Falling Cats, Parallel Parking, and Polarized Light

Robert W. Batterman[1]
Ohio State University

[1]I would like to thank Gordon Belot, Michael Berry, Rajendra Bhandari, Hans Boden, Roger Jones, and Mark Wilson for help in thinking about the topics discussed here. Thanks also to everyone I talked to who continue, as do I, to be fascinated by the simple experiment mentioned in section 1.1. Several anonymous referees offered extremely useful and insightful comments and criticisms. I am very pleased to have received the help.

Consider a system with a given complete set of state variables and dependent upon some set of parameters. Suppose you care about some quantity s that is a function of these variables and parameters. It turns out that in some instances, you can take the system on a round trip excursion in the abstract space of parameters and find that despite the fact that the state variables return to their initial values, and the fact that there is *no local rate of change for quantity* s, nevertheless, there is a *global* change in s's value at the end of the round trip. Things, in other words, are physically different. Furthermore, it turns out that you can explain the physical changes that appear as a result of these round trip excursions, by appeal to certain purely geometrical features of the abstract space in which the excursion can be parameterized. This is, *prima facie*, odd. What sort of role can geometrical/topological features of some abstract space play in *explaining* and providing *understanding* of "real" physical phenomena?

In some contexts, particularly those involving waves or wavefunctions, the failure to return to the same physical situation is attributed to what has been called a "geometric phase." The most important example of this is often called "Berry's phase" which was first discovered in studying the quantum mechanics of systems in situations where the adiabatic limit holds.

The understanding of geometric phases is related to a relatively recent controversy in the philosophical literature about how to understand the concept of gauge invariance. One aspect of this debate involves trying to understand the difference between the role of gauge potentials in classical physics (particularly, classical electromagnetism) where they appear to be nothing more than convenient mathematical constructs for generating *physically real* fields, and their role in quantum mechanics where it seems that they might very well have some sort of causal or physical relevance. The primary example discussed in the literature is the Aharonov-Bohm (AB) effect. See (Belot, 1998; Healey, 1997, 2001; Leeds, 1999). The AB effect, it turns out, is intimately related to Berry's phase.

This paper focuses on the explanatory value of the geometric structures that are the subject of this debate. Gauge structures appear in many places in physics and their geometric/topological properties often play important explanatory roles. In many cases issues about reifying these structures simply do not arise. One sees that genuine explanation of certain phenomena requires appeal to purely geometric or topological features of a relevant abstract space. The reason the debate rages in the context of quantum mechanics and electromagnetism—particularly in the AB effect—has to do with

certain metaphysical assumptions about the nature of spacetime which are absent in many applications where gauge invariance plays an important role.

Most of my attention here will focus on purely classical situations where round trip excursions are important. Examples include such diverse phenomena as why and how a cat can right itself when dropped with its legs up in the air, how a car can be parallel parked, and certain interference phenomena involving classical polarized light.

To get some sense of the ubiquity of this kind of geometrical aspect of round trip excursions in a space of parameters the next section considers some examples.

1 (An)holonomy: Some Examples

The failure of the physical situation to return completely to its original state upon a cycle of a parameter dependent system in parameter space is called an "anholonomy." Each such instance has the following form. Some quantity, s, characteristic of a system is "slaved" to certain variables $X_i, \{i = 1, 2, \ldots\}$ which are taken around some kind of loop in $\mathbf{X}$-space. If the values X_i return to their original values (that's what is meant by the loop), yet the slaved quantity s fails to return to its original value, the difference between the s values is the geometric phase or "anholonomy."[1]

1.1 Parallel Transport

Let's begin with a simple and familiar example. This is the parallel transport of a vector around a loop on the surface of the sphere. Consider the case where a vector tangent to the sphere at the north pole and to a given great circle follows that great circle down to the equator. It is then "parallel transported" along the equator (another great circle) to some other point, and then is finally taken back up to the north pole. See figure 1. Upon completion of its circuit on the sphere, parameterized by coordinates of longitude and latitude (X_1, X_2) the vector fails to return to its original "state." It is pointing in a different direction. This fact is called "holonomy" by the math-

[1]Many of the examples discussed here as well as a number of others are nicely presented in (Berry, 1991).

ematicians and "anholonomy" by the physicists.[2] The difference in angle between the initial and the final vectors at the north pole is proportional to the solid angle subtended by the circuit on the sphere and is independent of the particular coordinatization. It is a feature of the geometry of the surface of the sphere.

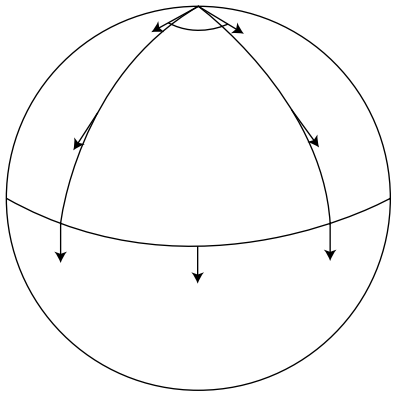


Figure 1: Parallel Transport Around a Sphere.

One can easily perform a very simple experiment which exhibits exactly the same phenomenon: Hold your arm out in front of you. Put out your thumb perpendicular to your arm so that it points up. Bring your arm up so that it is over your head. Next, bring your arm down to your side so that your thumb is now pointing backwards. Finally, bring your arm back in front of you. Your arm is pointing in the direction in which it started and your thumb is now pointing 90° from where it started. The direction of your thumb, just as the direction of the vector, fails to return to its initial place even though there has been no local rotation of your arm about its axis. Parallel transport is, in effect, defined in terms of the following restrictions: (1) the constant orthogonality between the the vector representing your thumb's direction and

[2]The terms "holonomy" and "anholonomy" derive from the classical mechanics of systems evolving under certain constraints. If the constraint is integrable and leads to a reduction in the number of degrees of freedom, it is called "holonomic." Nonintegrable constraints are called "anholonomic" or "nonholonomic." Geometers apparently do not respect this distinction calling anholonomies "holonomies." (Berry, 1990) takes this reversal of usage to be "a barbarism."

the radius vector from the center of the sphere (your arm's direction centered at your shoulder) and (2) the requirement that there be no twisting of the "thumb vector" about the radius vector.

1.2 Foucault's Pendulum

The kind of parallel transport just discussed features in Foucault's pendulum. Consider a "pendulum" that exhibits circular motion instead of the usual back and forth motion. (The latter can be understood as the superposition of two circular motions.) Consider figure 2.

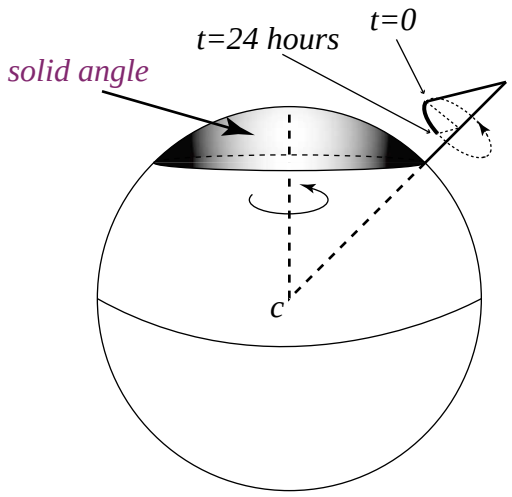


Figure 2: Foucault's Pendulum.

Suppose the pendulum bob has a period of one second about its axis of rotation. After one revolution of the earth (24 hours) about its axis the pendulum's axis clearly returns to the same position. However, pendulum bob has not returned to its initial position. That is to say, the "start" of the pendulum's rotation has shifted by a certain angle, called "Hannay's angle" which is equal to the solid angle subtended by the pendulum's axis of rotation around the globe. In this case the bob's starting position for its rotation about the pendulum axis is slaved to the rotation of the pendulum's axis itself.

1.3 Crystal Dislocations

And now for something (apparently) completely different. The simplest type of imperfection in a crystal is called an edge dislocation. Dislocations can

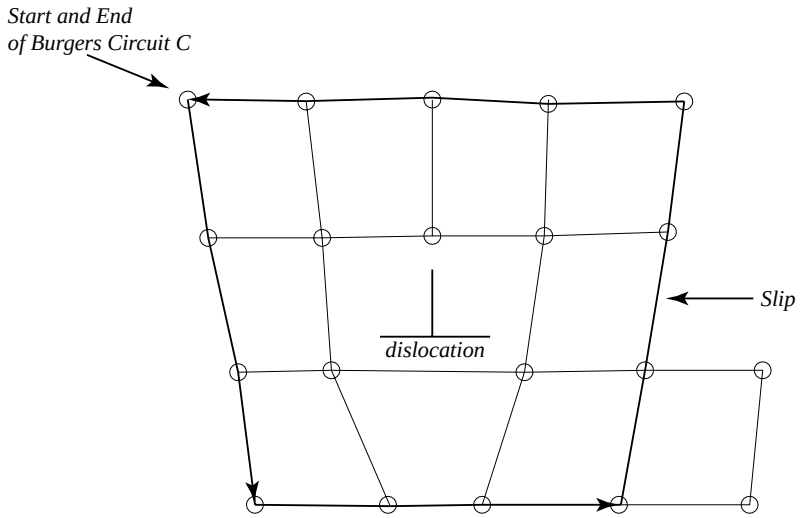


Figure 3: Edge Dislocation.

be produced by plane of atoms in a crystal lattice "slipping" over another plane in a way analogous to cards in a deck sliding over one another. If one ignores the edges of a crystal then such a slip doesn't change the perfection of the crystal structure since the entire plane just moves over, say, one lattice point. But most instances of slip are not global and affect only part of the slip plane leaving portions of it unaffected. Figure 3 gives an idea of what is going on here.[3]

Now imagine a circuit passing through lattice sites in the "good material"—that is, through parts of the crystal that lacks the imperfection.[4]

This is called a "Burgers circuit". In the figure it begins and ends at the site in the upper left. This curve, C, is associated with a circuit in an ideal (perfect) crystal that fails to close if and only if the Burgers circuit C

[3]See (Read, 1953) for a nice clear discussion. The fact that the lines connecting the lattice sites are not at right angles to one another reflects elastic strains in the material.

[4]In other words, such a circuit passes through lattice sites that, except for strains, look the same with respect to their nearest neighbors.

encircles a dislocation. This image is shown in figure 4.

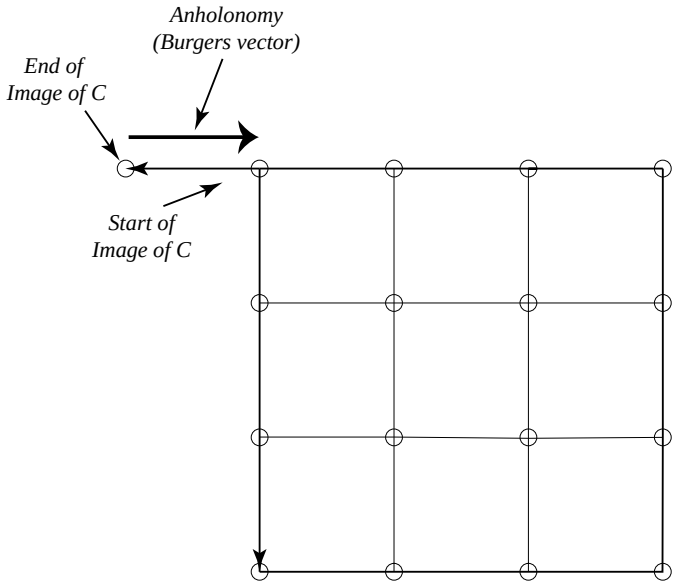


Figure 4: Image of the Burgers Circuit in Figure 3.

The anholonomy in this case is called the Burgers vector which is indicated in figure 4. In this case, it is the Burgers vector which is slaved to a set of discrete variables X_i that label the lattice sites of the crystal.

1.4 The Berry Phase

Recent interest in various anholonomies was sparked by Michael Berry's 1984 paper entitled "Quantal Phase Factors Accompanying Adiabatic Changes." Berry considered a nonrelativistic quantum system governed by a parameter dependent Hamiltonian $\hat{H}(\mathbf{X})$. He showed that if one transports the "system" adiabatically around a circuit C in parameter space ($\mathbf{X}$-space), the system will remain at every instant throughout this evolution in the same eigenstate for the Hamiltonian. Nevertheless, when the circuit C is completed, the system will have gained a circuit dependent "geometrical phase," $e^{i\gamma(C)}$ in addition to the dynamical phase, $e^{-iEt/\hbar}$, which is present in the evolution of any stationary state. The geometrical phase, known now as the "Berry Phase," was a truly remarkable discovery. It is a fundamental feature of quantum evolutions which had gone unnoticed by physicists working in

quantum mechanics for approximately 50 years![5] In the same paper Berry also showed that one can understand the famous AB effect—a quantum mechanical effect—as an instance of the geometrical or Berry phase. This will be discussed further in section 2 below.

1.5 Pancharatnam's Phase

Berry's work was presaged by the Indian physicist S. Pancharatnam (Pancharatnam, 1956) who discovered an analogous anholonomy while studying phase shifts in classical polarized light. Pancharatnam discovered an anholonomy in the phase of a light wave as it is taken through a cycle of polarization states. This phase shift is distinct from the shift associated with the free propagation of light over the same path. Here let me briefly describe Pancharatnam's phase. More details will be offered below in section 4.

In the classical theory of light one can completely represent the polarization states of a plane wave of light with a given wave vector $\mathbf{k}$ by points on the surface of a sphere called the "Poincaré sphere." (See figure 5.) The "north pole" of the sphere represents the state in which the light is right circularly polarized, the "south pole" represents left circularly polarized light, points along the "equator" represent different states of linear polarization, and all other points represent different states of elliptical polarization. Antipodal points represent orthogonal states of polarization.[6]

Pancharatnam considered the question of how to define the phase difference between two such light waves in different states of polarization. On physical grounds he argued that one ought to consider them to be completely in phase if, were one to allow them to interfere, the intensity of the resulting beam would be a maximum. In effect, this defines a conception of "distant parallelism"—a connection—on the Poincaré sphere. (Berry, 1987, p. 1402)

A consequence of Pancharatnam's definition is that "being in phase" is not transitive. That is, suppose a wave in polarization state $|A\rangle$ is in phase with a wave polarized in state $|B\rangle$. Further, suppose that $|B\rangle$ is in phase with $|C\rangle$. On Pancharatnam's conception, it doesn't follow that $|A\rangle$ is in phase with $|C\rangle$. In particular, if a light beam originally in state $|A\rangle$ is taken

[5]Actually, the geometrical phase is much more general than Berry's original paper shows. The evolutions need be neither adiabatic nor unitary as shown by (Samuel and Bhandari, 1988).

[6]Unless otherwise noted we will always assume that the light waves are completely polarized and of unit (normalized) intensity.

though a sequence of "ideal polarizers"[7] $|B\rangle$, $|C\rangle$, and then $|A\rangle$, the resultant beam $|A'\rangle$ generally will not be in phase with the initial beam. The difference in phase between $|A\rangle$ and $|A'\rangle$ is the anholonomy and is equal to -1/2 times the solid angle subtended by the spherical triangle ABC at the center of the Poincaré sphere. In this case the slaved variable is the phase of the light wave as it is taken around a loop in the polarization space—a circuit on the Poincaré sphere.

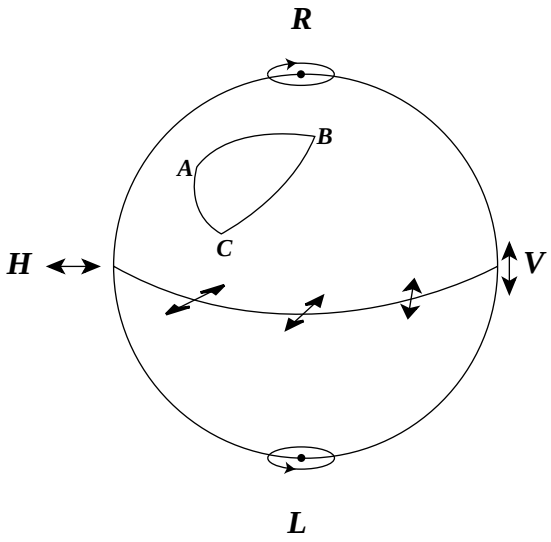


Figure 5: The Poincaré Sphere.

1.6 Falling Cats and Parallel Parking

A problem that has received a fair amount of attention in the literature on modern geometrical mechanics[8] concerns the description and explanation of the following seemingly paradoxical, yet commonplace fact: A cat when dropped at rest with its feet pointing up will (often, hopefully, if it's not too high ...) manage to right itself and land safely on its feet. Somehow

[7]This means that there is no loss of intensity as the beam is passed through the polarizer. (Equivalently, the polarizers in this idealization are represented by unitary transformations on the states.)

[8]See, (Montgomery, 1993) for a discussion.

the cat is able to rotate itself 180° on the way down *even though* (and this is the apparently paradoxical part) *it has zero angular momentum to start with, and by conservation of angular momentum, has zero angular momentum throughout its fall.* (Note, too, that the cat does not get any external purchase by scratching at the air.)

The apparent paradox comes from thinking of the cat as a rigid body for which one defines the angular momentum as the moment of inertia times the angular velocity. Of course, a cat is not a rigid body and is capable of exerting its muscles in such a way as to change its *shape*. We can think of the cat as having the same shape at the beginning of the fall (feet perpendicular to its body, approximately) as it does at the end of its fall. This suggests that we represent the cat's contortions in a space of shapes. As the cat twists itself around it changes its shape, eventually coming to have the same shape it began with. It executes a circuit or round trip in shape space, the end result of which is equivalent to a 180° rigid rotation in real space. Thus, its orientation in physical space is slaved to a set of variables describing its shape in shape space.[9] The anholonomy or geometric phase is the rotation of the cat in physical space.

In this situation we see that there is a constraint imposed upon the system—namely, the conservation of angular momentum. Geometrically, this constraint is a symmetry of the mechanical system. There are other types of constraints that do not arise from mechanical symmetries but which also lead to anholonomies. A paradigm example of this kind of anholonomic constraint is that of a wheel or ball being constrained to roll on a surface *without skidding.*

A car, for instance, is constrained (most of the time, one hopes) in this way. It can only move in the direction of its wheels and at a rate proportional to the angular velocity of the wheels. Crucially, the car cannot move (without skidding) perpendicular to its front-back orientation. Nevertheless, by executing familiar maneuvers one is able to move the car into a parking spot exactly perpendicular to this orientation! The car executes a series of motions involving changes in steering direction and direction of motion so that, without skidding, it moves in the perpendicular direction. It begins in a certain orientation or shape (parallel to the parking spot) and undergoes a circuit in the space of steering directions and directions of motion so that

[9]These variables will, for instance, describe the angles between the cats various body parts.

it returns to its initial orientation or shape but undergoes a motion impossible to achieve with either input motion alone. Its net displacement into the parking spot is the anholonomy.

2 The Aharonov-Bohm (AB) Effect

In 1959 Aharonov and Bohm predicted a peculiar quantum mechanical effect. For our purposes here we can discuss a simple gedanken experiment which illustrates the essential points. (See (Healey, 1997) for a detailed discussion of this experiment.) Consider a two slit experiment with electrons. Let $|\psi_1\rangle$ be the amplitude for passing through slit 1 and $|\psi_2\rangle$ be the amplitude for passing through slit 2. Then the probability density for arriving at the detector screen C is given by $\| \, |\psi_1\rangle + |\psi_2\rangle \, \|^2$. (See figure 6.) The effect of turning on the

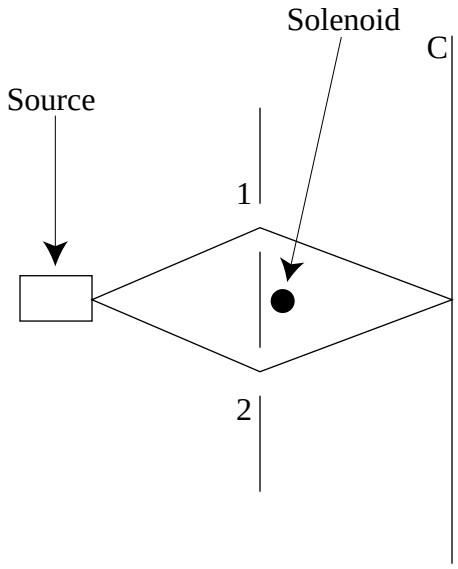


Figure 6: Two Slit Experiment with Solenoid.

solenoid current is to create a nonzero magnetic field within the solenoid (coming out of the page) and zero field elsewhere. Despite the fact that the electrons traversing the apparatus feel no magnetic field regardless of the magnetic flux through the solenoid, there is an observable difference in the interference pattern on the screen. This is the AB effect. The result of turning on the current leads to a new probability density for arriving at the

screen C given by $\||\psi_1\rangle + e^{iq\Phi}|\psi_2\rangle\|^2$, where q is the charge on the electron and Φ is the magnetic flux through the solenoid. (I'm oversimplifying somewhat here, see (Healey, 1997) for more details.) In other words, the interference pattern that appears on the screen will experience a shift when there is flux through the solenoid. The fact that the magnetic field is zero everywhere outside the cylinder has led a number of interpreters to wonder about the nonlocal effect that the field inside the solenoid may have on the distant electrons.

Some, including Aharonov and Bohm themselves, argue that this demonstrates that the magnetic vector potential $\mathbf{A}$ (which *does* change when there is current in the solenoid) is acting on the electrons as they traverse the apparatus. See (Healey, 1997) and (Belot, 1998) for discussions of various interpretive moves.

One way, stressed by Belot, of understanding the import of the AB effect relies on treating classical electromagnetism in the framework of gauge theories. On the traditional interpretation of electromagnetism (with no AB solenoid in the picture), the magnetic field $\mathbf{B}$ is defined to be the *curl* of the vector potential $\mathbf{A}$:

$$\mathbf{B} \equiv \nabla \times \mathbf{A}. \tag{1}$$

And, since $\nabla \times \mathbf{A} = \nabla \times (\mathbf{A} + \nabla\chi)$ for any sufficiently smooth function χ, we see that $\mathbf{B}$ is invariant under the transformation

$$\mathbf{A} \mapsto \mathbf{A}' = \mathbf{A} + \nabla\chi. \tag{2}$$

This is a gauge transformation. In classical electromagnetism $\mathbf{B}$ is, therefore, a gauge invariant quantity whereas $\mathbf{A}$ is not. Gauge invariance is often taken to be a necessary condition for a field quantity to be physically "real."[10] Hence for a fixed value of the electric field, $\mathbf{A}$ and $\mathbf{A}'$ can correspond to the same magnetic field—if they do, they lie on the same gauge orbit.

Now, if we consider the AB effect in this gauge framework, we see a problem. If we consider the field outside the solenoid, we find that there are potentials $\mathbf{A}$ and $\mathbf{A}'$ that correspond to the same magnetic field, yet lie on different gauge orbits. Belot's paper is an extended investigation of the consequences of this problem for interpreting electromagnetism.

In his original paper (Berry, 1984) Berry argues that the AB effect is a specific instance of his geometrical phase and hence an instance of anholon-

[10](Healey, 1997, p. 22), for instance, says "[b]ut there is reason to doubt that the magnetic vector potential is a physically real field, since $\mathbf{A}$ is not gauge-invariant ...."

omy. Roughly, the idea is that one can consider the experimental situation of figure 6 from a different perspective. Consider a single electron which traverses the upper path through slit 1 and is then brought back along the lower path through slit 2. It then traverses a circuit that encloses the solenoid, returning to its initial position in parameter space **X** which, in this case, is *real* space or spacetime. Upon return the wavefunction for the electron has picked up an additional phase that is directly proportional to the magnetic flux in the solenoid. In section 3 we will see why it is correct to consider this phase to be a function of the "geometry" of the situation.

Most interpreters argue that the AB effect shows us that the traditional interpretation of electromagnetism is untenable. On that view, *only* the electric and magnetic (or the electromagnetic) fields act on charged particles. And, the fields act on the particles *locally*. The AB effect can be understood as raising doubts about both these claims. Belot in effect subscribes to this view that the traditional interpretation is misguided although he puts the point slightly differently. He says that "[*U*]*ntil the discovery of the Aharonov-Bohm effect, we misunderstood what electromagnetism was telling us about our world.*" (Belot, 1998, p. 532)

3 (An)Holonomy and Fiber Bundles

The first examples of anholonomy considered above in section 1 were transparently geometrical in nature—they involved, explicitly the notion of parallel transport on the surface of a sphere. In this section, I would like to briefly discuss the general, natural mathematical theory for representing anholonomy. This is the theory of fiber bundles and it is here that we can see that all of the examples discussed above, including the AB effect, are really instances of a similar type of geometrical phenomenon. I will first describe the theory of fiber bundles using two simple examples: The cylinder and the Möbius strip. Following this I will discuss the more complicated case of the magnetic monopole which, as it will turn out, is intimately connected with the geometrical characterization of the Poincaré sphere.

3.1 The Cylinder and The Möbius Strip

Consider the space which is a cylinder of unit height and radius. This space, call it E, is the direct product of two spaces—a space M which in this case

is the circle $\mathbf{S}^1$, and a space F which in this case is the line segment $(0,1)$:

$$E = M \times F.$$

E (the cylinder) is called the "total space", $\mathbf{S}^1$ is called the "base space," and the line segment, $(0,1)$, is called the "fiber." See figure 7. There

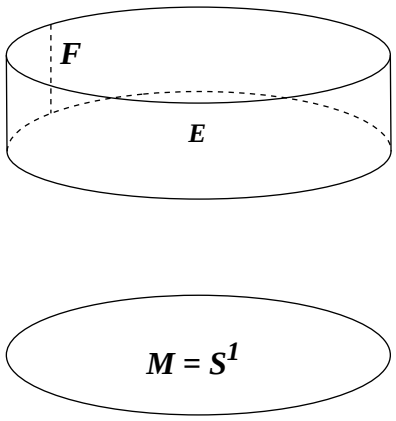


Figure 7: The Cylinder $E = M \times F$.

is a projection, π which maps the total space E onto the base space M. Suppose we cover the circle $\mathbf{S}^1$ with small neighborhoods U_α. Then for each neighborhood U_α, $\pi^{-1}(U_\alpha)$ is homeomorphic to $U_\alpha \times F$.[11] In other words, locally the total space E looks like a direct product. In fact, this is trivially the case for the cylinder since the entire cylinder (globally) just is itself the direct product $M \times F$.

Now consider the Möbius strip to be the total space. Here too the base space M is the circle $\mathbf{S}^1$ and the fiber F is the line segment $(0,1)$. See figure 8. If we take a small neighborhood U_α of some point in $\mathbf{S}^1$, then just as with the cylinder $\pi^{-1}(U_\alpha)$ will look like $U_\alpha \times F$. Globally, however, the Möbius strip E is not a direct product: It is twisted.

One can cover the base space M by small neighborhoods U_α with overlaps $U_i \cap U_j \equiv U_{ij}$. Suppose the bundle over each U_α is homeomorphic to $U_i \times F$. In order to sew all of these local bundles together to form the total space E we must have rules which enable the identification of the local trivializations

[11] A homeomorphism is a continuous map with a continuous inverse.

($U_i \times F$ and $U_j \times F$) above the intersections U_{ij}. These rules or transition functions (call them g_{ij}) are continuous maps from the intersections to a group G which acts on the fibers: $g_{ij} : U_{ij} \to G$. So a fiber bundle is a 5-tuple (E, π, F, G, M). In the case of the Möbius strip, one has $G = \{\pm 1\}$ where the element -1 acts on $F = (0, 1)$ by sending x to $1 - x$.

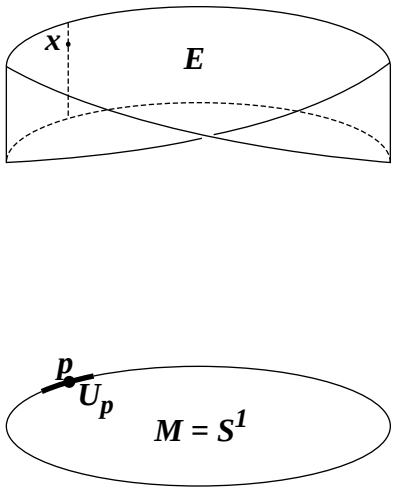


Figure 8: The Möbius Strip.

For instance, choose a point x on the fiber $\pi^{-1}(p)$ above a point p on the base space $\mathbf{S}^1$. Suppose that x is $3/4$ of the way "up" the fiber. See figure 8. Of course, x is a point in the total space E—the Möbius strip. In a local coordinate system of a neighborhood $U_p \times F$, x has coordinates: $(p, 3/4)$. If we choose a homeomorphism h from the fiber above p to the interval (0,1) and try to extend it continuously as we go around the circle from p back to p, we will see that fiber's "direction" will have been reversed. That is the image of the point x under this transformation will have coordinates in the neighborhood $U_p \times F$: $(p, 1/4)$. The point on the total space fails to return to its original position after a circuit in the base space. This is the anholonomy associated with the Möbius strip.

3.2 Reduction and Reconstruction

The mechanical examples of the falling cat and parallel parking are best understood in terms of fiber bundles. As mentioned in section 1.6 these systems evolve under certain kinds of constraints. Geometrical mechanics seeks

to understand the behaviors of such systems by exploiting the symmetries (e.g., angular momentum conservation) or constraints to *reduce* the phase space of the system to one with fewer degrees of freedom than the full phase space. In the case of falling cats, we can understand this reduction in the following way.

The full configuration space consists of the shape space together with a space of rotations and translations which orient the cat in physical space. Thus the base space of the fiber bundle is the shape space of the cat. The connection on the fiber bundle is sometimes called the "mechanical connection" and is the direct result of the symmetries of rigid body rotation. (Cendra et al., 2001; Montgomery, 1993) The group G of the fiber bundle is in this case the group of rigid spatial rotations and translations. In such situations it is fairly natural to think of the base space of the bundle (the shape space) as a space of control parameters. Thus the cat is able to control—via its muscles—its shape, resulting ultimately in the 180° rotation in real space.

Were one to observe the cat's evolution in shape space alone—that is, in a frame of reference attached to some part of the cat—then it would appear that there are "mysterious" (Coriolis and centrifugal) forces that are acting upon the cat. These forces can be understood in terms of the curvature of the mechanical connection. From this point of view, the problem is then to *reconstruct* the "full" motion, given the motions in the shape space—the reduced configuration space. Thus one aspect of the reconstruction problem is to determine the geometric phases or anholonomies. The reconstruction problem is, therefore, a sort of inverse of the program of reduction.

Pancharatnam was, in effect, engaged in a program of reconstruction. He realized that the representation of the polarization states of light in terms of the Poincaré sphere failed to capture all the observable features of polarized light. So, without really realizing it, he showed that one needs a nontrivial fiber bundle—the Hopf bundle—in order to tell the complete story.[12]

Whether one is most interested in reduction or reconstruction largely depends upon how the phenomenon in question presents itself. Sometimes, as in the case of the falling cat, we seek to understand the "full" behavior, and proceed by simplifying the dynamics given symmetries we observe. In other cases, such as the AB effect, it seems that we notice apparently anomalous behavior in (what turns out to be) the base space, and are concerned to explain or account for the presence of the anholonomies. The real difference

[12]The Hopf bundle is discussed in detail in the next section (3.3).

here is a difference in *attitude* towards the space in which the phenomenon takes place. In the case of the cat, we witness the behavior in real physical space, and seek to understand the behavior in terms of reduction focusing on the shape space. In the case of the AB effect, the phenomenon also apparently takes place in real physical space or spacetime, yet the full understanding seems to require our taking space or spacetime to be the base space of a fiber bundle and so we engage in the reconstruction.

3.3 The Magnetic Monopole

Let us now consider another example in somewhat more detail. Suppose, contrary to what is believed to be the case, there are magnetic monopoles. The upshot of this supposition is the nonzero divergence of the magnetic field $\mathbf{B}$, $\nabla \cdot \mathbf{B} \neq 0$. This has the consequence that one *cannot* write $\mathbf{B}$ as the curl of some *nonsingular* vector potential $\mathbf{A}$ as in equation (1) above. A magnetic monopole of strength g located at the origin has a radial magnetic field $\mathbf{B} = g\mathbf{r}/r^3$. The total magnetic flux through a sphere surrounding the monopole is $\Phi = 4\pi g$. Now, it is possible to find a vector potential $\mathbf{A}$ that gives the correct field $\mathbf{B}$ everywhere except on an infinite ray beginning at the origin which, without loss of generality, we take to be the negative z-axis. In spherical coordinates, (r, θ, ϕ), $\mathbf{A}$ is given by

$$A_r = A_\theta = 0, \quad A_\phi = \frac{g}{r}\frac{(1 - \cos\theta)}{\sin\theta}. \tag{3}$$

On the negative z-axis, $\theta = \pi$, and the vector potential is singular. This singularity in $\mathbf{A}$ along the entire negative z-axis is called a "Dirac string".

We could have chosen to have the Dirac string singularity along the positive z-axis (or, actually, along any radially increasing semi-infinite path beginning at the origin). Had we made the choice of the positive z-axis, then $\mathbf{A}$ would be given by

$$A_r = A_\theta = 0, \quad A_\phi = -\frac{g}{r}\frac{(1 + \cos\theta)}{\sin\theta}, \tag{4}$$

which is singular when $\theta = 0$.

(Wu and Yang, 1975) demonstrate how it is possible to define two vector potentials $\mathbf{A}^+$ and $\mathbf{A}^-$ each of which is nonsingular over a particular domain (respectively, R^+ and R^-) such that (i) their curls equal the magnetic field of the monopole in their respective regions and (ii) in their overlap region $R^\pm \equiv$

$R^+ \cap R^-$, $\mathbf{A}^+$ and $\mathbf{A}^-$ are related to one another by a gauge transformation. The regions are defined (for r fixed) by

$$\begin{aligned} R^+ &\equiv \{(\theta,\phi) : 0 \le \theta < \frac{\pi}{2} + \epsilon\} \\ R^- &\equiv \{(\theta,\phi) : \frac{\pi}{2} - \epsilon \le \theta \le \pi\}. \end{aligned}$$

See figure 9.

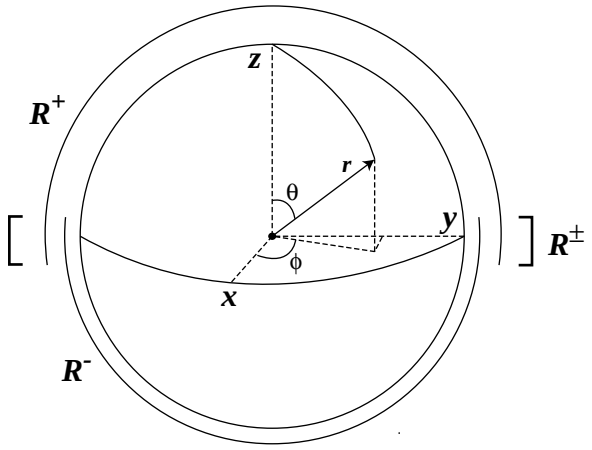


Figure 9: Construction of Nonsingular Vector Potentials for a Monopole.

The potentials, $\mathbf{A}^+$ and $\mathbf{A}^-$, in the domains R^+ and R^- are, respectively, just the potentials given by equations (3) and (4). In the overlap region $R^\pm$ they are related by the gauge transformation:

$$\begin{aligned} A_\phi^- &= A_\phi^+ - \frac{2g}{r\sin\theta} \\ A_\theta^- &= A_\theta^+ \\ A_r^- &= A_r^+ \end{aligned} \tag{5}$$

Geometrically, the above construction is an instance of what mathematicians call the "Hopf bundle." And the idea is that the vector potential is most naturally represented as the connection on the Hopf bundle. This is a fiber bundle with base space $\mathbf{S}^2$ (the surface of a sphere surrounding the

monopole at the origin) and fiber $U(1)$.[13] The Hopf bundle is similar to the Möbius strip in that both are nontrivial fiber bundles. The restricted bundle

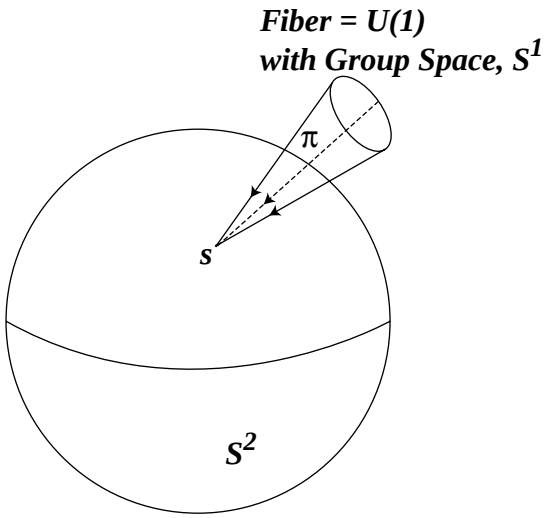


Figure 10: The Hopf (Monopole) Bundle.

over the upper region R^+ is homeomorphic to the direct product $R^+ \times U(1)$ and similarly the bundle over R^- is homeomorphic to $R^- \times U(1)$. However, globally, the bundle is *not* the product $\mathbf{S}^2 \times \mathbf{S}^1$; instead, it is $\mathbf{S}^3$. Figure 10 provides a way of visualizing some of this. As noted the total space E is the space $\mathbf{S}^3$. The projection π maps many points of $\mathbf{S}^3$—all those related to one another by multiplication by a member of the group $U(1)$—into a single point s in the base space $\mathbf{S^2}$.

Just as with the Möbius strip there are transition functions $g_\pm$ from the overlap regions (e.g. $R^\pm$) to the group $G = U(1)$ which enable the identification of points in the local trivializations $R^+ \times U(1)$ and $R^- \times U(1)$ above the overlap regions. That is,

$$g_\pm : (\theta, \phi, e^{i\alpha_-}) \mapsto (\theta, \phi, e^{if_\pm(\theta,\phi)} e^{i\alpha_+}), \tag{6}$$

where θ and ϕ are in the overlap region $R^\pm$.[14] These transition functions,

[13] $U(1)$ is the group of complex numbers of unit modulus—all numbers of the form $e^{i\alpha} = \cos\alpha + i\sin\alpha$. Furthermore, since $\cos^2\alpha + \sin^2\alpha = 1$, the space of these complex numbers is the circle $\mathbf{S}^1$. This example is a bit more complicated than the Möbius strip since the fiber is itself the structure group G. This is called a "Principal fiber bundle."

[14] The $f_\pm$ are winding numbers. They are topological invariants.

in fact, give rise to the gauge transformations such as that in equation (5) above.

Now consider an electron of charge q in the field of this monopole. Given the potentials $\mathbf{A}^+$ and $\mathbf{A}^-$ it follows that if the electron is slowly taken around a circuit on a sphere of radius r, its wavefunction will gain a phase which is proportional to the magnetic flux Φ through that area on the surface of the sphere defined by the circuit. In particular, suppose the circuit is defined by fixed r and θ but where ϕ ranges from 0 to 2π. The region, R^+ or R^-, in which the circuit lies determines the vector potential $\mathbf{A}^+$ or $\mathbf{A}^-$ to be used. Suppose it is in R^+. Then the change in phase $\Delta\gamma$ is given by

$$\begin{aligned} \Delta\gamma &= \frac{q}{\hbar c}\oint \mathbf{A}^+ \cdot \mathbf{dl} \\ &= \frac{q}{\hbar c}\int \left(\nabla \times \mathbf{A}^+\right) \cdot \mathbf{dS} \\ &= \frac{q}{\hbar c}\int \mathbf{B} \cdot \mathbf{dS} \\ &= \frac{q}{\hbar c}\Phi(r,\theta) \end{aligned} \tag{7}$$

It is clear that this phase change is purely a geometrical property since the flux is proportional to the solid angle of the circuit subtended by the circuit on the sphere. This is the physical manifestation of the anholonomy of the Hopf bundle.

There is a difference between the anholonomy of the Möbius strip and that of the monopole. The connection on the Möbius strip is flat. The monopole bundle, on the other hand, has curvature. *What is the relationship between the existence of anholonomy (nontrivial holonomy) and the curvature of the connection?* If the base space M is simply connected, and the connection is flat, then there is *no* nontrivial holonomy—no anholonomy.[15] Since the circle $\mathbf{S}^1$ is not simply connected—for instance, a path that loops k times around the circle cannot be deformed into one which goes around it l times ($l \neq k$) without leaving the circle—the Möbius strip can exhibit anholonomy despite its flatness. On the other hand, the base space for the monopole bundle is $\mathbf{S}^2$ which *is* simply connected—all paths that bound disks can be contracted to a point. The anholonomy in this case is due to the curvature of the connection.

[15] A topological space is simply connected if every loop in the space can be contracted to a point without leaving the space.

The connection here is, in fact, the vector potential $\mathbf{A}$—more precisely, it is the collection of local vector potentials $\mathbf{A}^+$ and $\mathbf{A}^-$.

Despite this difference there is an important sense in which the two situations are *strongly analogous*. The following conditional holds:

> If there is nontrivial holonomy or *anholonomy* and if the connection is *flat*, then *the base space must be nonsimply connected.*

The next section aims to show how this conditional yields a very strong geometric/topological analogy between the AB effect and the magnetic monopole.

3.4 Dirac Strings, The Wu-Yang Connection, and the AB Effect

We have just seen that it is possible to remove the singularity—the string—in the Dirac presentation of the monopole by moving to the fiber bundle formulation of Wu and Yang. Let us try to understand the relationship between these different formulations.

First note that in one important sense, the Dirac string singularity doesn't really "go away" when one employs the Wu-Yang construction as discussed above. In effect, the two situations are mathematically equivalent. As an analogy, consider a function in the complex plane, such as $\log z$, that has a branch cut singularity. One can view this function as being single-valued "without singularity" by representing the complex plane as a Riemann surface with multiple sheets. This is completely analogous to the Wu-Yang maneuver to "remove" the Dirac string and view the vector potential as a multiply connected field with components $\mathbf{A}^+$ and $\mathbf{A}^-$.[16]

Nevertheless, since the Dirac string can be moved anywhere (we saw, for instance, that it can be put on the positive or negative z-axes), Dirac realized that in some sense the string is "nonphysical." Furthermore, he showed that given the singularity of $\mathbf{A}$ (say, on the negative z-axis where $\theta = \pi$) and the fact that the wavefunction for a test particle (an electron) in the monopole field must be single-valued, then the following condition on $\Delta\gamma$ (see equation (7)) must obtain:

$$\Delta\gamma = 2N\pi.$$

[16]See (Moriyasu, 1983, p. 156).

If we let θ go from 0 to π in equation (7), then the change in phase $\Delta\gamma = 4\pi qg/\hbar c$.[17] This straightforwardly yields the Dirac quantization condition:[18]

$$qg = \frac{1}{2}N\hbar c. \tag{8}$$

(Wu and Yang, 1975) show how one can derive the Dirac quantization condition using the fiber bundle construction described above. In this context the quantization condition (8) emerges as a result of requiring the single-valuedness of the transition functions given by equation (6).

Were the Dirac quantization condition (8) not to obtain, it would be possible to discover the location of the Dirac string by performing AB-type experiments. That is, we could perform two-slit experiments of the sort described in section 2 at different locations and look for phase shifts due to the flux through the string! So, only if the quantization condition holds, will the ideal solenoid/string be undetectable. As a result, a number of investigators (agreeing with Dirac) have held that the monopole string is "merely" a mathematical singularity and has no physical significance whatsoever. (See, for instance, (Coleman, 1983; Ryder, 1985).)

But the issue here is really what counts as "physically significant" and whether there is really a distinction to be made *in this context* concerning the difference between the merely mathematical singularity of the Dirac string and a genuine singularity of physical significance. If there is no real distinction to be made, then we should begin to think about how "merely" mathematical/geometrical structures can play roles in *explaining* genuinely physical phenomena.

The AB effect is definitely detectable, and it remains for us to see to what extent there is an analogy between the solenoid in the AB experiment and the Dirac string. To get to this recall that Berry showed how the AB effect can be understood as an instance of anholonomy. Here I would like briefly to discuss the AB anholonomy in the language of fiber bundles.

One can think of classical electromagnetism for a static magnetic field as a $U(1)$ fiber bundle over the base space $\mathbb{R}^3$ or over Minkowski spacetime. If the electromagnetic field strength is zero at some point in $\mathbb{R}^3$ (the base space) then the connection on the bundle above that point will be flat. Furthermore, since each loop in the base space is contractible to a point (the space is

[17] The solid angle subtended by the sphere is 4π. "g" is the monopole charge.

[18] See for example (Ryder, 1985, pp. 413–417) for a discussion.

simply connected), the bundle theoretic nature of classical electromagnetism is trivial. There is, in other words, no anholonomy.

In a sense the AB effect demonstrates that this conclusion is too hasty. Recall that when there is current in the solenoid of the AB experiment, there is still no magnetic field (zero field strength) in any region outside the solenoid. Now, in the context we are considering—namely, that of classical electromagnetism with quantum mechanical particles—we can think of the solenoid as being impenetrable. For instance, the magnetic field inside the solenoid can have no effect whatsoever on any particle outside the solenoid. Thus, it is natural to idealize the solenoid to be infinitely long and infinitely thin. The line of magnetic flux, likewise, will be infinitely thin. Geometrically, one can then think of the base space as $\mathbb{R}^3 - \{z\text{-axis}\}$; or if we confine our attention to the plane of the experiment in figure 6, the base space is $\mathbb{R}^2 - \{\text{origin}\}$.

As we've seen in section 2, the magnetic field (and hence, the electric field) are both zero outside the solenoid. But if there is current in the solenoid, the vector potential $\mathbf{A}$ is nonzero in the region outside—which is the configuration or base space. On the other hand, the *curl* of $\mathbf{A}$, $\nabla \times \mathbf{A} = \mathbf{B} = \mathbf{0}$. We now have a $U(1)$ bundle over a nonsimply connected base space: $\mathbb{R}^2 - \{\text{origin}\}$. This fact is responsible for the AB effect. Despite the flatness of the connection on the bundle over this base space, there will be anholonomy.

To better understand the analogy between the monopole (with the Dirac string) and the AB effect we need to introduce the notion of a *contractible space*. A space $\mathbf{X}$ is contractible if there exists a family of maps

$$H_t : \mathbf{X} \to \mathbf{X} \quad \text{for} \quad 0 \leq t \leq 1$$

such that H_0 is the identity (i.e. for all $x \in \mathbf{X}$, $H_0(x) = x$) and such that H_1 is the constant map (i.e. there is a fixed point $p \in \mathbf{X}$ such that $H_1(x) = p$ for all $x \in \mathbf{X}$). The family H_t is called a contraction and it simultaneously shrinks all loops. Any space $\mathbf{X}$ for which such a contraction exists will therefore be simply connected. On the other hand, not all simply connected spaces are contractible. $\mathbf{S}^2$ is an example. While every loop on $\mathbf{S}^2$ is contractible to a point, you cannot shrink the entire space to a point which one could do *if* the space were contractible. You will get hung up on one of the poles! It is this feature that allows there to be anholonomies in round trip circuits on $\mathbf{S}^2$.

Now consider once again the monopole, with its simply connected base space $\mathbf{S}^2$ and fiber $U(1)$. We've seen that this is a nontrivial (Hopf) bundle—

it admits anholonomies. No nontrivial bundle over $\mathbf{S}^2$ admits a flat connection. Suppose though that we label the "south" pole of $\mathbf{S}^2$ as s. Even the nontrivial Hopf bundle becomes trivial if it is restricted to the complement of $\{s\}$:

$$\mathbf{S}^2 - \{s\}.$$

The Dirac string singularity is just a manifestation of the fact that this trivialization is not a possibility for the monopole bundle. Likewise, the AB solenoid idealized as a line removed from spacetime is a manifestation of the nontrivial bundle nature of electromagnetism on $\mathbb{R}^3 - \{z\text{-axis}\}$.

3.5 Anholonomies: A Distinction

It is worth pointing out a distinction between types of anholonomies.[19] It is not unreasonable to treat this as a distinction between topology and geometry. There are "topological phases" and "geometric phases." Consider again the Möbius strip—a fiber bundle with a *flat* connection. As we saw in section 3.1, a circuit around the nonsimply connected base space takes any point x on a fiber to $1 - x$. This is just a reflection about the midpoint on the fiber. Suppose we let the midpoint of the fiber be the origin, with points below it having negative values and points above it having positive values. Then the result of looping around the base space is just a change of sign $x \to -x$. The values of the anholonomy are, therefore, discrete—just a change of sign—and change *discontinuously* as a function of the shape of the circuit. The phase takes on discrete values as the circuit is completed. Call such an anholonomy "topological".

There are a number of instances where such purely topological phases have been show to be important physically. For instance, (Berry and Robbins, 1997, 2000) recently have provided an explanation for the spin-statistics connection for indistinguishable particles (the Pauli exclusion principle) in terms of a topological phases which is also a sign change associated with circuits in the projective plane.

By contrast, consider a connection with *curvature*, say, the Hopf (or monopole) bundle. Consider some loop in the base space S^2. Equation 7 tells us that the phase or anholonomy depends *continuously* on the shape of the circuit as it is proportional to the solid angle subtended by the circuit. Call this sort of anholonomy or phase "geometrical".

[19] Thanks to Michael Berry for helping me get clearer about this difference.

Oddly enough, the AB effect is a kind of hybrid topological and geometrical phase. The phase or anholonomy depends continuously on the flux in the solenoid, but (as in the case of the Möbius strip) it depends discontinuously upon the shape of the circuit. For example, two loops around gives an anholonomy twice that of one loop around for constant magnetic flux.

3.6 Summary

This section has aimed to do three things. First, it presents (in a relatively simple fashion) the theory of fiber bundles and applies this mathematical framework to some of the examples we have been considering. Second, it offers some brief remarks about how one might think of the fiber bundle formulation of various problems. We may consider them to be instances of a program of reduction or reconstruction. And third, it examines the analogy between the AB effect with its "physically real" (though highly idealized) solenoid, and the Dirac monopole with its "merely mathematical" solenoid—the Dirac string. The upshot of this discussion is that both the solenoid, idealized as a line missing from spacetime, and the Dirac string indicate the appropriateness of a bundle formulation of the various phenomena. More details about this will be offered below in section 5.2.

In the next section, I would like to return to Pancharatnam's anholonomy discovered in experiments concerning the polarization states of classical light. This, recall, is naturally represented in terms of a nontrivial bundle over the Poincaré sphere.

4 Polarization AB effect

In section 1.5 I noted that Pancharatnam discovered a surprising anholonomy in the phase of the light wave as the light is taken on a circuit in this polarization space. It turns out that, geometrically, this problem has the same structure as the magnetic monopole discussed above. In other words, a polarized light wave requires for its full specification the Hopf bundle over the Poincaré sphere. In this section I will more fully describe the representation of polarization states of classical light. Berry has claimed that the anholonomy discovered by Pancharatnam is "precisely analogous to the phase shift later predicted by Aharonov and Bohm ...." (Berry, 1987, p. 1404) I will discuss this claim.

4.1 The Spinor Representation of Polarization States

The state of a fully polarized wave ψ can be written as a two component *spinor*—a column vector of two complex elements which, in general, are functions of time:

$$\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix}.$$

These numbers represent the amplitudes for the wave to be in two orthogonal base states. For instance, suppose we have a wave propagating in the $\vec{z}$ direction. Choose for polarization base states the linearly polarized states $|X\rangle$ and $|Y\rangle$ in which the electric vector $\mathbf{E}$ vibrates, respectively in the $\vec{x}$ and $\vec{y}$ directions. Then in state $|X\rangle$, $\mathbf{E} = \vec{x}Ee^{i\omega t+\phi}$ and in $|Y\rangle$, $\mathbf{E} = \vec{y}Ee^{i\omega t+\phi}$.

We will restrict our attention to waves of unit intensity and to unitary (norm preserving), "polarization" transformations of these waves.[20] Not surprisingly a geometric interpretation (involving the Poincaré sphere) is available for the column vector representation $\psi = (c_1(t), c_2(t))$. In spherical coordinates[21] (and supposing time independence) the polarization state of a wave can be written (up to phase) as

$$\begin{pmatrix} \cos\frac{\theta}{2} \\ \\ \sin\frac{\theta}{2}e^{i\phi} \end{pmatrix}.$$

Consider two waves

$$\psi_1(\theta_1, \phi_1) = \begin{pmatrix} \cos\frac{\theta_1}{2} \\ \\ \sin\frac{\theta_1}{2}e^{i\phi_1} \end{pmatrix} \tag{9}$$

$$\psi_2(\theta_2, \phi_2) = e^{-i\alpha}\begin{pmatrix} \cos\frac{\theta_2}{2} \\ \\ \sin\frac{\theta_2}{2}e^{i\phi_2} \end{pmatrix}, \tag{10}$$

where α is a relative phase between the two waves.

If the two waves ψ_1 and ψ_2 are allowed to interfere, the intensity of the combined wave is proportional to:

$$|\psi_1 + \psi_2|^2 = 2 + 2Re\left(\psi_1^*\psi_2\right). \tag{11}$$

[20] See note 7.

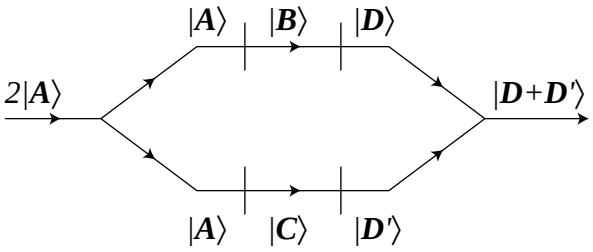
[21] θ and ϕ are, respectively, the polar angle and azimuth of a point on the surface of a sphere of unit radius. See figure 9.

Recall that Pancharatnam defines two waves to be in phase when their intensity is a maximum. This means that ψ_1 and ψ_2 are in phase if and only if

$$\begin{aligned} Re\,(\psi_1^*\psi_2) &> 0 \\ Im\,(\psi_1^*\psi_2) &= 0. \end{aligned} \tag{12}$$

This is Pancharatnam's connection, which as noted above, defines a notion of distant parallelism between polarization states. It is isomorphic to the Dirac connection on the monopole.

(a)



(b)

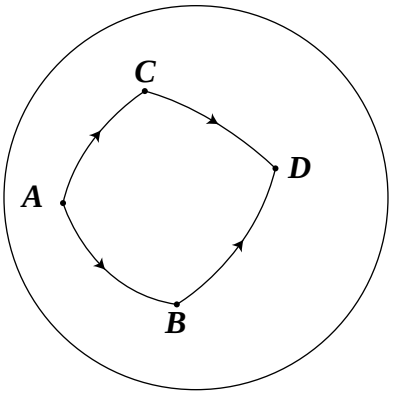


Figure 11: AB Effect on the Poincaré Sphere: (Berry, 1987, p. 1404) (a) Polarization History of a Light Wave. (b) Circuit on the Poincaré Sphere.

The phase difference α is gotten by substituting the column vectors (9)

and (10) into these last formulas.[22] This phase represents the "difference between the 'component of $|\psi_1\rangle$ along $|\psi_2\rangle$' and $|\psi_2\rangle$. In other words, α represents the phase difference between (i) the wave resulting from the passage of the wave in state $|\psi_1\rangle$ through a polarizer that allows only the state $|\psi_2\rangle$ to pass through and (ii) the reference wave in state $|\psi_2\rangle$." (Bhandari, 1997, p. 14)

The 1/2 solid angle result mentioned in section 1 follows from Pancharatnam's connection. That is, if through the use of various optical devices, a pure state of polarization is forced to trace out a closed circuit C on the Poincaré sphere, then the phase difference between the initial and final states—the anholonomy—is equal to 1/2 the solid angle subtended by C at the center of the sphere.

The analogy with the monopole and the Hopf bundle is precise: We imagine the existence of an abstract monopole (of strength -1/2) centered at the origin of the Poincaré sphere. (For details see (Berry, 1987; Brosseau, 1998; Morandi, 1992).) On this conception, the anholonomy due to a circuit C is (like the Dirac monopole) proportional to the flux through the surface of the sphere by the field generated by a monopole at its center.

(Berry, 1987) has called the Pancharatnam anholonomy the "Aharonov-Bohm" effect on the Poincaré sphere. He says that the

> result is precisely analogous to the phase shift later predicted by Aharonov and Bohm, according to which two electron beams develop a phase shift proportional to the magnetic flux they enclose. For polarized light the analogue of magnetic flux is the solid angle of the polygon [see figure 11] on the Poincaré sphere. This is also a flux, namely that of an abstract monopole of strength -1/2, situated at the centre of the sphere. (Berry, 1987, p. 1404)

In both situations, the phase shifts are represented by nontrivial topological features of their respective "state" spaces. A point of disanalogy is the fact, noted earlier in section 3.5, that the AB effect has both "topological" and "geometrical" aspects. In addition, a common attitude—one taken by Aharonov and Bohm but which I believe is ultimately mistaken—is that the vector potential (which changes when the flux changes) must be considered a *real* physical field. I think that the background intuition for this position depends upon the fact that the effect takes place in space or spacetime where

[22] See (Brosseau, 1998, pp. 131–135) for details of the calculations.

issues about causality come to the fore. It would be odd indeed to hold that the connection on the *abstract* state space—the Poincaré sphere—is *causally* responsible for Pancharatnam's phase. I think that this ought to give us pause. Perhaps the question about the "reality" of the connection is simply misguided. This is the subject of the next section.

5 Interpretations and Conclusions

5.1 Belot's Taxonomy of Interpretations

In section 2 I noted that an instructive way to think about classical (vacuum) electromagnetism is in the framework of gauge theories. Belot develops this point of view and offers three possible interpretations of electromagnetism from this perspective. On the one hand, one can think of the vector potential $\mathbf{A}$ as a real physical field. Suppose the electric field has value $\mathbf{E}$. The *state* of the electromagnetic field will then be different for different values of the vector potential. According to Belot

> [t]his gives us a literal, hence indeterministic, interpretation of the gauge-theoretic formulation of electromagnetism: each pair $(\mathbf{A}, \mathbf{E})$ satisfying $div\, \mathbf{E} = 0$ represents a distinct dynamical state of the ether, and two solutions, $\mathbf{A}(t)$ and $\mathbf{A}'(t) = \mathbf{A} + grad\, \Lambda(t)$, for the same initial data represent two physically distinct histories of the ether .... (Belot, 1998, pp. 541–542)

The indeterminism—more than one future state following from the same initial state—depends upon the fact that the Hamiltonian or dynamical law, because it is a function of the electric and magnetic fields only, is insensitive to the different values of the vector potential.[23] In other words, Belot assumes that on this interpretation we change only the notion of the state of the system, but leave the dynamics unchanged. In a sense this is the natural thing to want to do. After all, we are trying to give an interpretation of classical Maxwellian electromagnetism.

Despite this, it isn't clear to me that this is an appropriate assumption. The very notion of the state of a system (or field) cannot, in general, be divorced from the nature of the laws governing the system's (or field's) evolution. States and laws are correlative. What we take to be the relevant

[23] See equations (1) and (2).

state properties depends upon what we count as laws; and conversely, the nature of the laws is constrained by what features of the world are important for characterizing a system's state.[24] If, as the interpretation requires, the vector potential is a real physical field, then why think that the dynamics should be insensitive to the differences in values of the potential? At least some further argument seems to be required here.[25]

Something analogous, it seems would happen were we to consider a similar interpretation of the quantum state. We would have to treat each wavefunction $|\psi\rangle$ that differs from another $|\psi^{\alpha}\rangle$ by multiplication by a $U(1)$ phase—a gauge transformation—as a *distinct* state. However, because the Schrödinger evolution is insensitive to such phase differences, the same initial "state" can give rise to different future states. Hence, the interpretation renders the theory indeterministic. Note that this indeterminism is completely different from that typically associated with quantum mechanics. Typically the theory is taken to be indeterministic because one seems forced to give a probabilistic interpretation of the wavefunction. Once again, though, we might ask why it is, if the different phase values *are* to be physically relevant—that is, if they figure in the proper state description, that the dynamics shouldn't reflect this fact as well.

It seems very strange to me to maintain that a *dynamical* theory can be rendered indeterministic simply by opting for a literal interpretation of its gauge structure. It is considerably more plausible to note that the possibility of a literal interpretation simply indicates that the state description does not match up in an appropriate way with the dynamical law(s). There is no reason to take the different gauge related quantities to be relevant to the systems state if one is sure that the dynamical law is correct. On the other hand, if one has independent reason to take the gauge related quantities to be genuine or "real", then surely ones dynamics ought to be changed to reflect this fact. My view is that this is really what motivates the "traditional" interpretation for gauge theories, and not any real worry about indeterminism.

Belot's second possible interpretation for electromagnetism is the "traditional" interpretation in which *only* the electric and magnetic (or electromagnetic) fields are physically real. This is a gauge invariant interpretation

[24] I have discussed the relationships between laws and states at some length in my dissertation (Batterman, 1987).

[25] Note that the AB effect may provide some such further reasons. But, without such physical reasons there appears to be no good argument to move to a literal interpretation since the dynamics of the theory doesn't support it.

since the value of the magnetic field at any point in physical space is the same for any vector potentials related by the gauge transformation (equation (2)). A question of importance is the exact nature of the gauge invariance. Belot considers whether it is "simply gauge-invariant" or coarse-grained gauge-invariant." A simply gauge-invariant interpretation is one in which there is a bijection between gauge orbits and states of the electromagnetic field. A coarse-grained gauge-invariant interpretation is one in which the relation between gauge orbits and states may be many-one. Belot notes that if the topology of physical space is simply connected then the traditional interpretation is simply gauge invariant. But, if the topology is nontrivial (nonsimply connected), "...the traditional interpretation counts as coarse-grained gauge-invariant ...." (Belot, 1998, p. 543)

In a nutshell, for the case of electromagnetism, coarse-grained gauge invariance amounts to a claim to the effect that there is some information that is not encoded in the equivalence:

$$\mathbf{B} = \nabla \times \mathbf{A} = \nabla \times \mathbf{A}'. \tag{13}$$

In other words, despite this equivalence, $\mathbf{A}$ and $\mathbf{A}'$ are not related by the "gauge transformation" (2). Thus, on a coarse-grained gauge interpretation distinct gauge orbits $[\mathbf{A}] \neq [\mathbf{A}']$ can correspond to the same magnetic field. Such a situation is possible only if there is *no* function χ such that

$$\mathbf{A}' = \mathbf{A} + \nabla\chi.$$

This is exactly the situation that obtains in the AB effect.

The third interpretation restores the bijection between gauge orbits and states of the electromagnetic field in effect by taking into consideration the information missing in equation (13). It requires a reinterpretation of the notion of a gauge orbit. This interpretation has us in effect treat the phase space of electromagnetism as a pair consisting of the divergence free electric field and the gauge invariant "holonomy maps" defined in the following way. The "holonomy" around a loop γ in space given a vector potential, $\mathbf{A}$, is specified by the following gauge invariant integral:

$$h(\gamma) = e^{\oint i\mathbf{A}(\mathbf{x})\mathbf{dx}}. \tag{14}$$

Gauge invariance means that $h(\gamma) = h'(\gamma)$ if and only if $\mathbf{A}$ and $\mathbf{A}'$ are on the same gauge orbit. One can construct holonomy maps from the loops in

real space into the complex numbers of modulus one ($U(1)$). This becomes the configuration space for electromagnetism and together with its conjugate electric field at each point, we have a simply gauge invariant interpretation of electromagnetism even in those situations where real space is multiply connected.

Notice how this last interpretation can be understood in the fiber bundle formalism discussed in section 3. Having noted the presence of anholonomies in the AB effect, we engage in a program of *reconstruction* to provide a representation of these anholonomies. This involves treating the full space for representing electromagnetism as a nontrivial bundle over the base space—physical space.

As Belot and others (e.g., (Healey, 1997)) have noted, taking this third interpretive strategy for gauge theories brings with it a host of other problems. Since the full phase space now requires that we specify the (an)holonomies for each and every loop about a point in physical space, the state of the electromagnetic field is now rendered nonlocal. In order to say what the state of the field is at a given point we must refer to properties of loops in real space—loops that go through regions of space arbitrarily far from the given point.

Now Belot argues that the traditional, simply gauge invariant interpretation of classical electromagnetism is naturally to be preferred. And that

> [w]ithin the realm of classical physics, ..., [the traditional interpretation] is vindicated—there are no phenomena which allow one to distinguish between two gauge orbits $[A]$ and $[A']$ which correspond to the same magnetic field. Thus, there are no grounds internal to electromagnetism upon which to criticize the traditional interpretation. (Belot, 1998, pp 544–545)

And, of course, he is right about this. It is only when we start to think of electrons as quantum mechanical that we run into the various interpretive difficulties.

5.2 The Importance of Geometry

Nevertheless, I think that it is important to recognize that the interpretive moves that seem to be required to deal with the AB effect are also required to account for phenomena entirely "within the realm of classical physics." Naturally, I'm talking here about Pancharatnam's phase—the "polarization AB

effect" (section 4). In addition, the various examples discussed in section 1 (including the falling cat and parallel parking) likewise demand that we consider anholonomies as essential for the characterization of the phenomenon in question.

The ubiquity of anholonomy in both classical and quantum physics should lead us to consider certain methodological and ontological questions. Take the latter group of issues first. As I noted the AB effect has been much discussed in the literature largely because it seems to force one to adopt one of two rather strange interpretations. On the one hand, we can take the line, advocated by Aharonov and Bohm themselves and promulgated by Feynman, that the vector potential $\mathbf{A}$ is a *real physical field* thereby maintaining some kind of locality. The phase shift is to be explained by the local action of the vector potential on the wavefunctions of the electrons. This is a problematic interpretation since the vector potential is, as we've seen, not gauge invariant and there is some reason to hold that only gauge invariant quantities are genuine physical quantities. On the other hand, we can adopt Belot's third "holonomy" interpretation, noting as above, that the quantity in equation (14) is gauge invariant. But then this forces us to give up on locality.

There is some debate in the literature (see (Healey, 2001)) about whether one should adopt the holonomy interpretation or whether one should be a realist (= substantivalist) about the fiber bundle formulation of the theory. Healey opts for the nonlocal holonomy interpretation after arguing that the realist interpretation on which gauge potentials are *real* connections on fiber bundles fails to live up to the promise that the gauge potentials be locally defined objects.

I think that this debate is largely a red herring. Nothing like this seems to be relevant in any of the "classical" cases of anholonomy we have considered. The reason the debate rages at all, it seems to me, has to do with the nature of the base space, which in the cases Healey discusses, is "real" space or spacetime. This space is unlike the "abstract" space of shapes in the case of the falling cat or the "abstract" space of polarization states—the Poincaré sphere—in the case of the polarized light. The relevant difference is that consideration of spacetime carries with it many metaphysical commitments that are completely absent in the other cases. We need to worry about nonlocality, separability, etc., because of the odd relationships quantum mechanics has to relativity theory. But none of this arises in the other cases where surprising anholonomies appear.

Related ontological issues arise when we try to understand the nature of the "merely mathematical" Dirac strings in the magnetic monopole and in the abstract polarization monopole at the center of the Poincaré sphere. What sort of reality does one want to attach to these abstract features? How should one want to understand the relationship between the Dirac string and the real flux that is present in the AB effect? This is especially relevant once one recognizes that most discussions of the AB effect very quickly idealize the solenoid to an infinite line in space or spacetime. The flux, in this idealization, just is the abstract topological property of having space or spacetime be nonsimply connected.

One might ask[26] whether, e.g., it is obvious that the polarization monopole at the center of the Poincaré sphere is purely abstract and represents no real physical structure. In one sense, I think the answer here is: "Of course, it represents some physical structure." But in saying that I mean only that to speak of the monopole is simply to speak of a nonflat connection on the surface of the Poincaré sphere that explains and represents the Pancharatnam phase. Whether we should reify the monopole itself, or treat it as a purely formal/abstract object is, I believe, irrelevant. The issue is whether the idealizations—polarization monopole and nonsimply connected space in the AB effect—do better explanatory work than some less idealized description. I believe that the idealized descriptions do, in fact, do a better job. And, this leads us to the methodological questions.

From my perspective the most interesting issues *are* methodological and concern the *explanatory role* played by these abstract geometrical and topological features. The most remarkable feature of Berry's discovery of the geometric phase in quantum mechanics is the fact that topological and geometric structures of an *abstract* space of parameters can have observable, physical, but obviously noncausal "effects."

If one wants to understand the interference behavior of polarized light as the beam is taken through a series of polarizers and returned to its initial state, one must represent the polarization states using the full apparatus of the (nontrivial) Hopf bundle over the Poincaré sphere. Likewise, if one wants to understand how a cat can right itself when it is dropped with zero angular momentum, one must investigate its trajectories in an abstract space of shapes. The full understanding of such varied phenomena demands reference to nontrivial fiber bundles.

[26]In fact, an anonymous referee did ask.

For example, as we have seen in sections 1.5 and 4, Pancharatnam discovered experimentally the intransitivity of phases for classical polarized light. His physical criterion for when two beams in different polarization states are in phase—equations (11) and (12)—defines a connection on the Poincaré sphere. In effect, Pancharatnam was engaged in a program of reconstruction.[27] The modern characterization of this program is that he needed to appeal to the full, nontrivial, Hopf bundle in order to account for the phase he discovered. One might, I suppose, eschew this fiber bundle representation in favor of a set of statements about the polarization properties of closed loops on the Poincaré sphere, though I don't know why one would want to here.

In opting for the nonlocal holonomy interpretation (Belot's third interpretation) over a realist/substantivalist interpretation of fiber bundles in the context of the AB effect, Healey is in effect taking this latter line. The idea is that we do not need to (should not need to) appeal to a fiber bundle whose base space is nonsimply connected. Instead, keep the base space as simply connected Minkowski spacetime, and explain the AB effect by appeal to the effects of electromagnetic properties of closed loops in that spacetime. Again, I am really not sure what the advantage of this is. And, as I will now try to show, I think that there are some distinct disadvantages.

It is true that the gauge invariant content of electromagnetism is completely specified by the set of holonomy maps determined using equation (14) as discussed in section 5.1. And it is true that this works regardless of the topological structure of the base space—namely, spacetime. Nevertheless, I think that it leaves out important explanatory information that the fiber bundle formulation makes explicit. This is the explicit information concerning the topological nature of the base space. Healey's defense of his nonseparable holonomy thesis—that gauge potential are nonseparable holonomy properties—depends (in part) on the assumption that topological obstructions such as magnetic monopoles are *not* present. He says that

> [I]n the absence of magnetic monopoles one can compose a given loop enclosing a surface out of tinier and tinier loops around points on that surface. In the limit, the holonomy properties of a finite loop are determined by those of any infinitesimal loops that that compose it in this way. This gives what might be called

[27] See section 3.2.

> the loop supervenience of holonomy properties. (Healey, 2001, p. 450)

But where there are topological obstructions, say, monopoles, flux trapping in a superconductor[28], and (perhaps[29]) the AB effect, this justification will fail. Topological considerations play a crucial role.

It seems to me that for a full understanding of these anholonomies, one needs to appeal to the topology and geometry of the base space. The fiber bundle formulation makes that topology explicit. As I said above, I think that the question of the reality of the fiber bundle formulation versus the reality of the holonomy interpretation is largely an artifact of the discussion of cases in which the base space is real space or spacetime. If we take seriously the idea that topological features of various spaces (abstract or real) can play an explanatory role, then we can see how to unify a set of apparently diverse phenomena—namely all of the examples discussed in the pages above as well as many others.

In addition, appeal to topological features of the sort we are discussing can provide different and better explanations of the phenomena than one might otherwise have if one fails to mention them explicitly. For instance, one might think that the full, complete explanation for why the cat lands on its feet when dropped upside down is to be had by a detailed and complete Newtonian account of the forces acting on its various parts as it twists itself on the way down. What is the point of referring to its abstract shape space and to the mechanical connection on the fiber bundle?[30] I take it that in the context of the AB effect this question is analogous to the above question asking why we need the fiber bundle formulation given that specifying the holonomy properties of closed loops in spacetime will provide the explanation we are after.

The response to these questions involves pointing out that one thing we really want to understand is the ubiquity and universality of these types of phenomena. In the case of the falling cat, a question of interest concerns why, in general, cats behave in this fashion when dropped. Were we to explain every instance of this falling cat behavior by appeal to the detailed forces

[28]See (Moriyasu, 1983) for a discussion.

[29]I don't mean to beg any questions here. After all, part of what I'm trying to show is that it is most fruitful to treat the AB solenoid as an idealization that results in the multiple connectedness of the base space of a fiber bundle.

[30]This question was raised by an anonymous referee as well.

acting on the individual cats, we will achieve no answer to the question of why such behavior is generally to be expected.[31] In effect, each such account will be completely disjoint from the others and the general question about the ubiquitous pattern of behavior requires that we abstract from all the individual details of the particular cases. Such details simply get in the way. Referring to the mechanical connection on the bundle over the space of shapes provides the unification we desire. The geometric features enable us to understand the common behavior in a way that the individual Newtonian stories do not.

Similarly, in the AB effect, it appears that we will need to refer to different nonseparable holonomy properties for each case in which there is different flux running through the solenoid. The different cases are unified by the topological idealization of the solenoid as a string absent from spacetime which renders spacetime nonsimply connected. In this way we can understand why, for a given fixed magnetic flux, a loop that goes n times around the solenoid will have an anholonomy that is n times that of a loop that goes around once. This topological feature enables us to understand the common behavior in different AB experiments in a way that the individual appeals to nonseparable holonomy properties of closed loops in spacetime do not.

So, topological and geometric features of abstract (and real) spaces allow us to explain universal features of the world. This should not be too surprising if one recognizes that such explanations are quite prevalent in physics. For instance, the universality of critical phenomena is explained by appeal to topological features (fixed points, for instance) in an abstract space of Hamiltonians. (See (Batterman, 2002b) for detailed discussions.)

The acceptance of these types of explanations raises some additional worries about realism and about the role of idealizations. For example, how can it possibly be the case that appeal to an idealization such as the AB solenoid as a line missing from spacetime, provides a better explanation of genuine physical phenomena than can a less idealized, more "realistic" account where one doesn't idealize so severely? I've argued elsewhere (Batterman, 2002a) (and so will not rehearse the arguments here) that quite often, primarily when one is interested in explaining universal behavior, appeal to highly idealized models does, in fact, provide better explanations.

In sum, I think it is difficult to hold that the geometric and topological

[31] See (Batterman, 2002b, Chapters 3, 4) for a discussion of these types of explanatory why-questions.

features of the various spaces (particularly, in the cases such as the falling cat and polarized light) we have considered are causal features. Nevertheless, they play essential explanatory roles. If we recognize that similar abstract geometrical/topological properties are present in the AB effect, then it seems we ought to bracket the explanatory problematic from the metaphysical debates that appear to be driving the discussions in the literature. Questions about the reality of gauge potentials just do not seem to arise in many/most of the examples we have discussed. The suggestion here is that such questions may not matter much either when it comes to understanding such quantum effects as the AB effect.

References

R. W. Batterman. *Irreversibility, Statistical Mechanics, and the Nature of Physical States*. Ph.D. Dissertation, 1987.

R. W. Batterman. Asymptotics and the role of minimal models. *The British Journal for the Philosophy of Science*, 53(21–38), 2002a.

R. W. Batterman. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford Studies in the Philosophy of Science. Oxford University Press, New York, 2002b.

Gordon Belot. Understanding electromagnetism. *The British Journal for the Philosophy of Science*, 49:531–555, 1998.

M. V. Berry. Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London*, A392:45–57, 1984.

M. V. Berry. The adiabatic phase and Pacharatnam's phase for polarized light. *Journal of Modern Optics*, 34(11):1401–1407, 1987.

M. V. Berry. Anticipations of the geometric phase. *Physics Today*, 43(12): 34–40, 1990.

M. V. Berry. Bristol anholonomy calendar. In R. G. Chambers, J. E. Enderby, A. Keller, A. R. Lang, and J. W. Steeds, editors, *Sir Charles Frank OBE FRS, an eightieth birthday tribute*, pages 207–219. Adam Hilger, Bristol, 1991.

M. V. Berry and J. M. Robbins. Indistinguishability for quantum particles: Spin, statistics and the geometric phase. *Proceedings of the Royal Society of London*, A453:1771–1790, 1997.

M. V. Berry and J. M. Robbins. Spin-statistics connection and commutation relations. *American Institute of Physics*, CP545:3–15, 2000.

Rajendra Bhandari. Polarization of light and topological phases. *Physics Reports*, 281:1–64, 1997.

D. Bohm and Y. Aharonov. Significance of electromagnetic potentials in the quantum theory. *The Physical Review*, 115(3):485–491, 1959.

Christian Brosseau. *Fundamentials of Polarized Light: A Statistical Optics Approach*. John Wiley and Sons, New York, 1998.

Hernán Cendra, Jerrold E. Marsden, and Tudor S. Ratiu. Geometric mechanics, lagrangian reduction, and nonholonomic systems. In B Enguist and W. Schmid, editors, *Mathematics Unlimited—2001 and Beyond*, pages 221–273. Springer-Verlag, Berlin, 2001.

Sidney Coleman. The magnetic monopole fifty years later. In Antonino Zichichi, editor, *The Unity of the Fundamental Interactions*, volume 19 of *Proceedings of the Nineteenth Course of the International School of Subnuclear Physics*, pages 21–117. Plenum Press, 1983.

Richard Healey. Nonlocality and the Aharonov-Bohm effect. *Philosophy of Science*, 64:18–41, 1997.

Richard Healey. On the reality of gauge potentials. *Philosophy of Science*, 68(4):432–455, 2001.

Stephen Leeds. Gauges: Aharonov, Bohm, Yang, Healey. *Philosophy of Science*, 66:606–627, 1999.

Richard Montgomery. Gauge theory of the falling cat. In Michael J. Enos, editor, *Dynamics and Control of Mechanical Systems: The Falling Cat and Related Problems*, volume 1 of *Fields Institute Communications*, pages 193–218. American Mathematical Society, Providence, Rhode Island, 1993.

Giuseppe Morandi. *The Role of Topology in Classical and Quantum Physics*, volume m7 of *Lecture Notes in Physics*. Springer-Verlag, Berlin, 1992.

K. Moriyasu. *An Elementary Primer for Gauge Theory*. World Scientific, Singapore, 1983.

S. Pancharatnam. Generalized theory of interference, and its applications: Part I. Coherent results. *The Proceedings of the Indian Academy of Sciences*, 44(5):247–262, 1956.

W. T. Read. *Dislocations in Crystals*. McGraw-Hill, New York, 1953.

Lewis H. Ryder. *Quantum Field Theory*. Cambridge University Press, Cambridge, 1985.

Joseph Samuel and Rajendra Bhandari. General setting for Berry's phase. *Physical Review Letters*, 60(23):2339–2342, 1988.

T. T. Wu and C. N. Yang. Concept of nonintegrable phase factors and global formulation of gauge fields. *Physical Review D*, 12(12):3845–3857, 1975.