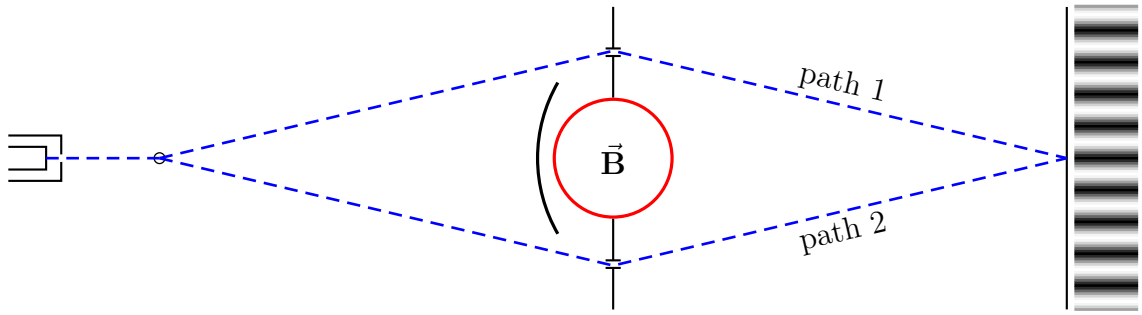
Aharonov–Bohm Effect, Magnetic Monopoles, and Charge Quantization

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 trajectory — the EM fields in some volume of space the particle never goes through do not affect it at all. But *in quantum mechanics, the 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

$$\oint_{\text{loop}} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r} \;=\; \iint_{\substack{\text{inside the loop} \\ \text{including the solenoid}}} \mathbf{B}(\mathbf{r}) \cdot d^2\mathbf{Area} \;=\; \Phi. \tag{1}$$

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

$$\mathbf{A}(\mathbf{r}) \;\to\; \mathbf{A}'(\mathbf{r}) \;=\; \mathbf{A}(\mathbf{r}) \;+\; \nabla\Lambda(\mathbf{r}) \;=\; 0 \quad \text{for some } \Lambda(\mathbf{r}), \tag{2}$$

but *globally* no single-valued $\Lambda(\mathbf{r})$ can gauge away the vector potential along both paths around the solenoid. Indeed,

$$\Delta\Lambda\big|_{\text{path 1}} \;=\; -\int_{\text{path 1}} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r}\,, \qquad \Delta\Lambda\big|_{\text{path 2}} \;=\; -\int_{\text{path 2}} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r}\,, \tag{3}$$

$$\begin{aligned} \Delta\Lambda\big|_{\text{path 1}} \;-\; \Delta\Lambda\big|_{\text{path 2}} \;&=\; -\int_{\text{path 1}} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r} \;+\; \int_{\text{path 2}} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r} \\ &=\; -\oint \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r} \;=\; -\Phi \;\neq\; 0. \end{aligned} \tag{4}$$

Instead, we have two separate gauge transforms — the $\Lambda_1(\mathbf{r})$ that gauges away $\mathbf{A}(\mathbf{r})$ along the path #1, and the $\Lambda_2(\mathbf{r})$ that gauges away $\mathbf{A}(\mathbf{r})$ along the path #2 — but they are different transforms, $\Lambda_1 \neq \Lambda_2$.

Let's relate these vector potentials and gauge transforms to the propagation amplitudes $U(\beta \leftarrow \alpha)$ from one point α to another point β — for example, from the electron gun to a point on the screen. By definition, the propagation amplitude during flight time t is

$$U(\beta \leftarrow \alpha) \;\stackrel{\text{def}}{=}\; \big\langle \mathbf{r}_\beta \big| \exp\!\big(-it\hat{H}/\hbar\big) \big| \mathbf{r}_\alpha \big\rangle \quad\Longrightarrow\quad \Psi(\mathbf{r}_\beta, t) \;=\; \iiint U(\beta \leftarrow \alpha)\Psi(\mathbf{r}_\alpha, t_0 = 0)\, d^3\mathbf{r}_\alpha\,. \tag{5}$$

For a charged particle, the wave function's phase depends on the gauge. Specifically, a gauge transform of the vector potential must be accompanied by a position-dependent phase

change of the wave function

$$\mathbf{A}'(\mathbf{r}) \;=\; \mathbf{A}(\mathbf{r}) \;+\; \nabla\Lambda(\mathbf{r}) \quad \text{while} \quad \Psi'(\mathbf{r}) \;=\; \Psi(\mathbf{r}) \times \exp\left(i\tfrac{q}{\hbar}\Lambda(\mathbf{r})\right) \quad \text{for the same } \Lambda(\mathbf{r}), \tag{6}$$

please see my previous set of notes for the explanation and details. To make this phase transform consistent with eq. (5) for the propagation amplitudes, such amplitudes should also change their phases according to

$$U'(\beta \leftarrow \alpha) \;=\; \exp\left(+i\tfrac{q}{\hbar}\Lambda(\beta)\right) \times U(\beta \leftarrow \alpha) \times \exp\left(-i\tfrac{q}{\hbar}\Lambda(\alpha)\right), \tag{7}$$

then we would have

$$\Psi(\mathbf{r}_\beta, t) \;=\; \iiint U(\beta \leftarrow \alpha)\Psi(\mathbf{r}_\alpha, 0)\, d^3\mathbf{r}_\alpha \quad \Longrightarrow \quad \Psi'(\mathbf{r}_\beta, t) \;=\; \iiint U'(\beta \leftarrow \alpha)\Psi'(\mathbf{r}_\alpha, 0)\, d^3\mathbf{r}_\alpha\,. \tag{8}$$

Now suppose along the electron's path from α to β there is no magnetic field but the vector potential happens to be non-zero, $\mathbf{B} = 0$ but $\mathbf{A} \neq 0$. Then *locally* the vector potential is gauge-equivalent to zero, meaning there exist some $\Lambda(\mathbf{r})$ such that

$$\mathbf{A}_0(\mathbf{r}) \;=\; \mathbf{A}(\mathbf{r}) \;+\; \nabla\Lambda(\mathbf{r}) \;=\; 0, \tag{9}$$

if not everywhere then at least throughout the neighborhood of the electron's path. Then comparing the propagation amplitude $U_{\mathbf{A}}(\beta \leftarrow \alpha)$ in presence of the vector potential with the similar amplitude $U_0(\beta \leftarrow \alpha)$ for $\mathbf{A}_0 \equiv 0$, we find

$$\begin{aligned} U_0(\beta \leftarrow \alpha) \;&=\; U_{\mathbf{A}}(\beta \leftarrow \alpha) \times \exp\left(\frac{iq}{\hbar}\left(\Lambda(\beta) - \Lambda(\alpha)\right)\right) \\ &=\; U_{\mathbf{A}}(\beta \leftarrow \alpha) \times \exp\left(\frac{iq}{\hbar}\int_\alpha^\beta \nabla\Lambda \cdot d\mathbf{r}\right) \\ &=\; U_{\mathbf{A}}(\beta \leftarrow \alpha) \times \exp\left(-\frac{iq}{\hbar}\int_\alpha^\beta \mathbf{A} \cdot d\mathbf{r}\right), \end{aligned} \tag{10}$$

and therefore

$$U_{\mathbf{A}}(\beta \leftarrow \alpha) \;=\; U_0(\beta \leftarrow \alpha) \times \exp\left(+\frac{iq}{\hbar} \int\limits_{\alpha}^{\beta} \mathbf{A} \cdot d\mathbf{r} \right). \tag{11}$$

This tells us that even when the vector potential $\mathbf{A}$ does not lead to a magnetic field in the region the electron travels through, it still has a non-trivial effect on the propagation amplitude, namely it changes its phase.

Note: if the $\mathbf{B}$ field vanishes along the electron's path but does not vanish somewhere else, then we can make the gauge-transformed potential $\mathbf{A}' = \mathbf{A} + \nabla\Lambda$ vanish along the path, but it would not vanish somewhere else. Consequently, the relation

$$\Lambda(\beta) \;-\; \Lambda(\alpha) \;=\; \int\limits_{\alpha}^{\beta} \nabla\Lambda \cdot d\mathbf{r} \;=\; -\int\limits_{\alpha}^{\beta} \mathbf{A} \cdot d\mathbf{r}$$

works only if we integrate $\mathbf{A} \cdot d\mathbf{r}$ along the electron path rather than some other line. In the context of eq. (11), this means that

$$U_{\mathbf{A}}(\beta \leftarrow \alpha) \;=\; U_0(\beta \leftarrow \alpha) \times \left(\frac{iq}{\hbar} \int\limits_{\text{electron's path}} \mathbf{A} \cdot d\mathbf{r} \right). \tag{12}$$

In the Aharonov–Bohm experiment, the electron can take two different paths from the same point α (the electron gun) to the same point β on the screen. The interference pattern on the screen follows from the net amplitude

$$U^{\text{net}}(\beta \leftarrow \alpha) \;=\; U^{\text{path 1}}(\beta \leftarrow \alpha) \;+\; U^{\text{path 2}}(\beta \leftarrow \alpha), \tag{13}$$

which depends on the phase difference between the amplitudes for each path,

$$\Delta\varphi(\beta) \;=\; \text{phase}\big(U^{\text{path 1}}(\beta \leftarrow \alpha)\big) \;-\; \text{phase}\big(U^{\text{path 2}}(\beta \leftarrow \alpha)\big). \tag{14}$$

Note that along both paths $\mathbf{B} = 0$ but $\mathbf{A} \neq 0$, which affects the phases of the each amplitude

according to eq. (12), specifically

$$\begin{aligned}
\text{phase}\left(U_{\mathbf{A}}^{\text{path 1}}(\beta \leftarrow \alpha)\right) &= \text{phase}\left(U_{0}^{\text{path 1}}(\beta \leftarrow \alpha)\right) + \frac{q}{\hbar}\int\limits_{\text{path 1}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r}, \\
\text{phase}\left(U_{\mathbf{A}}^{\text{path 2}}(\beta \leftarrow \alpha)\right) &= \text{phase}\left(U_{0}^{\text{path 2}}(\beta \leftarrow \alpha)\right) + \frac{q}{\hbar}\int\limits_{\text{path 2}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r}.
\end{aligned} \tag{15}$$

Consequently, the phase difference (14) is affected by the vector potential according to

$$\Delta\varphi_{\mathbf{A}} = \Delta\varphi_0 + \frac{q}{\hbar}\int\limits_{\text{path 1}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r} - \frac{q}{\hbar}\int\limits_{\text{path 2}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r} \tag{16}$$

where the difference between the two integrals $\int \mathbf{A}\cdot d\mathbf{r}$ over the two path is the magnetic flux through 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,

$$\int\limits_{\text{path 1}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r} - \int\limits_{\text{path 2}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r} = \oint\limits_{\text{closed loop}} \mathbf{A}(\mathbf{r})\cdot d\mathbf{r} = \Phi[\text{through the loop}], \tag{17}$$

which is basically the magnetic flux through the solenoid since outside the solenoid $\mathbf{B} = 0$. Thus

$$\Delta_{\mathbf{A}}\varphi(\beta) = \Delta_0\varphi(\beta) + \frac{q}{\hbar}\times\Phi\,, \tag{18}$$

which means that 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!

In the mathematical language, the Aharonov–Bohm effect feels the *cohomology* of the vector potential $\mathbf{A}(\mathbf{r})$. In a topologically trivial space — like the flat 3D space without any holes — specifying $\mathbf{A}(\mathbf{r})$ *modulo* gauge transforms $\mathbf{A}(\mathbf{r}) \to \mathbf{A}(\mathbf{r}) + \nabla\Lambda(\mathbf{r})$ is equivalent to specifying the magnetic field $\mathbf{B}(\mathbf{r}) = \nabla\times\mathbf{A}$. However, in spaces with holes the vector potential modulo $\nabla\Lambda(\mathbf{r})$ for *single-valued* $\Lambda(\mathbf{r})$ contains more information than the magnetic

field: In addition to $\mathbf{B}(\mathbf{r})$ for $\mathbf{r}$ outside the holes, the vector potential also knows the magnetic fluxes through the holes! Indeed, the integrals along closed loops

$$\oint_{\text{loop}} \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r} \;=\; \Phi(\text{loop}) \tag{19}$$

are gauge-invariant *for single-valued* $\Lambda(\mathbf{r})$, and when $\nabla \times \mathbf{A} \equiv 0$ everywhere outside the holes, then the fluxes (19) 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 $\mathbf{A}(\mathbf{r})$.

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 (18) 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

$$\Phi \;=\; \frac{2\pi\hbar}{q} \times \text{an integer}, \tag{20}$$

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

$$\Phi_1 \;=\; \frac{2\pi\hbar}{|q|} \tag{21}$$

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 measures only the fractional part of the flux in units of

$$\Phi_0 \;=\; \frac{2\pi\hbar}{2e} \;=\; 2.067\,833\,667(52)\times 10^{-15}\ \text{Wb} \quad (\text{Webers or Tesla}\times \text{m}^2), \tag{22}$$

or in Gauss units

$$\Phi_0 \;=\; \frac{2\pi\hbar c}{2e} \;=\; 2.067\,833\,667(52)\times 10^{-7}\ \text{Mx} \quad (\text{Maxwells or Gauss}\times \text{cm}^2). \tag{23}$$

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, the measuring the fractional part of the same flux Φ in two different units Φ_1 and Φ_2 with irrational Φ_1/Φ_2, we would be able to reconstruct the whole flux Φ and not just its fractional part. However, in reality all the 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

$$\Phi_u \;=\; \frac{2\pi\hbar}{e} \;=\; 2\Phi_0\,. \tag{24}$$

This universality is crucial to the very existence of magnetic monopoles, as we shall see in a moment.

Magnetic Monopoles and Charge Quantization

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; so long that the other pole is very far away. Outside the magnet itself, the magnetic field surrounding the pole in question is spherically symmetric

$$\mathbf{B}(r,\theta,\phi) \;=\; \frac{\mu_0 M}{4\pi r^2}\,\hat{\mathbf{r}}\,, \tag{25}$$

while inside the magnet there is magnetic flux $\mu_0 M$ 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 (25), so for all intents and purposes we have a magnetic monopole of magnetic charge M. In quantum mechanics however, one can also detect the Aharonov-Bohm effect due to the magnetic flux $\mu_0 M$ inside the magnet, and that would make the magnet itself detectable along its whole length. Moreover, in quantum field theory, the Aharonov–Bohm effect would disturb the free-wave modes of the charged fields — instead of the plane waves we would get eigenwaves of some place-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 $\mu_0 M$ is an integral multiplet of $2\pi\hbar/q$. 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,*

$$M \times q \;=\; \frac{2\pi\hbar}{\mu_0} \times \text{an integer} \tag{26}$$

in MKSA units; in the Gauss units this condition becomes

$$M \times q \;=\; \frac{\hbar c}{2} \times \text{an integer.} \tag{27}$$

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

A more rigorous argument was made by Paul A. M. Dirac himself years before the discovery of the Aharonov-Bohm effect. Instead of using a single vector potential $\mathbf{A}(\mathbf{r})$ throughout the whole space surrounding the monopole, Dirac divided the space into two overlapping regions and used a different potential in each region. However, the the two potentials are related by a gauge transform and thus are physically equivalent to each other.*

Specifically, in the spherical coordinates (r, θ, ϕ), the Northern region (N) span latitudes $0 \le \theta\pi - \epsilon$ — everything except a small neighborhood of the South pole, — while the Southern region (S) spans $\epsilon < \theta \le \pi$ — everything except a neighborhood of the North pole. The two regions overlap in a broad band around the equator. The vector potentials for the two regions are respectively:

$$\begin{aligned} \mathbf{A}_N(r,\theta,\phi) &= \frac{\mu_0 M}{4\pi}\,\frac{+1-\cos\theta}{r\sin\theta}\,\hat{\phi} \;=\; \frac{\mu_0 M}{4\pi}\,(+1-\cos\theta)\,(\nabla\phi), \\ \mathbf{A}_S(r,\theta,\phi) &= \frac{\mu_0 M}{4\pi}\,\frac{-1-\cos\theta}{r\sin\theta}\,\hat{\phi} \;=\; \frac{\mu_0 M}{4\pi}\,(-1-\cos\theta)\,(\nabla\phi), \end{aligned} \tag{28}$$

The two potentials are gauge-equivalent:

$$\mathbf{A}_N \;-\; \mathbf{A}_S \;=\; \frac{\mu_0 M}{4\pi}\,\frac{2}{r\sin\theta}\,\hat{\phi} \;=\; \frac{\mu_0 M}{2\pi}\,(\nabla\phi) \;=\; \nabla\left(\frac{\mu_0 M}{2\pi}\,\phi\right) \tag{29}$$

so they lead to the same magnetic field, namely (25). Indeed,

$$\begin{aligned} \nabla\times\mathbf{A}_{\mathrm{N\,or\,S}} &= \nabla\times\left(\frac{\mu_0 M}{4\pi}\,(\pm 1-\cos\theta)\nabla\phi\right) \\ &= \frac{\mu_0 M}{4\pi}\,(\nabla(\pm 1-\cos\theta))\times\nabla\phi \\ &= \frac{\mu_0 M}{4\pi}\,\frac{\sin\theta\,\hat{\theta}}{r}\times\frac{\hat{\phi}}{r\sin\theta} \\ &= \frac{\mu_0 M}{4\pi}\,\frac{\hat{\mathbf{r}}}{r^2}\,. \end{aligned} \tag{30}$$

The vector potentials (28) 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

$\star$ From the mathematical point of view, the Dirac monopole is a *gauge bundle*, a construction that generalizes multiple coordinate patches in Riemannian geometry. But Dirac himself did not use the bundle language, and you do not need it to understand my notes.

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 quantum mechanics of a charged particle, a gauge transform of the vector potential should be accompanied by a phase transform of the wave function according to

$$\mathbf{A}'(\mathbf{r}) \;=\; \mathbf{A}(\mathbf{r}) \;+\; \nabla\Lambda(\mathbf{r}) \quad \text{while} \quad \Psi'(\mathbf{r}) \;=\; \Psi(\mathbf{r}) \times \exp\left(i\tfrac{q}{\hbar}\Lambda(\mathbf{r})\right) \quad \text{for the same } \Lambda(\mathbf{r}). \tag{6}$$

Consequently, the two different gauge-equivalent vector potentials $\mathbf{A}_N(\mathbf{r})$ and $\mathbf{A}_S(\mathbf{r})$ should come with two different wave functions $\Psi_N(\mathbf{r})$ and $\Psi_S(\mathbf{r})$ in the corresponding regions of space, and in the overlap between the two regions, the $\Psi_N(\mathbf{r})$ and the $\Psi_S(\mathbf{r})$ should be related by the appropriate phase transform (6). Specifically,

$$\Lambda(r,\theta,\phi) \;=\; \frac{\mu_0 M}{2\pi}\,\phi\,, \tag{31}$$

$$\mathbf{A}_N(r,\theta,\phi) \;=\; \mathbf{A}_S(r,\theta,\phi) \;+\; \nabla\Lambda(r,\theta,\phi), \tag{32}$$

$$\Psi_N(r,\theta,\phi) \;=\; \Psi_S(r,\theta,\phi) \times \exp\left(i\,\frac{q}{\hbar}\,\Lambda(r,\theta,\phi)\right). \tag{33}$$

Note: the gauge-transform parameter Λ in eq. (31) is multi-valued since the longitude coordinate ϕ changes by 2π as we go around the equator. However, *multi-valued $\Lambda(\mathbf{r})$ are OK as long as both the EM potentials and the wave functions it relates are single-valued,* which means that

$$\text{both} \quad \nabla\Lambda \quad \text{and} \quad \exp\left(i\,\frac{q}{\hbar}\,\Lambda\right) \quad \text{must be single-valued.} \tag{34}$$

For the case at hand, $\nabla\phi$ and hence $\nabla\Lambda$ are single valued, but we need to check the phase

$$\exp\left(i\,\frac{q}{\hbar}\,\Lambda\right) \;=\; \exp\left(i\,\frac{qM\mu_0}{2\pi\hbar}\,\phi\right). \tag{35}$$

In general, the exponential $\exp(i\nu\phi)$ for a constant ν is a single-valued function of the angle ϕ if and only if ν is an integer, so the gauge transform (31)–(33) is allowed in quantum

mechanics if and only if

$$\frac{qM\mu_0}{2\pi\hbar} \text{ is an integer.} \tag{36}$$

Physically, this means that *a Dirac monopole of magnetic charge M may coexist with a quantum particle of electric charge q only when the charges obey the Dirac quantization condition*

$$M \times q \;=\; \frac{2\pi\hbar}{\mu_0} \times \text{an integer}. \tag{26}$$

In quantum field theory, for every existing *species* of a charged particle there are countless virtual particles of that species everywhere. Therefore, *if as much as a single magnetic monopole exist anywhere in the Universe, then the electric charges of all particle species must be quantized*,

$$q \;=\; \frac{2\pi\hbar}{\mu_0 M} \times \text{an integer}. \tag{37}$$

Historically, Dirac discovered the magnetic monopole while trying to explain the rather small *value* of the electric charge quantum e — in Gauss units,

$$e^2 \;\approx\; \frac{\hbar c}{137}\,. \tag{38}$$

The monopole gives us an excellent reason for the charge quantization in the first place, but alas it does not explain the value (38) of the quantum, and Dirac was quite disappointed.

BTW, in Gauss units, the electric and the magnetic charges have the same dimensionality. But in light of eqs. (27) and (38), they are quantized in rather different units, e for the electric charges and

$$\frac{\hbar c}{2e} \;\approx\; \frac{137}{2}\,e \tag{39}$$

for the magnetic charges. 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 (39). Also, the very existence of a single monopole would explain the electric charge quantization.

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 (39). More recently, several attempts to unify all the fundamental interactions within 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 (27)! 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.