PCCP

Cite this: *Phys. Chem. Chem. Phys.*, 2012, **14**, 6409–6432

www.rsc.org/pccp

PERSPECTIVE

Geometric phase and gauge connection in polyatomic molecules†‡

Curt Wittig*

Received 19th September 2011, Accepted 14th December 2011
DOI: 10.1039/c2cp22974a

Geometric phase is an interesting topic that is germane to numerous and varied research areas: molecules, optics, quantum computing, quantum Hall effect, graphene, and so on. It exists only when the system of interest interacts with something it perceives as exterior. An isolated system cannot display geometric phase. This article addresses geometric phase in polyatomic molecules from a gauge field theory perspective. Gauge field theory was introduced in electrodynamics by Fock and examined assiduously by Weyl. It yields the gauge field A^{μ}, particle–field couplings, and the Aharonov–Bohm phase, while Yang–Mills theory, the cornerstone of the standard model of physics, is a template for non-Abelian gauge symmetries. Electronic structure theory, including nonadiabaticity, is a non-Abelian gauge field theory with matrix-valued covariant derivative. Because the wave function of an isolated molecule must be single-valued, its global U(1) symmetry cannot be gauged, *i.e.*, products of nuclear and electron functions such as $\chi_n\psi_n$ are forbidden from undergoing local phase transformation on $\boldsymbol{R}$, where $\boldsymbol{R}$ denotes nuclear degrees of freedom. On the other hand, the synchronous transformations (first noted by Mead and Truhlar): $\psi_n \rightarrow \psi_n e^{i\zeta}$ and simultaneously $\chi_n \rightarrow \chi_n e^{-i\zeta}$, preserve single-valuedness and enable wave functions in each subspace to undergo phase transformation on $\boldsymbol{R}$. Thus, each subspace is compatible with a U(1) gauge field theory. The central mathematical object is Berry's adiabatic connection $i\langle n|\nabla n\rangle$, which serves as a communication link between the two subsystems. It is shown that additions to the connection according to the gauge principle are, in fact, manifestations of the synchronous ($e^{i\zeta}/e^{-i\zeta}$) nature of the ψ_n and χ_n phase transformations. Two important U(1) connections are reviewed: qA^{μ} from electrodynamics and Berry's connection. The gauging of SU(2) and SU(3) is reviewed and then used with molecules. The largest gauge group applicable in the immediate vicinity of a two-state intersection is U(2), which factors to U(1) × SU(2). Gauging SU(2) yields three fields, whereas U(1) is not gauged, as the result cannot be brought into registry with electronic structure theory, and there are other problems as well. A parallel with spontaneous symmetry breaking in electroweak theory is noted. Loss of SU(2) symmetry as the energy gap between adiabats increases yields the inter-related U(1) symmetries of the upper and lower adiabats, with spinor character imprinted in the vicinity of the degeneracy.

1. Introduction

The first item of business is to settle on what is meant by the term "phase" in the context of this article. We shall not be concerned with phase transitions, phase diagrams, and other such things. Rather, the phase to be discussed is the one that appears in an exponent, like the x in e^{ix}. Not to be underestimated, it has been known to vex even the most ardent of bookkeepers. Indeed, it can be subtle to the point of genuine difficulty.

The central theme is the geometric phase associated with the parallel transport of electron wave functions on molecular potential energy surfaces. Seminal papers by C. Alden Mead and Donald Truhlar (1979, 1982)[1,2] inspired renewed enthusiasm in a subject that had been unearthed decades earlier,[3–5] but for the most part had lain dormant. The large amount and high quality of the ensuing research has had a significant and lasting impact on chemical and molecular physics.

A simple manifestation, certainly one of the most common, arises with the conical intersection of two adiabatic potential energy surfaces (adiabats). The conical shape in the intersection region is a consequence of the off-diagonal Hamiltonian matrix elements in a diabatic basis being real. It is well known that encircling the intersection (degeneracy point) on either adiabat results in the wave function acquiring a phase whose magnitude is π.[6,7] This follows immediately in the toy model of

Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA. E-mail: wittig@usc.edu
† This article is part of a special collection of PCCP Perspective celebrating the *International Year of Chemistry*.
‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c2cp22974a

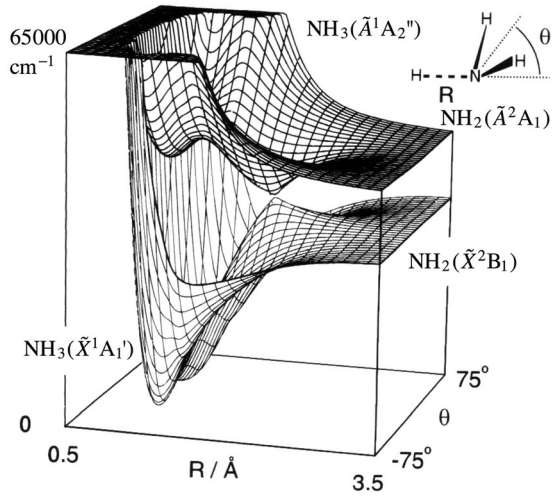

Fig. 1 The NH_3 $\tilde{X}^1A_1'$ and $\tilde{A}^1A_2''$ potential surfaces intersect conically at $\theta = 0$ and $R = 2$ Å. In the plane $\theta = 0$, $\tilde{A}^1A_2''$ correlates diabatically to NH_2 ($\tilde{X}^2B_1$), whereas $\tilde{X}^1A_1'$ correlates diabatically to NH_2 ($\tilde{A}^2A_1$). On the upper and lower adiabats, $\tilde{A}^1A_2''$ and $\tilde{X}^1A_1'$ correlate to NH_2 ($\tilde{A}^2A_1$) and NH_2 ($\tilde{X}^2B_1$), respectively. Geometric phase of $\pm\pi$ accrues for closed circuits on the lower and upper adiabats that encircle the degeneracy point at $\theta = 0$ and $R = 2$ Å. Adapted from ref. 9 and 10; the θ in the figure is unrelated to the θ in eqn (1.1).

a two-level system of diabats ϕ_1 and ϕ_2 coupled by the real matrix element H_{12} to yield the adiabats ψ_1 and ψ_2:[8]

$$\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix} = \begin{pmatrix}\cos\tfrac{1}{2}\theta & \sin\tfrac{1}{2}\theta\\ -\sin\tfrac{1}{2}\theta & \cos\tfrac{1}{2}\theta\end{pmatrix}\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix} \qquad (1.1)$$

where $\theta = \tan^{-1}(H_{12}/G)$, $G = \frac{1}{2}(H_{22} - H_{11})$, and it is assumed that ϕ_1 and ϕ_2 do not depend on θ.

As θ increases from 0, the Hamiltonian matrix elements vary. By the time $\theta = 2\pi$ is reached they have recovered their initial values. On the other hand, the adiabats have each undergone a sign change at the end of this circuit: $\psi_1(2\pi) = -\psi_1(0)$ and $\psi_2(2\pi) = -\psi_2(0)$. This spinor nature of two-state systems that have a degenerate point has been known for a long time. I am loath to reference specific early papers, as it can be traced to the very beginning of quantum mechanics, and even earlier in the mathematics literature.

Despite its utter simplicity, eqn (1.1) applies to a number of molecular systems, and illustrations of conical intersections such as the one shown in Fig. 1 abound. The sign change experienced by the adiabats is a geometric phase. Another phenomenon where polyatomic molecules display geometric phase is the Jahn–Teller effect,[3,4,11–15] in which an assigned symmetry of a degenerate state is often said to be broken spontaneously (an interesting use of language). As noted by a reviewer: "It is just the positions of the adiabatic minima that "lose" symmetry, but potential energy surfaces are not real anyway." This fascinating topic will not be discussed in this article.

Geometric phase that arises as a consequence of an adiabatic approximation is frequently referred to as Berry phase in deference to the seminal contributions of Michael Berry toward identifying and quantifying the relationship between adiabaticity, parallel transport, and geometric phase. Among other things, Berry showed that parallel transport in a three-dimensional (3D) space of slowly varying parameters that effect parallel transport leads to a geometric phase of $\pm\frac{1}{2}\Omega$ for a two-state degeneracy, where Ω is the solid angle subtended by a closed circuit in the parameter space, with the degeneracy point as the origin.[16] This becomes $\pm\pi$ for the conical intersection case indicated in eqn (1.1) and shown in Fig. 1, because Ω is always equal to 2π when the closed circuit contains the origin. When spin–orbit interaction participates, the imaginary part of H_{12} is not zero. However, for molecules comprised of light atoms, this is a small effect. Therefore, curves like the one shown in Fig. 1 are common.

Parameter spaces of dimension higher than three can be handled straightforwardly using differential geometry. However, 3D gets the message across and applies, at least as a first approximation, to most molecules. Thus, we will stick to 3D parameter spaces in which the parameters are nuclear degrees of freedom. Nowadays one can hardly attend a conference in the area of molecular dynamics without encountering the geometric phase engendered *via* the Born–Oppenheimer approximation (BOA). At the same time, the topic of geometric phase admits to intellectual depth that goes well beyond $\pm\frac{1}{2}\Omega$.[17,18]

A couple of decades ago I became interested in geometric phase, as did many scientists studying molecular dynamics. These interests have since evolved, albeit in fits and starts. During most of this time my perspective has been that of traditional chemical physics.[19–31] Recently, however, I began to look at molecular geometric phase from the perspective of gauge field theory. Others,[32–39] notably Mead and the Heidelberg group,[34–39] had done this earlier. These contributions are insightful and profound.

In Mead's monumental review,[36] molecular geometric phase is approached largely from the perspective of parallel transport. The work is masterful but mathematically sophisticated to an experimentalist like myself. Shortly thereafter, Pacher *et al.*[37] published a comprehensive article that included a discussion of gauge field theory aspects of the electronic structure problem. An important application is diabatization, which is not fully achievable beyond diatoms.[2,32–41] These and other publications have embraced aspects of gauge field theory, often through analogy.

Gauge field theory fits well with geometric phase. The insights it provides are compelling and far-reaching, often revealing relationships between systems that might otherwise appear disparate. Given that this is a "Perspective" article, what better venue could there be for a discussion of the gauge field theory perspective on geometric phase in polyatomic molecules! The article is aimed at the chemical physics (physical chemistry) community, including experimentalists. Others might benefit, but focus is needed and this is where it is placed. Nowadays the distinction between experiment and theory is blurred in the sense that experimentalists are comfortable with many aspects of theory and (especially) computation: electronic structure, wave packet propagation, and classical molecular dynamics simulations, to list a few. It is assumed, however, that the reader has had minimal contact with gauge field theory.

Do not be put off by the scary name and the fact that gauge field theory is used in particle physics. With the right approach, it is straightforward. Though most books in this area lean toward mathematical descriptions, this is not the case for the excellent texts by Aitchison and Hey (two volumes),[42,43] and by Guidry.[44] These texts treat the theory in a way that makes it accessible to a broad readership. They will be referred to frequently.

Gauge field theory is introduced through quantum mechanical redundancy. For example, the squared modulus of a particle's wave function ψ yields the probability density for finding the particle. The redundancy is that the phase transformation $\psi \rightarrow \psi e^{i\zeta(\boldsymbol{r},t)}$, where $\zeta(\boldsymbol{r},t)$ is an arbitrary function of $\boldsymbol{r}$ and t, does not affect the probability density. Of course this transformation must be applied simultaneously with another, *e.g.*, in order that the calculated momentum is not altered. This is how the electrodynamics gauge field, $A^{\mu} = (\phi,\boldsymbol{A})$, enters. The superscript will be explained later.

With A^{μ} in hand, the Aharonov–Bohm geometric phase is obtained.[45–47] The scalar ϕ and the vector $\boldsymbol{A}$ are the usual electromagnetic potentials, which together comprise the four-vector A^{μ}. The Aharonov–Bohm phase, especially the magnetic version, serves as an excellent introduction to molecular geometric phase. It is seen that A^{μ} is more fundamental than $\boldsymbol{E}$ and $\boldsymbol{B}$. It yields $\boldsymbol{E}$ and $\boldsymbol{B}$ *via* differentiation: $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ and $\boldsymbol{E} = -\nabla\phi - \partial_t\boldsymbol{A}$. Quantum mechanically, however, A^{μ} appends phase to wave functions in a way that $\boldsymbol{E}$ and $\boldsymbol{B}$ cannot. This is one reason that A^{μ} is more fundamental than $\boldsymbol{E}$ and $\boldsymbol{B}$. Another is that A^{μ} transforms as a Lorentz covariant four-vector, whereas $\boldsymbol{E}$ and $\boldsymbol{B}$ do not. This pedagogical material is essential for what follows.

A mathematical object called the covariant derivative emerges. It enables particle–field couplings to be turned on through a deft move involving the *gauge principle*, in which partial derivatives are replaced with their covariant derivative counterparts. The prescription is: (i) identify a global gauge symmetry (in electrodynamics this is multiplication by $e^{i\zeta}$, where ζ is a real constant), (ii) demand that the symmetry apply locally [in electrodynamics, $\zeta \rightarrow \zeta(\boldsymbol{r},t)$], and (iii) determine the resulting gauge fields. These enter the covariant derivative. The covariant derivative converts the system from one of global gauge invariance to one of local gauge invariance. This basic strategy extends well beyond electrodynamics, *e.g.*, to the other gauge symmetries of the standard model.

The covariant derivative is geometric in nature, containing what is referred to as the connection, whose integral yields geometric phase. Two examples of parallel transport (one classical and one quantum mechanical) strengthen intuition about traipsing on surfaces that are curved from the outset (classical), or result in an effective curvature through the presence of a gauge connection field (quantum mechanical). It is amusing that the classical case, which pertains mainly to general relativity,[48–51] has so much in common with the quantum case.[43]

We will see how parallel transport of classical and quantum vectors yields geometric phase. You might have seen demonstrations of parallel transport on a sphere, often using nothing more than hands and arms. The geometric phase in this case (in mathematics the holonomy) turns out to be the enclosed solid angle, Ω.[52] Thus, it comes as no surprise that the case of two adiabats, each having spinor character, yields $\pm\frac{1}{2}\Omega$. This material is as geometric as it gets. The principles and main results are presented, with details provided in the ESI.‡

The marriage of quantum mechanics and electromagnetism is a U(1) gauge field theory (unitary group of dimension one), as it follows from multiplication by $e^{iq\zeta(\boldsymbol{r},t)}$, where q is the electric charge. As with the time evolution operator e^{-iHt}, where the Hamiltonian is the generator of time evolution, it can be said that electric charge is the generator of the electrodynamics gauge transformation.

Following these primers, the strategy is extended to other gauge symmetries. In particle physics, SU(2) and SU(3) have been examined assiduously, as they are central to theories of the weak and strong forces. In polyatomic molecules these symmetries are relevant to intersections of two and three potential surfaces, respectively. Thus, the procedure for gauging them is presented in anticipation of their use later—SU(2) in Section 4, with the extension to SU(3) given in the Appendix. Comments about particle physics are unavoidable, but by no stretch of the imagination shall we enter this realm in a serious way.

As with electrodynamics, quantum mechanical redundancy leads to gauge fields. However, unlike electrodynamics, gauging SU(2) and SU(3) yields three and eight gauge fields, respectively. The number of gauge fields is equal to the number of generators of the group transformations. This is intuitive: the gauge fields need to act in concert with what would otherwise be an unacceptable phase transformation. We will see that analogous SU(n) gauge (connection) fields appear in the molecular case.

The driving force underlying the theory in all cases is redundancy and complementary gauge transformations of the two parts that act externally to one another. In the present context, "complementary" means that the two parts act in registry with one another, thereby enabling essential covariances and invariances to be achieved. The gauge connection field is the communication link between the two parts. Note that terms such as gauge field, gauge connection, gauge connection field, connection, curvature, and gauge curvature are used somewhat interchangeably with context dictating usage according to imagery. This is not my idea; the literature reads this way.

The above material is essential, albeit lengthier than most presentations of "background material". However, it has been my experience that sending the reader to myriad references, each with its own symbols, style, choice of conventions, and level of rigor, is a recipe for disaster. My own prejudices are bad enough. I have tried to make the article as self-contained as possible.

With these tools in hand, we turn to molecules, starting with Berry's adiabatic connection. The U(1) gauge connection $i\langle n|\nabla n\rangle$ is identified and discussed in terms of parallel transport. It is noted that the global U(1) symmetry of an isolated molecule cannot be gauged, as the molecule's total wave function must remain single-valued. This requirement cannot be met if the molecule couples to an external field. To make matters worse, this external field would have to be other than the electrodynamics gauge field. On the other hand, synchronous phase transformations of electron and nuclear wave functions [*e.g.*, $\psi_n \rightarrow \psi_n e^{i\zeta}$ and $\chi_n \rightarrow \chi_n e^{-i\zeta}$, respectively, where it is understood that $\zeta = \zeta(\boldsymbol{r},t)$] satisfy the isolated molecule ansatz.[1]

When the gauge principle is applied to an adiabat, say ψ_n, the phase transformation $\psi_n \rightarrow \psi_n e^{i\zeta}$ requires simultaneous addition of $\nabla\zeta$ to the gauge field. The system apparently obeys a U(1) gauge theory. It is shown that the addition to the gauge field is, in fact, due to ψ_n's partner χ_n, whose synchronous transformation is $\chi_n \rightarrow \chi_n e^{-i\zeta}$. The isolated molecule assumption is passed between ψ_n and χ_n *via* the gauge connection $i\langle n|\nabla n\rangle$.

The intersection of two adiabats is then examined. The largest global gauge symmetry group applicable there is U(2). It is viable as long as the system is restricted to the immediate vicinity of the degeneracy. At exact degeneracy any linear combination of the adiabats works. Therefore U(2) is an approximate symmetry in the immediate vicinity of the degeneracy. This restricted domain requirement differs from the case of particle physics, whose symmetries, whether exact or approximate, extend throughout all of causally related spacetime. In the molecular case, local variation is limited to a region where adiabats can be freely transformed within limits imposed by a physically motivated criterion.

The fact that U(2) factors to U(1) × SU(2) enables U(1) and SU(2) symmetries to be considered separately. SU(2) is gauged, knowing full well that this symmetry will go away as the system ventures from the immediate vicinity of the degeneracy. U(1) is relatively robust, as it involves multiplication by $e^{i\zeta}$. Yet, it is not gauged, as doing so would introduce egregious problems, including the fact that its gauge field would be incompatible with electronic structure theory. A parallel with electroweak spontaneous symmetry breaking is noted and discussed. The bottom line is that the BO gauge fields belong to SU(2). We will see how Berry's connection follows, and how spinor character is preserved in each adiabat, as it is imprinted through the topology engendered by the potentials. Cases of three adiabats follow along similar lines.[53–57]

A comment is in order about the terms spinor and isospinor. At a degeneracy, two adiabats comprise an isospinor doublet. Gauging SU(2) makes the symmetry local, with the caveat that the system must remain in the immediate vicinity of the degeneracy if SU(2) is to be applicable. On the other hand, each adiabat has spinor character due to the $\boldsymbol{R}$ space topology. This persists well after SU(2) gauge symmetry no longer applies, *i.e.*, well away from the degeneracy. It is also "iso" in the sense of its distinction from fermion spin $\frac{1}{2}$. To lessen confusion, the pair of adiabats is referred to as an isospin doublet, whereas each adiabat is said to have spinor character.

The electronic structure theory of polyatomic molecules, including nonadiabaticity, is a non-Abelian gauge field theory, subject to the isolated-molecule restriction that the global U(1) symmetry of an adiabat's total wave function, $\chi_n\psi_n$, is not gauged. In other words, $\chi_n\psi_n$ cannot undergo local phase transformation: $\chi_n\psi_n \rightarrow \chi_n\psi_n e^{i\zeta}$. Equivalently: ψ_n and χ_n can be phase transformed synchronously ($e^{i\zeta}/e^{-i\zeta}$) while meeting the isolated molecule requirement. The use of a U(1) gauge field theory to describe either ψ_n or χ_n is due to the complementary nature of their phase transformations. The matrix-valued covariant derivative is $\nabla + \boldsymbol{F}$,[37] where $\boldsymbol{F}_{mn} = \langle m|\nabla n\rangle$.

Much (but not all) of this has been discussed previously, and sophisticated methods have been developed for the needed electronic structure calculations. Take the Yarkony group for example. Though not light reading, the work is as thorough as it gets.[23–30,53–55,58] Likewise for the Heidelberg group, and so on. At the same time, it is impressive that so much qualitative understanding can be obtained through the route espoused herein.

Regarding nomenclature: natural units ($c = \hbar = 1$) are used, and Lorentz–Heaviside units ($\mu_0 = \varepsilon_0 = 1$) enable Maxwell's equations under vacuum to be expressed using $\boldsymbol{B}$ and $\boldsymbol{E}$ without factors of 4π or c. If four-vector notation is unfamiliar, trust that it is done correctly and/or check ref. 59. Briefly, special relativity demands that the value of $t^2 - r^2$ is preserved in going from one inertial reference frame to another *via* Lorentz transformation. It was Poincaré who pointed out that using *it* as a "fourth coordinate" provides the needed minus sign in the inner product of four-vectors that gives $t^2 - r^2$. This approach survived for most of a century, but has been replaced with one in which all vector components are real, whereas the basis vector for the fourth component is imaginary. This is embodied in a metric tensor, η, having only diagonal elements. There are two conventions: $\eta_{00} = -1$, $\eta_{11} = \eta_{22} = \eta_{33} = +1$, and $\eta_{00} = +1$, $\eta_{11} = \eta_{22} = \eta_{33} = -1$. Here, the latter is used, *i.e.*, spacetime (denoted x) has the "signature" $(+\ -\ -\ -)$.

Generic spacetime dummy indices are represented using Greek letters, with implied summation when repeated [up (contravariant), down (covariant)], *e.g.*, $a^\mu b_\mu = a^0 b_0 + a^1 b_1 + a^2 b_2 + a^3 b_3 = a^0 b^0 - a^1 b^1 - a^2 b^2 - a^3 b^3$. For the Euclidean spaces of the molecular case (Latin dummy indices) there is no difference between contravariant and covariant components. At times it is convenient to express a four-vector in terms of its zeroth and three-vector parts, *e.g.*, $x^\mu = (t,\boldsymbol{r})$ and $A^\mu = (\phi,\boldsymbol{A})$. The use of bras and kets is not rigorous. This is unlikely to cause confusion.

Finally, I ask that readers not take offense because their work should be referenced but is not. Unintended omissions are a shortcoming I have borne throughout life. Also, some groups have published so many papers that it is not feasible to reference all that are relevant. In such cases, representative and/or particularly relevant articles are listed.

2. Electrodynamics

According to the rules of standard non-relativistic quantum mechanics the probability of finding a particle whose wave function is ψ in an infinitesimal 3D volume is given by $|\psi|^2 d^3r$. It follows that ψ can be multiplied by a phase factor $e^{i\zeta(\boldsymbol{r},t)}$, where $\zeta(\boldsymbol{r},t)$ is an arbitrary real scalar function of position and time, without affecting this probability. Though not obvious *a priori*, we shall see that $\zeta(\boldsymbol{r},t)$ needs to be differentiable. At the same time, Hermitian operators that act on ψ yield observables that, in general, might be affected profoundly by the phase transformation $\psi \rightarrow \psi e^{i\zeta(\boldsymbol{r},t)}$. Therefore, if acting alone, this transformation is illegal, and egregiously so. Nonetheless, the fact remains that $\psi e^{i\zeta(\boldsymbol{r},t)}$ does just as well at locating the particle as does ψ. This is a redundancy in the description of nature. An infinite number of wave functions each give the same result when it comes to locating the particle.

The fact that multiplication by $e^{i\zeta(\boldsymbol{r},t)}$ affects the mathematical determination of a particle's momentum is a clue to the strategy that will be used to accommodate the phase transformation without compromising the physics. To ensure that nothing goes awry, a field that is already present must change along with the transformation $\psi \rightarrow \psi e^{i\zeta(\boldsymbol{r},t)}$. It turns out that this field is the four-vector potential $A^\mu = (\phi,\boldsymbol{A})$ that yields the magnetic and electric fields according to: $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ and $\boldsymbol{E} = -\nabla\phi - \partial_t \boldsymbol{A}$. It is referred to as the gauge field. Note that electromagnetism is also a redundant theory. Namely, adding the gradient of a scalar

to $\boldsymbol{A}$ (*i.e.*, $\boldsymbol{A} \rightarrow \boldsymbol{A} + \nabla\chi$) leaves $\boldsymbol{B}$ unaltered because the curl of a gradient is identically zero, while at the same time adding $-\partial_t\chi$ to ϕ leaves $\boldsymbol{E}$ unaltered.

The phase transformation of the wave function is referred to as a gauge transformation of the first kind, while the transformation of the gauge field is referred to as a gauge transformation of the second kind. We shall see that the redundancy of the wave function and the redundancy of the gauge field A^μ fit together perfectly. As long as they act in concert, an infinite number of wave functions and an infinite number of gauge fields yield correct answers. The first use of gauge transformations in physical theory was when the Danish physicist Ludwig Lorenz applied them to electromagnetism in the 1860's.

An appropriate tool for examining quantum geometric phase is gauge field theory. It is central to the theory of fundamental particles referred to as the standard model of physics.[60] We shall start with the gauge transformation in which a wave function is multiplied by a phase factor whose argument depends on spacetime coordinates. The basic strategy is then extended to other gauge symmetries. These tools are then applied to intersecting potential surfaces of molecules.

2.1. Consequences of redundancy

Given a particle wave function $\psi(\boldsymbol{r},t)$, it seems intuitively obvious that multiplication by $e^{i\alpha}$, where α is a constant, does not change the calculated outcome of a measurement. Indeed, this is taught as unassailable dogma. Yet, the operation of multiplication by $e^{i\alpha}$ takes effect instantaneously throughout all of space. How can a wave function's phase convention be established at different points in space at the same time? Is this not inconsistent with special relativity? Thus, even something as innocuous as multiplication by $e^{i\alpha}$ is not trivial. Such questions led scientists to think about local phase transformations.

Now consider $|\psi(\boldsymbol{r},t)e^{ia\zeta(\boldsymbol{r},t)}| = |\psi(\boldsymbol{r},t)|$. The real constant, a, is appended to $\zeta(\boldsymbol{r},t)$ for reasons that will soon be clear. Hereafter, the parentheses in $\psi(\boldsymbol{r},t)$ and $\zeta(\boldsymbol{r},t)$ are dropped (understood). Again, to ensure that $\psi \rightarrow \psi e^{ia\zeta}$ leaves the description of the physical system unaltered, simultaneous change must occur elsewhere, as multiplication by $e^{ia\zeta}$ itself incurs dire consequence. Without a partner transformation, ψ cannot be gauge transformed.

To obtain the Schrödinger equation that is compatible with the gauge transformation $\psi \rightarrow \psi e^{ia\zeta}$, we start by considering a free particle:

$$\left(-\frac{1}{2m}\nabla^2 - i\partial_t\right)\psi = 0. \tag{2.1}$$

Though an incorrect result is destined to follow, ψ is now replaced with $\psi e^{ia\zeta}$, yielding

$$e^{ia\zeta}\left(-\frac{1}{2m}(\nabla + ia\nabla\zeta)^2 + a\partial_t\zeta - i\partial_t\right)\psi = 0. \tag{2.2}$$

At this point, the multiplicative factor $e^{ia\zeta}$ can be cancelled if one wishes. However, we shall see in due course that the correct Schrödinger equation transforms covariantly, *i.e.*, it is multiplied by $e^{ia\zeta}$ when the wave function is gauge transformed according to: $\psi \rightarrow \psi e^{ia\zeta}$.

Because of the terms $ia\nabla\zeta$ and $a\partial_t\zeta$, eqn (2.2) describes a different physical situation than does eqn (2.1). Therefore these terms need to be eliminated. This requires a Schrödinger equation that already contains terms that make it possible to bring about the needed cancellation. The phase transformation of the wave function cannot simply be accompanied by additions of $-ia\nabla\zeta$ to ∇, and $-a\partial_t\zeta$ inside the large parentheses, as this eliminates the phase transformation altogether. The only option is that separate entities that are already present in the Schrödinger equation are altered in concert with the phase transformation.

Referring to eqn (2.2), to eliminate $ia\nabla\zeta$ a vector field must be present, whereas to eliminate $a\partial_t\zeta$ a scalar field must be present. Dealing first with $ia\nabla\zeta$, it is necessary that $\psi \rightarrow \psi e^{ia\zeta}$ be accompanied by a change in a vector field whose redundancy is in registry with that of the wave function. Thus, $-ia\boldsymbol{V}$ is added to ∇. The multiplicative factor $-ia$ is included as a matter of notational convenience, as seen below.

With ∇^2 in eqn (2.1) replaced by $(\nabla - ia\boldsymbol{V})^2$, the $ia\nabla\zeta$ in eqn (2.2) is eliminated by adding $\nabla\zeta$ to $\boldsymbol{V}$. In other words, $\psi \rightarrow \psi e^{ia\zeta}$ goes hand-in-hand with $\boldsymbol{V} \rightarrow \boldsymbol{V} + \nabla\zeta$. Applying these transformations to $(\nabla - ia\boldsymbol{V})\psi$ results in the needed cancellation:

$$(\nabla - ia(\boldsymbol{V} + \nabla\zeta))\psi e^{ia\zeta} = e^{ia\zeta}(\nabla - ia\boldsymbol{V})\psi. \tag{2.3}$$

Operating again with $(\nabla - ia(\boldsymbol{V} + \nabla\zeta))$ yields $e^{ia\zeta}(\nabla - ia\boldsymbol{V})^2\psi$.

To eliminate the $a\partial_t\zeta$ term in eqn (2.2) a scalar field must be present. Thus, $a\eta$ is added to the parentheses in eqn (2.1), and η is altered according to: $\eta \rightarrow \eta - \partial_t\zeta$. This results in the needed cancellation of the term $a\partial_t\zeta$ in eqn (2.2). In summary, when $-ia\boldsymbol{V}$ and $a\eta$ are included in eqn (2.1) it becomes

$$\left(\frac{1}{2m}(-i\nabla - a\boldsymbol{V})^2 + a\eta - i\partial_t\right)\psi = 0. \tag{2.4}$$

This equation is said to be gauge covariant (form invariant) because it transforms as the wave function, *i.e.*, when $\psi \rightarrow \psi e^{ia\zeta}$, the entire equation is multiplied by $e^{ia\zeta}$.

2.2. Link with electromagnetism

The assignments $\boldsymbol{V} = \boldsymbol{A}$, $\eta = \phi$, and $a = q$ are now made. The magnitude of the charge q is proportional to the strength of the coupling between the particle and the electromagnetic field, and its sign determines the direction of the force. If the particle does not have electric charge this coupling is zero, so forget about applying the gauge transformation discussed above to the wave function.

As mentioned earlier, the registry between quantum mechanics and electromagnetism is uncanny. The fields $\boldsymbol{B}$ and $\boldsymbol{E}$ are obtained from $\boldsymbol{A}$ and ϕ according to $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ and $\boldsymbol{E} = -\nabla\phi - \partial_t\boldsymbol{A}$. Adding $\nabla\zeta$ to $\boldsymbol{A}$ leaves $\boldsymbol{B}$ unaffected, while at the same time adding $-\partial_t\zeta$ to ϕ leaves $\boldsymbol{E}$ unaffected. Thus, the redundancy of the fields needed to satisfy quantum mechanical gauge covariance matches perfectly the redundancy of classical electromagnetism.

The group under consideration is referred to as U(1), which stands for unitary group of dimension one. It is commutative (Abelian) because $e^{iq\zeta_1}e^{iq\zeta_2} = e^{iq\zeta_2}e^{iq\zeta_1}$. You are familiar with generators of unitary transformations, *e.g.*, H is the generator of time evolution *via* e^{-iHt}. Here we have the unitary

transformation $e^{iq\zeta}$. Thus, it can be said that electric charge is the generator of the U(1) electrodynamics gauge transformation. This conclusion about charge being a generator of gauge transformation is, in fact, more general. It applies to weak and strong charges, though their gauge symmetries are considerably more involved than multiplication by a phase factor.

Because of q's role, it is sometimes referred to as the gauge coupling constant. In the present context this is more likely to confuse than enlighten, so we will stick to charge. Nonetheless, it is interesting that charge arises in the context of a gauge transformation. It follows from the symmetry principle that links particle dynamics to fields through their complementary redundancies. The concept of gauge coupling constants will become more intuitive in subsequent sections where gauge connections are encountered in the context of parallel transport, and in models of the weak and strong forces.

It follows that the gauge covariant Schrödinger equation is

$$\left(\frac{1}{2m}(-i\nabla - q\boldsymbol{A})^2 + q\phi - i\partial_t\right)\psi = 0. \qquad (2.5)$$

The overall gauge transformation affects both the wave function and the gauge field $A^\mu = (\phi,\boldsymbol{A})$. Not only have we obtained A^μ, it has emerged as more fundamental than $\boldsymbol{E}$ and $\boldsymbol{B}$. For a long time the potentials $\boldsymbol{A}$ and ϕ were believed to be mere conveniences for obtaining $\boldsymbol{E}$ and $\boldsymbol{B}$. This is what I was taught in school, and I fell for it.

The above result is not surprising. Electrically charged particles emanate fields, so it is not possible to have such particles in the absence of electromagnetism. Yet, the facile way in which the gauge field enters the Schrödinger equation and the synchrony of the gauge transformations of the first and second kind are impressive. The requirement of gauge covariance has eliminated the possibility that the particle can be free. Said differently, in order that a particle's wave function can undergo a U(1) gauge transformation the particle must carry electric charge. No fundamental particle can be free. It must couple to at least one field if we are to know of its existence. All known fundamental particles carry charge of one or more kinds (electric, weak, strong) that enable them to couple to their respective fields.

2.3. Aharonov–Bohm

It is straightforward to verify (*i.e.*, by direct substitution and using the Leibniz rule for differentiation of an integral) that the solution to eqn (2.5) has the general form:

$$\psi = \psi_{\boldsymbol{A}=0}\exp\left(iq\int_{r_0}^{r} \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right), \qquad (2.6)$$

where $\psi_{\boldsymbol{A}=0}$ is the solution to eqn (2.5) with $\boldsymbol{A} = 0$. The analogous exponential factor that contains the integral of $\mathrm{d}t\phi$ has been subsumed into $\psi_{\boldsymbol{A}=0}$. It will be dealt with shortly. Right now we shall deal exclusively with the magnetic part. It is assumed that $\boldsymbol{A}$ is static relative to a laboratory reference system, and that the particle travels slowly enough that it does not perceive significant time variation of $\boldsymbol{A}$.

In order to eliminate $\boldsymbol{A}$ through a gauge transformation (*i.e.*, $\nabla\zeta = -\boldsymbol{A}$) it cannot be associated with a local magnetic field. In other words, if $\boldsymbol{A}$ is to be expressed as the gradient of a scalar, it is necessary that $\nabla \times \boldsymbol{A} = 0$. In regions where $\nabla \times \boldsymbol{A} = 0$, the value of the integral of $\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}$ between two points is independent of the path. It depends only on the end points. Namely, $\mathrm{d}\boldsymbol{r}\cdot\nabla\zeta = \mathrm{d}\zeta$ integrates trivially, yielding $\zeta(r) - \zeta(r_0)$.

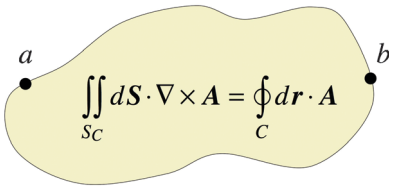

Fig. 2 In regions of zero curl, line integrals have values that depend only on the end points. For example, the value of the line integral from a to b is the opposite of the value of the line integral from b to a, regardless of the paths. When the curl is nonzero, closed circuits enclose flux, so the values of line integrals are path dependent.

Fig. 2 illustrates the fact that in regions where $\nabla \times \boldsymbol{A} = 0$ paths can be distorted without affecting values of line integrals. Alternatively, in regions where $\nabla \times \boldsymbol{A} \neq 0$ (*i.e.*, $\boldsymbol{B}$ is nonzero) the value of a given line integral depends on the path because closed circuits enclose flux. Applying Stokes' theorem to the integral in eqn (2.6) reveals straightaway the enclosed flux as indicated in Fig. 2.

Referring to eqn (2.6) the phase for a closed circuit C is

$$\gamma(\mathrm{C}) = q\oint_{\mathrm{C}} \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}. \qquad (2.7)$$

This is the geometric phase of the magnetic version of the Aharonov–Bohm effect.[45,46] Referring to Fig. 3(a), $\boldsymbol{B}$ is confined to (and distributed uniformly over) a solenoid of circular cross section. Because there is no magnetic flux outside the solenoid, the path can be distorted such that the path indicated in (b) can be chosen. Note that the upper arrow in (a) has reversed direction in going from (a) to (b) because it is the *difference* between the phases acquired on the upper and lower paths that is sought.

You might ask why the paths in Fig. 3 are drawn as classical trajectories. Should we not sum over many paths, as in the path integral formulation of quantum mechanics? The answer lies with the ability to distort paths in regions of zero flux. However many paths are chosen (say, above the solenoid), they can all be distorted to the same single path for the purpose of computing the phase.

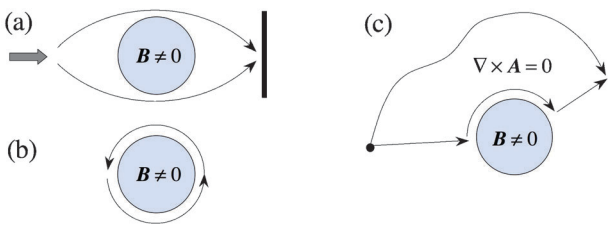

Fig. 3 (a) A particle wave is split, half going above the solenoid, half below. These components are combined at the wall, where they interfere. (b) In calculating phase, paths outside the solenoid can be distorted such that an equivalent path, as far as the phase difference is concerned, is the circular one. (c) The straight-line portions do not contribute because A is perpendicular to these paths. The phase is $q\Phi$ (where Φ is the enclosed flux) times the fraction of 2π in the arc.

Adiabatic separation can be introduced through the placement of the particle in a small box (*e.g.*, assign it to the ground state box eigenfunction) followed by slow transport of the box around the solenoid. Placing the particle in the small box drives home the issue of adiabaticity because a fast degree of freedom is clearly present, *i.e.*, the particle's location within the box.

The above case dealt with the magnetic Aharonov–Bohm effect, which is more popular than the electric version, possibly because it was first to be verified experimentally. The scalar potential ϕ, which was suppressed when dealing with the magnetic version, is now taken into account. Again, the workload is transferred to the wave function, and eqn (2.6) becomes

$$\psi = \psi_{\boldsymbol{A}=0,\phi=0} \exp\left\{ iq \left(\int_{\boldsymbol{r}_0,t_0}^{\boldsymbol{r},t} (\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A} - \mathrm{d}t\phi) \right) \right\}. \qquad (2.8)$$

As before, it is easily verified that this is the general solution by substituting it into eqn (2.5), and using the Leibniz rule for differentiation of an integral. Thus, the overall geometric phase is expressed as

$$\gamma = -q \int \mathrm{d}x^\mu A_\mu. \qquad (2.9)$$

In summary, the potentials ϕ and $\boldsymbol{A}$ have emerged as the central objects of electrodynamics, entering together as the four-vector gauge field $A^\mu = (\phi,\boldsymbol{A})$ [equivalently, $A_\mu = (\phi, -\boldsymbol{A})$]. There is no temptation to use $\boldsymbol{E}$ and $\boldsymbol{B}$, nor would they suffice were this attempted. The gauge field multiplied by the electric charge q is referred to as the gauge connection, though sometimes the field alone is referred to as the gauge connection. We shall see in Section 3 that it is responsible for the parallel transport of the wave function from one point to the next along a path.

2.4. Covariant derivative

The above results can be distilled into a compact algorithm. To see how this works, consider a charged particle that is present in an electromagnetic field. With no particle–field coupling, the Hamiltonian is: $H = p^2/2m + H_{\mathrm{EM}}$, where H_{EM} is the electromagnetic energy. Interaction is now turned on with the transformation: $\partial^\mu \rightarrow \partial^\mu + iqA^\mu$. The quantity to the right of the arrow is referred to as the covariant derivative, D^μ:

$$D^\mu = \partial^\mu + iqA^\mu \qquad (2.10)$$

You might wonder why D^μ is called a covariant derivative, as it is obviously contravariant. The term covariant has two meanings. When an equation or expression does not change its *form* under a transformation it is said to be covariant. For example, 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. In this context, these terms denote the transformation properties of the tensor components. Thus, both $D^\mu = \partial^\mu + iqA^\mu$ and $D_\mu = \partial_\mu + iqA_\mu$ are (Lorentz) covariant derivatives, even though they are expressed in terms of contravariant and covariant components, respectively. It would possibly make more sense to use the term *form invariant* rather than covariant when referring to how an expression or equation transforms. This will be done if it is likely to lessen confusion, but in general I will stick to the common usage described above.

The covariant derivative converts the system from one of global gauge invariance to one of local gauge invariance. 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 (2.10).

Because D_μ is Lorentz covariant, so is the commutator $[D_\mu,D_\nu]$. Expanding this commutator yields an important tensor:

$$[D_\mu,D_\nu] = [\partial_\mu + iqA_\mu, \partial_\nu + iqA_\nu] = iq(\partial_\mu A_\nu - \partial_\nu A_\mu) = iqF_{\mu\nu}. \qquad (2.11)$$

The term $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the (gauge invariant) electromagnetic field strength tensor. Maxwell's equations are expressed in terms of it, and $-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$ is the Lagrangian density for the free electromagnetic field. We shall encounter $F_{\mu\nu}$ later.

3. Geometry

In this section geometrical properties of the covariant derivative are discussed. Specifically, two cases will be examined: (i) how the covariant derivative takes into account curvature of the space itself (as arises in general relativity), and (ii) how an effective (induced) curvature arises through the presence of a gauge field such as the electromagnetic one. The idea is to provide a means of visualizing results that might otherwise seem overly algebraic.

In the last section it was shown that the electrodynamics covariant derivative turns on particle–field interactions through a neat algorithm (*i.e.*, $\partial^\mu \rightarrow D^\mu = \partial^\mu + iqA^\mu$) that follows from an algebraic derivation of its properties. However, everything fell into place so neatly that one cannot help but wonder if something more fundamental does not underlie the algebra. Not surprisingly, we shall see that the covariant derivative is of a geometric nature.

The two examples that serve to illustrate the principles are: (i) parallel transport of a classical vector on a possibly curved manifold; and (ii) parallel transport of the wave function of a particle whose charge is q in the presence of the electrodynamics gauge field A_μ. Much of the discussion presented here follows that given by Aitchison and Hey (volume II: Quantum Chromodynamics and Electroweak Theory, chapter 13).[61] A thorough yet accessible discussion of the classical case that places emphasis on physical (as opposed to mathematical) understanding is given by Schutz, chapter 6.[51] Texts on general relativity invariably cover classical parallel transport. Though our focus is on the quantum mechanical case, we shall first examine the classical case (which has been around since the time of Riemann) as it provides insights that apply to the quantum case. The electrodynamics gauge connection qA_μ has a great deal in common with the molecular counterpart discussed in Sections 5 and 6, in which the gauge connection is $i\langle n|\nabla n\rangle$.

3.1. Parallel transport of a classical vector

A vector $\vec{V}$ is expressed in terms of components and basis vectors: $\vec{V} = V^\alpha \vec{e}_\alpha$ (implied summation). Visualization is easiest using 2D and 3D examples, hence the arrows.

However, it is understood that these lower dimensional spaces are projected out of 4D spacetime, so Greek subscripts and superscripts are retained.

When $\vec{V}$ is differentiated, in addition to the differentiation of its components, the basis vectors $\vec{e}_\alpha$ need to be differentiated because in general they can vary from one point to the next. Thus, $\partial_\mu \vec{V}$ is given by

$$\partial_\mu \vec{V} = (\partial_\mu V^\alpha)\vec{e}_\alpha + V^\alpha(\partial_\mu \vec{e}_\alpha). \quad (3.1)$$

The term $\partial_\mu \vec{e}_\alpha$ is now expanded on the basis:

$$\partial_\mu \vec{e}_\alpha = \Gamma^\gamma{}_{\alpha\mu}\vec{e}_\gamma. \quad (3.2)$$

The symbol $\Gamma^\gamma{}_{\alpha\mu}$ is the expansion coefficient. Thus, eqn (3.1) becomes

$$\partial_\mu \vec{V} = (\partial_\mu V^\alpha)\vec{e}_\alpha + \Gamma^\gamma{}_{\alpha\mu}V^\alpha\vec{e}_\gamma. \quad (3.3)$$

Exchanging the labels of the repeated indices in the last term yields

$$\partial_\mu \vec{V} = (\partial_\mu V^\alpha + \Gamma^\alpha{}_{\gamma\mu}V^\gamma)\vec{e}_\alpha. \quad (3.4)$$

The parenthetic term is referred to as the covariant derivative of V^α:

$$D_\mu V^\alpha = \partial_\mu V^\alpha + \Gamma^\alpha{}_{\gamma\mu}V^\gamma. \quad (3.5)$$

The expansion coefficient $\Gamma^\alpha{}_{\gamma\mu}$ is called the connection (also affine connection and Christoffel symbol of the second kind). It is a field that exists throughout the space in which $\vec{V}$ is defined.§ It accounts for how the basis changes (evolves) from one point to the next. This change is determined by both the nature of the basis and the nature of the space through which (or surface on which) the path is taken, *i.e.*, whether it is inherently flat or curved.¶ The connection, as its name implies, connects (through the basis change) the components of a vector at one point to its components at a nearby point.

Thus, $D_\mu V^\alpha$ gives the rate of change of V^α, with everything referred to a single reference frame. It transforms as a tensor, whereas its ingredients $\partial_\mu V^\alpha$ and $\Gamma^\alpha{}_{\gamma\mu}V^\gamma$ do not individually transform as tensors. To see what is going on, note that $\partial_\mu V^\alpha$ is the rate of change of V^α along x^μ. However, if the basis changes between two points (say from P to P') the change of V^α involves one basis at P and another at P'. Consequently, it is not possible for $\partial_\mu V^\alpha$ to transform as a tensor, as it is not associated with a single reference frame. On the other hand, with $\Gamma^\alpha{}_{\gamma\mu}$ accounting for the basis change, the change of V^α can be evaluated in a single reference frame. This is the job of the covariant derivative $D_\mu V^\alpha$, which transforms as a mixed (one up, one down) tensor.

§ In general relativity space is curved by the presence of mass; without mass the space would be flat. The concept of space itself is intriguing. For example, truly empty space, *i.e.*, devoid of any object or field, cannot exist. A vector can be transported on a surface embedded in a higher dimensional space, and when this surface is inherently curved (*e.g.*, the surface of a sphere) we have a space that is curved. At the same time, the 3D space in which the sphere lies is flat.

¶ The basis change under consideration here is not one in which the coordinate system changes, as would be the case in going from Cartesian to spherical coordinates. Rather, basis vectors such as $\vec{e}_\theta$ and $\vec{e}_\phi$ in spherical coordinates change their directions according to the values of θ and ϕ. For example, $\vec{e}_\phi(\phi = 0)$ and $\vec{e}_\phi(\phi = \pi/2)$ are perpendicular to one another. This change occurs smoothly and continuously on the space.

If the connection is zero everywhere in infinitesimal closed circuits, the space is flat. On the other hand, if the connection is nonzero along a circuit, this does not guarantee that the space is curved, it simply raises the possibility. For example, nonzero contributions along a closed circuit might sum to zero upon completion of the circuit. This happens if the space is inherently flat but the basis is such that the connection does not vanish along the circuit, as discussed below.

3.2. Covariant derivative example

Suppose a vector changes from one point to the next. The covariant derivative enables us to account for the portion of this change that is due to the basis having changed. In general this change can be due to both the nature of the basis and the curved nature of the space. The covariant derivative enables us to isolate these two non-dynamical parts from any dynamical parts that might arise.

The fact that basis vectors can change from one location to the next is well known, even trivial. This is illustrated in Fig. 4 for an inherently flat space that is described using (r, θ) coordinates. The basis vectors $\vec{e}_r$ and $\vec{e}_\theta$ at P clearly differ from the basis vectors $\vec{e}_r{}'$ and $\vec{e}_\theta{}'$ at P'. The differential of $\vec{r} = x\boldsymbol{e}_1 + y\boldsymbol{e}_2 = r\cos\theta\boldsymbol{e}_1 + r\sin\theta\boldsymbol{e}_2$ (where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are x and y unit vectors) is now expressed in the $\vec{e}_r/\vec{e}_\theta$ basis:

$$d\vec{r} = dr(\cos\theta\,\boldsymbol{e}_1 + \sin\theta\,\boldsymbol{e}_2) + d\theta(-r\sin\theta\,\boldsymbol{e}_1 + r\cos\theta\,\boldsymbol{e}_2) \quad (3.6)$$

$$= dr\vec{e}_r + d\theta\vec{e}_\theta$$

Unlike $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, the basis vectors $\vec{e}_r$ and $\vec{e}_\theta$ depend on location, and $\vec{e}_\theta$ is not a unit vector, as it is proportional to r.

Now suppose a vector field is present on the 2D space indicated in Fig. 4. At point P it returns the vector $\vec{V}$, which is expanded on the $\vec{e}_r/\vec{e}_\theta$ basis: $\vec{V} = V^r\vec{e}_r + V^\theta\vec{e}_\theta$. The partial derivatives: $\partial_r\vec{V}$ and $\partial_\theta\vec{V}$, follow straightforwardly. Either serves to illustrate the relationship between this system and eqn (3.1)–(3.4). Choosing ∂_r, we write:

$$\partial_r\vec{V} = (\partial_r V^r)\vec{e}_r + V^r(\partial_r\vec{e}_r) + (\partial_r V^\theta)\vec{e}_\theta + V^\theta(\partial_r\vec{e}_\theta) \quad (3.7)$$

$$= (\partial_r V^\alpha)\vec{e}_\alpha + V^\alpha(\partial_r\vec{e}_\alpha) \quad (3.8)$$

$$= (\partial_r V^\alpha + \Gamma^\alpha{}_{\gamma r}V^\gamma)\vec{e}_\alpha. \quad (3.9)$$

This is eqn (3.4) with $\mu = r$. This example illustrates the fact that even an obviously inherently flat space can have nonzero connections throughout. In the present example we know *a priori* that this is not going to materialize as a curved manifold.

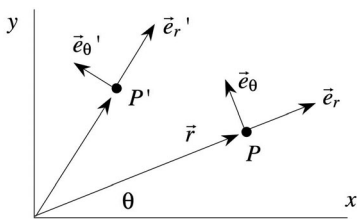

Fig. 4 In going from P to P', the basis rotates and $\vec{e}_\theta$ changes length (adapted from ref. 61).

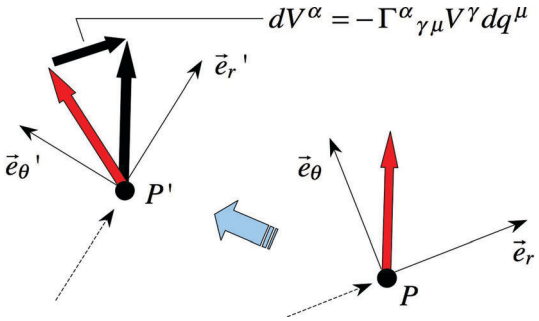

Fig. 5 A red arrow is parallel transported from P to a nearby point P'. Its orientation is unchanged relative to the local basis. The change relative to the original vector (vertical black arrow at P') is due to that of the local basis. The effect is exaggerated by making dV^α large (adapted from ref. 61).

3.3. Parallel transport

If nothing happens to $\vec{V}$ other than its transport from one point to another, the covariant derivative of V^α vanishes. After all, this was the basis for its construction in the first place. In this case eqn (3.4) becomes

$$0 = \partial_\mu V^\alpha + \Gamma^\alpha{}_{\gamma\mu}V^\gamma. \tag{3.10}$$

Multiplication by dq^μ and using $\partial_\mu V^\alpha dq^\mu \equiv dV^\alpha$ yields

$$dV^\alpha = -\Gamma^\alpha{}_{\gamma\mu}V^\gamma dq^\mu. \tag{3.11}$$

In parallel transport the change that a vector component undergoes is due to a change of basis. Fig. 5 illustrates this for the (r,θ) example. The red vector at P is moved to P' while remaining vertical, though turning black in the process. Its components in P' differ from those in P because the basis has changed. For example, the black vector's projection on $\vec{e}_r\,'$ at P' is larger than the red vector's projection on $\vec{e}_r$ at P. Had the vector at P been dragged to P' with its orientation with respect to the basis fixed, it would have arrived at P' as the red vector there. In this case, *i.e.*, of parallel transport, the differential element dV^α is due entirely to the basis change: $dV^\alpha = -\Gamma^\alpha{}_{\gamma\mu}V^\gamma dq^\mu$.

As mentioned earlier, the coefficients $\Gamma^\alpha{}_{\gamma\mu}$ connect the components of the vector at P to those at the nearby point P'. The differential element dV^α is not itself a vector component because the change [*i.e.*, $dV^\alpha = V^\alpha(P') - V^\alpha(P)$] involves different reference frames at P' and P. To have a differential element that transforms as a vector, it is necessary that the change it represents is evaluated with respect to a single basis.

Referring to Fig. 5, an observer moving with the reference frame knows that nothing of importance has happened to the vector, whereas an observer watching the red vector from afar simply sees that it has changed. Of course, it turns out that this change is due entirely to the basis having changed. The covariant derivative $D_\mu V^\alpha$ transforms as a mixed tensor, hence the term *covariant*. Neither $\partial_\mu V^\alpha$ nor $\Gamma^\alpha{}_{\gamma\mu}V^\gamma$ transform as tensors, despite their mathematical appearance and the fact that they are ingredients of a mixed tensor.

3.4. Closed circuit: classical case

An important relationship is revealed when a vector component (say V^α) is parallel transported around an infinitesimal closed circuit, as indicated in Fig. 6. The change experienced by V^α is

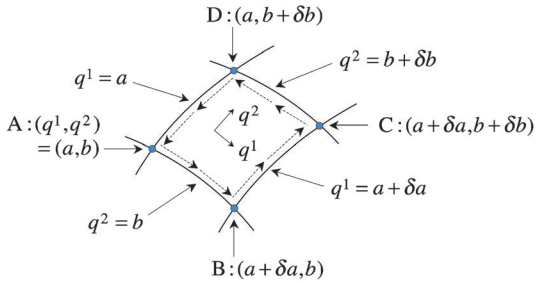

Fig. 6 The α component of $\vec{V}$ is parallel transported around the circuit ABCD. This yields what is called the Riemann curvature tensor.

calculated for each leg of the journey, and these contributions are summed to obtain the net change for the closed circuit. The math (which is an exercise in bookkeeping) is included in the ESI.‡ The resulting infinitesimal change of V^α for an enclosed area $dS^{\mu\sigma}$ is given by

$$dV^\alpha = R^\alpha{}_{\gamma\mu\sigma}V^\gamma dS^{\mu\sigma}. \tag{3.12}$$

The term $R^\alpha{}_{\gamma\mu\sigma}$ is called the Riemann curvature tensor. It is a complicated collection of connections and their derivatives (see the ESI‡ for its derivation). When all of its 256 elements (only 20 of which are independent) vanish the surface is flat at the location where it is evaluated.

This exercise has illustrated how the covariant derivative incorporates curvature *via* the field $\Gamma^\alpha{}_{\gamma\mu}$. If the manifold upon which a vector is parallel transported is curved, the vector points in a different direction at the end of a closed circuit than the direction it had before transport began. The geometric phase is the angular change upon completion of the circuit. In the (r,θ) example, the connection is nonzero *along* a closed circuit, but the Riemann curvature tensor vanishes, as the space is flat. The simplest example of parallel transport on a curved manifold is the case of a sphere. Eqn (3.12) is for the general case.

3.5. Parallel transport of ψ

The parallel transport of a quantum mechanical wave function has much in common with the classical case discussed above, but there are important differences. Here we shall examine the electrodynamics case discussed in Section 2. The wave function ψ varies with its location in spacetime by acquiring phase according to the integral of its gauge connection qA_μ, *e.g.*, eqn (2.9). This alters the apportionment of the wave function into its real and imaginary parts. In other words, parallel transport of the wave function through spacetime is mapped onto the 2D space of the wave function's real and imaginary parts. For a closed circuit and the magnetic version, to have nonzero net geometric phase the path must enclose magnetic flux. Otherwise contributions along the path sum to zero at the end. Here we shall explore the geometrical interpretation.

As in the classical case, multiplication of the covariant derivative of ψ, *i.e.*, $D_\mu\psi = \partial_\mu\psi + iqA_\mu\psi$, by dx^μ yields

$$D_\mu\psi dx^\mu = d\psi + iqA_\mu dx^\mu\psi. \tag{3.13}$$

where $d\psi = \partial_\mu\psi dx^\mu$ has been used.

In classical parallel transport, we saw that the connection $\Gamma^{\alpha}{}_{\gamma\mu}$ accounts for the basis changing smoothly and continuously along a spacetime path. In the present example of quantum parallel transport, the gauge connection is q times the gauge field A_μ. The imaginary unit i is present because it is a complex wave function that is undergoing change along the path. Specifically, the real and imaginary parts of ψ are transformed between themselves, as opposed to the classical case, where the components of a real vector are transformed among themselves. Thus, the change of ψ is represented using a 2D basis that apportions its real and imaginary parts.

One way to visualize the basis evolution that is counterpart to the one in the classical case is through consideration of the system's gauge invariance. To keep matters simple, the magnetic case is considered here. In the internal (complex) space of ψ the basis consists of $\vec{e}_{\mathrm{R}}$ and $\vec{e}_{\mathrm{I}}$, where R and I stand for real and imaginary. As the system is transported through space in the presence of $\boldsymbol{A}$, the phase of ψ changes. In general, it might undergo change brought about by dynamical processes as well, but here we are focusing on parallel transport.

At a given spatial location, $\boldsymbol{r}$, the wave function $\psi(\boldsymbol{r})$ can be expressed in terms of its real and imaginary parts according to: $\psi(\boldsymbol{r}) = \psi(\boldsymbol{r})_{\mathrm{R}} + i\psi(\boldsymbol{r})_{\mathrm{I}}$, as indicated in Fig. 7(a). We shall now consider the change of $\psi(\boldsymbol{r})$ that is brought about through the gauge transformation: $\psi(\boldsymbol{r}) \rightarrow \psi(\boldsymbol{r})\mathrm{e}^{iq\zeta(\boldsymbol{r})}$, specifically, how this enters the calculation of the differential element: $\mathrm{d}\psi = \psi(\boldsymbol{r} + \mathrm{d}\boldsymbol{r}) - \psi(\boldsymbol{r})$, where $\psi(\boldsymbol{r})$ has undergone gauge transformation: $(\psi(\boldsymbol{r})_{\mathrm{R}} + i\psi(\boldsymbol{r})_{\mathrm{I}})\mathrm{e}^{iq\zeta(\boldsymbol{r})}$. The gauge transformation, in effect, rotates the axes. As indicated in Fig. 7(b), $\psi(\boldsymbol{r})$ is now referenced to the new basis $\vec{e}_{\mathrm{R}}{}'$ and $\vec{e}_{\mathrm{I}}{}'$.

As mentioned above, in going from $\boldsymbol{r}$ to $\boldsymbol{r} + \mathrm{d}\boldsymbol{r}$, the wave function, in general, can undergo change due to dynamical processes, and we would like this change to be evaluated relative to the axes $\vec{e}_{\mathrm{R}}{}'$ and $\vec{e}_{\mathrm{I}}{}'$. Importantly, ζ changes according to: $\zeta(\boldsymbol{r}) \rightarrow \zeta(\boldsymbol{r} + \mathrm{d}\boldsymbol{r})$. Because ζ is associated with a rotation of axes, it follows that the change of ζ in going from $\boldsymbol{r}$ to $\boldsymbol{r} + \mathrm{d}\boldsymbol{r}$ causes the axes to undergo yet further rotation. This change of ζ is given by $\nabla\zeta\cdot\mathrm{d}\boldsymbol{r}$ and the further rotation of the axes it incurs is indicated in (c).

Following the gauge transformation, the wave function $\psi(\boldsymbol{r} + \mathrm{d}\boldsymbol{r})$ is referenced to the basis $\vec{e}_{\mathrm{R}}{}''/\vec{e}_{\mathrm{I}}{}''$, whereas $\psi(\boldsymbol{r})$ is referenced to the basis $\vec{e}_{\mathrm{R}}{}'/\vec{e}_{\mathrm{I}}{}'$. In order to compare quantities in a single reference frame, the angular change shown in (c) will be expressed relative to $\vec{e}_{\mathrm{R}}{}'/\vec{e}_{\mathrm{I}}{}'$. This is the work of the gauge field part of the covariant derivative. To see how this works, go back to the general expression for the wave function in terms of $\psi_{\boldsymbol{A}=0}$ and the integral of the gauge connection:

$$\psi = \psi_{\boldsymbol{A}=0}\exp\left(iq\int_{\boldsymbol{r}_0}^{\boldsymbol{r}}\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right). \tag{3.14}$$

We shall now go patiently through the evaluation of $\mathrm{d}\psi = \psi(\boldsymbol{r} + \mathrm{d}\boldsymbol{r}) - \psi(\boldsymbol{r})$ using eqn (3.14). To evaluate $\psi(\boldsymbol{r} + \mathrm{d}\boldsymbol{r})$, write:

$$\psi(\boldsymbol{r} + \mathrm{d}\boldsymbol{r}) = \exp\left(iq\int_{\boldsymbol{r}_0}^{\boldsymbol{r}+\mathrm{d}\boldsymbol{r}}\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right)\psi_{\boldsymbol{A}=0}(\boldsymbol{r} + \mathrm{d}\boldsymbol{r}) \tag{3.15}$$

$$= (1 + iq\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A})\exp\left(iq\int_{\boldsymbol{r}_0}^{\boldsymbol{r}}\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right)(\psi_{\boldsymbol{A}=0}(\boldsymbol{r}) + \nabla\psi_{\boldsymbol{A}=0}(\boldsymbol{r})\cdot\mathrm{d}\boldsymbol{r}). \tag{3.16}$$

Only the lowest order terms have been retained in the expansions of the exponential and $\psi_{\boldsymbol{A}=0}(\boldsymbol{r} + \mathrm{d}\boldsymbol{r})$. Again, retaining only the lowest order terms yields

$$\psi(\boldsymbol{r} + \mathrm{d}\boldsymbol{r}) = (\psi_{\boldsymbol{A}=0} + iq\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\psi_{\boldsymbol{A}=0} + \nabla\psi_{\boldsymbol{A}=0}\cdot\mathrm{d}\boldsymbol{r})\exp\left(iq\int_{\boldsymbol{r}_0}^{\boldsymbol{r}}\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right). \tag{3.17}$$

Note that all quantities are now evaluated at $\boldsymbol{r}$, on the $\vec{e}_{\mathrm{R}}{}'/\vec{e}_{\mathrm{I}}{}'$ basis. Subtracting $\psi(\boldsymbol{r})$ from this yields the differential element $\mathrm{d}\psi$ referenced to the basis $\vec{e}_{\mathrm{R}}{}'/\vec{e}_{\mathrm{I}}{}'$.

$$\mathrm{d}\psi = iq\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\psi + (\nabla\psi_{\boldsymbol{A}=0}\cdot\mathrm{d}\boldsymbol{r})\exp\left(iq\int_{\boldsymbol{r}_0}^{\boldsymbol{r}}\mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right). \tag{3.18}$$

In the case of parallel transport, the rightmost term in eqn (3.18) is zero because dynamical processes are excluded, by definition. In other words, $\nabla\psi_{\boldsymbol{A}=0} = 0$, and the entire change is due to the basis. Eqn (3.13) indicates that for the magnetic case the expression for the covariant derivative of

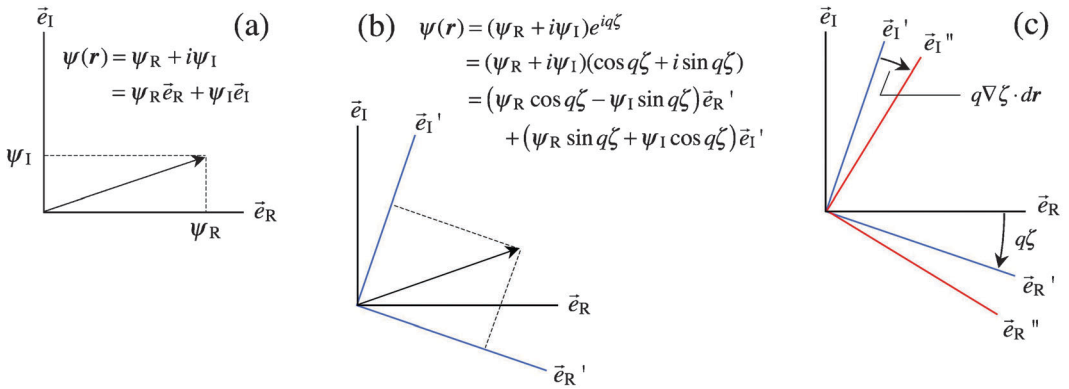

Fig. 7 The phase transformation $\psi \rightarrow \psi\mathrm{e}^{iq\zeta}$ is used to illustrate the different bases that arise in the calculation of the differential $\mathrm{d}\psi$. In (a) and (b) it is understood that all quantities are at location $\boldsymbol{r}$. (c) In going from $\boldsymbol{r}$ to $\boldsymbol{r} + \mathrm{d}\boldsymbol{r}$, ζ changes by $\nabla\zeta\cdot\mathrm{d}\boldsymbol{r}$, resulting in further rotation.

ψ is: $D\psi = \mathrm{d}\psi - iq\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{r}\psi$. For the case of parallel transport (*i.e.*, $D\psi = 0$), this expression is the same as eqn (3.18) with the rightmost term set equal to zero.

3.6. Closed circuit: quantum case

To end, ψ is parallel transported around a closed circuit. The procedure is similar to that used with $\vec{V}$. Because the covariant derivative vanishes, the infinitesimal change along the circuit is $\mathrm{d}\psi = -iqA_\mu\psi\mathrm{d}x^\mu$. The change for an infinitesimal closed circuit (see ESI‡) is

$$\mathrm{d}\psi = -iq(\partial_\mu A_\nu - \partial_\nu A_\mu)\psi\mathrm{d}S, \tag{3.19}$$

where $\mathrm{d}S$ is the infinitesimal surface area. The parenthetic term is recognized as a component of the electromagnetic field, *i.e.*, $F_{\mu\nu}$ in eqn (2.11). Eqn (3.19) can be integrated right away. For $\boldsymbol{A} \neq 0$ and $\phi = 0$, it yields

$$\begin{aligned}\psi &= \psi_0 \exp\left(iq \iint_{S_C} \mathrm{d}\boldsymbol{S}\cdot\nabla\times\boldsymbol{A}\right)\\ &= \psi_0 \exp\left(iq \oint_C \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}\right).\end{aligned} \tag{3.20}$$

The exponentials contain the magnetic version of the Aharonov–Bohm phase derived earlier by a less geometric route. The gauge connection induces curvature in which ψ does not recover its initial phase upon completion of the circuit.

To summarize this section, parallel transport of a classical vector yields a geometric phase due to curvature experienced along the path. The quantum version involves apportionment of the wave function's real and imaginary parts as the system progresses along the path in the presence of the gauge connection.

4. SU(2) and SU(3)

In Section 2 we saw how gauging U(1) leads to electrodynamics. In 1954, C. N. Yang and R. Mills suggested that the same strategy could be applied to other physical systems, raising the possibility of a field-theoretical model for fundamental particles.[62] In an attempt to place the neutron and proton on equal footing insofar as the strong force is concerned, they gauged SU(2), obtaining three independent gauge fields that are counterpart to the A^μ gauge field of electrodynamics. At the time the idea seemed reasonable. Quarks had not yet been discovered, so the fact that the neutron and proton are composites, each consisting of three quarks bound by gluons, was not known. Though the physics was wrong, the math was correct and even compelling.

Later their theory was applied at the quark level. It currently stands as the cornerstone of the standard model of physics. Here, we shall work through the gauging of SU(2). The SU(3) case—mathematically, a logical extension of SU(2)—is given in the Appendix. The former is important in the theory of the weak force,‖ whereas the latter is central to the theory of the strong force, quantum chromodynamics (QCD). The SU(2) and SU(3) gauge fields are germane to the geometric phases that arise when potential surfaces intersect or come into close proximity, namely, SU(2) and SU(3) for two and three potential surfaces, respectively.

The basic strategy is analogous to that used with U(1). Recall that gauging U(1) starts with the global gauge symmetry associated with multiplication of a wave function by $e^{i\alpha}$, where α is a real constant. When this symmetry is made local the gauge field A^μ emerges. Likewise, gauging SU(2) starts with a global SU(2) gauge symmetry, in which objects (states) can be rotated freely among themselves without changing the physical situation, *i.e.*, there is a degeneracy. Gauging SU(2) implies the presence of gauge fields that respond to the phase transformations accessible through the SU(2) generators. It is implicit that the physics accommodates the gauge principle, *i.e.*, the SU(2) field quanta (gauge bosons) must be "observable" experimentally. Indeed, the gauge bosons of both SU(2) (in combination with U(1) according to electroweak theory)‖ and SU(3) have been verified experimentally.

The fundamental SU(2) representation uses a 2D complex space. It is customary to organize states using the usual isospinor basis:

$$\psi_1 = \begin{pmatrix}1\\0\end{pmatrix}\quad \psi_2 = \begin{pmatrix}0\\1\end{pmatrix} \tag{4.1}$$

The prefix "iso" in isospinor is a reminder of the fact that the 3D space associated with the three generators of SU(2) is not related to physical space. This differs from fermion spin $\frac{1}{2}$, whose nonrelativistic Pauli representation enlists a mapping from 3D physical space to a complex plane, yielding the spinor.[63]

As mentioned above, in the seminal Yang and Mills paper, local gauge invariance was imposed in an attempt to put the neutron and proton on equal footing, and later the theory was used with up and down quarks. These quarks are not degenerate, as their masses differ by a few MeV (*i.e.*, 2.01 ± 0.14 *versus* 4.79 ± 0.16 MeV, respectively).[64] Despite this large fractional difference, the degeneracy assumption has been deemed acceptable because of the high energies characteristic of quark physics. For example, these "bare" quark masses are small compared to the nearly 1 GeV nucleon mass, nearly all of which arises from gluon exchange. The issue of exact *versus* approximate gauge symmetry is more important in particle physics than in studies of polyatomic molecules.

The mathematical statement of global SU(2) invariance applied to a doublet of fundamental particles is that the transformation:[61]

$$\psi^{\mathrm{iso}}(x)' = U\psi^{\mathrm{iso}}(x) = \exp(ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha})\psi^{\mathrm{iso}}(x), \tag{4.2}$$

alters the original state $\psi^{\mathrm{iso}}(x)$ while maintaining the degeneracy. The symbol $\boldsymbol{\sigma}$ denotes Pauli matrices, and the factor of $\frac{1}{2}$ is present because the generators of SU(2) are $\frac{1}{2}\sigma_i$. The unitary transformation U can be expressed using a 2×2 matrix or the exponential form shown. The vector $\boldsymbol{\alpha}$ consists of three phase parameters (real constants), one for each generator. The space on which the dot product in $\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}$ is taken is the 3D Euclidean parameter space of $\boldsymbol{\alpha}$, whose components are α_1, α_2, and α_3, whereas spacetime is denoted x. The parameter g is called the

‖ SU(2) does not stand alone in a theory of the weak force. It combines with U(1) in electroweak theory, whose gauge group is $\mathrm{U}(1)_y \times \mathrm{SU}(2)_\mathrm{L}$.

gauge coupling constant. It is present in theories of fundamental particles, but does not arise in the case of polyatomic molecules and the adiabatic (Born–Oppenheimer) approximation.

The phase transformation given by eqn (4.2) differs markedly from the global U(1) phase transformation. For example, instead of just one phase parameter, there are three. Consequently, there will be three independent gauge fields. The number of independent gauge fields is always equal to the number of generators of the group transformations. After all, when α_i varies locally, this must be countered through alteration of a field that is present. Despite the large differences between SU(2) and U(1), the easiest way to follow the development below is to refer to the analogous steps in the U(1) case.

The next step is the promotion of $\boldsymbol{\alpha}$ to its local version, $\boldsymbol{\alpha}(x)$. Recall the covariant derivative of electrodynamics: $D_\mu = \partial_\mu + iqA(x)_\mu$. The SU(2) counterpart is now obtained. Operating on eqn (4.2), where $\boldsymbol{\alpha} = \boldsymbol{\alpha}(x)$, with ∂_μ yields

$$\partial_\mu\psi^{\mathrm{iso}\prime} = \{ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot(\partial_\mu\boldsymbol{\alpha}(x))\}\exp(ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}(x))\psi^{\mathrm{iso}} + \exp(ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}(x))\partial_\mu\psi^{\mathrm{iso}}. \quad (4.3)$$

The partial derivative $\partial_\mu\psi^{\mathrm{iso}\prime}$ does not transform in the same manner as the isospinor ψ^{iso}. Namely, the second term on the right hand side transforms as the isospinor, whereas the first term on the right does not. To obtain a covariant derivative (*i.e.*, one that transforms as the isospinor) it is necessary that a gauge field is present. It is used to eliminate the first term on the right. The form of this covariant derivative is:[61]

$$D_\mu = \partial_\mu + ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}(x)_\mu. \quad (4.4)$$

Multiplication of D_μ and ∂_μ by 2×2 unit matrices is understood. The field $\boldsymbol{W}(x)_\mu$ is comprised of three independent gauge fields, $W_1(x)_\mu$, $W_2(x)_\mu$, and $W_3(x)_\mu$. In other words, $W_i(x)_\mu$ denotes the ith field in the 3D space of phase parameters (α_1, α_2, and α_3) with μ being its spacetime component. The matrix $\boldsymbol{\sigma}\cdot\boldsymbol{W}(x)_\mu$ is

$$\begin{pmatrix} W_3(x)_\mu & W_1(x)_\mu - iW_2(x)_\mu \\ W_1(x)_\mu + iW_2(x)_\mu & -W_3(x)_\mu \end{pmatrix}. \quad (4.5)$$

Thus, the three independent SU(2) gauge fields are revealed. It remains to demonstrate that the system is gauge covariant. Namely, when the covariant derivative acts on $\psi^{\mathrm{iso}\prime}$, it must yield a result in which the first term on the right hand side of eqn (4.3) is no longer present. This term needs to be eliminated through simultaneous addition to the gauge field. The fastest way to do this uses the U in eqn (4.2) and the requirement:

$$D'_\mu\psi^{\mathrm{iso}\prime} = (\partial_\mu + ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}'_\mu)(U\psi^{\mathrm{iso}}) = UD_\mu\psi^{\mathrm{iso}} = U(\partial_\mu + ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)\psi^{\mathrm{iso}} \quad (4.6)$$

Using the second and fourth terms: $(\partial_\mu + ig\frac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}'_\mu)(U\psi^{\mathrm{iso}}) = U(\partial_\mu + ig\frac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)\psi^{\mathrm{iso}}$, yields

$$(\partial_\mu U + ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}'_\mu U - Uig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)\psi^{\mathrm{iso}} = 0. \quad (4.7)$$

Now insert $U^{-1}U$ to the left of ψ^{iso} and operate to the left with U^{-1} to obtain

$$\boldsymbol{\sigma}\cdot\boldsymbol{W}'_\mu = (2i/g)(\partial_\mu U)U^{-1} + U(\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)U^{-1} = 0. \quad (4.8)$$

Barring a mathematical disaster, this equation is solvable for $\boldsymbol{W}'_\mu$.[61] Fortunately, it is not necessary to solve the equation because gauge covariance is assured as long as a solution *exists*. The most important result of this section is the emergence of the three independent SU(2) gauge fields indicated in eqn (4.4) and (4.5). We shall see that these equations play an important role in intersecting potential surfaces.

4.1. SU(2) tensor

Recall that the electromagnetic field strength tensor was obtained by taking the commutator of U(1) covariant derivatives: $[D_\mu,D_\nu] = iqF_{\mu\nu}$. The SU(2) analog is now obtained using the matrix-valued covariant derivative given by eqn (4.4):

$$[D_\mu,D_\nu] = [\partial_\mu + ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu, \partial_\nu + ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{W}_\nu]. \quad (4.9)$$

Minor algebra yields**

$$[D_\mu,D_\nu] = ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot(\partial_\mu\boldsymbol{W}_\nu - \partial_\nu\boldsymbol{W}_\mu - g\boldsymbol{W}_\mu\times\boldsymbol{W}_\nu) = ig\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{F}_{\mu\nu}. \quad (4.10)$$

In electrodynamics we saw that the tensor $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is useful. For example, Maxwell's equations and the Lagrangian density for free fields are expressed in terms of it. The tensor $\boldsymbol{F}_{\mu\nu}$ indicated in eqn (4.10) contains a similar term: $\partial_\mu\boldsymbol{W}_\nu - \partial_\nu\boldsymbol{W}_\mu$. However, the additional term: $-g\boldsymbol{W}_\mu\times\boldsymbol{W}_\nu$, does not have a counterpart in electrodynamics. As distant from molecular science as this might at first appear, it arises in the quest for potentials that are as diabatic as possible. For example, adapting an equation from Pacher *et al.*[37] to match the present nomenclature yields

$$\boldsymbol{F}_{\mu\nu} = \partial_\mu\boldsymbol{W}_\nu - \partial_\nu\boldsymbol{W}_\mu + [\boldsymbol{W}_\mu,\boldsymbol{W}_\nu]. \quad (4.11)$$

The commutator differs from $-g\boldsymbol{W}_\mu\times\boldsymbol{W}_\nu$ by a sign (Itzykson and Zuber also have a minus sign),[65] which may be due to spacetime *versus* Euclidean metric signatures, and a factor of $g/2$ that can be absorbed in the fields. This illustrates the close correspondence between the two cases. Though the physics is different, they share a common gauge theoretical structure.

For all derivative couplings to vanish, each element of $\boldsymbol{F}_{\mu\nu}$ must vanish. This can be understood intuitively. In electrodynamics, when the electromagnetic field strength tensor $\boldsymbol{F}_{\mu\nu}$ vanishes there can be no enclosed flux, so the gauge field can be eliminated through a gauge transformation. Likewise, in the BO case when $\boldsymbol{F}_{\mu\nu}$ vanishes the gauge field can be eliminated through a gauge transformation. In other words, nonadiabatic couplings can be eliminated, by definition yielding diabats. Unfortunately, this cannot be achieved in general because coupling to more distant states cannot be ignored. Great effort has been expended toward finding nearly diabatic representations.

4.2. Extension to SU(3)

Gauging SU(3) follows along similar lines. This group is central to the theory of the strong force, quantum chromodynamics (QCD). In the context of the BOA, SU(3) applies to

** Expanding the commutator yields:
$\partial_\mu\partial_\nu - \partial_\nu\partial_\mu + ig\frac{1}{2}(\boldsymbol{\sigma}\cdot\partial_\mu\boldsymbol{W}_\nu + (\boldsymbol{\sigma}\cdot\boldsymbol{W}_\nu)\partial_\mu - (\boldsymbol{\sigma}\cdot\boldsymbol{W}_\nu)\partial_\mu + (\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)\partial_\nu - \boldsymbol{\sigma}\cdot\partial_\nu\boldsymbol{W}_\mu - (\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)\partial_\nu) - g^2\frac{1}{4}((\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu)\cdot(\boldsymbol{\sigma}\cdot\boldsymbol{W}_\nu) - (\boldsymbol{\sigma}\cdot\boldsymbol{W}_\nu)\cdot(\boldsymbol{\sigma}\cdot\boldsymbol{W}_\mu))$.
The first two terms cancel, as do terms four and five, and six and eight. Now use the relation for Pauli matrices: $(\boldsymbol{\sigma}\cdot\boldsymbol{a})\cdot(\boldsymbol{\sigma}\cdot\boldsymbol{b}) = \boldsymbol{a}\cdot\boldsymbol{b} + i\boldsymbol{\sigma}\cdot\boldsymbol{a}\times\boldsymbol{b}$. Collecting nonzero terms gives eqn (4.10).

the intersection of three adiabats. Mathematically, it is a logical extension of SU(2), with interesting features such as eight independent gauge fields that evolve to two independent connections upon *complete* loss of SU(3) symmetry, *i.e.* no resulting SU(2) symmetry arising from SU(3) → SU(2). An overview of the QCD and BOA cases is given in the Appendix.

5. Adiabatic approximation

In the last section we saw how gauge fields enter mathematically when global SU(2) gauge symmetry is made local. This group (as well as its SU(3) extension) is richer than the U(1) of electrodynamics. For example, the fact that the SU(2) generators do not commute leads to phenomena that have no counterparts in electrodynamics, *e.g.*, non-commuting charges.

In the present section the geometric phase that follows from the adiabatic approximation is discussed, and aspects of both U(1) and SU(2) gauge symmetries are revealed. Following this, in Section 6 non-Abelian gauge groups that are germane to the BO case are discussed. An important issue is that of domains in $\boldsymbol{R}$ over which different gauge symmetries are applicable, including how these symmetries evolve, *e.g.*, the deterioration of a higher symmetry such as SU(2) as the system passes out of the immediate vicinity of a degeneracy. In particle physics no such issue arises because its gauge symmetries, whether exact or approximate, apply to all of causally related spacetime because they refer to fundamental particles.

In studies of polyatomic molecules, the identification of what is nowadays referred to as geometric phase took place half a century ago.[3–5] As mentioned earlier, Mead and Truhlar inspired renewed enthusiasm through the deep, fundamental understanding of this phenomenon that was manifest in their 1979 and 1981 articles.[1,2] It was Berry's 1984 paper,[16] however, that provided the most compelling theoretical model of its time of the relationship between adiabaticity and parallel transport in a space of slowly varying parameters that alter a Hamiltonian without inducing state changes. Given Berry's considerable expertise in general relativity,[66] the physical understanding and mathematical intricacies of parallel transport, curved spaces, connections, holonomy, *etc.* were no doubt second nature to him.

In the picture presented by Berry,[16] the parameters responsible for parallel transport are under complete external control. In marked contrast, a molecule's nuclear degrees of freedom, though lethargic on electron motion time scales, are fully quantum mechanical. This results in interplay between phases of electronic and nuclear wave functions. Mead and Truhlar were the first to point out that this interplay is germane to the issue of molecular geometric phase.[1] It is shown here that this seemingly simple matter goes to the very heart of molecular geometric phase.

5.1. Born–Oppenheimer

In electronic structure theory, the molecular Schrödinger equation is solved using the BOA, *i.e.*, electronic states (energies, wave functions) are obtained at specified values of nuclear coordinates, collectively denoted $\boldsymbol{R}$. In other words, $\boldsymbol{R}$ is a multidimensional vector in the space of the nuclear degrees of freedom (excluding translation and usually rotation). Thus, $\boldsymbol{R}$ serves as a parameter in the theory. The assumption that $\boldsymbol{R}$ varies slowly enough to accommodate adiabatic transport of electronic eigenfunctions is valid in the U(1) regime of Berry's connection, but it breaks down in regions where adiabats intersect, *i.e.*, where energy level spacing for electron *versus* nuclear degrees of freedom becomes comparable. Indeed, quite different gauge symmetry applies in the immediate vicinity of the intersection.

The phase progression of an electronic eigenfunction from one nuclear configuration to the next is not forthcoming from electronic structure calculations. These calculations are carried out at one $\boldsymbol{R}$, then another, and so on, and therefore relative phase is unknown. It must be determined separately. Thus, following Berry,[16] a reasonable ansatz is that this yet-to-be-determined phase, $\gamma_n(\boldsymbol{R},t)$, is included in the wave function Ψ_n for the nth electronic state (in our case the nth adiabat):

$$\Psi_n = |n\rangle e^{i\gamma_n} e^{-i\int dt\omega_n}. \tag{5.1}$$

Here and hereafter it is understood that $\gamma_n = \gamma_n(\boldsymbol{R},t)$, $|n\rangle = |n(\boldsymbol{R}(t))\rangle$, and ω_n is a slowly varying function of time, in keeping with the system's adiabaticity. An expression for γ_n is obtained by putting the Ψ_n in eqn (5.1) into the Schrödinger equation. Again, Berry considered externally controlled parameters, whereas we are using nuclear coordinates, denoted $\boldsymbol{R}$.

For an isolated nonlinear molecule, the number of degrees of freedom of the $\boldsymbol{R}$ space can be taken to be $3N - 6$, where N is the number of nuclei. The neglect of overall translation is rigorous, whereas the neglect of overall rotation is usually reasonable. In most cases, only a few nuclear degrees of freedom are needed for the determination of the geometric phase, *e.g.*, those that are used to "tune" Hamiltonian matrix elements until degeneracy is achieved. For example, if H_{12} is real (conical intersection), two nuclear degrees of freedom are needed: one to tune H_{12} until it vanishes and another to tune $H_{22} - H_{11}$ until it vanishes.

We shall take the $\boldsymbol{R}$ space to be 3D, with the understanding that it is, in general, a subspace of the large dimensional $(3N - 6)$ parameter space. This 3D space subsumes the 2D case of conical intersection. Namely, when $\mathrm{Im}H_{12}$ is everywhere zero on physical grounds, no tuning coordinate is needed to make it vanish, so in this case a plane in the 3D space is selected. Even polyatomic molecules having many nuclear degrees of freedom often submit to a 3D $\boldsymbol{R}$ space when nonadiabatic dynamics and geometric phases are sought. Coordinates that can be varied while maintaining degeneracy comprise what is called the intersection coordinate subspace (ICS).[27] Its dimension takes into account nuclear degrees of freedom that are held fixed to preserve a needed symmetry.

The state described by Ψ_n can change its energy and phase throughout its adiabatic transport, but it cannot undergo a transition "from one continuous (through geometry space) eigenvalue of the Hamiltonian to another", as noted by a reviewer. This is the essence of adiabatic change in quantum mechanics. Consequently, using eqn (5.1) with the Schrödinger equation yields

$$\begin{aligned}\omega_n\Psi_n &= i\,\dot{\Psi}_n \\ &= i\left(i\dot{\gamma}_n\Psi_n - i\omega_n\Psi_n + e^{i\gamma_n}e^{-i\int dt\omega_n}\frac{d|n\rangle}{dt}\right).\end{aligned} \tag{5.2}$$

Canceling the $\omega_n\Psi_n$ terms, using $\mathrm{d}|n\rangle/\mathrm{d}t = \dot{\boldsymbol{R}}\cdot|\nabla n\rangle$, and multiplying from the left by $\langle\Psi_n|$ and integrating over electron coordinates yields

$$i\mathrm{d}\gamma_n + \langle n|\nabla n\rangle\cdot\mathrm{d}\boldsymbol{R} = 0. \tag{5.3}$$

The term $\langle n|\nabla n\rangle$ is a diagonal vector matrix element that is denoted $\boldsymbol{F}_{nn}$. It is a vector in nuclear space and a matrix element in the space of electron functions. The fact that time is not present in eqn (5.3) underscores the geometric nature of γ_n. Integration over a closed circuit C yields the geometric phase, $\gamma_n(\mathrm{C})$, of the nth adiabat:

$$\gamma_n(\mathrm{C}) = \oint_{\mathrm{C}} \mathrm{d}\boldsymbol{R}\cdot i\boldsymbol{F}_{nn}. \tag{5.4}$$

Because of the fact that $\gamma_n(\mathrm{C})$ must be real, $\boldsymbol{F}_{nn}$ must be imaginary. Hereafter ψ_n is used in place of $|n\rangle e^{i\gamma_n}$ except when integration is implied, such as in $\langle n|\nabla n\rangle$. The only difference between ψ_n and Ψ_n is that the former lacks the evolution factor, $e^{-i\int \mathrm{d}t\omega_n}$.

The fact that $\boldsymbol{F}_{nn}$ must be imaginary also follows from the fact that $\langle m|i\nabla n\rangle$ is the matrix element of a Hermitian operator, which indicates that $\boldsymbol{F}_{mn}$ is anti-Hermitian, *i.e.*, $\boldsymbol{F}_{mn} = -F_{nm}{}^*$. Such terms ($m \neq n$) are nonadiabatic couplings that induce transitions between adiabats. The term $\boldsymbol{F}_{nn}$ is subtler, as $i\nabla$ acting on ψ_n returns phase through the parametric dependence of ψ_n on $\boldsymbol{R}$. This phase change is the only change that ψ_n can experience in $\boldsymbol{R}$ while remaining in the state ψ_n. Another way of saying this is that $\nabla\langle n|n\rangle = 2\mathrm{Re}\langle n|\nabla n\rangle = 0$, and therefore $\langle n|\nabla n\rangle$ is imaginary.

It can be said that dynamical phase (*i.e.*, the integral of $\mathrm{d}t\omega_n$) depends on how much time the trip requires, whereas geometric phase depends on the chosen route—a point made by Berry in his Overview article in ref. 17. From the discussion of parallel transport in Section 3, it follows that $i\boldsymbol{F}_{nn}$ is the connection on $\boldsymbol{R}$, *e.g.*, note that $\boldsymbol{F}_{nn}\cdot\mathrm{d}\boldsymbol{R} = \langle n|\mathrm{d}n\rangle$. The connection $i\boldsymbol{F}_{nn}$ on $\boldsymbol{R}$ is analogous to the connection qA_μ on x.

As mentioned earlier, the geometric phase associated with conical intersection accrues on a 2D $\boldsymbol{R}$ space. Integration over the phase discontinuity at the end of the 2π circuit of eqn (1.1) yields $\gamma_n(\mathrm{C}) = \pm\pi$. Alternatively, each adiabat can be gauge transformed, yielding the same result. When intersection requires a 3D $\boldsymbol{R}$ space, the geometric phase is $\pm\frac{1}{2}\Omega$, where Ω is the solid angle, subtended from the $\boldsymbol{R}$ space origin, of the closed circuit.[16]

Applying Stokes' theorem to eqn (5.4) yields

$$\gamma_n(\mathrm{C}) = \iint_{S_{\mathrm{C}}} \mathrm{d}\boldsymbol{S}\cdot\nabla\times i\boldsymbol{F}_{nn}. \tag{5.5}$$

This indicates that $\gamma_n(\mathrm{C})$ is impervious to the phase transformation: $\psi_n \rightarrow \psi_n e^{i\zeta}$, where ζ is an arbitrary scalar function of $\boldsymbol{R}$. Namely, $\nabla\times i\boldsymbol{F}_{nn}$ is unaffected by $i\boldsymbol{F}_{nn} \rightarrow i\boldsymbol{F}_{nn} - \nabla\zeta$ because $\nabla\times\nabla\zeta$ is identically zero. Eqn (5.5) shows that the geometric phase $\gamma_n(\mathrm{C})$ can be interpreted as the normal component of the flux: $\nabla\times i\boldsymbol{F}_{nn}$, integrated over the surface enclosed by C. This flux emanates from the degeneracy point, eliciting images of magnetic monopoles. The above results were derived in Berry's seminal 1984 paper.[16]

5.2. Phase discontinuity

Let us now discuss $i\boldsymbol{F}_{nn}$ in the context of U(1) gauge transformation. In Section 3 it was seen that curvature, either of the space itself, or induced through the presence of a gauge field, leads to geometric phase. For example, the change experienced by a local basis in parallel transporting $\vec{V}$ around a closed circuit is the angle between where $\vec{V}$ points at the start *versus* where it points at the end. This is a geometric phase. Likewise, the changing apportionment of a wave function into its real and imaginary components as the wave function is parallel transported over a closed circuit can yield a geometric phase, as discussed in the electrodynamics example of Section 3.

As noted above, $\langle n|\nabla n\rangle$ is purely imaginary. This can be satisfied if ψ_n is a single-valued function of $\boldsymbol{R}$ that is composed of a real function times $e^{i\alpha(\boldsymbol{R})}$, where $\alpha(\boldsymbol{R})$ is a scalar function of $\boldsymbol{R}$. In the conical intersection example of eqn (1.1), $\langle n|\nabla n\rangle$ vanishes everywhere on the circuit except where it closes. The functions used in eqn (1.1) are real, or so they appear. However, a subtle aspect arises when the circuit closes.

Fig. 8 illustrates parallel transport over a 2π circuit in the θ parameter that appears in eqn (1.1). Referring to Fig. 8(a), the phase of ψ_n changes abruptly by π when the circuit closes. Therefore, the adiabat must be complex. Specifically, it must vary as $e^{i\eta(\theta)}$, where $\eta(\theta)$ 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 $e^{i\eta(\theta)}$, with $\eta(\theta)$ going from 0 to π. Integration over this infinitesimal region gives $-\pi$. Referring to Fig. 8(b), we see that this phase discontinuity can also be dealt with through gauge transformation. With ψ_n multiplied by $e^{i\theta/2}$ 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. 8, 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 practical 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.

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. In the case of an adiabat, the electron and nuclear microcosms are not isolated from one another. They communicate through the connection field, as discussed in Section 5.3. 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 $qdr^\mu\boldsymbol{A}_\mu$ and $\mathrm{d}\boldsymbol{R}\cdot i\boldsymbol{F}_{nn}$ yield phases, as these are actions. However,

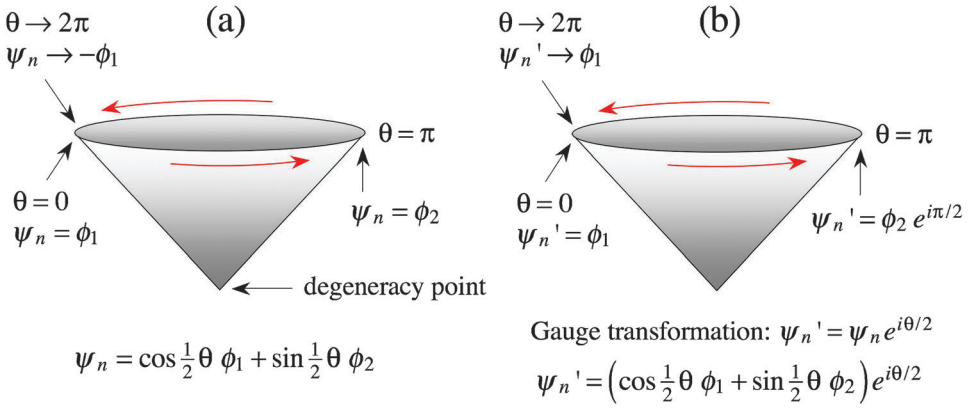

Fig. 8 The upper adiabat in the region of a conical intersection is indicated for the example given in eqn (1.1). The electronic wave function is parallel transported as θ goes from 0 to 2π, encircling the degeneracy point in the process. (a) The wave function is real everywhere except where the circuit 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, $\psi_n' = \psi_n e^{i\theta/2}$, 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.

the question remains: how can $i\boldsymbol{F}_{nn}$ be best visualized? The math leading to eqn (5.4) is beyond reproach, but how can the gauge connection field that follows from the adiabatic approximation be understood qualitatively?

5.3. Interpretation: gauge field theory

As a U(1) gauge field theory, the quantum mechanical redundancy of ψ_n with respect to local phase transformation on $\boldsymbol{R}$ implies the presence of the gauge field $i\langle n|\nabla n\rangle$. In the overall gauge transformation, $\psi_n \rightarrow \psi_n e^{i\zeta}$ is accompanied by the addition of $\nabla\zeta$ to $i\langle n|\nabla n\rangle$. No physical effect is incurred, because positive and negative $\nabla\zeta$ terms cancel. The addition of $\nabla\zeta$ to the gauge field is, in fact, achieved through a complementary redundancy in the partner system, in this case χ_n.

The invariances and covariances that are achieved through the gauge transformation in which ψ_n acts in concert with the gauge field are a manifestation of the isolated molecule assumption. By definition, an isolated molecule cannot couple to a gauge field. Otherwise it would not be isolated. Thus, by introducing the isolated molecule assumption, coupling to external fields is eliminated. With the molecule taken as isolated, its total wave function, ψ_{total}, must be single-valued. Thus, multiplication of ψ_{total} by $e^{i\zeta}$, where ζ varies locally, is forbidden. This simple rule ensures registry between the electronic structure and field theory pictures, as discussed below.

Until now emphasis has been on ψ_n. After all, it is assumed that ψ_n is not coupled to other adiabats, and the separation of the nuclear and electron degrees of freedom has rendered it 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 $\psi_{\text{total}} = \chi_n\psi_n$. It is significant that ψ_n and χ_n can undergo synchronous phase transformations on $\boldsymbol{R}$: $\psi_n \rightarrow \psi_n e^{i\zeta}$ and $\chi_n \rightarrow \chi_n e^{-i\zeta}$.[1] On the one hand, this seems like an interesting way to say nothing happened, as the phase of ψ_{total} is unaffected. On the other hand, because of this synchrony, ψ_n and χ_n can each obey the gauge principle. Let's now see how the gauge principle and the $e^{i\zeta}/e^{-i\zeta}$ synchrony are related.

The fact that $\psi_n \rightarrow \psi_n e^{i\zeta}$ is accompanied by the addition of $\nabla\zeta$ to $i\langle n|\nabla n\rangle$ means that χ_n is transformed such that the ∇ in $i\langle n|\nabla n\rangle$ becomes $\nabla - i\nabla\zeta$. This is due to the phase transformation: $\chi_n \rightarrow \chi_n e^{-i\zeta}$. To see how this works, recall that $\nabla(\chi_n\psi_n) = \psi_n\nabla\chi_n + \chi_n\nabla\psi_n$, and use left multiplication by ψ_n^* and integration over electron coordinates to write

$$(\nabla + \langle n|\nabla n\rangle)\chi_n. \tag{5.6}$$

The covariant derivative is $\nabla + \langle n|\nabla n\rangle$. Notice how closely this expression resembles the space part of its electrodynamics counterpart (*i.e.*, $\nabla - iq\boldsymbol{A}$) when it is written: $\nabla - i(i\langle n|\nabla n\rangle)$. Though the slow system is oblivious to the relatively rapid electron dynamics, it is understood that parallel transport and nuclear space dynamical processes are on equal footing, which makes sense in terms of the single-valuedness of ψ_{total} imposing $e^{i\zeta}/e^{-i\zeta}$ synchrony. For example, χ_n's phase transformation ($\chi_n \rightarrow \chi_n e^{-i\zeta}$) causes $\nabla\chi_n$ to become $e^{-i\zeta}(\nabla - i\nabla\zeta)\chi_n$, and this enters the gauge field, as indicated in Fig. 9(a). Indeed, the gauge field is invariant with respect to the simultaneous and synchronous gauge transformations of ψ_n and χ_n.

The bottom line is that U(1) gauge field theories work for χ_n and ψ_n because the gauge field links these systems, which have been created through the adiabatic separation, ensuring that overall gauge symmetry is preserved.

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. 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 the redundancy 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.

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.

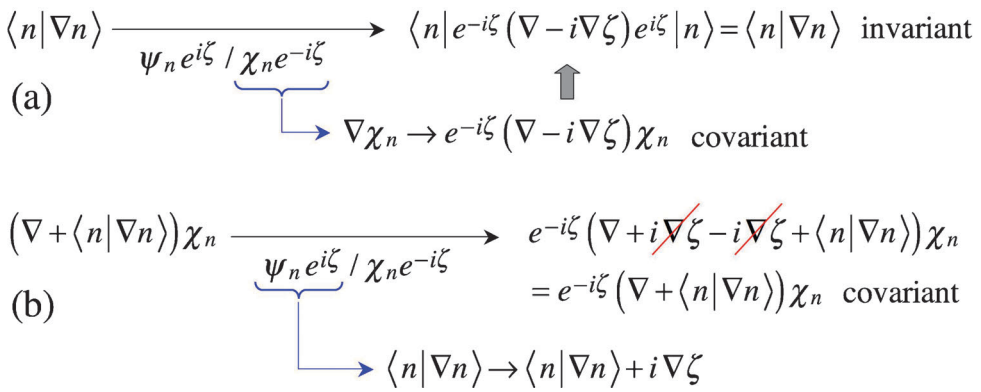

Fig. 9 (a) The gauge transformation $\psi_n \rightarrow \psi_n e^{i\zeta}$ requires the accompanying addition of $\nabla\zeta$ to the gauge field $i\langle n|\nabla n\rangle$. In fact, this is nothing more than $\chi_n \rightarrow \chi_n e^{-i\zeta}$ being manifest in the gauge field through the gradient. (b) Likewise, the gauge principle applied to $\chi_n \rightarrow \chi_n e^{-i\zeta}$ works because $\psi_n \rightarrow \psi_n e^{i\zeta}$ is manifest in the gauge field.

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. 9(b). The gauge field serves as the communication link between the fast and slow microcosms.

5.4. U(1) summary

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.

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. 10.

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 is no less a gauge field theory because of the $e^{i\zeta}/e^{-i\zeta}$ requirement. It is simply one that can be understood in detail.

6. Born–Oppenheimer non-Abelian gauge field theory

Non-Abelian gauge groups that are germane to the BOA, including its breakdown, are now examined. In Section 5, a simple expression for the closed-circuit geometric phase, $\gamma_n(C)$, was obtained, *i.e.*, eqn (5.4). By itself, eqn (5.4) leaves open the issue of what accrues upon parallel transport around a closed circuit. Namely, $\gamma_n(C)$ depends on the relation of the circuit to the topology: whether the nth adiabat intersects others, and whether transport encircles one or more intersection points.

Given the similarities between the nth adiabat and electrodynamics cases, it is not surprising that the gauge field $i\boldsymbol{F}_{nn}$ enters in more-or-less the same way as does qA^{μ}. For example, the redundancy of $\psi_n \rightarrow \psi_n e^{i\zeta}$ is matched to that of $i\boldsymbol{F}_{nn} + \nabla\zeta$. Nor is it surprising that their closed circuit integrations share a common relationship insofar as enclosed flux is concerned. In the case of electrodynamics, the covariant derivative is $\partial_{\mu} + iq\boldsymbol{A}_{\mu}$, whereas the molecular counterpart is $\partial_{\mu} + (\boldsymbol{F}_{nn})_{\mu}$. In electrodynamics, μ denotes spacetime, whereas in the molecular case it denotes 3D Euclidean $\boldsymbol{R}$ space. The fact that the parallel transport of an isolated adiabat obeys a U(1) gauge field theory is confirmed by the fact that $i\boldsymbol{F}_{nn}$ transforms as a U(1) gauge field.

The isolated molecule assumption and the consequent $e^{i\zeta}/e^{-i\zeta}$ synchrony requirement eliminates the possibility of local phase transformation of the $\chi_n\psi_n$ pair as a whole, *i.e.*, $\chi_n\psi_n \rightarrow \chi_n\psi_n e^{i\zeta}$ is forbidden. U(1) gauge theories nonetheless work for χ_n and ψ_n. It was shown that the simultaneous addition to the gauge field dictated by the gauge principle arises from the $e^{i\zeta}/e^{-i\zeta}$ synchrony. This illustrates how a gauge field theory works in a familiar, transparent system.

In this section, the gauge principle is applied to non-Abelian gauge groups in order to obtain a general expression for the gauge fields. As seen in Section 5, the gauge (connection) fields can be obtained through application of the gauge principle. These fields are then compared to those obtained directly from the Schrödinger equation.

6.1. Intersection of two adiabats

The case of two adiabats yields a well-known result. When off-diagonal Hamiltonian matrix elements are real, encircling a degeneracy point yields $\pm\pi$, and when these matrix elements are complex the phases are $\pm\frac{1}{2}\Omega$, where Ω is the solid angle subtended from the degeneracy point. Thus, on the one hand, the system behaves as a U(1) gauge field theory. On the other hand, phase of $\pm\pi$ is a signature of a spinor, which argues for SU(2) playing a role of some kind. To see how these facts come together, the degeneracy region is examined.

At an intersection of two adiabats, any normalized linear combination of their wave functions is a solution of the Schrödinger equation. In the strictest sense, the fact that the

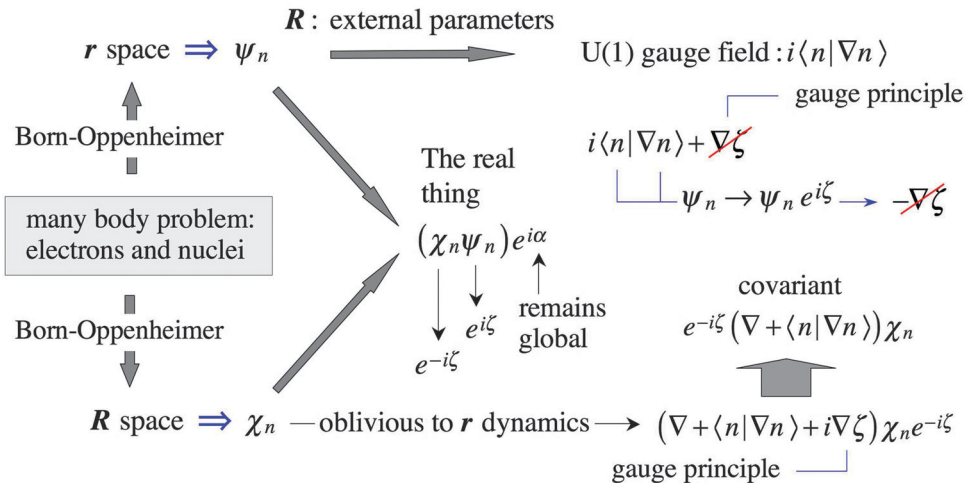

Fig. 10 When ψ_n perceives its adiabatic transport as effected by slowly varying external parameters, it is blind to χ_n. It obeys a U(1) theory. Addition of $\nabla\zeta$ to the gauge field is, in fact, a manifestation of ψ_{total}'s single-valuedness ($e^{i\zeta}/e^{-i\zeta}$ synchrony). Likewise, 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.

potentials intersect does not guarantee that the system can access this degeneracy. Specifically, treating the nuclear degrees of freedom quantum mechanically yields vibronic states whose energies need not lie at (or very close to) the degeneracy. However, this is usually not a large enough effect to lessen the utility of the qualitative picture. Thus, it is assumed that the degeneracy is accessible. Also, despite the fact that the BOA breaks down at the intersection, the adiabats serve as a convenient basis.

The largest global gauge symmetry group that can be applied to two degenerate adiabats is U(2), which factors to U(1) × SU(2). Recall that U(1) is multiplied by $e^{i\zeta}$, while SU(2) transformations enable 2D complex vectors to be continuously rotated in isospin space. Whereas U(1) appends a common phase to each member of the isospinor doublet, SU(2) appends phases of equal magnitude and opposite sign to the doublet members.

For U(2) to be applicable, the domain in ***R*** must be restricted to the immediate vicinity of the degeneracy. Except for the topological imprint it bestows, exact degeneracy is unimportant because it only exists as a subspace. For example, the ***g***–***h*** plane is 2D, whereas its origin is a point and therefore of dimension zero. Thus, in the ***g***–***h*** plane it is the 2D vicinity of the origin that is important. As pointed out earlier, degeneracy applies throughout the intersection coordinate subspace (ICS) but, by definition, not in the tuning coordinate subspace.

This use of a restricted domain in ***R*** over which a given gauge symmetry can be applied contrasts with the standard model, whose gauge symmetries, whether approximate or exact, apply throughout all of causally related spacetime. The standard model deals with fundamental particles that have no internal structure. The gauge symmetry of two particles that can be rotated one into the other has nothing to do with location in spacetime, as long as they remain in the time-like region. Approximate gauge symmetry arises when the particles of concern have different mass, as in the quark example given earlier (*i.e.*, up and down quarks have different masses: 2.01 ± 0.14 *versus* 4.79 ± 0.16 MeV, respectively,[64] and therefore they cannot be degenerate). On the other hand, with the BOA, it is not particle state vectors, but many-body wave functions that can be rotated one into the other, and this is valid as long as the states are sufficiently close to exact degeneracy to justify the latter's enlistment as a good approximation.

The issues addressed in this section are ones of lowered symmetry. In the electroweak theory of particle physics,[67] and in the Jahn–Teller effect,[68] symmetries are said to break spontaneously. When something in science is said to happen spontaneously, the cause is hidden from view and might be discovered later. For example, spontaneous emission, a term introduced by Einstein in 1916, follows straightaway from quantization of the electromagnetic field, which was introduced a decade later. In the Jahn–Teller effect the symmetry is that of the full Hamiltonian and it is the adiabatic minima that lose symmetry, while in electroweak theory, spontaneously broken symmetry is not yet fully understood.[69] In each case, "spontaneous" reflects understanding, not something fundamental.

To begin, we shall consider U(2) in the near-degeneracy region. Intuitively, slight deviations from exact degeneracy are not expected to be problematic. But how is: "slight deviations from exact degeneracy" to be construed in the present context? Fortunately, there is precedent. As mentioned above, in particle physics SU(2) has been applied to up and down quarks despite a significant mass difference (2.01 ± 0.14 *versus* 4.79 ± 0.16 MeV, respectively).[64] However, quark physics takes place at energies that exceed greatly the bare quark masses. Thus, the near-degeneracy assumption has been enlisted despite the fact that the theory is designed for exact symmetry.[70]

6.2. Gauging the SU(2) part

On the basis of the above arguments, it is assumed that global U(2) gauge symmetry is applicable in the immediate vicinity of the degeneracy. Its transformations can be represented as multiplication by $e^{i\zeta}$, where ζ is a real constant, of a 2 × 2 matrix that carries out global SU(2) transformations.

The mathematical difference between U(2) and SU(2) is that U(2) also has the unit matrix as a generator. It uses a complete basis of 2×2 Hermitian matrices, *e.g.*, the unit matrix plus the Pauli matrices.

Now consider gauging U(2) = U(1) × SU(2). When the system ventures from degeneracy enough that physical effects are incurred, U(1) and SU(2) must be looked at carefully. For example, whereas SU(2) symmetry goes away, U(1) is relatively robust. It is multiplication of the doublet by $e^{i\zeta}$, which has nothing to do *per se* with degeneracy.

Gauging SU(2) yields three gauge fields. This result is imported except for constants from Section 4. In contrast, U(1) is not gauged, as this would lead to egregious problems. When $e^{i\zeta}$ multiplies a doublet of adiabats it appends to each the same phase, and when it is permitted to vary locally this implies the presence of a gauge field. Were this gauge field present, it would be added to each diagonal of a 2×2 matrix like the one in eqn (4.5), with the same value at each of the two diagonal positions. It would then be impossible to satisfy known BO symmetries such as $\boldsymbol{F}_{mm} = -\boldsymbol{F}_{nn}$. Even worse is the specter of a gauge field that acts outside the realm of the $\boldsymbol{F}_{mn}$, and would require a new kind of charge. The bottom line is that the BO gauge fields belong to SU(2).

With U(1) out of the way, in the immediate vicinity of the degeneracy the BO covariant derivative D_μ is expressed in terms of the gradient and the SU(2) generators and fields. Adapting eqn (4.4) and (4.5) to the BO case, we write:

$$D_\mu = \partial_\mu \mathbf{1} - i \begin{pmatrix} W_3 & W_1 - iW_2 \\ W_1 + iW_2 & -W_3 \end{pmatrix}_\mu , \qquad (6.1)$$

where **1** denotes a 2×2 unit matrix. In going from eqn (4.4) and (4.5) to eqn (6.1) a factor of $\frac{1}{2}g$ has been subsumed into the fields. The $\frac{1}{2}$ is not important here, and the SU(2) gauge coupling constant g is ignored because in the BO case there is none. The use of the letter W in eqn (6.1) is to underscore the correspondence between the BO gauge fields and those in eqn (4.5).

Eqn (6.1) contains the BO gauge fields that arise through quantum mechanical redundancy. They are present in the immediate vicinity of the degeneracy and evolve to nonadiabatic interaction away from degeneracy. With the structure of the BO gauge fields in hand, we now turn to the progressive deterioration of SU(2) gauge symmetry as the system ventures from the degeneracy. First, however, two items need attention: (1) a few comments about electroweak theory are in order, as it parallels the BO case to an uncanny extent, at least mathematically; and (2) the BO gauge fields and covariant derivatives that have been alluded to in the above discussion are given.

6.3. Comments on electroweak theory

In electroweak theory, the electrodynamics and weak forces merge into a single force at extremely high energy: ~100 GeV (and corresponding short range of $\sim 10^{-8}$ Å), where the electrodynamics and weak forces would separately have comparable strength. The electroweak gauge group is $U(1)_y \times SU(2)_L$. The subscript L denotes the fact that transformations act only on left-handed fermions, *i.e.*, the particle's spin projection is antiparallel to its momentum. This experimental discovery prompted Pauli's famous comment: "I cannot believe that god is a weak left-hander." The subscript y denotes what is called weak hypercharge. Indeed, y is the generator of $U(1)_y$. In a similar vein, $SU(2)_L$ is referred to as weak isospin.

It is necessary to assign separate charges to the $U(1)_y$ and $SU(2)_L$ parts, and it turns out that electric charge q can be expressed in terms of y and weak isospin: $q = t_3 + y/2$ (in units of e), where t_3 is the third component of weak isospin. Without symmetry breaking, there are four massless gauge bosons. However, when symmetry is broken (*via* the Higgs mechanism), the surviving symmetry is $U(1)_{EM}$, where EM stands for electromagnetism. Three gauge bosons now have mass, whereas the fourth is massless (*i.e.*, the photon). Notice that the gauge group is not U(2), *i.e.*, weak hypercharge and weak isospin charge are kept separate. To the best of my knowledge, attempts to use U(2) have never caught on.

Going into this would stray and distract us from the main goal, to say nothing of the fact that the author has at best a cursory understanding of electroweak theory. This is an area of ongoing research. For example, in the unification regime why are there two gauge fields and not just one? Also, spontaneous symmetry breaking is another way of saying that there exists a symmetry that is hidden from view, so what is this symmetry.

Electroweak theory and the problem of two intersecting adiabats have in common the U(1) × SU(2) gauge group, including its breakdown, spontaneous or otherwise. This provides clues regarding the BO gauge fields. For example, the electroweak gauge group retains separate identities: $U(1)_y$ and $SU(2)_L$. Likewise, we considered U(1) and SU(2) separately, rather than simply gauging U(2), which would have proven untenable later on. This led to U(1)'s dismissal.

6.4. Covariant derivative

Consider two adiabats that intersect. To simplify matters, nonadiabatic couplings to all other adiabats are ignored. The general case, in which nonadiabatic couplings to all other adiabats are taken into account, has been discussed thoroughly by Pacher *et al.*[37] Now operate with ∇^2 on $\chi_n\psi_n$, multiply from the left by ψ_m^*, and integrate over electron coordinates to obtain

$$\langle \psi_m | \nabla^2 | \chi_n \psi_n \rangle = \left(\delta_{mn} \nabla^2 + 2\langle m | \nabla n \rangle \cdot \nabla + \langle m | \nabla^2 n \rangle \right) \chi_n . \qquad (6.2)$$

Straightforward manipulation of the parenthetic term yields:[37]

$$(\nabla + \boldsymbol{F}) \cdot (\nabla + \boldsymbol{F}) = \begin{pmatrix} \partial + F_{mm} & F_{mn} \\ F_{nm} & \partial + F_{nn} \end{pmatrix}_\mu \begin{pmatrix} \partial + F_{mm} & F_{mn} \\ F_{nm} & \partial + F_{nn} \end{pmatrix}_\mu , \qquad (6.3)$$

with implied summation over the nuclear degrees of freedom, denoted μ. Notice that the single operator $\boldsymbol{F}$ suffices to account for all nonadiabatic coupling. The covariant derivative is $\nabla + \boldsymbol{F}$. The operators ∇ and $\boldsymbol{F}$ are vectors in $\boldsymbol{R}$ and 2×2 matrices in the electron space, with $\boldsymbol{F}$ having matrix elements: $\boldsymbol{F}_{mn} = \langle m | \nabla n \rangle$. For three adiabats, $\boldsymbol{F}$ is 3×3, and so on for more adiabats.

6.5. Broken symmetry: SU(2) to U(1)

Here, the deterioration of SU(2) gauge symmetry as the energy gap between adiabats, $|E_m - E_n|$, increases is discussed. For small gaps, SU(2) is a good approximation. However, as the gap increases the system becomes non-degenerate, and eventually the adiabats lose their ability to communicate with one another through nonadiabatic transitions. Yet, the system retains memory of the intersection. For example, take the case of conical intersection. The topology of a cone is such that all of its curvature is at its apex. Thus, the curvature experienced in a closed circuit that encloses the origin does not depend on the distance from the origin. Spinor character is established near the degeneracy and it is not compromised by breakdown of SU(2) gauge symmetry.

Referring to eqn (6.1), when the system first departs from degeneracy, the gauge fields in the off-diagonal locations become the nonadiabatic couplings indicated in eqn (6.3). As the system departs yet further from degeneracy, the off-diagonal terms approach zero. In other words, SU(2) gauge symmetry goes away: the further the departure from degeneracy, the less valid is SU(2). To facilitate comparison, eqn (6.1) and the covariant derivative using a matrix from eqn (6.3) are shown below:

$$\partial_\mu - i\begin{pmatrix} W_3 & W_1 - iW_2 \\ W_1 + iW_2 & -W_3 \end{pmatrix}_\mu \quad \partial_\mu - i\begin{pmatrix} iF_{mm} & iF_{mn} \\ iF_{nm} & iF_{nn} \end{pmatrix}_\mu \tag{6.4}$$

Again, note that $\boldsymbol{F}_{mm} = -\boldsymbol{F}_{nn}$, and $\boldsymbol{F}_{mn} = -\boldsymbol{F}_{nm}$*.

In the electroweak case, the fields $W_1 \pm iW_2$ play an analogous role. Field quantization and symmetry breaking yield three massive gauge bosons. Two of these ($W^{\pm}$, arising from $W_1 \pm iW_2$) are electrically charged, which enables them to interconvert up and down quarks. For example, when a down quark emits W^- (which decays almost immediately to an electron and its antineutrino) it is converted to an up quark. This is β decay. Henri Becquerel discovered it in 1896, though the explanation came only much later. The fact that up and down quarks are electrically charged, differing by one unit of charge ($+2/3$ and $-1/3$, respectively), means that W^+ and W^- carry electric charge.

Analogy between the electroweak and BO cases is impressive. In the former, electrically charged gauge bosons $W^{\pm}$ arise through symmetry breaking of the electroweak gauge group, $U(1)_y \times SU(2)_L$, resulting in a surviving $U(1)_{EM}$ symmetry. These gauge bosons induce transitions between quarks, with their electrically charged nature playing a central role. In the BO case, coupling fields arise through BOA breakdown. When the $U(2) = U(1) \times SU(2)$ global gauge symmetry that applies in the immediate vicinity of the intersection breaks down, U(1) and SU(2) are no longer on equal footing. SU(2) is gauged, whereas U(1) is not. When SU(2) breaks down, eventually going away completely, it does so in a way that yields a pair of U(1) gauge symmetries, one for each adiabat. They are related: $\boldsymbol{F}_{mm} = -\boldsymbol{F}_{nn}$. These U(1) gauge symmetries are vestiges of SU(2). They are unrelated to the U(1) part of U(2).

6.6. Synopsis for Section 6

To summarize Section 6: in the immediate vicinity of the degeneracy, global U(2) gauge symmetry applies. It factors to $U(1) \times SU(2)$, thereby enabling these subgroups to be judged separately. In anticipation of lifting the degeneracy through the tuning coordinates, SU(2) is gauged, but U(1) is not. As the energy gap between adiabats increases from zero, the system begins to lose SU(2) symmetry and replace it with a pair of related U(1) symmetries, *i.e.*, one each for the upper and lower adiabats.

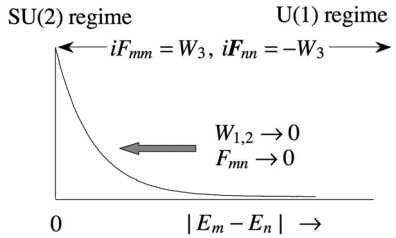

Fig. 11 As $|E_m - E_n|$ increases, the off-diagonals decrease. Diagonals are manifest in the U(1) regime as $i\boldsymbol{F}_{mm}$ and $i\boldsymbol{F}_{nn}$.

Referring to eqn (6.4), the coupling fields (off-diagonals) get smaller and eventually disappear, whereas W_3 and $-W_3$ correlate to $i\boldsymbol{F}_{mm}$ and $i\boldsymbol{F}_{nn}$. Far from the intersection, the system can be on one or the other adiabat without experiencing transitions between them. SU(2) symmetry no longer exists. This is the regime of Berry's connection. The sketch in Fig. 11 illustrates this correlation.

Triple intersections have been examined by a number of groups.[53–55,57,71] Application of gauge field theory to the triple intersection of adiabats follows along more-or-less the same lines as the SU(2) case, albeit with the richer mathematics that accompanies SU(3): eight generators and gauge fields, nine Lie algebras, and three different structure constants. A brief discussion of this is given in the Appendix.

7. Conclusions and summary

Geometric phase is an intellectually stimulating topic that is germane to numerous and varied research areas: molecules, optics, quantum computing, quantum Hall effect, graphene, and so on. It exists only when the system of interest interacts with something it perceives as exterior. Even with the celebrated Aharonov–Bohm effect, were the entire system considered (*e.g.*, including the solenoid in the magnetic version), it would have no geometric phase. However, when we limit our attention to the charged particle the phase appears. The same holds for molecules and the Born–Oppenheimer adiabatic separation. However valid it might be, it is not possible to eliminate the geometric phase that arises from intersecting adiabats, as this phase is a manifestation of the adiabatic separation.

Gauge field theory "brings to the table" a uniquely well-suited perspective as well as an arsenal of mathematical tools with which geometric phase can be examined. From its very beginning it has had profound and lasting impact on a number of fundamental scientific areas. The U(1) theory introduced by Vladimir Fock (and examined thoroughly by Herman Weyl) yields the gauge field A^μ, particle–field couplings, and the Aharonov–Bohm phase, while Yang–Mills theory, the cornerstone of the standard model of physics, serves as a template for non-Abelian gauge field theories. Importantly, these theories are applicable to the geometric

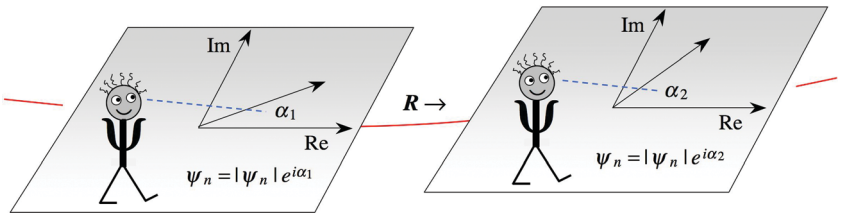

Fig. 12 The adiabat ψ_n is parallel transported on $\boldsymbol{R}$ (red path). It accrues geometric phase according to integration of its gauge connection $i\langle n|\nabla n\rangle$ along the path. The phase of ψ_n is displayed in ψ_n's internal space, which consists of its real and imaginary parts. If, at the end of a closed circuit, the integration yields a nonzero value, there is a net geometric phase.

phases that arise with intersecting potential surfaces. Aspects of gauge field theory that relate to molecular geometric phase were introduced into the chemical physics (physical chemistry) literature starting in the 1980's. Since then this approach has slowly but surely gained a modicum of acceptance.

7.1. Molecular electronic structure theory

Electronic structure theory describes, usually with good (or at least acceptable) accuracy, a broad range of molecular processes. In the vast majority of studies no reference is made to geometric phase, nor is its omission considered a matter for concern. For a single adiabat, integration over electron coordinates of the nuclear gradient operator acting on $\chi_n\psi_n$, *i.e.* $\langle n|\nabla|\chi_n\psi_n\rangle$, yields $\nabla + \boldsymbol{F}_{nn}$, where $\boldsymbol{F}_{nn} = \langle n|\nabla n\rangle$. This generalizes to the matrix-valued expression $\nabla + \boldsymbol{F}$ for electron spaces comprised of any number of adiabats, where the off-diagonal matrix elements $\boldsymbol{F}_{mn} = \langle m|\nabla n\rangle$ are nonadiabatic couplings. Likewise, the dressed Laplacian is $(\nabla+\boldsymbol{F})\cdot(\nabla+\boldsymbol{F})$. This material is often introduced at the beginning of a course in electronic structure theory.

When ψ_n is real, the diagonal term $\langle n|\nabla n\rangle$ vanishes. It is easily proved that the real part of $\langle n|\nabla n\rangle$ always vanishes, but it cannot be proved that $\langle n|\nabla n\rangle$ does likewise. Rather, it is always imaginary. This does not preclude it from being zero, but leaves open the possibility that it is nonzero. The closed circuit integration of $\langle n|\nabla n\rangle$ is, in fact, nonzero when the path subtends a solid angle Ω in a 3D $\boldsymbol{R}$ space whose origin is a degeneracy point for two adiabats. A subtle aspect of non-single-valued wave functions arises with the closing of the circuit. In the conical intersection example of eqn (1.1), the wave function is real throughout the circuit until the moment before the circuit closes. Only then does its complex character emerge, yielding geometric phase of $\pm\pi$. Nonetheless, in the majority of dynamics problems the use of $\langle n|\nabla n\rangle = 0$ is acceptable. When it fails, however, it does so in a big way.

7.2. Gauge connection

A molecular system's closed-circuit geometric phase can be determined through integration of $\mathrm{d}\boldsymbol{R}\cdot i\langle n|\nabla n\rangle = i\langle n|\mathrm{d}n\rangle$ over a closed path in $\boldsymbol{R}$. The term $i\langle n|\nabla n\rangle$ is the gauge connection, *i.e.*, Berry's adiabatic connection. It is somewhat analogous to the connection $\Gamma^{\gamma}{}_{\alpha\mu}$ that arises in the parallel transport of a classical vector over a path in spacetime. This analogy can only be taken so far, however, as ψ_n differs quite a bit from a classical vector.

In the case of a classical vector, the connection $\Gamma^{\gamma}{}_{\alpha\mu}$ transforms vector components among themselves along a path. On the other hand, the only change that ψ_n can undergo (while remaining in the state ψ_n) is a change of its phase: $\psi_n e^{i\alpha_1} \rightarrow \psi_n e^{i\alpha_2}$. Unlike the parallel transport of a classical vector, whose components are defined in the space in which transport is carried out, ψ_n's only "components" are its real and imaginary parts. Thus, as ψ_n is parallel transported along a path in $\boldsymbol{R}$ its phase variation is followed in its internal space (complex plane), as indicated in Fig. 12.

The U(1) molecular case has a great deal in common with the electrodynamics case, whose connection is qA_μ. In each, there is apparent curvature that affects the phase of the wave function. It is no coincidence that the covariant derivatives (respectively, $\nabla - i(i\boldsymbol{F})$ and, for the space part of the electrodynamics covariant derivative, $\nabla - iq\boldsymbol{A}$) each contain the imaginary unit, as this provides the means whereby the real and imaginary parts of the wave function are transformed between themselves during parallel transport. Using the easily visualized case of a closed circuit, we saw that the wave function's phase change is the geometric phase. It is said that the gauge connection "induces" curvature on the space insofar as the wave function is concerned: presumably to note the distinction from the parallel transport of a classical vector.

7.3. Gauge principle

The emergence of a molecular geometric phase and its striking similarity to the electrodynamics (Aharonov–Bohm) geometric phase are reasons enough to anticipate a gauge field theory description of the molecular system. Moreover, upon application of the gauge principle, it is seen that the "dressed" nuclear gradient operator $\nabla + \boldsymbol{F}$, as well as the Schrödinger equation for the nuclear degrees of freedom, transform covariantly. It is therefore appropriate to step back and examine what is going on in the context of gauge field theory.

As stated so aptly by Guidry:[72] "The root of a symmetry principle is the assumption that certain quantities are unobservable; this in turn implies an invariance under a related mathematical transformation, and the invariance under this transformation (if it is unitary, as is usually the case in quantum mechanics) implies a conservation law or selection rule." The gauge principle adheres to this. In electrodynamics the phase transformation $\psi \rightarrow \psi e^{iq\zeta}$ is partnered with $A_\mu \rightarrow A_\mu - \partial_\mu\zeta$. The synchrony with which these transformations act reflects a symmetry. It results in Lagrangians being invariant, and things like equations of motion transforming covariantly. The symmetry that underlies the gauge principle in this case is the redundancy of the wave function (with respect to phase transformation) acting in concert with the redundancy of the gauge field A_μ (with respect to the addition of the corresponding gradient $\nabla\zeta$). The wave function

does not really take on a new phase, nor does the gauge field acquire an extra piece. It is the symmetry—the registry between the two redundancies—that leads to invariance.

In electrodynamics the gauge field A_μ and particle charge q occupy center stage, with qA_μ serving as a communication link between the particle and the gauge field. The connection on spacetime is qA_μ, though A_μ by itself is often referred to as the connection. One sees why, as pointed out earlier, q is referred to in some quarters as a "gauge coupling constant". This usage adds little enlightenment in electrodynamics. It is more to the point with gauge symmetry groups such as SU(2), where the gauge coupling constant is a scalar and the "charges" carry the Lie algebra of the group and therefore do not commute.

There is no gauge coupling constant or charge in the molecular case. Instead, adiabatic separation creates a pair of spaces, each behaving as external to the other. The connection $i\langle n|\nabla n\rangle$ plays an analogous role to qA_μ. The quantum mechanical redundancy of ψ_n on $\boldsymbol{R}$ is partnered with another system that has complementary redundancy, with the gauge connection serving as a communication link between ψ_n and its partner χ_n. This of course implies that the other redundancy exists; otherwise forget about applying the gauge principle. Recall that in electrodynamics the gauge principle applies only to particles that carry electric charge. If the fundamental particle does not carry electric charge, the gauge principle cannot be applied to U(1) phase transformations. Simply stated: phase transformation is illegal.

The requirement that the total wave function must remain single-valued is satisfied with the synchronous transformations $\psi_n \rightarrow \psi_n e^{i\zeta}$ and simultaneously $\chi_n \rightarrow \chi_n e^{-i\zeta}$. With this synchrony taken into account everything falls into place. For example, $\langle n|\nabla n\rangle$ is unaffected (invariant) and $(\nabla + \langle n|\nabla n\rangle)\chi_n$ transforms covariantly. Thus, application of the gauge principle in the U(1) regime is explained. The gauge connection is the communication link in $\boldsymbol{R}$ between the χ_n and ψ_n subsystems. It is a consequence of the adiabatic separation. Though each space behaves externally to the other, it is impossible for them to be entirely divorced. Geometric phase is a manifestation of the adiabatic separation: a survivor, so to speak, of broken SU(2) gauge symmetry, as discussed below.

A related classical case is the one concerning Einstein and general relativity. Referring to Fig. 13, a ball is released in a stationary room that is present in a gravitational field. The effect recorded by an observer is indistinguishable from an effect recorded with the observer plus the room and its contents in free space (no gravity), but accelerated by a force whose strength is equal to that of the gravitational field. Likewise, in the molecular case a phase change of ψ_n is indistinguishable from an action that arises through the gauge field. In each of these examples the invariance is due to the symmetry. As it turns out, the gravity case is, in many ways, harder conceptually than the molecular one. For example, the gravitational field is a rank-2 tensor. Therefore its quantum is a spin-2 boson (graviton), which has not yet been observed experimentally, though there is currently great optimism at CERN. Quantization of molecular vibrations is obviously much easier.

7.4. SU(2) evolves to U(1)

Turning now to the degeneracy region, the largest global gauge symmetry group that is applicable in the immediate vicinity of a two-state intersection is U(2). It conveniently factors to U(1) × SU(2). Gauging SU(2) yields three independent gauge fields, whereas U(1) is not gauged, as doing so cannot be brought into registry with electronic structure theory, and there are other problems as well. All of this is done in the immediate vicinity of the degeneracy. Thus, the $\boldsymbol{R}$ space domain over which SU(2) applies is restricted, to be sure. In particle physics the domain over which its symmetries, whether exact or not, apply is all of causally related spacetime. Interestingly, this domain is also restricted in the sense that the time-like region is retained and the space-like region is rejected. This restriction has such firm physical basis, however, that its enlistment is automatic.

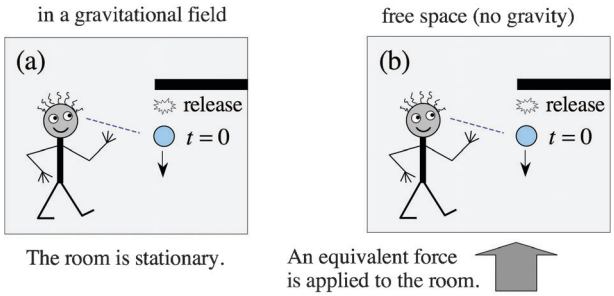

Fig. 13 (a) A person observes a ball drop in a room that is present in a gravitational field. (b) The situation in (a) is indistinguishable from one in which the room is in free space, with a force acting on the room that is equivalent to that of the gravitational field.

As the system ventures from the near-degeneracy region, global SU(2) gauge symmetry goes away, as it is based on the ability to freely transform the adiabats between themselves with no resulting change in the physical situation. A parallel with spontaneous symmetry breaking in electroweak theory was noted and discussed. The progressive loss of SU(2) gauge symmetry as the energy gap between adiabats increases explains the system's evolution from SU(2) symmetry to the pair of inter-related U(1) symmetries (*i.e.*, $\boldsymbol{F}_{nn} = -\boldsymbol{F}_{mm}$) that exist in the regime of Berry's connection. These apparent U(1) gauge symmetries are vestiges of SU(2). They are not related to the U(1) symmetry that entered with global U(2) [*i.e.*, U(1) × SU(2)] but was not gauged. Geometric phase exists throughout the entire regime of the transition from SU(2) to a pair of inter-related U(1) symmetries. It is manifest in both connections and couplings. Spinor character imprinted in the vicinity of the degeneracy is robust, from near-degeneracy to the U(1) regimes.

7.5. Finalé

Electronic structure theory, including nonadiabaticity, is a non-Abelian gauge field theory with matrix-valued covariant derivative. It is not merely analogous to a gauge field theory, it *is* a gauge field theory. When a molecule is treated as an isolated entity, its total wave function is single-valued. Consequently, its global U(1) symmetry cannot be gauged, and products of electron and nuclear functions are forbidden from undergoing local phase transformation, *i.e.*, $\chi_n\psi_n \rightarrow (\chi_n\psi_n)e^{i\zeta}$ is forbidden. This is consistent with electronic structure theory, including nonadiabaticity. Alternatively, the synchronous phase transformations: $\psi_n \rightarrow \psi_n e^{i\zeta}$ and $\chi_n \rightarrow \chi_n e^{-i\zeta}$, preserve the single-valuedness of the total wave function and

enable each subsystem to undergo phase transformation. Each obeys an "apparent" U(1) theory that, in fact, is a manifestation of the $e^{i\zeta}/e^{-i\zeta}$ synchrony.

In the two subsystems created by adiabatic separation, ψ_n sees the nuclear coordinates as a quiescent platform upon which it can undergo parallel transport, and χ_n is oblivious to the relatively rapid electron dynamics in $\boldsymbol{r}$. Each of these limits is described by a U(1) theory, with additions to the gauge connection dictated by the gauge principle. As mentioned above, it was shown that these additions are, in fact, manifestations of the synchronous $e^{i\zeta}/e^{-i\zeta}$ nature of the ψ_n and χ_n phase transformations on $\boldsymbol{R}$. This synchrony is another way of stating the isolated molecule assumption. A wonderful thing about the molecular case is that it illustrates the workings of a gauge field theory in a familiar system that has a transparent basis for the separation into two subsystems, each behaving externally to the other.

The article dealt with conceptual issues, complementing the arsenal of tools available through electronic structure theory. Calculations of geometric phases can be subtle, and are best left to the experts. Parallels were identified between seemingly disparate areas. Who would have thought that electroweak spontaneous symmetry breaking relates to Born–Oppenheimer gauge fields, or that general relativity would in any way prove germane. Most of the material that was presented is the work of others, often carried out in quite different context. A couple (to the best of my knowledge) new ideas were put forth: (i) a restricted $\boldsymbol{R}$ space domain accommodates SU(2) gauge symmetry for the case of two adiabats. This anticipates its deterioration as the energy gap increases, with the ultimate gauge symmetry being a pair of inter-related U(1) symmetries, one for each adiabat (*i.e.*, $\boldsymbol{F}_{nn} = -\boldsymbol{F}_{mm}$); likewise for SU($n$). (ii) The single valuedness of the molecule's total wave function implies $e^{i\zeta}/e^{-i\zeta}$ synchrony, as noted by Mead and Truhlar.[1] It has been shown that this underlies the gauge principle.

It is hoped that the article promotes deep understanding of geometric phase by embedding the molecular case in a broader theoretical framework. Features that might otherwise appear obtuse to experimentalists and theorists in physical chemistry and chemical physics can be appreciated as intuitive, *e.g.*, gauging away the coupling fields in the diabatization problem. Beyond the molecular realm, knowledge gained here may prove transferable, *e.g.*, the fascinating graphene system.[73] In any event, this "Perspective" article presents exactly that: the author's perspective on molecular geometric phase.

Appendix. Gauging SU(3)

The procedure for gauging SU(3) is outlined here; it is similar to the SU(2) case. The particle physics to which SU(3) applies is based on the fact that quarks have three "color" degrees of freedom: red, blue, and green (no relation to regular color). Color symmetry is exact. Each quark has a triplet that can be written as a column vector ψ. As with the 2D isospinors of eqn (4.1), a convenient basis is

$$\begin{pmatrix}1\\0\\0\end{pmatrix}\ \text{red}\quad \begin{pmatrix}0\\1\\0\end{pmatrix}\ \text{blue}\quad \begin{pmatrix}0\\0\\1\end{pmatrix}\ \text{green.} \tag{1}$$

Table 1 SU(3) has eight generators that are represented using the Gell-Mann matrices $\frac{1}{2}\lambda_i$. The Lie algebras are: $[\lambda_i,\lambda_j] = 2if^{ijk}\lambda_k$. The structure constants are: $f^{123} = 1; f^{147} = f^{165} = f^{246} = f^{257} = f^{345} = f^{376} = 1/2; f^{458} = f^{678} = \sqrt{3}/2$

λ_1: $\begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix}$ λ_2: $\begin{pmatrix}0&-i&0\\i&0&0\\0&0&0\end{pmatrix}$ λ_3: $\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix}$

λ_4: $\begin{pmatrix}0&0&1\\0&0&0\\1&0&0\end{pmatrix}$ λ_5: $\begin{pmatrix}0&0&-i\\0&0&0\\i&0&0\end{pmatrix}$ λ_6: $\begin{pmatrix}0&0&0\\0&0&1\\0&1&0\end{pmatrix}$

λ_7: $\begin{pmatrix}0&0&0\\0&0&-i\\0&i&0\end{pmatrix}$ λ_8: $\frac{1}{\sqrt{3}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&-2\end{pmatrix}$

The local phase transformation is given by: $\psi' = \exp(i(g_s/2)\boldsymbol{\lambda}\cdot\boldsymbol{\alpha}(x))\psi$. The eight SU(3) generators, $\frac{1}{2}\lambda_i$ (Table 1 lists the Gell-Mann matrix representation), combine with the eight independent phase parameters, α_i. Drawing on the treatment of SU(2), the covariant derivative is

$$D_\mu = \partial_\mu + ig_s\tfrac{1}{2}\boldsymbol{\lambda}\cdot\boldsymbol{A}(x)_\mu \tag{2}$$

where $\boldsymbol{A}(x)_\mu$ denotes the μ spacetime component of the eight independent fields $A_i(x)_\mu$.

The matrix $\boldsymbol{\lambda}\cdot\boldsymbol{A}(x)_\mu$ is given by:

$$\boldsymbol{\lambda}\cdot\boldsymbol{A}(x)_\mu = \begin{pmatrix} A_3+\frac{1}{\sqrt{3}}A_8 & A_1-iA_2 & A_4-iA_5 \\ A_1+iA_2 & -A_3+\frac{1}{\sqrt{3}}A_8 & A_6-iA_7 \\ A_4+iA_5 & A_6+iA_7 & -\frac{2}{\sqrt{3}}A_8 \end{pmatrix}_\mu \tag{3}$$

Triple intersection of adiabats

Triple intersections have been examined by a number of groups.[53–55,57,71] Comments are given here regarding the relationship between the gauge fields that arise *via* the BOA and those obtained by gauging SU(3). At the intersection, the adiabats can be transformed freely among themselves. As mentioned earlier, exact degeneracy is relevant only because of the topological imprint it bestows, *i.e.*, it is always of lower dimension than the space in which it plays a role. To facilitate comparison, eqn (2) is reproduced here alongside its BO counterpart:

$$\partial_\mu - i\begin{pmatrix} A_3+\frac{1}{\sqrt{3}}A_8 & A_1-iA_2 & A_4-iA_5 \\ A_1+iA_2 & -A_3+\frac{1}{\sqrt{3}}A_8 & A_6-iA_7 \\ A_4+iA_5 & A_6+iA_7 & -\frac{2}{\sqrt{3}}A_8 \end{pmatrix}_\mu \tag{4}$$

$$\partial_\mu - i\begin{pmatrix} iF_{mm} & iF_{mn} & iF_{mp} \\ iF_{nm} & iF_{nn} & iF_{np} \\ iF_{pm} & iF_{pn} & iF_{pp} \end{pmatrix}_\mu$$

The SU(3) case follows along the same lines as SU(2). There are two unique diagonal elements. With SU(3) symmetry compromised, off-diagonals are nonadiabatic couplings and diagonals are connections. Far from any degeneracy the diagonal SU(3) fields have become three U(1) fields, two of which are independent. When the Hamiltonian matrix elements are all real, the eight λ_i generators become five (*i.e.*, $A_2 = A_5 = A_7 = 0$). Matsika has discussed this recently, giving simple ways to get the signs for the various terms.[57]

Acknowledgements

This work was supported by grants from the U. S. National Science Foundation (CHE-0652830) and the U. S. Department of Energy, Office of Basic Energy Sciences (DE-SC0003976). I would like to thank the following individuals: Joel Bowman for alerting me to the rich history of this area, much of which I originally overlooked; Anna Krylov for asking that a qualitative picture of the gauge field be presented; Jordan Fine for critical reading of the manuscript from a graduate student perspective; John Stanton for a discussion concerning the deep meaning of geometric phase in molecules; and an anonymous reviewer for several helpful comments.

References

1 C. A. Mead and D. G. Truhlar, *J. Chem. Phys.*, 1979, **70**, 2284.
2 C. A. Mead and D. G. Truhlar, *J. Chem. Phys.*, 1982, **77**, 6090.
3 H. C. Longuet-Higgins, U. Öpik, M. H. L. Pryce and R. A. Sack, *Proc. R. Soc. London, Ser. A*, 1958, **244**, 1.
4 H. C. Longuet-Higgins, *Adv. Spectrosc.*, 1961, **2**, 429.
5 G. Herzberg and H. C. Longuet-Higgins, *Discuss. Faraday Soc.*, 1963, **35**, 77.
6 H. C. Longuet-Higgins, *Proc. R. Soc. London, Ser. A*, 1975, **344**, 147.
7 A. J. Stone, *Proc. R. Soc. London, Ser. A*, 1976, **351**, 141.
8 P. W. Atkins and R. S. Friedman, *Molecular Quantum Mechanics*, Oxford University Press, Oxford, 3rd edn, 1996.
9 M. N. R. Ashfold, R. N. Dixon, M. Kono, D. H. Mordaunt and C. L. Reed, *Philos. Trans. R. Soc. London, Ser. A*, 1997, **355**, 1659.
10 A. Bach, J. M. Hutchison, R. J. Holiday and F. F. Crim, *J. Chem. Phys.*, 2003, **118**, 7144.
11 H. A. Jahn and E. Teller, *Proc. R. Soc. London, Ser. A*, 1937, **161**, 220.
12 R. Englman, *The Jahn–Teller Effect in Molecules and Crystals*, Wiley, New York, 1972.
13 I. B. Bersuker and V. Z. Polinger, *Vibronic Interactions in Molecules and Crystals*, Springer, Berlin, 1989.
14 I. B. Bersuker, *The Jahn–Teller Effect*, Cambridge University Press, Cambridge, 2006.
15 *The Jahn–Teller Effect: Fundamentals and Implications for Physics and Chemistry*, ed. H. Köppel, D. R. Yarkony and H. Barentzen, Springer, Berlin, 2009.
16 M. V. Berry, *Proc. R. Soc. London, Ser. A*, 1984, **392**, 45.
17 *Geometric Phases in Physics*, ed. A. Shapere and F. Wilczek, World Scientific, Singapore, 1989.
18 A. Bohm, A. Mostafazadeh, H. Koizumi, Q. Niu and J. Zwanziger, *The Geometric Phase in Quantum Systems*, Springer Verlag, Berlin, 2003.
19 S. Xantheas, S. T. Elbert and K. Ruedenberg, *J. Chem. Phys.*, 1990, **93**, 7519.
20 G. J. Atchity, S. S. Xantheas and K. Ruedenberg, *J. Chem. Phys.*, 1991, **95**, 1862.
21 S. S. Xantheas, G. J. Atchity, S. T. Elbert and K. Ruedenberg, *J. Chem. Phys.*, 1991, **94**, 8054.
22 G. J. Atchity, K. Ruedenberg and A. Nanayakkara, *Theor. Chem. Acc.*, 1997, **96**, 195.
23 D. R. Yarkony, in *Modern Electronic Structure Theory*, ed. D. R. Yarkony, World Scientific, Singapore, 1995, pp. 642–721.
24 D. R. Yarkony, *Rev. Mod. Phys.*, 1996, **68**, 985.
25 D. R. Yarkony, *Theor. Chem. Acc.*, 1997, **98**, 197.
26 D. R. Yarkony, *Acc. Chem. Res.*, 1998, **31**, 511.
27 D. R. Yarkony, Conical Intersections: The New Conventional Wisdom, Feature Article, *J. Phys. Chem. A*, 2001, **105**, 6277.
28 *Conical Intersections: Electronic Structure, Dynamics, and Spectroscopy*, ed. W. Domcke, D. R. Yarkony and H. Köppel, World Scientific, Singapore, 2004.
29 D. R. Yarkony, *Conical Intersections: Their Description and Consequences*, in *Conical Intersections: Electronic Structure, Dynamics, and Spectroscopy*, ed. W. Domcke, D. R. Yarkony and H. Köppel, World Scientific, Singapore, 2004, pp. 41–127.
30 D. R. Yarkony, *Determination of Potential Energy Surface Intersections and Derivative Couplings in the Adiabatic Representation*, in *Conical Intersections: Electronic Structure, Dynamics, and Spectroscopy*, ed. W. Domcke, D.R. Yarkony and H. Köppel, World Scientific, Singapore, 2004, pp. 129–175.
31 H. J. Kupka, *Transitions in Molecular Systems*, Wiley-VCH, Weinheim, 2010.
32 F. Wilczek and A. Zee, *Phys. Rev. Lett.*, 1984, **52**, 2111.
33 J. Moody, A. Shapere and F. Wilczek, *Adiabatic Effective Lagrangians*, in *Geometric Phases in Physics*, ed. A. Shapere and F. Wilczek, World Scientific, Singapore, 1989, pp. 160–183.
34 C. A. Mead, *Phys. Rev. Lett.*, 1987, **59**, 161.
35 T. Pacher, C. A. Mead, L. S. Cederbaum and H. Köppel, *J. Chem. Phys.*, 1989, **91**, 7057.
36 C. A. Mead, *Rev. Mod. Phys.*, 1992, **64**, 51.
37 T. Pacher, L. S. Cederbaum and H. Köppel, *Adv. Chem. Phys.*, 1993, **84**, 293.
38 L. Cederbaum, *Born–Oppenheimer Approximation and Beyond*, in *Conical Intersections: Electronic Structure, Dynamics, and Spectroscopy*, ed. W. Domcke, D. R. Yarkony and H. Köppel, World Scientific, Singapore, 2004, pp. 3–40.
39 H. Köppel, *Diabatic Representation: Methods for the Construction of Diabatic Electronic States*, in *Conical Intersections Electronic Structure, Dynamics, and Spectroscopy*, ed. W. Domcke, D. R. Yarkony and H. Köppel, World Scientific, Singapore, 2004, pp. 175–204.
40 M. Baer, *Mol. Phys.*, 1980, **40**, 1011.
41 J. Stanton, *Mol. Phys.*, 2009, **107**, 1059, particularly the discussion at the end.
42 I. J. R. Aitchison and A. J. G. Hey, *Gauge Theories in Particle Physics, Volume I: From Relativistic Quantum Mechanics to QED*, Taylor and Francis, New York, 2004.
43 I. J. R. Aitchison and A. J. G. Hey, *Gauge Theories in Particle Physics, Volume II: QCD and the Electroweak Theory*, Taylor and Francis, New York, 2004.
44 M. Guidry, *Gauge Field Theories, an Introduction with Applications*, Wiley Interscience, New York, 1999.
45 Y. Aharonov and D. Bohm, *Phys. Rev.*, 1959, **115**, 485.
46 M. Peshkin and A. Tonomura, *The Aharonov–Bohm Effect*, Springer-Verlag, Lecture Notes in Physics 340, Heidelberg, 1989.
47 Y. Aharonov and D. Rohrlich, *Quantum Paradoxes*, Wiley VCH, Weinheim, Germany, 2005.
48 S. M. Carroll, *Spacetime and Geometry*, Addison Wesley, New York, 2004.
49 C. Misner, K. Thorne and J. Wheeler, *Gravitation*, W. H. Freeman and Co., New York, 1973.
50 W. Rindler, *Relativity: Special, General, and Cosmological*, Oxford University Press, Oxford, 2nd edn, 2006.
51 B. Schutz, *Geometric Methods of Mathematical Physics*, Cambridge University Press, Cambridge, 1999.
52 D. W. Henderson and D. Taimina, *Experiencing Geometry*, Pearson Prentice Hall, Upper Saddle River, New Jersey, 2005.
53 S. Han and D. R. Yarkony, *J. Chem. Phys.*, 2003, **119**, 11561.
54 M. S. Schuurman and D. R. Yarkony, *J. Chem. Phys.*, 2006, **124**, 124109.
55 M. S. Schuurman and D. R. Yarkony, *J. Phys. Chem. B*, 2006, **110**, 19031.
56 D. E. Manolopoulos and M. S. Child, *Phys. Rev. Lett.*, 1999, **82**, 2223.
57 S. Matsika, *Three-State Conical Intersections*, in *Conical Intersections: Theory, Computation and Experiment*, ed. W. Domcke, D. R. Yarkony and H. Köppel, World Scientific, Singapore, 2011.

58 *Conical Intersections: Theory, Computation and Experiment*, ed. W. Domcke, D. R. Yarkony and H. Köppel, World Scientific, Singapore, 2011.
59 B. Kusse and E. Westwig, *Mathematical Physics: Applied Mathematics for Scientists and Engineers*, Wiley, New York, 1998.
60 W. N. Cottingham and D. A. Greenwood, *An Introduction to the Standard Model of Particle Physics*, Cambridge University Press, Cambridge, 2nd edn, 2007.
61 I. J. R. Aitchison and A. J. G. Hey, *Gauge Theories in Particle Physics, Volume II: QCD and the Electroweak Theory*, Taylor and Francis, New York, 2004, ch. 13.
62 C. N. Yang and R. Mills, *Phys. Rev.*, 1954, **96**, 191.
63 R. Penrose and W. Rindler, *Spinors and Space-time Volume I. Two-Spinor Calculus and Relativistic Fields*, Cambridge Monographs on Mathematical Physics, Cambridge University Press, Cambridge, 1986.
64 C. T. H. Davies, C. McNeile, K. Y. Wong, E. Follana, R. Horgan, K. Hornbostel, G. P. Lepage, J. Shigemitsu and H. Trottier, *Phys. Rev. Lett.*, 2010, **104**, 132003.
65 C. Itzykson and J.-B. Zuber, *Quantum Field Theory*, Dover, New York, 1980, p. 565.
66 M. V. Berry, *Principles of Cosmology and Gravitation*, Cambridge University Press, Cambridge, 1976.
67 M. Guidry, *Gauge Field Theories, an Introduction with Applications*, Wiley Interscience, New York, 1999, ch. 9.
68 I. B. Bersuker has written informative pieces regarding history, misunderstandings, and current status of the Jahn–Teller effect. See the preface in ref. 14, and his chapter (pp. 3–24) in ref. 15.
69 Electroweak symmetry breaking cannot be judged to be finished without the Higgs boson, the quest for which is ongoing at the Large Hadron Collider (LHC) at CERN.
70 K. Huang, *Fundamental Forces of Nature: The Story of Gauge Fields*, World Scientific, Singapore, 2007.
71 J. D. Coe and T. J. Martinez, *J. Am. Chem. Soc.*, 2005, **127**, 4560.
72 M. Guidry, *Gauge Field Theories, an Introduction with Applications*, Wiley Interscience, New York, 1999, p. 157.
73 D. S. L. Abergel, V. Apalkov, J. Berashevich, K. Ziegler and T. Chakraborty, *Adv. Phys.*, 2009, **59**, 261.

