Quantum Mechanics of a Charged Particle

Consider a non-relativistic charged particle in electromagnetic fields $\mathbf{E}(\mathbf{x},t)$ and $\mathbf{B}(\mathbf{x},t)$. In classical mechanics, its motion is governed by the Lagrangian

$$L(\mathbf{x},\mathbf{v},t) \;=\; \frac{m\mathbf{v}^2}{2} \;-\; q\Phi(\mathbf{x},t) \;+\; \frac{q}{c}\,\mathbf{v}\cdot\mathbf{A}(\mathbf{x},t) \tag{1}$$

where q is the particle's electric charge and $\Phi(\mathbf{x},t)$ and $\mathbf{A}(\mathbf{x},t)$ are the scalar and the vector potentials of the EM fields evaluated at the particle's location $\mathbf{x}(t)$. In the Hamiltonian formalism, the particle's canonical momentum $\mathbf{p}$ is different from the usual kinematical momentum $\boldsymbol{\pi} = m\mathbf{v}$ — instead,

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

— while the Hamiltonian function is

$$H(\mathbf{x},\mathbf{p}) \;=\; \frac{\boldsymbol{\pi}^2}{2m} \;+\; q\Phi(\mathbf{x}) \;=\; \frac{1}{2m}\left(\mathbf{p} - \frac{q}{c}\mathbf{A}(\mathbf{x})\right)^2 \;+\; q\Phi(\mathbf{x}), \tag{3}$$

In quantum mechanics — *cf.* my notes on canonical quantization, — it's the canonical momentum operators $\hat{p}_i$ which obey the canonical commutation relations

$$[\hat{x}_i,\hat{x}_j] \;=\; 0, \quad [\hat{x}_i,\hat{p}_j] \;=\; i\hbar\delta_{ij}, \quad [\hat{p}_i,\hat{p}_j] \;=\; 0 \tag{4}$$

(at equal times or in the Schrödinger picture), while the kinematical momentum operators

$$\hat{\pi}_i \;\stackrel{\text{def}}{=}\; \hat{p}_i \;-\; \frac{q}{c}\hat{A}_i(\hat{\mathbf{x}}) \tag{5}$$

obey

$$[\hat{x}_i,\hat{\pi}_j] \;=\; i\hbar\delta_{ij} \quad \text{but} \quad [\hat{\pi}_i,\hat{\pi}_j] \;=\; \frac{i\hbar q}{c}\,\epsilon_{ijk}B_k(\hat{\mathbf{x}}). \tag{6}$$

The Hamiltonian operators is

$$\hat{H} \;=\; \frac{\hat{\boldsymbol{\pi}}^2}{2m} \;+\; q\Phi(\hat{\mathbf{x}}) \;=\; \frac{1}{2m}\left(\hat{\mathbf{p}} - \frac{q}{c}\mathbf{A}(\hat{\mathbf{x}})\right)^2 \;+\; q\Phi(\hat{\mathbf{x}}), \tag{7}$$

which leads to the Ehrenfest equations

$$\frac{d}{dt}\langle\hat{\mathbf{x}}\rangle \;=\; \frac{1}{m}\langle\hat{\boldsymbol{\pi}}\rangle\,, \qquad \frac{d}{dt}\langle\hat{\boldsymbol{\pi}}\rangle \;=\; q\,\langle\mathbf{E}(\hat{\mathbf{x}},t)\rangle \;+\; \frac{q}{2mc}\,\langle\hat{\boldsymbol{\pi}}\times\mathbf{B}(\hat{\mathbf{x}},t) \;-\; \mathbf{B}(\hat{\mathbf{x}},t)\times\boldsymbol{\pi}\rangle\,, \tag{8}$$

cf. homework set#4 (problem 2).

In the coordinate basis, it's the canonical momentum operators which act as

$$\hat{p}_i\psi(\mathbf{x},t)) \;=\; -i\hbar\frac{\partial}{\partial x_i}\psi(\mathbf{x},t) \tag{9}$$

— hence the commutation relations (4), — while the kinematical momentum operators act in a more complicated fashion as

$$\hat{\pi}_i\psi(\mathbf{x},t) \;=\; -i\hbar\frac{\partial}{\partial x_i}\psi(\mathbf{x},t) \;-\; \frac{q}{c}A_i(\mathbf{x},t)\psi(\mathbf{x},t) = \; -i\hbar\mathcal{D}_i\psi(\mathbf{x},t) \tag{10}$$

$$\text{for}\quad \vec{\mathcal{D}} \;\stackrel{\text{def}}{=}\; \nabla \;-\; \frac{iq}{\hbar c}\mathbf{A}(\mathbf{x},t), \tag{11}$$

and the Hamiltonian operator acts as

$$\hat{H}\psi(\mathbf{x},t) \;=\; -\frac{\hbar^2}{2m}\,\vec{\mathcal{D}}\cdot\vec{\mathcal{D}}\psi(\mathbf{x},t) \;+\; q\Phi(\mathbf{x},t)\psi(\mathbf{x},t). \tag{12}$$

The differential operators (11) are called the *covariant derivatives* because of the way they behave under gauge transforms of the EM potentials Φ and $\mathbf{A}$, as we shall see in a few pages.

Gauge Transforms

The electromagnetic fields $\mathbf{E}$ and $\mathbf{B}$ are invariant under *gauge transforms* of the potentials Φ and $\mathbf{A}$: Take an arbitrary function $\Lambda(\mathbf{x},t)$ and let

$$\mathbf{A}'(\mathbf{x},t) \;=\; \mathbf{A}(\mathbf{x},t) \;+\; \nabla\Lambda(\mathbf{x},t), \qquad \Phi'(\mathbf{x},t) \;=\; \Phi(\mathbf{x},t) \;-\; \frac{1}{c}\,\frac{\partial\Lambda(\mathbf{x},t)}{\partial t}, \tag{13}$$

then

$$\mathbf{E} \;=\; -\frac{1}{c}\,\frac{\partial\mathbf{A}}{\partial t} \;-\; \nabla\Phi, \qquad \mathbf{B} \;=\; \nabla\times\mathbf{A} \tag{14}$$

remain invariant,

$$\begin{aligned}
\mathbf{E}' &= -\frac{1}{c}\frac{\partial \mathbf{A}'}{\partial t} - \nabla\Phi' = -\frac{1}{c}\left(\frac{\partial \mathbf{A}}{\partial t} + \cancel{\frac{\partial}{\partial t}\nabla\Lambda}\right) - \left(\nabla\Phi - \cancel{\frac{1}{c}\nabla\frac{\partial}{\partial t}\Lambda}\right) \\
&= -\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} - \nabla\Phi = \mathbf{E}, \qquad (15) \\
\mathbf{B}' &= \nabla\times\mathbf{A}' = \nabla\times\mathbf{A} + \nabla\times\nabla\Lambda = \nabla\times\mathbf{A} + 0 = \mathbf{B}.
\end{aligned}$$

For a classical charged particle, the Lagrangian

$$L(\mathbf{x},\mathbf{v},t) = \frac{m\mathbf{v}^2}{2} - q\Phi(\mathbf{x},t) + \frac{q}{c}\,\mathbf{v}\cdot\mathbf{A}(\mathbf{x},t) \qquad (1)$$

is not *gauge invariant* — *i.e.*, is not invariant under the gauge transform (13), — but it changes by a total time derivative,

$$\begin{aligned}
\Delta L &= -q\Delta\Phi(\mathbf{x}(t),t) + \frac{q}{c}\,\mathbf{v}(t)\cdot\Delta\mathbf{A}(\mathbf{x}(t),t) \\
&= \frac{q}{c}\frac{\partial\Lambda}{\partial t}@(\mathbf{x}(t),t) + \frac{q}{c}\frac{d\mathbf{x}}{dt}\cdot\nabla\Lambda@(\mathbf{x},t) \qquad (16) \\
&= \frac{q}{c}\frac{d}{dt}\Lambda(\mathbf{x}(t),t).
\end{aligned}$$

Consequently, the action $S = \int L\,dt$ is not quite invariant, but

$$\Delta S = \int_{t_1}^{t_2}\Delta L\,dt = \frac{q}{c}\int_{t_1}^{t_2} d\Lambda(\mathbf{x}(t),t) = \frac{q}{c}\Big[\Lambda(\mathbf{x}_2,t_2) - \Lambda(\mathbf{x}_1,t_1)\Big] \qquad (17)$$

does not depend on the path from the initial point $\mathbf{x}(t_1) = \mathbf{x}_1$ to $\mathbf{x}(t_2) = \mathbf{x}_2$ but only on the initial and the final points of that path. Therefore, the path $\mathbf{x}(t)$ which used to minimize the action S before the gauge transform, continues to minimize the transformed action $S + \Delta S$. By the least action principle, this means that the classical path of the charged particle in given EM fields $\mathbf{E}(\mathbf{x},t)$ and $\mathbf{B}(\mathbf{x},t)$ is gauge-invariant, and so is the Euler–Lagrange equation

of motion for the classical path, namely

$$m\mathbf{a} \;=\; q\mathbf{E}(\mathbf{x},t) \;+\; \frac{q}{c}\,\mathbf{v}\times\mathbf{B}(\mathbf{x},t). \tag{18}$$

Now consider the quantum charged particle. In the WKB approximation,

$$\text{phase}\big(\Psi(\mathbf{x}_2,t_2)\big) \;-\; \text{phase}\big(\Psi(\mathbf{x}_1,t_1)\big) \;=\; \frac{1}{\hbar}\int_{t_1}^{t_2} L\,dt \quad \begin{array}{l}\text{along the classical path}\\ \text{from } (\mathbf{x}_1,t_1) \text{ to } (\mathbf{x}_2,t_2)\end{array} \tag{19}$$

cf. eq. (167) on page 32 of my notes on the WKB approximation. The RHS of eq. (19) is affected by a gauge transform according to eq. (17), thus

$$\text{phase}\big(\Psi(\mathbf{x}_2,t_2)\big) \;-\; \text{phase}\big(\Psi(\mathbf{x}_1,t_1)\big) \quad \text{changes by} \quad \frac{q}{\hbar c}\Big[\Lambda(\mathbf{x}_2,t_2) \;-\; \Lambda(\mathbf{x}_1,t_1)\Big]. \tag{20}$$

Consequently, *a gauge transform of EM potentials must be accompanied by a local phase transform of the wave function of a charged particle* according to

$$\Psi'(\mathbf{x},t) \;=\; \Psi(\mathbf{x},t)\times\exp\left(i\frac{q}{\hbar c}\,\Lambda(\mathbf{x},t)\right). \tag{21}$$

Although I've used the WKB approximation to derive this phase transform, it is actually exact. To see why, consider the kinematic momentum operator $\hat{\boldsymbol{\pi}} = -i\hbar\vec{\mathcal{D}}$ and its matrix elements $\langle\Psi_1|\,\hat{\boldsymbol{\pi}}\,|\Psi_2\rangle$. By the Ehrenfest equation

$$m\frac{d}{dt}\,\langle\hat{\mathbf{x}}\rangle \;=\; \langle\hat{\boldsymbol{\pi}}\rangle\,, \tag{22}$$

the expectation value of the $\hat{\boldsymbol{\pi}}$ must be gauge invariant in any quantum state $|\Psi\rangle$, so the matrix elements

$$\langle\Psi_1|\,\hat{\boldsymbol{\pi}}\,|\Psi_2\rangle \;=\; -i\hbar\int d^3\mathbf{x}\;\Psi_2^*\vec{\mathcal{D}}\Psi_1 \tag{23}$$

must be gauge invariant, although the differential operator $\vec{\mathcal{D}}$ itself is not invariant:

$$\vec{\mathcal{D}}' \;=\; \nabla \;-\; \frac{iq}{\hbar c}\mathbf{A}' \;=\; \nabla \;-\; \frac{iq}{\hbar c}\mathbf{A} \;-\; \frac{iq}{\hbar c}(\nabla\Lambda) \;=\; \vec{\mathcal{D}} \;-\; \frac{iq}{\hbar c}(\nabla\Lambda)\,. \tag{24}$$

However, we may compensate for the extra $\nabla\Lambda$ term here by changing the wave-function's phase exactly as in eq. (21): This makes $\vec{\mathcal{D}}'\Psi'(\mathbf{x},t)$ transform *covariantly*, *i.e.* exactly like

the $\Psi(\mathbf{x},t)$ itself:

$$\Psi' \;=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\Psi, \tag{21}$$

$$\begin{aligned}\nabla\Psi' \;&=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\nabla\Psi \;+\; \left(\nabla\exp\left(\frac{iq}{\hbar c}\Lambda\right) \;=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\frac{iq}{\hbar c}(\nabla\Lambda)\right)\Psi \\ &=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\left(\nabla\Psi \;+\; \frac{iq}{\hbar c}(\nabla\Lambda)\,\Psi\right),\end{aligned} \tag{25}$$

$$\begin{aligned}\vec{\mathcal{D}}'\Psi' \;&=\; \nabla\Psi' \;-\; \frac{iq}{\hbar c}\left(\mathbf{A}' \;=\; \mathbf{A}+\nabla\Lambda\right)\Psi' \\ &=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\left(\nabla\Psi \;+\; \cancel{\frac{iq}{\hbar c}(\nabla\Lambda)\,\Psi}\right) \;-\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\left(\frac{iq}{\hbar c}\mathbf{A}\Psi \;+\; \cancel{\frac{iq}{\hbar c}(\nabla\Lambda)\Psi}\right) \\ &=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\left(\nabla\Psi \;-\; \frac{iq}{\hbar c}\mathbf{A}\Psi\right) \;=\; \exp\left(\frac{iq}{\hbar c}\Lambda\right)\vec{\mathcal{D}}\Psi,\end{aligned} \tag{26}$$

and that's why $\vec{\mathcal{D}}$ is called the *covariant* derivative. Thanks to this covariance, the products like

$$\Psi_1^*(\mathbf{x},t)\,\vec{\mathcal{D}}\Psi_2(\mathbf{x},t) \tag{27}$$

are gauge invariant; indeed

$$\begin{aligned}\Psi_2'(\mathbf{x},t) \;&=\; e^{i(q/\hbar c)\Lambda}\,\Psi_2(\mathbf{x},t), \\ \vec{\mathcal{D}}\Psi_2'(\mathbf{x},t) \;&=\; e^{i(q/\hbar c)\Lambda}\,\vec{\mathcal{D}}\Psi_2(\mathbf{x},t), \\ \Psi_1'(\mathbf{x},t) \;&=\; e^{i(q/\hbar c)\Lambda}\,\Psi_1(\mathbf{x},t),\end{aligned} \tag{28}$$

hence

$$\Psi_1^{*\prime}\,\vec{\mathcal{D}}'\Psi_2' \;=\; \left(e^{i(q/\hbar c)\Lambda}\,\Psi_1\right)^*\left(e^{i(q/\hbar c)\Lambda}\,\vec{\mathcal{D}}\Psi_2\right) \;=\; \Psi_1^*\,\vec{\mathcal{D}}\Psi_2, \tag{29}$$

and that's what makes the matrix elements (23) of the $\hat{\boldsymbol{\pi}}$ operator gauge invariant,

$$\left\langle\Psi_1'\right|\hat{\boldsymbol{\pi}}'\left|\Psi_2'\right\rangle \;=\; -i\hbar\int d^3\mathbf{x}\,\Psi_1^{*\prime}\,\vec{\mathcal{D}}'\Psi_2' \;=\; -i\hbar\int d^3\mathbf{x}\,\Psi_1^*\,\vec{\mathcal{D}}\Psi_2 \;=\; \left\langle\Psi_1\right|\hat{\boldsymbol{\pi}}\left|\Psi_2\right\rangle. \tag{30}$$

Moreover, the products of covariant derivatives are also covariant,

$$\mathcal{D}_i'\mathcal{D}_j'\Psi' \;=\; e^{i(q/\hbar c)\Lambda}\times\mathcal{D}_i\mathcal{D}_j\Psi, \quad \mathcal{D}_i'\mathcal{D}_j'\mathcal{D}_k\Psi' \;=\; e^{i(q/\hbar c)\Lambda}\times\mathcal{D}_i\mathcal{D}_j\mathcal{D}_k\Psi, \quad \ldots, \tag{31}$$

so the matrix elements of product operators like kinetic energy $\hat{\pi}^2/2m$ are gauge-invariant,

$$\left\langle \Psi_1' \right| \frac{\hat{\pi}'^2}{2m} \left| \Psi_2' \right\rangle \;=\; \left\langle \Psi_1 \right| \frac{\hat{\pi}^2}{2m} \left| \Psi_2 \right\rangle , \tag{32}$$

etc., *etc.*.

Next, the Hamiltonian operator

$$\hat{H} \;=\; \frac{\hat{\pi}^2}{2m} \;+\; q\Phi(\hat{\mathbf{x}},t) \tag{7}$$

is not gauge invariant, and even its matrix elements are not gauge invariant for time-dependent gauge transforms due to $\Delta\Phi \neq 0$. However, the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{x},t) \;=\; \hat{H}\Psi(\mathbf{x},t) \;=\; -\frac{\hbar^2}{2m}\vec{\mathcal{D}}^2\Psi(\mathbf{x},t) \;+\; q\Phi(\mathbf{x},t)\Psi(\mathbf{x},t) \tag{33}$$

becomes covariant when we re-write it as

$$i\hbar\mathcal{D}_t\Psi(\mathbf{x},t) \;=\; -\frac{\hbar^2}{2m}\vec{\mathcal{D}}^2\Psi(\mathbf{x},t) \tag{34}$$

where

$$\mathcal{D}_t \;=\; \frac{\partial}{\partial t} \;+\; \frac{iq}{\hbar}\Phi(\mathbf{x},t) \tag{35}$$

is the covariant time derivative operator. That is, under the gauge transforms $\mathcal{D}_t\Psi(\mathbf{x},t)$ transforms exactly like the wave-function itself,

$$\mathcal{D}_t'\Psi'(\mathbf{x},t) \;=\; e^{i(q/\hbar c)\Lambda(\mathbf{x},t)} \times \mathcal{D}_t\Psi(\mathbf{x},t). \tag{36}$$

Indeed,

$$\begin{aligned}
\mathcal{D}_t'\Psi' \;&=\; \frac{\partial}{\partial t}\left(e^{i(q/\hbar c)\Lambda} \times \Psi\right) \;+\; \frac{iq}{\hbar}\left(\Phi' \;=\; \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t}\right) \times e^{i(q/\hbar c)\Lambda} \times \Psi \\
&=\; e^{i(q/\hbar c)\Lambda}\left(\frac{\partial\Psi}{\partial t} + \cancel{\frac{iq}{c}\frac{\partial\Lambda}{\partial t} \times \Psi}\right) \;+\; e^{i(q/\hbar c)\Lambda}\left(\frac{iq}{\hbar}\,\Phi \times \Psi \;-\; \cancel{\frac{iq}{c}\frac{\partial\Lambda}{\partial t} \times \Psi}\right) \\
&=\; e^{i(q/\hbar c)\Lambda}\left(\frac{\partial\Psi}{\partial t} \;+\; \frac{iq}{\hbar}\,\Phi \times \Psi\right) \;=\; e^{i(q/\hbar c)\Lambda} \times \mathcal{D}_t\Psi.
\end{aligned} \tag{37}$$

Consequently, both sides of the covariant Schrödinger equation (34) transform pick up exactly

the same extra phase under a gauge transform, so if the equation holds true in one gauge then it holds true in any other gauge.

Finally, the Hamiltonian (7) and hence the covariant Schrödinger equation (34) apply to spinless relativistic particles with no interactions besides EM. For spinning particles like the electrons, we may have additional terms due to magnetic dipole moments interacting with the **B** field, thus

$$\hat{H} \;=\; \frac{\hat{\boldsymbol{\pi}}^2}{2m} \;+\; q\Phi(\hat{\mathbf{x}},t) \;-\; \frac{gq}{2mc}\,\hat{\mathbf{S}}\cdot\mathbf{B}(\hat{\mathbf{x}},t) \tag{38}$$

and hence

$$i\hbar\mathcal{D}_t\Psi(\mathbf{x},t) \;=\; -\frac{\hbar^2}{2m}\,\vec{\mathcal{D}}^2\Psi(\mathbf{x},t) \;-\; \frac{gq}{2mc}\mathbf{B}(\mathbf{x},t)\cdot\hat{\mathbf{S}}\Psi(\mathbf{x},t). \tag{39}$$

The g in these formulae is the *gyromagnetic factor* of the particle species in question. For point-like elementary particles, the non-relativistic limit of the relativistic Dirac equation gives $g = 2$, which is a good approximation for the electrons or the muons. Although there are small corrections due to interactions between the electrons (or muons) with virtual photons, the leading-order perturbation theory producing

$$g \;=\; 2 \;+\; \frac{e^2}{\pi\hbar c} \;+\; O\left(\left(\frac{e^2}{\pi\hbar c}\right)^2\right) \;\approx\; 2 \;+\; \frac{1}{137\pi}\,. \tag{40}$$

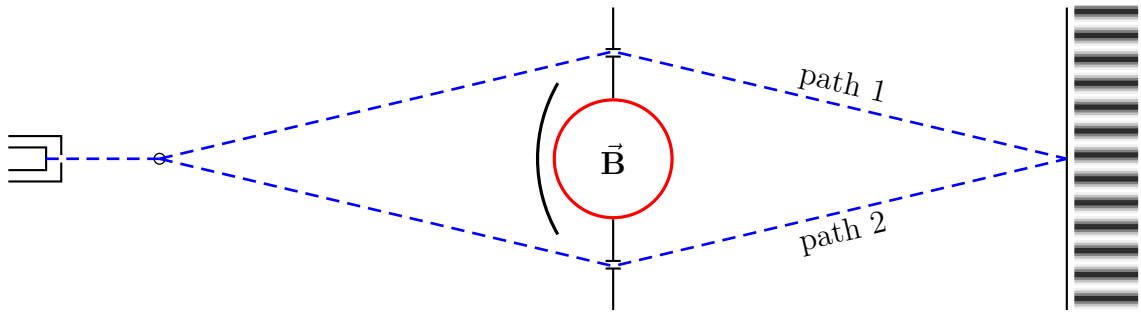
But the composite particles — like the nucleons made from quarks and gluons — may have rather different gyromagnetic factors; for example, the proton has $g \approx +5.58$ while the neutron has $g \approx -3.82$, or rather $qg \approx -3.82e$ despite $q = 0$.

Other non-minimal terms in the particle's Hamiltonian may include the electric dipole moment coupling $\hat{H} \supset -(d/S)\hat{\mathbf{S}}\cdot\mathbf{E}(\hat{\mathbf{x}})$, although fortunately the electrons, the protons, and the neutrons do not have electric dipoles. Also, there can be non-electromagnetic forces on the particle, such as nuclear forces on a proton or a neutron, but this gets us off the subject so let me stop here.

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 F 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

$$\underset{\text{loop}}{\oint} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} \;=\; \underset{\substack{\text{inside the loop}\\ \text{including the solenoid}}}{\iint} \mathbf{B}(\mathbf{x}) \cdot d^2\mathbf{Area} \;=\; F, \tag{41}$$

the magnetic flux through the solenoid. (Technically, F is the magnetic flux through the whole loop surrounding the solenoid, but since the $\mathbf{B}$ field outside the solenoid is negligible, the flux F comes from the solenoid itself.)

Locally, a curl-less vector potential is a gradient of some function, so it (the vector potential $\mathbf{A}(\mathbf{x})$) can be removed by a gauge transform,

$$\mathbf{A}(\mathbf{x}) \;\to\; \mathbf{A}'(\mathbf{x}) \;=\; \mathbf{A}(\mathbf{x}) \;+\; \nabla\Lambda(\mathbf{x}) \;=\; 0 \quad \text{for some } \Lambda(\mathbf{x}), \tag{42}$$

but *globally* no single-valued $\Lambda(\mathbf{x})$ can gauge away the vector potential along both paths around the solenoid. Instead, we have two separate gauge transforms — 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 — but they are different transforms, $\Lambda_1 \neq \Lambda_2$. To see how this works, let $\mathbf{x}_g$ be the electron gun's location while $\mathbf{x}_s$ is some point on the screen. *Along path #1* from $\mathbf{x}_g$ to $\mathbf{x}_s$, gauging away the vector potential calls for

$$d\Lambda_1(\mathbf{x}) \;=\; -\mathbf{A}(\mathbf{x}) \cdot d\mathbf{x}\,, \tag{43}$$

hence

$$\Lambda_1(\mathbf{x}_s) \;-\; \Lambda_1(\mathbf{x}_g) \;=\; \int\limits_{\text{path\#1}} d\Lambda_1 \;=\; -\int\limits_{\text{path\#1}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x}. \tag{44}$$

Likewise, *along path #2* from the same $\mathbf{x}_g$ to the same $\mathbf{x}_s$, gauging away the vector potential calls for

$$d\Lambda_2(\mathbf{x}) \;=\; -\mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} \tag{45}$$

and hence

$$\Lambda_2(\mathbf{x}_s) \;-\; \Lambda_2(\mathbf{x}_g) \;=\; \int\limits_{\text{path\#2}} d\Lambda_2 \;=\; -\int\limits_{\text{path\#2}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x}. \tag{46}$$

However, the integrals in eq. (44) and (46) are not equal to each other; instead

$$\int\limits_{\text{path\#1}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} - \int\limits_{\text{path\#2}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} \;=\; \oint\limits_{\mathcal{L}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} \tag{47}$$

where $\mathcal{L}$ is the closed loop made from path#1 from the electron gun $\mathbf{x}_g$ to the point $\mathbf{x}_s$ on the screen and then path#2 in reverse from $\mathbf{A}(\mathbf{x}_s)$ back to the electron gun $\mathbf{x}_g$. By the

Stokes theorem, the loop integral (47) is the magnetic flux through the loop $\mathcal{L}$, and since $\mathcal{L}$ surrounds the solenoid

$$\begin{aligned} \int\limits_{\text{path\#1}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} - \int\limits_{\text{path\#2}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \;&=\; \oint\limits_{\mathcal{L}} \mathbf{A}(\mathbf{x})\cdot d\mathbf{x} \\ &=\; F[\text{through } \mathcal{L}] \\ &=\; F[\text{through the solenoid}]. \end{aligned} \tag{48}$$

Consequently,

$$(\Lambda_1(\mathbf{x}_s) - \Lambda_1(\mathbf{x}_g)) \;-\; (\Lambda_2(\mathbf{x}_s) - \Lambda_2(\mathbf{x}_g)) \;=\; -F \;\neq\; 0, \tag{49}$$

which means we cannot possibly have the same single-valued $\Lambda_1(\mathbf{x}) \equiv \Lambda_2(\mathbf{x})$ gauge parameter for both paths.

The other key to the Aharonov–Bohm effect is the the local phase transform of the charged particle's wave function which must accompany the gauge transform of the vector potential,

$$\left.\begin{aligned} \Psi'(\mathbf{x}) \;&=\; \exp\left(\frac{iq}{\hbar c}\Lambda(\mathbf{x})\right)\,\Psi(\mathbf{x}) \\ \mathbf{A}'(\mathbf{x}) \;&=\; \mathbf{A}(\mathbf{x}) \;-\; \nabla\Lambda(\mathbf{x}) \end{aligned}\right\} \quad \text{for the same } \Lambda(\mathbf{x}). \tag{50}$$

Let's translate this local phase transform of the wave function to the language of the propagation amplitude (AKA the evolution kernel) $U(\mathbf{x}_2, \mathbf{x}_1)$ from one point $\mathbf{x}_1$ to another point $\mathbf{x}_2$. For example from the electron gun $\mathbf{x}_1 = \mathbf{x}_g$ to some particular point $\mathbf{x}_2 = \mathbf{x}_s$ on the screen. By definition, the propagation amplitude during flight time t is

$$U(\mathbf{x}_2, \mathbf{x}_1) \;\stackrel{\text{def}}{=}\; \langle \mathbf{x}_2 |\, \exp\bigl(-it\hat{H}/\hbar\bigr)\, |\mathbf{x}_1\rangle \;, \tag{51}$$

$$\Downarrow$$

$$\Psi(\mathbf{x}_2, t_2 = t) \;=\; \int U(\mathbf{x}_2, \mathbf{x}_1)\Psi(\mathbf{x}_1, t_1 = 0)\, d^3\mathbf{x}_1 \,. \tag{52}$$

When a gauge transform is accompanied by a local phase transform of the wave function as in eq. (50), the propagation amplitude also changes it's phase. Indeed, in order to keep

eq. (52) working in a new gauge, we need

$$U'(\mathbf{x}_2,\mathbf{x}_1) \;=\; \exp\left(+i\tfrac{q}{\hbar c}\Lambda(\mathbf{x}_2)\right)\times U(\mathbf{x}_2,\mathbf{x}_1)\times\exp\left(-i\tfrac{q}{\hbar c}\Lambda(\mathbf{x}_1)\right). \tag{53}$$

where the first phase factor changes the phase of the $\Psi(\mathbf{x}_2,t_2=t)$ while the second phase factor compensates for the changed phase of the $\Psi(\mathbf{x}_1,t_1=0)$, thus

$$\begin{aligned}\Psi'(\mathbf{x}_2,t) \;&=\int U'(\mathbf{x}_2,\mathbf{x}_1)\times\Psi'(\mathbf{x}_1,0)\,d^3\mathbf{x}_1\\ &=\int\exp\big(+i\tfrac{q}{\hbar c}\Lambda(\mathbf{x}_2)\big)U(\mathbf{x}_2,\mathbf{x}_1)\exp\big(-i\tfrac{q}{\hbar c}\Lambda(\mathbf{A}_1)\big)\times\exp\big(+i\tfrac{q}{\hbar c}\Lambda(\mathbf{x}_1)\big)\Psi(\mathbf{x}_1,0)\,d^3\mathbf{x}_1\\ &=\;\exp\big(+i\tfrac{q}{\hbar c}\Lambda(\mathbf{x}_2)\big)\times\int U(\mathbf{x}_2,\mathbf{x}_1)\Psi(\mathbf{x}_1,0)\,d^3\mathbf{x}_1\\ &=\;\exp\big(+i\tfrac{q}{\hbar c}\Lambda(\mathbf{x}_2)\big)\times\Psi(\mathbf{x}_2,t).\end{aligned} \tag{54}$$

In particular, suppose $\mathbf{B}\equiv 0$ along the electron's path from $\mathbf{x}_1$ to $\mathbf{x}_2$ but the vector potential does not vanish, $\mathbf{A}\neq 0$. Then *locally* the vector potential is gauge-equivalent to zero, meaning there exist some $\Lambda(\mathbf{x})$ such that

$$\mathbf{A}_0(\mathbf{x}) \;=\; \mathbf{A}(\mathbf{x}) \;+\; \nabla\Lambda(\mathbf{x}) \;=\; 0, \tag{55}$$

if not everywhere then at least throughout the neighborhood of the electron's path. Then comparing the propagation amplitude $U_{\mathbf{A}}(\mathbf{x}_2,\mathbf{x}_1)$ in presence of the vector potential with the similar amplitude $U_0(\mathbf{x}_2,\mathbf{x}_1)$ for $\mathbf{A}_0\equiv 0$, we find

$$\begin{aligned}U_0(\mathbf{x}_2,\mathbf{x}_1) \;&=\; U_{\mathbf{A}}(\mathbf{x}_2,\mathbf{x}_1)\times\exp\left(\frac{iq}{\hbar c}\left(\Lambda(\mathbf{x}_2)-\Lambda(\mathbf{x}_1)\right)\right)\\ &=\; U_{\mathbf{A}}(\mathbf{x}_2,\mathbf{x}_1)\times\exp\left(\frac{iq}{\hbar c}\int_{\mathbf{x}_1}^{\mathbf{x}_2}\nabla\Lambda\cdot d\mathbf{x}\right)\\ &=\; U_{\mathbf{A}}(\mathbf{x}_2,\mathbf{x}_1)\times\exp\left(\frac{-iq}{\hbar c}\int_{\mathbf{x}_1}^{\mathbf{x}_2}\mathbf{A}\cdot d\mathbf{x}\right),\end{aligned} \tag{56}$$

and therefore

$$U_{\mathbf{A}}(\mathbf{x}_2, \mathbf{x}_1) \;=\; U_0(\mathbf{x}_2, \mathbf{x}_1) \times \exp\left(\frac{iq}{\hbar c} \int\limits_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{A} \cdot d\mathbf{x} \right). \tag{57}$$

Thus, even when the vector potential $\mathbf{A}$ does not lead to a magnetic field in the region the electron travels through, it still manages to change the phase of its propagation amplitude.

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

$$\Lambda(\mathbf{x}_2) \;-\; \Lambda(\mathbf{x}_1) \;= \int\limits_{\mathbf{x}_1}^{\mathbf{x}_2} \nabla\Lambda \cdot d\mathbf{x} \;=\; -\int\limits_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{A} \cdot d\mathbf{x}$$

works only if we integrate $\mathbf{A} \cdot d\mathbf{x}$ along the electron path rather than some other line. In the context of eq. (57), this means that

$$U_{\mathbf{A}}(\mathbf{x}_2, \mathbf{x}_1) \;=\; U_0(\mathbf{x}_2, \mathbf{x}_1) \times \left(\frac{iq}{\hbar c} \int\limits_{\text{electron's path}} \mathbf{A} \cdot d\mathbf{x} \right). \tag{58}$$

In the Aharonov–Bohm experiment, the electron can take two different paths from the same point $\mathbf{x}_g$ (the electron gun) to the same point $\mathbf{x}_s$ on the screen. The interference pattern on the screen follows from the net amplitude

$$U^{\text{net}}(\mathbf{x}_s, \mathbf{x}_g) \;=\; U^{\text{path 1}}(\mathbf{x}_s, \mathbf{x}_g) \;+\; U^{\text{path 2}}(\mathbf{x}_s, \mathbf{x}_g), \tag{59}$$

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

$$\Delta\varphi(\mathbf{x}_s) \;=\; \text{phase}\big(U^{\text{path 1}}(\mathbf{x}_s, \mathbf{x}_g)\big) \;-\; \text{phase}\big(U^{\text{path 2}}(\mathbf{x}_s, \mathbf{x}_g)\big). \tag{60}$$

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

according to eq. (58), specifically

$$\begin{aligned}
\text{phase}\left(U_{\mathbf{A}}^{\text{path }1}(\mathbf{x}_s,\mathbf{x}_g)\right) &= \text{phase}\left(U_{0}^{\text{path }1}(\mathbf{x}_s,\mathbf{x}_g)\right) + \frac{q}{\hbar c}\int\limits_{\text{path }1}\mathbf{A}(\mathbf{x})\cdot d\mathbf{x},\\
\text{phase}\left(U_{\mathbf{A}}^{\text{path }2}(\mathbf{x}_s,\mathbf{x}_g)\right) &= \text{phase}\left(U_{0}^{\text{path }2}(\mathbf{x}_s,\mathbf{x}_g)\right) + \frac{q}{\hbar c}\int\limits_{\text{path }2}\mathbf{A}(\mathbf{x})\cdot d\mathbf{x}.
\end{aligned} \tag{61}$$

Consequently, the phase difference (60) is affected by the vector potential according to

$$\begin{aligned}
\Delta\varphi_{\mathbf{A}} &= \Delta\varphi_0 + \frac{q}{\hbar c}\int\limits_{\text{path }1}\mathbf{A}(\mathbf{x})\cdot d\mathbf{x} - \frac{q}{\hbar c}\int\limits_{\text{path }2}\mathbf{A}(\mathbf{x})\cdot d\mathbf{x}\\
&= \Delta\varphi_0 + \frac{q}{\hbar c}\times F,
\end{aligned} \tag{62}$$

where F is the magnetic flux through the solenoid, and the second equality follows from eq. (48).

For different points $\mathbf{x}_s$ on the screen we have different $\Delta\varphi_0(\mathbf{x}_s)$, that's why we see the interference pattern on the screen! The magnetic flux term in eq. (62) is the same for all points on the screen,

$$\Delta_{\mathbf{A}}\varphi(\mathbf{x}_s) = \Delta_0\varphi(\mathbf{x}_s) + \frac{q}{\hbar c}\times F\,, \tag{63}$$

so it *shifts the whole interference pattern along the screen!* 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!

In the mathematical language, the Aharonov–Bohm effect feels 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

$$\oint_{\text{loop}} \mathbf{A}(\mathbf{x}) \cdot d\mathbf{x} \;=\; F(\text{loop}) \tag{64}$$

are gauge-invariant *for single-valued* $\Lambda(\mathbf{x})$, and when $\nabla \times \mathbf{A} \equiv 0$ everywhere outside the holes, then the fluxes (64) 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 (63) 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

$$F \;=\; \frac{2\pi\hbar c}{q} \times \text{an integer}, \tag{65}$$

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

$$F_1 \;=\; \frac{2\pi\hbar c}{|q|} \tag{66}$$

where q is the electric charge of the particles used in the experiment. For example, a SQUID (Superconducting Quantum Interferometry Device — *cf.* my notes on superconductivity (page 8) — measures the magnetic flux through a hole surrounded by superconductors 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

$$F_0 \;=\; \frac{2\pi\hbar c}{2e} \;=\; 2.067\,833\,667(52)\times 10^{-7}\ \text{Mx} \quad (\text{Maxwells or Gauss}\times\text{cm}^2), \tag{67}$$

or in MKSA units

$$F_0 \;=\; \frac{2\pi\hbar}{2e} \;=\; 2.067\,833\,667(52)\times 10^{-15}\ \text{Wb} \quad (\text{Webers or Tesla}\times\text{m}^2). \tag{68}$$

Note that particles of different charges would measure the fractional part of the magnetic flux F 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 F in two different units F_1 and F_2 with irrational F_1/F_2, we would be able to reconstruct the whole flux F 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

$$F_u \;=\; \frac{2\pi\hbar c}{e} \;=\; 2F_0 \;=\; 4.135\,667\,3(1)\times 10^{-7}\text{Mx}. \tag{69}$$

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

Magnetic Monopoles and Charge Quantization

Thus far, the experimental physicists have seen plenty of electric charges but no magnetic charges: all the magnets we see around us are dipoles, quadrupoles, or higher multipoles, but no monopoles. But if tomorrow somebody does find a magnetic monopole, our theories can be extended to accommodate them. Classically, Maxwell equations (in Gauss units) would

become

$$\begin{aligned}
\nabla \cdot \mathbf{E} &= 4\pi\rho_{\rm el}\,, \\
\nabla \cdot \mathbf{B} &= 4\pi\rho_{\rm mag}\,, \\
\nabla \times \mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} - \frac{4\pi}{c}\mathbf{J}_{\rm mag}\,, \\
\nabla \times \mathbf{B} &= +\frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi}{c}\mathbf{J}_{\rm el}\,;
\end{aligned} \tag{70}$$

and a static magnetic monopole — a point-like magnetic charge M — would create a Coulomb-like magnetic field

$$\mathbf{B}(\mathbf{x}) = \frac{M}{r^2}\mathbf{n}_r\,, \tag{71}$$

Classically, the value of the magnetic charge M could be anything, but in the quantum theory M must be quantized: As Paul A. M. Dirac showed in 1937, for any magnetic monopole in the Universe and any electrically charged particle in the Universe, the product of their charges must be an integer multiple of $\frac{1}{2}\hbar c$ (in Gauss units),

$$M \times q = \tfrac{1}{2}\hbar c \times \text{an integer}, \tag{72}$$

Or in MKSA units,

$$M \times q = \frac{2\pi\hbar}{\mu_0} \times \text{an integer}. \tag{73}$$

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

Let me start with a heuristic argument for the magnetic charge quantization (72), and then I'll give you the real argument of Dirac. Heuristically, we can make a toy model of a magnetic monopole by looking at a pole 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. Near the pole in question — but outside the magnet itself — the magnetic field looks just like the monopole field (71); but inside the magnet, there is a very strong field feeding the magnetic flux $4\pi M$ to 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 (71), 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 $4\pi M$ 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 eigen-waves 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 $4\pi M$ 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,*

$$M \times q \;=\; \frac{\hbar c}{2} \times \text{an integer}. \tag{72}$$

The Dirac's argument starts with the motion of a charged quantum particle in the magnetic field (71) of a monopole. To write the Schrödinger equation, we need the vector potential $\mathbf{A}(\mathbf{x})$, but the vector potential of the monopole's field is singular. Or rather, it's not only singular at the monopole's location itself, but it has at least one singular point on any closed surface surrounding the monopole. (Since otherwise the Stokes theorem would require the net magnetic flux through the surface to be zero instead of $4\pi M$.) Thus, the singularities of the vector potential must form a whole string stretching from the monopole to the infinity, or perhaps multiple strings.

Dirac's solution to this problem was to divide the space surrounding the monopole into two overlapping regions and use a different vector potential potential in each region. However, 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:

$$\begin{aligned} \mathbf{A}_N(r,\theta,\phi) &= M\,\frac{+1-\cos\theta}{r\sin\theta}\,\mathbf{n}_\phi \;=\; M\,(+1-\cos\theta)\,(\nabla\phi), \\ \mathbf{A}_S(r,\theta,\phi) &= M\,\frac{-1-\cos\theta}{r\sin\theta}\,\mathbf{n}_\phi \;=\; M\,(-1-\cos\theta)\,(\nabla\phi), \end{aligned} \tag{74}$$

The two potentials are gauge-equivalent:

$$\mathbf{A}_N \;-\; \mathbf{A}_S \;=\; M\,\frac{2}{r\sin\theta}\,\mathbf{n}_\phi \;=\; 2M\,\nabla\phi \;=\; \nabla\,(2M\phi) \tag{75}$$

so they lead to the same magnetic field, namely (71). Indeed,

$$\begin{aligned} \nabla\times\mathbf{A}_{\mathrm{N\,or\,S}} &= \nabla\times\left(M\,(\pm1-\cos\theta)\nabla\phi\right) \\ &= M\,(\nabla(\pm1-\cos\theta))\times\nabla\phi \\ &= M\,\frac{\sin\theta\,\mathbf{n}_\theta}{r}\times\frac{\mathbf{n}_\phi}{r\sin\theta} \\ &= M\,\frac{\mathbf{n}_r}{r^2}\,. \end{aligned} \tag{76}$$

Each vector potential (74) 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

⋆ 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.

$\mathbf{A}_S(r,\theta,\phi)$ 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

$$\mathbf{A}'(\mathbf{x}) \;=\; \mathbf{A}(\mathbf{x}) \;+\; \nabla\Lambda(\mathbf{x}) \quad \text{while} \quad \Psi'(\mathbf{x}) \;=\; \Psi(\mathbf{x}) \times \exp\left(i\tfrac{q}{\hbar c}\Lambda(\mathbf{x})\right) \quad \text{for the same } \Lambda(\mathbf{x}). \tag{50}$$

Consequently, the two different gauge-equivalent vector potentials $\mathbf{A}_N(\mathbf{x})$ and $\mathbf{A}_S(\mathbf{x})$ should come with two different wave functions $\Psi_N(\mathbf{x})$ and $\Psi_S(\mathbf{x})$ in the corresponding regions of space, and in the overlap between the two regions, the $\Psi_N(\mathbf{x})$ and the $\Psi_S(\mathbf{x})$ should be related by the appropriate phase transform (50). Specifically,

$$\Lambda(r,\theta,\phi) \;=\; 2M\,\phi\,, \tag{77}$$

$$\mathbf{A}_N(r,\theta,\phi) \;=\; \mathbf{A}_S(r,\theta,\phi) \;+\; \nabla\Lambda(r,\theta,\phi), \tag{78}$$

$$\Psi_N(r,\theta,\phi) \;=\; \Psi_S(r,\theta,\phi) \times \exp\left(i\,\frac{q}{\hbar c}\,\Lambda(r,\theta,\phi)\right). \tag{79}$$

Note: the gauge-transform parameter Λ in eq. (77) is multivalued since the longitude coordinate ϕ changes by 2π as we go around the equator. However, *multivalued $\Lambda(\mathbf{x})$ are OK as long as both the EM potentials and the wave functions it relates are single-valued,* which means that

$$\text{both} \quad \nabla\Lambda \quad \text{and} \quad \exp\left(i\,\frac{q}{\hbar c}\,\Lambda\right) \quad \text{must be single-valued.} \tag{80}$$

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

$$\exp\left(i\,\frac{q}{\hbar c}\,\Lambda\right) \;=\; \exp\left(i\,\frac{2qM}{\hbar c}\,\phi\right). \tag{81}$$

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 (77)–(79) is allowed in quantum

mechanics if and only if

$$\frac{2qM}{\hbar c} \text{ is an integer.} \tag{82}$$

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*

$$M \times q \;=\; \frac{\hbar c}{2} \times \text{an integer.} \tag{72}$$

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*,

$$q \;=\; \frac{\hbar c}{2M} \times \text{an integer.} \tag{83}$$

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

$$e^2 \;\approx\; \frac{\hbar c}{137}\,. \tag{84}$$

The monopole gives us an excellent reason for the charge quantization in the first place, but alas it does not explain the value (84) 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. (72) and (84), they are quantized in rather different units, e for the electric charges and

$$\frac{\hbar c}{2e} \;\approx\; \frac{137}{2}\,e \tag{85}$$

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 (85). 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 (85). 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 (72)! 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*, §2.7.