Chapter 3.

Aharonov-Bohm Effect and Geometric Phase

$$e^{i(q/\hbar c)\oint dx^{\nu}\left(A_{\nu}+\partial_{\nu}\zeta\right)}$$

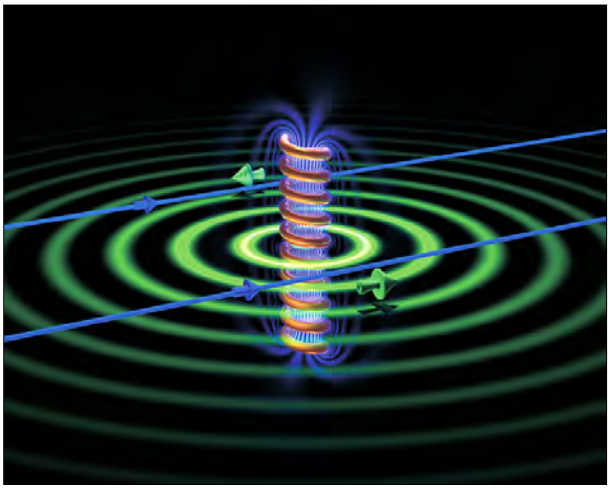
And all I wanted was a complex carrot.

"I have had my results for a long time but I do not yet know how I am to arrive at them."

Karl Friedrich Gauss

Contents

Introductory Comments



From Google Images

An interesting and far-reaching aspect of electrodynamics involves the potentials ϕ and $\vec{A}$ (together the four-vector A^{ν}) and their role in the quantum mechanics of charged particles. In this chapter we shall start by examining how phases of quantum mechanical particle waves are affected when these waves pass through regions in which the potentials are nonzero, whereas the force fields $\vec{E}$ and $\vec{B}$ are zero. The force fields and potentials are taken as static in what follows. The only time dependence is the one that arises from particle motion. This is assumed to be sufficiently slow that it can be disregarded insofar as force fields appearing in the particle's rest frame, as explained below.

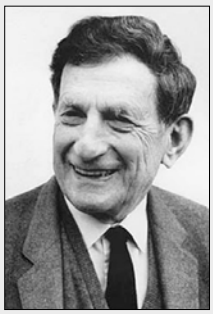
The Aharonov-Bohm effect is subtle. We shall see that the main idea is not difficult to grasp, though the devil is in the details. The effect relates directly to quantum electrodynamics (QED), and in so doing it is germane to the gauge field theory that underlies the standard model of physics. For example, it offers a glimpse into the weak and strong forces through analogy. For us, its significance includes the fact that it bears uncanny resemblance to the geometric phase that is encountered when a conical intersection of a polyatomic molecule comes into play. In analogy to the Aharonov-Bohm effect, a geometric phase accrues when the conical intersection is encircled through adiabatic transport (in the space of the nuclear degrees of freedom) around the intersection point. In fact, there is more than "uncanny resemblance;" there is registry. Consequently, the Aharonov-Bohm effect (hereafter referred to as the AB effect) is an excellent launching point for studies of conical intersections in molecules.

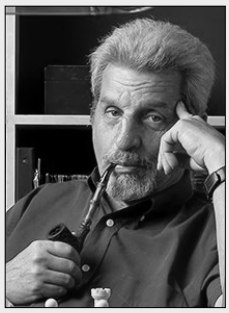
Like most scientific discoveries, the AB effect made its entrance amidst a number of precursor and complementary studies. It was not as original as it was "in the right place at the right time." Flux quantization in superconductivity, which is similar to the magnetic version of the AB effect, had been predicted by London, refined by others, and subsumed into the finished product that was delivered by Bardeen, Cooper, and Schrieffer in 1957: the BCS theory of superconductivity [26], for which they were awarded a Nobel Prize in Physics. Fairbanks verified flux quantization in 1961 [40]. Ehrenberg and Siday had published an equivalent result a decade earlier [41]. The prescient 1954 paper of Yang and Mills [42] generalized the U(1) gauge symmetry of quantum electrodynamics, as well as the AB effect, to SU(2), and did so five years before the paper of Aharonov and Bohm

was published. The 1954 Yang and Mills paper provided the mathematical foundation of what came to be known as the Standard Model of Physics.

The 1959 paper by David Bohm[1] and his graduate student Yakir Aharonov is about quantum mechanical effects that arise when particles pass through regions where the potentials ϕ and $\vec{A}$ are nonzero, whereas the so-called physical fields (the ones responsible for force) $\vec{E}$ and $\vec{B}$ are zero. It raised considerable interest and speculation, with debates that raged for years, some continuing to this day. I cannot explain why it created such a controversy, except to point out that such things happen from time to time in science.


David Bohm in his 30's


David Bohm later


Yakir Aharonov

1. The Main Issue

Let us begin with a short review of some material from Chapter 2. In classical physics, a particle that moves in vacuum in the presence of an electromagnetic field experiences a force described by the Lorentz force equation

$$\vec{F} = q\left(\vec{E} + \frac{\vec{v}}{c} \times \vec{B}\right), \tag{1.1}$$

where q is the charge of the particle, and $\vec{v}$ is its velocity.

For a long time it was believed that the potentials ϕ and $\vec{A}$ are simply a means of obtaining $\vec{E}$ and $\vec{B}$. The dogma – creed of the true believers – was that the potentials ϕ and $\vec{A}$ do not have physical significance of their own. Indeed, they serve only as conveniences for obtaining the physical fields $\vec{E}$ and $\vec{B}$, which are solely responsible for the electrodynamics forces. In fact, the potentials ϕ and $\vec{A}$ are not even unique.

---

[1] David Bohm (1917-1992) was a major figure in 20th century physics. He also led an extraordinary life: maligned during the McCarthy witch-hunt era (including being arrested and being fired by Princeton University) for his 1930's participation in political organizations such as the Young Communist League; moving from the U.S. to Brazil, then Israel, then England; contributing to the fledgling field of neuropsychology, including development of the holonomic theory of the brain; toward the end of his life suffering from severe depression. In 1959, he and his student Yakir Aharonov published their paper. Shortly thereafter, they learned that Ehrenberg and Siday had derived the same result a decade earlier. Consequently Bohm referred to the ESAB effect. This did not stick and the effect carries the names Aharonov and Bohm.

The fields $\vec{E}$ and $\vec{B}$ are obtained by differentiating the potentials, so they can be altered without affecting $\vec{E}$ and $\vec{B}$. These alterations are the gauge transformations discussed in Chapter 2. For example, with $\vec{B} = \nabla \times \vec{A}$, adding the gradient of a scalar to $\vec{A}$ does not affect $\vec{B}$.[2] Forces other than the electrodynamics force given by eqn (1.1) might be present as well. Their inclusion yields the equations of motion.

Alternatively, a Hamiltonian can be used to obtain the equations of motion. It is invariably expressed in terms of ϕ and $\vec{A}$. The equations of motion are obtained by using Hamilton's equations, and the resulting coupled equations contain derivatives of ϕ and $\vec{A}$. However, a bit of skilled manipulation converts these equations to ones in terms of $\vec{E}$ and $\vec{B}$. The Lorentz force given by eqn (1.1) is recovered, illustrating the consistency of the two approaches. Nothing could be more straightforward, at least from a classical perspective. That is what was believed a long time ago.

Quantum versus Classical Mechanics

Quantum mechanics, on the other hand, uses wave functions to describe particle waves and the dynamical processes they undergo. Sometimes the explicit use of wave functions can be avoided (or is not relevant, as with spin), but wave functions nonetheless lie at the heart of the theory. Phase is present from the outset: from the phase factor $\exp(-iE_n t/\hbar)$ of the n^{th} eigenstate, to the relative phases that enter the construction of a wave packet, to the geometric phases that are commonplace in molecular electronic structure. Even global phase plays a role. Invariance with respect to a global phase transformation yields charge conservation via Noether's theorem. Though not quantum mechanical *per se*, imagine where we would be were charge not conserved.

The wave function of a charged particle passing slowly through space (Fig. 1) can be expressed in such a way that the potentials ϕ and $\vec{A}$ appear in its phase. In the most elementary case, a free electron wave packet traveling in a region of constant potential energy, ($V = -e\phi$), acquires a phase $e\phi t/\hbar$. This follows directly from the gauge invariant Hamiltonian for an electron in vacuum: $H = p^2/2m - e\phi$, in which case the field free continuum eigenfunctions are each multiplied by $e^{ie\phi t/\hbar}$.

Figure 1(a) illustrates schematically how this phase can play a role in an experiment. A wave packet starts on the left. It is split (50/50) into components that proceed along the upper and lower paths, which pass through perfectly conducting cylinders. The upper and lower wave packets remain coherent with respect to one another. The voltages ϕ_1 and ϕ_2 are each held at zero until the packets are well inside the cylinders, at which time they are set at fixed values, say $\phi_1{}^0$ and $\phi_2{}^0$. Before the packets leave the cylinders, the voltages are turned back to zero. The upper and lower packets each acquire a phase that is brought about by the voltages $\phi_1{}^0$ and $\phi_2{}^0$. These phases are $e\phi_1{}^0\tau/\hbar$ and $e\phi_2{}^0\tau/\hbar$, where τ is the time that the voltage is on. Clearly the phases differ if $\phi_1{}^0 \neq \phi_2{}^0$.

[2] Speaking of uniqueness, note that $\vec{E}$ and $\vec{B}$ have a total of six components, whereas ϕ and $\vec{A}$ have a total of only four components.

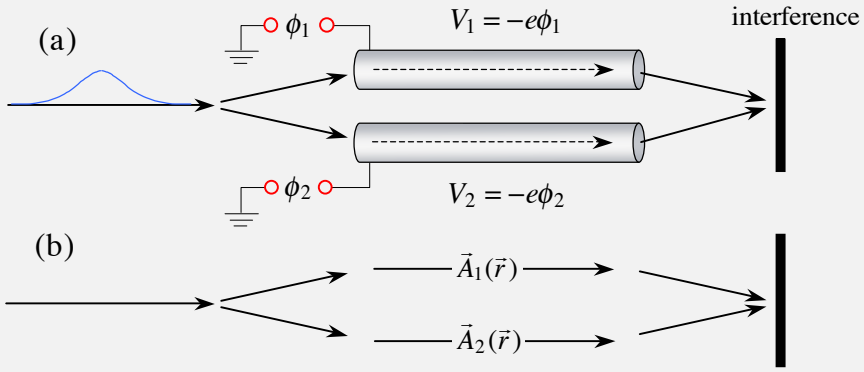


Figure 1. (a) An incident wave packet is split into two coherent wave packets that pass through perfectly conducting cylinders: regions where $\vec{E}$ is always zero aside from minor (unimportant) edge effects in the entry and exit regions. The electric potentials ϕ_1 and ϕ_2 are slowly turned on after the packets have entered the cylinders, reaching steady values of $\phi_1{}^0$ and $\phi_2{}^0$. They are slowly turned off before the packets exit the cylinders. Each of the packets acquires phase, and they are made to recombine at the screen labeled interference. The interference pattern changes as the difference between the potentials is varied. (b) Waves pass through regions where $\vec{B} = 0$, whereas $\vec{A}_1(\vec{r}) \neq 0$ and $\vec{A}_2(\vec{r}) \neq 0$. Of course, there exists some region where $\vec{B} \neq 0$, but the waves do not access this region.

The electric field, though it is always zero where the wave function is present, is not zero everywhere. It is nonzero in a region of space that the wave function is not allowed to enter. Instead of experiencing $\vec{E}$ directly (the local $\vec{E}$ field experienced by the wave function is zero) the particle experiences the potential ϕ associated with $\vec{E}$.

The situation shown in Fig. 1(b) is similar in spirit. Waves that follow the upper and lower paths acquire phases as they pass through the $\vec{A}_1(\vec{r}) \neq 0$ and $\vec{A}_2(\vec{r}) \neq 0$ regions, as discussed below. The sketch in Fig. 1(b) is ambiguous, however, because it says nothing about the location of the $\vec{B} \neq 0$ region. We will see that if it lies inside the region enclosed by the arrows interference is affected by the strength of $\vec{B}$. On the other hand, if it lies *outside* this region, varying $\vec{B}$ has no affect on the interference pattern.

The electric version of the AB effect is not as easily verified experimentally as the magnetic version. Perhaps for this reason, the effect that derives from the presence of the vector potential $\vec{A}$ is usually referred to as the AB effect. It is worth noting that there must be both magnetic and electric versions. As we know, in special relativity, electric and magnetic fields are transformed into each other through Lorentz boost. Likewise, components of the four-vector A^ν are mixed through Lorentz boost. Though it is interesting to ponder the electric and magnetic versions, both separately and together, here we shall consider just the magnetic one.

It is noteworthy that quantum mechanical waves respond to the potentials, which themselves are not gauge invariant. For example, if $\vec{A}$ is changed by the addition of $\nabla\zeta$, where ζ is a scalar function, the quantum mechanical wave must change accordingly in order that the gauge transformation: $\vec{A} \to \vec{A} + \nabla\zeta$, has no physical consequence. This

feature, in which $\vec{A}$ appears in a phase along a trajectory, distinguishes classical and quantum particles. As you might imagine, everything works out properly.

2. $\vec{E}$ and $\vec{B}$ versus ϕ and $\vec{A}$

When a particle wave packet passes through a region where $\vec{E} = 0$ and $\vec{B} = 0$, the Lorentz force is zero, so there cannot be an effect that has a classical counterpart.[3] Quantum mechanically, however, relative phases of wave packets can be manifest in interference phenomena that have no classical counterparts. Consequently, we need to look carefully at the roles played by potentials, for example, in the context of the Schrödinger equation.

The debate about which pair is more fundamental: $\vec{E}/\vec{B}$ or $\phi/\vec{A}$, will be avoided. The notion that one must choose between these options is what spawned and kept alive the original controversy over the AB effect. As mentioned earlier, before the 1959 paper, it had been accepted by the majority of scientists that potentials are conveniences that have no physical significance of their own. This is an interesting historical fact, given that a thorough reading of the literature most likely would have inclined them otherwise.

In any event, the AB effect showed that measurable effects could be attributed to the potentials. Consequently, it was suggested that privileged status should be assigned to ϕ and $\vec{A}$ rather than to $\vec{E}$ and $\vec{B}$. Not surprisingly, controversy ensued. Throughout the convoluted evolution of electrodynamics even great scientists got things wrong from time to time.

The question itself misses the mark. In the non-relativistic classical theory, the vector potential is indeed optional, a convenience. On the other hand, in electrodynamics and quantum mechanics the fundamental equations contain ϕ and $\vec{A}$. The physical fields $\vec{E}$

---

[3] It might seem that the quantum effect is non-local. After all, something happens because of a $\vec{B}$ field that is nonzero elsewhere but is zero in the region of the quantum system. The Lorentz force being zero might entice one to think that nothing can happen. However, such instinct is based on classical physics, which does not take into consideration interference of particle waves. The physics is indeed local. It is brought about through the gauge field $\vec{A}$.

Another example of a debate over local versus non-local physics is the Einstein-Podolsky-Rosen (EPR) paradox. Consider two atoms created by photodissociation of a homonuclear diatomic molecule. The atoms move in opposite directions in vacuum until they are far apart. Then one of them is detected. It is known *a priori* that one atom is formed in state a, whereas the other is formed in state b. The wave function ψ that describes the system gives equal probability of finding the detected particle in state a or b: $\psi \sim \psi_1(a)\psi_2(b) \pm \psi_1(b)\psi_2(a)$. The point is this. If particle 2 is detected in state a, particle 1 is instantly collapsed into state b. Information is not transferred to particle 1 at the speed of light. Rather, the collapse is instantaneous. In other words, the state of particle 1 is determined by an act that takes place far away from particle 1. This has been presented as a non-local effect. However, it is not mysterious, nor is it non-local. The state of the system must include the relative translational motion that connects the two atoms. The above expression for ψ does not include this if a and b are taken as internal states. The state of the system happens to be enormous, as it contains fragment relative translational motion.

and $\vec{B}$ can enter in terms like $\vec{\mu}\cdot\vec{E}$, but these arise from terms like $\vec{p}\cdot\vec{A}$. There is no getting around it; the potentials occupy center stage.

We shall now consider the effect brought about by $\vec{A}$ on a charged particle wave function. It is assumed that the particle does not interact with other particles, and that it experiences no potential other than $\vec{A}$. It is assumed that ϕ is zero. To distinguish effects due to $\vec{B}$ from effects due to $\vec{A}$, an experimental arrangement is conjured in which the region of space under consideration has $\vec{B} = \nabla\times\vec{A} = 0$, whereas $\vec{A} \neq 0$.

It turns out that the phase of an electron wave function is affected as it passes through the $\vec{A} \neq 0$ region. However, if this is all that happens (specifically, there is no boundary condition to be satisfied and the wave function only acquires a phase shift) there will be no observable effect because $|\psi|$ is unaffected. On the other hand, interference depends on relative phase. Thus, observable effects can arise in cases in which two waves that are coherent with respect to one another pass through a region of nonzero $\vec{A}$ and are then superposed. For example, in Fig. 1(b) each wave in general experiences a different $\vec{A}$. Following passage through their respective $\vec{A} \neq 0$ regions these waves are brought together such that interference is observed. Consequently, $|\psi_1+\psi_2|$ is affected by the relative phase between ψ_1 and ψ_2. Because such particle interference has no classical counterpart, effects arise due to the presence of $\vec{A}$ in regions where $\vec{B} = 0$. This is the basis of the AB effect.

3. Particle-on-a-Ring

Let us now examine a charged particle confined to a circular ring in a region where $\vec{B} = 0$ and $\vec{A} \neq 0$. The experimental arrangement suggested in the 1959 paper by Aharonov and Bohm will be analyzed later. The presence of nonzero $\vec{A}$ requires that there exists a region of space where $\vec{B} \neq 0$. However, we shall consider a case in which the particle wave never enters this region.

We know that a charged particle is affected by the presence of $\vec{A}$, for example, as seen in the kinetic energy operator

$$T = \frac{1}{2m}\left(\vec{p} - \frac{q}{c}\vec{A}\right)\cdot\left(\vec{p} - \frac{q}{c}\vec{A}\right). \tag{3.1}$$

where $\vec{p} = -i\hbar\nabla$ is the canonical momentum, and $\vec{\pi} = \vec{p} - (q/c)\vec{A}$ is the kinetic momentum. This is standard quantum mechanics with a classical rather than quantized electromagnetic field.

In the particle-on-a-ring example discussed below, we shall see that $\vec{A}$ lifts the twofold degeneracy of the field-free $e^{\pm in\phi}$ pairs. In other words, when $\vec{A} \neq 0$ the energies are not the same for $e^{in\phi}$ and $e^{-in\phi}$. For an observable effect to exist, it is essential that the region where $\vec{B}$ is nonzero lie *inside* the red ring shown in Fig. 2. Were the region of nonzero $\vec{B}$ *outside* the red ring, there would be no observable effect, despite the fact that $\vec{A}$ is nonzero on the ring. The reason for this will be made clear in subsequent sections. Following this exercise, a succinct way to account for $\vec{A}$'s presence is introduced. Name-

ly, in regions where $\vec{B} = 0$, the phase brought about by $\vec{A}$ can be obtained by integrating $\vec{A}$ over a conveniently chosen path.

These exercises enable the topological difference between *simply connected* and *non-simply-connected* domains to be introduced in a transparent manner. It is hoped that they enable the AB effect to be seen as unambiguous, even intuitive.

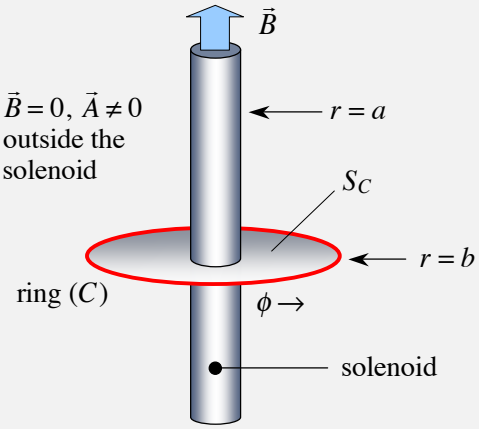


Figure 2. A particle moves on a circular ring of radius b (red). The field $\vec{B}$ is contained within a cylinder (solenoid) of radius a. Its uniform strength yields a flux $\Phi = \pi a^2 B$.

As mentioned earlier, an important application of the results obtained in the present chapter arises in the area of molecular electronic structure and associated nuclear dynamics. It is known that conical intersections of potential energy surfaces abound in nature and have important consequences: radiationless decay (internal conversion, intersystem crossing), chemical reaction pathways, etc. Something important called geometric phase accrues when a conical intersection is approached (notably, encircled). For example, geometric phase can affect vibrational wave functions profoundly. Without its inclusion they are often egregiously in error.

The phase associated with the AB effect is, for all practical purposes, the same as this molecular geometric phase. Therefore, as a matter of pedagogy, it makes no sense to approach conical intersections in electronic structure theory without first taking the time to understand the AB effect. In examining both the AB effect and conical intersections, we will see that intuition flows both ways. The latter can provide insight into the former.

The chapter ends with a summary that, among other things, puts to rest the paradox: If $\vec{E}$ and $\vec{B}$ are the true physical fields and ϕ and $\vec{A}$ are mere constructs that serve only to yield $\vec{E}$ and $\vec{B}$, how can ϕ and $\vec{A}$ be responsible for an experimental effect? We will see that this line of thought is inconsistent with quantum mechanics.

4. Solution for $\vec{A} \neq 0$

Referring to Fig. 2, assume that a magnetic field $\vec{B} = B\hat{z}$ of uniform strength is contained within a cylinder (solenoid) of radius a. A charged particle moves on a ring of radius b. Outside the cylinder $\vec{B} = 0$, so the total magnetic flux Φ is

$$\Phi = \iint_{S_C} d\vec{S} \cdot \vec{B} = \pi a^2 B = \iint_{S_C} d\vec{S} \cdot \nabla \times \vec{A}\,. \tag{4.1}$$

Application of Stokes' theorem yields the relationship between Φ and the closed line integral of $\vec{A}$ around the ring:[4]

$$\Phi = \oint_C dl \cdot \vec{A} = b \oint_C d\phi A_\phi = 2\pi b A_\phi .$$

Thus,

$$\boxed{A_\phi = \frac{\Phi}{2\pi b}} \tag{4.2}$$

The time independent Schrödinger equation, using $A_\phi = \Phi / 2\pi b$, $q = -e$, and the canonical momentum for the $\hat{\phi}$ direction, $-i\hbar\partial_{b\phi}$, is

$$E\psi = \frac{1}{2m}\left(\vec{p} + \frac{e}{c}\vec{A}\right)^2 \psi \qquad \vec{A} = \frac{\Phi}{2\pi b}\hat{\phi}$$

$$= \frac{1}{2m}\left(-\frac{\hbar^2}{b^2}\partial_\phi{}^2 + \left(\frac{e\Phi}{2\pi bc}\right)^2 - i\hbar\frac{e\Phi}{\pi b^2 c}\partial_\phi\right)\psi . \tag{4.3}$$

Because $\vec{A}$ is constant, $\vec{p}\cdot\vec{A} + \vec{A}\cdot\vec{p} = 2\vec{A}\cdot\vec{p}$. In the present case this reads

$$2\vec{A}\cdot\vec{p} = 2A_\phi p_\phi = 2(\Phi / 2\pi b)(-i\hbar\partial_{b\phi}) .$$

When this is multiplied by e/c, the result is recognized as the rightmost term inside the bracket in eqn (4.3). Equation (4.3) is tidied by introducing constants: $hc/e = \Phi_L$ is the London flux quantum that arises in the theory of superconductivity,[5] and $\hbar^2/2mb^2 = B_{rot}$ is a rotational constant. Thus, eqn (4.3) becomes

$$\partial_\phi{}^2\psi + i\,2(\Phi/\Phi_L)\partial_\phi\psi + C\psi = 0 , \tag{4.4}$$

where $C = E/B_{rot} - (\Phi/\Phi_L)^2$. The solution is obtained by introducing $\psi \propto e^{in\phi}$, in which case eqn (4.4) becomes

---

[4] Symmetry ensures that $\vec{A}$ has no ϕ or z dependence. Only radial dependence is allowed. Because $\nabla \times \vec{A} = 0$ outside the cylinder, A_z is constant for $r > a$ so it can be set to zero. Likewise, A_r is constant on the ring so it can be set to zero.

[5] In superconductivity the appropriate quantum is actually $\Phi_0 = \Phi_L/2$. The London quantum Φ_L was introduced before it was appreciated that current is carried by electron pairs. In a BCS superconductor, charge is carried by a quasiparticle called a Cooper pair. Its charge is – $2e$ instead of the electron charge – e in the case of a regular conductor.

$$n^2 + 2(\Phi/\Phi_L)n - C = 0. \tag{4.5a}$$

Thus,

$$n = -\Phi/\Phi_L \pm \sqrt{E_n / B_{rot}}\ . \tag{4.5b}$$

To satisfy the boundary condition: $\psi(0) = \psi(2\pi)$, n must be an integer. Consequently, the energy eigenvalues are (see Fig. 3)

$$E_n = B_{rot}\left(n + \Phi/\Phi_L\right)^2 . \tag{4.6}$$

For each value of n, the wave function (leaving aside normalization) is

$$\psi_n = e^{in\phi} . \tag{4.7}$$

Note that energies for n values that have the same magnitude but differ in sign are not degenerate, as they are when $\vec{A} = 0$. For example, putting $n = +2$ and -2 alternately into eqn (4.6) yields

$$E_{+2} = B_{rot}\left(2 + \Phi/\Phi_L\right)^2$$

$$E_{-2} = B_{rot}\left(-2 + \Phi/\Phi_L\right)^2 . \tag{4.8}$$

This difference is seen in Fig. 3 for the case $\Phi/\Phi_L = 1$; clearly $E_{+2} \to E_{-2}$ as $\Phi \to 0$.

The canonical and kinetic angular momenta differ. The former is obtained by differentiating eqn (4.7), which yields $n\hbar$. The latter includes the vector potential and is given by $\hbar\left(n + \Phi/\Phi_L\right)$. Clockwise and counterclockwise directions have different speeds for a given value of $|n|$. An interesting feature can be seen with eqns (4.6) and (4.8). Namely, energies are repeated if Φ/Φ_L is an integer. For example, with $\Phi/\Phi_L = 1$, $E_{+2} = 9B_{rot} = E_{-4}$. In the present context this appears as a curiosity, but in superconductivity it is important. If a large superconducting current flows on the ring it will produce a large magnetic field inside the ring. There need be no external source of magnetic field. The resulting flux Φ is quantized as a consequence of the particle-on-a-ring boundary condition. This is sufficiently interesting (and closely related to the AB effect) that it is presented and discussed separately in Appendix 3. *Flux Quantization in Superconductivity*.

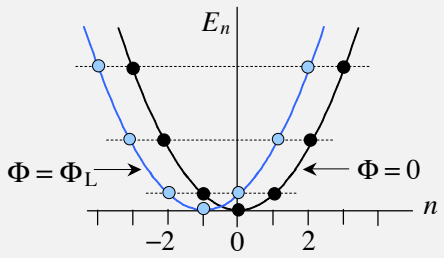


Figure 3. Plots of eqn (4.6) for Φ/Φ_L values of 0 and 1 are shown. Variation is parabolic. The blue curve: $\Phi/\Phi_L = 1$, is simply offset from the black curve.

The vector field $\vec{A}$ causes the electron wave's phase to change as the wave circulates on the ring. To satisfy the boundary condition, this phase change in turn results in higher or lower kinetic momentum.

The above example illustrates how a phase that has no classical counterpart has a significant effect on a quantum mechanical system.

5. Vector Potential Appears in a Phase Factor

Let us now return to the time dependent Schrödinger equation for a particle of charge q in a region where $\vec{A} \neq 0$ and $\vec{B} = 0$.

$$\left(\frac{1}{2m}\left(-i\hbar\nabla - \frac{q}{c}\vec{A}\right)^2\right)\psi = i\hbar\partial_t\psi \tag{5.1}$$

The scalar potential has been set to zero. The solution to eqn (5.1) is given by the $\vec{A} = 0$ solution times a phase factor. Specifically,

$$\psi = \psi_0 e^{ig(\vec{r})} \qquad g(\vec{r}) = \frac{q}{\hbar c}\int_{\vec{r}_0}^{\vec{r}} d\vec{r}\,'\cdot\vec{A}(\vec{r}\,') \,, \tag{5.2}$$

where ψ_0 is the $\vec{A} = 0$ solution.[6] To verify that the wave function given by eqn (5.2) is the solution to eqn (5.1), take the gradient of ψ:[7]

$$\nabla\left(\psi_0 e^{ig(\vec{r})}\right) = \left(\nabla\psi_0\right)e^{ig(\vec{r})} + \underbrace{i\left(\nabla g(\vec{r})\right)}_{(q/\hbar c)\vec{A}(\vec{r})}\psi_0 e^{ig(\vec{r})} \tag{5.3}$$

---

[6] The introduction of this form is not as *ad hoc* as it might appear. Phase accumulated along a path is given by the action, S, which is the integral of $dt\,L$:

$$S = \int dt\,L = \int dt\left(T + (q/c)\vec{v}\cdot\vec{A}\right) = S_0 + (q/c)\int d\vec{r}\,'\cdot\vec{A}(\vec{r}\,')$$

The subtlety is the Lagrangian: $T + (q/c)\vec{v}\cdot\vec{A}$, which can be found in Appendix 2. Beyond elementary systems, Lagrangians are often conjured to yield the equations of motion. It is straightforward (though algebraically tedious) to verify that this Lagrangian yields the correct equations of motion.

[7] The Leibniz integral rule for differentiation of an integral is:

$$\frac{d}{d\alpha}\int_{a(\alpha)}^{b(\alpha)} dx\, f(x,\alpha) = \frac{d\,b(\alpha)}{d\alpha} f\left(b(\alpha),\alpha\right) - \frac{d\,a(\alpha)}{d\alpha} f\left(a(\alpha),\alpha\right) + \int_{a(\alpha)}^{b(\alpha)} dx\,\frac{\partial}{\partial\alpha} f(x,\alpha)\,.$$

This result does not depend on the path taken between $\vec{r}_0$ and $\vec{r}$, just the integration end points. This is true in any region where $\nabla \times \vec{A} = 0$, as can be seen by applying Stokes' theorem. The underlying reason for the path independence is discussed in the next section. Using eqn (5.3), the kinetic momentum $\vec{\pi}$ operating on $\psi = \psi_0 e^{ig(\vec{r})}$ becomes

$$\begin{aligned} \vec{\pi}\psi &= \left(-i\hbar\nabla - (q/c)\vec{A}\right)\psi_0 e^{ig(\vec{r})} \\ &= e^{ig(\boldsymbol{r})}\left(-i\hbar\left(\nabla\psi_0\right) - i\hbar\left(i(q/\hbar c)\vec{A}\right)\psi_0 - (q/c)\vec{A}\psi_0\right) \\ &= -i\hbar\, e^{ig(\vec{r})}\left(\nabla\psi_0\right) \end{aligned} \tag{5.4}$$

The two terms that contain $\vec{A}$ cancel one another. To obtain $\vec{\pi}\cdot\vec{\pi}\psi$, operate on eqn (5.4) from the left with $\vec{\pi}\cdot = (-i\hbar\nabla - (q/c)\vec{A})\cdot$. A little algebra yields

$$\left(-i\hbar\nabla - (q/c)\vec{A}\right)^2 \psi = -\hbar^2\left(\nabla^2\psi_0\right)e^{ig(\vec{r})} \tag{5.5}$$

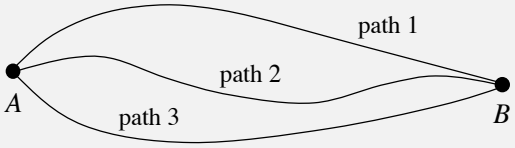
When this is put into eqn (5.1) the exponentials $e^{ig(\vec{r})}$ can be cancelled, leaving a Schrödinger equation for ψ_0:

$$-\frac{\hbar^2}{2m}\nabla^2\psi_0 = i\hbar\partial_t\psi_0 \tag{5.6}$$

Equation (5.6) confirms that ψ_0 is the solution to the Schrödinger equation obtained by setting $\vec{A} = 0$. These manipulations verify that the affect on the wave function brought about by the presence of $\vec{A}$ in a region where $\vec{B} = 0$ is accounted for with a phase factor. As mentioned earlier and discussed below, this result is valid only in regions where $\vec{B} = \nabla \times \vec{A} = 0$.

6. Topology

Stokes' theorem is now applied to a region where $\nabla \times \vec{A} = 0$ (Fig. 4). Because all closed line integrals of $\vec{A}$ are equal to zero in this region, the line integral between points *A* and *B* depends only on the end points, not the path between them. Were $\nabla \times \vec{A}$ not zero, circuit integrals would enclose flux, and integration along different paths would, in general, yield different results. In this case, eqn (5.2) would not be valid. This is why the material in the last section applies only to regions where $\nabla \times \vec{A} = 0$.



$$\nabla \times \vec{A} = 0 \Rightarrow \int_A^B d\vec{r}_1 \cdot \vec{A} + \int_B^A d\vec{r}_2 \cdot \vec{A} = 0$$

Figure 4. As long as $\nabla \times \vec{A} = 0$, all circuit integrals vanish (Stokes' theorem). In this case, integration between points depends on the values of $\vec{A}$ at the end points, not the path.

Because of the fact that only the end points count, the paths can be distorted to suit our convenience. For example, a complicated trajectory that passes around a solenoid can be replaced with a simpler one. This is the situation that exists in a *simply connected region*. A simply connected region is one in which a path that surrounds the region (in this case a path for a line integral) can be continuously deformed until it becomes a point.

As mentioned above, in a region where $\nabla \times \vec{A} \neq 0$, integration is over a surface that flux passes through, so Stokes' theorem yields $\Phi_{\text{encl}} = \oint d\vec{r} \cdot \vec{A}(\vec{r})$. Thus, a line integral between two points is not independent of the path, because different closed circuits enclose different amounts of flux. For example, when the path surrounds a cylinder that contains flux, we get the results that were obtained in the particle-on-a-ring section. In this case, the closed circuit integral encloses a *non-simply-connected* region.

7. Aharonov-Bohm

The groundwork has been laid. An experimental geometry close to that suggested in the original paper is shown in Fig. 5. A particle wave incident from the left is split into two equally intense components. The $\vec{B}$ field (out of the page) is present in a cylindrical solenoid. It is assumed that $\vec{B}$ is zero outside the cylinder. One component wave passes above the cylinder, while the other passes below it. Outside the cylinder, $\vec{A}$ is nonzero, so the waves experience $\vec{A}$, but not $\vec{B}$. It is assumed that passage is sufficiently slow that the fields are perceived by the particle as static. That is, no transitions between energy levels arise because of a time varying electromagnetic field. Of course a free electron has no energy levels other than spin, which is irrelevant in the present context. However, had we transported an electron in a box (*vide infra*, Section 9), the issue of transitions induced by a time varying electromagnetic field would have arisen.

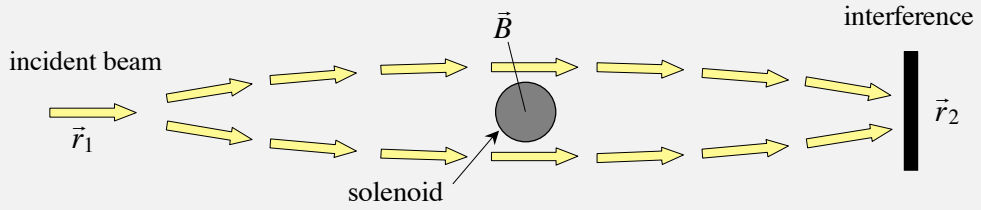


Figure 5. The $\vec{B}$ field (out of the page) is fully contained within a cylinder, whereas $\vec{A}$ is nonzero outside the cylinder. Waves that pass above and below are obtained by splitting a single wave incident from the left in order to ensure coherence. That is, with $\vec{B} = 0$, they retain a fixed phase with respect to one another throughout their trajectories. With $\vec{B} \neq 0$, a phase shift between them is acquired that depends on the strength of $\vec{B}$ and is detectable via interference at the place where the beams recombine. As the strength of $\vec{B}$ is varied, the interference pattern changes. Interestingly, the "center-of-mass" of the interference pattern is unaffected, as moving the center-of-mass would require a force.

From the electron's perspective, it is inevitable that $\vec{A}$ is, to some extent, time dependent because the electron enters and leaves the region of nonzero $\vec{A}$. This causes the electron to experience an electric field ($-\partial_{ct}\vec{A}$) and, from Maxwell's equations, a magnetic field. However, this has a negligible effect. Several authors have shown that the fields $\vec{E}$ and $\vec{B}$ thus generated are small enough to be inconsequential. Verification on your part is left as an exercise.

As shown earlier, in the region $r > a$, where a is the cylinder radius, $\vec{A}$ is given by

$$\vec{A} = \frac{\Phi}{2\pi r}\,\hat{\phi} \tag{7.1}$$

where $\hat{\phi}$ points counterclockwise.

At the points of closest approach to the cylinder, for the lower component wave, $\vec{A}$ points in the same direction as the wave's canonical momentum, while for the upper component wave $\vec{A}$ is directed oppositely to the wave's canonical momentum. Phase shifts are introduced to each of the waves, in one case advancing the wave's phase, in the other case retarding it.

After passing through the region near the cylinder, the waves are guided such that they recombine. Assuming that coherence has been maintained throughout the transit from $\vec{r}_1$ to $\vec{r}_2$, the interference pattern depends on the flux Φ. Thus, a reasonable experiment is to detect signals that are due to the recombining upper and lower waves, and monitor them as $\vec{B}$ is varied. This interference region is indicated by the thin black rectangle in Fig. 5. For example, this will result in a sinusoidal modulation of the signal at a given location. Several QuickTime movie clips that show how this works are available at the website:

http://rugth30.phys.rug.nl/quantummechanics/ab.htm.

The presence of $\vec{A}$ results in a phase difference, β, for waves that follow the two paths indicated in Fig. 5. The use of eqn (5.2) yields

$$\beta = \frac{q}{\hbar c}\left(\int_{\vec{r}_1}^{\vec{r}_2} d\vec{r}\cdot\vec{A}_{\text{lower}} - \int_{\vec{r}_1}^{\vec{r}_2} d\vec{r}\cdot\vec{A}_{\text{upper}} \right). \tag{7.2}$$

The limits on the second integral are now reversed to create a closed circuit. That is, β is the same as for a single wave that traverses the following circuit: (i) start at $\vec{r}_1$, (ii) pass below the cylinder, (iii) turn around at $\vec{r}_2$, (iv) pass above the cylinder, and (v) arrive back at $\vec{r}_1$. Thus, β is given by

$$\boxed{\beta = \oint d\vec{r}\cdot(q/\hbar c)\vec{A}\,.} \tag{7.3}$$

The right hand side of eqn (7.3) is the closed circuit integral of something with units of momentum: $(q/c)\vec{A}$), divided by $\hbar$. Because $\nabla\times\vec{A} = 0$ outside the cylinder, paths between $\vec{r}_1$ and $\vec{r}_2$ can be deformed as we please. For example, deformations that capitalize on the symmetry are obtained by using semicircular paths of radius $a + \varepsilon$, where the solenoid radius is a, and ε is small: one path above the cylinder and the other below (Fig. 6). Integration over the closed circuit then yields

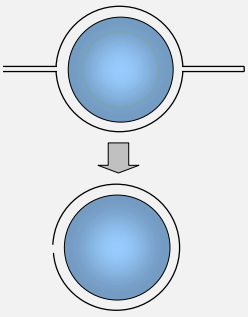


Figure 6. The path indicated in Fig. 5 can be distorted without affecting the outcome. Upper: Semicircles are formed around the solenoid. Lower: The straight parts cancel leaving a circular path, open on the left as a reminder that there is no boundary condition.

$$\beta = (q/\hbar c)\Phi = 2\pi\Phi/\Phi_{\text{L}} \tag{7.4}$$

This result is the same as the one obtained using the particle-on-a-ring model.

The integration in eqn (7.3) is written for a closed circuit to underscore the relationship to the particle-on-a-ring model. Keep in mind, however, that there is no boundary condition to be met here, in contrast to the particle-on-a-ring case. The circuit C is complete in the sense that the mathematics can be written as a closed line integral, but there is no bound eigenstate. Thus, $\vec{A}$ does not affect the energy of the waves that pass above and below the cylinder. A phase shift arises because $\vec{A}$ appends phases to the waves, and this manifests in the interference region indicated in Fig. 5.

8. Extensions and Generalizations

The potentials ϕ and $\vec{A}$ have emerged as central objects in the quantum theory. There is no temptation to use $\vec{E}$ and $\vec{B}$, nor would they suffice were this attempted. Without doubt, the quantum mechanical wave of a particle having charge q acquires phase when it passes through a region of nonzero ϕ and/or $\vec{A}$:

$$\exp\left(-i(q/\hbar)\int dt\,\phi\right) \tag{8.1}$$

and

$$\exp\left(i(q/\hbar c)\int d\vec{r}\cdot\vec{A}\right) \tag{8.2}$$

It is not a coincidence that the expressions given in eqns (8.1) and (8.2) can be combined into a single expression by using the four-vector: $A^\mu = (A^0, A^1, A^2, A^3) = (\phi, \vec{A}) = (\phi, A_x, A_y, A_z) = (A_0, -A_1, -A_2, -A_3)$.

$$\exp\left(-i(q/\hbar c)\int \underbrace{dx^\mu A_\mu}_{\text{or } dx_\mu A^\mu}\right) = \exp\left(-iS/\hbar\right) \text{ where } S \text{ is the action} \tag{8.3}$$

We will continue using arrows and Latin superscripts and subscripts to represent three-vectors and their components, while four-vector components will be represented using Greek superscripts and subscripts. Until now, we have referred to the $\vec{E}/\vec{B}$ pair as *fields* and the $\phi/\vec{A}$ pair as *potentials*. Of course, the potentials are also fields according to the mathematical definition of a field. Now that we have, for the most part, abandoned $\vec{E}/\vec{B}$ in favor of $\phi/\vec{A}$, the potentials will be referred to as fields. For example, the gauge field of electromagnetism is A^μ. Note: If this is your first encounter with phase and (covariant/contravariant) tensor algebra, it is hard to avoid sign errors. Do not be disheartened if this becomes exasperating. It is just math.

The product $dx^\mu A_\mu$ $(= dx_\mu A^\mu)$ is a Lorentz scalar. Thus, the central object in the theory is invariant with respect to Lorentz transformation. This is something you may wish to dwell on for a while. The other important requirement is gauge invariance. Physical

quantities should not depend on a choice of gauge. A schematic illustration of the gauge field and how gauge invariance works is given in Fig. 7. The plane denoted x^μ is a schematic representation of 4D spacetime. From each spacetime point a strand called a *fiber* extends upward. In Fig. 7(a), the blue dots represent the gauge field associated with spacetime points. The gauge field in general varies in spacetime, so the points are shown at different heights. The spacetime points are of course densely packed, as they must be if they are to represent a continuous spacetime. The collection of fibers is called a *bundle*. The mathematical term *fiber bundle* brings to mind the bundles of fibers that are used in fiber-optic communications.

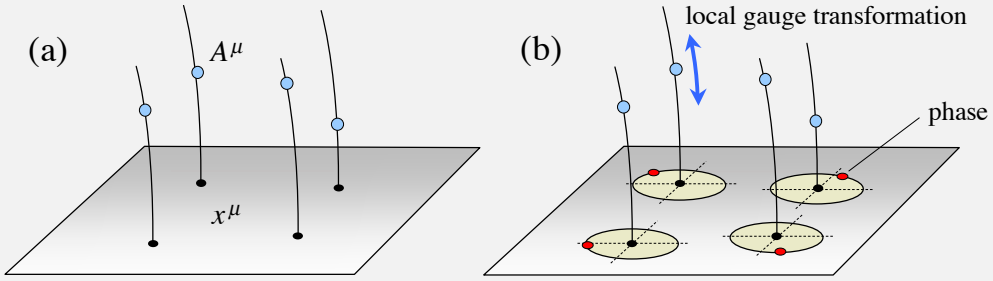


Figure 7. This figure is highly schematic. (a) The gauge field A^μ is distributed over spacetime. Its value at each spacetime point is indicated by a blue dot above the spacetime plane, x^μ. (b) A local gauge transformation changes, in a correlated manner, both the gauge field and the phase factor of a state vector. The former slides on its fiber; the latter is represented by red dots and yellow circles. Adapted from K. Huang: *Fundamental Forces of Nature: The Story of Gauge Fields*.

The last few sections have shown that the gauge field appears in both the kinetic momentum operator and the wave function. This is essential if gauge invariance is to be satisfied. Referring to Fig. 7(b), a local gauge transformation applied to the gauge field is represented schematically by displacements of the blue dots along their respective fibers. At the same time, the wave function must change in a correlated manner by acquiring a phase factor to ensure that the physical world remains oblivious to our choice of gauge. This phase factor is indicated with red dots and yellow circles, with phase angles defined relative to the axes shown. Though Fig. 7 is highly schematic, it gives one a sense of the registry that must exist between the gauge transformation of the field A^μ and the gauge transformation of the wave function.

To see how this works insofar as wave functions and observables are concerned, let us add to $\vec{A}$ the gradient of a scalar: $\vec{A} \rightarrow \vec{A} + \nabla\zeta$. We know that this does not affect $\vec{B}$, because $\vec{B} = \nabla \times \vec{A}$ and the curl of any gradient is identically zero. Also, $\partial_{ct}\nabla\zeta$ must be added to $\vec{E}$ to cancel the contribution that arises from changing $\vec{A}$. In other words, $\vec{E}$ remains $-\nabla\phi - \partial_{ct}\vec{A}$, where ϕ and $\vec{A}$ are the original potentials, as the gauge transformation gives: $-\nabla\phi - \partial_{ct}(\vec{A} + \nabla\zeta) + \partial_{ct}\nabla\zeta$. The $\nabla\zeta$ terms cancel, leaving the original $\vec{E}$.

The affect on ψ of adding $\nabla\zeta$ to $\vec{A}$ is given by

$$\psi \rightarrow \tilde{\psi} = \psi e^{i(q/\hbar c)\zeta} . \tag{8.4}$$

When the gauge transformed kinetic momentum operator $\tilde{\vec{\pi}}$ acts on $\tilde{\psi}$ given by eqn (8.4) we find that nothing has changed. In other words, the gauge transformation does not affect the physical world. We would have a strange theory on our hands otherwise. To see how this works, operate on $\tilde{\psi}$ with $\tilde{\vec{\pi}}$:

$$\tilde{\vec{\pi}}\tilde{\psi} = \left(-i\hbar\nabla - (q/c)\vec{A} - (q/c)\nabla\zeta\right)\psi e^{i(q/\hbar c)\zeta}$$

$$= e^{i(q/\hbar c)\zeta}\Big(-i\hbar\nabla \underbrace{- i\hbar\, i(q/\hbar c)\nabla\zeta}_{\text{these cancel}} - (q/c)\vec{A} \underbrace{- (q/c)\nabla\zeta}_{\text{these cancel}}\Big)\psi \qquad (8.5)$$

$$= e^{i(q/\hbar c)\zeta}\left(-i\hbar\nabla - (q/c)\vec{A}\right)\psi \qquad (8.6)$$

The term $\nabla\zeta$ has disappeared, and ζ appears only in the phase factor. Applying $\tilde{\vec{\pi}}$ to eqn (8.6) shows that the kinetic energy is unaffected. When this is used with the time dependent Schrödinger equation the exponential factor $e^{i(q/\hbar c)\zeta}$ multiplies (from the left) all terms in the Schrödinger equation, so it can be canceled, and the gauge transformation does not appear. In other words, though $\vec{A}$ is changed by $\nabla\zeta$, and the electron wave function acquires a potentially complicated phase factor due to the gauge transformation, the observables are unaffected.

Suppose one decides to simplify matters by choosing $\nabla\zeta = -\vec{A}$. In this case, $\vec{A} + \nabla\zeta$ vanishes, leaving just $\vec{p} = -i\hbar\nabla$. However, $\vec{p}$ acting on $\psi e^{i(q/\hbar c)\zeta}$ gives

$$\vec{p}\left(\psi e^{i(q/\hbar c)\zeta}\right) = e^{i(q/\hbar c)\zeta}\vec{p}\psi + (iq/\hbar c)(-i\hbar\nabla\zeta)(e^{i(q/\hbar c)\zeta}\psi).$$

The last term is $-e^{i(q/\hbar c)\zeta}(q/c)\vec{A}\psi$. Referring to eqn (8.5), in this case it is the third and fourth terms that cancel. The second term is now $(-i\hbar)(iq/\hbar c)(-\vec{A}) = -(q/c)\vec{A}$, so nothing has changed.

As a consequence of the gauge transformation, $\tilde{\psi}$ is related to ψ by

$$\tilde{\psi} = U\psi\,, \qquad (8.7)$$

where[8]

$$\boxed{U = e^{i(q/\hbar c)\zeta}\,.} \qquad (8.8)$$

You are undoubtedly familiar, from classical and quantum mechanics, with generators of transformations. The Hamiltonian is the generator of time evolution; momentum is the generator of spatial displacement; the discrete generators of point groups; and so on. Here

[8] For U to have meaning in quantum mechanics it must be single valued. This requires that the space on which U operates is simply connected so the integral of $\nabla\zeta$ depends on the end points.

we have a transformation: $U = e^{i(q/\hbar c)\zeta}$, whose generator is charge. Thus, it can be said that *charge is the generator of gauge transformation*. This conclusion is far-reaching, applying to charges other than the electric one considered here.

The transformation U is multiplication by a phase factor whose argument is a function of spatial location for $\vec{A} \neq 0$ and $\phi = 0$. For each spatial location, it is a rotation in the complex plane. This unitary gauge group is labeled U(1). It is the gauge group of quantum electrodynamics (QED). Other aspects of QED can prove quite challenging, but not its gauge group.

Covariant Derivative

The above results can be distilled into a neat algorithm. To see how this works, consider a charged particle in the presence of an electromagnetic field. We begin by assuming that the system is expressed in terms of the charged particle and the field, but without any coupling between them. In other words, the Hamiltonian is $p^2/2m + H_{\text{EM}}$, where H_{EM} is the electromagnetic energy that was discussed in Chapter 2. The interaction is now turned on. The electrodynamics interaction is achieved by promoting the canonical momentum from $p^\mu = i\hbar\partial^\mu$ to the momentum: $i\hbar\partial^\mu - (q/c)A^\mu$. In other words,

$$\partial^\mu \rightarrow \partial^\mu + i(q/\hbar c)A^\mu = D^\mu \tag{8.9}$$

The symbol D^μ denotes what is called a covariant derivative. You might wonder why something like D^μ is called a *covariant* derivative, as it is obviously contravariant. A person might find this baffling, and rightly so.

$$D^\mu = \partial^\mu + i(q/\hbar c)A^\mu$$
$$\partial^0 \rightarrow D^0 = \partial^0 + i(q/\hbar c)\phi$$
$$\partial^j \rightarrow D^j = \partial^j + i(q/\hbar c)A^j$$

Is it safe to come out yet?

The confusion arises because the term covariant has two meanings. To make matters worse, authors tend to mingle them without mentioning the distinction – leaving that to context. This passes muster once you know what is going on, but it can be confusing on first encounter. When an equation or expression does not change its *form* under a Lorentz transformation it is said to be Lorentz covariant. For example: four-vectors like p^μ are Lorentz covariant; Maxwell's equations and the Dirac equation are Lorentz covariant; the Schrödinger equation is not Lorentz covariant; and so on. The covariant derivative D^μ falls into this category. However, we also deal with covariant and contravariant components of tensors. This is different. In this context, these terms denote the transformation properties of the tensor components. With this distinction in mind, we see that D^μ and D_μ are each Lorentz covariant, whereas D^μ transforms as a contravariant four-vector, while D_μ transforms as a covariant four-vector.

The covariant derivative converts the system from one of global gauge invariance to one of local gauge invariance, and when it undergoes a Lorentz transformation it mixes electric and magnetic interactions according to the requirements of special relativity. Without it the system is not invariant with respect to a local gauge transformation. Local

gauge invariance is assured through the substitution given by eqn (8.9). This is referred to as minimal substitution (or minimal coupling) because it is the least complicated way to ensure gauge invariance. In other words, without the gauge field A^μ there is no straightforward mechanism for ensuring gauge invariance. This idea was illustrated in Fig. 7, where it was pointed out that if A^μ changes, so must the wave function's phase factor.

Unitary Transformation

When quantization of the electromagnetic field was examined in the Chapter 2, we saw how gauge transformation works in electromagnetic theory. To further explore how gauge invariance works in the quantum mechanics of particles interacting with fields, let us revisit the material surrounding eqns (8.4) – (8.8). Namely, a gauge transformation is applied to the vector potential $\vec{A}$ by adding the gradient of a scalar: $\vec{A} \to \vec{A} + \nabla\zeta$, while at the same time the wave function undergoes: $\psi \to \tilde{\psi} = \psi e^{i(q/\hbar c)\zeta}$. These are called gauge transformations of the second and first kind, respectively – terms introduced by Pauli.

The easiest way to handle manipulations is by using the unitary transformation given by eqn (8.8): $U = e^{i(q/\hbar c)\zeta}$. Let us start by applying this transformation to the kinetic momentum: $\vec{\pi} = \vec{p} - (q/c)\vec{A}$, where $\vec{p} = -i\hbar\nabla$.

$$\tilde{\vec{\pi}} = U\vec{\pi}U^\dagger \tag{8.10a}$$

$$= U\left(\vec{p} - (q/c)\vec{A}\right)U^\dagger \tag{8.10b}$$

$$= U\vec{p}U^\dagger - (q/c)\vec{A} \tag{8.10c}$$

$$= \vec{p} - (q/c)\vec{A} - (q/c)\nabla\zeta \tag{8.11}$$

Obviously, $\vec{\pi}$ can be recovered from $\tilde{\vec{\pi}}$ by using $\vec{\pi} = U^\dagger\tilde{\vec{\pi}}U$. Now the Schrödinger equation is transformed from one involving H and ψ to one involving $\tilde{H}$ and $\tilde{\psi}$ by operating from the left with U:

$$H\psi = i\hbar\partial_t\psi \implies UHU^\dagger U\psi = i\hbar\partial_t(U\psi) \implies \tilde{H}\tilde{\psi} = i\hbar\partial_t\tilde{\psi} \tag{8.12}$$

where $\tilde{H} = UHU^\dagger$ is

$$\tilde{H} = \frac{1}{2m}\left(U\vec{\pi}U^\dagger \cdot U\vec{\pi}U^\dagger\right) \tag{8.13a}$$

$$= \frac{1}{2m}\left(U\left(\vec{p} - (q/c)\vec{A}\right)U^\dagger \cdot U\left(\vec{p} - (q/c)\vec{A}\right)U^\dagger\right) \tag{8.13b}$$

$$= \frac{1}{2m}\left(\vec{p} - (q/c)\vec{A} - (q/c)\nabla\zeta\right)\cdot\left(\vec{p} - (q/c)\vec{A} - (q/c)\nabla\zeta\right) \tag{8.13c}$$

$$= \frac{\tilde{\vec{\pi}}\cdot\tilde{\vec{\pi}}}{2m} \tag{8.14}$$

When this acts on $\tilde{\psi}$ the original Schrödinger equation is recovered.

Exercise: Verify that the above statement is true.

Because $\tilde{H}\tilde{\psi} = i\hbar\partial_t\tilde{\psi}$ is the same as $H\psi = i\hbar\partial_t\psi$, each form gives identical physical results. These manipulations have been carried out using the vector potential $\vec{A}$, with the electric potential ϕ set to zero. Let us now see how things work using $\phi \neq 0$. With $\vec{A} = 0$ and $\phi \neq 0$, apply the gauge transformations: $\phi \to \phi + \phi^0$ and $U = \exp(-i(q/\hbar)\int dt\phi^0)$ to the Schrödinger equation:

$$\tilde{H}\tilde{\psi} = \left(\frac{p^2}{2m} + q\phi + q\phi^0\right)\tilde{\psi} = i\hbar\partial_t\tilde{\psi}$$

$$= i\hbar\partial_t\left(\psi\exp\left(-i(q/\hbar)\int dt\phi^0\right)\right)$$

$$= \exp\left(-i(q/\hbar)\int dt\phi^0\right)\left(i\hbar\partial_t + q\phi^0\right)\psi\,. \tag{8.15}$$

The exponential multiplies the left and right hand sides, so it cancels. The $q\phi^0$ terms also cancel, leaving

$$\left(\frac{p^2}{2m} + q\phi\right)\psi = i\hbar\partial_t\psi\,. \tag{8.16}$$

Again, we see that $\tilde{H}\tilde{\psi} = i\hbar\partial_t\tilde{\psi}$ is the same as $H\psi = i\hbar\partial_t\psi$. Referring to eqn (8.3), it is now clear how the four-vector form for the unitary transformation works.

$$\boxed{\begin{aligned} U &= \exp\left(-i(q/\hbar c)\int dx^\mu A_\mu\right) \\ &= \exp\left(-i(q/\hbar)\int dt\phi + i(q/\hbar c)\int d\vec{r}\cdot\vec{A}\right) \end{aligned}} \tag{8.17}$$

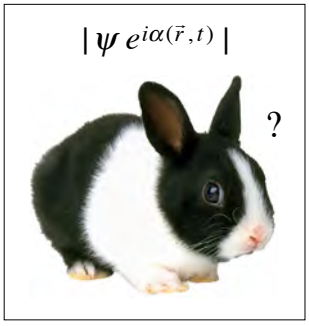
Notice that the + sign in front of the last integral on the right is due to the fact that $d\vec{r}\cdot\vec{A} = -dr^i A_i$.

Local Gauge Invariance

Gauge invariance is now examined from a different perspective. Specifically, the requirement of local gauge invariance is seen to lead to the minimal coupling prescription: $\vec{p} \rightarrow \vec{p} - (q/c)\vec{A}$ and the presence of $q\phi$, and all that follows from it. In other words, the requirement of local gauge invariance ensures that the charged particle and the electromagnetic field couple to one another. In fact, it dictates the form this coupling must assume. From this perspective, the requirement of local gauge invariance is the driving force, and quantum electrodynamics is a consequence. Do not underestimate the depth and utility of this *gauge principle*. It reveals an essential difference between classical and quantum electrodynamics. Classically when $\vec{A}$ and ϕ appear they can be converted to expressions containing $\vec{E}$ and $\vec{B}$. With $\vec{E}$ and $\vec{B}$ in hand, $\vec{A}$ and ϕ are no longer needed. Quantum mechanically this is not the case because phase enters in a way that has no classical counterpart. Without local gauge invariance quantum mechanics is inconsistent.

The reason for assigning this privileged status to the gauge principle can be illustrated using the free particle Schrödinger equation. To see how this works we consider a particle in what we at first take to be field-free space, and demand that quantum mechanics satisfies local gauge invariance. This reveals why and how the electromagnetic field must be included in the equations of motion.

A solution of the Schrödinger equation for a particle yields a wave function whose overall phase is immaterial insofar as the particle's probability density. Despite the fact that we routinely use complex functions like e^{ikx}, measurements of probability density tell us about the magnitude $|\psi|$ rather than ψ. Overall phase does not enter.



Everyone knows that multiplication of a total wave function by $e^{i\alpha}$, where α is a real constant, has no affect. This is an example of global phase. Invariance with respect to a global phase (gauge) transformation gives rise to conservation of charge. Let us now promote the global phase α, which is independent of space and time, to $\alpha(\vec{r},t)$, which depends on space and time. Though $\psi \rightarrow \psi e^{i\alpha}$ is now $\psi \rightarrow \psi e^{i\alpha(\vec{r},t)}$, the magnitude of ψ is unaffected, because $|\psi e^{i\alpha}| = |\psi e^{i\alpha(\vec{r},t)}|$. Applying this gauge transformation to the wave function, the Schrödinger equation is

$$i\hbar\partial_t\left(\psi e^{i\alpha(\vec{r},t)}\right) = \frac{1}{2m}\left(-i\hbar\nabla\right)^2 \psi e^{i\alpha(\vec{r},t)}. \tag{8.18}$$

Hereafter, α will be understood to mean $\alpha(\vec{r},t)$. A bit of algebra yields

$$i\hbar\partial_t\psi = \left(\hbar\partial_t\alpha + \frac{1}{2m}\left(-i\hbar\nabla + \hbar\nabla\alpha\right)^2\right)\psi \tag{8.19}$$

The only way this expression can recover the original Schrödinger equation is to alter the original Schrödinger equation. The original Schrödinger equation must be augmented. Specifically, it must contain one or more terms that enable cancellation of the terms $\hbar\nabla\alpha$ and $\hbar\partial_t\alpha$ that have entered through differentiation of the gauge (phase) transformed wave function, as seen in eqn (8.19). In other words, the term $-i\hbar\nabla$ in eqn (8.18) must be augmented in a way that enables $\hbar\nabla\alpha$ in eqn (8.19) to be eliminated. Likewise, $\hbar\partial_t\alpha$ needs to be eliminated when the scalar potential is included, though we shall focus here on the vector potential.

This is achieved by including in the Schrödinger equation a vector field that can have $-\hbar\nabla\alpha$ added to it, and in so doing cancel the $\hbar\nabla\alpha$ in eqn (8.19). The presence of this vector field at the outset is used to ensure that no physical effect arises in the equations because of the gauge transformation applied to ψ. Let us call this vector field $-(q/c)\vec{A}$. In other words, we are now starting with $-i\hbar\nabla-(q/c)\vec{A}$ in eqn (8.18) instead of just $-i\hbar\nabla$. Consequently, we are also starting with $(-i\hbar\nabla-(q/c)\vec{A}+\hbar\nabla\alpha)$ in eqn (8.19) instead of just $(-i\hbar\nabla+\hbar\nabla\alpha)$. If the quantity $(\hbar c/q)\nabla\alpha$ is now added to $\vec{A}$, all is well:

$$-i\hbar\nabla-(q/c)\vec{A}+\hbar\nabla\alpha \Rightarrow -i\hbar\nabla-(q/c)\left(\vec{A}+(\hbar c/q)\nabla\alpha\right)+\hbar\nabla\alpha = -i\hbar\nabla-(q/c)\vec{A}\,. \quad (8.20)$$

This shows that a local gauge transformation of the wave function can be used as long as it is accompanied by a change of the gauge field: $\vec{A}\rightarrow\vec{A}+(\hbar c/q)\nabla\alpha$. Consistency with previous notation is achieved by using $\alpha=(q/\hbar c)\zeta$. In other words, $\psi\rightarrow\psi e^{i(q/\hbar c)\zeta}$ must be accompanied by $\vec{A}\rightarrow\vec{A}+\nabla\zeta$.

Application of the gauge principle has resulted in a requirement that a vector field be present in the quantum mechanical equations of motion. Furthermore, the requirement that no physical consequence results from the transformation: $\vec{A}\rightarrow\vec{A}+\nabla\zeta$ implies that, for the classical counterpart, forces due to $\vec{A}$ manifest in its curl. But this is nothing more that electromagnetism: $\vec{B}=\nabla\times\vec{A}$.

The term $\hbar\nabla\alpha$ in eqn (8.19) has now been eliminated, but $\hbar\partial_t\alpha$ in eqn (8.19) must also be eliminated. On the basis of the above introduction of the vector field $\vec{A}$, the manner in which this works can be anticipated. Namely, if the original Schrödinger equation contains the potential energy $q\phi$, this needs to undergo the transformation $\phi\rightarrow\phi-\partial_t\zeta$ in order to cancel the term $\hbar\partial_t\alpha$ (which is equal to $(q/c)\partial_t\zeta$) in eqn (8.19). This maneuver leaves the electric field unchanged:

$$\vec{E}=-\nabla\phi-\partial_{ct}\vec{A} \Rightarrow -\nabla\left(\phi-\partial_{ct}\zeta\right)-\partial_{ct}\left(\vec{A}+\nabla\zeta\right)=-\nabla\phi+\nabla\cancel{\partial_{ct}\zeta}-\partial_{ct}\vec{A}-\cancel{\partial_{ct}\nabla\zeta}\,. \quad (8.21)$$

The complete picture is now in place. The requirement of local gauge invariance implies the presence of the vector and scalar quantities $\vec{A}$ and ϕ (equivalently, A^μ), the electromagnetic potentials. The gauge invariant Schrödinger equation is

$$i\hbar\partial_t\psi = \left(\frac{1}{2m}\left(-i\hbar\nabla-(q/c)\vec{A}\right)^2+q\phi\right)\psi\,. \quad (8.22)$$

The role of electric charge is subtle. It is interesting that q multiplies the fields $\vec{A}$ and ϕ, given that q is a property of the particle. For example, the interaction of $\vec{A}$ with the charged particle is already partly in place with $(q/c)\vec{A}$. The dynamical part of the coupling arises from $\vec{p}$ interacting with $(q/c)\vec{A}$. Thus, terms like $\vec{p}\cdot(q/c)\vec{A}$ should not be thought of simply as a particle (described by $\vec{p}$) interacting with a field $(q/c)\vec{A}$.

The four-vector A^μ is called a *gauge field* because it ensures local gauge invariance. We were able to work through the AB effect without invoking the gauge principle because the math is straightforward. However, life is not as easy with nature's weak and strong forces, where the gauge principle is invaluable. We will not go into these topics. We shall see, however, that application of the gauge principle to the Born-Oppenheimer approximation offers great advantage.

The Gauge Principle

The transformation: $\psi \to \tilde{\psi} = \psi e^{i(q/\hbar c)\zeta}$

implies the presence of $(q/c)\vec{A}$ and $q\phi$,

namely, the gauge field A^μ.

$$\vec{A} \to \tilde{\vec{A}} = \vec{A} + \nabla\zeta$$

$$\phi \to \tilde{\phi} = \phi - \partial_{ct}\zeta$$

9. Variation on the Main Theme

We have seen that a monoenergetic electron that passes through a region where $\vec{A} \neq 0$ acquires a phase: $\gamma = -(e/\hbar c)\int d\vec{r} \cdot \vec{A}$. Though the highly idealized situation indicated in Fig. 5 enlists 1D (albeit along a curved trajectory) monoenergetic waves of uniform magnitude, the same phase would be appended to 1D *wave packets*, because each of its *k*-components acquires the same phase.

Electron in a Box

Taking this a step further, suppose an electron is contained inside a hollow box and this box-plus-electron combination is transported around the solenoid. A stationary box has a set of single particle energies E_n and eigenfunctions ψ_n. For a box whose dimensions are sufficiently small, the energies E_n are large relative to any influence that $\vec{A}$ might have on them. Likewise, as long as the box is transported slowly, the eigenvalues and eigenfunctions of the stationary box are expected to hold, to a high degree of accuracy, in the absence of degeneracy. This is the regime of interest. It is assumed that the box has sides whose lengths differ such that there is no degeneracy, and that the electron is in one of the eigenstates of the box.

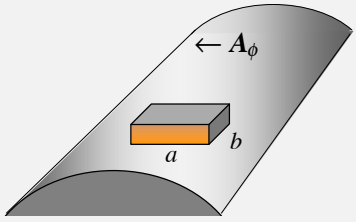


Figure 8. An electron is in a box that is positioned outside a solenoid that contains the flux Φ. The long sides of the box have lengths *a* and *b*.

As mentioned above, as long as the box is transported *slowly*, it is reasonable to expect the solutions for the stationary box to remain good approximations. For example, the motion of the box will distort *ever so slightly* the electron wave functions ψ_n, causing them to take on a mixed character in the basis of stationary box functions. However, as long as the energy separation between levels remains large compared to any energy imparted to the particle due to motion of the box, this is a minor effect that can be ignored. The electron wave functions acquire the phase γ as the box is transported, because this depends only on the line integral of $\vec{A}$ along the path taken by the box.

Now consider this from the perspective of an adiabatic separation. The environment inside the box has associated with it a characteristic time scale that is dictated by the box dimensions and the particle mass (in this case that of an electron). This time scale is rapid relative to the one on which external parameters are varied. The large difference in time scales justifies the adiabatic approximation. Electron dynamics inside the box average out insofar as their contribution to the phase γ is concerned. The only relevant coordinate for computing the phase is the location of the box.

10. Born-Oppenheimer Approximation and Local Gauge Invariance

The gauge principle is now applied to Born-Oppenheimer states of polyatomic molecules. There is an important distinction between this case and the examples in the last few sections. There, control of external parameters was absolute. For example, the electron-in-a-box follows adiabatically the slow variation of the box location without influencing it in turn. There is no wave function for the external parameters, as they are completely under our control.

With the Born-Oppenheimer approximation (BOA), the Schrödinger equation for the nuclear degrees of freedom can be solved for each of the adiabatic potential energy surfaces (adiabats). The BOA is great for solving the electron problem and then solving the nuclear problem, thereby obtaining vibrational wave functions. Its breakdown is taken into account by computing couplings due to terms that are omitted from the Hamiltonian in making the BOA. We shall see that the BOA also leads to an interesting phase. This phase is the molecular counterpart of the AB phase. It is discussed in the present section and the next.

Phase and Flux

Momentum operators for nuclear degrees of freedom play a central role in nonadiabatic dynamics because nuclear motions couple adiabats. The Born-Oppenheimer approximation is most accurate when the potential surfaces under consideration are separated by large amounts of energy. On the other hand, when the potential surfaces lie in close proximity to one another, notably in regions of the nuclear space where they cross, the approximation breaks down.

To use nuclear momentum operators in a Schrödinger equation for the nuclei, it is necessary to integrate over the electron coordinates. For example, consider the action of $\vec{p} = -i\hbar\nabla$ on $\chi\psi$, where χ and ψ are nuclear and electron wave functions, respectively, and it is understood that ∇ operates only with respect to nuclear coordinates. Next, $-i\hbar\nabla\chi\psi$ is multiplied from the left by ψ, and integration is carried out over the electron coordinates. Thus, the effective nuclear kinetic momentum operator, $\boldsymbol{\pi}$, is defined as

$$\boldsymbol{\pi}\chi = \langle\psi|\boldsymbol{p}\chi\psi\rangle . \tag{10.1}$$

Keep in mind that ψ depends parametrically on the nuclear coordinates. Here and hereafter bold type is used to indicate nuclear degrees of freedom.

Despite the fact that the wave function $\chi\psi$ (say, a single adiabat ψ and a single vibrational level χ) is single-valued, χ and ψ are not, in general, separately single-valued. We may wish to make each of them single-valued for convenience. However, they need not start out this way, and it is not *necessary* that they be single-valued. In other words, as long as χ and ψ form a single-valued $\chi\psi$, they can each be non-single-valued.

For example, with a single-valued $\chi\psi$, a gauge transformation of the electronic wave function: $\psi \to \tilde{\psi} = \psi e^{i\eta}$ can be used to obtain a single-valued $\tilde{\psi}$. It is understood that η is a function of external parameters $\boldsymbol{R}$, which in the present case are the nuclear coordinates. Note that a gauge transformation applied to the electron wave function must be manifest in $\tilde{\chi}$ to ensure that $\tilde{\chi}\tilde{\psi}$ is single valued. In other words, $\tilde{\psi} = \psi e^{i\eta}$ requires $\tilde{\chi} = \chi e^{-i\eta}$. In addition, it is a good idea to keep track of terms that account for transitions, *i.e.*, $\langle \psi' | \boldsymbol{p}\chi\psi \rangle$, as these play a central role in the theory of nonadiabatic transitions. All of this is well known.

Let us now return to local gauge invariance and examine the role it plays in the present context. Any time a wave function is subjected to a local gauge transformation, a vector field must be present. It enables dynamical effects that would be incurred through the wave function's phase transformation to be counteracted. In other words, the overall gauge transformation (of the wave function and the vector field) must not result in additional physics. This was discussed *ad nauseam* in Section 8.

In the present case, the requirement of local gauge invariance implies the existence of a vector field that we shall call $\boldsymbol{f}$. As in Section 8, this vector field is present in such a way that it can have something added to it to ensure that no physical effect enters the mathematics via the gauge transformation. The "something" that needs to be added is the gradient of a scalar because it is integrable. That is, because $d\boldsymbol{R}\cdot\nabla\eta = d\eta$, the integral of $d\boldsymbol{R}\cdot\nabla\eta$ depends only on the end points, and therefore no flux is enclosed in a circuit integral of $d\boldsymbol{R}\cdot\nabla\eta$. The circuit integral of the gauge field $\boldsymbol{f}$ in general encloses flux, whereas the circuit integral of $d\boldsymbol{R}\cdot\nabla\eta$ does not.[9] This is how gauge invariance is achieved. This procedure was used in Section 8, where the gauge principle was applied to electrodynamics.

An important distinction between the present case and the one discussed in section 8 is the space upon which the phase varies. In Section 8, the space was that of the electron. Here, however, the space is that of the nuclear degrees of freedom. Consequently, the gauge field will appear in the dynamical equations for the nuclei, but not the dynamical equations for the electrons. This is a switch. Here, the nuclei move in a field created by the electrons, whereas in Section 8 the electrons moved in the field A^ν.

In electrodynamics, the charge of the particle couples to the gauge field A^ν, and the momentum of the particle participates in this coupling through terms like $\vec{p}\cdot(q/c)\vec{A}$. In the Born-Oppenheimer case, the gauge field must likewise be proportional to a momentum in order that its line integral yields an action. This gauge field can be obtained through consideration of how nuclear momentum is coupled to electron degrees of freedom, of course averaged over electron degrees of freedom. Thus, eqn (10.1) identifies the gauge field: $\boldsymbol{f} = \langle \psi | \boldsymbol{p}\psi \rangle$, through the operation

---

[9] To see how this works, recall electromagnetism:

$$\oint d\vec{l}\cdot\vec{A} = \int d\vec{S}\cdot\nabla\times\vec{A} = \int d\vec{S}\cdot\vec{B} = \Phi .$$

The addition of $\nabla\eta$ to $\vec{A}$ results in the term: $\oint d\vec{l}\cdot\nabla\eta$. However, this vanishes because $\nabla\times\nabla\eta = 0$. Therefore the circuit integral of $\nabla\eta$ does not enclose flux.

$$\boldsymbol{\pi}\chi = \langle\psi|\boldsymbol{p}\chi\psi\rangle = \big(\boldsymbol{p} + \underbrace{\langle\psi|\boldsymbol{p}\psi\rangle}_{\boldsymbol{f} = -i\hbar\langle\psi|\nabla\psi\rangle}\big)\chi\,. \tag{10.2}$$

Invariance (actually covariance) with respect to gauge transformation is verified by the substitutions: $\tilde{\chi} = \chi e^{-i\eta}$ and $\tilde{\boldsymbol{f}} = \boldsymbol{f} + \hbar\nabla\eta$.[10] This differs slightly from the phase that was used to gauge transform the wave function in the AB case. There, $e^{i(q/\hbar c)\zeta}$ was used, with the sign in the exponent taken to be positive, whereas here (with $e^{-i\eta}$) the sign in the exponent is taken to be negative. The reason for using $e^{-i\eta}$ here is because χ describes nuclei. In other words, $\tilde{\chi} = \chi e^{-i\eta}$ implies for the corresponding gauge transformed electron wave function $\tilde{\psi} = \psi e^{i\eta}$ to ensure single-valuedness of $\tilde{\chi}\tilde{\psi}$. In any event, the signs assigned to the phases of $\tilde{\chi}$ and $\tilde{\psi}$ are arbitrary as long as they differ

The desired phase is obtained by closed circuit integration of the gauge field. As before, a closed circuit is chosen because it is easy to interpret. In the AB effect, the phase β was obtained through integration of $(q/c)\vec{A}$ over a path, followed by division by $\hbar$. Here, the gauge field is $\boldsymbol{f}$, so the integration is

$$\begin{aligned}\gamma(C) &= -\frac{1}{\hbar}\oint_C d\boldsymbol{R}\cdot\boldsymbol{f} \\ &= i\oint_C d\boldsymbol{R}\cdot\langle\psi|\nabla\psi\rangle\,.\end{aligned} \tag{10.3}$$

Applying Stokes' theorem yields

$$\gamma(C) = i\iint_{S_C} d\boldsymbol{S}\cdot\nabla\times\langle\psi|\nabla\psi\rangle\,. \tag{10.4}$$

With $\gamma(C)$ expressed as a surface integral, it is clear that it is invariant with respect to the transformation: $\boldsymbol{f}\to\boldsymbol{f}+\hbar\nabla\eta$. In other words, $\nabla\times(\boldsymbol{f}+\hbar\nabla\eta) = \nabla\times\boldsymbol{f}$ because the curl of the gradient of any scalar is identically zero.

Geometry issues loom. Stokes' theorem presupposes a 3D $\boldsymbol{R}$-space. In general, however, the number of participating nuclear degrees of freedom is not limited to three. To deal with spaces having more than three dimensions, differential geometry is used, as this generalizes Stokes' theorem to an arbitrary number of dimensions. This works nicely, but if you are not familiar with differential geometry, it will take a while to acquire a useful level of facility. We shall therefore stick to 3D. From eqn (10.4), we see that $\gamma(C)$ is a flux through the surface enclosed by C (Fig. 11) in the space of the nuclear coordinates that account for the adiabatic transport of the electronic wave function. The phase $\gamma(C)$ that we obtained in the AB effect – and now here by using local gauge invariance – is called the geometric phase.

---

[10] $\tilde{\boldsymbol{\pi}}\tilde{\chi} = \big(\boldsymbol{p}+\boldsymbol{f}+\hbar\nabla\eta\big)\chi e^{-i\eta} = e^{-i\eta}\big(\boldsymbol{p}+(-i\hbar)(-i\nabla\eta)+\boldsymbol{f}+\hbar\nabla\eta\big)\chi$. The two $\nabla\eta$ terms inside the large parentheses cancel, leaving $e^{-i\eta}(\boldsymbol{p}+\boldsymbol{f})\chi$.

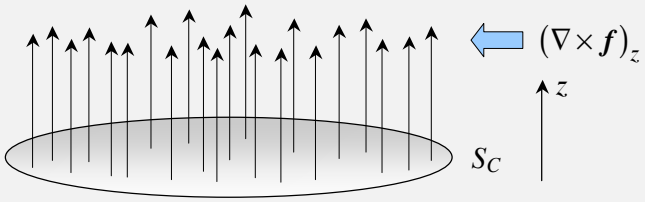


Figure 11. A flux $\nabla \times \boldsymbol{f}$ passes through the surface S_C. Integration of its component along the surface normal (indicated with arrows), and division by $-\hbar$, yields the phase $\gamma(C)$.

The expression for the geometric phase given by eqn (10.4) is in terms of a single adiabat ψ. However, the presence of one or more additional adiabats is implicit in the gradient of ψ, namely, the fact that $\nabla\psi \neq 0$. In other words, suppose ψ is composed of two diabats, and this composition is a function of the nuclear coordinates. Then $\nabla\psi$ will in general be nonzero. However, the fact that ψ is composed of two diabats implies that there is another adiabat that is composed of the same two diabats. This marks a distinction with electrodynamics. There the gauge group is U(1). Transitions do not take place between electron states.

Recall the expression derived earlier and given by eqn (7.3) for the AB phase

$$\beta = (q/\hbar c)\oint d\vec{r}\cdot\vec{A}\ . \tag{10.5}$$

The fact that the $\gamma(C)$ of the present section has the same form as the above β comes as no surprise, as these phases arise in the same way. In electrodynamics, charge q combines with the vector field A^μ to form a term that, at least in some ways, has the character of a momentum. In the present case, the nuclei undergo motion in the field presented by the electrons. If the electron wave function does not vary with the nuclear coordinates, there is no coupling and the geometric phase vanishes. It is the variation of ψ with $\boldsymbol{R}$ that constitutes the gauge field.

Surface Integral

To evaluate the surface integral of $\nabla\times\langle\psi|\nabla\psi\rangle$ in eqn (10.4), the operator $\nabla\times$ needs to be expressed on a basis. Using the vector identity: $\nabla\times(h\nabla g) = (\nabla h)\times(\nabla g)$, where h and g are scalars, and choosing the n^{th} adiabat yields $\nabla\times\langle n|\nabla n\rangle = \langle\nabla n|\times|\nabla n\rangle$. Adding closure enables eqn (10.4) to be written:

$$\gamma_n(C) = i\iint_{S_C} d\vec{S}\cdot\sum_{m\neq n}\langle\nabla n|m\rangle\times\langle m|\nabla n\rangle\ . \tag{10.6}$$

The $m = n$ term is excluded because $\langle\nabla n|n\rangle\times\langle n|\nabla n\rangle = 0$. Note also that $\langle n|\nabla n\rangle$ is imaginary, as $\langle n|\vec{p}n\rangle$ is real, being the diagonal matrix element of a Hermitian operator. Con-

sequently, $\langle \nabla n | n \rangle \times \langle n | \nabla n \rangle$ is real, and therefore its integration must vanish because $\gamma_n(C)$ must be real. No matter how you look at it, the $m = n$ term is excluded.

A common maneuver is to express $\langle m | \nabla n \rangle$ in terms of gradients of H by taking the gradient of $H|n\rangle = E_n|n\rangle$ and operating from the left with $\langle m|$. This yields

$$\langle m|\nabla H|n\rangle + \langle m|H|\nabla n\rangle = \langle m|\nabla E_n|n\rangle + \langle m|E_n|\nabla n\rangle . \tag{10.7}$$

The first term on the right hand side vanishes, leaving

$$\langle m|\nabla n\rangle = \frac{\langle m|\nabla H|n\rangle}{E_n - E_m} . \tag{10.8}$$

Using this with eqn (10.6) yields

$$\gamma_n(C) = -\iint_{S_C} d\vec{S}\cdot\vec{V}_n$$

$$\vec{V}_n = \mathrm{Im}\sum_{m\neq n} \frac{\langle n|\nabla H|m\rangle \times \langle m|\nabla H|n\rangle}{(E_n - E_m)^2} . \tag{10.9}$$

An advantage of eqn (10.9) over the path integral is that it is relatively easy to compute gradients of the Hamiltonian and evaluate the matrix elements.

Two States and a Degeneracy

The geometric nature of $\gamma(C)$ has emerged, for example as illustrated in Fig. 11. Manipulations that further highlight the geometric nature of $\gamma(C)$ are now carried out. To begin, consider two states and their $\boldsymbol{R} = 0$ degeneracy, as with two intersecting adiabats. Because only two states are involved, the sum in eqn (10.9) reduces to one term:

$$\vec{V}_n = \mathrm{Im}\left(\frac{\langle n|\nabla H|m\rangle \times \langle m|\nabla H|n\rangle}{(E_n - E_m)^2} \right). \tag{10.10}$$

A form for the Hamiltonian is used that can be evaluated readily using eqn (10.10). The most general form for a 2×2 Hamiltonian matrix is given by

$$H = \tfrac{1}{2}\vec{\sigma}\cdot\vec{R}$$

$$= \tfrac{1}{2}\left(\sigma_1 X + \sigma_2 Y + \sigma_3 Z\right)$$

$$= \tfrac{1}{2}\begin{bmatrix} Z & X - iY \\ X + iY & -Z \end{bmatrix} \tag{10.11}$$

The σ_i are Pauli matrices, the parameters X, Y, and Z are real, and the factor of ½ is there because it turns out to be convenient. Energies are relative to the $\boldsymbol{R} = 0$ degeneracy, so there is no need to include the 2×2 unit matrix that is required if we to have a complete basis in the space of 2×2 complex matrices. The eigenvalues of the matrix in eqn (10.11) (including the ½) are $\pm R/2$, where $R^2 = X^2+Y^2+Z^2$. The parameter-space vector $\boldsymbol{R}$ changes slowly, causing the eigenvalues and eigenvectors to evolve adiabatically, as indicated in Fig. 12.

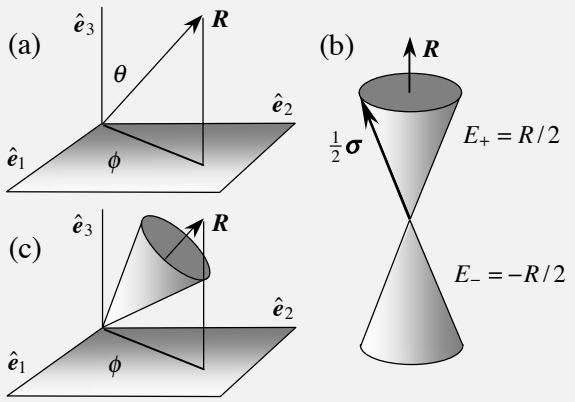


Figure 12. (a) $\boldsymbol{R} = X\hat{e}_1 + Y\hat{e}_2 + Z\hat{e}_3$. (b) $\boldsymbol{\sigma}$ is referenced to $\boldsymbol{R}$; eigenvalues of eqn (10.11) are $\pm R/2$. (c) $\boldsymbol{\sigma}$ is referenced to $\boldsymbol{R}$ throughout its evolution, facilitating the evaluation of matrix elements, *i.e.,* independent of where $\boldsymbol{R}$ points. The angles θ and ϕ vary as the tip of $\boldsymbol{R}$ traces a path in its parameter space.

You might wonder why the form used in eqn (10.11) has been chosen. What insight prompted its enlistment? The answer is that when two states are expressed in a two-state basis, with a degeneracy at $\boldsymbol{R} = 0$, the system has a spin-½ representation. Consequently, it is known *a priori* that the wave function undergoes a sign reversal in 2π. This corresponds to $Y = 0$ in eqn (10.11). Our two-state model with $Y \neq 0$ is more general so let us see what happens.

Referring to Fig. 12(c), the natural quantization direction for $\vec{\sigma}$ is $\boldsymbol{R}$. As $\boldsymbol{R}$ varies $\vec{\sigma}$ follows. The matrix element in eqn (10.10), is evaluated using the gradient of $H = \vec{\sigma}\cdot\boldsymbol{R}/2$, namely, $\nabla H = \boldsymbol{\sigma}/2$. The notation $|n\rangle = |+\rangle$ and $|m\rangle = |-\rangle$ is introduced in deference to the spin representation. The states follow adiabatically the evolution of $\boldsymbol{R}$, as indicated in Fig. 12(c). Because $\vec{\sigma}$ is referenced to $\boldsymbol{R}$, the matrix element in eqn (10.10) is easily evaluated:

$$\langle +|\nabla H|-\rangle = \tfrac{1}{2}\left(\langle +|\sigma_{1'}|-\rangle\hat{e}_{1'} + \langle +|\sigma_{2'}|-\rangle\hat{e}_{2'} + \langle +|\sigma_{3'}|-\rangle\hat{e}_{3'}\right)$$

$$= \tfrac{1}{2}\left(\hat{e}_{1'} - i\hat{e}_{2'}\right). \tag{10.12}$$

The primes indicate a reference system that retains its orientation with respect to $\boldsymbol{R}$ throughout the adiabatic evolution. The above matrix element was evaluated using the standard representation

$$|-\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad |+\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{10.13}$$

The matrix element $\langle + | \nabla H | - \rangle$ is put into eqn (10.10), and $E_+ - E_- = R$ is used to write

$$\vec{V}_+ = \mathrm{Im}\left(\frac{(\hat{e}_{1'} - i\hat{e}_{2'}) \times (\hat{e}_{1'} + i\hat{e}_{2'})}{4R^2} \right) \tag{10.14a}$$

$$= \frac{\hat{e}_{3'}}{2R^2} \tag{10.14b}$$

Thus, $\vec{V}_+$ points in the $\boldsymbol{R}$ direction, which is, by definition, $\hat{e}_{3'}$. As $\boldsymbol{R}$ evolves, its tip traces a path, returning to its original position. Using eqn (10.14) with eqn (10.9) gives

$$\begin{aligned} \gamma_+(C) &= -\iint_{S_C} d\vec{S} \cdot \frac{\hat{e}_{3'}}{2R^2} \\ &= -\iint_{S_C} R^2 d\Omega \, \hat{e}_{3'} \cdot \frac{\hat{e}_{3'}}{2R^2} \\ &= -\tfrac{1}{2}\Omega \end{aligned} \tag{10.15}$$

Clearly $\gamma_-(C) = +\frac{1}{2}\Omega$. Thus, the phase for the closed circuit C is

$$\boxed{\gamma_\pm(C) = \mp \tfrac{1}{2}\Omega} \tag{10.16}$$

The solid angle Ω subtended from the $\boldsymbol{R} = 0$ degeneracy point (Fig. 13) can take on a continuous range of values between 0 and 2π, dictated by the path of $\boldsymbol{R}$. When the matrix elements of H are real ($Y = 0$), $\boldsymbol{R}$ completes its circuit in the $\hat{e}_3\hat{e}_1$ plane. Because this plane contains $\boldsymbol{R} = 0$, it follows that $\gamma_{\pm}(C) = \mp\pi$. This is what happens in the case of a *conical intersection*. For the general case of a complex Hamiltonian matrix, an infinite number of planes contain $\boldsymbol{R} = 0$. This is relevant to molecules that have an odd number of electrons. This is discussed at some length in Appendix 4.

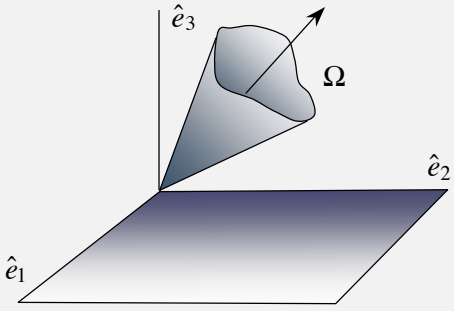
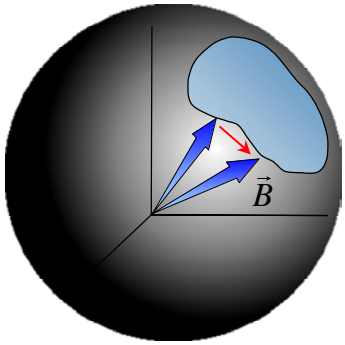


Figure 13. The closed circuit C defines a solid angle Ω in a 3D $\boldsymbol{R}$-space.

Spin in a Magnetic Field

In light of the connection that has just been made to spin-½, let us now consider the case of an arbitrary spin in a magnetic field. If $\vec{B}$ slowly changes its laboratory direction such that it traces out a path that returns it to its original orientation (Fig. 14), the Hamiltonian varies accordingly. The spin eigenstates follow adiabatically, remaining referenced to $\vec{B}$ throughout its slow evolution. Upon completion of the path, $\vec{B}$ has returned to its original orientation, but has the spin state also recovered its original form, or has it acquired a phase due to its adiabatic passage around the path?

Figure 14. The magnitude of $\vec{B}$ is constant as the direction of $\vec{B}$ traces a closed path. The tip of the thick blue arrow traces a closed path (light blue) on the sphere's surface.



From the previous section, we know how spin-½ behaves, so we now examine integer or odd-half-integer spin interacting with $\vec{B}$. The gradient of $H = K\vec{S}\cdot\vec{B}$ is $K\vec{S}$, and the energies are KBm. Thus, eqn (10.9) for a given m is

$$\vec{V}_m = \mathrm{Im}\sum_{m"\neq m}\frac{\langle m|\,\vec{S}\,|m"\rangle\times\langle m"|\,\vec{S}\,|m\rangle}{B^2(m-m")^2}. \qquad (10.17)$$

Exercise: Show that the gradient of $H = K\vec{S}\cdot\vec{B}$ is $K\vec{S}$. Hint: define the parameter space $\boldsymbol{R}$ in terms of $\vec{B}$.

Referring to eqn (10.17), the use of a double prime avoids confusion with the rotated axes, which are labeled with a single prime. The matrix elements are easily evaluated:

$$\langle m|\,\vec{S}\,|m"\rangle = \langle m|\,S_{1'}\hat{e}_{1'} + S_{2'}\hat{e}_{2'} + S_{3'}\hat{e}_{3'}\,|m"\rangle$$

$$= \langle m|\,S_{1'}\,|m\pm 1\rangle\hat{e}_{1'} + \langle m|\,S_{2'}\,|m\pm 1\rangle\hat{e}_{2'}$$

$$= \tfrac{1}{2}\Big(\langle m|\,S^{+} + S^{-}|m\pm 1\rangle\hat{e}_{1'} - i\langle m|\,S^{+} - S^{-}|m\pm 1\rangle\hat{e}_{2'}\Big). \qquad (10.18)$$

Nonzero matrix elements have $m" = m \pm 1$. Some algebra yields $\vec{V}_m = \big(m/B^2\big)\hat{e}_{3'}$. As with eqn (10.15), integration over $d\vec{S}$ is trivial because $\vec{V}_m$ is parallel to $\vec{S}$. For a surface enclosed by C, integration yields the geometric phase

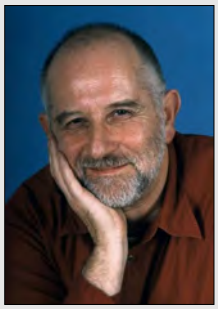
$$\boxed{\gamma_m(C) = -m\Omega\,.} \qquad (10.19)$$

This differs from eqn (10.16) in that $\mp$ ½ has been replaced by $-m$. When m is an integer and Ω is 2π, $\gamma_m(C)$ is an integer multiple of 2π. In this case the wave function does not change sign when $\Omega = 2\pi$. On the other hand, when m is odd-half-integer, the wave function changes sign when $\Omega = 2\pi$.

11. Molecular Geometric Phase

We have seen that one of the ways in which adiabatic change is manifest is through the geometric phase it engenders. The origins of what nowadays falls under the heading of geometric phase go back a long way: a theorem due to Wigner and Von Neumann, the Jahn-Teller effect (Peierls distortion), the Aharonov-Bohm effect, flux quantization in superconductivity, Pancharatnam and Berry phases, etc. Geometric phase serves as an essential ingredient in theories of intersecting potential energy surfaces and associated non-adiabatic dynamics. Appendix 5 describes the classical geometric phase that arises from the parallel transport of a vector on the surface of a sphere, for example, as in the case of the Foucault pendulum. You might find this interesting.

Quantum mechanically, adiabatic evolution means that the state does not change its quantum numbers. For example, nodes in a vibrational wave function cannot be created or annihilated as the curvature of the potential is varied adiabatically. This would require a precipitous change of the curvature of the wave function.



Michael Berry (Bristol)

Reinhard Schinke (Göttingen)

In this section, the approach introduced by Michael Berry is presented. Many of the (adiabatic) geometric phase results derived by Berry have already been obtained via local gauge invariance. The close relationships among the Aharonov-Bohm effect and its progenitors, gauge field theory, and the material in this section are stunning. It is amusing to note the resemblance between Michael Berry and a good friend of ours, and an excellent scientist, Reinhard Schinke.

General Situation

A state whose ket is $|n\rangle$ is assumed to evolve adiabatically under the influence of external parameters that shall be denoted collectively as $\boldsymbol{R}$. Given that our interest lies with polyatomic molecules, what we shall refer to as a "value" of $\boldsymbol{R}$ specifies a set of values: one each for all of the nuclear degrees of freedom. In other words, one value of $\boldsymbol{R}$ is a point in the multidimensional space of the nuclear degrees of freedom.

Berry took a more general approach in the sense that his $\boldsymbol{R}$ comprises a set of external parameters that are under one's complete control. The external parameters in Berry's model, being under complete external control, are not quantum mechanical. In the case of polyatomic molecules, we shall deal with the quantum nature of the nuclear degrees of freedom, which turns out to be of great importance.

It is assumed that the parameters of the $\boldsymbol{R}$-space vary on a time scale that is slow relative to the one on which dynamical processes associated with $|n\rangle$ transpire. As $\boldsymbol{R}$ under-

goes its relatively slow evolution, the time independent Schrödinger equation yields energy eigenvalues and eigenfunctions at each instant of time according to

$$H(\boldsymbol{R})|n(\boldsymbol{R})\rangle = E_n(\boldsymbol{R})|n(\boldsymbol{R})\rangle . \tag{11.1}$$

Parenthetic $\boldsymbol{R}$ is to emphasize the parametric dependence on the nuclear degrees of freedom. It is of course understood that $H(\boldsymbol{R})$ and $|n(\boldsymbol{R})\rangle$ depend on electron coordinates.

Think of this slow change of $E_n(\boldsymbol{R})$ and $|n(\boldsymbol{R})\rangle$ as a series of snapshots – one at each value of $\boldsymbol{R}$ throughout the system's adiabatic evolution. In electronic structure theory eqn (11.1) is solved for one value of $\boldsymbol{R}$, then another, and so on. However, these calculations provide no information about the relative phases of the solutions.

For such a system to adhere to nonrelativistic quantum mechanics, it must satisfy the Schrödinger equation, or something close to it, at all points in time throughout its adiabatic evolution:

$$H(\boldsymbol{R})|\psi_n\rangle = i\frac{d}{dt}|\psi_n\rangle . \tag{11.2}$$

Here and hereafter $\hbar = 1$ shall be used.

The use of a total derivative on the right hand side is necessary in order to account for the adiabatic evolution. You will see that the geometric phase does not depend on the amount of time required to complete the adiabatic change. However, it is necessary to take into account changes that take place in the $\boldsymbol{R}$-space, and this is the reason for using the total derivative instead of a partial derivative.

The state vector $|\psi_n\rangle$ includes a phase factor for the usual Schrödinger eigenstate phase, as well as an additional phase $\gamma_n(\boldsymbol{R})$:

$$|\psi_n\rangle = \exp\left(-i\int_0^t \omega_n(\boldsymbol{R})\,dt\right) e^{i\gamma_n(\boldsymbol{R})}\,|n(\boldsymbol{R})\rangle . \tag{11.3}$$

The phase $\gamma_n(\boldsymbol{R})$ (hereafter denoted γ_n) is due to adiabatic evolution on $\boldsymbol{R}$. The rest is standard nonrelativistic quantum mechanics. The ansatz given by eqn (11.3) was used by Berry in his 1984 paper. The term γ_n is an admission of our ignorance of how phase evolves on $\boldsymbol{R}$. This phase varies according to changes in $\boldsymbol{R}$, not how long it takes these changes to happen. For example, if $\boldsymbol{R}$ remains constant over a long stretch of time, γ_n does not change during this period.

An equation for γ_n is obtained by putting the $|\psi_n\rangle$ given by eqn (11.3) into eqn (11.2).[11] This yields three terms on the right hand side, one of which cancels the $\omega_n|\psi_n\rangle$ term on the left hand side, leaving

$$\frac{d\gamma_n}{dt}|\psi_n\rangle = i\exp\left(-i\int_0^t \omega_n dt\right) e^{i\gamma_n} \underbrace{\frac{d}{dt}|n\rangle}_{|\nabla n\rangle\cdot\frac{d\boldsymbol{R}}{dt}}. \tag{11.4}$$

The gradient is with respect to the parameters that constitute the $\boldsymbol{R}$-space. Now multiply eqn (11.4) from the left with $\langle\psi_n|$ and use $\langle\psi_n|\psi_n\rangle = 1$ on the left hand side to write

$$\frac{d\gamma_n}{dt} = i\langle\psi_n|\exp\left(-i\int_0^t \omega_n dt\right) e^{i\gamma_n}|\nabla n\rangle\cdot\frac{d\boldsymbol{R}}{dt}. \tag{11.5}$$

Next, use eqn (11.3) to replace $\langle\psi_n|$, in the above expression, yielding

$$\frac{d\gamma_n}{dt} = i\langle n|\nabla n\rangle\cdot\frac{d\boldsymbol{R}}{dt}. \tag{11.6}$$

The dt's in the denominators are now canceled. As anticipated, we see that time does not play an explicit role in determining the phase. Equation (11.6) thus becomes

$$d\gamma_n = i\langle n|\nabla n\rangle\cdot d\boldsymbol{R}, \tag{11.7}$$

and integration along a path in parameter space yields

$$\boxed{\gamma_n = \int_{\boldsymbol{R}_0}^{\boldsymbol{R}} d\boldsymbol{R}\cdot i\langle n|\nabla n\rangle.} \tag{11.8}$$

---

[11] Putting eqn (11.3) into eqn (11.2) yields

$$H|\psi_n\rangle = \omega_n|\psi_n\rangle = i\frac{d}{dt}\exp\left(-i\int_0^t \omega_n\, dt\right) e^{i\gamma_n}|n\rangle.$$

The right hand side is equal to $-\frac{d\gamma_n}{dt}|\psi_n\rangle + \omega_n|\psi_n\rangle + i\exp\left(-i\int_0^t \omega_n\, dt\right) e^{i\gamma_n}\frac{d}{dt}|n\rangle$.

The two $\omega_n|\psi_n\rangle$ terms cancel one another leaving eqn (11.4).

It is understood that γ_n is equal to the difference between the initial and final values: $\gamma_n = \gamma_n(\boldsymbol{R}) - \gamma_n(\boldsymbol{R}_0)$.

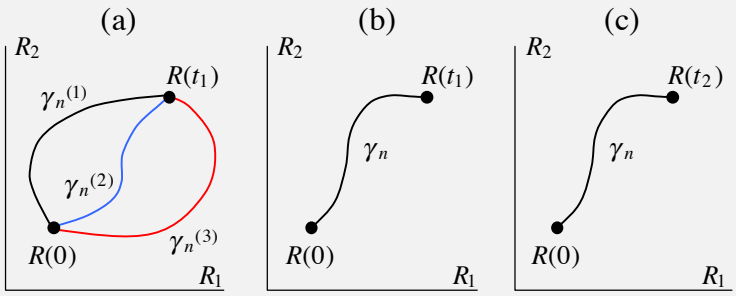


Figure 15. (a) Different paths result in different $\gamma_n^{(i)}$ values, though the time is the same for each path. (b and c) Different traversals of the same path require different amounts of time. Nonetheless, the same phase is accumulated.

Equation (11.8) shows that the phase γ_n depends on the path, not how long it takes to traverse it. Referring to Fig. 15(a), different paths that end at the same point in general give different phases, despite the fact that the time elapsed in traversing them is the same. Thus, γ_n cannot be written as an explicit function of time. Referring to Figs. 15(b) and (c), using different amounts of time to traverse the same path yields the same phase. Said differently, for a given path, the phase accumulated in going from $\boldsymbol{R}_0$ to $\boldsymbol{R}$ is the same whether passage is carried out slowly or "less slowly," as long as it is done adiabatically. Given that the state retains its quantum numbers, the only change that it can undergo is its phase, as illustrated in Fig. 16.

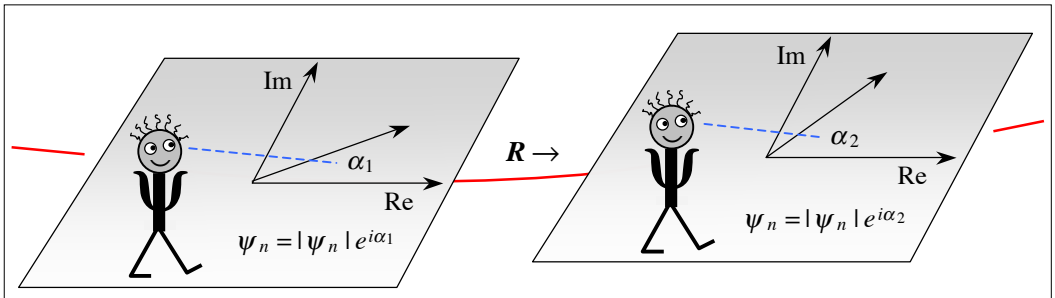


Figure 16. The adiabat ψ_n undergoes adiabatic evolution on $\boldsymbol{R}$ (red path). It accrues geometric phase according to integration of its gauge connection $i\langle n|\nabla n\rangle$ along the path. The phase of ψ_n is displayed in ψ_n's internal space, which consists of its real and imaginary parts. If, at the end of a closed circuit, the integration yields a nonzero value, there is a net geometric phase.

If $\boldsymbol{R}$ returns to its initial value via a closed path C, the geometric phase $\gamma_n(C)$ is that of a completed circuit:

$$\gamma_n(C) = \oint_C d\boldsymbol{R} \cdot i\langle n|\nabla n\rangle . \tag{11.9}$$

Equation (11.9) reveals the gauge field. In Section 7 we saw that the coupling between a charged particle and a steady electromagnetic field goes hand-in-hand with the Aharonov-Bohm phase: $\int d\vec{r} \cdot (q/c)\vec{A}$. In the present case, the phase is $\int d\boldsymbol{R} \cdot i\langle n|\nabla n\rangle$, so identification is clear: $i\langle n|\nabla n\rangle$ is the gauge field.

If $|n\rangle$ is single-valued its differentiation can be carried out with impunity along C. We will see that $|n\rangle$ can be assigned an arbitrary, parameter dependent phase without changing $\gamma_n(C)$, ensuring that a single-valued wave function can be used. Even if $|n\rangle$ begins life not single-valued it can be made single-valued. Alternatively, if the phase is chosen such that $d\boldsymbol{R} \cdot \langle n|\nabla n\rangle$ is zero along the path, it will be necessary to deal with a discontinuity upon completion of the circuit. This is evident from the gauge invariance of γ_n.

There is a trade-off in which $d\boldsymbol{R} \cdot \langle n|\nabla n\rangle$ can be made to vanish along C but at the expense of a wave function that is not single-valued. Alternatively, a single-valued wave function can be used, but at the expense of nonzero $d\boldsymbol{R} \cdot \langle n|\nabla n\rangle$ along C. The geometric phase $\gamma_n(C)$ is the same in either case. Because $\gamma_n(C)$ must be real, $\langle n|\nabla n\rangle$ must be imaginary.

Geometric phase can be represented as a line integral of a current in parameter space, and through Stokes' theorem as a flux density passing normal to the surface enclosed by C. Though eqns (11.9) and (11.10) are for a closed circuit, geometric phase accumulates along any path in $\boldsymbol{R}$-space. We shall focus on closed circuits, however, because manifestations of geometric phase are clear. The quantity $d\boldsymbol{R} \cdot \langle n|\nabla n\rangle$ is referred to as a *connection* (a term used frequently in general relativity) because it relates the wave function to its increment: $d\boldsymbol{R} \cdot \langle \psi|\nabla \psi\rangle = \langle \psi|d\psi\rangle$.

This result is similar to what happens with classical vector transport on the surface of a sphere (Appendix 5), where the classical geometric phase is given by the enclosed solid angle. In the previous section we saw that quantum geometric phase is proportional to the enclosed solid angle in $\boldsymbol{R}$-space.

Choosing Phase

We have seen that there are different ways to obtain $\gamma(C)$. For example, if the integrand vanishes along C, $\gamma(C)$ can be calculated at the close of the circuit, where the wave function is discontinuous. The integral vanishing along C means $\langle \psi|d\psi\rangle = 0$. That is, ψ is orthogonal to $d\psi$. This is not surprising. For example, in time independent nondegenerate perturbation theory the first order correction to a wave function is always orthogonal to the wave function. An alternate way to obtain $\gamma(C)$ involves the use of a gauge transformation to make the wave function single-valued on C.

These points are illustrated with an example in which the elements of a real 2×2 electronic Hamiltonian matrix depend parametrically on an angle in $\boldsymbol{R}$-space. Let us start by arranging H to suit our purposes:

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{bmatrix} \tag{11.11}$$

$$= \tfrac{1}{2}\mathrm{Tr}H + R\begin{bmatrix}\cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha\end{bmatrix}. \tag{11.12}$$

It is understood that a unit matrix multiplies the term ½ TrH . Parameterization of the Hamiltonian matrix in terms of an angle α, and the resulting adiabats in terms of the half-angle $\alpha/2$, is explained in the box below. The eigenvectors ψ_1 and ψ_2 are those of the angle matrix in eqn (11.12). Assuming that the H_{ij} matrix elements in eqn (11.12) are in a ϕ_1/ϕ_2 basis, the normalized eigenvectors are

$$\begin{bmatrix}\psi_1 \\ \psi_2\end{bmatrix} = \begin{bmatrix}\cos\alpha/2 & \sin\alpha/2 \\ -\sin\alpha/2 & \cos\alpha/2\end{bmatrix}\begin{bmatrix}\phi_1 \\ \phi_2\end{bmatrix}. \tag{11.13}$$

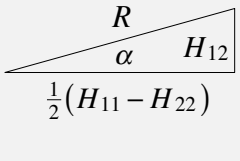
Half Angle

Write H_{11} and H_{22} as: $H_{11} = \frac{1}{2}(H_{11}+H_{22}) + \frac{1}{2}(H_{11}-H_{22})$

$H_{22} = \frac{1}{2}(H_{11}+H_{22}) - \frac{1}{2}(H_{11}-H_{22})$

The Hamiltonian matrix, minus the trace term, is parameterized in terms of an angle α using the triangle and on the right.

$$R\begin{bmatrix}\dfrac{(H_{11}-H_{22})}{2R} & \dfrac{H_{12}}{R} \\ \dfrac{H_{12}}{R} & -\dfrac{(H_{11}-H_{22})}{2R}\end{bmatrix} = R\begin{bmatrix}\cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha\end{bmatrix}$$



Eigenvalues of the angle matrix are ± 1. Using these with the characteristic equation gives the eigenfunctions, *e.g.*, for $+R$:

$$\frac{c_1}{c_2} = \frac{\sin\alpha}{1-\cos\alpha} \text{ ... (some juggling) ... } \frac{\cos(\alpha/2)}{\sin(\alpha/2)}$$

Therefore, $\psi(+R) = \phi_1\cos(\alpha/2) + \phi_2\sin(\alpha/2)$. Likewise, using $-R$ yields

$$\psi(-R) = -\phi_1\sin(\alpha/2) + \phi_2\cos(\alpha/2)$$

In electronic structure theory the ψ's are adiabats and the ϕ's are diabats. Equation (11.13) is a spin representation. Consequently, $\psi_1(\psi_2)(2\pi) = -\psi_1(\psi_2)(0)$. Also,

$$\partial_\alpha\psi_1(\alpha) = \tfrac{1}{2}\psi_2(\alpha) \tag{11.14}$$

and

$$\partial_\alpha\psi_2(\alpha) = -\tfrac{1}{2}\psi_1(\alpha)\,. \tag{11.15}$$

Note that $\langle\psi_1|\nabla\psi_1\rangle = \langle\psi_2|\nabla\psi_2\rangle = 0$ everywhere on C except at $\alpha = 2\pi$, *i.e.*, at the point where the wave function's phase is discontinuous. In this case, the geometric phase is calculated at the close of the circuit. For example,

$$\gamma_1(C) = i\oint_C d\boldsymbol{R}\cdot\langle\psi_1|\nabla\psi_1\rangle \tag{11.16}$$

$$= \int_{2\pi-\varepsilon/2}^{2\pi+\varepsilon/2} d\alpha\, i\psi_1^*\partial_\alpha\psi_1 \tag{11.17}$$

At $\alpha = 2\pi$, ψ_1 undergoes a phase change of π, so we need to account for its complex character in the infinitesimal region at the end of the circuit. Thus, we write $\partial_\alpha(\psi_1 e^{i\phi}) = i\psi_1\partial_\alpha\phi$, with ϕ accounting for the sign change of ψ_1. Setting $\delta\phi = \pi$ for small $\delta\alpha$ [*i.e.*, the ε in eqn (11.17)] yields $\gamma(C) = -\pi$.

Alternatively, the wave function can be gauge transformed. In this case, it is multiplied throughout by $e^{i\alpha/2}$ in order to make it single-valued.[12] Using $\langle\tilde{\psi}_1|\nabla\tilde{\psi}_1\rangle = i\nabla\alpha/2 = i\hat{e}_\alpha/2$ in eqn (11.17) yields

$$\gamma(C) = i\oint_C \left(d\alpha\,\hat{e}_\alpha\right)\cdot\left(i\tfrac{1}{2}\hat{e}_\alpha\right) \tag{11.18}$$

$$= -\tfrac{1}{2}\int_0^{2\pi} d\alpha \tag{11.19}$$

$$= -\pi \tag{11.20}$$

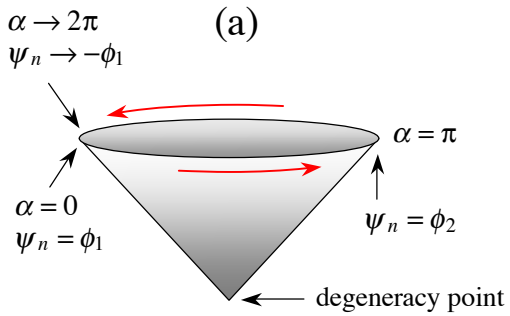
These two options are illustrated in Fig. 17.

---

[12] To verify that the transformed wave function is single-valued, multiply ψ_1 in eqn (11.13) by $e^{i\alpha/2}$ and write $\cos\alpha/2$ and $\sin\alpha/2$ in terms of exponentials:

$$\tilde{\psi}_1 = e^{i\alpha/2}\left(\tfrac{1}{2}\left(e^{i\alpha/2}+e^{-i\alpha/2}\right)\phi_1 - i\tfrac{1}{2}\left(e^{i\alpha/2}-e^{-i\alpha/2}\right)\phi_2\right) = \tfrac{1}{2}\left((e^{i\alpha}+1)\phi_1 - i(e^{i\alpha}-1)\phi_2\right).$$

This is clearly single-valued in α.



$$\psi_n = \cos\tfrac{1}{2}\alpha\ \phi_1 + \sin\tfrac{1}{2}\alpha\ \phi_2$$

Gauge transformation: $\psi_n' = \psi_n e^{i\alpha/2}$

$$\psi_n' = \left(\cos\tfrac{1}{2}\alpha\ \phi_1 + \sin\tfrac{1}{2}\alpha\ \phi_2\right) e^{i\alpha/2}$$

Figure 17. The upper adiabat in the region of a conical intersection is indicated for the example given by eqn (11.12). The electronic wave function is parallel transported as α goes from 0 to 2π, encircling the degeneracy point in the process. (a) The wave function is real everywhere, except where the circuit closes. As 2π is approached, the wave function is necessarily complex, yielding $-\pi$ upon integration through the phase discontinuity. (b) The gauge-transformed wave function, $\psi_n' = \psi_n e^{i\alpha/2}$, is complex and single-valued everywhere on the circuit. The evaluation of the integral of $i\langle n|\nabla n\rangle$ yields $-\pi$, and there is no phase discontinuity at the end.

Figure 17 illustrates adiabatic (parallel) transport over a 2π circuit of the parameter α in eqn (11.13). Referring to Fig. 17(a), the phase of the adiabat ψ_n changes abruptly by π when the circuit closes. Therefore ψ_n is complex. In other words, it must vary as $e^{i\eta(\alpha)}$, where $\eta(\alpha)$ is nonzero only in the infinitesimal region at the end of the circuit. Because the adiabat is real from 0 to $2\pi^{(-)}$, where the minus sign indicates the infinitesimal region before the circuit closes, $\langle n|\nabla n\rangle$ vanishes throughout this region. However, in the infinitesimal region where the circuit closes, the adiabat varies as $e^{i\eta(\alpha)}$, with $\eta(\alpha)$ going from 0 to π. Integration over this infinitesimal region gives $-\pi$.

Referring to Fig. 17(b), we see that this phase discontinuity can also be dealt with by using gauge transformation. With ψ_n multiplied by $e^{i\alpha/2}$ the adiabat is single-valued everywhere, including where the circuit closes. Integration yields $-\pi$ straightaway, and no phase discontinuity is encountered upon closing the circuit.

In each of the two cases indicated in Fig. 17, the wave function is complex. It is just a matter of how its complex character is distributed. In (a) its complex character is concentrated at the end of the circuit, whereas in (b) it is distributed throughout the circuit. In most applications it is desirable to use single-valued wave functions. Single-valuedness is also preferred on conceptual grounds. For example, though Stokes' theorem works with a vector field whose contribution is concentrated at the end of a closed circuit, it is not as easily visualized as with a vector field that is single-valued.

Thus, the equivalence of the approaches has been demonstrated. The geometric phase value of $-\pi$ (which could have been written down immediately from $\gamma(C) = \Omega/2$) has been obtained by integration on C both with and without a single-valued wave function.

This illustrates that the gauge transformation affects how the calculation of the geometric phase is carried out, but not the result.

In the above, $e^{i\alpha/2}$ ensured single-valuedness, and in so doing added what appears to be an angular momentum of ½ (of course, ½ $\hbar$) to the nuclear dynamics. Suffice it to say that α depends on the Hamiltonian in a manner such that as α goes from 0 to 2π the system moves around a conical intersection. The ½ is an *intrinsic angular momentum* that arises because of the ψ_1 / ψ_2 degeneracy. It is present at energies well away from the degeneracy point.

Non-single-valued wave functions arise when systems are not isolated. As mentioned earlier, no fundamental particle is isolated, nor is a system that is coupled to external parameters isolated. In the case of an adiabat, the electron and nuclear microcosms are not isolated from one another. They communicate through the connection field. Therefore they can separately have non-single-valued wave functions, whereas the total wave function needs to be taken as single-valued if the molecule is to be treated as isolated.

It is intuitive that integrations of quantities such as $(q/c)dx^{\nu}A_{\nu}$ and $d\boldsymbol{R}\cdot i\langle n|\nabla n\rangle$ yield phases, as these are actions. The question nonetheless remains: How can $\langle n|\nabla n\rangle$ be visualized? The math is beyond reproach, but how can the gauge connection field that follows from the adiabatic approximation be understood qualitatively?

12. Interpretation in Terms of Gauge Field Theory

From the perspective of a U(1) gauge field theory, a local phase transformation of the adiabat ψ_n on $\boldsymbol{R}$ requires the presence of a gauge field that undergoes a concerted transformation. We have seen that this gauge field turns out to be $i\langle n|\nabla n\rangle$. Specifically, the gauge principle requires that $\psi_n \to \psi_n e^{i\zeta}$ be accompanied by the addition of $\nabla\zeta$ to $i\langle n|\nabla n\rangle$. No physical effect is introduced into the mathematics when this prescription is followed, because it results in positive and negative $\nabla\zeta$ terms that cancel.

In this section, we shall see that the addition of $\nabla\zeta$ to the gauge field is the direct result of a correlated local phase transformation of ψ_n's partner wave function χ_n. This accounts for how and why the gauge principle works in this system. For a gauge field theory to apply it is necessary that there exist correlated gauge transformations. In quantum electrodynamics, local phase transformation of a particle's wave function is paired with gauge transformation of the field A^ν of electromagnetism. There is a common redundancy insofar as $|\psi_n|$ being impervious to the phase transformation $\psi_n \to \psi_n e^{i\zeta}$, and the fields $\vec{B}$ and $\vec{E}$ being impervious to the correlated changes of A^ν incurred through $\psi_n \to \psi_n e^{i\zeta}$. In the adiabatic case, the partnered gauge transformations are each of the first kind: local phase transformations. They are not independent. A phase transformation of one wave function imposes on the partnered wave function a correlated phase transformation.

The fact that invariances and covariances are achieved through the gauge transformation in which ψ_n is altered in concert with the gauge field $i\langle n|\nabla n\rangle$ is traced to the isolated molecule assumption. An isolated molecule cannot couple *as a whole* to a gauge field.[13] The molecule's total wave function must therefore be single-valued. That is, multiplication of the total wave function by $e^{i\zeta}$, where ζ varies locally, is forbidden. For a given adiabat described by ψ_n and a given vibrational level described by χ_n, the total wave function $\chi_n\psi_n$ cannot be gauge transformed. This ensures registry between the electronic structure and field theory pictures, as discussed below.

Until now emphasis has been mainly on ψ_n. It is assumed that ψ_n is not coupled to other adiabats, and the adiabatic separation of the nuclear and electron degrees of freedom has rendered ψ_n blind, insofar as dynamical processes are concerned, to the relatively lethargic gestures of the nuclei. Nonetheless, there are two quantum mechanical systems: electrons with wave function ψ_n, and nuclear degrees of freedom with wave function χ_n. They come together as $\psi_{total} = \chi_n\psi_n$.

It is significant that ψ_n and χ_n can each undergo phase transformations as long as they are correlated: $\psi_n \to \psi_n e^{i\zeta}$ and at the same time $\chi_n \to \chi_n e^{-i\zeta}$. On the one hand, this seems like an interesting way to say that nothing happened, as the phase of ψ_{total} is unaffected. On the other hand, because of this synchrony, ψ_n and χ_n can separately obey the gauge principle. Let us now see how the gauge principle and the $e^{i\zeta}$ / $e^{-i\zeta}$ synchrony are related.

---

[13] The electromagnetic gauge field that brings about absorption and emission of radiation acts on the charged constituents of the molecule.

Correlated Phase Transformations

The gauge principle tells us that the phase transformation $\psi_n \to \psi_n e^{i\zeta}$ must be accompanied by the addition of $\nabla\zeta$ to the gauge field $i\langle n|\nabla n\rangle$. Let us put this to the side for the time being, with the understanding that we will return to it later. Right now, let us concentrate on the fact that $\psi_n \to \psi_n e^{i\zeta}$ must be accompanied by $\chi_n \to \chi_n e^{-i\zeta}$. We shall see that this partnership manifests as the apparent addition of $\nabla\zeta$ to the gauge field $i\langle n|\nabla n\rangle$.

To see what is going on, first operate on the total wave function with a gradient in the space of nuclear coordinates:

$$\nabla\left(\chi_n\psi_n\right) = \psi_n\nabla\chi_n + \chi_n\nabla\psi_n . \tag{12.1}$$

Left multiplication by ψ_n^* and integration over electron coordinates yields the effective $\boldsymbol{R}$-space operator

$$\left(\nabla + \langle n|\nabla n\rangle\right)\chi_n . \tag{12.2}$$

The covariant derivative is thus identified: $\nabla + \langle n|\nabla n\rangle$. Introduction of the correlated phase transformations $\chi_n \to \chi_n e^{-i\zeta}$ and $\psi_n \to \psi_n e^{i\zeta}$ yields

$$\left(\nabla + \langle n|e^{-i\zeta}\nabla e^{i\zeta}|n\rangle\right)\chi_n e^{-i\zeta} = e^{-i\zeta}\left(\nabla - i\nabla\zeta + \langle n|\nabla|n\rangle + i\nabla\zeta\right)\chi_n \tag{12.3}$$

$$= e^{-i\zeta}\left(\nabla + \langle n|\nabla n\rangle\right)\chi_n . \tag{12.4}$$

The operation transforms covariantly.

Notice that there has been no seemingly *ad hoc* addition to the gauge field according to the prescription of the gauge principle. Only the isolated molecule assumption has been enlisted. From the perspective of ψ_n it appears that an addition has been made to the gauge field, because ψ_n knows nothing about χ_n. The gauge field in a sense is a communication link between the two systems. The gauge field passes information to ψ_n: nothing bad is going to happen with the physics as a consequence of ψ_n dressing itself with a fancy phase. The situation with χ_n is similar. It undergoes the phase transformation $\chi_n \to \chi_n e^{-i\zeta}$ and discovers that the gauge field has been altered in concert in such a way that nothing bad happens with the physics. This is illustrated in Fig. 18.

Equations (12.3) and (12.4) illustrate the mechanism whereby expression (12.2) transforms covariantly. It is true that the nuclei are oblivious to rapid electron dynamical processes, and the electron system is oblivious to the lethargic gestures of the nuclei. Nonetheless, the electronic state's adiabatic evolution and $\boldsymbol{R}$-space dynamical processes are symbiotic. All of this follows from the single-valuedness of ψ_{total}, which imposes $e^{i\zeta}/e^{-i\zeta}$ synchrony.

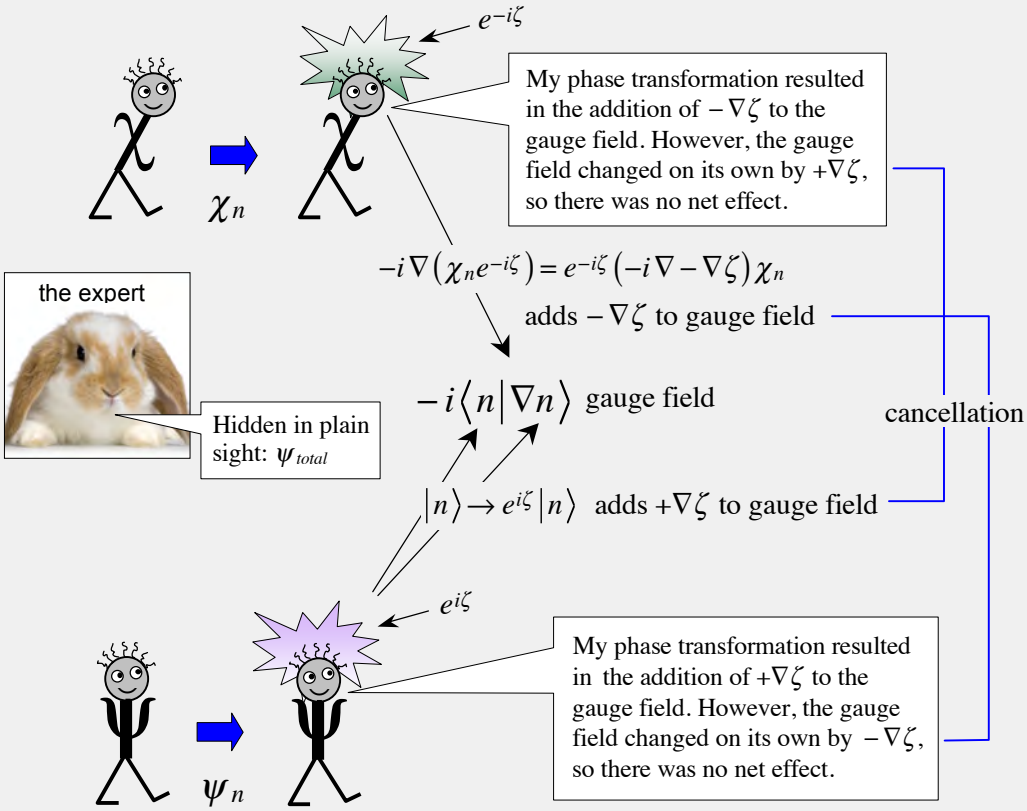


Figure 18. Wave function χ_n is slow, unable to discern ψ_n's frantic gesticulations. She senses only how ψ_n follows her movements on average. She knows that her phase transformation affects her kinetic momentum, which creates a major problem. Miraculously, however, the gauge field apparently undoes this effect. The gauge field tries to convey to χ_n what really happened, but is unable to get the message across. Wave function ψ_n sees χ_n as lethargic to the point of comatose. She knows that her phase transformation affects the gauge field by adding $+\nabla\zeta$, which will affect her geometric phase on $\boldsymbol{R}$. Miraculously, however, the gauge field apparently has undone this effect. The gauge field tries to convey to ψ_n what happened, but fails to get across.

We began our foray into gauge field theory by identifying the perfect registry that exists between redundancies in quantum mechanics and electromagnetism. In the present section we have done something similar. It is not possible to gauge a molecule as a whole, which in effect is the isolated molecule ansatz. Complementary phase transformations take place in the fast and slow spaces created by the adiabatic separation. The covariant derivative is comprised of a gradient on the nuclear space plus the gauge field. The latter is a matrix element in the electron space and a vector in the nuclear space.

Registry is perfect. It is no coincidence that $\nabla + \langle n|\nabla n\rangle$ resembles its electrodynamics counterpart: $\partial_\nu - i(q/c)A_\nu$. The bottom line is that apparent U(1) gauge field theories work for χ_n and ψ_n. The gauge field links these systems, which are created through the adiabatic separation, ensuring that overall gauge symmetry is preserved.

Analogy with Gravity

Einstein noted that a ball falling in a room that is present in a gravitational field behaves in the same way as if the room were in gravity-free space, but accelerated by an equivalent external force. The room is an inertial frame insofar as its contents. This led ultimately to the curved spacetime of general relativity, in which curvature is due to the presence of a field whose source is mass. This is analogous to a local phase transformation of ψ_n on $\boldsymbol{R}$ and the role of the gauge field. Namely, a local phase change appended to a particle wave function is not distinguishable from the same phase change incurred through the presence of a field in which the particle moves.

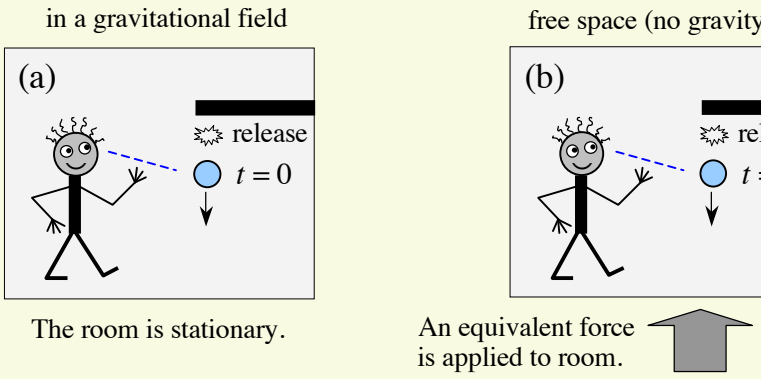


Figure 19. (a) A person observes a ball drop in a room in a gravitational field. (b) The situation in (a) is indistinguishable from one in which the room is in gravity-free space, with the force acting on the room equivalent to that of the gravitational field.

The situation in which local phase transformation is balanced by an addition to the gauge field can be likened to a perception on ψ_n's part that derives from the adiabatic approximation. Likewise, χ_n, being unaware of ψ_n, notices that its phase change is balanced by an addition to the gauge field, as indicated in Fig. 18. The gauge field serves as the communication link between the fast and slow microcosms.

13. Summary

- The Aharonov-Bohm effect is a double slit experiment. If a measurement is carried out in which the passage of a charged particle causes the deflection of a device that contains $\vec{B}$ or $\vec{E}$, interference is eliminated because one of the slits in the double slit arrangement is blocked. Thus, the Aharonov-Bohm effect is a *requirement* of the quantum theory. Without it quantum mechanics would not be consistent.
- The gauge field per unit charge $\vec{A}$ is a momentum of sorts that acts on the electron as an external parameter. In the AB effect, $\vec{A}$ is not quantized. In Chapter 2, we saw how to quantize the transverse electromagnetic field (photons), whereas the $\vec{A}$ encountered here is longitudinal (its curl vanishes, not its divergence). This $\vec{A}$ can be quantized in an interesting way when it is the immediate consequence of quantized electron momentum. This is also discussed in Appendix 3: *Flux Quantization in Superconductivity*.
- In the Aharonov-Bohm effect the electron passes slowly through the region where the gauge field is nonzero. Were this passage rapid, the electron would sense significant transient electric and magnetic fields. For example, in the magnetic Aharonov-Bohm effect, where the electron passes a solenoid that contains $\vec{B}$, the electron would sense an electric field $-\partial_{ct}\vec{A}$. This can also be made clear by placing the electron in a box and transporting the box around the cylinder, first slowly and then quickly. The electromagnetic fields thus generated have transverse components whose quanta are photons. Thus, transitions between levels can take place, assuming of course that the transient field has Fourier components whose frequencies match those of the electron in the box.
- In electrodynamics (Aharonov-Bohm effect), the gauge field times q is the connection whose integral gives the geometric phase. The magnetic version has been verified experimentally. In the molecular case, the connection $i\langle n|\nabla n\rangle$ advances the adiabat on $\boldsymbol{R}$. In the case of two intersecting adiabats, spinor character of each adiabat accounts for its geometric phase. The spinor character is a consequence of the topology engendered by the intersecting potentials. How it is manifest in a system's geometric phase depends on the chosen path, notably, enclosing the degeneracy versus not enclosing it.
- The adiabatic approximation does a great deal of "gauge fixing," with topological and gauge theoretical consequences. The fact that ψ_{total} is single-valued eliminates uncorrelated phase transformations of ψ_n and χ_n. This is in accord with the fundamental equations of electronic structure theory. Phase transformation of ψ_n (or χ_n) is compatible with the gauge principle, as this subsumes the $e^{i\zeta}/e^{-i\zeta}$ synchrony. The overall picture is summarized in Fig. 20.
- A nice thing about the adiabatic approximation is that it permits one to look into why a gauge field theory works in a familiar setting that can be understood without undue difficulty. Molecular quantum mechanics, with its $e^{i\zeta}/e^{-i\zeta}$ synchrony requirement, is no less of a gauge field theory than QED. It is one that can be understood in detail.

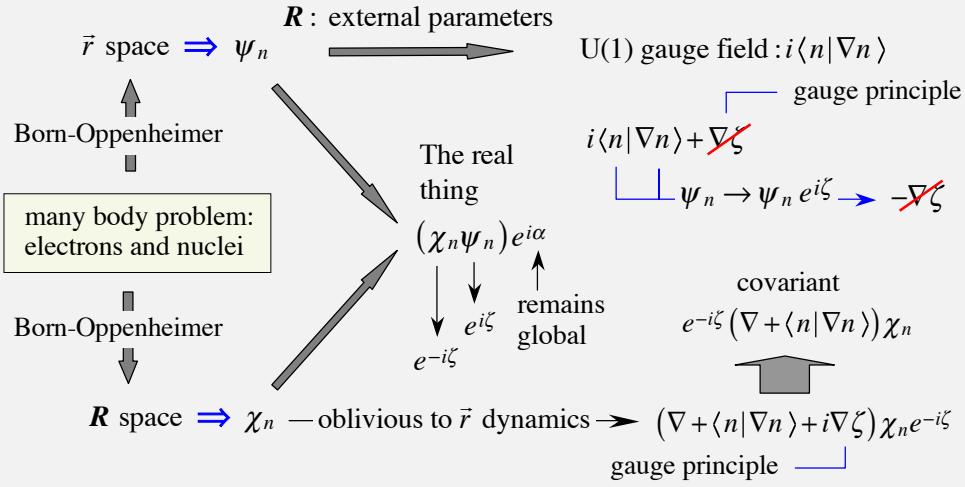


Figure 20. When ψ_n perceives its adiabatic transport as effected by slowly varying external parameters, it is blind to χ_n. It obeys a U(1) theory. Addition of $\nabla\zeta$ to the gauge field is, in fact, a manifestation of ψ_{total}'s single-valuedness ($e^{i\zeta}$ / $e^{-i\zeta}$ synchrony). Likewise, with χ_n oblivious to ψ_n, it obeys a U(1) theory: $\chi_n \rightarrow \chi_n e^{-i\zeta}$ and the simultaneous addition of $-\nabla\zeta$ to the gauge field. The fact that the product $\chi_n\psi_n$ cannot be gauged explains the relationship between the apparent invariance / covariance of the limiting cases and electronic structure theory.

Magnetic Monopole: A Mathematical Curiosity

If $\vec{A} = \nabla_\alpha \zeta = \hat{e}_\alpha / 2\rho$ is a vector potential, its curl is a magnetic field. But $\nabla \times \vec{A}$ is zero everywhere *except the origin*. To obtain $\vec{B}$, use Stokes' theorem:

$$\oint_C d\vec{l} \cdot \nabla\zeta = \oint_C (\rho\, d\alpha\, \hat{e}_\alpha) \cdot (\hat{e}_\alpha / 2\rho) = \pi = \iint_C d\vec{S} \cdot \vec{B}$$

The fact that the circuit integral of $\vec{l} \cdot \nabla\zeta$ is π, whereas $\vec{B} = \nabla \times \nabla\zeta$ vanishes everywhere except the origin means that $\vec{B}$ is located at the origin. It is proportional to a delta function at the origin – a monopole!

Bibliography and References

These are germane to Chapter 3 and Appendices 3-5.

1. J. D. Jackson, *Classical Electrodynamics*, Third Edition (Wiley, New York, 1999).
2. E. M. Purcell and D. J. Morin, *Electricity and Magnetism*, Second Edition (McGraw Hill, New York, 2013).
3. I. J. R. Aitchison and A. J. G. Hey, *Gauge Theories in Particle Physics*, Volumes I and II (Taylor and Francis, New York, 2004).
4. M. Guidry, *Gauge Field Theories: An Introduction with Applications* (Wiley Interscience, New York, 1999).
5. G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, 7th Edition (Academic Press, New York, 2012).
6. B. Kusse and E. Westwig, *Mathematical Physics: Applied Mathematics for Scientists and Engineers* (Wiley, New York, 2006).
7. D. W. Henderson and D. Taimina, *Experiencing Geometry* (Pearson Prentice Hall, Upper Saddle River, New Jersey, 2005).
8. B. Schutz, *Geometric Methods of Mathematical Physics*, Cambridge University Press, Cambridge, 1999).
9. H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*, Third Edition (Pearson, New York, 2002).
10. J. B. Marion and S. T. Thornton, *Classical Dynamics*, Fourth Edition (Harcourt, New York, 1995).
11. M. Peshkin and A. Tonomura, *The Aharonov-Bohm Effect* (Springer Verlag, Heidelberg, 1989).
12. Y. Aharonov and D. Rohrlich, *Quantum Paradoxes* (Wiley VCH, Weinheim, Germany, 2005).
13. A. Shapere and F. Wilczek, *Geometric Phases in Physics* (World Scientific, Singapore, 1989).
14. A. Bohm, A. Mostafazadeh, H. Koizumi, Q. Niu, and J. Zwanziger, *The Geometric Phase in Quantum Systems* (Springer Verlag, Berlin, 2003).
15. W. Domcke, D. R. Yarkony, and H. Köppel, *Conical Intersections: Electronic Structure, Dynamics, and Spectroscopy* (World Scientific, Singapore, 2004).
16. G. Baym, *Lectures on Quantum Mechanics* (Addison Wesley, New York, 1990).
17. J. J. Sakurai, *Modern Quantum Mechanics* (Addison Wesley, New York, 1994).
18. D. J. Griffiths, *Introduction to Quantum Mechanics* (Prentice Hall, Upper Saddle River, NJ, 1995).
19. F. Schwabl, *Quantum Mechanics* (Springer Verlag, Berlin, 1995).
20. E. Merzbacher, *Quantum Mechanics*, Third Edition (Wiley, New York, 1998).
21. J. R. Schrieffer, *Theory of Superconductivity* (Addison Wesley, Frontiers in Physics, Fourth Printing, 1988).

22. M. V. Berry, *Quantal phase factors accompanying adiabatic changes*, Proc. R. Soc. London A **392**, 45-57 (1984).
23. Y. Aharonov and D. Bohm, *Significance of electromagnetic potentials in the quantum theory*, Phys. Rev. **115**, 485-490 (1959).
24. W. Meissner and R. Ochsenfeld, Naturwiss. **21**, 787 (1933).
25. H. Kammerlingh Onnes, Comm. Phys. Lab. Univ. Leiden, nos. 119, 120, 122 (1911).
26. J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **106**, 162 (1957); **108**, 1175 (1957).
27. F. London, Superfluids, Vols. I and II (Wiley, New York, 1950).
28. L. N. Cooper, Phys. Rev. 104, 1189 (1956).
29. F. London and H. London, Proc. Roy. Soc. (London) **A216**, 71 (1935); Physica, **2**, 341 (1935).
30. L. Onsager, Phys. Rev. Lett. **7**, 50 (1961).
31. V. L. Ginsburg and L. D. Landau, J. Exptl. Theoret. Phys. (USSR) **20**, 1064 (1950).
32. B. S. Deaver, Jr., and W. M. Fairbank, Phys. Rev. Lett. **7**, 43 (1961).
33. M. V. Berry, J. Phys. A: Math. Gen. **30**, 8355 (1997).
34. B. D. Josephson, Phys. Rev. Lett. **1**, 251 (1962).
35. H. Frölich, Phys. Rev. **79**, 845 (1950).
36. J. Bardeen, Rev. Mod. Phys. **23**, 261 (1951).
37. M. R. Schafroth, Helv. Phys. Acta **24**, 645 (1951).
38. J. Preskill, *Chromatic Aberrations: Yang and Mills meet Aharonov and Bohm*, Lecture notes, 1993.
39. N. Byers and C. N. Yang, Phys. Rev. Lett. **7**, 46 (1961).
40. B. S. Deaver, Jr. and W. M. Fairbank, *Experimental Evidence for Quantized Flux in Superconducting Cylinders*, Phys. Rev. Lett. **7**, 43 (1961).
41. W. Ehrenberg and R. E. Siday, *The Optical Properties of Axially Symmetric Magnetic Prisms*, Proc. Phys. Soc. **59**, 1036 (1949).
42. C. N. Yang and R. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, Phys. Rev. **96**, 191-195 (1954).

Exercises

1. Following eqn (5.6) is the statement: "… this result is valid only in regions where $\vec{B} = \nabla \times \vec{A} = 0$." Prove that this is true.

2. Referring to Fig. 2, show that the only nonzero component of $\vec{A}$ throughout the region $r > a$ is A_ϕ.

3. In the electric version of the AB effect, the particle wave components are each sent through a hollow conducting tube as shown in the figure (under construction).

3. Consider the gauge transformation: $\vec{A} \rightarrow \vec{A} + \nabla\chi$ and $\psi \rightarrow \psi e^{iq\chi/\hbar}$. The position operator $\vec{r}$ is invariant with respect to this transformation. Moreover, the canonical momentum operator $-i\hbar\nabla$ retains the same form it had before the transformation, thereby satisfying $[x, p] = -i\hbar$. Show that the expectation value of p is not invariant with respect to the gauge transformation.

5. Show that $\boldsymbol{\pi}$ is invariant with respect to: $\boldsymbol{A} \rightarrow \boldsymbol{A} + \nabla\zeta(\boldsymbol{r})$, $\psi \rightarrow \psi \exp(iq\zeta(\boldsymbol{r})/\hbar)$.

6. Consider the gauge transformation: $\boldsymbol{A} \rightarrow \boldsymbol{A} + \nabla\zeta(\boldsymbol{r})$ and $\psi \rightarrow \psi e^{iq\zeta(\boldsymbol{r})/\hbar}$. Note that $c = 1$ is used. The operator $\boldsymbol{\pi} = \boldsymbol{p} - q\boldsymbol{A} - q\nabla\zeta(\boldsymbol{r})$ acting on the new wave function is the same as $\boldsymbol{\pi} = \boldsymbol{p} - q\boldsymbol{A}$ acting on ψ. Why would one use the former? It has one more term than the latter and therefore involves more math. Is it not more complicated? Discuss this in terms of the Born-Oppenheimer approximation.

7. In this problem, you will consider a charged particle that is present in a time independent, spatially homogeneous magnetic field: how this appears in the Hamiltonian and some likely consequences of its presence.

(a) Show that the vector potential $\boldsymbol{A}$ is given by $-\frac{1}{2}(\boldsymbol{r} \times \boldsymbol{B})$.

(b) Assume $\boldsymbol{B} = B\hat{z}$. A charged particle is subject to a potential energy V that is important overall but plays no role in the problem under consideration. In other words, we are only interested in effects due to $\boldsymbol{B}$. Derive a Hamiltonian and put it in a compact form in terms of $H_0 = p^2/2m + V$ and H', where H' is due to $\boldsymbol{B}$.

(c) The term linear in B is paramagnetism. It is responsible for what is called the normal (no spin) Zeeman effect. The term proportional to B^2 is diamagnetism. Assume that B is small so its effects can be treated as perturbations. In the small-B limit, the linear term dominates. Use a sensible basis and write the eigenvalues for this limit. Then add the quadratic term and explain how it can be treated. Comment on the result.

8. Following eqn (10.18) is the statement: "Some algebra yields $\boldsymbol{V}_m = (m/B^2)\hat{\boldsymbol{e}}_{3'}$." Fill in these steps.

9. Consider the Zeeman effect for a one-electron-type atom having $l = 1$ and $s = ½$, *e.g.*, atomic hydrogen with $n = 2$. The Zeeman Hamiltonian is $H_Z = A(\boldsymbol{l} + 2\boldsymbol{s}) \cdot \boldsymbol{B}$, where A is a negative real constant and $\boldsymbol{B}$ is the external magnetic field. Calculate the geometric phase for a closed circuit that encloses the $\boldsymbol{B} = 0$ degeneracy point. Do this for the $|1,+\rangle$ and $|0,+\rangle$ states, where 1 and 0 denote m_l values, and + denotes $m_s = +½$. The geometric phases for the other H_Z eigenstates can be inferred. Results were obtained in the text for integer and odd-half-integer spin, where the gradient of the Hamiltonian is proportional to the spin vector. Here, $H_Z = A(\boldsymbol{l} + 2\boldsymbol{s}) \cdot \boldsymbol{B}$ is examined to see if the interesting factor of two has an effect, and if so, what is it.

10. The $n = 2$ state of atomic hydrogen has an eightfold degeneracy, not counting nuclear spin. Neglect nuclear spin. That is, $\sum (2l + 1)(2s + 1)$ for $s = 1/2$ and $l = 0$ and 1 is equal to 8. The summation is over l values. An H atom is present in a steady electric field $\boldsymbol{E} = E_0 \mathbf{e}_z$. To begin, neglect spin-orbit interaction. The Stark Hamiltonian is $H_S = -\boldsymbol{\mu} \cdot \boldsymbol{E} = e\boldsymbol{r} \cdot \boldsymbol{E}$, where $\boldsymbol{\mu} = q\boldsymbol{r} = -e\boldsymbol{r}$ is the electric dipole moment operator. Thus, $H_S = erE_0 \cos\theta$. When $\boldsymbol{E} = 0$, the $n = 2$ levels are degenerate, but when $\boldsymbol{E} \neq 0$, this degeneracy is lifted for $2p_z$ and $2s$, which mix, creating a large dipole moment. This problem should be treated as a project that might require several days.

(a) The magnitude of $\boldsymbol{E}$ remains constant, but $\boldsymbol{E}$ now slowly changes its orientation relative to the laboratory reference frame. Its tip traces a closed circuit C whose surface contains the $\boldsymbol{E} = 0$ degeneracy point. Make a picture that shows how the dipole moment follows $\boldsymbol{E}$. Discuss this system in the context of geometric phase. How much geometric phase has accrued upon completion of the circuit? Is it possible to deduce the answer without resorting to lengthy calculation? If so, how? What is the answer?

(b) Now introduce spin-orbit interaction. Examine the limits of large and small E_0 and calculate the energies. In these respective limits, the Stark effect and spin-orbit interaction dominate, and the weaker of the two interactions can be treated perturbatively. Again, discuss the system within the context of geometric phase.

(c) Discuss qualitatively the case $\boldsymbol{E} \neq 0$ and $\boldsymbol{B} \neq 0$, where $\boldsymbol{B}$ is a magnetic field; ignore spin-orbit interaction. Each field can be varied independently, both magnitude and orientation. What are the external parameters? What is the dimension of the $\boldsymbol{R}$-space?

12. In Appendix 5, it was stated in the box entitled *Coriolis Force* that $x(t)$ and $y(t)$ have a phase difference of $\pi/2$. Show that this is true.

13. An astronaut goes to Mars and brings along a Foucault pendulum. Once on the planet, the astronaut measures the precession angle and finds that in 24 hours the pendulum has precessed 180° in the clockwise direction. Where is the pendulum located relative to the Martian equator and North Pole?

14. Consider vector transport on the surface of a sphere (Appendix 5). Using a model sphere, review the derivation of the Gauss-Bonnet formula and the holonomy of a triangle. Extend this to a closed four-sided path, which obviously can be made from two triangles joined by a common side. Now extend it to a regular polygon of arbitrary shape.

15. (Courtesy of Phil Pechukas) The adiabatic theorem has to do with motion under a time dependent Hamiltonian, and it says (roughly) once in an eigenstate, always in an eigenstate, provided the Hamiltonian changes sufficiently slowly. Let $H(\lambda)$, $0 \le \lambda \le 1$, be a Hamiltonian that depends smoothly on a parameter λ. The adiabatic theorem considers the problem

$$i\hbar\,\dot{\psi}_\varepsilon(t) = H_\varepsilon(t)\,\psi_\varepsilon(t) \tag{i}$$

where $H_\varepsilon(t) = H(\varepsilon t)$, and then looks at the limit $\varepsilon \to 0$. Suppose instead we set $\varepsilon = 1$ and vary $\hbar$, looking at the problem:

$$i\hbar\,\dot{\phi}_\hbar(t) = H(t)\,\phi_\hbar(t) \tag{ii}$$

as $\hbar \to 0$. Let $\hbar_0$ denote the real Planck's constant. Show that $\phi_\hbar(t) = \psi_{\hbar/\hbar_0}(\hbar_0 t/\hbar)$. Thus, the adiabatic limit is equivalent to the quasi-classical limit $\hbar \to 0$.

16. (Courtesy of Phil Pechukas) A magnetic field $B(t)$ lies in the x-y plane and slowly precesses around the z-axis: $B_z(t) = 0$; $B_x(t) = B\cos\omega t$; $B_y(t) = B\sin\omega t$. The Hamiltonian for a spin-1/2 particle in the field looks like this:

$$H(t) = \frac{\Delta}{2}\begin{pmatrix} 0 & e^{-i\omega t} \\ e^{i\omega t} & 0 \end{pmatrix} \tag{i}$$

where the energy splitting Δ is proportional to the field strength B. Solve the Schrödinger equation

$$i\hbar\begin{pmatrix} \dot{c}_1(t) \\ \dot{c}_2(t) \end{pmatrix} = H(t)\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix} \tag{ii}$$

exactly for the case $c_1(0) = c_2(0) = 1$; using 1 instead of $1/\sqrt{2}$ simplifies manipulations and is fixed trivially later. Hint: define $\tilde{\boldsymbol{c}}$ by writing

$$\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix} = \begin{pmatrix} e^{-i\alpha t} & 0 \\ 0 & e^{i\alpha t} \end{pmatrix}\begin{pmatrix} \tilde{c}_1(t) \\ \tilde{c}_2(t) \end{pmatrix} \tag{iii}$$

By a clever choice of α you will get an equation for

$$\begin{pmatrix} \dot{\tilde{c}}_1(t) \\ \dot{\tilde{c}}_2(t) \end{pmatrix} \tag{iv}$$

that you know how to solve. Show that the exact solution stays close to the $+\Delta/2$ eigenvector of $H(t)$ for all time, provided ω is sufficiently small. What is the condition that ω must satisfy to be considered sufficiently small?