the norms arise from the metric on X and the Killing form on the Lie algebra $\mathcal{G}$ of G. Integrals of such objects over manifolds like X will be introduced in Chapter 4. The physically interesting field configurations $(\boldsymbol{\omega}, \phi)$, are those which (at least locally) minimize the value of the action functional. The Calculus of Variations provides necessary conditions (the Euler-Lagrange differential equations) that must be satisfied by such minima. The Euler-Lagrange equations for the action $A(\boldsymbol{\omega}, \phi)$, are the appropriate field equations (the "equations of motion") of our gauge theory. One can, of course, generalize the model we have described by including more than one matter field.

From the point-of-view of physics, one is generally interested only in **finite action** configurations $(\boldsymbol{\omega}, \phi)$, i.e., those for which

$$A(\boldsymbol{\omega}, \phi) < \infty.$$

This is assured if X is compact, but otherwise one must assume some sort of appropriate asymptotic behavior for the terms in the integrand. Such asymptotic conditions turn out to have profound topological consequences (e.g., the existence of "topological charge") and investigating this link between topology and asymptotics is one of our primary objectives.

These then are the basic ingredients required to build a classical gauge theory. There is a special case that has gotten a great deal of attention, particularly in the mathematical community. When $c_1 = c_2 = 0$ in the action $A(\boldsymbol{\omega}, \phi)$, (so that, effectively, there are no matter fields present), then one can think of A as depending only on $\boldsymbol{\omega}$ and, in this case, it is referred to as the

Yang-Mills action and written

$$\mathcal{YM}(\boldsymbol{\omega}) = c \int_X \| F_{\boldsymbol{\omega}} \|^2.$$

The Euler-Lagrange equations for $\mathcal{YM}$ are called the **Yang-Mills equations** and can be written

$$d^{\boldsymbol{\omega}} {}^* F_{\boldsymbol{\omega}} = 0,$$

where ${}^* F_{\boldsymbol{\omega}}$ is the Hodge dual of $F_{\boldsymbol{\omega}}$ and $d^{\boldsymbol{\omega}}$ is the covariant exterior derivative.

Remark: We have not yet defined the Hodge dual or the covariant exterior derivative in sufficient generality to cover the context in which we now find ourselves, but the generalization is easy and will be provided in Chapter 4. We will also find that, quite independently of the action, the field $F_{\boldsymbol{\omega}}$ also satisfies a purely geometrical constraint known as the **Bianchi identity**

$$d^{\boldsymbol{\omega}} F_{\boldsymbol{\omega}} = 0.$$

These last two equations lie at the heart of what is called pure **Yang-Mills theory**. The impact of this subject on low dimensional topology is discussed at some length in [**N4**]. Although this special case may not appear to be in the spirit of our announced intention here to model interactions we will find that it leads (through a process known as "dimensional reduction") directly to the particular interactions of most interest to us in Section 2.5.

2.2 Electromagnetic Fields

The prototypical example of a gauge theory is classical electrodynamics. Although our real interests lie elsewhere, this example will provide a nice warm-up in familiar territory and so we will describe it in some detail. Keep in mind that our intention in this chapter is primarily motivational so we will feel free to adopt a rather casual attitude, occasionally anticipating concepts and results that are introduced carefully only later in the text.

The arena within which electrodynamics is done is **Minkowski space-time** $\mathbb{R}^{1,3}$ (assuming gravitational effects are neglected). As a differentiable manifold, $\mathbb{R}^{1,3}$ is just $\mathbb{R}^4$. Rather than the usual Riemannian metric on $\mathbb{R}^4$, however, we define on $\mathbb{R}^{1,3}$ the semi-Riemannian metric given, relative to standard coordinates x^0, x^1, x^2, x^3 on $\mathbb{R}^4$ by $\eta_{\alpha\beta} dx^\alpha \otimes dx^\beta$, where

$$\eta_{\alpha\beta} = \begin{cases} 1, & \alpha = \beta = 0 \\ -1, & \alpha = \beta = 1, 2, 3. \\ 0, & \alpha \neq \beta \end{cases}$$

Remark: For the moment we will require very little of the geometry of $\mathbb{R}^{1,3}$ and its physical significance. Simply think of the elements of $\mathbb{R}^{1,3}$ as "events" whose standard coordinates represent the time (x^0) and spatial (x^1, x^2, x^3) coordinates by which the event is identified in some fixed, but arbitrary inertial frame of reference. The entire history of a (point) object can then be identified with a continuous sequence of events (i.e., a curve) in $\mathbb{R}^{1,3}$ called its "world-line." Finally, since the differentiable structure of $\mathbb{R}^4$ is just its natural structure as a real vector space (Example #3, page 4), each tangent space $T_p(\mathbb{R}^{1,3})$ is canonically identified with $\mathbb{R}^4$ itself. Since the components of the semi-Riemannian metric we have introduced are the same at every $p \in \mathbb{R}^{1,3}$ one can think of $\mathbb{R}^{1,3}$ simply as the vector space $\mathbb{R}^4$ equipped with the **Minkowski inner product** $g(v,w) = \eta_{\alpha\beta} v^\alpha w^\beta = v^0 w^0 - v^1 w^1 - v^2 w^2 - v^3 w^3$. A vector v in $\mathbb{R}^{1,3}$ is said to be **spacelike**, **timelike**, or **null** if $g(v,v)$ is $< 0, > 0$, or $= 0$, respectively. The physical origin of the terminology will emerge as we procced.

We introduce a matrix

$$\eta = (\eta_{\alpha\beta}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

and note that its inverse $\eta^{-1} = (\eta^{\alpha\beta})$ is, in fact, equal to η.

Now, to build our gauge theory we begin by letting X denote some open submanifold of $\mathbb{R}^{1,3}$ (the charges creating our electromagnetic field live in $\mathbb{R}^{1,3}$ and we intend to carve out their worldlines and deal only with the source free Maxwell equations on the resulting open submanifold of $\mathbb{R}^{1,3}$). The gauge group G is $U(1)$ so we consider a principal $U(1)$-bundle

$$U(1) \hookrightarrow P \xrightarrow{\mathcal{P}} X$$

over X and a connection $\boldsymbol{\omega}$ on it (we consider first the pure Yang-Mills theory in which matter fields are absent). Since $U(1)$ is Abelian, all brackets in the Lie algebra $u(1) = \operatorname{Im} \mathbb{C}$ are zero so the curvature $\boldsymbol{\Omega}$ of $\boldsymbol{\omega}$ is given by $\boldsymbol{\Omega} = d\boldsymbol{\omega}$. If $s : V \longrightarrow P$ is a local cross-section, then we may write the local gauge potential and field strength as

$$\begin{aligned} \boldsymbol{\mathcal{A}} &= s^* \boldsymbol{\omega} \ = -\boldsymbol{i}\, \boldsymbol{A} \\ \boldsymbol{\mathcal{F}} &= s^* \boldsymbol{\Omega} = d\, \boldsymbol{\mathcal{A}} \ = -\boldsymbol{i}\, d\, \boldsymbol{A} = -\boldsymbol{i}\, \boldsymbol{F}, \end{aligned}$$

where $\boldsymbol{A}$ and $\boldsymbol{F}$ are real-valued forms on V (the minus sign is conventional). If $s_i : V_i \longrightarrow P$ and $s_j : V_j \longrightarrow P$ are two such local cross-sections with $V_j \cap V_i \neq \emptyset$ and if $g_{ij} : V_j \cap V_i \longrightarrow U(1)$ is the corresponding transition function, then

$$\boldsymbol{\mathcal{A}}_j = {g_{ij}}^{-1} \boldsymbol{\mathcal{A}}_i \, g_{ij} + {g_{ij}}^{-1} \, dg_{ij} = \boldsymbol{\mathcal{A}}_i \ + {g_{ij}}^{-1} \, dg_{ij}$$

and

$$\boldsymbol{\mathcal{F}}_j = {g_{ij}}^{-1}\, \boldsymbol{\mathcal{F}}_i\, g_{ij} = \boldsymbol{\mathcal{F}}_i$$

on $V_j \cap V_i$ because $U(1)$ is Abelian. In particular, the local field strengths, since they agree on any intersections of their domains, piece together to give a *globally defined field strength* 2-form $\boldsymbol{\mathcal{F}}$ on X. This is a peculiarity of Abelian gauge theories and one should note that, even here, the potentials $\boldsymbol{\mathcal{A}}$ do *not* agree on the intersections of their domains and so do not give rise to a globally defined object on X. Indeed, since the transition function g_{ij} is a map into $U(1)$ it can be written as

$$g_{ij}(x) = e^{-\boldsymbol{i}\Lambda_{ij}(x)}$$

so that ${g_{ij}}^{-1} dg_{ij} = -\boldsymbol{i} d\Lambda_{ij}$ and $\boldsymbol{\mathcal{A}}_j = \boldsymbol{\mathcal{A}}_i - \boldsymbol{i} d\Lambda_{ij}$. Equivalently,

$$\boldsymbol{A}_j = \boldsymbol{A}_i + d\Lambda_{ij},$$

which is the traditional form for the relationship between two "vector potentials."

Relative to standard coordinates x^0, x^1, x^2, x^3 on $\mathbb{R}^{1,3}$ we can write, for any $s : V \longrightarrow P$,

$$\boldsymbol{\mathcal{A}} = \mathcal{A}_\alpha dx^\alpha = -\boldsymbol{i} A_\alpha dx^\alpha$$

and

$$\boldsymbol{\mathcal{F}} = \frac{1}{2}\mathcal{F}_{\alpha\beta} dx^\alpha \wedge dx^\beta = -\frac{1}{2}\boldsymbol{i} F_{\alpha\beta} dx^\alpha \wedge dx^\beta,$$

where

$$\begin{aligned}
\mathcal{F}_{\alpha\beta} &= \partial_\alpha \boldsymbol{\mathcal{A}}_\beta - \partial_\beta \boldsymbol{\mathcal{A}}_\alpha + [\boldsymbol{\mathcal{A}}_\alpha, \boldsymbol{\mathcal{A}}_\beta] \\
&= \partial_\alpha \boldsymbol{\mathcal{A}}_\beta - \partial_\beta \boldsymbol{\mathcal{A}}_\alpha \qquad \text{(because } U(1) \text{ is Abelian)} \\
&= -\boldsymbol{i}(\partial_\alpha A_\beta - \partial_\beta A_\alpha) \\
&= -\boldsymbol{i} F_{\alpha\beta}.
\end{aligned}$$

The $F_{\alpha\beta}$ are skew-symmetric in α and β. To make some contact with the notation used in physics we define functions E^1, E^2, E^3 and B^1, B^2, B^3 by

$$F_{i0} = E_i$$

and

$$F_{ij} = \varepsilon_{ijk} B^k$$

where $i, j, k = 1, 2, 3$ and ε_{ijk} is the Levi-Civita symbol (1 if ijk is an even permutation of 123, -1 if ijk is an odd permutation of 123, and 0 otherwise). Thus,

$$(F_{\alpha\beta}) = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & B^3 & -B^2 \\ E^2 & -B^3 & 0 & B^1 \\ E^3 & B^2 & -B^1 & 0 \end{pmatrix}$$

and

$$\begin{aligned}\boldsymbol{F} = \frac{1}{2} F_{\alpha\beta}\, dx^\alpha \wedge dx^\beta &= -E^1 dx^0 \wedge dx^1 - E^2 dx^0 \wedge dx^2 - E^3 dx^0 \wedge dx^3 \\ &\quad + B^3 dx^1 \wedge dx^2 - B^2 dx^1 \wedge dx^3 + B^1 dx^2 \wedge dx^3 \\ &= (E^1 dx^1 + E^2 dx^2 + E^3 dx^3) \wedge dx^0 \\ &\quad + B^3 dx^1 \wedge dx^2 + B^1 dx^2 \wedge dx^3 + B^2 dx^3 \wedge dx^1.\end{aligned}$$

One is to think of $\vec{E} = (E^1, E^2, E^3)$ and $\vec{B} = (B^1, B^2, B^3)$ as the "electric field" and the "magnetic field," respectively, that correspond to $\boldsymbol{\mathcal{F}}$ (the justification for thinking this way will appear shortly).

Next we introduce functions $F^{\alpha\beta}$ on X defined by

$$F^{\alpha\beta} = \eta^{\alpha\gamma}\eta^{\beta\delta} F_{\gamma\delta}, \qquad \alpha, \beta = 0, 1, 2, 3$$

(classically this is referred to as "raising the indices" with the Minkowski metric). Thus, for example, $F^{01} = \eta^{0\gamma}\eta^{1\delta} F_{\gamma\delta} = \eta^{00}\eta^{11} F_{01} = (1)(-1)\, F_{01} = -F_{01} = E^1$ and $F^{12} = \eta^{1\gamma}\eta^{2\delta} F_{\gamma\delta} = \eta^{11}\eta^{22} F_{12} = (-1)(-1) F_{12} = F_{12} = B^3$, etc., so

$$(F^{\alpha\beta}) = \begin{pmatrix} 0 & E^1 & E^2 & E^3 \\ -E^1 & 0 & B^3 & -B^2 \\ -E^2 & -B^3 & 0 & B^1 \\ -E^3 & B^2 & -B^1 & 0 \end{pmatrix}.$$

The Hodge dual ${}^*\boldsymbol{F}$ of $\boldsymbol{F}$ is defined to be the 2-form on X whose standard components are given by

$${}^*F_{\alpha\beta} = \frac{1}{2}\varepsilon_{\alpha\beta\gamma\delta}\, F^{\gamma\delta}, \quad \alpha, \beta = 0, 1, 2, 3.$$

Writing these out one finds that

$$({}^*F_{\alpha\beta}) = \begin{pmatrix} 0 & B^1 & B^2 & B^3 \\ -B^1 & 0 & E^3 & -E^2 \\ -B^2 & -E^3 & 0 & E^1 \\ -B^3 & E^2 & -E^1 & 0 \end{pmatrix}$$

so that

$$\begin{aligned}{}^*F &= \frac{1}{2}\, {}^*F_{\alpha\beta}\, dx^\alpha \wedge dx^\beta \\ &= (-B^1 dx^1 - B^2 dx^2 - B^3 dx^3) \wedge dx^0 \\ &\quad + E^3 dx^1 \wedge dx^2 + E^1 dx^2 \wedge dx^3 + E^2 dx^3 \wedge dx^1.\end{aligned}$$

One verifies that

$${}^{**}\boldsymbol{F} = -\boldsymbol{F} \qquad (\text{on } \mathbb{R}^{1,3}).$$

We also define

$$^*\boldsymbol{\mathcal{F}} = -\boldsymbol{i}\,{}^*\boldsymbol{F}.$$

Finally, one can also "raise the indices" of $^*\boldsymbol{F}$ and define $^*F^{\alpha\beta} = \eta^{\alpha\gamma}\,\eta^{\beta\delta}\,{}^*F_{\alpha\beta}$ so that

$$(^*F_{\alpha\beta}) = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & E^3 & -E^2 \\ B^2 & -E^3 & 0 & E^1 \\ B^3 & E^2 & -E^1 & 0 \end{pmatrix}.$$

One can think of the Hodge dual as the 2-form obtained by formally replacing $\vec{B}$ by $\vec{E}$ and $\vec{E}$ by $-\vec{B}$.

All of this apparently *ad hoc* notation will eventually be seen to fit naturally into the general scheme of things. For the time being the reader may wish to regard all of it as simply a useful bookkeeping device. For example, one has the following easily verified formulas:

$$\begin{aligned} \frac{1}{2}\,F_{\alpha\beta}\,F^{\alpha\beta} &= |\vec{B}|^2 - |\vec{E}|^2 \\ \frac{1}{4}\,F_{\alpha\beta}\,{}^*F^{\alpha\beta} &= \vec{E}\cdot\vec{B} \end{aligned}$$

for the two scalar invariants normally associated with $\boldsymbol{F}$ in classical electromagnetic theory.

But what is the justification for all of these references to classical electromagnetic theory? We began by looking at an arbitrary connection $\boldsymbol{\omega}$ on an arbitrary principal $U(1)$-bundle over an open submanifold of Minkowski spacetime and have deviously interjected things we have called "electric fields" and "magnetic fields." Is there any reason to believe that these objects have anything whatever to do with what physicists call "electric fields" and "magnetic fields?" The answer lies in the Yang-Mills equations. We are, after all, not really interested in arbitrary connections, but only in the stationary values of the Yang-Mills action (page 56). In our present circumstances we will find that this action can be written

$$\mathcal{YM}(\boldsymbol{\omega}) = \int_X -\frac{1}{4}F_{\alpha\beta}F^{\alpha\beta}dx^0dx^1dx^2dx^3$$

and the corresponding Yang-Mills equations are

$$d\,{}^*\boldsymbol{F} = 0,$$

while the Bianchi identity is

$$d\boldsymbol{F} = 0$$

(in the Abelian case, covariant exterior derivatives are just ordinary exterior derivatives). In standard coordinates these read

$$\partial_\alpha F^{\alpha\beta} = 0, \quad \beta = 0, 1, 2, 3$$

and

$$\partial_\alpha {}^*F^{\alpha\beta} = 0, \quad \beta = 0, 1, 2, 3,$$

respectively. Now, the remarkable part is that, if one writes these out in terms of the $\vec{E}$'s and $\vec{B}$'s we introduced earlier the result is

$$\vec{\nabla} \times \vec{B} - \frac{\partial \vec{E}}{\partial x^0} = \vec{0} \quad \text{and} \quad \vec{\nabla} \cdot \vec{E} = 0 \qquad (d^*\boldsymbol{F} = 0)$$

and

$$\vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial x^0} = \vec{0} \quad \text{and} \quad \vec{\nabla} \cdot \vec{B} = 0 \qquad (d\boldsymbol{F} = 0),$$

respectively, where $\vec{\nabla} = (\partial_1, \partial_2, \partial_3)$ is the usual gradient operator and the $\times$ and $\cdot$ refer to the cross product and dot product on $\mathbb{R}^3$. These, of course, are the source free Maxwell equations in their usual guise.

Remark: One should note that the second pair of Maxwell equations corresponds to the Bianchi identity and is, in this sense, purely geometrical, i.e., due to the fact that we have chosen to model our fields as connections on principal bundles. The proper way to look at this, however, is the other way around. We are able to build a model of an electromagnetic field as a connection on a principal bundle only because of this second pair of Maxwell equations. As a matter of terminology, one says that a differential form whose exterior derivative is zero is **closed**. Thus, the source free Maxwell equations assert that both $\boldsymbol{F}$ and ${}^*\boldsymbol{F}$ are closed.

Before moving on let us make one more bit of notational contact with physics. The local gauge potentials $\boldsymbol{\mathcal{A}} = -\boldsymbol{i}\boldsymbol{A} = -\boldsymbol{i}A_\alpha dx^\alpha$ satisfy $d\boldsymbol{A} = \boldsymbol{F}$, i.e., $F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha$, which we rewrite as follows: Define $A^\alpha = \eta^{\alpha\gamma} A_\gamma$ for $\alpha = 0, 1, 2, 3$ (i.e., "raise the indices" of $\boldsymbol{A}$) and write

$$(A^0, A^1, A^2, A^3) = \left(V, \vec{A}\right),$$

where $V = A^0 = A_0$ and $\vec{A} = (A^1, A^2, A^3) = (-A_1, -A_2, -A_3)$. Then a brief calculation shows that $d\boldsymbol{A} = \boldsymbol{F}$ becomes

$$\vec{E} = -\frac{\partial \vec{A}}{\partial x^0} - \vec{\nabla} V \quad \text{and} \quad \vec{B} = \vec{\nabla} \times \vec{A}$$

which are the usual expressions from physics for the electric and magnetic fields in terms of the scalar and vector potentials.

We would now like to write out two concrete examples (Coulomb fields and Dirac monopoles) which illustrate all of this apparatus and which, more importantly, make clear the difference between magnetic charge (which is "topological") and electric charge (which is not). We will build the examples by specifying local gauge potentials and allowing these to determine (by the way they are related on the intersections of their domains) the transition

functions and therefore the bundle on which the corresponding connections are defined. In order to fully appreciate the topological nature of what is going on here, however, we must briefly anticipate some material that we will treat carefully only somewhat later in the text.

Principal $U(1)$-bundles over any manifold X are classified up to equivalence by a certain de Rham cohomology class of X called the first Chern class of X. Very briefly, here's what these words mean: As mentioned earlier, a differential form $\boldsymbol{\varphi}$ whose exterior derivative $d\boldsymbol{\varphi}$ is zero is said to be closed. $\boldsymbol{\varphi}$ is said to be **exact** if it is the exterior derivative of some form $\boldsymbol{\psi}$ of degree one less ($\boldsymbol{\varphi} = d\boldsymbol{\psi}$). According to the Poincaré Lemma (Theorem 4.4.2), every exact form is closed ($d^2 = 0$), but the converse is not true. Two closed forms $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ are said to be **cohomologous** if they differ by an exact form ($\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2 = d\boldsymbol{\psi}$). For each degree k this is an equivalence relation and the set of equivalence classes, which admits a natural real vector space structure, is denoted $H^k_{\text{deR}}(X)$ and called the k^{th} **de Rham cohomology group** of X.

Now let $G \hookrightarrow P \xrightarrow{\mathcal{P}} X$ be any principal G-bundle over X (with G a matrix Lie group). The **first Chern class** $c_1(P)$ of the bundle is an element of $H^2_{\text{deR}}(X)$ defined as follows: Choose *any* connection $\boldsymbol{\omega}$ on the bundle (we will prove later that connections exist on any principal bundle). Let $\boldsymbol{\Omega}$ be the curvature of $\boldsymbol{\omega}$. For any local cross-section s let $\boldsymbol{\mathcal{F}} = s^*\boldsymbol{\Omega}$ be the local field strength. These generally depend on the choice of s and do not agree on the intersections of their domains (except in the Abelian case). Indeed, if s^g is another cross-section, then $\boldsymbol{\mathcal{F}}^g = g^{-1}\boldsymbol{\mathcal{F}}g$ on the intersection. Notice, however, that, since the trace of a matrix is invariant under conjugation, trace $\boldsymbol{\mathcal{F}}^g =$ trace $\boldsymbol{\mathcal{F}}$ on the intersection and so these piece together to give a globally defined 2-form on X which we will denote simply trace $\boldsymbol{\mathcal{F}}$. *A priori* trace $\boldsymbol{\mathcal{F}}$ is complex-valued, but one can show that, in fact, it takes values in Im $\mathbb{C}$ so that $\frac{\boldsymbol{i}}{2\pi}$ trace $\boldsymbol{\mathcal{F}}$ is real-valued (the $\frac{1}{2\pi}$ actually forces the integrals of this 2-form over compact, oriented surfaces in X to be integers—not obvious, but true). It also happens that this 2-form is closed and so determines a cohomology class

$$c_1(P) = \left[\frac{\boldsymbol{i}}{2\pi}\text{trace } \boldsymbol{\mathcal{F}}\right] \in H^2_{\text{deR}}(X).$$

Now, the remarkable part of all this is that this cohomology class (*not* the 2-form, but its cohomology class) does not depend on the initial choice of the connection $\boldsymbol{\omega}$ from which it arose. It is therefore a characteristic of the bundle itself and not of the connection. Indeed, $c_1(P)$ is the simplest example of what is called a **characteristic class** for the bundle. We'll encounter one more example of such a thing (the second Chern class) in Section 2.5.

Now let us specialize to the case of $U(1)$-bundles. Here we have seen that there is a globally defined field strength $\boldsymbol{\mathcal{F}}$ for any connection. Moreover, since $u(1)$ consists of 1×1 matrices, the trace just picks out the sole entry in this matrix so trace $\boldsymbol{\mathcal{F}} = -\boldsymbol{i}\boldsymbol{F}$ and therefore

$$c_1(P) = \left[\frac{\boldsymbol{i}}{2\pi}(-\boldsymbol{i}\boldsymbol{F})\right] = \frac{1}{2\pi}\,[\boldsymbol{F}] \qquad (G = U(1)).$$

In particular, the first Chern class for a principal $U(1)$-bundle over an open submanifold of Minkowski spacetime is just $\frac{1}{2\pi}$ times the cohomology class of the globally defined (electromagnetic) field $\boldsymbol{F}$ on X. Since principal $U(1)$-bundles over any manifold are classified up to equivalence by their first Chern classes (Appendix E, [**FU**]), we conclude that the $U(1)$-bundle on which an electromagnetic field $\boldsymbol{F}$ is to be modeled as a connection is uniquely determined by the cohomology class of $\boldsymbol{F}$.

Now we return to our concrete examples. First, the Coulomb field, i.e., a static, purely electric field of a point charge which we assume to be located at the (x^1, x^2, x^3)-origin in $\mathbb{R}^{1,3}$. Thus, the worldline of our source is the x^0-axis in $\mathbb{R}^{1,3}$ so we take

$$X = \mathbb{R}^{1,3} - \{(x^0, 0, 0, 0) \in \mathbb{R}^{1,3} : x^0 \in \mathbb{R}\}.$$

Define $\boldsymbol{A} = A_\alpha dx^\alpha = (-n/\rho)dx^0$, where n is an integer and $\rho > 0$ with $\rho^2 = (x^1)^2 + (x^2)^2 + (x^3)^2$ (we measure "charge" in multiples of the charge of the electron so it is an integer). A simple calculation shows that the functions $F_{\alpha\beta} = \partial_\alpha A_\beta - \partial_\beta A_\alpha,\ \alpha, \beta = 0, 1, 2, 3$, are given by

$$(F_{\alpha\beta}) = \frac{n}{\rho^3}\begin{pmatrix} 0 & -x^1 & -x^2 & -x^3 \\ x^1 & 0 & 0 & 0 \\ x^2 & 0 & 0 & 0 \\ x^3 & 0 & 0 & 0 \end{pmatrix}.$$

Thus,

$$\boldsymbol{F} = \frac{n}{\rho^3}(x^1 dx^1 + x^2 dx^2 + x^3 dx^3) \wedge dx^0$$

so

$$\vec{B} = \vec{0} \quad \text{and} \quad \vec{E} = \frac{n}{\rho^3}\vec{r}, \quad \vec{r} = (x^1,\ x^2,\ x^3).$$

This is, of course, the classical Coulomb field that we wish to describe.

The critical observation here is this: Our Coulomb potential $\boldsymbol{A} = (-n/\rho)dx^0$ is defined and satisfies $d\boldsymbol{A} = \boldsymbol{F}$ *globally* on all of X so that $\boldsymbol{F}$ is exact on X and its cohomology class $[\boldsymbol{F}] \in H^2_{\text{deR}}(X)$ is zero. Thus, the $U(1)$-bundle on which $\boldsymbol{F}$ is modeled by a connection with field strength $\boldsymbol{\mathcal{F}} = -\boldsymbol{i}\boldsymbol{F}$ has first Chern class zero and so must be the trivial bundle. This is true for *any* charge n so that, in particular, the electric charge of the source is not encoded in the topology of this bundle (we will find that the situation is quite different for a Dirac magnetic monopole).

Here's another way to look at this: Somewhat later we will calculate the cohomology of $X = \mathbb{R}^{1,3} - \{(x^0, 0, 0, 0) \in \mathbb{R}^{1,3} : x^0 \in \mathbb{R}\}$ and find that

$$H^k_{\text{deR}}(X) = \begin{cases} \mathbb{R}, & k = 0, 2 \\ 0, & \text{otherwise} \end{cases}.$$

Now, the 2-form $\boldsymbol{F}$ describing the Coulomb field on X is cohomologically trivial, i.e., $[\boldsymbol{F}] = 0 \in H^2_{\mathrm{deR}}(X)$. After learning how to integrate 2-forms over 2-dimensional manifolds we will find that this implies that the integral of $\boldsymbol{F}$ over any closed, smoothly embedded surface in some $x^0 =$ constant slice must be zero. In particular, these integrals do not detect any "enclosed charge." The rest of the 2-dimensional cohomology of X is to be found in the dual ${}^*\boldsymbol{F} = \frac{1}{2}\, {}^*F_{\alpha\beta}\, dx^\alpha \wedge dx^\beta$, where

$$({}^*F_{\alpha\beta}) = \frac{n}{\rho^3} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & x^3 & -x^2 \\ 0 & -x^3 & 0 & x^1 \\ 0 & x^2 & -x^1 & 0 \end{pmatrix}.$$

Thus,
$${}^*\boldsymbol{F} = \frac{n}{\rho^3}\,(x^1\, dx^2 \wedge dx^3 - x^2\, dx^1 \wedge dx^3 + x^3\, dx^1 \wedge dx^2).$$

Notice that, on the 2-sphere $\rho = 1$ in any $x^0 =$ constant slice of X, ${}^*\boldsymbol{F}$ reduces to n times $x^1\, dx^2 \wedge dx^3 - x^2\, dx^1 \wedge dx^3 + x^3\, dx^1 \wedge dx^2$ on S^2. This, as we shall see, is what is called the standard volume form of S^2 and its integral over S^2 is the area 4π of S^2 (Section 4.6). Thus, the integral of ${}^*\boldsymbol{F}$ over this sphere is $4\pi n$, which is nonzero and this implies that ${}^*\boldsymbol{F}$ cannot be cohomologically trivial. Thus, $[{}^*\boldsymbol{F}]$ generates $H^2_{\mathrm{deR}}(X)$. Furthermore, we will show that the integral of ${}^*\boldsymbol{F}$ over any 2-sphere surrounding $\{(x^0, 0, 0, 0) : x^0 \in \mathbb{R}\}$ is the same so these integrals "detect" the charge n enclosed by the sphere. Over any 2-sphere that does not enclose the x^0-axis, the integral of ${}^*\boldsymbol{F}$ is zero (this will follow from "Stokes' Theorem").

The electric charge n is not "topological" because all Coulomb fields are represented by connections on the trivial bundle—the charge is not encoded in the topology of the bundle. The situation is quite different for a Dirac monopole (which is not surprising since the Hodge dual for 2-forms on X essentially interchanges "electric" and "magnetic" so, for a magnetic charge, $[\boldsymbol{F}]$ will play the role that $[{}^*\boldsymbol{F}]$ played for the Coulomb field). In more detail, we once again let $X = \mathbb{R}^{1,3} - \{(x^0, 0, 0, 0) \in \mathbb{R}^{1,3} : x^0 \in \mathbb{R}\}$. For the moment we let g denote an arbitrary real number (to be thought of as the magnetic "charge" of the monopole whose worldline is the x^0-axis). We are interested in the field $\boldsymbol{F} = \frac{1}{2}\, F_{\alpha\beta}\, dx^\alpha \wedge dx^\beta$, where

$$(F_{\alpha\beta}) = \frac{g}{\rho^3} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & x^3 & -x^2 \\ 0 & -x^3 & 0 & x^1 \\ 0 & x^2 & -x^1 & 0 \end{pmatrix}.$$

Thus,
$$\boldsymbol{F} = \frac{g}{\rho^3}\,(x^1\, dx^2 \wedge dx^3 - x^2\, dx^1 \wedge dx^3 + x^3\, dx^1 \wedge dx^2)$$

so

$$\vec{E} = \vec{0} \quad \text{and} \quad \vec{B} = \frac{g}{\rho^3}\,\vec{r}, \quad \vec{r} = (x^1, x^2, x^3).$$

Since $\boldsymbol{F}$ is independent of x^0 this defines a 2-form on any $x^0 =$ constant slice. Moreover, expressed in terms of standard spherical coordinates (ρ, φ, θ) on such a slice,

$$\boldsymbol{F} = g\,\sin\varphi\, d\varphi \wedge d\theta.$$

Notice that this is independent of ρ and so may be further restricted to the copy $\rho = 1$ of S^2 in, say, the $x^0 = 0$ slice of X. Now, if the monopole field on X is the field strength of some connection on a $U(1)$-bundle over X, then its restrictions would likewise be field strengths for connections on $U(1)$-bundles over the submanifolds (restriction means pullback by the inclusion map and the inclusion of a restricted bundle is a bundle map). Henceforth, we will concentrate on these restrictions.

For the field $\boldsymbol{F}$ under consideration there is no globally defined potential $\boldsymbol{A}$ satisfying $d\boldsymbol{A} = \boldsymbol{F}$ (pages 2–3 of [**N4**]), but there are the usual local potentials. Specifically, we define $\boldsymbol{A}_N$ and $\boldsymbol{A}_S$ on $U_N = S^2 - \{(0,0,0,-1)\}$ and $U_S = S^2 - \{(0,0,0,1)\}$, respectively, by

$$\boldsymbol{A}_N = g(1 - \cos\varphi)\, d\theta$$

and

$$\boldsymbol{A}_S = -g(1 + \cos\varphi)\, d\theta.$$

Then, on their respective domains, these satisfy $d\boldsymbol{A}_N = \boldsymbol{F}$ and $d\boldsymbol{A}_S = \boldsymbol{F}$. Consider now the corresponding $u(1)$-valued forms (we identify $u(1)$ with $\operatorname{Im}\mathbb{C}$):

$$\begin{aligned} \boldsymbol{\mathcal{A}}_N = -\boldsymbol{i}\boldsymbol{A}_N &= -\boldsymbol{i}g(1 - \cos\varphi)\, d\theta \\ \boldsymbol{\mathcal{A}}_S = -\boldsymbol{i}\boldsymbol{A}_S &= \quad \boldsymbol{i}g\ (1 + \cos\varphi)\, d\theta \\ \boldsymbol{\mathcal{F}} = -\boldsymbol{i}\,\boldsymbol{F} &= -\boldsymbol{i}g\,\sin\varphi\, d\varphi \wedge d\theta \end{aligned}$$

If our monopole field is to be modeled by a connection on some principal $U(1)$-bundle over S^2, then that bundle could be trivialized over U_N and U_S and so would have a transition function $g_{SN} : U_N \cap U_S \longrightarrow U(1)$ for which $\boldsymbol{\mathcal{A}}_N = {g_{SN}}^{-1}\,\boldsymbol{\mathcal{A}}_S g_{SN} + {g_{SN}}^{-1} dg_{SN}$. But notice that $\boldsymbol{\mathcal{A}}_N - \boldsymbol{\mathcal{A}}_S = -2g\boldsymbol{i}d\theta$ so

$$\boldsymbol{\mathcal{A}}_N = \boldsymbol{\mathcal{A}}_S - 2g\boldsymbol{i}d\theta = e^{2g\theta\boldsymbol{i}}\boldsymbol{\mathcal{A}}_S e^{-2g\theta\boldsymbol{i}} + e^{2g\theta\boldsymbol{i}} d(e^{-2g\theta\boldsymbol{i}})$$

and this gives

$$g_{SN}(\varphi, \theta) = e^{-2g\theta\boldsymbol{i}}.$$

Next we observe that the only values of the constant g that are of any interest are given by

$$g = \frac{n}{2}, \qquad n \in \mathbb{Z}.$$

One can understand this in a number of ways, all of which are instructive. From the point of view of physics, it is just the Dirac quantization condition (page 7, [**N4**]). On the other hand, if $g_{SN}(\varphi, \theta) = e^{-2g\theta\boldsymbol{i}}$ is really the

transition function for a principal $U(1)$-bundle $U(1) \hookrightarrow P \longrightarrow S^2$, then that bundle is characterized by its first Chern class $c_1(P)$. As we pointed out earlier, the integral of $c_1(P)$ over S^2 is an integer, called the **first Chern number** of the bundle. But a simple calculation gives

$$\int_{S^2} c_1(P) = \frac{1}{2\pi} \int_{S^2} \boldsymbol{F} = \frac{g}{2\pi} \int_{S^2} \sin\varphi d\varphi \wedge d\theta = 2\,g$$

(we assume S^2 has its standard orientation). Thus, $2g \in \mathbb{Z}$. From yet another perspective, the restriction of $g_{SN}(\varphi, \theta) = e^{-2g\theta \boldsymbol{i}}$ to the equatorial circle S^1 in $U_N \cap U_S$ (i.e., $e^{\boldsymbol{i}\theta} \longrightarrow (e^{\boldsymbol{i}\theta})^{-2g}$) would be the "characteristic map" whose homotopy type determines the bundle (page 228, [**N4**]) and this map is not even well-defined (single-valued) on S^1 unless $2g$ is an integer. However you choose to view the situation, we will henceforth restrict our attention to the following forms:

$$\begin{aligned} \boldsymbol{\mathcal{A}}_N &= -\frac{1}{2}\, n\boldsymbol{i}\,(1 - \cos\varphi)\, d\theta \\ \boldsymbol{\mathcal{A}}_S &= \quad \frac{1}{2}\, n\boldsymbol{i}\,(1 + \cos\varphi) d\theta \\ \boldsymbol{\mathcal{F}} &= -\frac{1}{2}\, n\boldsymbol{i} \sin\varphi\; d\varphi \wedge d\theta. \end{aligned}$$

For each fixed integer n, the potentials $\boldsymbol{\mathcal{A}}_N$ and $\boldsymbol{\mathcal{A}}_S$ uniquely determine a connection $\boldsymbol{\omega}_n$ on the principal $U(1)$-bundle

$$U(1) \hookrightarrow P_n \xrightarrow{\mathcal{P}_n} S^2$$

whose transition function g_{SN} is given by

$$g_{SN}(\varphi, \theta) = e^{-n\theta \boldsymbol{i}}.$$

The globally defined field strength on S^2 is $\boldsymbol{\mathcal{F}}$ and represents the field of a Dirac monopole of "magnetic charge" n (which is the Chern number of the bundle). Since the Chern number is a topological characteristic of the bundle, magnetic charge is directly encoded into the topology of the bundle and is therefore an instance of a "topological charge."

Although it is not really necessary to do so (because the transition functions and local gauge potentials contain all of the relevant information about the bundles and the connections), it is possible to describe these Dirac monopoles more explicitly. We will outline such a description.

$$\boxed{n = 1}$$

This gives the natural connection on the complex Hopf bundle

$$U(1) \hookrightarrow S^3 \xrightarrow{\mathcal{P}_1} S^2,$$

where $\mathcal{P}_1$ is the restriction to $S^3 \subseteq \mathbb{C}^2$ of the map

$$\mathcal{P}_1\left(z^1, z^2\right) = \left(z^1\bar{z}^2 + \bar{z}^1 z^2, -\boldsymbol{i} z^1\bar{z}^2 + \boldsymbol{i}\bar{z}^1 z^2, \left|z^1\right|^2 - \left|z^2\right|^2\right).$$

$\boldsymbol{\omega}_1$ is given by

$$\boldsymbol{\omega}_1 = \boldsymbol{i}\iota^*(\mathrm{Im}(\bar{z}^1 dz^1 + \bar{z}^2 dz^2)),$$

where $\iota : S^3 \hookrightarrow \mathbb{C}^2$ is the inclusion.

$$\boxed{n = -1}$$

This gives the natural connection on the alternate version of the complex Hopf bundle

$$U(1) \hookrightarrow S^3 \xrightarrow{\mathcal{P}_{-1}} S^2,$$

where $\mathcal{P}_{-1}$ is the restriction to $S^3 \subseteq \mathbb{C}^2$ of the map

$$\mathcal{P}_{-1}\left(z^1, z^2\right) = \left(z^1\bar{z}^2 + \bar{z}^1 z^2, \boldsymbol{i} z^1\bar{z}^2 - \boldsymbol{i}\bar{z}^1 z^2, \left|\, z^1\,\right|^2 - \left|\, z^2\,\right|^2\right).$$

$\boldsymbol{\omega}_{-1}$ is given by

$$\boldsymbol{\omega}_{-1} = \boldsymbol{i}\iota^*\left(\mathrm{Im}\left(\bar{z}^1 dz^1 + \bar{z}^2 dz^2\right)\right) = \boldsymbol{\omega}_1,$$

where $\iota : S^3 \hookrightarrow \mathbb{C}^2$ is the inclusion.

$$\boxed{n = 0}$$

This gives the flat connection on the trivial bundle

$$U(1) \hookrightarrow S^2 \times U(1) \xrightarrow{\mathcal{P}_0} S^2,$$

where $\mathcal{P}_0$ is the projection onto the first factor. $\boldsymbol{\omega}_0$ is given by

$$\boldsymbol{\omega}_0 = \pi^*\boldsymbol{\Theta},$$

where $\pi : S^2 \times U(1) \longrightarrow U(1)$ is the projection onto the second factor and $\boldsymbol{\Theta}$ is the Cartan 1-form on $U(1)$.

$$\boxed{n > 1}$$

Denote by $U(1) \hookrightarrow P_n \xrightarrow{\mathcal{P}_n} S^2$ the principal $U(1)$-bundle over S^2 with transition function $g_{SN}(\varphi, \theta) = e^{-n\theta\boldsymbol{i}}$. We can identify P_n explicitly as a manifold as follows: Identify the discrete group $\mathbb{Z}_n$ of integers modulo n with the following subgroup of $U(1)$:

$$\mathbb{Z}_n = \{e^{2k\pi\boldsymbol{i}/n} : k = 0, 1, \ldots, n-1\}.$$

Then $\mathbb{Z}_n$ acts on S^3 on the right (because $U(1)$ does). We let $S^3/\mathbb{Z}_n$ be the orbit space, e.g., $S^3/\mathbb{Z}_2 = \mathbb{RP}^3$. One can provide $S^3/\mathbb{Z}_n$ with a manifold

structure in the same way as for $\mathbb{RP}^3$. The Hopf map $\mathcal{P}_1 : S^3 \longrightarrow S^2$ carries each orbit of the usual $U(1)$-action on S^3 to a point in S^2 so it does the same for a $\mathbb{Z}_n$-orbit. Moreover, each point of S^2 is the image under $\mathcal{P}_1$ of a $\mathbb{Z}_n$-orbit (indeed, of many $\mathbb{Z}_n$-orbits). Thus, $\mathcal{P}_1$ descends to a surjective map

$$P_n : S^3/\,\mathbb{Z}_n \longrightarrow S^2.$$

Also note that the usual $U(1)$-action on S^3 carries any $\mathbb{Z}_n$-orbit onto another $\mathbb{Z}_n$-orbit which is inside the same $U(1)$-orbit (and so has the same image under $\mathcal{P}_n$). Thus, the $U(1)$-action on S^3 descends to a $U(1)$- action on $S^3/\mathbb{Z}_n$ which preserves the fibers of $\mathcal{P}_n$. Local triviality of $\mathcal{P}_n : S^3/\mathbb{Z}_n \longrightarrow S^2$ follows so that

$$U(1) \hookrightarrow S^3/\,\mathbb{Z}_n \xrightarrow{\mathcal{P}_n} S^2$$

is a principal $U(1)$-bundle and the transition function g_{SN} is given by $g_{SN}(\varphi, \theta) = e^{-n\,\theta\,\boldsymbol{i}}$. Since a bundle is determined by its transition functions, we have an explicit model for $U(1) \hookrightarrow P_n \xrightarrow{\mathcal{P}_n} S^2$. In particular, $P_n \cong S^3/\,\mathbb{Z}_n$.

Remark: $P_n = S^3/\,\mathbb{Z}_n$ is an example of a "lens space."

Since the transition function for $U(1) \hookrightarrow S^3/\,\mathbb{Z}_n \xrightarrow{\mathcal{P}_n} S^2$ is $g_{SN}(\varphi,\, \theta) = e^{-n\,\theta\,\boldsymbol{i}}$, $\boldsymbol{\mathcal{A}}_N = -\frac{1}{2} n\boldsymbol{i}\,(1 - \cos\varphi)\,d\theta$ and $\boldsymbol{\mathcal{A}}_S = \frac{1}{2}\,n\boldsymbol{i}(1 + \cos\varphi)\,d\theta$ are gauge potentials on this bundle. The connection $\boldsymbol{\omega}_n$ they determine can be described as follows: The $\mathrm{Im}\,\mathbb{C}$-valued 1-form $\tilde{\boldsymbol{\omega}}_n$ on $\mathbb{C}^2$ given by $\tilde{\boldsymbol{\omega}}_n = \boldsymbol{i}\,n\,\mathrm{Im}(\bar{z}^1 dz^1 + \bar{z}^2 dz^2)$ restricts to an $\mathrm{Im}\,\mathbb{C}$-valued 1-form $\iota^*\tilde{\boldsymbol{\omega}}_n$ on S^3 that is invariant under the $\mathbb{Z}_n$-action and so descends to an $\mathrm{Im}\,\mathbb{C}$-valued 1-form $\boldsymbol{\omega}_n$ on $S^3/\,\mathbb{Z}_n$. $\boldsymbol{\omega}_n$ is the required connection on $U(1) \hookrightarrow S^3/\,\mathbb{Z}_n \xrightarrow{\mathcal{P}_n} S^2$.

$$\boxed{n < -1}$$

Here the construction is exactly the same as above for $n > 1$ except that $\mathcal{P}_1$ is replaced by $\mathcal{P}_{-1}$.

2.3 Spin Zero Electrodynamics

We consider next our first example of a gauge theory in which matter fields are present. The gauge field will be an electromagnetic field of the type discussed in Section 2.2. The matter field coupled to this gauge field will represent a charged particle experiencing the effects of the electromagnetic field. According to the General Framework described in Section 2.1, the description of such a particle will require a vector space $\mathcal{V}$ and a representation $\rho : U(1) \longrightarrow GL(\mathcal{V})$ of $U(1)$ on $\mathcal{V}$. Now, in physics, charged particles have wavefunctions with a certain number of complex components, the number of such components being determined by the particle's spin s. Specifically, s is an element of $\{0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots\}$ and the wavefunction of a particle with spin s

has $2s+1$ components. The simplest case is that of a particle of spin 0 (e.g., a π^- meson) for which the wavefunction takes values in $\mathbb{C}$. This is the case we consider here.

Remark: Certain technical complications arise when $s > 0$. For example, an electron has $s = \frac{1}{2}$ and, according to Dirac, the corresponding matter field is defined, not on a $U(1)$-bundle over X, but on a certain $SL(2,\mathbb{C})$-bundle over X called a spinor bundle. Nevertheless, an electron responds to an electromagnetic, i.e., a $U(1)$-gauge, field. To fit this into our General Framework would require "splicing" the two bundles together into a single bundle on which both objects may be thought to live. We will return to these issues in the next section.

We will adopt the notation of Section 2.1. Thus, X is an open submanifold of $\mathbb{R}^{1,3}$ (standard coordinates x^0, x^1, x^2, x^3) and $\boldsymbol{\omega}$ is a connection on a principal $U(1)$-bundle

$$U(1) \hookrightarrow P \xrightarrow{\mathcal{P}} X$$

over X with curvature $\boldsymbol{\Omega}$. For any cross-section s we write

$$s^*\boldsymbol{\omega} = \boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}_\alpha\, dx^\alpha = -\boldsymbol{i}\,\boldsymbol{A} = -\boldsymbol{i}A_\alpha\, dx^\alpha$$

and

$$s^*\boldsymbol{\Omega} = \boldsymbol{\mathcal{F}} = \frac{1}{2}\mathcal{F}_{\alpha\beta}\, dx^\alpha \wedge dx^\beta = -\boldsymbol{i}\,\boldsymbol{F} = -\frac{1}{2}\boldsymbol{i}F_{\alpha\beta}\, dx^\alpha \wedge dx^\beta,$$

where $\mathcal{F}_{\alpha\beta} = \partial_\alpha \mathcal{A}_\beta - \partial_\beta \mathcal{A}_\alpha = -\boldsymbol{i}(\partial_\alpha A_\beta - \partial_\beta A_\alpha)$. Now we let $\mathcal{V} = \mathbb{C}$ (as a 2-dimensional real vector space). The usual inner product on $\mathbb{C} = \mathbb{R}^2$ is given by

$$\langle z_1, z_2\rangle = \frac{1}{2}(z_1\bar{z}_2 + z_2\bar{z}_1).$$

For each integer n we define a representation $\rho_n : U(1) \longrightarrow GL(\mathbb{C})$ by

$$\rho_n(g)(z) = g \cdot z = g^n z,$$

where we identify an element of $U(1)$ with a complex number of modulus 1 (these are, in fact, all of the "irreducible" representations of $U(1)$ on $\mathbb{C}$). Notice that ρ_n is orthogonal with respect to $\langle\ ,\ \rangle$ since

$$\begin{aligned}
\langle \rho_n(g)(z_1), \rho_n(g)(z_2)\rangle &= \langle g^n z_1, g^n z_2\rangle \\
&= \frac{1}{2}\left((g^n z_1)\,(\overline{g^n z_2}\,) + (g^n z_2)\,(\overline{g^n z_1}\,)\right) \\
&= \frac{1}{2}\left(g^n(z_1\bar{z}_2)g^{-n} + g^n(z_2\bar{z}_1)g^{-n}\right) \\
&= \frac{1}{2}\left(z_1\bar{z}_2 + z_2\bar{z}_1\right) \\
&= \langle z_1, z_2\rangle.
\end{aligned}$$

The potential function $U : \mathbb{C} \longrightarrow \mathbb{R}$ (#7 of the General Framework) is taken to be

$$U(z) = \frac{1}{2}m\langle z, z\rangle = \frac{1}{2}m|\, z\,|^2 = \frac{1}{2}mz\bar{z}$$

where $m \geq 0$ is a constant. Since ρ_n is orthogonal with respect to $\langle\ ,\ \rangle$, U is invariant under the action of $U(1)$ on $\mathbb{C}$, as required.

A matter field of type ρ_n is a $\mathbb{C}$-valued map ϕ on P that is equivariant with respect to the actions of $U(1)$ on P and $\mathbb{C}$, i.e., that satisfies

$$\phi(p \cdot g) = g^{-n}\phi(p)$$

for each $p \in P$ and $g \in U(1)$. The connection $\boldsymbol{\omega}$ determines a covariant exterior derivative $d^{\boldsymbol{\omega}}\phi$ of ϕ and the action $A(\boldsymbol{\omega}, \phi)$ contains an appropriate squared norm $\|\, d^{\boldsymbol{\omega}}\phi\|^{\,2}$ arising from the Minkowski metric on X and the invariant inner product $\langle\ ,\ \rangle$ on $\mathbb{C}$. We will defer until later the general procedure for constructing such norms and simply indicate at this point what the result is for the case under consideration and why it is the "natural" choice. For any local cross-section s, $s^*\phi = \phi \circ s$ and, for convenience, we will write $\phi = \phi(x^0, x^1, x^2, x^3)$ for the standard coordinate representation for $s^*\phi$. The corresponding coordinate expression for the pullback of $d^{\boldsymbol{\omega}}\phi$ is

$$(\partial_\alpha\phi - \boldsymbol{i}n\, A_\alpha\phi)\, dx^\alpha$$

(Example #1, page 52). Each $\partial_\alpha\phi - \boldsymbol{i}n\, A_\alpha\phi$ is a function with values in $\mathbb{C}$. The invariant inner product $\langle\ ,\ \rangle$ on $\mathbb{C}$ and the Minkowski inner product $(g(v, w) = v^0w^0 - v^1w^1 - v^2w^2 - v^3w^3 = \eta_{\alpha\beta}v^\alpha w^\beta = \eta^{\alpha\beta}v_\alpha w_\beta$, where $v_\alpha = \eta_{a\alpha}v^a$ and $w_\beta = \eta_{b\beta}w^b$) combine to give the squared norm

$$\begin{aligned}
\|\, d^{\boldsymbol{\omega}}\phi\,\|^2 &= \eta^{\alpha\beta}(\partial_\alpha\phi - \boldsymbol{i}nA_\alpha\phi)\Big(\overline{\partial_\beta\phi - \boldsymbol{i}nA_\beta\phi}\Big) \\
&= \eta^{\alpha\beta}(\partial_\alpha\phi - \boldsymbol{i}nA_\alpha\phi)\ (\partial_\beta\bar{\phi} + \boldsymbol{i}nA_\beta\bar{\phi}) \quad (A_\beta \text{ is real}) \\
&= \ (\partial_\alpha\phi - \boldsymbol{i}nA_\alpha\phi)\ (\eta^{\alpha\beta}(\partial_\beta\bar{\phi}) + \boldsymbol{i}n(\eta^{\alpha\beta}A_\beta)\bar{\phi}) \\
&= (\partial_\alpha\phi - \boldsymbol{i}nA_\alpha\phi)\ (\partial^\alpha\bar{\phi} + \boldsymbol{i}nA^\alpha\bar{\phi}),
\end{aligned}$$

where we have written

$$\partial^\alpha = \eta^{\alpha\beta}\partial_\beta, \qquad \alpha = 0, 1, 2, 3$$

(so that $\partial^0 = \partial_0,\ \partial^1 = -\partial_1,\ \partial^2 = -\partial_2$ and $\partial^3 = -\partial_3$) and

$$A^\alpha = \eta^{\alpha\beta}A_\beta, \qquad \alpha = 0, 1, 2, 3.$$

Although the expression we have given for $\|d^{\boldsymbol{\omega}}\phi\|^2$ is local and appears to depend on the choice of the cross-section s, it is, in fact, gauge invariant and therefore determines a globally defined, real-valued function on X. To see

this we observe the following: We have already seen (page 59) that a gauge transformation g can be written $g(x) = e^{-\boldsymbol{i}\Lambda(x)}$ and has the following effects on the potential $\boldsymbol{A}$ and the matter field ϕ:

$$\begin{aligned} \boldsymbol{A} &\longrightarrow \boldsymbol{A}^g = \boldsymbol{A} + d\Lambda \\ \phi &\longrightarrow \phi^g = g^{-1} \cdot \phi = e^{\boldsymbol{i}\, n\Lambda}\phi. \end{aligned}$$

Thus,

$$\begin{aligned} \partial_\alpha \phi^g - \boldsymbol{i} n (A^g)_\alpha \phi^g &= \partial_\alpha (e^{\boldsymbol{i}\, n\Lambda}\phi) - \boldsymbol{i} n (A_\alpha + \partial_\alpha \Lambda)(e^{\boldsymbol{i}\, n\Lambda}\phi) \\ &= e^{\boldsymbol{i}\, n\Lambda}\partial_\alpha \phi + \boldsymbol{i} n e^{\boldsymbol{i}\, n\Lambda}(\partial_\alpha \Lambda)\phi - \boldsymbol{i} n A_\alpha e^{\boldsymbol{i}\, n\Lambda}\phi \\ &\qquad - \boldsymbol{i} n (\partial_\alpha \Lambda) e^{\boldsymbol{i}\, n\Lambda}\phi \\ &= e^{\boldsymbol{i}\, n\Lambda}(\partial_\alpha \phi - \boldsymbol{i} n A_\alpha \phi) \end{aligned}$$

and, similarly,

$$\partial^\alpha \overline{\phi^g} + \boldsymbol{i} n (A^g)^\alpha \overline{\phi^g} = e^{-\boldsymbol{i}\, n\Lambda}(\partial^\alpha \bar{\phi} + \boldsymbol{i} n A^\alpha \bar{\phi}).$$

Consequently,

$$\begin{aligned} &(\partial_\alpha \phi^g - \boldsymbol{i} n (A^g)_\alpha \phi^g)\left(\partial^\alpha \overline{\phi^g} + \boldsymbol{i} n (A^g)^\alpha \overline{\phi^g}\right) \\ &\quad = (\partial_\alpha \phi - \boldsymbol{i} n A_\alpha \phi)\left(\partial^\alpha \bar{\phi} + \boldsymbol{i} n A^\alpha \bar{\phi}\right) \end{aligned}$$

as required. With this and an appropriate choice of normalizing and coupling constants, we take our action functional to be

$$\begin{aligned} A(\boldsymbol{\omega}, \phi) = \int_X \Big[&-\frac{1}{4} F_{\alpha\beta} F^{\alpha\beta} + \frac{1}{2}(\partial_\alpha \phi - \boldsymbol{i} n A_\alpha \phi)\ (\partial^\alpha \bar{\phi} + \boldsymbol{i} n A^\alpha \bar{\phi}) \\ &+ \frac{1}{2} m \phi \bar{\phi} \Big]\, dx^0 dx^1 dx^2 dx^3. \end{aligned}$$

The corresponding Euler-Lagrange equations are

$$\begin{aligned} (\partial_\alpha - \boldsymbol{i} n A_\alpha)\ (\partial^\alpha - \boldsymbol{i} n A^\alpha)\, \phi + m^2 \phi &= 0 \\ \partial_\alpha F^{\alpha\beta} &= 0, \quad \beta = 0, 1, 2, 3 \end{aligned}$$

(see [**Bl**]). The first of these is the **Klein-Gordon equation** (for a spin zero particle of mass m and charge n interacting with a gauge field $\boldsymbol{\mathcal{F}} = -\boldsymbol{i}\boldsymbol{F} = -\boldsymbol{i} d\boldsymbol{A}$ determined by the local gauge potentials $\boldsymbol{\mathcal{A}} = -\boldsymbol{i} A_\alpha dx^\alpha$). The second equation is equivalent to $d^* \boldsymbol{F} = 0$ (see page 62). Since $-\boldsymbol{i}\boldsymbol{F}$ is the pullback of a curvature form, the Bianchi identity gives $d\boldsymbol{F} = 0$ also so $\boldsymbol{F}$ satisfies Maxwell's equations.

Remark: Of course, it was our stated intention to model a spin zero particle in an electromagnetic field, but the point here is that the electromagnetic

nature of the field $\boldsymbol{F}$ that appears in the action $A(\boldsymbol{\omega}, \phi)$ is necessitated by the Euler-Lagrange equations. We do not have to impose Maxwell's equations "by hand." Also note that conjugating the Klein-Gordon equation gives

$$(\partial_\alpha + \boldsymbol{i}nA_\alpha)(\partial^\alpha + \boldsymbol{i}nA_\alpha)\bar{\phi} + m^2\bar{\phi} = 0.$$

Thus, if ϕ represents a Klein-Gordon field of mass m and charge n, $\bar{\phi}$ represents a Klein-Gordon field of mass m and charge $-n$. Physicists view ϕ and $\bar{\phi}$ as wavefunctions for a particle/antiparticle pair.

Taking $\boldsymbol{A}$ to be zero above (i.e., "turning off" the electromagnetic field) gives the **free Klein-Gordon equation**

$$\partial_\alpha\partial^\alpha\phi + m^2\phi = 0$$

for a spin zero particle.

Remark: Free objects in physics are also subject to field equations. For example, a free particle of mass m in Newtonian mechanics satisfies $\frac{d}{dt}(m\vec{v}) = \vec{0}$, while in quantum mechanics its wavefunction ψ satisfies the Schroedinger equation

$$-\frac{1}{2m}\vec{\nabla}^2\psi = \boldsymbol{i}\frac{\partial\psi}{\partial t}.$$

Note also that if ϕ satisfies the free Klein-Gordon equation, then so does $\bar{\phi}$.

We will not pursue the business of seeking solutions to these equations and sorting out their physical significance. Indeed, it would seem that no plausible physical interpretation of such solutions exists outside the context of quantum field theory and for this the only service we can perform for our reader is a referral to the physics literature (e.g., **[Gui]**, **[Ry]**, or **[Wein]**). We will, however, spend a few moments describing an alternative approach to the Klein-Gordon equation which more clearly exposes the philosophical underpinnings of modern gauge theory. Physicists refer to this approach as "minimal coupling" and it begins with the free particle equation

$$\partial_\alpha\partial^\alpha\phi + m^2\phi = 0.$$

This equation itself might be arrived at by "quantizing" the classical relativistic relation $E^2 = \vec{p}^2 + m^2$ between energy and momentum ("quantization" is the mystical process of replacing classical quantities such as these with "corresponding operators" that act on the wavefunction). One then "couples" the particle to the field by replacing the "ordinary derivatives" ∂_α by "covariant derivatives" $\partial_\alpha - \boldsymbol{i}nA_\alpha$ involving the potentials A_α for the field. This, of course, does give the full Klein-Gordon equation, but the motivation is no doubt obscure. Our search for motivation leads to the very early days of quantum mechanics.

In old-fashioned (non-relativistic) quantum mechanics a charged particle (of mass m and charge n) in an electromagnetic field has a wavefunction ψ that satisfies the Schroedinger equation

$$\frac{1}{2m}\left(-\boldsymbol{i}\vec{\nabla}-n\vec{A}\right)^2\ \psi=\left(\boldsymbol{i}\frac{\partial}{\partial t}-nV\right)\psi, \tag{2.3.1}$$

where $(V,\vec{A})=(A^0,A^1,A^2,A^3)=(A_0,-A_1,-A_2,-A_3)$ and $\boldsymbol{A}=A_\alpha dx^\alpha$ satisfies $d\boldsymbol{A}=\boldsymbol{F}$ (see page 63).

Remark: The meaning of $(-\boldsymbol{i}\vec{\nabla}-n\vec{A})^2$ as an operator on ψ is as follows:

$$\begin{aligned}\left(-\boldsymbol{i}\vec{\nabla}-n\vec{A}\right)^2\psi&=\left(-\boldsymbol{i}\vec{\nabla}-n\vec{A}\right)\cdot\left(-\boldsymbol{i}\vec{\nabla}-n\vec{A}\right)\psi\\&=\left(-\boldsymbol{i}\vec{\nabla}-n\vec{A}\right)\cdot\left(-\boldsymbol{i}\vec{\nabla}\psi-n\ \psi\vec{A}\right)\\&=-\vec{\nabla}^2\psi+n\boldsymbol{i}\vec{\nabla}\cdot(\psi\vec{A})+n\boldsymbol{i}\vec{A}\cdot(\vec{\nabla}\psi)\\&\qquad+n^2|\,\vec{A}\,|^2\psi.\end{aligned}$$

Now, ψ takes values in $\mathbb{C}$ and so, at each point, has a modulus and a phase ($\psi=re^{\boldsymbol{i}\theta}$). There is some arbitrariness in the phase, however, since, if a is an element of $U(1)$ (identified with a complex number of modulus 1), then $a\psi$ satisfies (2.3.1) whenever ψ does (being a constant, a just slips outside of all the derivatives in (2.3.1)). Moreover, $a\psi$ differs from ψ only in the phase factor (since $|a|=1$) and so $|a\psi|^2=|\psi|^2$. Since all of the physically significant probabilities in quantum mechanics depend only on this squared modulus, ψ and $a\psi$ should represent the same physical object. This freedom to alter the phase of ψ is quite restricted, however. Since a must be constant, any phase shift in the wavefunction must be accomplished at all spatial locations simultaneously. Such a global phase shift, however, violates both the spirit and the letter of relativistic law (you can't do anything "at all spatial locations simultaneously"). Nevertheless, it is difficult to shake the feeling that the physical significance of ψ "should" persist under some sort of phase shift (again, because squared moduli will be unaffected). Notice that the relativistic objection to a phase shift would disappear if one allowed the phase to shift independently at each spacetime point, i.e., if one replaced ψ by $a\psi$, where a is now a function of (x,y,z,t) taking values in $U(1)$. The problem with this, of course, is that, if a is not constant, it will not simply "slip outside" of all the derivatives in (2.3.1). Indeed, product rules will generate all sorts of new terms that do not cancel so there is no reason to suppose that $a\psi$ will even be a solution to (2.3.1). A bit (actually, quite a bit) of vector calculus will, in fact, establish the following: Let ψ be a solution to (2.3.1) and let $\Lambda(x,y,z,t)$ be any smooth, real-valued function. Then

$$\psi'=e^{\boldsymbol{i}\,n\,\Lambda}\psi$$

is a solution to

$$\frac{1}{2m}\left(-\boldsymbol{i}\vec{\nabla} - n\left(\vec{A} + \vec{\nabla}\Lambda\right)\right)^2 \psi' = \left(\boldsymbol{i}\frac{\partial}{\partial t} - n\left(V - \frac{\partial \Lambda}{\partial t}\right)\right)\psi'. \qquad (2.3.2)$$

This last result is interesting from a number of different points of view. Physicists in the early part of this century found in it some confirmation of their long held belief that the nonuniqueness of the potential for an electromagnetic field and the nonuniqueness of the phase for a wavefunction were both matters of no physical consequence. The reasoning went something like this: In equation (2.3.1) the electromagnetic field to which ψ is responding is described by the potential $(V, \vec{A})$. The corresponding 1-form is $\boldsymbol{A} = A_\alpha dx^\alpha$, where $A_0 = V$ and $(A_1, A_2, A_3) = -\vec{A}$ (see page 63). In equation (2.3.2), $(V, \vec{A})$ is replaced by $(V - \frac{\partial \Lambda}{\partial t}, \vec{A} + \vec{\nabla}\Lambda)$ and the corresponding 1-form is $\boldsymbol{A} - d\Lambda$ which, of course, describes the same electromagnetic field $\boldsymbol{F} = d\boldsymbol{A} = d(\boldsymbol{A} - d\Lambda)$ since $d^2 = 0$ (see page 63). Classically, the potential was regarded as a convenient computational device, but with no physical significance of its own outside of the fact that it gives rise via differentiation to the electromagnetic field. Thus, (2.3.1) and (2.3.2) should, in some sense, be "equivalent" (describe the same physics). Now, the solutions to (2.3.1) and (2.3.2) are in one-to-one correspondence ($\psi \leftrightarrow e^{\boldsymbol{i}\, n\Lambda}\psi$) and the corresponding functions differ only by the phase factor $e^{\boldsymbol{i}\, n\Lambda}$. Since $|\psi|^2 = |e^{\boldsymbol{i}\, n\Lambda}\psi|^2$ at each point, all of the usual probabilities calculated for ψ and $e^{\boldsymbol{i}\, n\Lambda}\psi$ in quantum mechanics are the same so that these should be two descriptions of the same physical object. Everything seems to fit together quite nicely.

This placid scene was disturbed in the late 1950s when Aharonov and Bohm [**AB**] suggested that, while the phase of a single charge may well be unmeasurable, the *relative* phase of two charged particles that interact should have observable consequences. Their proposed experiment (later confirmed by Chambers in 1960) is described on page 6 of [**N4**]. The potential had now to be taken seriously as a physical field and not merely a mathematical contrivance. This being the case, one is no longer free to regard a phase shift $\psi \longrightarrow e^{\boldsymbol{i}\, n\Lambda}\psi$ as devoid of physical content. Indeed, physicists have gone to quite the other extreme and elevated the invariance of electrodynamics under such local phase transformations to the status of a basic physical principle (this is an instance of the so-called "gauge principle" which lies at the heart of modern gauge theory). To see more clearly the consequences of such a principle we consider the special cases of (2.3.1) and (2.3.2) in which the electromagnetic field is "turned off."

If the electromagnetic field is zero one may choose the potential $\boldsymbol{A}$ for which $(V, \vec{A}) = (0, \vec{0})$ so that (2.3.1) becomes the usual free Schroedinger equation

$$\frac{1}{2m}(-\boldsymbol{i}\vec{\nabla})^2\,\psi = \boldsymbol{i}\frac{\partial}{\partial t}\,\psi. \qquad (2.3.3)$$

On the other hand, the same (trivial) electromagnetic field is described by any potential of the form $(0 - \frac{\partial \Lambda}{\partial t}, \vec{0} + \vec{\nabla}\Lambda)$. Begging the indulgence of the reader we would now like to call this $(V, \vec{A})$ and write (2.3.2) as

$$\frac{1}{2m}\left(-\boldsymbol{i}\,\vec{\nabla} - n\,\vec{A}\right)^2 \psi' = \left(\boldsymbol{i}\frac{\partial}{\partial t} - nV\right)\psi'. \tag{2.3.4}$$

Now suppose one adopts as a basic physical principle that electrodynamics should be invariant under local phase shifts. In particular, then equations (2.3.3) and (2.3.4), with $(V, \vec{A}) = (-\frac{\partial \Lambda}{\partial t}, \vec{\nabla}\Lambda)$, are equivalent. Physicists take the following rather remarkable view of this: Even in a vacuum (electromagnetic field zero) the requirement of local phase shift invariance (gauge invariance) necessitates the existence of a physical field $\boldsymbol{A}$ (not the electromagnetic field – that's zero – but the gauge potential field) whose task is to counteract, or balance, the effects of any phase shift and keep the physics invariant. Algebraically, this happens by adding to the operators that appear in the Schroedinger equation terms whose sole purpose is to cancel the extra stuff you get from product rules for $e^{\boldsymbol{i}\,n\,\Lambda}\psi$ when Λ is not constant. The marvelous thing about this way of viewing the situation is that it provides a completely mindless way of ensuring gauge invariance in all sorts of contexts. One need only fudge into the differential operators whatever it takes to cancel the offending extra terms that arise from the product rule. In electrodynamics it works just as well when the field is not turned off. Indeed, "turning the field on," i.e., coupling a particle to an electromagnetic field, can be viewed in exactly the same light. A free particle satisfies (2.3.3). To couple the particle to an electromagnetic field $\boldsymbol{F}$, write $\boldsymbol{F} = d\boldsymbol{A}$, where $\boldsymbol{A} = A_\alpha\, dx^\alpha, (A_0, A_1, A_2, A_3) = (A^0, -A^1, -A^2, -A^3) = (V, \vec{A})$ and make the substitutions

$$-\boldsymbol{i}\,\vec{\nabla} \longrightarrow -\boldsymbol{i}\,\vec{\nabla} - n\,\vec{A} \quad \text{and} \quad \boldsymbol{i}\frac{\partial}{\partial t} \longrightarrow \boldsymbol{i}\frac{\partial}{\partial t} - n\,V$$

to obtain

$$\frac{1}{2m}\left(-\boldsymbol{i}\,\vec{\nabla} - n\,\vec{A}\right)^2 \psi = \left(\boldsymbol{i}\frac{\partial}{\partial t} - n\,V\right)\psi$$

which is (2.3.1). The solutions to (2.3.1) will describe the wavefunction of the charged particle coupled to the field $\boldsymbol{F}$. The description is only one among many possible descriptions, each differing from the others by a choice of phase (gauge) determined by the particular $\boldsymbol{A}$ one has chosen for the representation $\boldsymbol{F} = d\boldsymbol{A}$.

Finally, we show that the operator substitutions described above assume a more elegant (and more familiar) form if we recast them in relativistic notation. Here we let

$$(\partial_0, \partial_1, \partial_2, \partial_3) = \left(\frac{\partial}{\partial x^0}, \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3}\right) = \left(\frac{\partial}{\partial t}, \vec{\nabla}\right)$$

and

$$(\partial^0, \partial^1, \partial^2, \partial^3) = (\partial_0, -\partial_1, -\partial_2, -\partial_3) = \left(\frac{\partial}{\partial t}, -\vec{\nabla}\right)$$

so

$$\begin{aligned}\left(\boldsymbol{i}\frac{\partial}{\partial t} - n\,V,\ -\boldsymbol{i}\,\vec{\nabla} - n\,\vec{A}\right) &= \boldsymbol{i}\,\left(\frac{\partial}{\partial t} + \boldsymbol{i}\,n\,V, -\vec{\nabla} + \boldsymbol{i}\,n\,\vec{A}\right) \\ &= \boldsymbol{i}\,\left(\partial^0 + \boldsymbol{i}\,n\,V,\ \partial^1 + \boldsymbol{i}\,n\,A^1,\right. \\ &\qquad \left.\partial^2 + \boldsymbol{i}\,n\,A^2, \partial^3 + \boldsymbol{i}\,n\,A^3\right).\end{aligned}$$

Since

$$\left(\boldsymbol{i}\,\frac{\partial}{\partial t}, -\boldsymbol{i}\vec{\nabla}\right) = \boldsymbol{i}\,\left(\frac{\partial}{\partial t}, -\vec{\nabla}\right) = \boldsymbol{i}\,\left(\partial^0, \partial^1, \partial^2, \partial^3\right),$$

the substitutions above simply amount to

$$\partial^\alpha \longrightarrow \partial^\alpha + \boldsymbol{i}\,n\,A^\alpha, \quad \alpha = 0, 1, 2, 3,$$

or, lowering the indices once again,

$$\partial_\alpha \longrightarrow \partial_\alpha + \boldsymbol{i}\,n\,A_\alpha, \quad \alpha = 0, 1, 2, 3$$

(the sign is + rather than − because of our decision to include the conventional minus sign in $\boldsymbol{\mathcal{A}} = -\boldsymbol{i}\,\boldsymbol{A}$).

We will conclude our discussion of the Klein-Gordon equation by describing its two most important invariance properties: gauge invariance and Lorentz invariance. We have already shown that the action $A(\boldsymbol{\omega}, \phi)$ is invariant under gauge transformations and it follows that the same is true of its Euler-Lagrange equations. Nevertheless, a direct proof is instructive. Thus, we consider the equation

$$(\partial_\alpha - \boldsymbol{i}\,n\,A_\alpha)\,(\partial^\alpha - \boldsymbol{i}\,n\,A^\alpha)\phi + m^2\phi = 0 \tag{2.3.5}$$

and a gauge transformation $g(x) = e^{-\boldsymbol{i}\,\Lambda(x)}$. The corresponding changes in the potential and the matter field are, as usual,

$$\boldsymbol{A} \longrightarrow \boldsymbol{A}^g = \boldsymbol{A} + d\,\Lambda$$

and

$$\phi \longrightarrow \phi^g = g^{-1} \cdot \phi = e^{\boldsymbol{i}\,n\,\Lambda}\phi.$$

Our objective is to show that, if ϕ satisfies (2.3.5), then

$$(\partial_\alpha - \boldsymbol{i}\,n\,(A^g)_\alpha)\,(\partial^\alpha - \boldsymbol{i}\,n\,(A^g)^\alpha)\,\phi^g + m^2\,\phi^g = 0. \tag{2.3.6}$$

First, we compute

$$
\begin{aligned}
(\partial^\alpha - \boldsymbol{i}\, n\, (A^g)_\alpha)\, \phi^g &= (\partial^\alpha - \boldsymbol{i}\, n\, (A^\alpha + \partial^\alpha \Lambda))\, \left(e^{\boldsymbol{i}\, n\, \Lambda} \phi\right) \\
&= \partial^\alpha\, (e^{\boldsymbol{i}\, n \Lambda} \phi) - \boldsymbol{i}\, n\, A^\alpha\, e^{\boldsymbol{i}\, n\, \Lambda} \phi - \boldsymbol{i}\, n\, \partial^\alpha \Lambda\, e^{\boldsymbol{i}\, n\, \Lambda} \phi \\
&= e^{\boldsymbol{i}\, n\, \Lambda} \partial^\alpha\, \phi + \boldsymbol{i}\, n\, e^{\boldsymbol{i}\, n\, \Lambda} \partial^\alpha \Lambda\, \phi \\
&\qquad - \boldsymbol{i}\, n\, A^\alpha\, e^{\boldsymbol{i}\, n\, \Lambda}\, \phi - \boldsymbol{i}\, n\, \partial^\alpha \Lambda\, e^{\boldsymbol{i}\, n\, \Lambda} \phi \\
&= e^{\boldsymbol{i}\, n\, \Lambda} (\partial^\alpha\, \phi - \boldsymbol{i}\, n\, A^\alpha\, \phi) \\
&= e^{\boldsymbol{i}\, n\, \Lambda} (\partial^\alpha - \boldsymbol{i}\, n\, A^\alpha)\, \phi.
\end{aligned}
$$

In the same way,

$$
(\partial_\alpha - \boldsymbol{i}\, n\, (A^g)_\alpha) \left(e^{\boldsymbol{i}\, n\, \Lambda} \phi\right) = e^{\boldsymbol{i}\, n\, \Lambda} (\partial_\alpha - \boldsymbol{i}\, n\, A^\alpha)\, \phi
$$

so

$$
\begin{aligned}
&(\partial_\alpha - \boldsymbol{i}\, n\, (A^g)_\alpha)\, (\partial^\alpha - \boldsymbol{i}\, n\, (A^g)^\alpha)\, \phi^g \\
&\quad = (\partial_\alpha - \boldsymbol{i}\, n\, (A^g)_\alpha) \left(e^{\boldsymbol{i}\, n\, \Lambda} (\partial^\alpha - \boldsymbol{i}\, n\, A^\alpha)\, \phi\right) \\
&\quad = e^{\boldsymbol{i}\, n\, \Lambda}\, (\partial_\alpha - \boldsymbol{i}\, n\, A_\alpha)(\partial^\alpha - \boldsymbol{i}\, n\, A^\alpha)\, \phi.
\end{aligned}
$$

From this it is clear that (2.3.5) implies (2.3.6). The Klein-Gordon equation is gauge invariant.

We wish to show next that the Klein-Gordon equation is "relativistically invariant." Roughly, this means that the equation has the same mathematical form in all inertial frames of reference, but the precise meaning of such a statement in general will require some discussion. In this section we will be content to spell out explicitly what is being asserted for the Klein-Gordon equation alone. When we turn to the Dirac equation in the next section we will describe precisely what is meant by "relativistic invariance" in general.

Thus far we have written the Klein-Gordon equation (2.3.5) only in standard coordinates x^0, x^1, x^2, x^3 for $\mathbb{R}^{1,3}$, i.e., only in one fixed inertial frame of reference. To emphasize this fact we will now write the derivatives ∂_α and ∂^α explicity as $\partial/\partial x^\alpha$ and $\eta^{\alpha\beta}\, \partial/\partial x^\beta$ so that (2.3.5) becomes

$$
\begin{aligned}
\eta^{\alpha\beta}\ \left(\frac{\partial}{\partial x^\alpha} - \boldsymbol{i}\, n\, A_\alpha(x^0, \ldots, x^3)\right)\ \left(\frac{\partial}{\partial x^\beta} - \boldsymbol{i}\, n\, \boldsymbol{A}_\beta(x^0,\right. \\
\left. \ldots, x^3)\right) \phi\, (x^0, \ldots, x^3) + m^2\, \phi\, (x^0, \ldots, x^3) = 0.
\end{aligned}
\tag{2.3.7}
$$

The basic postulate of Special Relativity is that one inertial frame of reference is as good as another. The coordinates y^0, y^1, y^2, y^3 for $\mathbb{R}^{1,3}$, supplied by another such frame of reference are assumed to be related to x^0, x^1, x^2, x^3 by

$$
y^\alpha = \Lambda^\alpha{}_\beta x^\beta, \quad \alpha = 0, 1, 2, 3,
$$

where $\Lambda = (\Lambda^{\alpha}{}_{\beta})$ is an element of the so-called "proper, orthochronous Lorentz group" $\mathcal{L}_+^{\uparrow} = \{\Lambda : \Lambda^{\top}\,\eta\,\Lambda = \eta,\, \det\Lambda = 1, \Lambda^0{}_0 \geq 1\}$, where η is on page 58.

Remark: The "general Lorentz group" $\mathcal{L}$ consists of those Λ that satisfy $\Lambda^{\top}\,\eta\,\Lambda = \eta$, and these relate arbitrary orthonormal bases in $\mathbb{R}^{1,3}$. The "orthochronous" condition $\Lambda^0{}_0 \geq 1$ ensures that Λ does not reverse time orientations, i.e., $x^0 \geq 0$ implies $y^0 = \Lambda^0{}_{\alpha}\,x^{\alpha} \geq 0$, for timelike or null vectors, while $\det\Lambda = 1$ ("proper") then guarantees that Λ does not reverse the orientation of the spatial coordinates (x^1, x^2, x^3).

The inverse of Λ is written $\Lambda^{-1} = (\Lambda_{\alpha}{}^{\beta})$ so that

$$x^{\beta} = \Lambda_{\alpha}{}^{\beta}\,y^{\alpha}, \quad \beta = 0, 1, 2, 3.$$

Furthermore, one has

$$\Lambda^{\alpha}{}_{\gamma}\,\Lambda^{\beta}{}_{\delta}\,\eta^{\gamma\delta} = \eta^{\alpha\beta}, \quad \alpha, \beta = 0, 1, 2, 3$$

and

$$\Lambda_{\alpha}{}^{\gamma}\,\Lambda_{\beta}{}^{\delta}\,\eta^{\alpha\beta} = \eta^{\gamma\delta}, \quad \gamma, \delta = 0, 1, 2, 3$$

(physical motivation for and basic properties of the Lorentz group are discussed in some detail in the Introduction and first three sections of Chapter 1 in [**N3**]).

The precise assertion we are making about the Klein-Gordon equation is as follows: If $\phi = \phi(x^0, \ldots, x^3)$ is a solution to (2.3.7) and if we define $\hat{\phi} = \hat{\phi}(y^0, \ldots, y^3)$ by

$$\hat{\phi}\,(y^0, \ldots, y^3) = \phi\,(\Lambda_{\alpha}{}^{0}\,y^{\alpha}, \ldots, \Lambda_{\alpha}{}^{3}\,y^{\alpha}),$$

then $\hat{\phi}$ satisfies

$$\eta^{\alpha\beta}\left(\frac{\partial}{\partial y^{\alpha}} - \boldsymbol{i}\,n\hat{A}_{\alpha}(y^0, \ldots, y^3)\right)\left(\frac{\partial}{\partial y^{\beta}} - \boldsymbol{i}\,n\hat{A}_{\beta}(y^0, \ldots, y^3)\right)\hat{\phi}\,(y^0, \ldots, y^3) + m^2\hat{\phi}\,(y^0, \ldots, y^3) = 0. \tag{2.3.8}$$

where $\boldsymbol{A} = \hat{A}_{\alpha}\,dy^{\alpha}$ so that $\hat{A}_{\alpha} = \boldsymbol{A}\left(\frac{\partial}{\partial y^{\alpha}}\right) = \boldsymbol{A}\left(\frac{\partial}{\partial x^{\beta}}\,\frac{\partial x^{\beta}}{\partial y^{\alpha}}\right) = \Lambda_{\alpha}{}^{\beta}\boldsymbol{A}\left(\frac{\partial}{\partial x^{\beta}}\right) = \Lambda_{\alpha}{}^{\beta}A_{\beta}$.

Remark: The essential, but rather camouflaged, issue here is that it is up to us to *specify* what the wavefunction is to be in the new coordinates (i.e., how ϕ transforms under $\mathcal{L}_+^{\uparrow}$) and that, in order to satisfy the requirements of special relativity, we must do this in such a way that it satisfies "the same equation in the new coordinate system." In this case, the wavefunction transforms trivially

(physicists say, as a scalar), i.e., we simply rewrite ϕ in the new coordinates, but there is no reason to expect such simplicity in general (compare this, for example, with the transformation of a vector field on $\mathbb{R}^3$ under rotation of the coordinate system).

Having spelled out exactly what we mean by relativistic invariance in this case, the proof is just a simple calculation. First observe that

$$\begin{aligned}
&\left(\frac{\partial}{\partial y^\beta} - \boldsymbol{i}\, n\hat{A}_\beta(y^0,\ldots,y^3)\right)\hat{\phi}\,(y^0,\ldots,y^3)\\
&\quad= \frac{\partial}{\partial y^\beta}\,\hat{\phi}\,(y^0,\ldots,y^3) - \boldsymbol{i}\, n\hat{A}_\beta\,(y^0,\ldots,y^3)\,\hat{\phi}\,(y^0,\ldots,y^3)\\
&\quad= \frac{\partial}{\partial y^\beta}\,\phi\,(\Lambda_a{}^0\, y^a,\ldots) - \boldsymbol{i}\, n\hat{A}_\beta\,(y^0,\ldots)\,\phi\,(\Lambda_a{}^0\, y^a,\ldots)\\
&\quad= \frac{\partial}{\partial x^\delta}\,\phi\,(\Lambda_a{}^0 y^a,\ldots)\frac{\partial x^\delta}{\partial y^\beta} - \boldsymbol{i}\, n\,\Lambda_\beta{}^\delta A_\delta\,(\Lambda_a{}^0 y^a,\ldots)\,\phi\,(\Lambda_a{}^0 y^a,\ldots)\\
&\quad= \Lambda_\beta{}^\delta\left(\frac{\partial}{\partial x^\delta} - \boldsymbol{i}\, nA_\delta(\Lambda_a{}^0\, y^a,\ldots)\right)\phi\,(\Lambda_a{}^0\, y^a,\ldots).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left(\frac{\partial}{\partial y^\alpha} - \boldsymbol{i}\, n\hat{A}_\alpha(y^0,\ldots,y^3)\right)\left(\frac{\partial}{\partial y^\beta} - \boldsymbol{i} n\hat{A}_\beta(y^0,\ldots,y^3)\right)\hat{\phi}\,(y^0,\ldots,y^3)\\
&\quad= \Lambda_\alpha{}^\gamma\,\Lambda_\beta{}^\delta\left(\frac{\partial}{\partial x^\gamma} - \boldsymbol{i}\, nA_\gamma\ (\Lambda_a{}^0\, y^a,\ldots)\right)\left(\frac{\partial}{\partial x^\delta} - \boldsymbol{i}\, nA_\delta(\Lambda_a{}^0\, y^a,\ldots)\right)\\
&\qquad\qquad\times\,\phi\,(\Lambda_a{}^0\, y^a,\ldots).
\end{aligned}$$

Since $\eta^{\alpha\beta}\,\Lambda_\alpha{}^\gamma\,\Lambda_\beta{}^\delta = \eta^{\gamma\delta}$, we have

$$\begin{aligned}
&\eta^{\alpha\beta}\left(\frac{\partial}{\partial y^\alpha} - \boldsymbol{i}\, n\hat{A}_\alpha(y^0,\ldots,y^3)\right)\left(\frac{\partial}{\partial y^\beta} - \boldsymbol{i}\, n\hat{A}_\beta(y^0,\ldots,y^3)\right)\hat{\phi}(y^0,\ldots,y^3)\\
&\quad= \eta^{\gamma\delta}\left(\frac{\partial}{\partial x^\gamma} - \boldsymbol{i}\, nA_\gamma(x^0,\ldots,x^3)\right)\left(\frac{\partial}{\partial x^\delta} - \boldsymbol{i}\, nA_\delta(x^0,\ldots,x^3)\right)\phi(x^0,\ldots,x^3)
\end{aligned}$$

and from this it is clear that (2.3.7) implies (2.3.8).

2.4 Spin One-Half Electrodynamics

Electrons, protons and neutrons (unlike the π mesons of Section 2.3) have spin $s = \frac{1}{2}$ and so, according to the generally accepted scheme of (nonrelativistic) quantum mechanics, should have a wavefunction with $2\,(\frac{1}{2}) + 1 = 2$ components. We intend to say just a few words on the rationale behind this

and then refer those who are interested to the elegant, and quite accessible, account of these phenomena in Volume III of *The Feynman Lectures on Physics* [**Fey**].

The classical Bohr picture of an atom (negatively charged electrons revolving around a positively charged nucleus) suggests that an orbiting electron actually constitutes a tiny current loop. Such a current loop produces a magnetic field which, at large distances, is the same as that of a magnetic dipole (located at the center of the loop and perpendicular to the plane of the loop). Such a dipole has a magnetic moment (a vector describing its orientation and strength). Consequently, an electron in an atom has associated with it an "orbital magnetic moment." Now, magnetic moments behave in interesting and well-understood ways when subjected to external magnetic fields. In 1922 (just before the advent of quantum mechanics), Stern and Gerlach carried out an experiment designed to detect these effects for the orbital magnetic moment of an electron. From the point of view of classical physics (the only point of view available at the time), the results were quite shocking. Somewhat later, the quantum mechanics of Schroedinger and Heisenberg, when applied to the orbital magnetic moment of the electron, provided a qualitative, but not quantitative explanation of the outcome. Finally, it was suggested by Uhlenbeck and Goudsmit that this discrepancy (and various others associated with the anomalous Zeeman effect and the splitting of certain spectral lines) could be accounted for if one assumed that the electron had associated with it an additional magnetic moment, not arising from its orbital motion, but rather from a sort of "intrinsic" angular momentum or "spinning" of the electron. The suggestion was not that an electron actually spins on some axis in the same way that the earth does on its, but rather that it possesses some intrinsic property (called "spin") that manifests itself in an external magnetic field by mimicing the behavior of the magnetic moment of a spinning charged ball. This intrinsic magnetic moment vector, however, must be of a rather peculiar sort that one could only encounter in quantum mechanics. The Stern-Gerlach experiment suggested that its component in *any* spatial direction could take on only one of two possible values (± 1 with the proper choice of units). This has the following consequence. Let us select (arbitrarily) some direction in space (say, the z-direction of some coordinate system). The intrinsic magnetic moment of an electron has, at each point, a z-component σ_z that can take on one of the two values ± 1. Which value it has will determine how the electron responds to certain magnetic fields and so a complete description of the electron's wavefunction must contain this information. More precisely, the wavefunction must be regarded as a function of not only x, y, z and t, but of σ_z as well.

$$\psi = \psi(x, y, z, t, \sigma_z)$$

However, since σ_z can assume only the two values ± 1, such a wavefunction is equivalent to a *pair* of functions $\psi_1(x, y, z, t) = \psi(x, y, z, t, 1)$ and $\psi_2(x, y, z, t) = \psi(x, y, z, t, -1)$. It is convenient to put these two together into

a column vector and adopt the point of view that an electron (or any spin one-half particle) has a two-component wavefunction

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

The probabilistic interpretations of the wavefunctions in quantum mechanics now run as follows: For each fixed t and each region $R \subseteq \mathbb{R}^3$, $\int_R \psi_1 \bar{\psi}_1$ is the probability at time t that the particle will be detected in R with its spin vector directed in the *positive* z-direction; $\int_R \psi_2 \bar{\psi}_2$ is the probability that it will be found in R with its spin vector in the *negative* z-direction; $\int_R \psi_1 \bar{\psi}_1 + \psi_2 \bar{\psi}_2$ is the probability that it will be found in R at all.

Pauli formulated a theory of the electron along the lines suggested above and, although this work was later superceded by that of Dirac, it provides an instructive warm-up and we will spend a few moments outlining some of its essential features. We will consider only stationary states $\Psi(x, y, z, t) = \psi(x, y, z)\, e^{-\boldsymbol{i}\omega t}$ and will focus our attention on the spatial part $\psi(x, y, z)$ (a proper treatment of time dependence should be relativistic, which Pauli's theory was not). Thus, we are interested in the two-component object

$$\psi(x, y, z) = \begin{pmatrix} \psi_1(x, y, z) \\ \psi_2(x, y, z) \end{pmatrix}.$$

Here x, y and z presumably represent "standard" coordinates in $\mathbb{R}^3$ and we (along with Pauli) will require that our theory be independent of which particular oriented, orthonormal basis for $\mathbb{R}^3$ gives rise to these coordinates, i.e., that it be invariant under the rotation group $SO(3)$. We are therefore led to ask the following question: How is the two-component wavefunction $\left(\begin{smallmatrix} \psi_1 \\ \psi_2 \end{smallmatrix}\right)$ transformed if the coordinate system is subjected to the rotation corresponding to some element of $SO(3)$?

To answer this question let us first review the manner in which such issues are addressed in classical physics and then decide what, if any, modifications are required by quantum mechanics. If the coordinate system is subjected to a rotation $R \in SO(3)$, then the state of the electron will be described by a pair $\left(\begin{smallmatrix} \psi_1' \\ \psi_2' \end{smallmatrix}\right)$, where each ψ_i' is a function of the new coordinates x', y' and z'. Now, it is conceivable that each ψ_i' is simply ψ_i expressed in terms of these new coordinates. This is, indeed, what we found to be the case for the Klein-Gordon equation (see page 83). However, it is also conceivable that the dependence of $\left(\begin{smallmatrix} \psi_1' \\ \psi_2' \end{smallmatrix}\right)$ on $\left(\begin{smallmatrix} \psi_1 \\ \psi_2 \end{smallmatrix}\right)$ is more analogous to the transformation law for an ordinary vector, or tensor field on $\mathbb{R}^3$. We will not prejudge this issue here, but will simply make a few tentative assumptions about this dependence. We assume, for example, that ψ_1' and ψ_2' are linear functions of ψ_1 and ψ_2. The reason is that, presumably, any differential equations one might arrive at for

the wavefunction will be generalizations of the (linear) Schroedinger equation and should (at least as a first guess) themselves be linear. Thus, for some 2×2 complex matrix $T = T(R)$ we have

$$\begin{pmatrix} \psi_1' \\ \psi_2' \end{pmatrix} = T(R) \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}$$

(here it is understood that both sides have been written in terms of one of the two coordinate systems, $x,\ y,\ z$ or $x',\ y'$ and z'). Assuming also (for the moment) that the wavefunction is uniquely determined in each coordinate system we find that a rotation by $R_2 \in SO(3)$ followed by a rotation by $R_1 \in SO(3)$ must have the same effect as the rotation $R_1 R_2 \in SO(3)$. Thus, we must have

$$T(R_1 R_2) = T(R_1) T(R_2).$$

Similarly, each $T(R)$ must be invertible and satisfy

$$T(R^{-1}) = (T(R))^{-1}$$

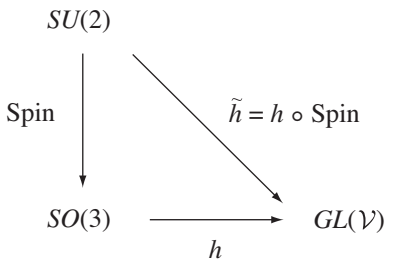
What we find then is that the rule T which associates with every $R \in SO(3)$ the corresponding transformation matrix $T(R)$ for $\left(\begin{smallmatrix} \psi_1 \\ \psi_2 \end{smallmatrix}\right)$ is a homomorphism into the group of invertible, 2×2, complex matrices. Identifying this latter group with $GL(\mathbb{C}^2)$ we find that T is a representation of $SO(3)$ on $\mathbb{C}^2$. The reason this information is useful is that all of the representations of $SO(3)$ are known. These are usually described somewhat indirectly as follows: In Appendix A of [**N4**] it is shown that $SU(2)$ is the (double) covering group of $SO(3)$. More precisely, there exists a smooth, surjective group homomorphism

$$\text{Spin} : SU(2) \longrightarrow SO(3)$$

with kernel $\pm \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and with the property that each point of $SO(3)$ has an open neighborhood V whose inverse image under Spin is a disjoint union of (two) open sets in $SU(2)$, each of which is mapped diffeomorphically onto V by Spin. Now, consider a representation

$$h : SO(3) \longrightarrow GL(\mathcal{V})$$

of $SO(3)$. Composing with Spin then gives a representation of $SU(2)$.
Every representation of $SO(3)$ "comes from" a representation of $SU(2)$. The converse is not true, however. That is, a given representation $\tilde{h} : SU(2) \longrightarrow GL(\mathcal{V})$ of $SU(2)$ will not induce a representation of $SO(3)$ unless $\tilde{h}(-g) = \tilde{h}(g)$ for every $g \in SU(2)$. The representations of $SU(2)$ that do *not* satisfy this condition are sometimes referred to in the physics literature as "2-valued representations of $SO(3)$," although they are not representations of $SO(3)$ at all, of course.



Now, it is quite easy to write out some rather obvious representations of $SU(2)$. Let $\mathbb{C}[z_1, z_2]$ be the vector space of all polynomials with complex coefficients in the two unknowns z_1 and z_2. For each $k = 0, 1, \ldots$, let $\mathcal{V}_k$ be the subspace consisting of those polynomials that are homogeneous of degree k:

$$c_0 {z_1}^k + c_1 {z_1}^{k-1} z_2 + \cdots + c_k {z_2}^k.$$

A basis for $\mathcal{V}_k$ consists of all polynomials ${z_1}^{k-r} {z_2}^r$, $r = 0, 1, \ldots, k$. Each $g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ in $SU(2)$ gives rise to a linear transformation on $\mathcal{V}_k$ which carries ${z_1}^{k-r} {z_2}^r$ onto $(z_1')^{k-r}(z_2')^r$, where

$$\begin{pmatrix} z_1' \\ z_2' \end{pmatrix} = \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

These linear transformations are clearly invertible and so the assignment, to each $g \in SU(2)$, of the corresponding element of $GL(\mathcal{V}_k)$ is a representation of $SU(2)$, usually denoted

$$D^{\frac{k}{2}} : SU(2) \longrightarrow GL(\mathcal{V}_k),$$

and called the **spin-j representation**, where $j = \frac{k}{2}$. One can show (see [**vdW**]) that each of these representations is *irreducible* (i.e., that there is no proper subspace of $\mathcal{V}_k$ that is invariant under every $D^{\frac{k}{2}}(g)$, $g \in SU(2)$) and that every irreducible representation of $SU(2)$ with complex representation space is equivalent to one of these (two representations $D_1 : G \longrightarrow GL(\mathcal{V}_1)$ and $D_2 : G \longrightarrow GL(\mathcal{V}_2)$ of a group G are *equivalent* if it is possible to choose bases for $\mathcal{V}_1$ and $\mathcal{V}_2$ so that, for each $g \in G$, the matrices of $D_1(g)$ and $D_2(g)$ are the same). Furthermore, any representation of $SU(2)$ can be constructed from these irreducible representations by forming finite direct sums (the *direct sum* of $D_1 : G \longrightarrow GL(\mathcal{V}_1)$ and $D_2 : G \longrightarrow GL(\mathcal{V}_2)$ is the representation $D_1 \oplus D_2 : G \longrightarrow GL(\mathcal{V}_1 \oplus \mathcal{V}_2)$ defined by $(D_1 \oplus D_2)(g)(v_1, v_2) = (D_1(g)(v_1), D_2(g)(v_2))$. In effect, we now have all of the representations of $SU(2)$.

The polynomials have now served their purpose and it will be convenient to note that $\mathcal{V}_k$ has complex dimension $k+1$ and so can be identified with $\mathbb{C}^{k+1}$ by identifying ${z_1}^{k-r}{z_2}^r$, $r = 0, 1, \ldots, k$, with the standard basis for $\mathbb{C}^{k+1}$. The linear transformations $D^{\frac{k}{2}}(g)$ can therefore be identified with $(k+1)\times(k+1)$ complex matrices. For example, $k = 0$ gives the trivial representation of $SU(2)$ on $\mathbb{C}$ ($D^0(g) = (1)$ for each $g \in SU(2)$), while $k = 1$ gives the identity representation of $SU(2)$ on $\mathbb{C}^2$ ($D^{\frac{1}{2}}(g) = g$ for every $g \in SU(2)$). Note that $D^{\frac{1}{2}}$ is the only irreducible representation of $SU(2)$ on $\mathbb{C}^2$. The only other way to get a representation of $SU(2)$ on $\mathbb{C}^2$ is to form the direct sum of two copies of D^0:

$$(D^0 \oplus D^0)(g) = (1) \oplus (1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This, of course, leaves everything in $\mathbb{C}^2$ alone. Finally notice that $D^{\frac{1}{2}}(-g) = -D^{\frac{1}{2}}(g)$ and $(D^0 \oplus D^0)(-g) = (D^0 \oplus D^0)(g)$ so only this second example descends to a representation of $SO(3)$ (the trivial representation of $SO(3)$ on $\mathbb{C}^2$).

The situation we have just described would seem to present us with something of a dilemma. To ensure the rotational invariance of Pauli's theory of the electron we were led to seek a representation of $SO(3)$ on $\mathbb{C}^2$ that would transform the two-component wavefunctions $\binom{\psi_1}{\psi_2}$ when the coordinate system is rotated. We find now that there is only one such ($D^0 \oplus D^0$) and this is the trivial representation. Under this representation the transformed wavefunction $\binom{\psi_1'}{\psi_2'}$ would simply be $\binom{\psi_1}{\psi_2}$ written in terms of the new coordinates. However, this is clearly not consistent with the phenomenon (spin one-half) which led us to two-component wavefunctions in the first place. Recall that $\psi_1(x, y, z) = \psi(x, y, z, 1)$ and $\psi_2(x, y, z) = \psi(x, y, z, -1)$, where ± 1 are the possible z-components of the intrinsic magnetic moment of the electron. A rotation which reverses the direction of the z-axis must interchange ψ_1 and ψ_2 and so cannot leave $\binom{\psi_1}{\psi_2}$ unchanged. Thus, $D^0 \oplus D^0$ is not consistent with the structure we are attempting to model. Must we conclude then that Pauli's proposal is doomed to failure?

To extricate ourselves from this dilemma we must understand that there is an essential feature of quantum mechanics that requires an adjustment in the classical picture we painted earlier (pages 86–87). Our conclusion that the transformation matrices $T(R)$ satisfy $T(R_1R_2) = T(R_1)T(R_2)$ and $(T(R))^{-1} = T(R^{-1})$ and therefore give rise to a representation of $SO(3)$ followed from the assumption that the wavefunction is uniquely determined in each coordinate system. This, however, is not (quite) the case. For example, both $\pm\binom{\psi_1}{\psi_2}$ represent the same state of our electron since an overall sign change has no effect on the probabilities described earlier (page 86) and all of

the physical content of the wavefunction is contained in such probabilities. It follows, in particular, that for a given $R \in SO(3)$, the transformation matrix $T(R)$ is determined only up to sign. Physicists would call T a "2-valued representation" of $SO(3)$. This makes no sense, of course, but we have just seen exactly how one can make sense of it. The appropriate tactic is to "go to the covering space," i.e., to represent a rotation, not by an element of $SO(3)$, but rather by two elements of $SU(2)$ and seek the transformation law for $\left(\begin{smallmatrix}\psi_1\\ \psi_2\end{smallmatrix}\right)$ among the representations of $SU(2)$. No longer being constrained to select a representation of $SU(2)$ that descends to $SO(3)$, we have available one more option, i.e., $D^{\frac{1}{2}}$. With this choice each $g = \left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right)$ in $SU(2)$ would transform the wavefunction $\left(\begin{smallmatrix}\psi_1\\ \psi_2\end{smallmatrix}\right)$ as follows:

$$\begin{pmatrix} \psi_1' \\ \psi_2' \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}.$$

Notice, in particular, that if $g = \left(\begin{smallmatrix}0 & -\boldsymbol{i}\\ -\boldsymbol{i} & 0\end{smallmatrix}\right)$, then Spin $(\pm g)$ is the rotation about the x-axis through π (Appendix A, [**N4**]) and therefore reverses the z-axis. Since

$$\begin{pmatrix} 0 & -\boldsymbol{i} \\ -\boldsymbol{i} & 0 \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = -\boldsymbol{i} \begin{pmatrix} \psi_2 \\ \psi_1 \end{pmatrix}$$

and $-\boldsymbol{i}\left(\begin{smallmatrix}\psi_2\\ \psi_1\end{smallmatrix}\right)$ represents the same state as $\left(\begin{smallmatrix}\psi_2\\ \psi_1\end{smallmatrix}\right)$, the representation $D^{\frac{1}{2}}$, unlike $D^0 \oplus D^0$, is at least consistent with our proposed model of spin one-half.

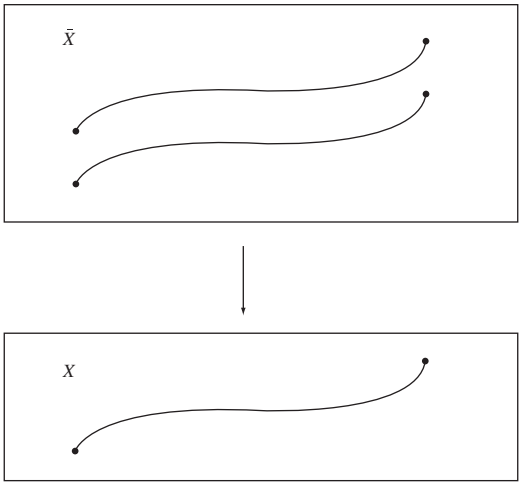
The program we have just described can leave one with the queasy feeling of ambiguity. Suppose that one is given a frame in $\mathbb{R}^3$ and wants to rotate by $R \in SO(3)$ to a new frame and thereby a new representation of the wavefunction. If Spin $(\pm g) = R$, then $D^{\frac{1}{2}}(g)\left(\begin{smallmatrix}\psi_1\\ \psi_2\end{smallmatrix}\right)$ and $D^{\frac{1}{2}}(-g)\left(\begin{smallmatrix}\psi_1\\ \psi_2\end{smallmatrix}\right) = -D^{\frac{1}{2}}(g)\left(\begin{smallmatrix}\psi_1\\ \psi_2\end{smallmatrix}\right)$ are physically equivalent and that's fine, but it is difficult not to ask oneself, "But, really, which is it?" The answer, interestingly enough, resides in the topologies of $SU(2)$, $SO(3)$ and the spinor map. One is forced to regard a rotation of frames in $\mathbb{R}^3$ not as an instantaneous jump from one to the other, but as a physical process that begins with one frame and continuously rotates the axes to the new position. The transformation law for the wavefunction depends not only on the end result of the rotation, but also on "how you got there." For example, the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}$$

represents a rotation through t radians about the x-axis. As t varies from 0 to 2π one has a continuous sequence of rotations (i.e., a curve $R_1(t)$ in

$SO(3)$) that represents the physical process of rotating a frame through one complete turn (360°) about its x-axis (identify each element of $SO(3)$ with the configuration of the axes that would result from applying that rotation to the initial configuration). The curve $R_2(t)$ in $SO(3)$ defined by the same formula, but with $0 \leq t \leq 4\pi$ represents a rotation about the x-axis through 720°. Both R_1 and R_2 begin and end with the same configuration, but there is a real difference, both physically and mathematically.

Spin : $SU(2) \longrightarrow SO(3)$ is a covering space (Exercise A.13, [**N4**]) and covering spaces have the property that curves in the covered space lift uniquely to curves in the covering space once an initial point is selected (Corollary 1.5.13, [**N4**]).



In particular, given a curve in $SO(3)$ (representing a continuous rotation of one frame into another) and a choice of either g or $-g$ above the initial point (frame), there is a uniquely determined element of $SU(2)$ above the terminal point that represents the transformation law that gives the wavefunction in the new frame. No ambiguity at all!

For the curves R_1 and R_2 in $SO(3)$, both of which begin and end at the identity, one can write out the lifts g_1 and g_2 starting at $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ explicitly. Both are given by

$$\begin{pmatrix} \cos\frac{t}{2} & -\boldsymbol{i}\sin\frac{t}{2} \\ -\boldsymbol{i}\sin\frac{t}{2} & \cos\frac{t}{2} \end{pmatrix}$$

but with $0 \leq t \leq 2\pi$ for g_1 and $0 \leq t \leq 4\pi$ for g_2 (see Appendix A, [**N4**]). But notice that g_1 is a path in $SU(2)$ from $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ to $-\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, whereas g_2 begins and

ends at $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$. Thus, a rotation of the frame through 360° changes the sign of the wavefunction, but a rotation through 720° leaves the sign unchanged, even though both rotations begin and end with the same configuration of the axes. Mathematically, the difference between R_1 and R_2 is that they represent two different homotopy classes in $\pi_1(SO(3)) \cong \mathbb{Z}_2$. R_2 is nullhomotopic since it is Spin $\circ g_2$ and g_2 is a loop at $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$ in $SU(2) \cong S^3$, but R_1 is not because it lifts to a path g_1 from $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$ to $-\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$, in $SU(2)$.

The next step in this program would be to look at the differential equations proposed by Pauli for $\left(\begin{smallmatrix}\psi_1\\\psi_2\end{smallmatrix}\right)$ and decide whether or not they assume the same form when the coordinate system is rotated and the wavefunction is transformed by $D^{\frac{1}{2}}$ (they do!). Since Pauli's theory was eventually abandoned (because it is not relativistically invariant) we shall not pursue this here, but will instead turn to the profoundly successful alternative proposed by Dirac.

Remark: Before moving forward with our discussion of spin one-half we observe that the arguments we presented (on pages 85–86) for modeling such a particle with a wavefunction having two components really had nothing specific to do with spin one-half. The essential feature that led to the doubling of the number of components was the existence of an "internal structure" that could be represented by a parameter (σ_z for the electron) that could assume precisely two values. There are other examples of this sort of thing, e.g., the "isotopic spin" parameter of a nucleon which determines (in the absence of electromagnetic fields) whether the particle is a proton or a neutron. The two components of the nucleon wavefunction then represent the "proton part" and the "neutron part" of the doublet. Indeed, it was this example that provided the initial motivation for Yang-Mills theory ([**YM**]).

Dirac set out to construct a relativistically invariant equation that would be satisfied by the wavefunction of a spin one-half particle. He reasoned that his equation should, in some sense, "imply" the Klein-Gordon equation since, as we have noted, this is just the quantized version of the relativistic energy-momentum relation $E^2 = \vec{p}^2 + m^2$. However, he also sought to remedy certain physical problems associated with the appearance of the second time derivative in the Klein-Gordon equation (for a discussion of these see Chapter 6 of [**Hol**]). Thus, Dirac sought a first order linear equation which, upon iteration, yielded the Klein-Gordon equation. Somewhat more precisely, Dirac was in the market for a first order differential operator

$$\not{\partial} = \gamma^0\partial_0 + \gamma^1\partial_1 + \gamma^2\partial_2 + \gamma^3\partial_3 = \gamma^\alpha\partial_\alpha \tag{2.4.1}$$

($\partial_\alpha = \frac{\partial}{\partial x^\alpha},\ \alpha = 0, 1, 2, 3$) such that, when applied to the equation

$$\not{\partial}\phi = -\boldsymbol{i}m\phi, \tag{2.4.2}$$

the result is the free Klein-Gordon equation

$$\eta^{\alpha\beta}\partial_\alpha\partial_\beta\phi = -m^2\phi. \tag{2.4.3}$$

The problem is to determine the γ^α, $\alpha = 0, 1, 2, 3$, so that this will be the case. Apply the operator $\not{D}$ in (2.4.1) to both sides of (2.4.2).

$$\begin{aligned} \not{D}\,(\not{D}\,\phi) &= \not{D}\,(-\boldsymbol{i}m\phi) \\ (\gamma^\alpha\partial_\alpha)\,(\gamma^\beta\partial_\beta\phi) &= -\boldsymbol{i}m\,\not{D}\phi \\ \gamma^\alpha\gamma^\beta\partial_\alpha\,(\partial_\beta\phi) &= -\boldsymbol{i}m\,(-\boldsymbol{i}m\phi) \\ (\gamma^\alpha\gamma^\beta\partial_\alpha\partial_\beta)\,\phi &= -m^2\phi \end{aligned} \tag{2.4.4}$$

This will agree with (2.4.3) if

$$\gamma^\alpha\gamma^\beta\partial_\alpha\partial_\beta = \eta^{\alpha\beta}\partial_\alpha\partial_\beta,$$

i.e., if

$$\gamma^\alpha\gamma^\beta = \eta^{\alpha\beta}, \quad \alpha, \beta = 0, 1, 2, 3. \tag{2.4.5}$$

Now, (2.4.5) clearly cannot be satisfied if the γ^α are taken to be numbers (none can be 0 since $\eta^{\alpha\alpha} = \pm 1$, but $\gamma^\alpha\gamma^\beta = 0$ if $\alpha \neq \beta$). Dirac's idea was to allow ϕ to have more than one complex component and interpret (2.4.5) as matrix equations (one for each $\alpha, \beta = 0, 1, 2, 3$). Specifically, if

$$\phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}$$

and $\partial_\beta\phi$ is computed entrywise, then $\not{D}\phi = \gamma^\alpha\partial_\alpha\phi$ would require the γ^α to have n columns (the reason we do not immediately follow Pauli's lead and take $n = 2$ will become clear shortly). To iterate the operator and define $\not{D}(\not{D}\phi)$ requires that the γ^α have n rows as well. Now, (2.4.5) must be interpreted as matrix equations

$$\gamma^\alpha\gamma^\beta = \eta^{\alpha\beta}\,\mathrm{id}, \quad \alpha, \beta = 0, 1, 2, 3,$$

where id is the $n \times n$ identity matrix. Notice, however, that each $\eta^{\alpha\beta}$ id is a symmetric matrix. To accommodate this fact we observe that, because $\partial_\alpha\partial_\beta\phi = \partial_\beta\partial_\alpha\phi$, (2.4.4) can be written

$$\frac{1}{2}\left(\gamma^\alpha\gamma^\beta + \gamma^\beta\gamma^\alpha\right)\partial_\alpha\partial_\beta\phi = -m^2\phi$$

so that we might just as well have written (2.4.5) as

$$\frac{1}{2}\left(\gamma^\alpha\gamma^\beta + \gamma^\beta\gamma^\alpha\right) = \eta^{\alpha\beta}$$

and so the matrix conditions we are currently trying to satisfy may be taken to be

$$\gamma^\alpha\gamma^\beta + \gamma^\beta\gamma^\alpha = 2\eta^{\alpha\beta}\mathrm{id}, \quad \alpha, \beta = 0, 1, 2, 3. \tag{2.4.6}$$

Finding square matrices that satisfy (2.4.6) is actually a problem familiar to algebraists. What we are looking for here is a matrix representation of the Clifford algebra of $(\mathbb{R}^4, \langle\ ,\ \rangle)$, where $\langle\ ,\ \rangle$ is the Minkowski inner product. There are many solutions, but the smallest n for which such matrices can be found is $n = 4$. Thus, Pauli's choice of $n = 2$ cannot succeed in this context (we will eventually have to sort out how to reconcile this with the arguments we presented earlier to the effect that a spin one-half wavefunction should have two components). It is, in fact, easy to write down a set of 4×4 matrices satisfying conditions (2.4.6). Letting

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{pmatrix}, \text{ and } \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

be the Pauli spin matrices and taking σ_0 to be the 2×2 identity matrix, one easily verifies the usual commutation relations

$$\begin{aligned} \sigma_i^2 &= \sigma_0, & & i = 1, 2, 3 \\ \sigma_i \sigma_j &= -\sigma_j \sigma_i, & & i, j = 1, 2, 3, \quad i \neq j. \end{aligned} \tag{2.4.7}$$

Now, we define

$$\gamma^0 = \begin{pmatrix} 0 & \sigma_0 \\ \sigma_0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\gamma^1 = \begin{pmatrix} 0 & -\sigma_1 \\ \sigma_1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\gamma^2 = \begin{pmatrix} 0 & -\sigma_2 \\ \sigma_2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \boldsymbol{i} \\ 0 & 0 & -\boldsymbol{i} & 0 \\ 0 & -\boldsymbol{i} & 0 & 0 \\ \boldsymbol{i} & 0 & 0 & 0 \end{pmatrix}$$

$$\gamma^3 = \begin{pmatrix} 0 & -\sigma_3 \\ \sigma_3 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

It is now a simple matter to verify that the conditions in (2.4.6) are satisfied by these matrices, e.g.,

$$\gamma^1\gamma^2+\gamma^2\gamma^1$$
$$=\begin{pmatrix}0&-\sigma_1\\ \sigma_1&0\end{pmatrix}\begin{pmatrix}0&-\sigma_2\\ \sigma_2&0\end{pmatrix}+\begin{pmatrix}0&-\sigma_2\\ \sigma_2&0\end{pmatrix}\begin{pmatrix}0&-\sigma_1\\ \sigma_1&0\end{pmatrix}$$
$$=\begin{pmatrix}-\sigma_1\sigma_2&0\\ 0&-\sigma_1\sigma_2\end{pmatrix}+\begin{pmatrix}-\sigma_2\sigma_1&0\\ 0&-\sigma_2\sigma_1\end{pmatrix}$$
$$=\begin{pmatrix}-(\sigma_1\sigma_2+\sigma_2\sigma_1)&0\\ 0&-(\sigma_1\sigma_2+\sigma_2\sigma_1)\end{pmatrix}$$
$$=\begin{pmatrix}0&0\\ 0&0\end{pmatrix}[4pt]=2\eta^{12}\begin{pmatrix}1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\end{pmatrix}$$

and

$$\gamma^1\gamma^1+\gamma^1\gamma^1=2\gamma^1\gamma^1=2\begin{pmatrix}0&-\sigma_1\\ \sigma_1&0\end{pmatrix}\begin{pmatrix}0&-\sigma_1\\ \sigma_1&0\end{pmatrix}$$
$$=2\begin{pmatrix}-\sigma_1^2&0\\ 0&-\sigma_1^2\end{pmatrix}=-2\begin{pmatrix}\sigma_0&0\\ 0&\sigma_0\end{pmatrix}$$
$$=2\eta^{11}\begin{pmatrix}1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\end{pmatrix}$$

etc.

Remark: There are many other possible choices for γ^0, γ^1, γ^2 and γ^3, many of which are used in the physics literature. Indeed, for any nonsingular matrix B, one can replace each γ^α by $B\gamma^\alpha B^{-1}$ and obtain another set of "Dirac matrices" satisfying (2.4.6). Conversely, one can show (see, e.g., pages 104–106 of [**Gre**]) that any set of 4×4 matrices satisfying (2.4.6) differs from our choice by such a similarity transformation. Algebraically, this means that, up to equivalence, there is only one representation of the Clifford algebra of $\mathbb{R}^{1,3}$ by 4×4 matrices. The choice we have made is called the **Weyl**, or **chiral representation**.

With γ^0, γ^1, γ^2 and γ^3 the 4×4 matrices described above, the wavefunction for our spin one-half particle has four components

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}$$

and the **Dirac equation** (expressed in standard coordinates on $\mathbb{R}^{1,3}$) is

$$\not{D}\phi = -\boldsymbol{i}m\phi$$
$$\gamma^\alpha \partial_\alpha \phi = -\boldsymbol{i}m\phi$$

$$\begin{pmatrix} 0 & 0 & \partial_0 - \partial_3 & -\partial_1 + \boldsymbol{i}\partial_2 \\ 0 & 0 & -\partial_1 - \boldsymbol{i}\partial_2 & \partial_0 + \partial_3 \\ \partial_0 + \partial_3 & \partial_1 - \boldsymbol{i}\partial_2 & 0 & 0 \\ \partial_1 + \boldsymbol{i}\partial_2 & \partial_0 - \partial_3 & 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}$$

$$= -\boldsymbol{i}m \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}.$$

We will discuss the relativistic invariance of the Dirac equation shortly, but first we would like to show that, in the massless ($m = 0$) case, there is a simpler solution to the problem of finding a first order operator $D = \gamma^\alpha \partial_\alpha$ which, when applied to the equation

$$D\phi = 0$$

yields the $m = 0$ Klein-Gordon equation

$$\eta^{\alpha\beta}\partial_\alpha \partial_\beta \phi = 0.$$

For this purpose we write this last equation as

$$\partial_0 \partial_0 \phi = \delta^{ij}\partial_i \partial_j \phi. \tag{2.4.8}$$

Now, $D\phi = 0$ can be written

$$\gamma^0 \partial_0 \phi = -\gamma^i \partial_i \phi,$$

or, assuming γ^0 is invertible,

$$\partial_0 \phi = -\mu^i \partial_i \phi \quad (\mu^i = (\gamma^0)^{-1}\gamma^i,\ i = 1, 2, 3).$$

Then

$$\begin{aligned} \partial_0 \partial_0 \phi &= \partial_0(-\mu^i \partial_i \phi) = -\mu^i \partial_0 \partial_i \phi \\ &= -\mu^i \partial_i \partial_0 \phi = -\mu^i \partial_i(-\mu^j \partial_j \phi) \end{aligned}$$

$$\partial_0 \partial_0 \phi = \mu^i \mu^j \partial_i \partial_j \phi. \tag{2.4.9}$$

Now compare (2.4.8) and (2.4.9). Again, $\mu^i\mu^j = \delta^{ij}$ cannot be satisfied by numbers so we rewrite (2.4.9) as

$$\partial_0\partial_0\phi = \frac{1}{2}(\mu^i\mu^j + \mu^j\mu^i)\partial_i\partial_j\phi, \tag{2.4.10}$$

and seek matrices satisfying

$$\mu^i\mu^j + \mu^j\mu^i = 2\delta^{ij}\ \mathrm{id}, \quad i,j = 1,2,3. \tag{2.4.11}$$

In the terminology of algebra, matrices satisfying the conditions (2.4.11) constitute a matrix representation of the Clifford algebra of $(\mathbb{R}^3,\ \langle\ ,\ \rangle)$, where $\langle\ ,\ \rangle$ is the usual positive definite inner product on $\mathbb{R}^3$. Notice that the conditions (2.4.11) coincide with the commutation relations (2.4.7) for the Pauli spin matrices $\sigma_1,\ \sigma_2$ and σ_3. In particular, there is now a 2×2 solution to the problem so, in the massless case, our wavefunction can have two components

$$\phi = \begin{pmatrix}\phi_3\\ \phi_4\end{pmatrix}$$

(the reason for the peculiar numbering will become clear soon). One obtains an operator $D = \gamma^\alpha\partial_\alpha$ of the required type by taking, for example,
$\gamma^0 = \sigma_0 = \left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$ and $\gamma^i = \mu^i = -\sigma_i,\ i = 1,\ 2,\ 3$. Thus,

$$\begin{aligned} D &= \begin{pmatrix}1&0\\0&1\end{pmatrix}\partial_0 + \begin{pmatrix}0&-1\\-1&0\end{pmatrix}\partial_1 + \begin{pmatrix}0&\boldsymbol{i}\\-\boldsymbol{i}&0\end{pmatrix}\partial_2 + \begin{pmatrix}-1&0\\0&1\end{pmatrix}\partial_3\\ &= \begin{pmatrix}\partial_0 - \partial_3 & -\partial_1 + \boldsymbol{i}\partial_2\\ -\partial_1 - \boldsymbol{i}\partial_2 & \partial_0 + \partial_3\end{pmatrix}. \end{aligned}$$

The corresponding equation $D\phi = 0$ then becomes

$$\begin{pmatrix}\partial_0 - \partial_3 & -\partial_1 + \boldsymbol{i}\partial_2\\ -\partial_1 - \boldsymbol{i}\partial_2 & \partial_0 + \partial_3\end{pmatrix}\begin{pmatrix}\phi_3\\ \phi_4\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \tag{2.4.12}$$

and is known as the **Weyl neutrino equation**. Notice that this is just what one would obtain from the Dirac equation with $m = 0$ and a wave- function of the form

$$\begin{pmatrix}0\\0\\\phi_3\\\phi_4\end{pmatrix}.$$

For future reference we note also that the $m = 0$ Dirac equation for a wave-function of the form

$$\begin{pmatrix}\phi_1\\\phi_2\\0\\0\end{pmatrix}$$

reduces to

$$\begin{pmatrix} \partial_0 + \partial_3 & \partial_1 - \boldsymbol{i}\partial_2 \\ \partial_1 + \boldsymbol{i}\partial_2 & \partial_0 - \partial_3 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{2.4.13}$$

and that the coefficient matrix in (2.4.13) is the formal conjugate, transposed inverse of the coefficient matrix in (2.4.12). The significance of these observations will emerge in our discussion of the transformation properties of the wavefunctions and the corresponding invariance properties of the equations.

Now we return to the issue of the Lorentz invariance of the Dirac equation. The problem, as it was for the Klein-Gordon equation (page 82), is to show that the Dirac equation has the same form in any other coordinate system $y^0,\ y^1,\ y^2,\ y^3$ for $\mathbb{R}^{1,3}$, related to the standard coordinates $x^0,\ x^1,\ x^2,\ x^3$ by

$$y^\alpha = \Lambda^\alpha{}_\beta x^\beta, \qquad \alpha = 0, 1, 2, 3,$$

where $\Lambda = (\Lambda^\alpha{}_\beta) \in \mathcal{L}_+^\uparrow$. However, since the Dirac wavefunction has four complex components, this will require finding a representation $T : \mathcal{L}_+^\uparrow \longrightarrow GL(\mathbb{C}^4)$ of $\mathcal{L}_+^\uparrow$ on $\mathbb{C}^4$ which, if taken to be the transformation law for the wavefunction, preserves the form of the Dirac equation (cf., the discussion of the two-component Pauli theory on pages 86–93). As was the case for $SO(3)$ in the Pauli theory, it so happens that all of the representations $\mathcal{L}_+^\uparrow$ are known, that they are most conveniently described in terms of a two- fold covering group of $\mathcal{L}_+^\uparrow$ and that, because of the nature of a quantum mechanical wavefunction, it is actually the representations of the covering group that do *not* descend to $\mathcal{L}_+^\uparrow$ that turn out to be of most interest. We begin with a brief summary of the relevant results (see Chapter 3 for more details).

Identifying $\mathcal{L}_+^\uparrow$ with a subset of $\mathbb{R}^{16}$ one finds that it is a submanifold diffeomorphic to $SO(3) \times \mathbb{R}^3$ and so is a 6-dimensional Lie group containing $SO(3)$ as a closed subgroup. We denote by $SL(2, \mathbb{C})$ the group of 2×2 complex matrices with determinant one. Identifying $SL(2, \mathbb{C})$ with a subset of $\mathbb{C}^4 = \mathbb{R}^8$ one finds that it is a submanifold diffeomorphic to $S^3 \times \mathbb{R}^3$ and so it is also a 6-dimensional (simply connected) Lie group. Note that $SU(2)$ is a closed subgroup of $SL(2, \mathbb{C})$. Now, we have already described a two-fold covering map

$$\text{Spin}\colon SU(2) \longrightarrow SO(3)$$

and we wish now to show that this is, in fact, the restriction to $SU(2)$ of a two-fold covering map of $SL(2, \mathbb{C})$ onto $\mathcal{L}_+^\uparrow$, also denoted

$$\text{Spin}\colon SL(2, \mathbb{C}) \longrightarrow \mathcal{L}_+^\uparrow.$$

The construction of this map is carried out in detail in Section 1.7 of **[N3]** so we will be brief. $\mathbb{R}^{1,3}$ can be identified with the linear space $\mathcal{H}$ of 2×2

complex Hermitian matrices

$$x = \begin{pmatrix} x^0 + x^3 & x^1 - \boldsymbol{i}x^2 \\ x^1 + \boldsymbol{i}x^2 & x^0 - x^3 \end{pmatrix}$$
$$= x^0\sigma_0 + x^1\sigma_1 + x^2\sigma_2 + x^3\sigma_3 = x^\alpha\sigma_\alpha,$$

where the squared norm is taken to be the determinant. Observe that each of the coordinates x^α can be expressed as

$$x^\alpha = \frac{1}{2}\text{trace}(\sigma_\alpha x)$$

so that

$$x = \sum_{\alpha=0}^{3} \frac{1}{2}\text{trace}(\sigma_\alpha x)\sigma_\alpha.$$

Now, for each $g \in SL(2,\mathbb{C})$ we define a map $\Lambda_g : \mathcal{H} \longrightarrow \mathcal{H}$ by

$$\Lambda_g(x) = gx\bar{g}^\top.$$

Note that $\Lambda_g(x)$ is, indeed, in $\mathcal{H}$ because $\overline{\Lambda_g(x)}^\top = (\overline{gx\bar{g}^\top})^\top = (\bar{g}\bar{x}g^\top)^\top = gx\bar{g}^\top = \Lambda_g(x)$. Also note that, for $g \in SU(2) \subseteq SL(2,\mathbb{C})$, $\bar{g}^\top = g^{-1}$. Now, Λ_g is surely linear and satisfies $\det(\Lambda_g(x)) = \det(gx\bar{g}^\top) = \det(x)$ so it preserves the Minkowski inner product on $\mathcal{H} = \mathbb{R}^{1,3}$. Thus, Λ_g is an orthogonal transformation and, with a bit more work (page 77, [**N3**]), one can show that it is proper and orthochronous. Now let

$$\Lambda_g(x) = y^\alpha\sigma_\alpha.$$

Then

$$y^\alpha = \frac{1}{2}\text{trace}(\sigma_\alpha\Lambda_g(x)) = \frac{1}{2}\text{trace}(\sigma_\alpha gx\bar{g}^\top)$$
$$= \frac{1}{2}\text{trace}(\sigma_\alpha g(x^\beta\sigma_\beta)\bar{g}^\top)$$
$$= \frac{1}{2}\text{trace}(\sigma_\alpha g\sigma_\beta\bar{g}^\top)x^\beta$$
$$= \Lambda^\alpha{}_\beta x^\beta$$

where

$$\Lambda^\alpha{}_\beta = \frac{1}{2}\text{trace}(\sigma_\alpha g\sigma_\beta\bar{g}^\top), \qquad \alpha,\beta = 0,1,2,3.$$

Thus, $(\Lambda^\alpha{}_\beta)_{\alpha,\beta=0,1,2,3}$ is in $\mathcal{L}_+^\uparrow$ and we define Spin: $SL(2,\mathbb{C}) \longrightarrow \mathcal{L}_+^\uparrow$ by

$$\text{Spin}(g) = (\Lambda^\alpha{}_\beta)_{\alpha,\,\beta\,=0,1,2.3}$$

for each $g \in SL(2,\mathbb{C})$. One then shows that Spin is a two-fold covering group for $\mathcal{L}_+^\uparrow$ and that its restriction to $SU(2)$ agrees with the map of the same

name discussed on page 88. Before proceeding we will record one additional fact that we will need shortly. Notice that

$$\Lambda^{\alpha}{}_{\beta} = \frac{1}{2}\text{trace}(\sigma_\alpha g \sigma_\beta \bar{g}^\top) = \frac{1}{2}\text{trace}(\sigma_\beta(\bar{g}^\top \sigma_\alpha g))$$

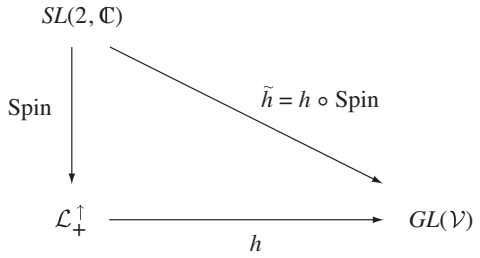
so that

$$\bar{g}^\top \sigma_\alpha g = \sum_{\beta=0}^{3} \Lambda^{\alpha}{}_{\beta}\sigma_\beta. \tag{2.4.14}$$

Now we proceed just as we did for $SO(3)$ in our discussion of the Pauli theory. Consider a representation

$$h : \mathcal{L}_+^\uparrow \longrightarrow GL(\mathcal{V})$$

of $\mathcal{L}_+^\uparrow$. Composing with Spin then gives a representation of $SL(2,\mathbb{C})$.



Thus, every representation of $\mathcal{L}_+^\uparrow$ comes from a representation of $SL(2,\mathbb{C})$, but conversely, a representation of $SL(2,\mathbb{C})$ will descend to a representation of $\mathcal{L}_+^\uparrow$ if and only if it takes the same value at $\pm g$ for each $g \in SL(2,\mathbb{C})$ (all others are of the "2-valued" variety).

The representations of $SL(2,\mathbb{C})$ are all known and are described in some detail in Section 3.1 of **[N3]**. We will limit ourselves to a brief discussion of just those items that are relevant to our present context. There are a few obvious representations of $SL(2,\mathbb{C})$ on $\mathbb{C}^n$. By sending every element of $SL(2,\mathbb{C})$ to the $n \times n$ identity matrix one obtains the trivial representation which leaves everything in $\mathbb{C}^n$ fixed. When $n = 2$ one also has the identity representation which sends each $g = \left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ in $SL(2,\mathbb{C})$ to the linear transformation on $\mathbb{C}^2$ defined by

$$\begin{pmatrix} z^1 \\ z^2 \end{pmatrix} \longrightarrow \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} z^1 \\ z^2 \end{pmatrix} = \begin{pmatrix} \alpha z^1 + \beta z^2 \\ \gamma z^1 + \delta z^2 \end{pmatrix}.$$

This representation is generally denoted

$$D^{(\frac{1}{2},0)} : SL(2,\mathbb{C}) \longrightarrow GL(\mathbb{C}^2)$$

and called the **left-handed spinor representation** of $SL(2,\mathbb{C})$. Identifying the linear transformation $D^{(\frac{1}{2},0)}(g)$ on $\mathbb{C}^2$ with its matrix relative to the

standard basis for $\mathbb{C}^2$ one can write

$$D^{(\frac{1}{2},0)}(g) = g$$

for each $g \in SL(2,\mathbb{C})$. Another, somewhat less obvious, representation

$$D^{(0,\frac{1}{2})} : SL(2,\mathbb{C}) \longrightarrow GL(\mathbb{C}^2)$$

of $SL(2,\mathbb{C})$ on $\mathbb{C}^2$ sends each g to the linear transformation whose matrix relative to the standard basis is the inverse of the conjugate transpose $\bar{g}^\top$ of g (called the **right-handed spinor representation** of SL $(2,$ $\mathbb{C})$). For simplicity we again write

$$D^{(0,\frac{1}{2})}(g) = (\bar{g}^\top)^{-1}.$$

Remarks: The map that sends g to (the linear transformation on $\mathbb{C}^2$ whose matrix is) $(g^\top)^{-1}$ is also a representation of $SL(2,\mathbb{C})$, but it is equivalent to $D^{(\frac{1}{2},0)}$. The reason is that, if $B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, then, for any $g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ in $SL(2,\mathbb{C})$,

$$BgB^{-1} = (g^\top)^{-1}.$$

It follows that $D^{(0,\frac{1}{2})}$ is equivalent to the **conjugation representation** $g \longrightarrow \bar{g}$. $D^{(\frac{1}{2},0)}$ and $D^{(0,\frac{1}{2})}$ are, however, *not* equivalent as representations of $SL(2,\mathbb{C})$. To see this, suppose there were a matrix B such that $BgB^{-1} = (\bar{g}^\top)^{-1}$ for each $g \in SL(2,\mathbb{C})$. Then, in particular, we would have $\text{trace}(g) = \text{trace}(\bar{g}^\top)^{-1}$ for each $g \in SL(2,\mathbb{C})$. This, however, is not true for $g = \begin{pmatrix} -2\boldsymbol{i} & 0 \\ 0 & \frac{1}{2}\boldsymbol{i} \end{pmatrix} \in SL(2,\mathbb{C})$ since $(\bar{g}^\top)^{-1} = \begin{pmatrix} -\frac{1}{2}\boldsymbol{i} & 0 \\ 0 & 2\boldsymbol{i} \end{pmatrix}$. It is interesting to note, however, that $g \longrightarrow g$ and $g \longrightarrow (\bar{g}^\top)^{-1}$ *are* equivalent as representations of $SU(2)$ since there $\bar{g}^\top = g^{-1}$ so $(\bar{g}^\top)^{-1} = g$.

There is a precise sense in which all of the representations of $SL(2,\mathbb{C})$ can be constructed from $D^{(\frac{1}{2},0)}$ and $D^{(0,\frac{1}{2})}$. Rather than describing this procedure in general we will be content to illustrate it in the only case of real interest to us, i.e., the representations of $SL(2,\mathbb{C})$ on $\mathbb{C}^4$. The most obvious way to construct a representation on $\mathbb{C}^4$ from two representations on $\mathbb{C}^2$ is by forming their direct sum. For example, we can define the representation

$$D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})} : SL(2,\mathbb{C}) \longrightarrow GL(\mathbb{C}^4)$$

by

$$D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}(g) = \begin{pmatrix} g & 0 \\ 0 & (\bar{g}^\top)^{-1} \end{pmatrix},$$

where all of the entries are 2×2 matrices and we are again identifying a linear transformation on $\mathbb{C}^4$ with its matrix relative to the standard basis. Similarly, one can define the direct sum of any such pair. A somewhat less obvious procedure for building a representation on $\mathbb{C}^4$ is the tensor product of two representations on $\mathbb{C}^2$. For instance, the representation

$$D^{(\frac{1}{2},\frac{1}{2})} = D^{(\frac{1}{2},0)} \otimes D^{(0,\frac{1}{2})} : SL(2,\mathbb{C}) \longrightarrow GL(\mathbb{C}^4)$$

can be described as follows: For each $g = \left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right) \in SL(2,\mathbb{C})$, we define

$$D^{(\frac{1}{2},\frac{1}{2})}(g) = \begin{pmatrix} \alpha(\bar{g}^\top)^{-1} & \beta(\bar{g}^\top)^{-1} \\ \gamma(\bar{g}^\top)^{-1} & \delta(\bar{g}^\top)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha\bar{\delta} & -\alpha\bar{\gamma} & \beta\bar{\delta} & -\beta\bar{\gamma} \\ -\alpha\bar{\beta} & \alpha\bar{\alpha} & -\beta\bar{\beta} & \beta\bar{\alpha} \\ \gamma\bar{\delta} & -\gamma\bar{\gamma} & \delta\bar{\delta} & -\delta\bar{\gamma} \\ -\gamma\bar{\beta} & \gamma\bar{\alpha} & -\delta\bar{\beta} & \delta\bar{\alpha} \end{pmatrix}.$$

Note that $D^{(\frac{1}{2},\frac{1}{2})}(-g) = D^{(\frac{1}{2},\frac{1}{2})}(g)$ so $D^{(\frac{1}{2},\frac{1}{2})}$ descends to a representation of $\mathcal{L}_+^\uparrow$ on $\mathbb{C}^4$ (also denoted $D^{(\frac{1}{2},\frac{1}{2})}$). One can show that $D^{(\frac{1}{2},\frac{1}{2})}$ is equivalent to the natural (vector) representation of $\mathcal{L}_+^\uparrow$ on $\mathbb{R}^{1,3}$. Similarly, one can define such representations as $D^{(1,0)} = D^{(\frac{1}{2},0)} \otimes D^{(\frac{1}{2},0)}$. One can show that these are irreducible, whereas such things as $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$, of course, are not. From our point of view the important fact is that we have just described *all* of the representations of $SL(2,\mathbb{C})$ on $\mathbb{C}^4$ (up to equivalence). Proving the relativistic invariance of the Dirac equation therefore amounts to searching among these few representations of $SL(2,\mathbb{C})$ on $\mathbb{C}^4$ for one which, if adopted as the transformation law for the 4-component Dirac wavefunction, will lead to a transformed wavefunction that satisfies the same (Dirac) equation in the transformed coordinate system.

Begin with the Dirac equation in standard coordinates $x = (x^0, x^1, x^2, x^3)$ on $\mathbb{R}^{1,3}$.

$$\Big(\gamma^\beta \partial_\beta + \boldsymbol{i}m\Big)\phi(x) = 0. \tag{2.4.15}$$

Introduce new coordinates $y = (y^0, y^1, y^2, y^3)$ on $\mathbb{R}^{1,3}$ by

$$y^\alpha = \Lambda^\alpha{}_\beta x^\beta, \quad \alpha = 0,1,2,3,$$

where $\Lambda = (\Lambda^\alpha{}_\beta) \in \mathcal{L}_+^\uparrow$ and define $\hat{\partial}_\alpha = \frac{\partial}{\partial y^\alpha}$, $\alpha = 0,1,2,3$. Then

$$\partial_\beta = \frac{\partial}{\partial x^\beta} = \frac{\partial y^\alpha}{\partial x^\beta}\frac{\partial}{\partial y^\alpha} = \Lambda^\alpha{}_\beta \hat{\partial}_\alpha.$$

Let $g \in SL(2,\mathbb{C})$ be such that Spin $(\pm g) = (\Lambda^{\alpha}{}_{\beta})$. Thus,

$$\Lambda^{\alpha}{}_{\beta} = \frac{1}{2}\text{trace}\left(\sigma_{\alpha} g \sigma_{\beta} \bar{g}^{\top}\right), \qquad \alpha, \beta = 0, 1, 2, 3.$$

Our objective is to find a representation

$$\rho : SL(2,\mathbb{C}) \longrightarrow GL(\mathbb{C}^4)$$

such that, if

$$\hat{\phi}(y) = \rho(g)(\phi(\Lambda^{-1} y)),$$

then (2.4.15) implies

$$\left(\gamma^{\alpha}\hat{\partial}_{\alpha} + \boldsymbol{i}m\right)\hat{\phi}(y) = 0. \tag{2.4.16}$$

We begin by simply rewriting (2.4.15) in the new coordinates

$$\gamma^{\beta}\partial_{\beta}\phi + \boldsymbol{i}m\phi = 0$$

$$\gamma^{\beta}\left(\Lambda^{\alpha}{}_{\beta}\,\hat{\partial}_{\alpha}\right)\left((\rho(g))^{-1}\hat{\phi}(y)\right) + \boldsymbol{i}m(\rho(g))^{-1}\hat{\phi}(y) = 0$$

$$\gamma^{\beta}(\rho(g))^{-1}\Lambda^{\alpha}{}_{\beta}\hat{\partial}_{\alpha}\hat{\phi}(y) + (\rho(g))^{-1}\left(\boldsymbol{i}m\hat{\phi}(y)\right) = 0.$$

Multiply through by $\rho(g)$ to obtain

$$\rho(g)\gamma^{\beta}(\rho(g))^{-1}\Lambda^{\alpha}{}_{\beta}\hat{\partial}_{\alpha}\hat{\phi}(y) + \boldsymbol{i}m\hat{\phi}(y) = 0$$

$$\left(\left(\rho(g)\gamma^{\beta}(\rho(g))^{-1}\Lambda^{\alpha}{}_{\beta}\right)\hat{\partial}_{\alpha} + \boldsymbol{i}m\right)\hat{\phi}(y) = 0$$

which will agree with (2.4.16) if and only if

$$\rho(g)\gamma^{\beta}(\rho(g))^{-1}\Lambda^{\alpha}{}_{\beta} = \gamma^{\alpha}. \tag{2.4.17}$$

This then is the condition that our representation ρ must satisfy in order to preserve the form of the Dirac equation. We obtain a more convenient form of this condition as follows:

$$(\rho(g))\gamma^{\beta}(\rho(g))^{-1}\Lambda^{\alpha}{}_{\beta} = \gamma^{\alpha}$$

$$(\rho(g))\gamma^{\beta}(\rho(g))^{-1} = \Lambda_{\alpha}{}^{\beta}\gamma^{\alpha}$$

$$\gamma^{\beta}(\rho(g))^{-1} = (\rho(g))^{-1}\Lambda_{\alpha}{}^{\beta}\gamma^{\alpha}$$

$$\gamma^\beta = (\rho(g))^{-1}\Lambda_\alpha{}^\beta\gamma^\alpha(\rho(g))$$

$$\gamma^\beta = \Lambda_\alpha{}^\beta\Big((\rho(g))^{-1}\gamma^\alpha(\rho(g))\Big)$$

$$\Lambda^\alpha{}_\beta\gamma^\beta = (\rho(g))^{-1}\gamma^\alpha\rho(g)$$

$$(\rho(g))^{-1}\gamma^\alpha\rho(g) = \Lambda^\alpha{}_\beta\gamma^\beta, \quad \alpha = 0,1,2,3. \tag{2.4.18}$$

At this point one need only check each of the representations ρ of $SL(2,\mathbb{C})$ on $\mathbb{C}^4$ described earlier in the hope of finding one that satisfies (2.4.18). One's hopes are not dashed. The winner is $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$, as we now show. With $\gamma^0, \gamma^1, \gamma^2$ and γ^3 as on page 96 and

$$\rho(g) = \begin{pmatrix} g & 0 \\ 0 & (\bar{g}^\top)^{-1} \end{pmatrix}$$

we compute

$$(\rho(g))^{-1}\gamma^0\rho(g) = \begin{pmatrix} 0 & g^{-1}\sigma_0(\bar{g}^\top)^{-1} \\ \bar{g}^\top\sigma_0 g & 0 \end{pmatrix} \tag{2.4.19}$$

and, for $i = 1,2,3$,

$$(\rho(g))^{-1}\gamma^i\rho(g) = \begin{pmatrix} 0 & -g^{-1}\sigma_i(\bar{g}^\top)^{-1} \\ \bar{g}^\top\sigma_i g & 0 \end{pmatrix}. \tag{2.4.20}$$

On the other hand,

$$\begin{aligned}\Lambda^\alpha{}_\beta\gamma^\beta &= \Lambda^\alpha{}_0\gamma^0 + \Lambda^\alpha{}_1\gamma^1 + \Lambda^\alpha{}_2\gamma^2 + \Lambda^\alpha{}_3\gamma^3 \\ &= \Lambda^\alpha{}_0\begin{pmatrix} 0 & \sigma_0 \\ \sigma_0 & 0 \end{pmatrix} + \Lambda^\alpha{}_1\begin{pmatrix} 0 & -\sigma_1 \\ \sigma_1 & 0 \end{pmatrix} \\ &\quad + \Lambda^\alpha{}_2\begin{pmatrix} 0 & -\sigma_2 \\ \sigma_2 & 0 \end{pmatrix} + \Lambda^\alpha{}_3\begin{pmatrix} 0 & -\sigma_3 \\ \sigma_3 & 0 \end{pmatrix}\end{aligned}$$

so

$$\Lambda^\alpha{}_\beta\gamma^\beta = \begin{pmatrix} 0 & \Lambda^\alpha{}_0\sigma_0 - \sum_{i=1}^3 \Lambda^\alpha{}_i\sigma_i \\ \sum_{\beta=0}^3 \Lambda^\alpha{}_\beta\sigma_\beta & 0 \end{pmatrix}. \tag{2.4.21}$$

Now, for any $\alpha = 0,1,2,3$, the 21-block in $(\rho(g))^{-1}\gamma^\alpha\rho(g)$ is $\bar{g}^\top\sigma_\alpha g$ and the 21-block in $\Lambda^\alpha{}_\beta\gamma^\beta$ is $\sum_{\beta=0}^3 \Lambda^\alpha{}_\beta\sigma_\beta$ and these are equal by (2.4.14). The equality of the 12-blocks in (2.4.18) follows by taking inverses on both sides of (2.4.14).

More precisely, one observes that

$$(\bar{g}^\top \sigma_\alpha g)^{-1} = g^{-1}\sigma_\alpha^{-1}(\bar{g}^\top)^{-1} = g^{-1}\sigma_\alpha(\bar{g}^\top)^{-1}$$

and, from a brief calculation that we will leave for the reader,

$$\left(\sum_{\beta=0}^{3}\Lambda^\alpha{}_\beta\sigma_\beta\right)^{-1} = \begin{cases} \Lambda^0{}_0\sigma_0 - \sum_{i=1}^{3}\Lambda^0{}_i\sigma_i, & \alpha = 0 \\ -\Lambda^\alpha{}_0\sigma_0 + \sum_{i=1}^{3}\Lambda^\alpha{}_i\sigma_i, & \alpha = 1,2,3 \end{cases}.$$

With this we have established the Lorentz invariance of the Dirac equation.

The emergence of the representation $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$ as the appropriate transformation law for a Dirac wavefunction has interesting and important consequences that we will briefly explore. Let us write the Dirac wavefunction ϕ as

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix} = \begin{pmatrix} \phi_L \\ \phi_R \end{pmatrix},$$

where $\phi_L = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$ and $\phi_R = \begin{pmatrix} \phi_3 \\ \phi_4 \end{pmatrix}$. Since ϕ transforms according to $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$, ϕ_L and ϕ_R transform according to $D^{(\frac{1}{2},0)}$ and $D^{(0,\frac{1}{2})}$, respectively. Furthermore, the Dirac equation becomes a pair of coupled equations for ϕ_L and ϕ_R:

$$\gamma^\alpha\partial_\alpha\phi = -\boldsymbol{i}m\phi$$

$$\begin{pmatrix} 0 & \sigma_0\partial_0 - \sum_{i=1}^{3}\sigma_i\partial_i \\ \sigma_0\partial_0 + \sum_{i=1}^{3}\sigma_i\partial_i & 0 \end{pmatrix}\begin{pmatrix} \phi_L \\ \phi_R \end{pmatrix} = -\boldsymbol{i}m\begin{pmatrix} \phi_L \\ \phi_R \end{pmatrix}$$

$$\begin{cases} \left(\sigma_0\partial_0 - \sum_{i=1}^{3}\sigma_i\partial_i\right)\phi_R = -\boldsymbol{i}m\phi_L \\ \left(\sigma_0\partial_0 + \sum_{i=1}^{3}\sigma_i\partial_i\right)\phi_L = -\boldsymbol{i}m\phi_R \end{cases}. \tag{2.4.22}$$

In particular, in the massless ($m = 0$) case one obtains two uncoupled equations

$$\left(\sigma_0\partial_0 - \sum_{i=1}^{3}\sigma_i\partial_i\right)\phi_R = 0 \tag{2.4.23}$$

$$\left(\sigma_0\partial_0 + \sum_{i=1}^{3}\sigma_i\partial_i\right)\phi_L = 0 \tag{2.4.24}$$

which are, of course, just equations (2.4.12) and (2.4.13). To understand the significance of ϕ_R and ϕ_L in general, the relationship between our current model of spin one-half particles as 4-component objects and our earlier (2-component) view of spin one-half (pages 85–86) and just what (2.4.23) and (2.4.24) have to do with neutrinos (see page 100) we must discuss yet another symmetry (invariance property) of the Dirac equation.

The Dirac equation is invariant under the proper, orthochronous Lorentz group $\mathcal{L}_+^\uparrow$ because we were able to find a ("2-valued") representation of $\mathcal{L}_+^\uparrow$ which, if taken to be the transformation law for the wavefunction, led to a transformed wavefunction that satisfied the Dirac equation in the transformed coordinate system. Now we wish to consider a coordinate transformation, called **spatial inversion**, that does not correspond to an element of $\mathcal{L}_+^\uparrow$. Its matrix is

$$\pi = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

and its effect is simply to switch the orientation of the spatial coordinate system. Note that, since π^2 is the 4 $\times$4 identity matrix, π generates a group of coordinate transformations which we may denote $\mathbb{Z}_2$, since that's what it is isomorphic to. In order to prove the invariance of the Dirac equation under spatial inversions we will find a representation $\rho\colon \mathbb{Z}_2 \longrightarrow GL(\mathbb{C}^4)$ which, if taken to be the transformation law for ϕ, leads to a transformed wavefunction that satisfies the Dirac equation in the transformed coordinate system. Note that, since $(\gamma^0)^2$ is the 4 $\times$4 identity matrix, the assignments

$$\begin{aligned} \mathrm{id} &\longrightarrow \mathrm{id} \\ \pi &\longrightarrow \gamma^0 \end{aligned}$$

define a representation ρ of $\mathbb{Z}_2$ on $\mathbb{C}^4$. We will show that this does the job.

Begin with the Dirac equation (2.4.15) in standard coordinates $x = (x^0, x^1, x^2, x^3)$ on $\mathbb{R}^{1,3}$. Introduce new coordinates $y = (y^0, y^1, y^2, y^3)$ on $\mathbb{R}^{1,3}$ by $y^\alpha = \Lambda^\alpha{}_\beta x^\beta,\ \alpha = 0, 1, 2, 3$, where $\Lambda = (\Lambda^\alpha{}_\beta)$ is in $\mathbb{Z}_2$. Define $\hat{\partial}_\alpha = \partial/\partial y^\alpha,\ \alpha = 0, 1, 2, 3$. Then $\partial_\beta = \Lambda^\alpha{}_\beta \hat{\partial}_\alpha$. Now define $\hat{\phi}(y)$ by

$$\hat{\phi}\,(y) = \rho(\Lambda)(\phi\,(\Lambda^{-1}y)).$$

Substituting into (2.4.15) as on pages 106–107 gives

$$\Big((\rho(\Lambda)\gamma^\beta(\rho(\Lambda))^{-1}\,\Lambda^\alpha{}_\beta)\hat{\partial}_\alpha + \boldsymbol{i}m\Big)\hat{\phi}(y) = 0$$

which will be the Dirac equation if

$$(\rho(\Lambda))\gamma^\beta(\rho(\Lambda))^{-1}\Lambda^\alpha{}_\beta = \gamma^\alpha, \quad \alpha = 0,1,2,3,$$

i.e., if

$$(\rho(\Lambda))^{-1}\gamma^\alpha\rho(\Lambda) = \Lambda^\alpha{}_\beta\gamma^\beta, \quad \alpha = 0,1,2,3 \tag{2.4.25}$$

(see page 107). Now, (2.4.25) is obviously satisfied if $\Lambda = \mathrm{id}$ so we need only verify that it is also satisfied if $\Lambda = \pi$. In this case, (2.4.25) becomes

$$\gamma^0\gamma^\alpha\gamma^0 = \pi^\alpha{}_\beta\gamma^\beta, \qquad \alpha = 0,1,2,3,$$

i. e.,

$$\gamma^0\gamma^0\gamma^0 = \gamma^0$$

and

$$\gamma^0\gamma^i\gamma^0 = -\gamma^i \quad i = 1,2,3.$$

Since these are all easy to verify directly we have established the invariance of the Dirac equation under spatial inversion.

Notice, in particular, that if we write our Dirac wavefunction $\phi = \left(\begin{smallmatrix}\phi_L\\ \phi_R\end{smallmatrix}\right)$ as on page 109, then, under a spatial inversion, it transforms as follows:

$$\phi = \begin{pmatrix}\phi_L\\ \phi_R\end{pmatrix} \longrightarrow \gamma^0\phi = \begin{pmatrix}0 & \sigma_0\\ \sigma_0 & 0\end{pmatrix}\begin{pmatrix}\phi_L\\ \phi_R\end{pmatrix} = \begin{pmatrix}\phi_R\\ \phi_L\end{pmatrix}.$$

Switching the orientation of the spatial coordinate axes interchanges ϕ_L and ϕ_R and one therefore thinks of these two components as having opposite "handedness" or "chirality." In particular, in the $m = 0$ case one can regard (2.4.23) (respectively, (2.4.24)) as equations for a particle that is massless, spin one-half and "left-handed" (respectively, "right-handed"). Since the discovery that, in β-decay processes (in which a neutrino is emitted), parity is not conserved, these equations have been regarded as potential models for the neutrino. Lee and Yang ([**LY**]) suggested (2.4.23), but Feynman and Gell-Mann ([**FG-M**]) showed that the experimental evidence suggests (2.4.24) ("neutrinos spin to the left"). In the general (massive) case one thinks of a spin one-half particle as having two chiral components ϕ_L and ϕ_R, each of which has two additional components representing the possible spin states.

One final invariance property of the Dirac equation is worthy of note. If θ is some real *constant* so that $e^{\boldsymbol{i}\theta}$ is in $U(1)$, then

$$(\gamma^\alpha\partial_\alpha + \boldsymbol{i}m)\phi = 0$$

obviously implies

$$(\gamma^\alpha \partial_\alpha + \boldsymbol{i}m)(e^{\boldsymbol{i}\theta}\phi) = 0.$$

The Dirac equation is therefore invariant under the global $U(1)$-action $\phi \longrightarrow e^{\boldsymbol{i}\theta}\phi$. Just as was the case for the Schroedinger and Klein-Gordon equations, elevating this global symmetry to a local gauge symmetry in which θ is a function of x^0, x^1, x^2 and x^3 requires the presence of an electromagnetic gauge potential. In the physics literature such potentials are included by "minimal coupling," i.e., by replacing the ordinary derivatives ∂_α in the Dirac equation by "covariant derivatives" $\partial_\alpha + \boldsymbol{i}n\boldsymbol{A}_\alpha$ (see pages 76–80).

The problem of molding all of the information we have assembled thus far into a gauge theory model of the type described in Section 2.1 is complicated by a number of issues. Recall that in the spin zero case (Section 2.2) we began with an electromagnetic field (i.e., a connection on a $U(1)$-bundle over spacetime) and the various representations of $U(1)$ on $\mathbb{C}$ and, from them, constructed complex scalar fields, an action and the corresponding Euler-Lagrange equations. The resulting Klein-Gordon equations happened to be Lorentz invariant. Our point of departure in this section has been to insist at the outset on the relativistic invariance of the free particle equations. This led us to a wavefunction taking its values in $\mathbb{C}^4$ whose *external* symmetry (Lorentz invariance) was expressed in the form of a transformation law corresponding to a specific representation of the double cover $SL(2,\mathbb{C})$ of $\mathcal{L}_+^\uparrow$. This suggests that, in a corresponding gauge theory model, the wavefunction is a matter field on some $SL(2,\mathbb{C})$-bundle over spacetime. However, electrons are coupled to electromagnetic fields and these are connections on $U(1)$-bundles, not $SL(2,\mathbb{C})$-bundles, over spacetime. Furthermore, gauge invariance refers specifically to the *internal* symmetry of a particle reflected in the behavior of its wavefunction under changes in the local gauge potentials for the electromagnetic field so this notion also "lives" in a $U(1)$-bundle. To build a proper gauge theory model for Dirac electrons coupled to electromagnetic fields will require the "splicing together" of the external $SL(2,\mathbb{C})$-bundle and the internal $U(1)$-bundle into a single $SL(2,\mathbb{C}) \times U(1)$-bundle on which both the electron and the electromagnetic field may be thought to live. This turns out to be a relatively simple thing to do and we will outline the construction shortly.

A more delicate, and much more interesting, obstacle is one that we could evade altogether by simply continuing to restrict our attention to the spacetime $\mathbb{R}^{1,3}$ and its open submanifolds. From the perspective of the workaday world of particle physics this would be an entirely reasonable choice since it amounts to ignoring gravitational effects and these are generally negligible in elementary particle interactions in the laboratory. From the perspective of topology (and nonperturbative quantum field theory), however, such a choice would "evade" the best part. We will conclude this section with a brief synopsis of the issues involved in describing spin one-half particles that live in more

general spacetimes where gravitational effects are not neglected. We will deal with these issues in detail in the remaining chapters of the book.

A spacetime is a 4-dimensional (second countable, Hausdorff) manifold X with a "Lorentz metric" $\boldsymbol{g}$ (this is a semi-Riemannian metric with the property that each tangent space $T_x(X)$ has a basis $\{e_0,\ e_1,\ e_2,\ e_3\}$ for which $\boldsymbol{g}(x)(e_\alpha, e_\beta) = \eta_{\alpha\beta}$). Thus, each $T_x(X)$ with its inner product $\boldsymbol{g}(x)$ can be identified with $\mathbb{R}^{1,3}$. The general Lorentz group $\mathcal{L}$ therefore acts on the orthonormal bases of each $T_x(X)$. We are, however, only interested in bases related by elements of $\mathcal{L}_+^\uparrow$. Although one can isolate such a collection of bases at each $T_x(X)$ individually just by selecting an isomorphism onto $\mathbb{R}^{1,3}$, an unambiguous choice over the entire manifold X is possible only if X is assumed orientable and "time orientable." We will discuss this latter condition in more detail in Chapter 3; essentially, one assumes the existence of a vector field $\boldsymbol{V}$ on X that is timelike ($\boldsymbol{g}(x)(\boldsymbol{V}(x), \boldsymbol{V}(x))) > 0$ for each $x \in X$) and so makes a smooth selection over X of a timelike direction at each point that we may (arbitrarily) decree "future-directed." We will adopt both of these assumptions and thereby obtain, at each point, a family of oriented, time oriented, orthonormal bases for the tangent space related by elements of $\mathcal{L}_+^\uparrow$.

Remark: We will find that compact spacetimes necessarily violate certain rather basic notions of causality (they contain closed timelike curves). For this reason we will henceforth restrict our attention to the noncompact variety.

Now, Lorentz invariance means invariance under $\mathcal{L}_+^\uparrow$. When $X = \mathbb{R}^{1,3}$, bundles over X are trivial so gauge fields, matter fields, etc., can all be identified with objects defined on X. Furthermore, each tangent space can be canonically identified with X itself so $\mathcal{L}_+^\uparrow$ acts on X. In the general case, none of this is true. In particular, choosing Lorentz frames and acting by $\mathcal{L}_+^\uparrow$ on such frames cannot take place globally on all of X, but only point by point. To describe all of this precisely we will build (in Chapter 3) the "oriented, time oriented, orthonormal frame bundle" of X. This is a principal $\mathcal{L}_+^\uparrow$-bundle

$$\mathcal{L}_+^\uparrow \hookrightarrow \mathcal{L}(X) \xrightarrow{\mathcal{P}_\mathcal{L}} X$$

over X whose fibers consist of the oriented, time oriented, orthonormal bases for the tangent spaces to X. The action of $\mathcal{L}_+^\uparrow$ on $\mathcal{L}(X)$ will simply carry one such basis onto another (above the same point in X) and one can make sense of "Lorentz invariance" for matter fields associated with this bundle by some representation of $\mathcal{L}_+^\uparrow$.

The frame bundle $\mathcal{L}_+^\uparrow \hookrightarrow \mathcal{L}(X) \longrightarrow X$ exists for every oriented, time oriented spacetime and so presents no real obstacle to our program. However, there is an obstacle (or, rather, "obstruction"). The fibers of $\mathcal{L}(X)$ are all isomorphic to $\mathcal{L}_+^\uparrow$, but the Dirac wavefunction did not arise from a representation of $\mathcal{L}_+^\uparrow$ on $\mathbb{C}^4$. Rather, it was determined by the representation $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$

of $SL(2,\mathbb{C})$ on $\mathbb{C}^4$. Thus, to regard a Dirac electron as a matter field in the sense of Section 2.1 it will be necessary to "globalize" over all of X the double cover

$$\begin{array}{ccc} SL\,(2,\mathbb{C}) & & \\ \Big\downarrow & \text{Spin} & \\ \mathcal{L}_+^{\uparrow} & & \end{array}$$

The frame bundle provides a copy of $\mathcal{L}_+^{\uparrow}$ above every $x \in X$ so what we need is an $SL(2,\mathbb{C})$-bundle

$$SL(2,\mathbb{C}) \hookrightarrow S(X) \xrightarrow{\mathcal{P}_S} X \tag{2.4.26}$$

over X and a map of $S(X)$ onto $\mathcal{L}(X)$ that is, in effect, the spinor map of $\mathcal{P}_S^{-1}(x)$ onto $\mathcal{P}_{\mathcal{L}}^{-1}(x)$ for each $x \in X$. More precisely, a **spinor structure** for X consists of a principal $SL(2,\mathbb{C})$-bundle (2.4.26) over X and a map

$$\lambda : S(X) \longrightarrow \mathcal{L}(X)$$

such that

$$\mathcal{P}_{\mathcal{L}}(\lambda(p)) = \mathcal{P}_S(p)$$

and

$$\lambda(p \cdot g) = \lambda(p) \ \cdot \ \mathrm{Spin}(g)$$

for all $p \in S(X)$ and all $g \in SL(2,\mathbb{C})$. The following diagram therefore commutes.

$$\begin{array}{ccccc} S(X) \times SL(2,\mathbb{C}) & \xrightarrow{\ \bullet\ } & S(X) & \xrightarrow{\ \mathcal{P}_S\ } & X \\ \Big\downarrow{\scriptstyle \lambda \times \mathrm{Spin}} & & \Big\downarrow{\scriptstyle \lambda} & & \Big\updownarrow{\scriptstyle \mathrm{id}_X} \\ \mathcal{L}(X) \times \mathcal{L}_+^{\uparrow} & \xrightarrow[\ \bullet\ ]{} & \mathcal{L}(X) & \xrightarrow[\ \mathcal{P}_{\mathcal{L}}\ ]{} & X \end{array}$$

It is at this point that we encounter our obstruction. Not every oriented, time oriented spacetime X admits a spinor structure and, for those which do not, it is simply not possible to define the Dirac wavefunction for a massive, spin one-half particle. Confidence in the Dirac equation is such that this is generally regarded as adequate justification for dismissing as physically unacceptable any spacetime without a spinor structure. From our point of view the interesting part of all of this is that the existence or nonexistence of a spinor structure is a purely topological question about X. We will show that there is a certain Čech cohomology class $w_2(X) \in \check{\mathrm{H}}^2(X; \mathbb{Z}_2)$, called the "second Stiefel-Whitney class" of X, the vanishing of which is a necessary and sufficient condition for the existence of a spinor structure on X : $w_2(X)$ is the obstruction to the existence of a spinor structure (see Section 6.5).

We will show that all of the spacetimes of interest to us do, in fact, admit spinor structures. With this the construction of a gauge theory model for a free Dirac electron proceeds as follows: X is a (noncompact) spacetime manifold with $w_2(X) = 0$. The vector space $\mathcal{V}$ in which the wavefunction will take its values is $\mathbb{C}^4$. The (external) symmetry group G is $SL(2, \mathbb{C})$ and the representation ρ of $SL(2, \mathbb{C})$ on $\mathbb{C}^4$ is

$$\rho = D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})} : SL(2, \mathbb{C}) \longrightarrow GL(\mathbb{C}^4).$$

One can define an inner product $\langle\ ,\ \rangle$ on $\mathbb{C}^4$ relative to which this representation is orthogonal as follows: First define the "twisted" Hermitian form H: $\mathbb{C}^4 \times \mathbb{C}^4 \longrightarrow \mathbb{C}$ by

$$\mathrm{H}((z_1, \ldots, z_4), (w_1, \ldots, w_4)) = z_1\bar{w}_3 + z_2\bar{w}_4 + z_3\bar{w}_1 + z_4\bar{w}_2.$$

Regarding $z,\ w \in \mathbb{C}^4$ as column matrices, this is equivalent to

$$\mathrm{H}(z,\ w) = z^\top \gamma^0 w,$$

where

$$\gamma^0 = \begin{pmatrix} 0 & \sigma_0 \\ \sigma_0 & 0 \end{pmatrix}$$

as on page 96. A simple computation shows that

$$\mathrm{H}(\rho(g)(z), \rho(g)(w)) = \mathrm{H}(z, w)$$

for all $g \in SL(2, \mathbb{C})$. The required inner product on $\mathbb{C}^4$ is then given by

$$\langle z, w \rangle = \frac{1}{2}(\mathrm{H}(z, w) + \mathrm{H}(w, z)).$$

The principal $SL(2,\mathbb{C})$-bundle over X is a spinor bundle

$$SL(2,\mathbb{C}) \hookrightarrow S(X) \xrightarrow{\mathcal{P}_S} X.$$

A **free Dirac electron** is a corresponding matter field, i.e., a smooth map

$$\phi : S(X) \longrightarrow \mathbb{C}^4$$

that is equivariant, i.e., satisfies

$$\phi(p \cdot g) = g^{-1} \cdot \phi(p) = \begin{pmatrix} g^{-1} & 0 \\ 0 & \bar{g}^{\top} \end{pmatrix} \phi(p).$$

Remarks: These electrons are "free" in the sense that they are not coupled to an electromagnetic field. Such an electromagnetic field does not live on the spinor bundle, but rather (as a connection) on a $U(1)$-bundle over X. Shortly we will describe how to splice the spinor bundle and the $U(1)$-bundle together into a single $SL(2,\mathbb{C}) \times U(1)$-bundle on which an interaction can be described. Notice, however, that our "free" electron is not entirely free if the underlying spacetime X represents a nontrivial gravitational field. Such influences enter these considerations in the form of a connection on the spinor bundle that is essentially the lift of the canonical (Levi-Civita) connection on the frame bundle (see Section 3.3). With this and the potential function $U : \mathbb{C}^4 \longrightarrow \mathbb{R}$ given by $U(z) = \frac{1}{2}m\|z\|^2 = \frac{1}{2}m\langle z, z\rangle$ one can write down an action whose Euler-Lagrange equations constitute the general spacetime version of the Dirac equation. The details are available in Section 6.4 of [**Bl**]. Needless to say, when $X = \mathbb{R}^{1,3}$ and the matter field is pulled back by the standard global cross-section of the (necessarily) trivial spinor bundle, this reduces to $(\gamma^\alpha \partial_\alpha + \boldsymbol{i}m)\phi = 0$.

A free Dirac electron can, of course, also be regarded as a cross-section of the vector bundle associated to $SL(2,\mathbb{C}) \hookrightarrow S(X) \longrightarrow X$ by the representation $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$ (see pages 49–50). More generally, if ρ is any representation of $SL(2,\mathbb{C})$ on some $\mathbb{C}^k$, then an equivariant $\mathbb{C}^k$-valued map on $S(X)$ (or, equivalently, a cross-section of the associated vector bundle $S(X) \times_\rho \mathbb{C}^k$) is called a k-component **spinor field** of type ρ on X. 4-component spinor fields of type $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$ are generally called **Dirac spinor** fields. 2-component spinor fields of type $D^{(\frac{1}{2},0)}$, or $D^{(0,\frac{1}{2})}$ are called **Weyl spinor** fields.

Finally, we outline the "splicing" procedure for building from the spinor bundle (where electrons live) and a $U(1)$-bundle (where electromagnetic fields live) a single $SL(2,\mathbb{C}) \times U(1)$-bundle (where both live and therefore can interact). Consider, in general, two principal bundles over the base X:

$$G_1 \hookrightarrow P_1 \xrightarrow{\mathcal{P}_1} X$$
$$G_2 \hookrightarrow P_2 \xrightarrow{\mathcal{P}_2} X$$

Let

$$P_1 \circ P_2 = \{(p_1, p_2) \in P_1 \times P_2 : \mathcal{P}_1(p_1) = \mathcal{P}_2(p_2)\}.$$

Then $P_1 \circ P_2$ is a submanifold of $P_1 \times P_2$ and will be the total space of the spliced bundle. Define

$$\mathcal{P}_{12} : P_1 \circ P_2 \longrightarrow X$$

by

$$\mathcal{P}_{12}(p_1, p_2) = \mathcal{P}_1(p_1) = \mathcal{P}_2(p_2).$$

Then $\mathcal{P}_{12}$ is a smooth map of $P_1 \circ P_2$ onto X. Define a smooth right action of $G_1 \times G_2$ on $P_1 \circ P_2$ by

$$(p_1, p_2) \cdot (g_1, g_2) = (p_1 \cdot g_1, p_2 \cdot g_2).$$

Then

$$G_1 \times G_2 \hookrightarrow P_1 \circ P_2 \xrightarrow{\mathcal{P}_{12}} X$$

is a smooth principal $G_1 \times G_2$-bundle over X.

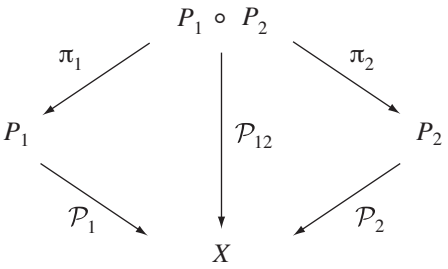
Next we define maps $\pi_1 : P_1 \circ P_2 \longrightarrow P_1$ and $\pi_2 : P_1 \circ P_2 \longrightarrow P_2$ by $\pi_i(p_1, p_2) = p_i,\ i = 1, 2$. Letting e_1 and e_2 denote the identities in G_1 and G_2 we identify $\{e_1\} \times G_2$ with G_2 and $G_1 \times \{e_2\}$ with G_1. We then have principal bundles

$$G_2 \hookrightarrow P_1 \circ P_2 \xrightarrow{\pi_1} P_1$$

and

$$G_1 \hookrightarrow P_1 \circ P_2 \xrightarrow{\pi_2} P_2$$

for which the following diagram commutes:



Now, suppose we have connections $\boldsymbol{\omega}_1$ on $\mathcal{P}_1 : P_1 \longrightarrow X$ and $\boldsymbol{\omega}_2$ on $\mathcal{P}_2 : P_2 \longrightarrow X$. Identify $\mathcal{G}_1$ with $\mathcal{G}_1 \times \{0\} \subseteq \mathcal{G}_1 \oplus \mathcal{G}_2$ and $\mathcal{G}_2$ with $\{0\} \times \mathcal{G}_2 \subseteq \mathcal{G}_1 \oplus \mathcal{G}_2$. Then $\pi_1^* \boldsymbol{\omega}_1$ is a connection on $\pi_2 : P_1 \circ P_2 \longrightarrow P_2$, $\pi_2^* \boldsymbol{\omega}_2$ is a connection on $\pi_1 : P_1 \circ P_2 \longrightarrow P_1$ and $\pi_1^* \boldsymbol{\omega}_1 \oplus \pi_2^* \boldsymbol{\omega}_2$ is a connection on $\mathcal{P}_{12} : P_1 \circ P_2 \longrightarrow X$.

Finally, let $\mathcal{V}$ be a vector space and let $\rho_1 : G_1 \longrightarrow GL(\mathcal{V})$ and $\rho_2 : G_2 \longrightarrow GL(\mathcal{V})$ be two representations that satisfy

$$\rho_1(g_1) \circ \rho_2(g_2) = \rho_2(g_2) \circ \rho_1(g_1) \tag{2.4.27}$$

for all $g_1 \in G_1$ and $g_2 \in G_2$. Then we can define

$$\rho_1 \times \rho_2 : G_1 \times G_2 \longrightarrow GL(\mathcal{V})$$

by

$$(\rho_1 \times \rho_2)(g_1, g_2) = \rho_1(g_1) \circ \rho_2(g_2) = \rho_2(g_2) \circ \rho_1(g_1)$$

and obtain a representation of $G_1 \times G_2$ on $\mathcal{V}$ with associated left action on $\mathcal{V}$ given by

$$(g_1, g_2) \cdot \xi = (\rho_1(g_1) \circ \rho_2(g_2))(\xi).$$

Remark: ρ_1 and ρ_2 must commute, i.e., satisfy (2.4.27), to ensure that $\rho_1 \times \rho_2$ is a representation:

$$\begin{aligned}(\rho_1 \times \rho_2)\ ((g_1, g_2)\ (g_1', g_2')) &= (\rho_1 \times \rho_2)\ ((g_1 g_1', g_2 g_2')) = \\ \rho_1(g_1 g_1') \circ \rho_2(g_2 g_2') &= \rho_1(g_1) \circ \rho_1(g_1') \circ \rho_2(g_2) \circ \rho_2(g_2') \\ &= \rho_1(g_1) \circ \rho_2(g_2) \circ \rho_1(g_1') \circ \rho_2(g_2') \\ &= (\rho_1 \times \rho_2)(g_1, g_2) \circ (\rho_1 \times \rho_2)\ (g_1',\ g_2').\end{aligned}$$

Now, a matter field on $G_1 \times G_2 \hookrightarrow P_1 \circ P_2 \xrightarrow{\mathcal{P}_{12}} X$ associated with $\rho_1 \times \rho_2$: $G_1 \times G_2 \longrightarrow GL(\mathcal{V})$ is a map $\phi : P_1 \circ P_2 \longrightarrow \mathcal{V}$ satisfying

$$\phi((p_1, p_2) \cdot (g_1, g_2)) = (g_1^{-1},\ g_2^{-1}) \cdot \phi\,(p_1, p_2),$$

i.e.,

$$\phi\,((p_1 \cdot g_1, p_2 \cdot g_2)) = \left(\rho_1\left(g_1^{-1}\right)\right)\left(\left(\rho_2(g_2^{-1})\right)\ (\phi\,(p_1, p_2))\right).$$

Now we apply this construction to the following special case. Begin with an oriented, time oriented spacetime X and a spinor bundle

$$SL(2,\ \mathbb{C}) \hookrightarrow S(X) \xrightarrow{\mathcal{P}_S} X$$

on X (this will be $G_1 \hookrightarrow P_1 \xrightarrow{\mathcal{P}_1} X$). Let $\boldsymbol{\omega}_1$ be the spinor connection on $S(X)$ referred to in the Remark on page 117. Take $\mathcal{V} = \mathbb{C}^4$ and let ρ_1 be the representation $D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$ of $SL(2, \mathbb{C})$ on $\mathbb{C}^4$. Thus,

$$(\rho_1(g_1)) \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix} = \begin{pmatrix} g_1 & 0 \\ 0 & \left(\bar{g}_1^{\top}\right)^{-1} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix}$$

for each g_1 in $SL(2, \mathbb{C})$. Next let $U(1) \hookrightarrow P \xrightarrow{\mathcal{P}_2} X$ be some principal $U(1)$-bundle over X and let $\boldsymbol{\omega}_2$ be a connection on it (representing

some electromagnetic field to which the Dirac electron will respond). Take $\rho_2\colon U(1) \longrightarrow GL(\mathbb{C}^4)$ to be the representation given by

$$(\rho_2\,(g_2)) \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix} = g_2 \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix} = \begin{pmatrix} g_2 z_1 \\ \vdots \\ g_2 z_4 \end{pmatrix}$$

for each $g_2 \in U(1)$. Note that $\rho_1(g_1) \circ \rho_2(g_2) = \rho_2(g_2) \circ \rho_1(g_1)$ as required in (2.4.27). Thus, we have a representation

$$\rho_1 \times \rho_2 : SL(2,\mathbb{C}) \times U(1) \longrightarrow GL(\mathbb{C}^4)$$

given by

$$\begin{aligned}(\rho_1 \times \rho_2)(g_1, g_2) \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix} &= \rho_1(g_1) \circ \rho_2\,(g_2) \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix} \\ &= \begin{pmatrix} g_1 & 0 \\ 0 & \left(\bar{g}_1^{\top}\right)^{-1} \end{pmatrix} \begin{pmatrix} g_2 z_1 \\ \vdots \\ g_2 z_4 \end{pmatrix} \\ &= g_2 \begin{pmatrix} g_1 & 0 \\ 0 & \left(\bar{g}_1^{\top}\right)^{-1} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_4 \end{pmatrix}.\end{aligned}$$

Now we splice the two bundles together to obtain

$$SL\,(2,\mathbb{C}) \times U(1) \hookrightarrow S(X) \circ P \longrightarrow X.$$

A **Dirac electron** (coupled to the $U(1)$-potential $\boldsymbol{\omega}_2$) is then a smooth map $\phi : S(X) \circ P \longrightarrow \mathbb{C}^4$ satisfying

$$\phi\,((p_1, p_2) \cdot (g_1, g_2)) = g_2^{-1} \begin{pmatrix} g_1^{-1} & 0 \\ 0 & \bar{g}_1^{\top} \end{pmatrix} \phi\,(p_1, p_2),$$

i.e.,

$$\phi\,(\,p_1 \cdot g_1,\, p_2 \cdot g_2\,) = \begin{pmatrix} (g_1\, g_2)^{-1} & 0 \\ 0 & (\overline{g_1\, g_2})^{\top} \end{pmatrix} \phi\,(\,p_1,\, p_2),$$

where $g_1\, g_2$ is the entrywise product of $g_1 = e^{\boldsymbol{i}\theta_1}$ with $g_2 \in SL(2,\mathbb{C})$. With the equipment now available one can write down an action functional whose Euler-Lagrange equations describe the interaction of a massive, spin one-half

particle with an electromagnetic field (the details are available in Section 7.2 of [**Bl**]).

As one final illustration of this technique we will sketch an analogous construction for the interaction of a nucleon with a classical Yang-Mills field (for the details, see Section 7.3 of [**Bl**]). Here we face the same problem as in the case of a Dirac electron coupled to an electromagnetic field. A Yang-Mills field is given by a connection on a principal $SU(2)$-bundle over spacetime (Section 6.3 of [**N4**]), whereas a nucleon (proton/neutron) is a massive, spin one-half particle and therefore lives in a spinor bundle. There is an additional complication, however. A nucleon is a proton/neutron doublet, i.e., its wavefunction has a proton component and a neutron component and so must take its values in $\mathcal{V} = \mathbb{C}^4 \oplus \mathbb{C}^4 = \mathbb{C}^8$.

We begin then with an oriented, time oriented spacetime X and a spinor bundle

$$SL(2,\mathbb{C}) \hookrightarrow S(X) \xrightarrow{\mathcal{P}_S} X.$$

$\boldsymbol{\omega}_1$ is again the spinor connection referred to in the Remark on page 117. Now take $\mathcal{V} = \mathbb{C}^4 \oplus \mathbb{C}^4$, which we identify with the set of $\binom{v_1}{v_2}$ with $v_1, v_2 \in \mathbb{C}^4$. Letting $\rho = D^{(\frac{1}{2},0)} \oplus D^{(0,\frac{1}{2})}$ we define $\rho_1 : SL(2,\mathbb{C}) \longrightarrow \mathbb{C}^4 \ \oplus \ \mathbb{C}^4$ by

$$(\,\rho_1(\,g_1))(v) = (\,\rho_1(\,g_1)) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} (\rho(g_1))(v_1) \\ (\rho(g_1))(v_2) \end{pmatrix}.$$

Now let $SU(2) \hookrightarrow P \xrightarrow{\mathcal{P}_2} X$ be some principal $SU(2)$-bundle over X and $\boldsymbol{\omega}_2$ some connection on it (representing the Yang-Mills potential to which the nucleon is coupled). Define $\rho_2 : SU(2) \longrightarrow GL(\mathbb{C}^4 \oplus \mathbb{C}^4)$ as follows: For each

$$g_2 = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

in $SU(2)$,

$$(\,\rho_2(\,g_2\,))(v) = (\,\rho_2(\,g_2\,)) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

$$= \begin{pmatrix} \alpha v_1 + \beta v_2 \\ \gamma v_1 + \delta v_2 \end{pmatrix}.$$

A simple calculation shows that $\rho_1(\,g_1) \circ \rho_2(\,g_2\,) = \rho_2(\,g_2) \circ \rho_1(\,g_1)$ so we have a representation

$$\rho_1 \times \rho_2 : SL(2,\mathbb{C}) \times SU(2) \longrightarrow GL(\mathbb{C}^4 \oplus \mathbb{C}^4).$$

Letting

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_4 \\ w_1 \\ \vdots \\ w_4 \end{pmatrix}$$

we have

$$(\rho_1 \times \rho_2(g_1, g_2))(v) = \begin{pmatrix} (\rho(g_1))(\alpha v_1 + \beta v_2) \\ (\rho(g_1))(\gamma v_1 + \delta v_2) \end{pmatrix}$$

$$= \begin{pmatrix} \begin{pmatrix} g_1 & 0 \\ 0 & (\bar{g}_1^{\top})^{-1} \end{pmatrix} & \begin{pmatrix} \alpha z_1 + \beta w_1 \\ \vdots \\ \alpha z_4 + \beta w_4 \end{pmatrix} \\ \begin{pmatrix} g_1 & 0 \\ 0 & (\bar{g}_1^{\top})^{-1} \end{pmatrix} & \begin{pmatrix} \gamma z_1 + \delta w_1 \\ \vdots \\ \gamma z_4 + \delta w_4 \end{pmatrix} \end{pmatrix}.$$

A **nucleon field** (coupled to an $SU(2)$-Yang-Mills potential $\boldsymbol{\omega}_2$) is then a smooth map $\phi : S(X) \circ P \longrightarrow \mathbb{C}^4 \oplus \mathbb{C}^4$ such that

$$\phi\,((p_1, p_2) \cdot (g_1, g_2)) = (g_1^{-1}, g_2^{-1}) \cdot \phi(p_1, p_2).$$

Writing $\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$ we have

$$\begin{pmatrix} \phi_1(p_1 \cdot g_1, p_2 \cdot g_2) \\ \phi_2(p_1 \cdot g_1, p_2 \cdot g_2) \end{pmatrix} = \begin{pmatrix} (\rho(g_1^{-1}))(\alpha\phi_1(p_1 \cdot g_1, p_2 \cdot g_2) \\ \qquad\qquad +\beta\phi_2(p_1 \cdot g_1, p_2 \cdot g_2)) \\ (\rho(g_1^{-1}))(\gamma\phi_1(p_1 \cdot g_1, p_2 \cdot g_2) \\ \qquad\qquad +\delta\phi_2(p_1 \cdot g_1, p_2 \cdot g_2)) \end{pmatrix}.$$

ϕ_1 is called the **proton component** of ϕ, while ϕ_2 is its **neutron component**. Note the "mixing" of the proton and neutron components due to the $SU(2)$-action.

2.5 $SU(2)$-Yang-Mills-Higgs Theory on $\mathbb{R}^n$

The motivation behind our final example of a classical gauge theory is a process known as "dimensional reduction." The example itself is of profound significance to physics, arises naturally from the pure Yang-Mills theory on $\mathbb{R}^4$ that was the subject of [**N4**], provides an extraordinary insight into the true nature of the Dirac magnetic monopole and gives rise to even deeper

connections between physics and topology than those we have encountered thus far. We will begin by simply enumerating, without motivation, the eight items required (in Section 2.1) for the construction of a classical gauge theory. Then we will review the structure of pure $SU(2)$- Yang-Mills theory on $\mathbb{R}^4$ and show how our example arises from it. Finally, we will describe a number of the remarkable properties of the model, both physical and mathematical.

The base manifold is $X = \mathbb{R}^n$ with its usual orientation and Riemannian metric. For the vector space $\mathcal{V}$ we take the Lie algebra $su(2)$ of 2×2 complex matrices that are skew-Hermitian and tracefree. The (positive definite) inner product on $su(2)$ is given by $\langle A, B\rangle = -2$ trace (AB) ($-$trace(AB) is the Killing form of $su(2)$ and the 2 is a matter of convenience). The Lie group G is taken to be $SU(2)$ and $\rho : SU(2) \longrightarrow GL(su(2))$ is the adjoint representation

$$ad_g(A) = gAg^{-1}$$

for all $g \in SU(2)$ and $A \in su(2)$. Note that $\langle ad_g(A), ad_g(B)\rangle = \langle gAg^{-1}, gBg^{-1}\rangle = -2$ trace$((gAg^{-1})(gBg^{-1})) = -2$ trace$(g(AB)g^{-1}) = -2$ trace$(AB) = \langle A, B\rangle$, as required. Since every bundle over $\mathbb{R}^n$ is trivial, we will trivialize at the outset and take

$$SU(2) \hookrightarrow \mathbb{R}^n \times SU(2) \xrightarrow{\mathcal{P}} \mathbb{R}^n$$

as our bundle, where the right action of $SU(2)$ on $\mathbb{R}^n \times SU(2)$ is given by

$$p \cdot g = (x, h) \cdot g = (x, hg)$$

for all $p = (x, h) \in \mathbb{R}^n \times SU(2)$ and all $g \in SU(2)$. There is a natural global cross-section $s : \mathbb{R}^n \longrightarrow \mathbb{R}^n \times SU(2)$ given by

$$s(x) = (x, e)$$

where, for convenience, we write e for the identity element in $SU(2)$. Any other global cross-section then has the form

$$\begin{aligned} s^g : \mathbb{R}^n &\longrightarrow \mathbb{R}^n \times SU(2) \\ s^g(x) &= s(x) \cdot g(x) \\ &= (x, e) \cdot g(x) \\ &= (x, g(x)) \end{aligned}$$

for some smooth map $g : \mathbb{R}^n \longrightarrow SU(2)$ (Exercise 4.3.5 of [**N4**]). The cross-section s^g gives rise to an automorphism of the bundle (i.e., a global gauge transformation) in the usual way (see page 343 of [**N4**]):

$$\begin{aligned} s(x) \cdot h &\longrightarrow s^g(x) \cdot h \\ (x, e) \cdot h &\longrightarrow (x, g(x)) \cdot h \\ (x, h) &\longrightarrow (x, g(x)h). \end{aligned}$$

Thus, one can identify a gauge transformation with a smooth map $g : \mathbb{R}^n \longrightarrow SU(2)$ which multiplies in the fibers on the left.

Because of the triviality of the bundle, any connection $\boldsymbol{\omega}$ on $SU(2) \hookrightarrow \mathbb{R}^n \times SU(2) \longrightarrow \mathbb{R}^n$ is uniquely determined by its gauge potential

$$\boldsymbol{\mathcal{A}} = s^*\boldsymbol{\omega},$$

which is an $su(2)$-valued 1-form on $\mathbb{R}^n$. Furthermore, any $su(2)$-valued 1-form on $\mathbb{R}^n$ is the pullback by s of a unique connection on $SU(2) \hookrightarrow \mathbb{R}^n \times SU(2) \longrightarrow \mathbb{R}^n$ (page 333, [**N4**]). Thus, we may restrict our attention entirely to globally defined gauge potentials $\boldsymbol{\mathcal{A}}$ on $\mathbb{R}^n$. Relative to standard coordinates on $\mathbb{R}^n$ we write

$$\boldsymbol{\mathcal{A}} = s^*\boldsymbol{\omega} = \mathcal{A}_\alpha dx^\alpha,$$

where each $\mathcal{A}_\alpha,\ \alpha = 1, \ldots, n$, takes values in $su(2)$ (see (2.5.2) below). A gauge transformation $g : \mathbb{R}^n \longrightarrow SU(2)$ gives a new gauge potential

$$\boldsymbol{\mathcal{A}}^g = (s^g)^*\boldsymbol{\omega}$$

related to $\boldsymbol{\mathcal{A}}$ by

$$\boldsymbol{\mathcal{A}}^g = g^{-1}\boldsymbol{\mathcal{A}}g + g^{-1}dg,$$

where dg is the entrywise exterior derivative of $g : \mathbb{R}^n \longrightarrow SU(2)$ and the products are matrix products (we will do an explicit calculation of this sort for the "t' Hooft-Polyakov monopole" somewhat later). The curvature $\boldsymbol{\Omega}$ of $\boldsymbol{\omega}$ is likewise uniquely determined by the field strength

$$\boldsymbol{\mathcal{F}} = s^*\boldsymbol{\Omega} = d\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{A}} \wedge \boldsymbol{\mathcal{A}} = \frac{1}{2}\mathcal{F}_{\alpha\beta}dx^\alpha \wedge dx^\beta,$$

where

$$\mathcal{F}_{\alpha\beta} = \partial_\alpha\mathcal{A}_\beta - \partial_\beta\mathcal{A}_\alpha + [\mathcal{A}_\alpha, \mathcal{A}_\beta], \quad \alpha, \beta = 1, \ldots, n.$$

A gauge transformation $g : \mathbb{R}^n \longrightarrow SU(2)$ gives a new field strength

$$\boldsymbol{\mathcal{F}}^g = (s^g)^*\boldsymbol{\Omega}$$

related to $\boldsymbol{\mathcal{F}}$ by

$$\boldsymbol{\mathcal{F}}^g = g^{-1}\boldsymbol{\mathcal{F}}g.$$

In this context a matter field is a smooth $su(2)$-valued map Φ on $\mathbb{R}^n \times SU(2)$ that satisfies

$$\begin{aligned}\Phi(\,p \cdot g) &= g^{-1} \cdot \Phi(p)\\ \Phi(\,(x,h) \cdot g) &= ad_{g^{-1}}(\Phi(x,h))\\ \Phi\ (x, hg\ ) &= g^{-1}\Phi(x,h)g\end{aligned}$$

for all $(x,h) \in \mathbb{R}^n \times SU(2)$ and all $g \in SU(2)$. When, as in this case, $\mathcal{V}$ is the Lie algebra $\mathcal{G}$ of the structure group G and ρ is the adjoint representation of G on $\mathcal{G}$, a matter field is referred to as a **Higgs field**. The triviality of the

bundle in our present circumstances allows us to identify the Higgs field with its pullback by the global cross-section s:

$$\phi = s^*\mathbf{\Phi} = \mathbf{\Phi} \circ s$$
$$\phi(x) = \mathbf{\Phi}(x, e).$$

Under a gauge transformation $g : \mathbb{R}^n \longrightarrow SU(2)$,

$$\phi^g = (s^g)^*\mathbf{\Phi} = g^{-1}\phi g$$

because

$$\begin{aligned}((s^g)^*\mathbf{\Phi})(x) &= \mathbf{\Phi}(s^g(x)) \\ &= \mathbf{\Phi}(x, g) \\ &= \mathbf{\Phi}(x, eg) \\ &= g^{-1}\mathbf{\Phi}(x, e)g \\ &= g^{-1}\phi(x)g.\end{aligned}$$

The next item on the agenda (#7 of Section 2.1) is the potential function $U : su(2) \longrightarrow \mathbb{R}$. This plays a rather peculiar role in the story we wish to tell. Initially we adopt what is called the **Georgi-Glashow potential** $U : su(2) \longrightarrow \mathbb{R}$ given by

$$U(A) = \frac{\lambda}{8}(||A||^2 - 1)^2,$$

where $\lambda \geq 0$ is a constant and $\| A \|^2 = \langle A, A \rangle = -2 \text{ trace}(A^2)$, noting that $U(g \cdot A) = U(gAg^{-1}) = U(A)$ as required. Shortly, however, we will take λ to be zero and retain only a vestige of the potential in the form of an asymptotic boundary condition that it imposes on the Higgs field ϕ (see pages 132–133).

In order to describe the appropriate action (#8 of Section 2.1) for our example we must anticipate a few results on differential forms that will be proved later (Chapter 4). We will content ourselves with just a brief description of those particular items required for the example. We have already introduced real- and vector-valued 0-forms (pages 10 and 12), 1-forms (pages 9 and 12), and 2-forms (pages 11 and 12). k-forms, for integers $k \geq 3$, are defined analogously and all of the familiar algebraic and analytic operations on 0-, 1-, and 2-forms extend to this more general context. For example, a real-valued 3-form on a manifold X is a map $\boldsymbol{\alpha}$ that assigns to each $p \in X$ a real-valued trilinear function $\boldsymbol{\alpha}_p$ on $T_p(X) \times T_p(X) \times T_p(X)$ that is skew-symmetric (changes sign when-ever two of its arguments are interchanged) and smooth in the sense that, for any $\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3 \in \mathcal{X}(X)$, the function $\boldsymbol{\alpha}(\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3)$ on X defined by $(\boldsymbol{\alpha}(\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3))(p) = \boldsymbol{\alpha}_p(\boldsymbol{V}_1(p), \boldsymbol{V}_2(p), \boldsymbol{V}_3(p))$ is in $C^\infty(X)$. These arise, for example, as exterior derivatives of 2-forms and wedge products of 1-forms and 2-forms, or of three 1-forms (all of which will be defined carefully in Chapter 4).

In general, the set of k-forms on an n-dimensional manifold X is denoted $\Lambda^k(X)$ and admits a natural $C^\infty(X)$-module structure (and so in particular, is a real vector space). If (U, φ) is a chart for X with coordinate functions

$x^1, \ldots, x^n$, then any $\boldsymbol{\alpha} \in \Lambda^k(X)$ has a local coordinate expression

$$\boldsymbol{\alpha} = \frac{1}{k!}\, \alpha_{i_1 \cdots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}$$

(summation over $i_1, \ldots, i_k = 1, \ldots, n$), where each $\alpha_{i_1 \cdots i_k}$ is in $C^\infty(U)$. The exterior derivative $d\boldsymbol{\alpha}$ of $\boldsymbol{\alpha} \in \Lambda^k(X)$ is an element of $\Lambda^{k+1}(X)$ which, locally, is given by

$$d\boldsymbol{\alpha} = \frac{1}{k!}(d\alpha_{i_1 \cdots i_k}) \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}.$$

There are no nonzero k-forms on X if $k > n$ and, for $0 \leq k \leq n$, we will show that the dimension of $\Lambda^k(X)$ as a $C^\infty(X)$-module is $\binom{n}{k}$. Since $\binom{n}{n-k} = \binom{n}{k}$, the modules $\Lambda^k(X)$ and $\Lambda^{n-k}(X)$ are isomorphic. We will find that, when X is oriented and has a metric (Riemannian or semi-Riemannian), then there is a natural isomorphism

$$^* : \Lambda^k(X) \longrightarrow \Lambda^{n-k}(X),$$

called the Hodge star operator. Moreover, $\dim \Lambda^n(X) = \binom{n}{n} = 1$ and, when X is oriented and has a metric, there is a distinguished generator for $\Lambda^n(X)$ called the metric volume form and denoted **vol** (in standard coordinates on $\mathbb{R}^n$ this is just $dx^1 \wedge \cdots \wedge dx^n$). In particular, for any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Lambda^k(X)$, $\boldsymbol{\alpha} \wedge {}^*\boldsymbol{\beta}$ is in $\Lambda^n(X)$ and so is a multiple, by some element of $C^\infty(X)$, of **vol**. We denote this element of $C^\infty(X)$ by $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$:

$$\boldsymbol{\alpha} \wedge^* \boldsymbol{\beta} = \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \mathbf{vol}.$$

This defines an inner product on $\Lambda^k(X)$ and, when $\boldsymbol{\beta} = \boldsymbol{\alpha}$, we will write $\langle \boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle = \| \boldsymbol{\alpha} \|^2$. In the Riemannian case we will find that ${}^{**}\boldsymbol{\beta} = (-1)^{k(n-k)}\boldsymbol{\beta}$ and that it follows from this that the Hodge star operator is actually an isometry. k-forms that vanish outside of a compact set can be integrated over k-dimensional manifolds. Indeed, we will find that they can even be integrated over k-dimensional regions with a sufficiently smooth $(k-1)$-dimensional "boundary" and it is in this context that we will prove a version of Stokes' Theorem relating the integral of a $(k-1)$-form $\boldsymbol{\alpha}$ over this boundary to the integral of $d\boldsymbol{\alpha}$ over the region it bounds.

Much of what we have just said about real-valued forms extends easily to vector-valued forms by simply doing everything (evaluation at tangent vectors, exterior derivative, Hodge star, etc.) componentwise relative to some basis for the vector space (we will show that it all turns out to be independent of the choice of basis). There are a few troublesome items (e.g., wedge products) that we will treat carefully in Chapter 4 and simply illustrate here for the particular vector space of interest in our example i.e., $su(2)$. We take as a basis for $su(2)$ the set $\{T_1, T_2, T_3\}$, where

$$T_1 = -\frac{1}{2}\boldsymbol{i}\sigma_1, \quad T_2 = -\frac{1}{2}\boldsymbol{i}\sigma_2, \quad T_3 = -\frac{1}{2}\boldsymbol{i}\sigma_3,$$

and

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

are the Pauli spin matrices. It is easy to see that $\{T_1, T_2, T_3\}$ is orthonormal with respect to the inner product $\langle A, B \rangle = -2 \operatorname{trace}(AB)$ on $su(2)$. Then any $su(2)$-valued k-form $\boldsymbol{\varphi}$ on $\mathbb{R}^n$ (e.g, $\boldsymbol{\mathcal{A}}$ or $\boldsymbol{\mathcal{F}}$) can be regarded as a matrix of complex k-forms

$$\boldsymbol{\varphi} = \boldsymbol{\varphi}^a T_a = -\frac{1}{2}\boldsymbol{\varphi}^a(\boldsymbol{i}\sigma_a) = -\frac{1}{2}\begin{pmatrix} \boldsymbol{\varphi}^3\boldsymbol{i} & \boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\boldsymbol{i} \\ -\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\boldsymbol{i} & -\boldsymbol{\varphi}^3\boldsymbol{i} \end{pmatrix} \tag{2.5.1}$$

where $\boldsymbol{\varphi}^1$, $\boldsymbol{\varphi}^2$, and $\boldsymbol{\varphi}^3$ are in $\Lambda^k(\mathbb{R}^n)$. Alternatively one can write

$$\boldsymbol{\varphi}^a = \frac{1}{k!}\varphi^a_{i_1\cdots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}, \quad a = 1, \ldots, n$$

and then

$$\begin{aligned} \boldsymbol{\varphi} = \boldsymbol{\varphi}^a T_a &= \left(\frac{1}{k!}\varphi^a_{i_1\cdots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}\right) T_a \\ &= \frac{1}{k!}\left(\varphi^a_{i_1\cdots i_k} T_a\right) dx^{i_1} \wedge \cdots \wedge dx^{i_k} \end{aligned}$$

$$\boldsymbol{\varphi} = \frac{1}{k!}\,\varphi_{i_1\cdots i_k}\, dx^{i_1} \wedge \cdots \wedge dx^{i_k}, \tag{2.5.2}$$

where

$$\begin{aligned} \varphi_{i_1\cdots i_k} &= \varphi^a_{i_1\cdots i_k} T_a \\ &= -\frac{1}{2}\begin{pmatrix} \varphi^3_{i_1\cdots i_k}\boldsymbol{i} & \varphi^2_{i_1\cdots i_k} + \varphi^1_{i_1\cdots i_k}\boldsymbol{i} \\ -\varphi^2_{i_1\cdots i_k} + \varphi^1_{i_1\cdots i_k}\boldsymbol{i} & -\varphi^3_{i_1\cdots i_k}\boldsymbol{i} \end{pmatrix} \end{aligned} \tag{2.5.3}$$

is a smooth map into $su(2)$ for each $i_1 \ldots,\ i_k = 1, \ldots, n$.

We will, in Chapter 4, discuss various natural ways of defining a wedge product for matrix-valued forms such as these. From our point of view at the moment, the most useful such notion can be described as follows: The wedge product of two complex-valued forms is obtained by multiplying the forms as if they were complex numbers, but with real and imaginary parts multiplied by the ordinary wedge product of real-valued forms, i.e.,

$$(\boldsymbol{\varphi}^1 + \boldsymbol{\varphi}^2\boldsymbol{i}) \wedge (\boldsymbol{\psi}^1 + \boldsymbol{\psi}^2\boldsymbol{i}) = (\boldsymbol{\varphi}^1 \wedge \boldsymbol{\psi}^1 - \boldsymbol{\varphi}^2 \wedge \boldsymbol{\psi}^2) + (\boldsymbol{\varphi}^1 \wedge \boldsymbol{\psi}^2 + \boldsymbol{\varphi}^2 \wedge \boldsymbol{\psi}^1)\boldsymbol{i}$$

Now, if $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are both $su(2)$-valued and written as in (2.5.1), then $\boldsymbol{\varphi} \wedge \boldsymbol{\psi}$ is obtained by simply forming their matrix product, with entries multipled by

the complex wedge product described above. We will illustrate the procedure with an example that will also allow us to write out our action functional. We consider a k-form $\boldsymbol{\varphi}$ with values in $su(2)$ and written in the form (2.5.1) and will compute $\boldsymbol{\varphi} \wedge {}^*\boldsymbol{\varphi}$, where ${}^*\boldsymbol{\varphi}$ is the Hodge dual of $\boldsymbol{\varphi}$, computed componentwise (i.e., entrywise). Thus,

$$\boldsymbol{\varphi} \wedge {}^*\boldsymbol{\varphi} = \frac{1}{4}\begin{pmatrix} \boldsymbol{\varphi}^3\,\boldsymbol{i} & \boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\boldsymbol{i} \\ -\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\boldsymbol{i} & -\boldsymbol{\varphi}^3\boldsymbol{i} \end{pmatrix}\begin{pmatrix} {}^*\boldsymbol{\varphi}^3\,\boldsymbol{i} & {}^*\boldsymbol{\varphi}^2 + {}^*\boldsymbol{\varphi}^1\boldsymbol{i} \\ -{}^*\boldsymbol{\varphi}^2 + {}^*\boldsymbol{\varphi}^1\boldsymbol{i} & -{}^*\boldsymbol{\varphi}^3\boldsymbol{i} \end{pmatrix}$$

$$= \frac{1}{4}\begin{pmatrix} -\boldsymbol{\varphi}^3 \wedge {}^*\boldsymbol{\varphi}^3 & (\boldsymbol{\varphi}^3\,\boldsymbol{i}) \wedge ({}^*\boldsymbol{\varphi}^2 + {}^*\boldsymbol{\varphi}^1\,\boldsymbol{i}) \\ \quad +(\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\,\boldsymbol{i}) & \quad -(\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\,\boldsymbol{i}) \\ \quad\quad \wedge(-{}^*\boldsymbol{\varphi}^2 + {}^*\boldsymbol{\varphi}^1\,\boldsymbol{i}) & \quad\quad \wedge({}^*\boldsymbol{\varphi}^3\,\boldsymbol{i}) \\ (-\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\,\boldsymbol{i}) & (-\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\,\boldsymbol{i}) \\ \quad \wedge({}^*\boldsymbol{\varphi}^3\,\boldsymbol{i}) - (\boldsymbol{\varphi}^3\,\boldsymbol{i}) & \quad\quad \wedge({}^*\boldsymbol{\varphi}^2 + {}^*\boldsymbol{\varphi}^1\,\boldsymbol{i}) \\ \quad\quad \wedge(-{}^*\boldsymbol{\varphi}^2 + {}^*\boldsymbol{\varphi}^1\,\boldsymbol{i}) & \quad\quad -\boldsymbol{\varphi}^2 \wedge {}^*\boldsymbol{\varphi}^3 \end{pmatrix}.$$

Each entry is computed in the same way. For example,

$$\begin{aligned} -\boldsymbol{\varphi}^3 \wedge {}^*\boldsymbol{\varphi}^3 + (\boldsymbol{\varphi}^2 + \boldsymbol{\varphi}^1\boldsymbol{i}) \wedge (-{}^*\boldsymbol{\varphi}^2 \wedge {}^*\boldsymbol{\varphi}^1\boldsymbol{i}) \\ = -\|\boldsymbol{\varphi}^3\|^2\mathbf{vol} - \|\boldsymbol{\varphi}^2\|^2\mathbf{vol} - \|\boldsymbol{\varphi}^1\|^2\mathbf{vol} \\ + \ (\langle\boldsymbol{\varphi}^2, \boldsymbol{\varphi}^1\rangle\mathbf{vol} - \langle\boldsymbol{\varphi}^1,\ \boldsymbol{\varphi}^2\rangle\mathbf{vol})\boldsymbol{i} \\ = -(\|\boldsymbol{\varphi}^1\|^2 + \|\boldsymbol{\varphi}^2\|^2 + \|\boldsymbol{\varphi}^3\|^2)\mathbf{vol} \end{aligned}$$

and similarly for the rest. In particular, the $(2,2)$-entry is the same so

$$-2\mathrm{trace}(\boldsymbol{\varphi} \wedge {}^*\boldsymbol{\varphi}) = (\|\boldsymbol{\varphi}^1\|^2 + \|\boldsymbol{\varphi}^2\|^2 + \|\boldsymbol{\varphi}^3\|^2)\mathbf{vol}.$$

We define

$$\|\boldsymbol{\varphi}\|^2 = \|\boldsymbol{\varphi}^1\|^2 + \|\boldsymbol{\varphi}^2\|^2 + \|\boldsymbol{\varphi}^3\|^2$$

so that

$$-2\mathrm{trace}(\boldsymbol{\varphi} \wedge {}^*\boldsymbol{\varphi}) = \|\boldsymbol{\varphi}\|^2\mathbf{vol}. \tag{2.5.4}$$

Remark: More generally, if $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are any two $su(2)$-valued k-forms on $\mathbb{R}^n$ and $\{T_1, T_2, T_3\}$ is *any* orthonormal basis for $su(2)$ and if we write $\boldsymbol{\varphi} = \boldsymbol{\varphi}^a\, T_a$ and $\boldsymbol{\psi} = \boldsymbol{\psi}^a\, T_a$, then we can define $\langle\boldsymbol{\varphi}, \boldsymbol{\psi}\rangle = \langle\boldsymbol{\varphi}^1, \boldsymbol{\psi}^1\rangle + \langle\boldsymbol{\varphi}^2, \boldsymbol{\psi}^2\rangle + \langle\boldsymbol{\varphi}^3, \boldsymbol{\psi}^3\rangle$ and show, as above, that

$$-2\mathrm{trace}(\boldsymbol{\varphi} \wedge {}^*\boldsymbol{\psi}) = \langle\boldsymbol{\varphi}, \boldsymbol{\psi}\rangle\ \mathbf{vol}.$$

With the machinery we have assembled thus far we can write out the **Yang-Mills-Higgs action functional** $A(\boldsymbol{\mathcal{A}}, \phi)$ for our example:

$$\begin{aligned} A(\boldsymbol{\mathcal{A}}, \phi) \quad &= \int_{\mathbb{R}^n} \Big(- \operatorname{trace}(\boldsymbol{\mathcal{F}} \wedge {}^*\boldsymbol{\mathcal{F}}) - \operatorname{trace}(d^{\mathcal{A}}\,\phi \wedge {}^*d^{\mathcal{A}}\phi) \\ &\qquad\qquad + \frac{\lambda}{8}\,{}^*(\|\phi\|^2 - 1)^2\Big) \\ &= \frac{1}{2}\int_{\mathbb{R}^n} \Big(\|\boldsymbol{\mathcal{F}}\|^2 + \|d^{\mathcal{A}}\phi\|^2 \\ &\qquad\qquad + \frac{\lambda}{4}\,(\|\phi\|^2 - 1)^2\Big)\; dx^1 \wedge \cdots \wedge dx^n, \end{aligned} \tag{2.5.5}$$

where $d^{\mathcal{A}}\phi = d\phi + [\boldsymbol{\mathcal{A}}, \phi]$ is the covariant exterior derivative of the Higgs field ϕ. We observe first that $A(\boldsymbol{\mathcal{A}}, \phi)$ is gauge invariant, i.e., that the effect of a gauge transformation $g : \mathbb{R}^n \longrightarrow SU(2)$ ($\boldsymbol{\mathcal{A}} \longrightarrow \boldsymbol{\mathcal{A}}^g$, $\boldsymbol{\mathcal{F}} \longrightarrow \boldsymbol{\mathcal{F}}^g$ and $\phi \longrightarrow \phi^g$) is to leave the integral unchanged. We have already seen that $\boldsymbol{\mathcal{F}}^g = g^{-1}\,\boldsymbol{\mathcal{F}} g$ and $\phi^g = g^{-1}\phi g$ and we show now that

$$d^{\boldsymbol{\mathcal{A}}^g}\phi^g = g^{-1}\Big(d^{\boldsymbol{\mathcal{A}}}\,\phi\Big)g \tag{2.5.6}$$

as well (we will need to use a few simple algebraic properties of d, e.g., the product rule for matrix products, that will be proved in Chapter 4). Indeed, $d^{\boldsymbol{\mathcal{A}}}\,\phi = d\phi + [\boldsymbol{\mathcal{A}}, \phi]$ implies

$$g^{-1}\Big(d^{\,\boldsymbol{\mathcal{A}}}\,\phi\Big)g = g^{-1}\,d\phi g + g^{-1}[\boldsymbol{\mathcal{A}}, \phi]g$$

and

$$\begin{aligned} d^{\,\boldsymbol{\mathcal{A}}^g}\phi^g &= d\phi^g + [\boldsymbol{\mathcal{A}}^g, \phi^g] = d\Big(g^{-1}\phi g\Big) + \Big[g^{-1}\boldsymbol{\mathcal{A}}g + g^{-1}dg, g^{-1}\,\phi g\Big] \\ &= d\Big(g^{-1}\,\phi g\Big) + \Big[g^{-1}\boldsymbol{\mathcal{A}}g, g^{-1}\,\phi g\Big] + \Big[g^{-1}dg, g^{-1}\,\phi g\Big] \\ &= g^{-1}[\boldsymbol{\mathcal{A}}, \phi]g + d\Big(g^{-1}\,\phi g\Big) + \Big[g^{-1}dg, g^{-1}\,\phi g\Big] \\ &= g^{-1}[\boldsymbol{\mathcal{A}}, \phi]g + g^{-1}d(\phi g) + dg^{-1}\,\phi g \\ &\qquad + g^{-1}\,dg\,g^{-1}\phi g - g^{-1}\phi\,g\,g^{-1}dg \\ &= g^{-1}[\boldsymbol{\mathcal{A}}, \phi]g + g^{-1}\,\phi dg + g^{-1}\,d\phi g \\ &\qquad + dg^{-1}\,\phi g + g^{-1}\,dg\,g^{-1}\,\phi g - g^{-1}\,\phi dg \\ &= g^{-1}[\boldsymbol{\mathcal{A}}, \phi]g + g^{-1}\,d\phi g + dg^{-1}\,\phi g + g^{-1}\,dg\,g^{-1}\,\phi g \\ &= g^{-1}[\boldsymbol{\mathcal{A}}, \phi]g + g^{-1}d\phi g = g^{-1}\Big(d^{\,\boldsymbol{\mathcal{A}}}\,\phi\Big)g \end{aligned}$$

because

$$\begin{aligned} g^{-1}g \ = id &\Longrightarrow g^{-1}\,dg + dg^{-1}\,g = 0 \\ &\Longrightarrow g^{-1}\,dg = -dg^{-1}\,g \\ &\Longrightarrow g^{-1}\,dg\,g^{-1}\,\phi g = -dg^{-1}\,\phi g. \end{aligned}$$

Gauge invariance of the action will therefore follow if we can show that $\|g^{-1}\,\boldsymbol{\varphi} g\|^2 = \|\boldsymbol{\varphi}\|^2$ for any $su(2)$-valued form $\boldsymbol{\varphi}$. But if we write $\boldsymbol{\varphi} = \boldsymbol{\varphi}^a\,T_a$, then $g^{-1}\,\boldsymbol{\varphi} g = \boldsymbol{\varphi}^a\,(g^{-1}\,T_a\,g)$ and, since $\langle g^{-1}Ag, g^{-1}Bg\rangle = \langle A, B\rangle$, $\{g^{-1}T_1 g,$ $g^{-1}T_2 g, g^{-1}T_3 g\}$ is also an orthonormal basis for $su(2)$ so this is clear (see the Remark on page 130).

We are interested in finite action, stationary configurations $(\boldsymbol{\mathcal{A}}, \phi)$, i.e., solutions to the Euler-Lagrange equations for the action (2.5.5) for which $A(\boldsymbol{\mathcal{A}}, \phi) < \infty$. As it happens, no such solutions exist when $n > 4$ (see **[JT]**). When $n = 2$ such solutions do exist and they are called **vortices**. These are studied exhaustively in **[JT]**, but we will have no more to say about them. When $n = 4$ any such solution is gauge equivalent to a pure Yang-Mills field $(\lambda = 0, \phi = 0)$ of the type discussed in **[N4]** (we will briefly review this material shortly). Our primary concern is with the case $n = 3$ where finite action, stationary configurations are, for reasons we hope to make clear, called **monopoles**. In fact, we intend to discuss only a special case in which just a vestige of the Georgi-Glashow potential survives. This special case arises in the following way: The requirement that $A(\boldsymbol{\mathcal{A}}, \phi) < \infty$ implies that, as $|x| \longrightarrow \infty$ in $\mathbb{R}^3$,

$$\left\|\boldsymbol{\mathcal{F}}\right\| \longrightarrow 0 \tag{2.5.7}$$

$$\left\|d^{\mathcal{A}}\,\phi\right\| \longrightarrow 0 \tag{2.5.8}$$

and, at least if $\lambda \neq 0$,

$$\left\|\phi\right\| \longrightarrow 1. \tag{2.5.9}$$

Indeed, it is shown in Chapter 4, Sections 10–15, of **[JT]** that each of these limits is achieved uniformly. Now, when $\lambda = 0$ there is no reason to suppose that finite action implies $\|\phi\| \longrightarrow 1$ as $|x| \longrightarrow \infty$. However, **[JT]** also shows that, even in this case, there is some constant $c \geq 0$ such that $\|\phi\| \longrightarrow c$ uniformly as $|x| \longrightarrow \infty$ and that if $c \neq 0$, one can rescale in $\mathbb{R}^3$ to obtain a new configuration $(\boldsymbol{\mathcal{A}}'(x), \phi'(x)) = (c^{-1}\,\boldsymbol{\mathcal{A}}(c^{-1}\,x),\ c^{-1}\,\phi(c^{-1}x))$ which is a finite action stationary point for the action A with $\lambda = 0$ and satisfies $\|\phi'\| \longrightarrow 1$ uniformly as $|x| \longrightarrow \infty$. In effect, one loses nothing, even in the $\lambda = 0$ case, by restricting attention to those configurations for which (2.5.7), (2.5.8) *and* (2.5.9) are satisfied. This, then, is precisely what we intend to.

Remark: Before abandoning the self-interaction term, however, we point out some general features of the action (2.5.5). In particular, we note that there are some obvious absolute minima. Indeed, $A(\boldsymbol{\mathcal{A}}, \phi)$ is obviously zero whenever $\boldsymbol{\mathcal{A}} = 0$ and $\phi = \phi_0$ is a constant in $su(2)$ with $\|\phi_0\| = 1$. Such an absolute minimum is called a *ground state* for the system. The corresponding quantum state of lowest energy is called a *vacuum state* and physicists perform perturbation calculations about such vacuum states. The point here is that such vacuum states are not unique (as long as $\boldsymbol{\mathcal{A}} = 0$, any $\phi_0 \in S^2 \subseteq su(2)$ will give rise to such a state). A specific choice of such a ϕ_0 is said to *break the symmetry* from $SU(2)$ to $U(1)$ and we wish to very briefly explain the terminology (see Section 10.3 of **[Bl]** for more details). A gauge transformation $g : \mathbb{R}^3 \longrightarrow SU(2)$ acts on ϕ by $\phi \longrightarrow \phi^g = g^{-1}\,\phi g$. If the ground state is to be gauge invariant, then we must have $g^{-1}\,\phi_0\,g = \phi_0$ and this occurs only for g in the isotropy subgroup of ϕ_0 in $SU(2)$ under the adjoint action. Now, we claim that this isotropy subgroup is a copy $U(1)$. To see this, identify $su(2)$ with $\mathbb{R}^3$ and $SU(2)\,/\,\pm\,\mathrm{id}$ with $SO(3)$ (page 88). Then the adjoint action of $SU(2)$ on $su(2)$ corresponds to the natural action of $SO(3)$ on $\mathbb{R}^3$ (this is proved, although not stated in these terms in Appendix A to **[N4]**). But this natural action of $SO(3)$ on $\mathbb{R}^3$ (rotation) is transitive on $S^2 \subseteq \mathbb{R}^3$ (page 27) Now, if $H = \{g \in SU(2) : g^{-1}\,\phi_0\,g = \phi_0\}$, then (since $SU(2)$ is compact), $SU(2)\,/\,H \cong S^2$ (Remark on page 27). Thus, $\dim H = 1$. But H is closed in $SU(2)$ so it too is compact and $H \cong U(1)$ (S^1 is the only compact, connected 1-manifold; see Section 5–11 of **[N1]**). The ground states of our $SU(2)$ Yang-Mills-Higgs theory are therefore invariant only under a $U(1)$ subgroup of $SU(2)$. This is an instance of the phenomenon of *spontaneous symmetry breaking* in which a field theory with an exact symmetry group G (e.g., $SU(2)$) gives rise to ground states that are invariant only under some proper subgroup H (e.g., $U(1)$) of G.

With these few remarks behind us we now turn to the case ($\lambda = 0$) of most interest to us. More precisely, we consider the action

$$\begin{aligned} A(\boldsymbol{\mathcal{A}}, \phi) &= \int_{\mathbb{R}^3} \Big(-\operatorname{trace}(\boldsymbol{\mathcal{F}} \wedge {}^*\boldsymbol{\mathcal{F}}) - \operatorname{trace}(d^{\mathcal{A}}\phi \wedge {}^*d^{\mathcal{A}}\phi) \Big) \\ &= \frac{1}{2} \int_{\mathbb{R}^3} \Big(\|\boldsymbol{\mathcal{F}}\|^2 + \|d^{\mathcal{A}}\,\phi\|^2 \Big)\, dx^1 \wedge dx^2 \wedge dx^3 \end{aligned} \tag{2.5.10}$$

and take as our configuration space

$$C = \left\{ (\boldsymbol{\mathcal{A}}, \phi) : A(\boldsymbol{\mathcal{A}}, \phi) < \infty, \lim_{R \longrightarrow \infty} \sup_{|x| \geq R} \big| 1 - \|\phi\| \big| = 0 \right\} \tag{2.5.11}$$

(the second condition reflecting our decision to restrict attention to finite action configurations for which $\|\phi\| \longrightarrow 1$ as $|x| \longrightarrow \infty$ in $\mathbb{R}^3$). Writing out the Euler-Lagrange equations for the action (2.5.11) yields what are called

the **Yang-Mills-Higgs (YMH) equations**

$$\begin{cases} {}^*d^{\mathcal{A}}\,{}^*\boldsymbol{\mathcal{F}} = [d^{\mathcal{A}}\phi, \phi] \\ {}^*d^{\mathcal{A}}\ {}^*d^{\mathcal{A}}\phi = 0 \end{cases} . \tag{2.5.12}$$

The configuration $(\boldsymbol{\mathcal{A}}, \phi)$ must also satisfy the following *Bianchi identities*:

$$\begin{cases} d^{\mathcal{A}}\,\boldsymbol{\mathcal{F}} = 0 \\ d^{\mathcal{A}}\,d^{\mathcal{A}}\phi = [\boldsymbol{\mathcal{F}}, \phi] \end{cases} . \tag{2.5.13}$$

Thus we are looking for solutions to (2.5.12) that live in C. We shall find some interesting ones, but not by studying (2.5.12) directly. As it happens, there is a simpler set of first order equations whose solutions necessarily also satisfy (2.5.12) and, in fact, give the absolute minima of the action (2.5.10).

Remark: This is entirely analogous to the situation encountered in [**N4**], where the (anti-) self-dual equations on $\mathbb{R}^4$ gave the absolute minima of the Yang-Mills action. Shortly we will review that situation and find that there is a closer connection than simple analogy.

The best way to see where these equations come from is as follows: We denote by $\langle\ ,\ \rangle$ the inner product we have defined on $su(2)$-valued forms on $\mathbb{R}^3$ (Remark, page 130). Notice that, on $\mathbb{R}^3$, $\boldsymbol{\mathcal{F}}$ and ${}^*d^{\mathcal{A}}\phi$ are both 2- forms and (since the metric on $\mathbb{R}^3$ is Riemannian), $\|d^{\mathcal{A}}\phi\|^2 = \|{}^*d^{\mathcal{A}}\phi\|^2$. Now notice that

$$\begin{aligned} \|\boldsymbol{\mathcal{F}}\|^2 + \left\|d^{\mathcal{A}}\phi\right\|^2 &= \|\boldsymbol{\mathcal{F}}\|^2 + \left\|{}^*d^{\mathcal{A}}\phi\right\|^2 \\ &= \langle \boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{F}}\rangle + \left\langle {}^*d^{\mathcal{A}}\phi,\ {}^*d^{\mathcal{A}}\phi \right\rangle \\ &= \left\langle \boldsymbol{\mathcal{F}} - {}^*d^{\mathcal{A}}\phi, \boldsymbol{\mathcal{F}} - {}^*d^{\mathcal{A}}\phi \right\rangle + 2\left\langle \boldsymbol{\mathcal{F}}, {}^*d^{\mathcal{A}}\phi \right\rangle \\ &= \left\| \boldsymbol{\mathcal{F}} - {}^*d^{\mathcal{A}}\phi \right\|^2 + 2\left\langle \boldsymbol{\mathcal{F}}, {}^*d^{\mathcal{A}}\phi \right\rangle \end{aligned}$$

and, similarly,

$$\|\boldsymbol{\mathcal{F}}\|^2 + \left\|d^{\mathcal{A}}\phi\right\|^2 = \left\|\boldsymbol{\mathcal{F}} + {}^*d^{\mathcal{A}}\phi\right\|^2 - 2\left\langle \boldsymbol{\mathcal{F}}, {}^*d^{\mathcal{A}}\phi \right\rangle.$$

It follows that $(\boldsymbol{\mathcal{A}}, \phi)$ achieves an absolute minimum when

$$\boldsymbol{\mathcal{F}} = \pm {}^*d^{\mathcal{A}}\phi. \tag{2.5.14}$$

These are the (**Bogomolny**) **monopole equations** and any configuration $(\boldsymbol{\mathcal{A}}, \phi)$ which satisfies them also satisfies the YMH equations (2.5.12) (either observe that an absolute minimum for $A(\boldsymbol{\mathcal{A}}, \phi)$ is necessarily a stationary value, or simply substitute $\boldsymbol{\mathcal{F}} = \pm {}^*d^{\mathcal{A}}\phi$ into (2.5.12) and use the Bianchi identities (2.5.13)). For the record we write out these equations in standard coordinates on $\mathbb{R}^3$, using a formula for the Hodge dual of a 1-form on $\mathbb{R}^3$

that we will prove in Chapter 4. With $\boldsymbol{\mathcal{A}} = \mathcal{A}_k\, dx^k$ and $\boldsymbol{\mathcal{F}} = \frac{1}{2}\,\mathcal{F}_{ij}\, dx^i \wedge dx^j$, (2.5.14) becomes

$$\mathcal{F}_{ij} = \pm \sum_{k=1}^{3} \epsilon_{ijk}(\partial_k\, \phi + [\mathcal{A}_k, \phi]), \quad i, j = 1, 2, 3, \tag{2.5.15}$$

where ϵ_{ijk} is the Levi-Civita symbol (anti-symmetric in $i\, j\, k$ and $\epsilon_{123} = 1$).

We will have something to say shortly about why these are called "monopole" equations. First we wish to give some indication of how the example before us now actually arises quite naturally out of the pure Yang-Mills theory on $\mathbb{R}^4$ discussed in **[N4]**. This is the $n = 4$ case of the model we have constructed when, in the action (2.5.5), $\lambda = 0$ and $\phi = 0$. In this special case the action (which now depends only on $\boldsymbol{\mathcal{A}}$) is called the **Yang-Mills-action** and written

$$\begin{aligned} \mathcal{YM}(\boldsymbol{\mathcal{A}}) &= -\int_{\mathbb{R}^4} \operatorname{trace}(\boldsymbol{\mathcal{F}} \wedge {}^*\boldsymbol{\mathcal{F}}) \\ &= \frac{1}{2}\int_{\mathbb{R}^4} \|\boldsymbol{\mathcal{F}}\|^2\, dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4. \end{aligned} \tag{2.5.16}$$

The Euler-Lagrange equations now become the **Yang-Mills equations**

$$d^{\mathcal{A}}\, {}^*\boldsymbol{\mathcal{F}} = 0, \tag{2.5.17}$$

while the Bianchi identities become

$$d^{\mathcal{A}}\, \boldsymbol{\mathcal{F}} = 0. \tag{2.5.18}$$

Now notice that if we have a potential $\boldsymbol{\mathcal{A}}$ for which the field strength $\boldsymbol{\mathcal{F}}$ is **self-dual** (**SD**)

$${}^*\boldsymbol{\mathcal{F}} = \boldsymbol{\mathcal{F}}, \tag{2.5.19}$$

or **anti-self-dual** (**ASD**)

$${}^*\boldsymbol{\mathcal{F}} = -\boldsymbol{\mathcal{F}}, \tag{2.5.20}$$

then the Bianchi identity (2.5.18) implies that $\boldsymbol{\mathcal{F}}$ necessarily satisfies the Yang-Mills equations (2.5.17). Such potentials $\boldsymbol{\mathcal{A}}$ are called $SU(2)$ **instantons** on $\mathbb{R}^4$ and, in **[N4]**, a number of examples (called the **BPST instantons**) were described. Written in quaternionic notation (i.e., identifying $\mathbb{R}^4$ with $\mathbb{H}$ and $su(2)$ with the Lie algebra $\operatorname{Im}\mathbb{H}$ of pure imaginary quaternions) these can be written

$$\boldsymbol{\mathcal{A}}_{\lambda,n}(q) = \operatorname{Im}\left(\frac{\bar{q} - \bar{n}}{\lambda^2 + |q - n|^2} dq\right), \tag{2.5.21}$$

where $\lambda > 0$ and $n \in \mathbb{H}$ are parameters called the **scale** and **center** of the instanton, respectively. The corresponding field strengths are

$$\boldsymbol{\mathcal{F}}_{\lambda,n}(q) = \frac{\lambda^2}{(\lambda^2 + |q - n|^2)^2}\, d\bar{q} \wedge dq. \tag{2.5.22}$$

These are ASD and satisfy

$$\frac{1}{2}\,\|\boldsymbol{\mathcal{F}}_{\lambda,n}(q)\|^2 = \frac{48\lambda^2}{(\lambda^2+|q-n|^2)^4} \tag{2.5.23}$$

and

$$\mathcal{YM}(\boldsymbol{\mathcal{A}}_{\lambda,n}) = \int_{\mathbb{R}^4} \frac{48\,\lambda^2}{(\lambda^2+|q-n|^2)^4}\,dq^0\,dq^1\,dq^2\,dq^3 = 8\pi^2. \tag{2.5.24}$$

Notice that all of these potentials have the same Yang-Mills action (total field strength). For a fixed center n, $\|\boldsymbol{\mathcal{F}}_{\lambda,n}(q)\|^2$ has a maximum value of $96/\lambda^2$ at $q=n$ and, as $\lambda \longrightarrow 0$, this maximum value approaches infinity in such a way that the integrals over $\mathbb{R}^4$ remain constant at $8\pi^2$.

The fact that all of these BPST potentials $\boldsymbol{\mathcal{A}}_{\lambda,n}$ have the same Yang-Mills action has a much more profound significance than may be apparent at first glance. It was shown in **[N4]** (see also Examples 3 and 4, pages 40–42, of Chapter 1) that each $\boldsymbol{\mathcal{A}}_{\lambda,n}$ is the pullback to $\mathbb{R}^4$ via stereographic projection of a connection $\boldsymbol{\omega}_{\lambda,n}$ on the Hopf bundle $SU(2) \hookrightarrow S^7 \longrightarrow \mathbb{HP}^1$. Thinking of $\mathbb{HP}^1 \cong S^4$ as the one-point compactification of $\mathbb{R}^4$, one can reverse one's point of view here and say that, by virtue of the asymptotic behavior implicit in the fact that $\mathcal{YM}(\boldsymbol{\mathcal{A}}_{\lambda,n}) < \infty$, each $\boldsymbol{\mathcal{A}}_{\lambda,n}$ "extends to the point at infinity." Indeed, Uhlenbeck's Removable Singularities Theorem asserts that, if $\boldsymbol{\mathcal{A}}$ is an $SU(2)$ gauge potential on $\mathbb{R}^4$ with $\mathcal{YM}(\boldsymbol{\mathcal{A}}) < \infty$, then $\boldsymbol{\mathcal{A}}$ always "extends to the point at infinity" in the sense that there exists a unique $SU(2)$-bundle over S^4 and a connection $\boldsymbol{\omega}$ on it such that $\boldsymbol{\mathcal{A}} = (s \circ \varphi^{-1})^*\boldsymbol{\omega}$, where s is a cross-section of the bundle defined on the complement of some point in S^4 and φ is a stereographic projection to $\mathbb{R}^4$. The specific bundle to which the potential $\boldsymbol{\mathcal{A}}$ "extends" is, moreover, determined by the value of $\mathcal{YM}(\boldsymbol{\mathcal{A}})$. It is essential to understand precisely what is being asserted here so, before returning to the Bogomolny monopole equations, we will elaborate.

According to the Classification Theorem (page 34), the set of equivalence classes of principal $SU(2)$-bundles over S^4 is in one-to-one correspondence with the elements of the homotopy group $\pi_3(SU(2)) \cong \pi_3(S^3) \cong \mathbb{Z}$. There are various ways of associating with each $SU(2)$-bundle over S^4 an integer that characterizes it up to equivalence, but from our point of view the most useful of these arises in the theory of characteristic classes. We have already had one brief encounter with characteristic classes (the first Chern class, in Section 2.2) and will return to the general theory in Chapter 6. For the present we will content ourselves with a brief, informal description of those aspects of the subject relevant to characterizing a principal $SU(2)$-bundle

$$SU(2) \hookrightarrow P \xrightarrow{\mathcal{P}} S^4$$

over S^4 up to equivalence.

Choose any connection $\boldsymbol{\omega}$ on the bundle and let $\boldsymbol{\Omega}$ denote its curvature (we will prove in Chapter 3 that connections exist on any smooth principal

bundle). The local field strengths $\tilde{\boldsymbol{\mathcal{F}}} = s^* \boldsymbol{\omega}$ for various cross-sections generally do not agree on the intersections of their domains and so do not piece together into a globally defined 2-form on all of S^4. Indeed, we know that if s^g is another cross-section, then $\tilde{\boldsymbol{\mathcal{F}}}^g = g^{-1}\tilde{\boldsymbol{\mathcal{F}}}\, g$ on the intersection. However, certain algebraic combinations of the local field strengths can be found which *do* agree on the intersections and so do piece together into globally defined forms on S^4. Essentially, all that is required is a symmetric, multilinear function on $su(2)$ that is ad-invariant, i.e., takes the same values at A and $g^{-1}Ag$ for all $A \in su(2)$ and $g \in SU(2)$. The trace is an obvious choice and this gave rise, in Section 2.2, to the first Chern class which, for $U(1)$-bundles over spacetime, is the cohomology class of the electromagnetic field strength. For $SU(2)$-bundles, however, the local field strengths take values in $su(2)$ where everything has trace zero so this will not get us very far. The trace of the product would be the next likely candidate and this gets us very far indeed. We define a 4-form on all of S^4 by decreeing that, relative to any local cross-section, it is given by

$$\frac{1}{8\pi^2}\text{trace}(\tilde{\boldsymbol{\mathcal{F}}} \wedge \tilde{\boldsymbol{\mathcal{F}}})$$

(the $\frac{1}{8\pi^2}$ is a normalizing constant whose purpose we will describe shortly). Now, *a priori* this is a complex 4-form on S^4, but it can be shown to be real-valued (because $\tilde{\boldsymbol{\mathcal{F}}}$ is skew-Hermitian). Being a 4-form on the 4-dimensional manifold S^4, it is also closed (i.e., has exterior derivative zero) and therefore determines a de Rham cohomology class

$$c_2(P) = \frac{1}{8\pi^2}\left[\text{trace}(\tilde{\boldsymbol{\mathcal{F}}} \wedge \tilde{\boldsymbol{\mathcal{F}}})\right] \in H^4_{\text{de R}}(S^4). \tag{2.5.25}$$

Remarkably, this cohomology class does not depend on the initial choice of the connection $\boldsymbol{\omega}$ from which it arose; it is a characteristic class for the bundle. $c_2(P)$ is called the **second Chern class** of the bundle. Any representative of the class $c_2(P)$ is a 4-form on the compact 4-manifold S^4 and so can be integrated over S^4. It will follow from Stokes' Theorem that two forms in the same cohomology class have the same integral so, in effect, we may integrate $c_2(P)$ over S^4. The result is written

$$c_2(P)[S^4] = \int_{S^4} c_2(P) = \frac{1}{8\pi^2}\int_{S^4} \text{trace}(\tilde{\boldsymbol{\mathcal{F}}} \wedge \tilde{\boldsymbol{\mathcal{F}}}) \tag{2.5.26}$$

and called the **second Chern number** of the bundle (physicists call $-c_2(P)[S^4]$ the **topological charge** or **instanton number** of the bundle). The $\frac{1}{8\pi^2}$ ensures that $c_2(P)[S^4]$ is an *integer* (not obvious, but true) and it is this integer that labels the equivalence classes of $SU(2)$-bundles over S^4. More precisely, two principal $SU(2)$-bundles over S^4 are equivalent if and only if their second Chern numbers are equal.

To relate this to our gauge potentials on $\mathbb{R}^4$ recall that there is a stereographic projection φ, defined at all but one point of S^4, that is an orientation preserving (conformal) diffeomorphism onto $\mathbb{R}^4$. The integrals we define will be invariant under such maps and unaffected by the omission of one point so $c_2(P)[S^4]$ can be computed by integrating over $\mathbb{R}^4$ the pullback of $\frac{1}{8\pi^2}$ trace$(\tilde{\boldsymbol{\mathcal{F}}} \wedge \tilde{\boldsymbol{\mathcal{F}}})$ by φ. Pullback commutes with trace and the wedge product and $(\varphi^{-1})^* \tilde{\boldsymbol{\mathcal{F}}} = (\varphi^{-1})^* (s^* \Omega) = (s \circ \varphi^{-1})^* \boldsymbol{\Omega} = \boldsymbol{\mathcal{F}}$ is a gauge potential on $\mathbb{R}^4$. Thus,

$$c_2(P)[S^4] = \frac{1}{8\pi^2} \int_{\mathbb{R}^4} \text{trace}(\boldsymbol{\mathcal{F}} \wedge \boldsymbol{\mathcal{F}}). \tag{2.5.27}$$

Now here's the good part: If $\boldsymbol{\mathcal{A}}$ is an ASD potential on $\mathbb{R}^4$, then ${}^*\boldsymbol{\mathcal{F}} = -\boldsymbol{\mathcal{F}}$ so

$$\begin{aligned} \mathcal{YM}(\boldsymbol{\mathcal{A}}) &= \int_{\mathbb{R}^4} -\text{trace}(\boldsymbol{\mathcal{F}} \wedge {}^*\boldsymbol{\mathcal{F}}) \\ &= \int_{\mathbb{R}^4} \text{trace}(\boldsymbol{\mathcal{F}} \wedge \boldsymbol{\mathcal{F}}) \qquad (ASD). \end{aligned} \tag{2.5.28}$$

If this is finite, the Removable Singularities Theorem assures us that $\boldsymbol{\mathcal{A}}$ extends to some principal $SU(2)$-bundle over S^4. The Chern number of this bundle can be computed from any connection on it so we might as well use the extension of $\boldsymbol{\mathcal{A}}$. Then, comparing (2.5.27) and (2.5.28) gives

$$c_2(P)[S^4] = \frac{1}{8\pi^2} \mathcal{YM}(\boldsymbol{\mathcal{A}}) \qquad (ASD). \tag{2.5.29}$$

The Yang-Mills action of $\boldsymbol{\mathcal{A}}$ (i.e., its total field strength) is directly encoded in the topology of the bundle to which $\boldsymbol{\mathcal{A}}$ extends as its Chern number. Since the Yang-Mills action is determined by the asymptotic behavior of the field strength $\boldsymbol{\mathcal{F}}$ as $|x| \longrightarrow \infty$ in $\mathbb{R}^4$, we find that it is this asymptotic behavior that determines the bundle to which $\boldsymbol{\mathcal{A}}$ extends. This phenomenon of boundary conditions on physical fields manifesting themselves as topology will be a recurrent theme here. Notice, in particular, that, since the Chern number of an $SU(2)$-bundle over S^4 is an integer, the possible asymptotic boundary conditions for finite action, ASD potential on $\mathbb{R}^4$ fall into countably many, discrete "topological types."

Remark: We have already seen that all of the BPST potentials $\boldsymbol{\mathcal{A}}_{\lambda,n}$ are ASD and have Yang-Mills action $8\,\pi^2$. They must, of course, have the same Yang-Mills action since they all extend to the same bundle, i.e., the Hopf bundle (which we now see has Chern number 1).

In order to establish contact with the Bogomolny monopole equations we consider an arbitrary potential $\hat{\boldsymbol{\mathcal{A}}} = \hat{\mathcal{A}}_\alpha \, dx^\alpha$ on $\mathbb{R}^4$ with field strength $\hat{\boldsymbol{\mathcal{F}}} = \frac{1}{2} \hat{\mathcal{F}}_{\alpha\beta} \, dx^\alpha \wedge dx^\beta$ and use (6.4.4) of **[N4]** to write the components of the Hodge

dual ${}^*\hat{\boldsymbol{\mathcal{F}}}$ as

$$ {}^*\hat{\mathcal{F}}_{\alpha\beta} = \frac{1}{2}\sum_{\gamma,\delta=1}^{4} \epsilon_{\alpha\beta\gamma\delta}\,\hat{\mathcal{F}}_{\alpha\beta}, \quad \alpha,\beta = 1,2,3,4, $$

where $\epsilon_{\alpha\beta\gamma\delta}$ is the Levi-Civita symbol (totally anti-symmetric in $\alpha\,\beta\,\gamma\,\delta$ with $\epsilon_{1234} = 1$). Then the (anti-) self-duality equations take the form

$$ \hat{\mathcal{F}}_{\alpha\beta} = \mp\frac{1}{2}\sum_{\gamma,\delta=1}^{4} \epsilon_{\alpha\beta\gamma\delta}\,\hat{\mathcal{F}}_{\alpha\beta}, \quad \alpha,\beta = 1,2,3,4. \tag{2.5.30} $$

Using i, j and k for indices taking the values 1, 2 and 3 one finds (by just writing them out) that all of these equations are contained in the following:

$$ \hat{\mathcal{F}}_{ij} = \pm\sum_{k=1}^{3} \epsilon_{ijk}\hat{\mathcal{F}}_{k4}, \quad i,j = 1,2,3, \tag{2.5.31} $$

where ϵ_{ijk} is totally anti-symmetric in ijk and $\epsilon_{123} = 1$. For example, taking $\alpha = 3$ and $\beta = 4$ in (2.5.30) gives

$$ \begin{aligned} \hat{\mathcal{F}}_{34} &= \mp\frac{1}{2}\sum_{\gamma,\delta=1}^{4} \epsilon_{34\gamma\delta}\hat{\mathcal{F}}_{\gamma\delta} = \mp\frac{1}{2}\left[\epsilon_{3412}\hat{\mathcal{F}}_{12} + \epsilon_{3421}\hat{\mathcal{F}}_{21}\right] \\ &= \mp\frac{1}{2}\left[2\epsilon_{3412}\hat{\mathcal{F}}_{12}\right] = \mp\,\epsilon_{3412}\hat{\mathcal{F}}_{12} = \pm\,\epsilon_{4123}\hat{\mathcal{F}}_{12} \\ &= \pm\,\epsilon_{123}\hat{\mathcal{F}}_{12} \end{aligned} $$

which is equivalent to

$$ \hat{\mathcal{F}}_{12} = \pm\,\epsilon_{123}\hat{\mathcal{F}}_{34} $$

and this is (2.5.31) with $i = 1$ and $j = 2$. Similarly, taking $\alpha = 1$ and $\beta = 2$ in (2.5.30) gives

$$ \hat{\mathcal{F}}_{12} = \mp\,\epsilon_{1234}\hat{\mathcal{F}}_{34} = \pm\,\epsilon_{123}\hat{\mathcal{F}}_{34} $$

which is also (2.5.31) with $i = 1$ and $j = 2$.

Thus, the (anti-) self-duality equations on $\mathbb{R}^4$ take the form (2.5.31). Finite action solutions to these equations are the instantons discussed earlier. We now wish to seek solutions to (2.5.31) that are *static*, i.e., for which the $\hat{\mathcal{A}}_\alpha$ are independent of x^4. Naturally, no such solution can have finite action on $\mathbb{R}^4$ (unless it is zero) and so cannot be an instanton. For $\hat{\mathcal{A}}_\alpha$'s that are independent

of x^4, (2.5.31) becomes

$$\begin{aligned}\hat{\mathcal{F}}_{ij} &= \pm \sum_{k=1}^{3} \epsilon_{ijk} \hat{\mathcal{F}}_{k4} \\ &= \pm \sum_{k=1}^{3} \epsilon_{ijk} \left(\partial_k \hat{\mathcal{A}}_4 - \partial_4 \hat{A}_k + \left[\hat{\mathcal{A}}_k, \hat{\mathcal{A}}_4 \right] \right) \\ \hat{\mathcal{F}}_{ij} &= \pm \sum_{k=1}^{3} \epsilon_{ijk} \left(\partial_k \hat{\mathcal{A}}_4 + \left[\hat{\mathcal{A}}_k, \hat{\mathcal{A}}_4 \right] \right), \quad i,j = 1,2,3 \end{aligned} \tag{2.5.32}$$

(you may wish to glance back at (2.5.15) if you're wondering where all of this is going).

Now we "reduce to $\mathbb{R}^3$" as follows: Fix some value x_0^4 of x^4 and consider the submanifold $\mathbb{R}^3 \times \{x_0^4\}$ of $\mathbb{R}^4$ (which we henceforth identify with $\mathbb{R}^3$). Restrict our trivial $SU(2)$-bundle over $\mathbb{R}^4$ to this $\mathbb{R}^3$ and obtain a trivial $SU(2)$-bundle over $\mathbb{R}^3$. Let $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ be the restrictions to $\mathbb{R}^3$ of $\hat{\mathcal{A}}_1, \hat{\mathcal{A}}_2$ and $\hat{\mathcal{A}}_3$ (all of which are assumed independent of x^4). Then

$$\boldsymbol{\mathcal{A}} = \mathcal{A}_1 dx^1 + \mathcal{A}_2 dx^2 + \mathcal{A}_3 dx^3 = \mathcal{A}_i dx^i$$

is a gauge potential on $\mathbb{R}^3$ and the corresponding field strength $\boldsymbol{\mathcal{F}} = \frac{1}{2}\mathcal{F}_{ij} dx^i \wedge dx^j$ has components $\mathcal{F}_{ij} = \partial_i \mathcal{A}_j - \partial_j \mathcal{A}_i + [\mathcal{A}_i, \mathcal{A}_j]$ that are just the restrictions to $\mathbb{R}^3$ of the $\hat{\mathcal{F}}_{ij}$, $i,j = 1,2,3$. Note that $\hat{\mathcal{A}}_4$ does not enter into either the potential or the field strength on $\mathbb{R}^3$. However, if we define $\phi : \mathbb{R}^3 \longrightarrow su(2)$ by

$$\phi = \hat{\mathcal{A}}_4 \,|\, \mathbb{R}^3,$$

then the time independent, (anti-) self-duality equations (2.5.32) on $\mathbb{R}^4$ become (when restricted to $\mathbb{R}^3$) the Bogomolny monopole equations

$$\mathcal{F}_{ij} = \pm \sum_{k=1}^{3} \epsilon_{ijk} (\partial_k \phi + [\mathcal{A}_k,\ \phi]), \quad i,j = 1,2,3. \tag{2.5.33}$$

Thus, an $SU(2)$ Bogomolny monopole on $\mathbb{R}^3$ is essentially just a static, (anti-) self-dual gauge potential on $\mathbb{R}^4$. We intend to describe the geometry and topology of these monopoles in more detail, but first we will exhibit a concrete example that will, in some sense, bring us full circle. We began (in Chapter 0, [**N4**]) by thinking about Dirac's magnetic monopoles and finding that they were most naturally modeled by connections on $U(1)$-bundles over S^2. In particular, the monopole of lowest strength was identified with the natural connection on the complex Hopf bundle. Seeking generalizations we looked at the natural connection on the quaternionic Hopf bundle and found lurking there a BPST instanton. Now we find that a "static instanton" is to be identified with a particular type of Yang-Mills-Higgs field on $\mathbb{R}^3$ which we

have (for reasons that no doubt remain obscure) called a monopole. To understand the terminology and to see Dirac monopoles in an entirely new light we will describe now the **t'Hooft-Polyakov-Prasad-Sommerfield monopole,** which is an exact solution to the equations (2.5.33). For this we will need some notation. First, in $\mathbb{R}^3$ we will write

$$\begin{aligned} x &= (x^1, x^2, x^3) \\ r &= |x| = \sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2} \\ n^i &= x^i/r, \;\; i = 1, 2, 3 \quad (r \neq 0) \\ d\vec{x} &= (dx^1, dx^2, dx^3) \end{aligned}$$

and, in $su(2)$,

$$\begin{aligned} T_a &= -\frac{1}{2}\, \boldsymbol{i}\, \sigma_a, \quad a = 1, 2, 3 \\ \vec{T} &= (T_1, T_2, T_3). \end{aligned}$$

In addition we set

$$\begin{aligned} \vec{n} \cdot \vec{T} &= n^1\, T_1 + n^2\, T_2 + n^3\, T_3 = n^a\, T_a \\ \vec{n} \times \vec{T} &= (n^2\, T_3 - n^3\, T_2, n^3\, T_1 - n^1\, T_3, n^1\, T_2 - n^2\, T_1) \end{aligned}$$

and

$$\begin{aligned} \left(\vec{n} \times \vec{T}\right) \cdot d\vec{x} = (n^2\, T_3 - n^3\, T_2)\, dx^1 + (n^3\, T_1 - n^1\, T_3)\, dx^2 \\ + (n^1\, T_2 - n^2\, T_1)\, dx^3. \end{aligned}$$

Our objective is to find $(\boldsymbol{\mathcal{A}}(x), \phi(x))$ which satisfies (2.5.33) and has the required asymptotic behavior ($\|\phi\| \longrightarrow 1$ as $r \longrightarrow \infty$).

Although the monopole equations (2.5.33) are substantially less complicated that the full Yang-Mills-Higgs equations, they are still far beyond the means of elementary techniques. To reduce the level of difficulty a bit more requires a guess (physicists prefer the term *Ansatz*) as to the form one might expect for a solution. Here's the one that worked for t'Hooft and Polyakov (some rationale for the *Ansatz* is discussed in Section 4.2 of [**GO**]): We will seek functions $f(r)$ and $h(r)$ satisfying

$$\frac{f(r)}{r} \longrightarrow 1 \quad \text{as} \quad r \longrightarrow \infty \quad \text{and} \quad f(0) = 0, \tag{2.5.34}$$

and

$$h\,(r) \longrightarrow 0 \quad \text{as} \quad r \longrightarrow \infty \quad \text{and} \quad h(0) = 1, \tag{2.5.35}$$

such that

$$\phi\,(x) = \frac{f(r)}{r} \left(\vec{n} \cdot \vec{T}\right) = \frac{f(r)}{r^2}\, x^a\, T_a \tag{2.5.36}$$

and

$$\begin{aligned} \boldsymbol{\mathcal{A}}(x) &= \frac{1-h(r)}{r}\left(\vec{n}\times\vec{T}\right)\cdot d\vec{x} \\ &= \frac{1-h(r)}{r^2}\sum_{a=1}^{3}\epsilon_{aij}\,x^j\,dx^i\,T_a \end{aligned} \tag{2.5.37}$$

give a solution $(\boldsymbol{\mathcal{A}},\phi)$ to (2.5.33).

Remark: The conditions $f(0)=0$ and $h(0)=1$ are intended to give $\phi(x)$ and $\boldsymbol{\mathcal{A}}(x)$ a chance of being smooth at the origin.

Substituting (2.5.36) and (2.5.37) into (2.5.33) (with the minus sign) gives the following system of ordinary differential equations for $f(r)$ and $h(r)$:

$$\begin{cases} r\,\dfrac{dh}{dr} = -hf \\ r\,\dfrac{df}{dr} = f-(h^2-1). \end{cases} \tag{2.5.38}$$

Now make the change of variable

$$-f = 1+rF, \quad h = rH$$

to obtain the new system

$$\begin{cases} \dfrac{dH}{dr} = HF \\ \dfrac{dF}{dr} = H^2 \end{cases}$$

for F and H. Observe that it follows from these equations that F^2-H^2 is constant since

$$\frac{d}{dr}\left(F^2-H^2\right) = 2F\,\frac{dF}{dr} - 2H\,\frac{dH}{dr} = 2FH^2 - 2H(HF) = 0.$$

But since, for $r>0$, $F^2-H^2 = \dfrac{(1+f)^2}{r^2} - \dfrac{h^2}{r^2} = \dfrac{1}{r^2} + \dfrac{2}{r}\left(\dfrac{f}{r}\right) + \left(\dfrac{f}{r}\right)^2 - \dfrac{1}{r^2}\,h^2$ and this is to approach 1 as $r\longrightarrow\infty$, we must have

$$F^2 - H^2 = 1.$$

To get solutions f and h satisfying $f(0)=0$ and $h(0)=1$ we take for F and H satisfying $F^2-H^2=1$ the following:

$$F(r) = -\coth r \quad \text{and} \quad H(r) = \operatorname{csch} r.$$

Then

$$f(r) = r\coth r - 1 \quad \text{and} \quad h(r) = r\operatorname{csch} r$$

and one can verify directly that these are smooth (even at $r=0$, where they take the required boundary values) and satisfy (2.5.38). Thus, our field

configuration $(\boldsymbol{\mathcal{A}}, \phi)$ is given by

$$\phi(x) = \left(\coth r - \frac{1}{r}\right) \vec{n} \cdot \vec{T} = \frac{1}{r}\left(\coth r - \frac{1}{r}\right) x^a T_a \tag{2.5.39}$$

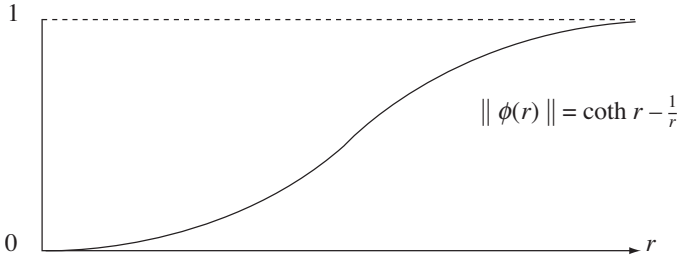
$$\begin{aligned} \boldsymbol{\mathcal{A}}(x) &= \left(\frac{1}{r} - \operatorname{csch} r\right)\left(\vec{n} \times \vec{T}\right) \cdot d\vec{x} \\ &= \frac{1}{r}\left(\frac{1}{r} - \operatorname{csch} r\right) \sum_{a=1}^{3} \epsilon_{aij}\, x^j\, dx^i\, T_a \end{aligned} \tag{2.5.40}$$

Remark: The same calculation for (2.5.33) with the plus sign gives the same result except that the sign of ϕ is changed.

The exact solution $(\boldsymbol{\mathcal{A}}, \phi)$ given by (2.5.39) and (2.5.40) is remarkable for a number of reasons (quite aside from the fact that it *is* an exact solution which is more than one generally has a right to expect). The form of the thing itself is extraordinary for its "mixing" of the spatial and internal directions. For example, (2.5.39) describes a Higgs field which, in the x^a-direction in $\mathbb{R}^3$, $a = 1, 2, 3$, has only an internal T^a-component. It is, in some strange way, "radial" (Polyakov called it a "hedgehog" solution). Another feature, and one that will be particularly significant quite soon, is that, despite appearances, the component functions of ϕ and $\boldsymbol{\mathcal{A}}$ are smooth (in fact, real analytic), even at the origin. For example,

$$\begin{aligned} \coth r - \frac{1}{r} &= \frac{1}{r}\left(\frac{r\cosh r}{\sinh r} - 1\right) = \frac{1}{r}\left(\frac{r + \frac{1}{2!}r^3 + \frac{1}{4!}r^5 + \cdots}{r + \frac{1}{3!}r^3 + \frac{1}{5!}r^5 + \cdots} - 1\right) \\ &= \frac{1}{r}\left(\frac{1 + \frac{1}{2!}r^2 + \frac{1}{4!}r^4 + \cdots}{1 + \frac{1}{3!}r^2 + \frac{1}{5!}r^4 + \cdots} - 1\right) \\ &= \frac{1}{r}\left(1 + \left(\frac{1}{2!} - \frac{1}{3!}\right)r^2 + \cdots - 1\right) \end{aligned}$$

(by long division) and this is indeed analytic at $r = 0$. Computing the derivative of $\coth r - \frac{1}{r}$ one finds that it is positive. Moreover, $\lim_{r \longrightarrow \infty}(\coth r - \frac{1}{r}) = 1$ so one obtains the following picture of $\|\phi\|$:



The configuration $(\boldsymbol{\mathcal{A}}, \phi)$ is a globally defined, smooth object on all of $\mathbb{R}^3$.

But how did configurations such as this come to be called "monopoles"? This is not entirely clear from the "hedgehog" form in which we currently have the solution written, but will become clear after an appropriate (local) gauge transformation. We intend to define a gauge transformation on an open subset of $\mathbb{R}^3$ which, at each point in space, rotates the Higgs field (in the internal space $su(2)$ at that point) from its "radial" direction to a direction parallel to the internal T_3-axis. One should keep in mind that a gauge transformation is assumed to change only the appearance, not the physics of field configurations. Our gauge transformation will be a map g from an open subset of $\mathbb{R}^3$ into $SU(2)$ and its effect on ϕ and $\boldsymbol{\mathcal{A}}$ will be, as usual,

$$\begin{aligned} \phi &\longrightarrow \phi^g = g^{-1}\phi g \\ \boldsymbol{\mathcal{A}} &\longrightarrow \boldsymbol{\mathcal{A}}^g = g^{-1}\,\boldsymbol{\mathcal{A}} g + g^{-1}dg. \end{aligned}$$

Specifically, in terms of spherical coordinates (r, φ, θ) on $\mathbb{R}^3$, we let

$$g(x) = \begin{pmatrix} \cos\dfrac{\varphi}{2} & -e^{-\boldsymbol{i}\theta}\sin\dfrac{\varphi}{2} \\ e^{\boldsymbol{i}\theta}\sin\dfrac{\varphi}{2} & \cos\dfrac{\varphi}{2} \end{pmatrix}. \tag{2.5.41}$$

To compute ϕ^g we proceed as follows:

$$\begin{aligned} \phi^g = g^{-1}\phi g &= g^{-1}\left(\frac{f(r)}{r^2}x^a T_a\right) g = \frac{f(r)}{r^2}\left(g^{-1}\left(x^a T_a\right) g\right) \\ &= \frac{f(r)}{r^2}x^a\left(g^{-1}T_a g\right) = -\frac{1}{2}\boldsymbol{i}\frac{f(r)}{r^2}x^a\left(g^{-1}\sigma_a g\right). \end{aligned}$$

Now, notice that, for any $g = \left(\begin{smallmatrix} \alpha & \beta \\ -\bar{\beta} & \alpha \end{smallmatrix}\right)$, $\alpha \in \mathbb{R}$, in $SU(2)$, $g^{-1} = \left(\begin{smallmatrix} \alpha & -\beta \\ \bar{\beta} & \alpha \end{smallmatrix}\right)$ and

$$g^{-1}\sigma_1 g = g^{-1}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} g = \begin{pmatrix} -\alpha\bar{\beta} - \alpha\beta & \alpha^2 - \beta^2 \\ -\bar{\beta}^2 + \alpha^2 & \alpha\bar{\beta} + \alpha\beta \end{pmatrix}$$

$$g^{-1}\sigma_2 g = \boldsymbol{i}g^{-1}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} g = \boldsymbol{i}\begin{pmatrix} \alpha\bar{\beta} - \alpha\beta & -\alpha^2 - \beta^2 \\ \bar{\beta}^2 + \alpha^2 & -\alpha\bar{\beta} + \alpha\beta \end{pmatrix}$$

$$g^{-1}\sigma_3 g = g^{-1}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} g = \begin{pmatrix} \alpha^2 - \beta\bar{\beta} & 2\alpha\beta \\ 2\alpha\bar{\beta} & \beta\bar{\beta} - \alpha^2 \end{pmatrix}$$

Now, with $g = \left(\begin{smallmatrix} \alpha & \beta \\ -\bar{\beta} & \alpha \end{smallmatrix}\right) = \begin{pmatrix} \cos\frac{\varphi}{2} & -e^{-\boldsymbol{i}\theta}\sin\frac{\varphi}{2} \\ e^{\boldsymbol{i}\theta}\sin\frac{\varphi}{2} & \cos\frac{\varphi}{2} \end{pmatrix}$ we have

$$\begin{aligned}
-\alpha\bar{\beta} - \alpha\beta &= -\alpha(\bar{\beta} + \beta) = -\cos\frac{\varphi}{2}(2\mathrm{Re}(\beta)) \\
&= 2\cos\frac{\varphi}{2}\sin\frac{\varphi}{2}\cos\theta = \sin\varphi\cos\theta = \frac{x^1}{r} \\
\alpha^2 - \beta^2 &= \cos^2\frac{\varphi}{2} - e^{-2\boldsymbol{i}\theta}\sin^2\frac{\varphi}{2} \\
&= \cos^2\frac{\varphi}{2} - (\cos 2\theta - \boldsymbol{i}\sin 2\theta)\sin^2\frac{\varphi}{2} \\
&= \left(\cos^2\frac{\varphi}{2} - \cos 2\theta\sin^2\frac{\varphi}{2}\right) + \boldsymbol{i}\sin 2\theta\sin^2\frac{\varphi}{2} \\
\boldsymbol{i}\big(\alpha\bar{\beta} - \alpha\beta\big) &= \boldsymbol{i}\alpha\big(\bar{\beta} - \beta\big) = \boldsymbol{i}\alpha(-2\boldsymbol{i}\,\mathrm{Im}(\beta)) \\
&= 2\alpha\,\mathrm{Im}\,(\beta) = 2\cos\frac{\varphi}{2}(\sin\theta\sin\frac{\varphi}{2}) \\
&= \sin\varphi\sin\theta = \frac{x^2}{r} \\
-\alpha^2 - \beta^2 &= (-\cos^2\frac{\varphi}{2} - \cos 2\theta\sin^2\frac{\varphi}{2}) + \boldsymbol{i}\sin 2\theta\sin^2\frac{\varphi}{2} \\
\alpha^2 - \beta\bar{\beta} &= \cos^2\frac{\varphi}{2} - \sin^2\frac{\varphi}{2} \\
&= \cos\varphi = \frac{x^3}{r}
\end{aligned}$$

$$\begin{aligned}
2\alpha\beta &= 2\cos\frac{\varphi}{2}\left(-e^{-\boldsymbol{i}\theta}\sin\frac{\varphi}{2}\right) = -e^{-\boldsymbol{i}\theta}\sin\varphi \\
&= -\cos\theta\sin\varphi + \boldsymbol{i}\sin\theta\sin\varphi.
\end{aligned}$$

We need just the (1,1) and (1,2) entries of ϕ^g so we compute as follows: The (1,1) entry of $x^a(g^{-1}\sigma_a g)$ is

$$\begin{aligned}
&x^1(\sin\varphi\cos\theta) + x^2(\sin\varphi\sin\theta) + x^3\cos\varphi \\
&= x^1\left(\frac{x^1}{r}\right) + x^2\left(\frac{x^2}{r}\right) + x^3\left(\frac{x^3}{r}\right) = \frac{r^2}{r} = r.
\end{aligned}$$

The real part of the (1,2) entry of $x^a(g^{-1}\sigma_a g)$ is

$$\begin{aligned}
&x^1(\cos^2\frac{\varphi}{2} - \cos 2\theta \sin^2\frac{\varphi}{2}) + x^2(-\sin 2\theta \sin^2\frac{\varphi}{2}) + x^3(-\cos\theta\sin\varphi)\\
&\quad= (r\sin\varphi\cos\theta)\cos^2\frac{\varphi}{2} - (r\sin\varphi\cos\theta)\cos 2\theta\sin^2\frac{\varphi}{2}\\
&\qquad - (r\sin\varphi\sin\theta)\sin 2\theta\sin^2\frac{\varphi}{2} - r\cos\varphi\cos\theta\sin\varphi\\
&\quad= r\sin\varphi\cos\theta\left(\frac{1}{2}+\frac{1}{2}\cos\varphi\right)\\
&\qquad - r\sin\varphi\cos\theta(1-2\sin^2\theta)\left(\frac{1}{2}-\frac{1}{2}\cos\varphi\right)\\
&\qquad - 2r\sin\varphi\sin^2\theta\cos\theta\left(\frac{1}{2}-\frac{1}{2}\cos\varphi\right) - r\cos\varphi\cos\theta\sin\varphi\\
&\quad= \frac{1}{2}r\sin\varphi\cos\theta + \frac{1}{2}r\sin\phi\cos\varphi\cos\theta\\
&\qquad - (r\sin\varphi\cos\theta - 2r\sin\varphi\cos\theta\sin^2\theta)\left(\frac{1}{2}-\frac{1}{2}\cos\varphi\right)\\
&\qquad - 2r\sin\varphi\cos\theta\sin^2\theta\left(\frac{1}{2}-\frac{1}{2}\cos\varphi\right) - r\cos\varphi\cos\theta\sin\varphi\\
&\quad= \frac{1}{2}r\sin\varphi\cos\theta + \frac{1}{2}r\sin\varphi\cos\varphi\cos\theta\\
&\qquad - (r\sin\varphi\cos\theta)\left(\frac{1}{2}-\frac{1}{2}\cos\varphi\right) - r\sin\varphi\cos\varphi\cos\theta\\
&\quad= \frac{1}{2}r\sin\varphi\cos\theta - \frac{1}{2}r\sin\varphi\cos\varphi\cos\theta\\
&\qquad - \frac{1}{2}r\sin\varphi\cos\theta + \frac{1}{2}r\sin\varphi\cos\varphi\cos\theta\\
&\quad= 0.
\end{aligned}$$

The imaginary part of the (1,2) entry of $x^a(g^{-1}\sigma_a g)$ is shown to be zero in the same way so the (1,2) entry itself is zero. Since ϕ^g takes values in $su(2)$ we have

$$\phi^g = -\frac{1}{2}\boldsymbol{i}\frac{f(r)}{r^2}\begin{pmatrix} r & 0\\ 0 & -r\end{pmatrix} = -\frac{1}{2}\boldsymbol{i}\frac{f(r)}{r}\begin{pmatrix} 1 & 0\\ 0 & -1\end{pmatrix}$$

$$\phi^g = \left(\coth r - \frac{1}{r}\right)T_3. \tag{2.5.42}$$

Thus, as promised, our Higgs field is now aligned along the third isospin axis T_3 in the internal space at each point of $\mathbb{R}^3$. As to the gauge potential $\boldsymbol{\mathcal{A}}^g$, we will again need the (1,1) and (1,2) entries in $g^{-1}\boldsymbol{\mathcal{A}}g + g^{-1}dg$. First observe that

$$
\begin{aligned}
g^{-1}\boldsymbol{\mathcal{A}}g &= g^{-1}\left(\frac{1-h(r)}{r^2}\sum_{a=1}^{3}\epsilon_{aij}x^j dx^i T_a\right)g \\
&= \frac{1-h(r)}{r^2}\sum_{a=1}^{3}(\epsilon_{aij}x^j dx^i)\left(g^{-1}T_a g\right) \\
&= -\frac{1}{2}\boldsymbol{i}\frac{1-h(r)}{r^2}\sum_{a=1}^{3}(\epsilon_{aij}x^j dx^i)\left(g^{-1}\sigma_a g\right) \\
&= -\frac{1}{2}\boldsymbol{i}\frac{1-h(r)}{r^2}\Big[(x^3dx^2 - x^2dx^3)\left(g^{-1}\sigma_1 g\right) \\
&\qquad\qquad + (x^1dx^3 - x^3dx^1)(g^{-1}\sigma_2 g) \\
&\qquad\qquad + \ (x^2dx^1 - x^1dx^2)(g^{-1}\sigma_3 g)\Big].
\end{aligned}
$$

Notice that the (1,1) entry of $g^{-1}\boldsymbol{\mathcal{A}}g$ is zero:

$$
\begin{aligned}
&-\frac{1}{2}\boldsymbol{i}\frac{1-h(r)}{r^2}\left[(x^3dx^2 - x^2dx^3)\left(\frac{x^1}{r}\right) + (x^1dx^3 - x^3dx^1)\left(\frac{x^2}{r}\right)\right. \\
&\qquad\qquad \left. +(x^2dx^1 - x^1dx^2)\left(\frac{x^3}{r}\right)\right] \\
&= -\frac{1}{2}\boldsymbol{i}\frac{1-h(r)}{r^3}\Big[x^1x^3dx^2 - x^1x^2dx^3 + x^1x^2dx^3 - x^2x^3dx^1 \\
&\qquad\qquad + x^2x^3dx^1 - x^1x^3dx^2\Big] \\
&= 0.
\end{aligned}
$$

Thus, the (1,1) entry of $\boldsymbol{\mathcal{A}}^g$ is the same as the (1,1) entry of $g^{-1}dg$ (and so, in particular, does not depend on the gauge potential $\boldsymbol{\mathcal{A}}$).
Let $g = \left(\begin{smallmatrix} \alpha & \beta \\ -\bar{\beta} & \alpha \end{smallmatrix}\right)$ with α real and depending only on φ and β depending only on φ and θ. Then $g^{-1} = \left(\begin{smallmatrix} \alpha & -\beta \\ \bar{\beta} & \alpha \end{smallmatrix}\right)$ and

$$
dg = \begin{pmatrix} \alpha_\varphi d\varphi & \beta_\varphi d\varphi + \beta_\theta d\theta \\ -\bar{\beta}_\varphi d\varphi - \bar{\beta}_\theta d\theta & \alpha_\varphi d\varphi \end{pmatrix}
$$

so

$$
g^{-1}dg = \begin{pmatrix} \alpha & -\beta \\ \bar{\beta} & \alpha \end{pmatrix}\begin{pmatrix} \alpha_\varphi d\varphi & \beta_\varphi d\varphi + \beta_\theta d\theta \\ -\bar{\beta}_\varphi d\varphi - \bar{\beta}_\theta d\theta & \alpha_\varphi d\varphi \end{pmatrix}
$$

The (1,1) entry of $g^{-1}dg$ is

$$\begin{aligned}
&\alpha\alpha_\varphi d\varphi + \beta\bar{\beta}_\varphi d\varphi + \beta\bar{\beta}_\theta d\theta \\
&\quad = \cos\frac{\varphi}{2}\left(-\frac{1}{2}\sin\frac{\varphi}{2}\right)d\varphi \\
&\qquad + \left(-e^{-\boldsymbol{i}\theta}\sin\frac{\varphi}{2}\right)\left(-e^{\boldsymbol{i}\theta}\frac{1}{2}\cos\frac{\varphi}{2}\right)d\varphi \\
&\qquad + \left(-e^{-\boldsymbol{i}\theta}\sin\frac{\varphi}{2}\right)\left(-\boldsymbol{i}e^{\boldsymbol{i}\theta}\sin\frac{\varphi}{2}\right)d\theta \\
&\quad = -\frac{1}{2}\sin\frac{\varphi}{2}\cos\frac{\varphi}{2}d\varphi + \frac{1}{2}\sin\frac{\varphi}{2}\cos\frac{\varphi}{2}d\varphi + \boldsymbol{i}\sin^2\frac{\varphi}{2}d\theta \\
&\quad = \boldsymbol{i}\sin^2\frac{\varphi}{2}d\theta \\
&\quad = \frac{1}{2}\boldsymbol{i}(1-\cos\varphi)d\theta.
\end{aligned}$$

Thus, $\boldsymbol{\mathcal{A}}^g$ has the form

$$\begin{aligned}
\boldsymbol{\mathcal{A}}^g &= \begin{pmatrix} \frac{1}{2}\boldsymbol{i}(1-\cos\varphi)d\theta & \text{——} \\ \text{——} & \text{——} \end{pmatrix} \\
&= -\frac{1}{2}\begin{pmatrix} (-(1-\cos\varphi)d\theta)\boldsymbol{i} & \text{——} \\ \text{——} & \text{——} \end{pmatrix}
\end{aligned}$$

which we write as

$$\boldsymbol{\mathcal{A}}^g = -\frac{1}{2}\begin{pmatrix} -\frac{1}{2}(2)\boldsymbol{i}(1-\cos\varphi)d\theta & \text{——} \\ \text{——} & \text{——} \end{pmatrix}. \tag{2.5.43}$$

A similar calculation for the (1,2) entry gives $(\boldsymbol{\mathcal{A}}^g)^1$ and $(\boldsymbol{\mathcal{A}}^g)^2$ as shown below.

$$\begin{aligned}
(\phi^g)^1 &= (\phi^g)^2 = 0 \\
(\phi^g)^3 &= \frac{f(r)}{r} = \coth r - \frac{1}{r} \\
(\boldsymbol{\mathcal{A}}^g)^1 &= -h(r)(\sin\theta\, d\varphi + \cos\theta\sin\varphi\, d\theta) \\
&= -r\,\mathrm{csch}\, r(\sin\theta\, d\varphi + \cos\theta\sin\varphi\, d\theta) \qquad (2.5.44) \\
(\boldsymbol{\mathcal{A}}^g)^2 &= h(r)(\cos\theta\, d\varphi - \sin\theta\sin\varphi\, d\theta) \\
&= r\,\mathrm{csch}\, r(\cos\theta\, d\varphi - \sin\theta\sin\varphi\, d\theta) \\
(\boldsymbol{\mathcal{A}}^g)^3 &= -(1-\cos\varphi)d\theta.
\end{aligned}$$

Now for the good part. We have already observed that $(\boldsymbol{\mathcal{A}}, \phi)$ is a globally defined, smooth field configuration on all of $\mathbb{R}^3$. As $r \longrightarrow \infty$ (i.e., as seen from a distance), the Higgs field approaches the constant value T_3 (because $\dfrac{f(r)}{r} \longrightarrow 1$), whereas $(\boldsymbol{\mathcal{A}}^g)^1$ and $(\boldsymbol{\mathcal{A}}^g)^2$ approach 0 (because $h(r) \longrightarrow 0$), while $(\boldsymbol{\mathcal{A}}^g)^3$, which does not depend on r, remains fixed at $-(1-\cos\varphi)d\theta$. As seen from infinity the potential function $\boldsymbol{\mathcal{A}}$ in this gauge assumes the form

$$-\frac{1}{2}\begin{pmatrix} -\frac{1}{2}(2)\boldsymbol{i}(1-\cos\varphi)d\theta & 0 \\ 0 & \frac{1}{2}(2)\boldsymbol{i}(1-\cos\varphi)d\theta \end{pmatrix}$$

which is essentially just the $\operatorname{Im}\mathbb{C}$-valued 1-form

$$-\frac{1}{2}(2)\boldsymbol{i}(1-\cos\varphi)d\theta$$

and this is precisely the potential for a Dirac monopole of magnetic charge 2 (see page 68). Seen from afar, the (nonsingular) t' Hooft-Polyakov monopole looks like a Dirac monopole. The "string singularity" (page 4, [**N4**]) of the Dirac monopole now shows up in the fact that our gauge transformation g (defined in terms of spherical coordinates on $\mathbb{R}^3$) is singular on the nonpositive x^3-axis.

Remark: The essential point to keep in mind here is that, in classical electrodynamics, monopoles are (but certainly need not be) "put in by hand," whereas, in $SU(2)$ Yang-Mills-Higgs theory, they arise of their own accord as solutions to the field equations.

Since the t'Hooft-Polyakov monopole satisfies $\boldsymbol{\mathcal{F}} = {}^*d^{\mathcal{A}}\phi$ one can compute the field strength $\boldsymbol{\mathcal{F}}$ entirely from $d^{\mathcal{A}}\phi$. In gauge s^g,

$$\begin{aligned}
d^{\mathcal{A}}\phi &= d\phi + [\boldsymbol{\mathcal{A}}, \phi] \\
&= d\left(\frac{f(r)}{r}T_3\right) + \left[\boldsymbol{\mathcal{A}}^a T_a, \frac{f(r)}{r}T_3\right] \\
&= d\left(\frac{f(r)}{r}\right)T_3 + \frac{f(r)}{r}\boldsymbol{\mathcal{A}}^a[T_a, T_3] \\
&= d\left(\frac{f(r)}{r}\right)T_3 + \frac{f(r)}{r}\left[\boldsymbol{\mathcal{A}}^1[T_1, T_3] + \boldsymbol{\mathcal{A}}^2[T_2, T_3] + \boldsymbol{\mathcal{A}}^3[T_3, T_3]\right] \\
&= d\left(\frac{f(r)}{r}\right)T_3 + \frac{f(r)}{r}[-\boldsymbol{\mathcal{A}}^1 T_2 + \boldsymbol{\mathcal{A}}^2 T_1] \\
&= \frac{f(r)}{r}(\boldsymbol{\mathcal{A}}^2 T_1 - \boldsymbol{\mathcal{A}}^1 T_2) + \left(\frac{\partial}{\partial r}\frac{f(r)}{r}\right)dr\, T_3 \\
&= \left(\coth r - \frac{1}{r}\right)(\boldsymbol{\mathcal{A}}^2 T_1 - \boldsymbol{\mathcal{A}}^1 T_2) + \left(\frac{1}{r^2} - \operatorname{csch}^2 r\right)dr\, T_3
\end{aligned}$$

$$d^{\mathcal{A}}\phi = \left(\coth r - \frac{1}{r}\right)(r\,\mathrm{csch}\,r)\big[(\cos\theta\,d\varphi - \sin\theta\sin\varphi\,d\theta)T_1 + (\sin\theta\,d\varphi + \cos\,\theta\,\sin\,\varphi\,d\theta)T_2\big] + \left(\frac{1}{r^2} - \mathrm{csch}^2\,r\right)dr\,T_3 \tag{2.5.45}$$

Let us return to an arbitrary configuration $(\boldsymbol{\mathcal{A}}, \phi)$ satisfying the monopole equations $\boldsymbol{\mathcal{F}} = \pm{}^*d^{\mathcal{A}}\phi$. We have already seen (page 134) that for such a configuration

$$\|\,\boldsymbol{\mathcal{F}}\,\|^2 + \left\|\,d^{\mathcal{A}}\phi\,\right\|^2 = \pm 2\left\langle \boldsymbol{\mathcal{F}}, {}^*d^{\mathcal{A}}\phi\right\rangle$$

so

$$\begin{aligned} A(\boldsymbol{\mathcal{A}}, \phi) &= \frac{1}{2}\int_{\mathbb{R}^3}\left(\|\,\boldsymbol{\mathcal{F}}\,\|^2 + \|\,d^{\mathcal{A}}\phi\|^2\right)\,dx^1\wedge dx^2\wedge dx^3 \\ &= \pm\int_{\mathbb{R}^3}\left\langle \boldsymbol{\mathcal{F}}, {}^*d^{\mathcal{A}}\phi\right\rangle dx^1\wedge dx^2\wedge dx^3 \\ &= \mp\int_{\mathbb{R}^3} 2\,\mathrm{trace}\big(\boldsymbol{\mathcal{F}}\wedge {}^{**}d^{\mathcal{A}}\phi\big) \\ &= \mp\int_{\mathbb{R}^3} 2\,\mathrm{trace}\big(\boldsymbol{\mathcal{F}}\wedge d^{\mathcal{A}}\phi\big) \end{aligned}$$

which we now write as

$$A(\boldsymbol{\mathcal{A}}, \phi) = \mp\int_{\mathbb{R}^3}\mathrm{Tr}\big(\boldsymbol{\mathcal{F}}\wedge d^{\mathcal{A}}\phi\big) \qquad \text{(monopole)}, \tag{2.5.46}$$

where $\mathrm{Tr} = 2\,\mathrm{trace}$. Computing this integral for the t'Hooft-Polyakov monopole (by writing it as $\int_{\mathbb{R}^3}\mathrm{Tr}({}^*d^{\mathcal{A}}\phi\wedge d^{\mathcal{A}}\phi) = -\int_{\mathbb{R}^3}\|\,d^{\mathcal{A}}\phi\,\|^2\,dx^1\wedge dx^2\wedge dx^3$ and using (2.5.45)) gives a value of 4π. For any configuration $(\boldsymbol{\mathcal{A}}, \phi)\in\mathcal{C}$ satisfying the monopole equations (2.5.14) we define the **monopole number** $N(\boldsymbol{\mathcal{A}}, \phi)$ by

$$N(\boldsymbol{\mathcal{A}}, \phi) = \frac{1}{4\pi}\int_{\mathbb{R}^3}\mathrm{Tr}(\boldsymbol{\mathcal{F}}\wedge d^{\mathcal{A}}\phi). \tag{2.5.47}$$

There are alternative ways of computing $N(\boldsymbol{\mathcal{A}}, \phi)$ (several of which are discussed below) that make it clear that $N(\boldsymbol{\mathcal{A}}, \phi)$ is actually an integer. Indeed, such an integer-valued monopole number can be defined in a much more general context. In [**JT**] it is shown that $\frac{1}{4\pi}\int_{\mathbb{R}^3}\mathrm{Tr}(\boldsymbol{\mathcal{F}}\wedge d^{\mathcal{A}}\phi)$ is well-defined and integer-valued for any $(\boldsymbol{\mathcal{A}}, \phi)\in\mathcal{C}$ that is a critical point for the action A given

by (2.5.10). Then **[Groi2]** shows that it is not even necessary to assume $(\boldsymbol{\mathcal{A}}, \phi)$ is a critical point. More precisely, if

$$A(\boldsymbol{\mathcal{A}}, \phi) = \frac{1}{2} \int_{\mathbb{R}^3} \left(\| \boldsymbol{\mathcal{F}} \|^2 + \| d^{\mathcal{A}} \phi \|^2 \right) dx^1 \wedge dx^2 \wedge dx^3$$

and

$$C = \left\{ (\boldsymbol{\mathcal{A}}, \phi) : A(\boldsymbol{\mathcal{A}}, \phi) < \infty, \lim_{R \longrightarrow \infty} \sup_{|x| \geq R} \left| 1 - \| \phi \| \right| = 0 \right\},$$

then, for any $(\boldsymbol{\mathcal{A}}, \phi) \in \mathcal{C}$,

$$N(\boldsymbol{\mathcal{A}}, \phi) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \mathrm{Tr} \left(\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}} \phi \right)$$

is a well-defined integer called the **monopole number** of $(\boldsymbol{\mathcal{A}}, \phi)$ (even though we do not assume that $(\boldsymbol{\mathcal{A}}, \phi)$ is a solution to the monopole equations).

Remark: The results of **[Groi2]** are actually much more general than this, but these will do for our purposes. Much of what we have to say also extends to the $\lambda > 0$ case which we abandoned on page 133 (see **[JT]** and **[Groi1]**).

We will now describe some alternative ways of computing $N(\boldsymbol{\mathcal{A}}, \phi)$ that shed much light on its topological significance. First we think of $\int_{\mathbb{R}^3}$ as $\lim_{R \longrightarrow \infty} \int_{|x| \leq R}$ and apply Stokes' Theorem to each $\int_{|x| \leq R}$. For this we first note that

$$\mathrm{Tr} \left(\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}} \phi \right) = d\big(\mathrm{Tr}(\phi \boldsymbol{\mathcal{F}}) \big), \tag{2.5.48}$$

where $\phi \boldsymbol{\mathcal{F}}$ is a matrix product. To see this note that $\mathrm{Tr}(d^{\mathcal{A}} \phi) = \mathrm{Tr}(d\phi + [\boldsymbol{\mathcal{A}}, \phi]) = \mathrm{Tr}(d\phi) + \mathrm{Tr}([\boldsymbol{\mathcal{A}}, \phi]) = \mathrm{Tr}(d\phi)$ because any commutator has trace zero ($\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$). Now, using a few simple computational facts that we will establish in Chapter 4,

$$\begin{aligned} d\left(\mathrm{Tr}(\phi \boldsymbol{\mathcal{F}})\right) &= \mathrm{Tr}\big(d(\phi \boldsymbol{\mathcal{F}})\big) = \mathrm{Tr}\big(d^{\mathcal{A}}(\phi \boldsymbol{\mathcal{F}})\big) \\ &= \mathrm{Tr}(\phi d^{\mathcal{A}} \boldsymbol{\mathcal{F}} + d^{\mathcal{A}} \phi \wedge \boldsymbol{\mathcal{F}}) \qquad \text{(“Product Rule”)} \\ &= \mathrm{Tr}(d^{\mathcal{A}} \phi \wedge \boldsymbol{\mathcal{F}}) \qquad \text{(Bianchi identity)} \\ &= \mathrm{Tr}(\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}} \phi), \end{aligned}$$

where the last equality follows from the fact that $d^{\mathcal{A}} \phi \wedge \boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}} \phi$ differ by a bracket, which has trace zero. This proves (2.5.48) and Stokes' Theorem (Section 4.7) gives

$$\int_{|x| \leq R} \mathrm{Tr}\Big(\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}} \phi \Big) = \int_{|x| \leq R} d\left(\mathrm{Tr}(\phi \boldsymbol{\mathcal{F}}) \right) = \int_{|x| = R} \mathrm{Tr}(\phi \boldsymbol{\mathcal{F}}).$$

Writing S^2_R for the set of points in $\mathbb{R}^3$ with $|\,x\,| = R$ we obtain

$$N(\boldsymbol{\mathcal{A}}, \phi) = \frac{1}{4\pi}\int_{\mathbb{R}^3} \mathrm{Tr}\,(\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}}\phi) = \lim_{R\longrightarrow\infty} \frac{1}{4\pi}\int_{|\,x\,|\leq R} \mathrm{Tr}\,(\boldsymbol{\mathcal{F}} \wedge d^{\mathcal{A}}\phi)$$

$$N\,(\boldsymbol{\mathcal{A}}, \phi) = \lim_{R\longrightarrow\infty} \frac{1}{4\pi}\int_{S^2_R} \mathrm{Tr}\,(\phi\boldsymbol{\mathcal{F}}). \tag{2.5.49}$$

Since $\lim_{R\longrightarrow\infty} \sup_{|\,x\,|\geq R} |\,1 - \|\,\phi\,\|\,| = 0$, there exists an $R_0 < \infty$ such that $\|\,\phi(x)\,\| > \frac{1}{2}$ for all x with $|\,x\,| > R_0$. For any such x we define

$$\hat{\phi}\,(\,x\,) = \|\,\phi\,(x)\,\|^{-1}\phi\,(x)$$

and, if $R > R_0$,

$$\hat{\phi}_R = \hat{\phi}|_{S^2_R}.$$

These are smooth maps on their domains and they map into $S^2_{su(2)}$ (the unit 2-sphere in $su(2) \cong \mathbb{R}^3$). One can show (see **[JT]** and **[Groi2]**) that ϕ can be replaced by $\hat{\phi}$ in (2.5.49), i.e.,

$$\begin{aligned} N(\boldsymbol{\mathcal{A}}, \phi) \;&=\; \lim_{R\longrightarrow\infty} \frac{1}{4\pi}\int_{S^2_R} \mathrm{Tr}(\,\hat{\phi}\boldsymbol{\mathcal{F}}\,) \\ &= \lim_{R\longrightarrow\infty} \frac{1}{4\pi}\int_{S^2_R} \|\phi\|^{-1}\mathrm{Tr}(\phi\boldsymbol{\mathcal{F}}). \end{aligned} \tag{2.5.50}$$

The reason for preferring the maps $\hat{\phi}$ and $\hat{\phi}_R$ can be seen as follows: Each $\hat{\phi}_R$ can be regarded as a map from S^2 to S^2 and so determines an element $[\hat{\phi}_R]$ of the homotopy group $\pi_2(S^2)$. Since $\hat{\phi}$ is smooth for $R > R_0$, $\hat{\phi}_R$ varies smoothly with $R > R_0$ so this homotopy class is *independent of* $R > R_0$ and we will denote it simply $[\hat{\phi}]$. We claim that $[\hat{\phi}]$ is also *gauge invariant*, i.e., that if $g : \mathbb{R}^3 \longrightarrow SU(2)$ is a gauge transformation and $\phi^g = g^{-1}\phi g$, then, on $|\,x\,| > R_0$, $\|\,\phi^g\,\|^{-1}\phi^g$ is well-defined and homotopic to $\|\,\phi\,\|^{-1}\phi$. It is well-defined because $\|\,\phi^g\,\| = \|\,g^{-1}\phi g\,\| = \|\,\phi\,\|$ which is nonzero on $|\,x\,| > R_0$. On the other hand, since $\mathbb{R}^3$ is contractible, g is homotopic to the map that sends all of $\mathbb{R}^3$ the identity e in $SU(2)$ (Exercise 2.3.6, **[N4]**). Thus, on $|\,x\,| > R_0$, $\|\,\phi^g\,\|^{-1}\phi^g = \|\,\phi\,\|^{-1}(g^{-1}\phi g)$ is homotopic to $\|\,\phi\,\|^{-1}(e^{-1}\phi e) = \|\,\phi\,\|^{-1}\phi$ so $[\phi^g] = [\phi]$ as required.

Each map $\hat{\phi}_R$ can be regarded as a map from S^2 to S^2 and therefore has a Brouwer degree $\deg(\hat{\phi}_R)$ (see Section 5.7 or Section 3.4 of **[N4]**). Since the various maps $\hat{\phi}_R, R > R_0$, determine the same homotopy class in $\pi_2(S^2)$, they have the same degree. Remarkably, this degree actually coincides with

the monopole number, i.e.,

$$N(\boldsymbol{\mathcal{A}}, \phi) = \deg\left(\hat{\phi}_R\right) = \deg\left(\| \phi \|^{-1} \phi|_{S^2_R}\right) \quad (R > R_0) \tag{2.5.51}$$

(see **[JT]** and **[Groi2]**). Notice that $\deg(\| \phi \|^{-1}\phi|_{S^2_R})$ depends only on ϕ and, indeed, only on its asymptotic behavior. Even more, it depends only on the "homotopy type of its asymptotic behavior" (if you get my drift). The monopole number distinguishes "homotopy classes" of Higgs fields. These classes are stable in the sense that a continuous perturbation of the field cannot change the class (physicists would say that an infinite potential barrier separates fields with different monopole numbers). Mathematically, there is a natural topology on the configuration space $\mathcal{C}$ with path components labeled by the integers and such that two configurations lie in the same path component if and only if they have the same monopole number (see **[Groi2]**).

We point out that there is an explicit integral formula for calculating the degrees (monopole numbers) in (2.5.51) that is sometimes more manageable than those in earlier formulas:

$$N(\boldsymbol{\mathcal{A}}, \phi) = -\frac{1}{4\pi} \int_{S^2_R} \mathrm{Tr}\left(\hat{\phi} d\hat{\phi} \wedge d\hat{\phi}\right) \quad (R > R_0). \tag{2.5.52}$$

We'll do this calculation for the t' Hooft-Polyakov monopole. From (2.5.39) and the fact that $\| \phi(x) \| = \coth r - \frac{1}{r}$ we conclude that, on S^2_R $(R > R_0)$,

$$\hat{\phi}(x) = \vec{n} \cdot \vec{T} = \frac{x^a}{R} T_a = -\frac{1}{2R} \begin{pmatrix} x^3 \boldsymbol{i} & x^2 + x^1 \boldsymbol{i} \\ -x^2 + x^1 \boldsymbol{i} & -x^3 \boldsymbol{i} \end{pmatrix}.$$

Thus,

$$\begin{aligned} & d\hat{\phi} \wedge d\hat{\phi} \\ &= \frac{1}{4R^2} \begin{pmatrix} dx^3 \boldsymbol{i} & dx^2 + dx^1 \boldsymbol{i} \\ -dx^2 + dx^1 \boldsymbol{i} & -dx^3 \boldsymbol{i} \end{pmatrix} \begin{pmatrix} dx^3 \boldsymbol{i} & dx^2 + dx^1 \boldsymbol{i} \\ -dx^2 + dx^1 \boldsymbol{i} & -dx^3 \boldsymbol{i} \end{pmatrix} \\ &= -\frac{1}{2R^2} \begin{pmatrix} dx^1 \wedge dx^2 \boldsymbol{i} & -dx^1 \wedge dx^3 + dx^2 \wedge dx^3 \boldsymbol{i} \\ dx^1 \wedge dx^3 + dx^2 \wedge dx^3 \boldsymbol{i} & -dx^1 \wedge dx^2 \boldsymbol{i} \end{pmatrix} \end{aligned}$$

and so

$$\begin{aligned} \hat{\phi}\, d\hat{\phi} \wedge d\hat{\phi} = & \frac{1}{4R^3} \begin{pmatrix} x^3 \boldsymbol{i} & x^2 + x^1 \boldsymbol{i} \\ -x^2 + x^1 \boldsymbol{i} & -x^3 \boldsymbol{i} \end{pmatrix} \\ & \times \begin{pmatrix} dx^1 \wedge dx^2 \boldsymbol{i} & -dx^1 \wedge dx^3 + dx^2 \wedge dx^3 \boldsymbol{i} \\ dx^1 \wedge dx^3 + dx^2 \wedge dx^3 \boldsymbol{i} & -dx^1 \wedge dx^2 \boldsymbol{i} \end{pmatrix}. \end{aligned}$$

The (1,1) entry is

$$\frac{1}{4R^3}\left[\left(-x^3dx^1\wedge dx^2+x^2dx^1\wedge dx^3-x^1dx^2\wedge dx^3\right)\right.$$
$$\left.+\left(x^1dx^1\wedge dx^3+x^2dx^2\wedge dx^3\right)\boldsymbol{i}\right].$$

and the (2,2) entry is

$$\tfrac{1}{4R^3}\left[\left(x^2dx^1\wedge dx^3-x^1dx^2\wedge dx^3-x^3dx^1\wedge dx^2\right)\right.$$
$$\left.-\left(x^1dx^1\wedge dx^3+x^2dx^2\wedge dx^3\right)\boldsymbol{i}\right].$$

Thus,

$$\begin{aligned}\mathrm{Tr}\left(\hat{\phi}d\hat{\phi}\wedge d\hat{\phi}\right)&=2\,\mathrm{trace}\left(\hat{\phi}d\hat{\phi}\wedge d\hat{\phi}\right)\\&=-\frac{1}{R^3}\left(x^1dx^2\wedge dx^3-x^2dx^1\wedge dx^3+x^3dx^1\wedge dx^2\right).\end{aligned}$$

We will learn how to integrate such a 2-form over S^2_R in Chapter 4 (indeed, we will find that the restriction of $x^1dx^2\wedge dx^3-x^2dx^1\wedge dx^3+x^3dx^1\wedge dx^2$ to S^2 is just the standard volume (i.e., area) form on S^2). Once the machinery is all in hand we will find that one can calculate such things by simply doing what comes natural. In this case, one introduces spherical coordinates

$$\begin{aligned}x^1&=R\sin\varphi\cos\theta\\x^2&=R\sin\varphi\sin\theta\\x^3&=R\cos\varphi\end{aligned}$$

with $0\le\theta\le 2\pi$ and $0\le\varphi\le\pi$, computes

$$\begin{aligned}x^1dx^2\wedge dx^3&=(R\sin\varphi\cos\theta)\,d\,(R\sin\varphi\sin\theta)\wedge d\,(R\cos\varphi)\\&=R^3\sin\varphi\cos\theta(\cos\varphi\sin\theta d\varphi+\sin\varphi\cos\theta d\theta)\wedge(-\sin\varphi d\varphi)\\&=-R^3\sin^3\varphi\cos^2\theta d\theta\wedge d\varphi\end{aligned}$$

and similarly for the remaining terms. The result is

$$\begin{aligned}\int_{S^2_R}\mathrm{Tr}\left(\hat{\phi}\,d\hat{\phi}\wedge d\hat{\phi}\right)&=-\int_{(0,\pi)\times(0,2\pi)}\sin\phi\,d\phi\wedge d\theta\\&=-\int_0^{2\pi}\int_0^{\pi}\sin\phi\,d\phi d\theta\\&=-4\pi.\end{aligned}$$

Thus, for the t' Hooft-Polyakov monopole,

$$N(\boldsymbol{\mathcal{A}}, \phi) = -\frac{1}{4\pi} \int_{S^2_R} \mathrm{Tr}\Big(\hat{\phi}\, d\hat{\phi} \wedge d\hat{\phi}\Big) = 1.$$

Remark: Changing the sign of ϕ (but leaving $\boldsymbol{\mathcal{A}}$ alone) gives a configuration which is a solution to $\boldsymbol{\mathcal{F}} = {}^* d^{\mathcal{A}} \phi$ (see page 144) and has monopole number -1.

The behavior of the Higgs field on large 2-spheres S^2_R therefore captures the topological type of the configuration. There is yet another way of seeing this that is reminiscent of our earlier experience with instantons (where the topological type was to be found in the Chern class of a certain bundle). To see this we again fix

$$(\boldsymbol{\mathcal{A}}, \phi) \in \mathcal{C} = \left\{ (\boldsymbol{\mathcal{A}}, \phi) : A(\boldsymbol{\mathcal{A}}, \phi) < \infty, \lim_{R \longrightarrow \infty} \sup_{|x| \geq R} |1 - \|\phi\|\,| = 0 \right\}$$

and select $R_0 < \infty$ such that $\|\phi(x)\| > \frac{1}{2}$ on $|x| > R_0$. Define

$$\hat{\phi} = \|\phi\|^{-1} \phi : \mathbb{R}^3 - \{x : |x| \leq R_0\} \longrightarrow S^2_{su(2)}$$

and, for each $R > R_0$,

$$\hat{\phi}_R = \hat{\phi}\Big|_{S^2_R} : S^2_R \longrightarrow S^2_{su(2)}.$$

Now we fix some $R > R_0$. ϕ is the pullback by the standard cross-section of an equivariant map $\boldsymbol{\Phi} : \mathbb{R}^3 \times SU(2) \longrightarrow su\ (2)$. The restriction of the trivial $SU(2)$-bundle over $\mathbb{R}^3$ to S^2_R is the trivial $SU(2)$-bundle over S^2_R:

$$SU(2) \hookrightarrow S^2_R \times SU(2) \xrightarrow{\mathcal{P}} S^2_R.$$

Let $\boldsymbol{\Phi}_R = \boldsymbol{\Phi}_{S^2_R \times SU(2)}$ and $\hat{\boldsymbol{\Phi}}_R = \|\boldsymbol{\Phi}_R\|^{-1} \boldsymbol{\Phi}_R$. Both are equivariant and $\hat{\boldsymbol{\Phi}}_R$ takes values in $S^2_{su(2)}$. Furthermore, $\hat{\phi}_R$ is the pullback to S^2_R by the standard cross-section of $\hat{\boldsymbol{\Phi}}_R$. Thus, $\hat{\phi}_R$ is the standard gauge representation for a Higgs field on the trivial $SU(2)$-bundle over S^2_R with values in $S^2_{su(2)}$.

Now, select some $\phi_0 \in S^2_{su(2)}$ (a "ground state" for the "virtual potential"; see pages 132–133). The isotropy subgroup of ϕ_0 (with respect to the adjoint action of $SU(2)$ on $su(2)$) is a copy of $U(1)$ in $SU(2)$ (pages 132–133). One can show that $\hat{\boldsymbol{\Phi}}_R^{-1}(\phi_0)$ is a submanifold of $S^2_R \times SU(2)$ (because $\hat{\boldsymbol{\Phi}}_R$ is a submersion at each point of $\hat{\boldsymbol{\Phi}}_R^{-1}(\phi_0)$) and, furthermore

(i) for each $x \in S^2_R$, $\mathcal{P}^{-1}(x) \cap \hat{\boldsymbol{\Phi}}_R^{-1}(\phi_0) \neq \emptyset$, and

(ii) for $p \in \hat{\mathbf{\Phi}}_R^{-1}(\phi_0)$ and $g \in SU(2)$,

$$p \cdot g \in \hat{\mathbf{\Phi}}_R^{-1}(\phi_0) \quad \text{iff} \quad g \in U(1) \qquad \text{(isotropy subgroup of } \phi_0\text{)}.$$

From these it follows that

$$\mathcal{P}\Big|\hat{\mathbf{\Phi}}_R^{-1}(\phi_0) : \hat{\mathbf{\Phi}}_R^{-1}(\phi_0) \longrightarrow S_R^2$$

is a principal $U(1)$-bundle over S_R^2 (where the action of $U(1)$ on $\hat{\mathbf{\Phi}}_R^{-1}(\phi_0)$ is just the original $SU(2)$-action on $S_R^2 \times SU(2)$, but with p restricted to $\hat{\mathbf{\Phi}}_R^{-1}(\phi_0)$ and g restricted to $U(1) \subseteq SU(2)$). This $U(1)$-bundle over S_R^2 is called a **reduction** of the structure group of $SU(2) \hookrightarrow S_R^2 \times SU(2) \xrightarrow{\mathcal{P}} S_R^2$ to $U(1)$. Recall that principal $U(1)$-bundles over spheres are characterized up to equivalence by their 1st Chern number (see pages 63–64 and 68 of Section 2.2). The result of interest to us is the following: *The 1st Chern number of*

$$U(1) \hookrightarrow \hat{\mathbf{\Phi}}_R^{-1}(\phi_0) \xrightarrow{\mathcal{P}\left|\hat{\mathbf{\Phi}}_R^{-1}(\phi_0)\right.} S_R^2$$

is the monopole number $N(\boldsymbol{\mathcal{A}}, \phi)$ *of the configuration* $(\boldsymbol{\mathcal{A}}, \phi)$.

Since the Chern class can be computed from any connection on the bundle, the proof amounts to finding such a connection that arises naturally from the original Yang-Mills-Higgs potential. We will briefly illustrate how this is done (the following is a special case of Proposition 6.4, Chapter II, of [**KN1**]): We have a connection on $SU(2) \hookrightarrow \mathbb{R}^3 \times SU(2) \longrightarrow \mathbb{R}^3$. Its restriction to $SU(2) \hookrightarrow S_R^2 \times SU(2) \xrightarrow{\mathcal{P}} S_R^2$ is a connection which we will denote $\boldsymbol{\omega}$. Since $U(1)$ is a subgroup of $SU(2)$, $u(1)$ is a subalgebra of $su(2)$. Now, we have an ad-invariant, positive definite inner product

$$\langle A, B\rangle = -2\text{trace}(AB)$$

on $su(2)$. Let $u(1)^\perp$ be the $\langle\ ,\ \rangle$-orthogonal complement of $u(1)$ in $su(2)$. Then $su(2) = u(1) \oplus u(1)^\perp$ and $u(1)^\perp$ is also ad-invariant (we let $\mu : SU(2) \longrightarrow GL(u(1)^\perp)$ be the induced representation). If $\iota : \hat{\mathbf{\Phi}}_R^{-1}(\phi_0) \hookrightarrow S_R^2 \times SU(2)$ is the inclusion map, then $\iota^*\boldsymbol{\omega}$ is an $su(2)$-valued 1-form on $\hat{\mathbf{\Phi}}_R^{-1}(\phi_0)$ and therefore splits

$$\iota^*\boldsymbol{\omega} = \boldsymbol{\omega}_0 + \boldsymbol{\gamma},$$

where $\boldsymbol{\omega}_0$ is $u(1)$-valued and $\boldsymbol{\gamma}$ is $u(1)^\perp$-valued. It's easy to see that $\boldsymbol{\omega}_0$ is a connection 1-form on

$$U(1) \hookrightarrow \hat{\mathbf{\Phi}}_R^{-1}(\phi_0) \xrightarrow{\mathcal{P}\left|\hat{\mathbf{\Phi}}_R^{-1}(\phi_0)\right.} S_R^2$$

(and $\boldsymbol{\gamma}$ is "tensorial of type μ"). Computing the 1^{st} Chern number of this bundle from $\boldsymbol{\omega}_0$ gives the expression (2.5.52) for $N(\boldsymbol{\mathcal{A}}, \phi)$.

2.6 Epilogue

We hope by now to have satisfied the curiosity of those who may have wondered how such apparently abstruse mathematical notions as spinor structures and characteristic classes might arise in the study of the world around us. We will have one more serious encounter with this in the Appendix, but it is time now to put aside the informal, heuristic, discussions that have characterized this chapter and deal honestly with these notions for their own sake. The remainder of the book is intended to do just that. Certainly, one need not demand any physical motivation to study and appreciate the rather beautiful mathematics to follow. Nevertheless, it is pro-foundly satisfying that the physical motivation exists and, as the concepts are made precise and the theorems are rigorously proved, we recommend a periodic dip in the murkier waters of physics (the journal *Communications in Mathematical Physics* is fine for browsing). It lends perspective.

http://www.springer.com/978-1-4419-7894-3

Topology, Geometry and Gauge fields
Interactions
Naber, G.L.
2011, XII, 420 p., Hardcover
ISBN: 978-1-4419-7894-3