Covariant Derivatives in Quantum Mechanics, Aharonov–Bohm Effect, and Magnetic Monopoles

Covariant Derivatives in Quantum Mechanics

In my my notes on the local phase symmetry, I have defined the covariant derivative of a charged field $\phi(x)$ as $D_\mu\phi(x) = \partial_\mu\phi(x) + iqA_\mu(x)\phi(x)$. In 3D-vector notations and in Gauss units, this definition becomes

$$\mathbf{D} \;=\; \nabla \;-\; \frac{iq}{\hbar c}\mathbf{A}(\mathbf{x},t), \qquad D_t \;=\; \frac{\partial}{\partial t} \;+\; \frac{iq}{\hbar}A^0(\mathbf{x},t). \tag{1}$$

In this section, we shall see how these covariant derivatives fit into ordinary quantum mechanics of a charged particle.

A classical charged particle in EM background has canonical momentum $\mathbf{p}$ different from the ordinary kinematic momentum $\vec{\pi} = m\mathbf{v}$, namely

$$\mathbf{p} \;=\; m\mathbf{v} \;+\; \frac{q}{c}\mathbf{A}(\mathbf{x}), \tag{2}$$

hence classical Hamiltonian

$$H(\mathbf{x},\mathbf{p}) \;=\; \frac{m}{2}\mathbf{v}^2 \;+\; qA^0(\mathbf{x}) \;=\; \frac{1}{2m}\left(\mathbf{p} \;-\; \frac{q}{c}\mathbf{A}(\mathbf{x})\right)^2 \;+\; qA^0(\mathbf{x}). \tag{3}$$

In quantum mechanics, this translates to the Hamiltonian operator

$$\hat{H} \;=\; \frac{1}{2m}\left(\hat{\mathbf{p}} \;-\; \frac{q}{c}\mathbf{A}(\hat{\mathbf{x}})\right)^2 \;+\; qA^0(\hat{\mathbf{x}}), \tag{4}$$

where $\hat{\mathbf{p}}$ is the canonical momentum operator which obeys the canonical commutation relations with the coordinate operator $\hat{\mathbf{x}}$,

$$[\hat{x}_i,\hat{x}_j] \;=\; 0, \quad [\hat{p}_i,\hat{p}_j] \;=\; 0, \quad [\hat{x}_i,\hat{p}_j] \;=\; i\hbar\delta_{ij}\,. \tag{5}$$

Consequently, in the coordinate basis for the wave functions, the canonical momentum op-

erator $\hat{\mathbf{p}}$ acts as a gradient, or rather

$$\hat{\mathbf{p}}\psi(\mathbf{x}) \;=\; -i\hbar\nabla\psi(\mathbf{x}). \tag{6}$$

As to the kinematic momentum $\vec{\pi} = m\mathbf{v}$, in quantum mechanics it's defined as

$$\hat{\vec{\pi}} \;=\; \hat{\mathbf{p}} \;-\; \frac{q}{c}\mathbf{A}(\hat{\mathbf{x}}), \tag{7}$$

so in the coordinate basis it acts as

$$\hat{\vec{\pi}}\psi(\mathbf{x}) \;=\; -i\hbar\nabla\psi(\mathbf{x}) \;-\; \frac{q}{c}\mathbf{A}(\mathbf{x})\psi(\mathbf{x}) \;=\; -i\hbar\left(\nabla - \frac{iq}{\hbar c}\mathbf{A}(\mathbf{x})\right)\psi(\mathbf{x}) \;=\; -i\hbar\mathbf{D}\psi(\mathbf{x}) \tag{8}$$

where $\mathbf{D}$ is the covariant space derivative (1). Consequently, in the coordinate basis, the Hamiltonian operator (4) acts as

$$\hat{H}\psi(\mathbf{x}) \;=\; \frac{-\hbar^2}{2m}\,\mathbf{D}^2\psi(\mathbf{x}) \;+\; qA^0(\mathbf{x})\psi(\mathbf{x}), \tag{9}$$

where the vector potential $\mathbf{A}$ hides inside the covariant derivative $\mathbf{D}$.

Now consider the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\,|\psi\rangle \;=\; \hat{H}\,|\psi\rangle \tag{10}$$

which in the coordinate basis becomes

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{x},t) \;=\; \frac{-\hbar^2}{2m}\,\mathbf{D}^2\psi(\mathbf{x},t) \;+\; qA^0(\mathbf{x},t)\psi(\mathbf{x},t). \tag{11}$$

Moving the second term on the RHS to the LHS of the equation, we get

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{x},t) \;-\; qA^0(\mathbf{x},t)\psi(\mathbf{x},t) \;=\; \frac{-\hbar^2}{2m}\,\mathbf{D}^2\psi(\mathbf{x},t) \tag{12}$$

where the LHS amounts to

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{x},t) \;-\; qA^0(\mathbf{x},t)\psi(\mathbf{x},t) \;=\; i\hbar\left(\frac{\partial}{\partial t} \;+\; \frac{iq}{\hbar}A^0(\mathbf{x},t)\right)\psi(\mathbf{x},t) \;=\; i\hbar D_t\psi(\mathbf{x},t), \tag{13}$$

D_t being the covariant time derivative (1).

Thus, we arrive at the covariant Schrödinger equation

$$i\hbar D_t\psi(\mathbf{x},t) \;=\; \frac{-\hbar^2}{2m}\,\mathbf{D}^2\psi(\mathbf{x},t). \tag{14}$$

In this form, neither electric potential A^0 nor the magnetic potential $\mathbf{A}$ are manifest in this equation; instead, they are hiding inside the covariant derivatives D_t and $\mathbf{D}$.[⋆]

Note that the covariant derivatives are covariant only when the fields or wave-functions on which they act undergo local phase transforms simultaneously with the gauge transform of the EM potentials. Specifically, the covariant equations (14) require a gauge transform

$$\mathbf{A}'(\mathbf{x},t) \;=\; \mathbf{A}(\mathbf{x},t) \;+\; \nabla\Lambda(\mathbf{x},t), \quad A^{0\prime}(\mathbf{x},t) \;=\; A^0(\mathbf{x},t) \;-\; \frac{1}{c}\frac{\partial}{\partial t}\Lambda(\mathbf{x},t) \tag{16}$$

to be accompanies by the local phase transform of the wave function according to

$$\psi'(\mathbf{x},t) \;=\; \exp\left(\frac{iq}{\hbar c}\Lambda(\mathbf{x},t)\right)\times\psi(\mathbf{x},t). \tag{17}$$

In the next section of these notes we shall see how this phase transform of the wave function gives rise to the Aharonov–Bohm effect. But the Aharonov–Bohm effect is best described in terms of the propagation amplitude — also called the evolution kernel — and the way it transforms under EM gauge transforms. The propagation amplitude $U(y \leftarrow x)$ is defined as an amplitude of a particle initially at point $\mathbf{x}$ at time x^0 to reach the point $\mathbf{y}$ at

---

⋆ As written, eq. (14) applies to a spinless non-relativistic charged particle. For a non-relativistic charged particle of spin $= \frac{1}{2}$ — like an electron or a proton — it becomes

$$i\hbar D_t\psi(\mathbf{x},t) \;=\; -\frac{\hbar^2}{2m}\,\mathbf{D}^2\psi(\mathbf{x},t) \;-\; \frac{gq\hbar}{4mc}\mathbf{B}(\mathbf{x},t)\cdot\vec{\sigma}\psi(\mathbf{x},t) \tag{15}$$

where $\sigma_x, \sigma_y, \sigma_z$ are the Pauli matrices acting on the spin and g is the gyromagnetic factor. (For the electron, $g \approx 2$.) Since the magnetic field $\mathbf{B}$ is gauge-invariant, eq. (15) is just as covariant as eq. (14).

a later time y^0; in Dirac notations

$$U(y \leftarrow x) \;=\; \big\langle \mathbf{y}, y^0 | \mathbf{x}, x^0 \big\rangle_{\text{Heisenberg}} \;=\; \langle \mathbf{y} | \exp\left(-i\frac{y^0 - x^0}{\hbar}\hat{H}\right) | \mathbf{x} \rangle_{\text{Schroedinger}} \,. \tag{18}$$

Consequently, given the wave function (in the coordinate basis) at time x^0, the propagation amplitude gives us the wave function at a later time y^0 as

$$\psi(\mathbf{y}, y^0) \;=\; \int d^3\mathbf{x}\, U(y \leftarrow x)\psi(\mathbf{x}, x^0). \tag{19}$$

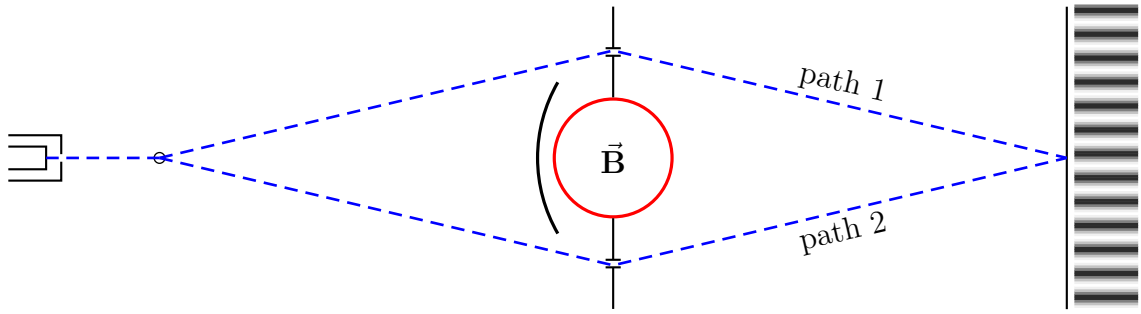
Now let's apply this formula to the propagation amplitude and the wave function of a charged particle. Under a gauge transform of the EM potentials, the wave function changes its phase according to eq. (17). Hence, to compensate the phase change of $\psi(\mathbf{x}, x^0)$ inside the integral (19), the propagation amplitude $U(y \leftarrow x)$ must change its phase by an opposite factor $\exp\big(-(iq/\hbar c)\Lambda(\mathbf{x}, x^0)\big)$. At the same time, to change the phase of $\psi(\mathbf{y}, y^0)$ on the LHS of eq. (19), the $U(y \leftarrow x)$ must change its phase by $\exp\big(+(iq/\hbar c)\Lambda(\mathbf{y}, y^0)\big)$. Altogether, under a gauge transform of the EM potentials, the propagation amplitude of a charged particle changes its phase by

$$U'(y \leftarrow x) \;=\; U(y \leftarrow x) \times \exp\left(\frac{iq}{\hbar c}\big(\Lambda(y) - \Lambda(x)\big)\right). \tag{20}$$

Aharonov–Bohm Effect

In classical mechanics, the motion of a charged particle depends only on the electric and magnetic tension fields $\mathbf{E}$ and $\mathbf{B}$; the potentials A^0 and $\mathbf{A}$ do not have any direct effect. Also, the motion depends only on the $\mathbf{E}$ and $\mathbf{B}$ fields along the particle's world-line — the EM fields in some volume of space the particle never goes through do not affect it at all. But *in quantum mechanics, interference between two trajectories a charged particle might take depends on the magnetic field between the trajectories, even if along the trajectories themselves* $\mathbf{B} = 0$. This effect was first predicted by Werner Ehrenberg and Raymond E. Siday in 1949, but their paper was not noticed until the effect was re-discovered theoretically by David Bohm and Yakir Aharonov in 1959 and then confirmed experimentally by R. G. Chambers in 1960.

Consider the following idealized experiment: Take a two-slit electron interference setup, and put a solenoid between the two slits as shown below:



The solenoid is thin, densely wound, and very long, so the magnetic field outside the solenoid is negligible. Inside the solenoid there is a strong $\mathbf{B}$ field, but the electrons do not go there; instead, they fly outside the solenoid along paths 1 and 2. But despite $\mathbf{B} = 0$ along both paths, the magnetic flux Φ inside the solenoid affects the interference pattern between the two paths.

The key to the Aharonov–Bohm effect is the vector potential $\mathbf{A}$. Outside the solenoid $\mathbf{B} = \nabla \times \mathbf{A} = 0$ but $\mathbf{A} \neq 0$ because for any closed loop surrounding the solenoid we have a non-zero integral

$$\oint_{\text{loop}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} \;=\; \oint_{\substack{\text{inside the loop} \\ \text{including the solenoid}}} \mathbf{B}(\mathbf{x}) \cdot d^2\mathbf{Area} \;=\; \Phi. \tag{21}$$

Locally, $\nabla \times \mathbf{A} = 0$ makes the vector potential a gradient of some function so we may gauge it away:

$$\mathbf{A}(\mathbf{x}) \;\to\; \mathbf{A}'(\mathbf{x}) \;=\; \mathbf{A}(\mathbf{x}) \;+\; \nabla\Lambda(\mathbf{x}) \;=\; 0 \quad \text{for some } \Lambda(\mathbf{x}), \tag{22}$$

but *globally* no single-valued $\Lambda(\mathbf{x})$ can gauge away the vector potential along both paths around the solenoid. Indeed, consider two points — the electron gun at $\mathbf{x}_0$ and some point on the screen at $\mathbf{y}$, and let

$$\Delta\Lambda \;=\; \Lambda(\mathbf{y}) \;-\; \Lambda(\mathbf{x}_0). \tag{23}$$

Then using two different electron's paths from $\mathbf{x}_0$ to $\mathbf{y}$ gives two different values of the $\Delta\Lambda$:

$$\Delta\Lambda(\text{path}) \;=\; \int\limits_{\text{path}} \nabla\Lambda\cdot d\mathbf{x} \;=\; -\int\limits_{\text{path}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x}$$
$$\text{since } \mathbf{A} + \nabla\Lambda \;=\; 0, \tag{24}$$

$$\Delta\Lambda(\text{path\#1}) \;-\; \Delta\Lambda(\text{path\#2}) \;= -\int\limits_{\text{path\#1}} \mathbf{A}\cdot d\mathbf{x} \;+\; \int\limits_{\text{path\#2}} \mathbf{A}\cdot d\mathbf{x}$$
$$= \;-\oint \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;=\; -\Phi \;\neq\; 0, \tag{25}$$

which is utterly impossible for any single-valued $\Lambda(\mathbf{x})$. Instead, we have two separate gauge transforms parametrized by two different $\Lambda(\mathbf{x})$: the $\Lambda_1(\mathbf{x})$ that gauges away $\mathbf{A}(\mathbf{x})$ along the path #1, and the $\Lambda_2(\mathbf{x})$ that gauges away $\mathbf{A}(\mathbf{x})$ along the path #2, thus

$$\begin{aligned} \nabla\Lambda_1(\mathbf{x}) \;&=\; -\mathbf{A}(\mathbf{x}) \text{ [along path\#1]},\\ \nabla\Lambda_2(\mathbf{x}) \;&=\; -\mathbf{A}(\mathbf{x}) \text{ [along path\#2]},\\ \text{and}\quad \Lambda_1(\mathbf{x}) \;&\neq \Lambda_2(\mathbf{x}). \end{aligned} \tag{26}$$

In quantum mechanics, a gauge transform affects not only the vector potential but also the phase of a charged particle's wave function and hence the propagation amplitudes, *cf.* eq. (20). So consider an electron traveling along some path from the electron gun at $\mathbf{x}_0$ to some point $\mathbf{y}$ on the screen through a region where these is no magnetic field, $\mathbf{B} = 0$, but the vector potential does not vanish. We assume this $\mathbf{A}(\mathbf{x})$ to be time-independent, so we may gauge it away using a time-independent $\Lambda(\mathbf{x})$ without raising an electric potential, $\mathbf{A}' = \mathbf{A} + \nabla\Lambda = 0$ while $A^{0\prime} = A^0 = 0$. Gauging away the vector potential also changes the phase of the evolution amplitude according to

$$U_0(\mathbf{y}\leftarrow\mathbf{x}_0) \;=\; U_{\mathbf{A}}(\mathbf{y}\leftarrow\mathbf{x}_0)\times\exp\left(\frac{iq}{\hbar c}\big(\Lambda(\mathbf{y})-\Lambda(\mathbf{x}_0)\big)\right) \tag{27}$$

where $U_{\mathbf{A}}$ is the initial amplitude in presence of the vector potential $\mathbf{A}(\mathbf{x})$, U_0 is the amplitude

resulting from gauging $\mathbf{A}$ away, and

$$\Lambda(\mathbf{y}) \;-\; \Lambda(\mathbf{x}_0) \;=\int\limits_{\mathbf{x}_0}^{\mathbf{y}} \nabla\Lambda(x)\cdot d\mathbf{x} \;=\; -\int\limits_{\mathbf{x}_0}^{\mathbf{y}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \quad \langle\!\langle\, \text{since } \mathbf{A}' = \mathbf{A} + \nabla\Lambda = 0 \,\rangle\!\rangle. \tag{28}$$

Consequently,

$$U_0(\mathbf{y} \leftarrow \mathbf{x}_0) \;=\; U_{\mathbf{A}}(\mathbf{y} \leftarrow \mathbf{x}_0) \times \exp\left(-\frac{iq}{\hbar c}\int\limits_{\mathbf{x}_0}^{\mathbf{y}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x}\right), \tag{29}$$

or equivalently

$$U_{\mathbf{A}}(\mathbf{y} \leftarrow \mathbf{x}_0) \;=\; U_0(\mathbf{y} \leftarrow \mathbf{x}_0) \times \exp\left(+\frac{iq}{\hbar c}\int\limits_{\mathbf{x}_0}^{\mathbf{y}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x}\right). \tag{30}$$

Thus, given the amplitude U_0 in the total absence of a vector potential, turning on a pure-gauge vector potential changes the amplitude's phase according to eq. (30).

In the Aharonov–Bohm experiment we have two different paths from the same point $\mathbf{x}_0$ (the electron gun) to the same point $\mathbf{y}$ on the screen. Along each path $\mathbf{B} = 0$ but $\mathbf{A} \neq 0$, and the amplitudes depend on the vector potential according to eq. (30):

$$\begin{aligned} U_{\mathbf{A}}^{\text{path 1}}(\mathbf{y} \leftarrow \mathbf{x}_0) \;&=\; U_0^{\text{path 1}}(\mathbf{y} \leftarrow \mathbf{x}_0) \times \exp\left(+\frac{iq}{\hbar c}\int\limits_{\text{path 1}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x}\right), \\ U_{\mathbf{A}}^{\text{path 2}}(\mathbf{y} \leftarrow \mathbf{x}_0) \;&=\; U_0^{\text{path 2}}(\mathbf{y} \leftarrow \mathbf{x}_0) \times \exp\left(+\frac{iq}{\hbar c}\int\limits_{\text{path 2}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x}\right). \end{aligned} \tag{31}$$

The interference pattern on the screen depends on the phase difference

$$\Delta\varphi(\mathbf{y}) \;=\; \arg\Big(U^{\text{path 1}}(\mathbf{y} \leftarrow \mathbf{x}_0)\Big) \;-\; \arg\Big(U^{\text{path 2}}(\mathbf{y} \leftarrow \mathbf{x}_0)\Big) \tag{32}$$

between the two amplitudes. In light of eqs. (31), this phase difference depends on the vector

potential $\mathbf{A}$ as

$$\Delta_{\mathbf{A}}\varphi(\mathbf{y}) \;=\; \Delta_0\varphi(\mathbf{y}) \;+\; \frac{q}{\hbar c}\int\limits_{\text{path 1}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;-\; \frac{q}{\hbar c}\int\limits_{\text{path 2}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x}\,. \tag{33}$$

Moreover, the difference between the two integrals here is nothing but the magnetic flux Φ inside the solenoid! Indeed, consider a closed loop around the solenoid that first follows path 1 from the electron gun to the screen and then goes back to the electron gun along path 2 (in reverse). For this loop,

$$\int\limits_{\text{path 1}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;-\; \int\limits_{\text{path 2}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;=\; \oint\limits_{\text{closed loop}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;=\; \Phi, \tag{34}$$

hence

$$\Delta_{\mathbf{A}}\varphi(\mathbf{y}) \;=\; \Delta_0\varphi(\mathbf{y}) \;+\; \frac{q}{\hbar c}\times\Phi. \tag{35}$$

Thus, even though $\mathbf{B} = 0$ along both paths an electron might take from the gun to the screen, the quantum interference between the paths depends on the magnetic flux in the solenoid!

Now consider the mathematical side of the Aharonov–Bohm effect — the *cohomology* of the vector potential $\mathbf{A}(\mathbf{x})$. In a topologically trivial space — like the flat 3D space without any holes — specifying $\mathbf{A}(\mathbf{x})$ *modulo* gauge transforms $\mathbf{A}(\mathbf{x}) \to \mathbf{A}(\mathbf{x}) - \nabla\Lambda(\mathbf{x})$ is equivalent to specifying the magnetic field $\mathbf{B}(\mathbf{x}) = \nabla\times\mathbf{A}$. However, in spaces with holes the vector potential modulo $\nabla\Lambda(\mathbf{x})$ for *single-valued* $\Lambda(\mathbf{x})$ contains more information than the magnetic field: In addition to $\mathbf{B}(\mathbf{x})$ for $\mathbf{x}$ outside the holes, the vector potential also knows the magnetic fluxes through the holes! Indeed, the integrals along closed loops

$$\oint\limits_{\text{loop}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;=\; \Phi(\text{loop}) \tag{36}$$

are gauge-invariant *for single-valued* $\Lambda(\mathbf{x})$, and when $\nabla\times\mathbf{A} \equiv 0$ everywhere outside the holes, then the fluxes (36) depend only on the topologies of the loops in question — which hole(s) they surround and how many times. In math, such integrals are called *cohomologies* of the one-form $\mathbf{A}(\mathbf{x})$.

In classical mechanics, the motion of a charged particle depends on the magnetic field $\mathbf{B}$ in the region of space through which the particle travels, and it does not care about any cohomologies of the vector potential $\mathbf{A}$. But in quantum mechanics, the Aharonov–Bohm effect makes quantum interference sensitive to the cohomologies that the classical mechanics does not see. Specifically, when the space has some holes through which the particle does not get to travel — like the solenoid (and a bit of space around it) in the AB experiment — the interference between alternative paths on different sides of a hole depends on the cohomology of $\mathbf{A}$ for that hole — *i.e.*, the magnetic flux through the hole.

To be precise, the interference between two paths depends on the phase difference (35) only modulo 2π — changing the phase by $2\pi n$ for some integer n would not affect the interference at all. Consequently, the Aharonov–Bohm effect is un-detectable for

$$\Phi \;=\; \frac{2\pi\hbar c}{q} \times \text{an integer}, \tag{37}$$

or in other words, the AB effect measures only the fractional part of the magnetic flux through the solenoid in units of

$$\Phi_1 \;=\; \frac{2\pi\hbar c}{|q|} \tag{38}$$

where q is the electric charge of the particles used in the experiment. For example, a SQUID (SuperConducting Quantum Interferometry Device) measures the magnetic flux through a hole surrounded by superconductor using Aharonov–Bohm–like interference of the Cooper pairs in the superconductor. Since a Cooper pair has electric charge $-2e$, a SQUID measure only the fractional part of the flux in units of

$$\Phi_{\text{squid}} \;=\; \frac{2\pi\hbar c}{2e} \;=\; 2.067\,833\,667(52) \times 10^{-7}\ \text{Mx} \quad (\text{Maxwells or Gauss} \times \text{cm}^2). \tag{39}$$

Note that particles of different charges would measure the fractional part of the magnetic flux Φ in different units! Thus, were Nature kind enough to provide us with two particle species with an irrational charge ratio q_1/q_2, then in principle we could have measured the whole magnetic flux Φ and not just its fractional part in some units.[⋆] However, in reality

⋆ To be precise, we could have measure the fractional parts of Φ in different units Φ_1 and Φ_2, but for irrational ratio Φ_1/Φ_2 this would have allowed us to reconstruct the whole flux Φ.

all electric charges are integral multiplets of the fundamental charge units e. Consequently, the AB effect using any existing particle species can measure only the fractional part of the magnetic flux in universal units

$$\Phi_u \;=\; \frac{2\pi\hbar c}{e} \;=\; 2\Phi_{\rm squid}\,. \tag{40}$$

Mathematically, this reduction of our ability to measure the cohomology of the $\mathbf{A}$ field is related to the compactification of the gauge symmetry group when all charges are integer multiples of e. Indeed, consider a generic gauge transform parametrized by $\Lambda(\mathbf{x},t)$ and let

$$u(\mathbf{x},t) \;=\; \exp\bigl(i(e/\hbar c)\Lambda(\mathbf{x},t)\bigr) \in U(1). \tag{41}$$

The $U(1)$ here is a special case of $U(N)$ — the group of complex unitary $N\times N$ matrices. For $N=1$, such a matrix is simply a unimodular complex number u; in other words, the u's in eq. (41) live on a unit circle in the complex plane. As a group, the $U(1)$ is the group of phase symmetries, where changing the phase by $2\pi\times$ an integer has no effect whatsoever.

Taking a spacetime derivative of eq. (41) we get

$$\partial_\mu u(x) \;=\; u(x)\times\frac{ie}{\hbar c}\,\partial_\mu\Lambda(x), \tag{42}$$

hence the gauge transform of the 4-vector field $A_\mu(x)$ can be restated in terms of $u(x)$ as

$$A'_\mu(x) \;=\; A_\mu(x) \;+\; \frac{\hbar c}{e}\times iu^{-1}(x)\partial_\mu u(x). \tag{43}$$

At the same time, a charged field $\Psi(x)$ of charge $q=n\times e$ transforms as

$$\Psi'(x) \;=\; \Psi(x)\times\exp\left(\frac{inq}{\hbar c}\Lambda(x)\right) \;=\; \Psi(x)\times\left[\exp\left(\frac{iq}{\hbar c}\Lambda(x)\right)\right]^n \;=\; \Psi(x)\times u^n(x). \tag{44}$$

Thus, if all the fields have integer charges n in units of e, then any single-valued unimodular $u(x)$ parametrizes a single-valued gauge/phase transform, even if $\Lambda(x)=(\hbar c/e)\arg(u(x))$ happens to be multi-valued!

In a topologically trivial spacetime, one can write down a single-valued $\Lambda(x)$ for any single-valued $u(x)$, but this is not true in a spacetime with holes. For example, let's focus on time-independent gauge transforms and consider a space with a cylindrical hole (the solenoid in the AB experiment); in cylindrical coordinates (ρ, ϕ, z), the points outside the hole have $\rho > \rho_h$. The angle coordinate ϕ is multi-valued modulo 2π, but $\exp(i\phi)$ is single valued. In a space without the hole, $\exp(i\phi)$ would be ill-defined along the axis, but outside the hole it's a perfectly well-defined single-valued function of $\mathbf{x}$. Consequently, letting

$$\Lambda(\rho,\phi,z) \;=\; \frac{\hbar c}{e}\times\phi \quad\Longleftrightarrow\quad u(\rho,\phi,z) \;=\; e^{+i\phi} \tag{45}$$

would give us a multi-valued $\Lambda(\mathbf{x})$ but a single-valued $u(\mathbf{x})$. In the gauge theory with integral charges only, such gauge transforms are legitimate — as long as all the charged fields and the $\mathbf{A}(\mathbf{x})$ transform in a single-valued fashion, we don't care if the $\Lambda(\mathbf{x})$ parameter itself is single-valued or multi-valued.

However, the magnetic fluxes through holes in space are not invariant under gauge transforms with multi-valued Λ's. Instead, they change by integral multiplets of the Aharonov–Bohm flux unit (40). Indeed, for a gauge transform (45), the vector potential outside the hole changes to

$$\mathbf{A}'(\mathbf{x}) \;=\; \mathbf{A}(\mathbf{x}) \;+\; \frac{\hbar c}{e}\nabla\phi \;=\; \mathbf{A}(\mathbf{x}) \;+\; \frac{\hbar c}{e}\,\frac{\mathbf{n}_\phi}{\rho} \tag{46}$$

where $\mathbf{n}_\phi$ is the unit vector in the ϕ direction. Consequently, the magnetic flux through the hole changes by

$$\Phi' \;-\; \Phi \;=\; \frac{\hbar c}{e}\oint \nabla\phi\cdot d\mathbf{x} \;=\; \frac{\hbar c}{e}\int_0^{2\pi} d\phi \;=\; \frac{2\pi\hbar c}{e} \;\equiv\; \Phi_u\,. \tag{47}$$

Likewise, we may change the flux by any integer multiple k of the flux unit Φ_u using

$$u(\mathbf{x}) \;=\; e^{ik\phi} \quad\text{(single valued for integer } k) \quad\Longrightarrow\quad \Phi' \;=\; \Phi \;+\; k\times\Phi_u\,. \tag{48}$$

Consequently, specifying the vector potential $\mathbf{A}(\mathbf{x})$ modulo gauge transforms with single-valued $u(\mathbf{x})$ phases would give us fluxes through holes in space only modulo Φ_u; in other

words, we can find the fractional parts of those fluxes (in units of Φ_u) but not the whole parts. The Aharonov–Bohm effect measures precisely these data — the fractional parts of the fluxes through holes. The whole parts of the fluxes are not detectable because they are gauge-dependent in the theory with a compact group $U(1)$ of local phase symmetries.

Magnetic Monopoles

The easiest way to visualize a magnetic monopole is by considering a pole of a long, thin magnet or an end point of a long, thin solenoid. Let us choose our coordinates such that the pole is at the origin and the magnet goes along the negative z semi-axis. Then, in spherical coordinates (r, θ, ϕ), everywhere outside the magnet

$$\mathbf{B}(r, \theta, \phi) \;=\; \frac{m}{r^2}\, \mathbf{n}_r \tag{49}$$

where $\mathbf{n}_r$ us the unit vector in the radial direction, while inside the magnet there is magnetic flux $4\pi m$ towards the pole.

Suppose the magnet is infinitely thin, infinitely long and does not interact with the rest of the universe except through the magnetic field it carries. Classically, all one can observe under such circumstances is the magnetic field (49), so for all intents and purposes we have a magnetic monopole of magnetic charge $M = m$. (In Gauss units; in rationalized units the magnetic charge is $M = 4\pi m$.) In quantum mechanics however, one can also detect the Aharonov-Bohm effect due to the magnetic flux $4\pi m$ inside the magnet, and that would make the magnet itself detectable along its whole length. Moreover, in quantum field theory the AB effect would disturb the free-wave modes of the charged fields — instead of the plane waves we would get eigen-waves of some $\mathbf{x}$-dependent differential operator. This would give rise to a Casimir effect — a finite and detectable change of the net zero point energy. For a long thin magnet this Casimir energy would be proportional to the magnet's length, so the magnet would behave as a string with finite tension force T. Consequently, the two poles of the magnet would not be able to separate from each other to infinite distance and acts as independent magnetic monopoles. Instead, the North pole and the South pole would pull each other with a finite force T no matter how far they get from each other.

However, the Aharonov–Bohm effect disappears when the magnetic flux $4\pi m$ is an integral multiplet of $2\pi\hbar c/q$. Consequently, for an infinitely thin magnet there would not be any Casimir effect, hence no string tension, and the poles would be allowed to move independently from each other as if they were separate magnetic monopoles. Since this can happen only when the magnetic flux is not detectable by the AB effect, this gives rise to the Dirac's quantization condition: *For all magnetic monopoles in the universe and for all electrically-charged particles in the universe,*

$$M \times q \;=\; (\tfrac{1}{2}\hbar c) \times \text{an integer} \tag{50}$$

in Gauss units; in rationalized $\hbar = c = 1$ units, this condition reads

$$M \times q \;=\; 2\pi \times \text{an integer}. \tag{51}$$

Consequently, *if there is a magnetic monopole anywhere in the universe, all electrical charges must be quantized.*

A more rigorous argument was made by P. A. M. Dirac himself years before the discovery of the Aharonov-Bohm effect. Instead of using just one vector potential $\mathbf{A}(\mathbf{x})$ to describe the magnetic field of a monopole, Dirac have used two potentials $\mathbf{A}_N(\mathbf{x})$ and $\mathbf{A}_S(\mathbf{x})$ related by a gauge transform. From the mathematical point of view, Dirac monopole is a *gauge bundle*, a construction that generalizes multiple coordinate patches in Riemannian geometry.

Most Riemannian manifolds cannot be covered by a single coordinate system without singularities or multi-valuedness. Instead, one covers the manifold with several overlapping patches and uses different coordinate systems for each patch. This is OK as long as: (1) the patches overlap their neighbors and collectively cover the whole manifold; (2) each patch has a single-valued non-singular coordinate system; (3) in the overlap regions, the coordinate systems of the overlapping patches map onto each other without singularities, *i.e.*, the derivatives $\partial x^\mu_{(1)}/\partial x^\nu_{(2)}$ are all finite and the matrix of those derivatives has a non-zero determinant (the Jacobian).

In a gauge bundle, different patches covering a manifold have not only different coordinate systems but also different gauges for the $A^\mu(x)$ and charged fields. But in the overlap

regions, all fields from the overlapping patches are related to each other by a gauge transform, *eg.,*

$$A_\mu^{(2)}(x) \;=\; A_\mu^{(1)}(x) \;-\; \partial_\mu \Lambda^{1,2}(x), \quad \text{each } \Psi_a^{(2)}(x) \;=\; \Psi_a^{(1)}(x) \times \exp\bigl(i q_a \Lambda^{1,2}(x)\bigr) \tag{52}$$

in the overlap between patches 1 and 2. Eq. (52) is written for the abelian gauge symmetry, but there are suitable generalizations to the non-abelian symmetry groups. In string theory, abelian and non-abelian gauge bundles on curved 6D manifolds play a very important role in obtaining effective four-dimensional theories from the ten-dimensional superstring.

Dirac himself did not use the gauge bundle language, he simply divided the space outside the monopole itself into two overlapping regions and wrote different but gauge-equivalent vector potentials for each region. In spherical coordinates (r, θ, ϕ), the Northern region (N) spans latitudes $0 \le \theta < \pi - \epsilon$ while the Southern region (S) spans $\epsilon < \theta \le \pi$; the two regions overlap in a broad band around the equator. The vector potentials for the two regions are

$$\begin{aligned} \mathbf{A}_N(r,\theta,\phi) \;&=\; m(+1-\cos\theta)\nabla\phi \;=\; m\,\frac{+1-\cos\theta}{r\,\sin\theta}\,\mathbf{n}_\phi\,, \\ \mathbf{A}_S(r,\theta,\phi) \;&=\; m(-1-\cos\theta)\nabla\phi \;=\; m\,\frac{-1-\cos\theta}{r\,\sin\theta}\,\mathbf{n}_\phi\,; \end{aligned} \tag{53}$$

The two potentials are gauge-equivalent:

$$\mathbf{A}_N \;-\; \mathbf{A}_S \;=\; \frac{2m}{r\,\sin\theta}\,\mathbf{n}_\phi \;=\; 2m\nabla\phi \tag{54}$$

so they lead to the same magnetic field, namely (49). Indeed,

$$\begin{aligned} \nabla \times \mathbf{A}_{\mathrm{N\,or\,S}} \;&=\; \nabla \times (m(\pm 1-\cos\theta)\nabla\phi) \\ &=\; m\,(\nabla(\pm 1-\cos\theta)) \times \nabla\phi \\ &=\; m\,\frac{\sin\theta\,\mathbf{n}_\theta}{r} \times \frac{\mathbf{n}_\phi}{r\sin\theta} \\ &=\; m\,\frac{\mathbf{n}_r}{r^2}\,. \end{aligned} \tag{55}$$

The vector potentials (53) may be analytically continued to the entire 3D space (except the monopole point $r = 0$) itself, but such continuations are singular. The $\mathbf{A}_N(r,\theta,\phi)$ has a

so-called "Dirac string" of singularities along the negative z semi-axis ($\theta = \pi$), while the $\mathbf{A}_S(r, \theta, \phi)$ has a similar Dirac string of singularities along the positive z semi-axis ($\theta = 0$). To make a non-singular picture of the monopole field, Dirac used both vector potentials $\mathbf{A}_N$ and $\mathbf{A}_S$ but restricted each potential to the region of space where it is not singular. The two regions overlap, and in the overlap we may use either $\mathbf{A}_N$ or $\mathbf{A}_S$, whichever we like.

In QFT or even in quantum mechanics, a gauge transform of the vector potential should be accompanied by a phase transform of the charged fields or charged particles' wave functions. Consequently, for each charged species we must use different charged $\Psi^N(\mathbf{x})$ and $\Psi^S(\mathbf{x})$ in the Northern and Southern regions; in the overlap $\epsilon < \theta < \pi - \epsilon$, the two fields for the same species are related according to eq. (52). For the gauge transform (54) in question,

$$\Psi_a^N(r, \theta, \phi) \;=\; \Psi_a^S(r, \theta, \phi) \cdot \exp\left(2iqm\phi\right). \tag{56}$$

Both $\Psi^N(\mathbf{x})$ and $\Psi^S(\mathbf{x})$ are single-valued functions of $\mathbf{x}$ everywhere they are defined. In the overlap region both functions are defined and both are single valued, so the phase factor $\exp(2iqm\phi)$ in eq. (56) must be single valued. This single-valuedness requires integer $2q \times m$, hence the Dirac quantization condition (50).

Note: in the rationalized $\hbar = c = 1$ units, the magnetic charge of the monopole is $M = 4\pi m$, so the Dirac quantization condition reads

$$q \times M \;=\; 2\pi \times \text{an integer}. \tag{57}$$

In Gauss units, the magnetic charge is $M = m$ (without the 4π factor), but the phase in eq. (56) associated with the gauge transform (54) is $(2qm/\hbar c) \times \phi$, so the Dirac quantization condition becomes

$$q \times M \;=\; \frac{\hbar c}{2} \times \text{an integer}. \tag{58}$$

In Gauss units the magnetic charges and the electric charges have the same dimensionality. However, the quanta of the two charges are quite different: The electric charges of

all free particles are quantized in units of e; hence, according to eq. (58), all the magnetic charges should be quantized in units of

$$\frac{\hbar c}{2e} \approx \frac{137}{2} e. \tag{59}$$

Of course, as far as the Quantum ElectroDynamics is concerned, the monopoles do not have to exist at all. But if they do exist, their charges must be quantized in units of (59). Furthermore, if as much as one magnetic monopole exist anywhere in the universe then the electric charges of all free particles must be exactly quantized. Historically, Dirac discovered the magnetic monopole while trying to explain *the value* of the electric charge quantum e; instead, he found a reason for the charge quantization, but no explanation for $e^2 \approx \hbar c/137$, and he was quite disappointed.

Today, we have other explanations of the electric charge quantization; in particular the Grand Unification of strong, weak and electromagnetic interactions at extremely high energies produces quantized electrical charges. Curiously, the same Grand Unified Theories also predict that there **are** magnetic monopoles with charges (59). More recently, several attempts to unify all the fundamental interactions withing the context of the String Theory also gave rise to magnetic monopoles, with charges quantized in units of $N\hbar c/2e$, where N is an integer such as 3 or 5. It was later found that in the same theories, there were superheavy particles with fractional electric charges e/N, so the monopoles in fact had the smallest non-zero charges allowed by the Dirac condition (58)! Nowadays, most theoretical physicists believe that any fundamental theory that provides for exact quantization of the electric charge should also provide for the existence of magnetic monopoles, but this conjecture has not been proved (yet).

Suggested Reading: J. J. Sakurai, *Modern Quantum Mechanics*, §2.6.