Simplified Path Integral Approach to the Aharonov-Bohm Effect

William O. Straub
Pasadena, California 91104
March 26, 2014

Abstract

In classical electrodynamics the vacuum is defined as a region where there are no electric or magnetic fields. In such a region, a charged particle (such as an electron) will feel no effect — the Lorentz force is zero. The space external to a perfect (i.e., infinite) solenoid can be considered an electromagnetic vacuum, since $\vec{E}$ and $\vec{B}$ vanish there. While a non-zero vector potential $\vec{A}$ does exist outside the solenoid, it can exert no influence on the particle, and as such cannot be directly detected or quantified classically. However, in 1959 Aharonov and Bohm predicted that a vector field would exert a purely quantum-mechanical effect on the phase of the particle's wave function, which in principle should be detectable. The predicted phase shift was not observed experimentally until 1986, when Tonomura brilliantly verified the effect using a microscopic solenoid. This paper provides a simplified explanation of the Aharonov-Bohm prediction using a path-integral approach that is suitable for the advanced undergraduate.

1. Introduction

The presence of an electric or magnetic field is easy to detect — one simply introduces a charged particle (fixed or moving at some initial velocity) and observes its behavior. If it accelerates or deflects, then there's an electromagnetic field nearby; the greater the deflection, the stronger the field. The electromagnetic field is, of course, composed of the electric field $\vec{E}$ and the magnetic field $\vec{B}$, either singly or in combination. But Maxwell's equations reveal a deeper phenomenon associated with these fields, one that was not experimentally demonstrated until relatively recently. It involves the concept of a *potential*, which comes in two forms — the 3-component *vector potential* $\vec{A}$ and the scalar potential Φ, both of which are generally time-dependent functions of space. These potentials arise from a consideration of the differential forms of the two homogeneous Maxwell's equations,

$$\vec{\nabla} \cdot \vec{B} = 0$$

$$\vec{\nabla} \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t}$$

which can be solved directly to give

$$\vec{B} = \vec{\nabla} \times \vec{A}$$

$$\vec{E} = -\vec{\nabla}\Phi - \frac{1}{c}\frac{\partial \vec{A}}{\partial t}$$

The four potential quantities combine into what is called the *four-potential* $A_\mu (= \Phi, \vec{A})$ which, unlike either $\vec{E}$ or $\vec{B}$, transforms like a Lorentz four-vector. It is in this sense that the four-potential is more fundamental than the electric and magnetic fields.

1. Classical Invisibility of the Potential

The scalar potential Φ is familiar from high school electromagnetism, where it is commonly identified with the potential difference or voltage V across a battery. But the vector potential $\vec{A}$ normally does not make its appearance until much later; indeed, many undergraduate science courses ignore it completely. Perhaps much of the reason for this stems from a preoccupation with the electric and magnetic fields themselves, which are usually the only "solutions" sought in undergraduate physics and engineering classes. When $\vec{A}$ is finally introduced, it's generally because some problems are easier to solve by first calculating the potential, then using $\vec{B} = \vec{\nabla} \times \vec{A}$ (in the same way that the scalar potential is used calculate $\vec{E}$).

But there are other reasons why $\vec{A}$ doesn't get more immediate attention. For one (unlike the magnetic field itself) the vector potential is impossible to observe classically. Another reason involves the related fact that, like the scalar potential, the vector potential does not have a unique representation in any given application. For example, the potential difference $\Delta V = V - V_0$ of a charged sphere depends on how the *ground potential* V_0 is defined; like potential energy, the only thing that's ever measured is the difference between two energy levels, as there is no *absolute* reference. The same is true for $\vec{A}$. To see this, we consider a simultaneous change in Φ and $\vec{A}$ given by the *gauge transformation*

$$\Phi' = \Phi + \frac{\partial \lambda}{\partial x^0}$$
$$\vec{A}' = \vec{A} - \vec{\nabla}\lambda$$

where $\lambda(x,t)$ is an *arbitrary* function of the spacetime coordinates. Substituting these new potentials into the expressions $\vec{B}' = \vec{\nabla} \times \vec{A}'$ and $\vec{E}' = -\vec{\nabla}\Phi' - \partial \vec{A}'/\partial x^0$ shows that

$$\vec{E}' = \vec{E}$$
$$\vec{B}' = \vec{B}$$

Thus, the electric and magnetic fields remain unchanged under a gauge transformation of the fields, and by this we say that Maxwell's equations are *gauge invariant*. Because of the arbitrariness of the gauge parameter $\lambda(x)$, the four-potential A_μ has no unique mathematical definition. Very often, a clever choice of the gauge parameter can be used to simply the calculation of $\vec{E}$ and $\vec{B}$. Several examples are the *Coulomb gauge* and the *Lorentz gauge*, but we will not be needing them in what follows.

For many years after Maxwell first set down his famous equations, the gauge property of the potentials was generally looked upon as only a useful computational device, even a mathematical oddity, and little real meaning was ascribed to the potentials themselves (the vector potential $\vec{A}$ in particular). Indeed, because it could not actually be seen or detected, it took a back seat to the magnetic field, which was considered the only real field. To give a concrete example of this, consider an ideal solenoid of very long (essentially infinite) length. When a current is sent through the coiled wire, a magnetic field is set up *within* the solenoid, while the magnetic field external to the coil is zero. If an observer now directs a moving particle of charge q outside the coil, the particle will not experience any deflection (recall the Lorentz force law $\vec{F} = q/c\,\vec{v} \times \vec{B}$), because the magnetic field is zero there. But the vector potential exterior to the solenoid is *not* zero; in fact, there is one non-zero component about the cylindrical axis of the solenoid given by

$$A_\phi = \frac{B r_s^2}{2r}$$

where r_s is the radius of the solenoid and r is the radial distance from the axis. Nevertheless, an observer sees the particle proceeding on its merry way, and concludes that no field is present. The vector potential thus escapes detection.

2. The Aharonov-Bohm Prediction

It was not until 1959 that a method was devised for demonstrating the physical reality of the vector potential. Actually, it was little more than a thought experiment because it could only be demonstrated mathematically. David Bohm and his graduate student Yakir Aharonov, both at the University of Bristol at the time, showed theoretically that the vector potential should be detectable when applied to the famous double-slit experiment. The proposed setup consisted of a coherent charged particle source and the usual double slit and distant detector screen. By firing the particles one at a time at the slit, each particle's wave function interferes with itself, resulting in the usual interference pattern at the detector. The Aharonov-Bohm *ansatz* was to imagine placing a tiny solenoid immediately behind and between the two slits, where presumably the solenoid's external vector potential would induce a measurable phase shift in the pattern.

3. Path Integral Approach

The mathematical reasoning we will use for the predicted shift is equivalent but different than that employed by Aharonov and Bohm, but it will provide additional insight into the strange world of the four-potential. It will also give us the chance to use a particularly powerful quantum tool, which is the *path integral* approach originally pioneered by physicist Richard Feynman in his 1942 Princeton doctoral dissertation. The path integral provides a particularly clear and elegant solution to the thought experiment that Aharonov and Bohm proposed.

The path integral is a mathematical way of describing the probability amplitude that a particle will go from Point A to Point B in some finite period of time. It says that the particle can traverse *any* intermediate path on its way to Point B. This means that, after leaving Point A, the particle can execute literally any of an *infinite* number of paths, crazy or otherwise, until it arrives at Point B. More amazingly, each path that the particle can take is just as important and logical as any other, *including* the classical, direct path from A to B. Does the particle really take all these paths? Mathematically, the answer is a definite yes. In reality, we can never really know just what the particle does. But for each possible path there is a corresponding probability amplitude, and these amplitudes can interfere with one another constructively or destructively. The crazy paths tend to be the ones that get canceled out by destructive interference, while the logical, "classical" paths reinforce one another. While the path integral is necessarily infinite-dimensional, it can in principle be calculated in closed form; for simple problems, such as the free particle and the harmonic oscillator, the calculation is straightforward and agrees perfectly with the demands of quantum mechanics.

Path integrals can also be applied to *fields*, in which case they give the probability amplitude that a field will propagate or transition from one field to another over a specified period of time. But the Aharonov-Bohm thought experiment described here will only be concerned with charged *particles* (like electrons) going from Point A (their source) through the double slit to Point B (the detector).

The path integral for a particle going from Point $x = A$ to Point $x = B$ when the slits are not present is expressed by

$$I = \int \mathscr{D}x(t) \exp\left(\frac{iS(x,t)}{\hbar}\right)$$

Here, I represents the probability amplitude for the overall process; it is a complex number, and its conjugate square represents the real probability that the particle will leave A and arrive at B. The quantity $\mathscr{D}x$ is shorthand for the infinite path sequence $\int dx_1\, dx_2\, dx_3 \ldots$, where each $x = x(t)$ represents a given path and the single integral sign represents an infinite number of them, one integral for each path. The quantity S in the exponential term is the *action*, which is defined as the integral of the lagrangian density $\mathscr{L}$ over time. Since we'll be considering the motion of classical charged particles moving with small velocities, we'll employ the non-relativistic form of the lagrangian. Recall that the lagrangian for a free particle of mass m is just $\mathscr{L}_0 = 1/2\, m(dx/dt)^2$. For a free particle having a charge q in the presence of a vector potential, it becomes

$$\mathscr{L} = \mathscr{L}_0 + \frac{q}{c} A_i \frac{dx^i}{dt}$$

(Note that we have set the scalar potential term $A_0 = \Phi$ to zero, because this field is absent in the Aharonov-Bohm setup.) The path now integral looks like

$$I = \int \mathscr{D}x(t) \exp\left(\frac{iS_0}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_0^t A_i \frac{dx^i}{dt} dt\right) \tag{3.1}$$

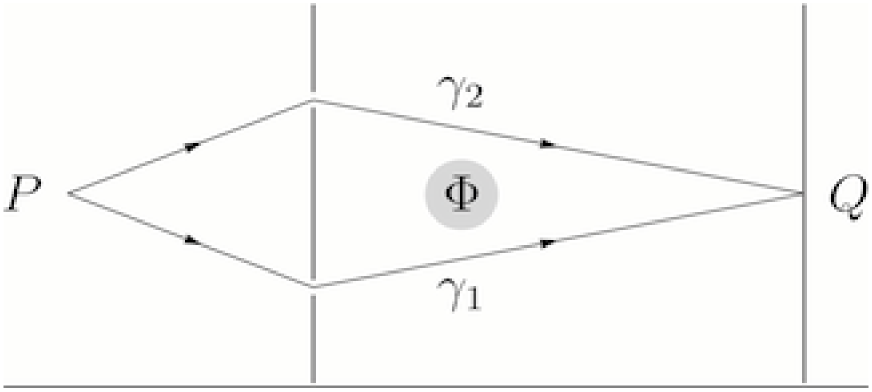
where $S_0 = \int_0^t \mathscr{L}_0 dt$. Lastly, note that the integral over time of the vector potential term becomes a line integral over space:

$$\int_0^t A_i \frac{dx^i}{dt} dt = \int_A^B A_i dx^i$$

so we have, finally,

$$I = \int \mathscr{D}x(t) \exp\left(\frac{iS_0}{\hbar}\right) \exp\left(\frac{iq}{\hbar c} \int_A^B A_i\, dx^i\right)$$

Armed now with the path integral, let us proceed to the Aharonov-Bohm experiment itself.



4. Derivation of the Aharonov-Bohm Phase Shift

Consider the above plan view of the classic double-slit experiment, consisting of a source P of identical particles each having a charge q (electrons will do nicely), an impenetrable screen with two closely-spaced, narrow slits, and a detector Q. In practice, the slit widths will be on the order of several microns, separated by a section of screen of comparable dimension. Such small distances are necessary in view of the quantum nature of the double slit experiment itself. In addition, we place a solenoid (Φ) immediately behind the slit separation. The solenoid itself must be extremely small, and the ends must be shielded in such a way as to prevent any "fringe" interaction with the charged particles.

We begin with the solenoid in place but with the current turned off, so $\vec{A} = 0$ outside the solenoid. We now start firing charged particles at the slits at a rate slow enough to prevent the particles from interacting with one another. Some of the individual particles will pass through Slit 1 (path γ_1)and some through Slit 2 (path γ_2), but in general each particle will seem to go through *both* slits — this is, after all, the nature of the quantum weirdness of the double slit experiment! We can associate a path integral for each slit:

$$I_1 = \int_1 \mathscr{D}x \, \exp\left(\frac{iS_0}{\hbar}\right)$$
$$I_2 = \int_2 \mathscr{D}x \, \exp\left(\frac{iS_0}{\hbar}\right)$$

where the subscripts refer to the paths taken by the particles through the two slits. The paths are constrained to go through their respective slits, but they're free-particle path integrals, and we know the solution to such integrals. For an unconstrained particle going from point x_A to point x_B, it's

$$I = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{im(x_B - x_A)^2}{2t}\right)$$

(The exponential term is a phase factor that will disappear when we take the conjugate square of I, so we'll ignore it in the subsequent analysis.) In view of this, I_1 and I_2 can differ only in phase, and so we write

$$I_1 = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right)$$
$$I_2 = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_2}{\hbar}\right) \tag{4.1}$$

where the quantities θ_1 and θ_2 are phase constants reflecting the path constraints. The total path integral is then $I = I_1 + I_2$, which we can write as

$$I = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2} \exp\left(\frac{i\theta_1}{\hbar}\right) \left[1 + \exp\left(\frac{i(\theta_2 - \theta_1)}{\hbar}\right)\right]$$

The quantity $(\theta_2 - \theta_1)/\hbar \equiv \Delta$ represents the *phase difference* between the combined paths; the detector at any given point will see constructive interference when $\Delta = 2n\pi$ and destructive interference when $\Delta = (2n+1)\pi$, where n is an integer. The path integral I is a probability amplitude, and to get the probability we have to take the conjugate square of this quantity. It is not difficult to show that

$$I^*I = |I|^2 = \frac{2m}{\pi\hbar t}\cos^2\left(\frac{1}{2}\Delta\right) \tag{4.2}$$

Thus, the path integral approach to the double-slit experiment explains the sinusoidal interference pattern seen at the detector.

We now start the current running into the solenoid. What effect, if any, can we expect? The presence of the solenoid induces a longitudinal magnetic field inside the solenoid, but this magnetic field is restricted to its *interior*; thus, there is no external magnetic field to affect the charged particles as they pass through the slits. As far as they're concerned, they see the same vacuum as they did when the solenoid was turned off. But in view of (3.1) there's now an additional term in the path integral, and we can no longer use the free-particle solution. Or can we? From the symmetry of the set-up we can assume with negligible error that the line integral $\int A_i\, dx^i$ is independent of the path taken from one point to another, so we can take this term out of the path integral and treat it as a phase coefficient. For path n $(= 1, 2)$, we have

$$\begin{aligned} I_n &= \int_n \mathscr{D}x(t)\exp\left(\frac{iS_0}{\hbar}\right)\exp\left(\frac{iq}{\hbar c}\int_n A_i dx^i\right) \\ &= \exp\left(\frac{iq}{\hbar c}\int_n A_i dx^i\right)\int_n \mathscr{D}x(t)\exp\left(\frac{iS_0}{\hbar}\right) \\ &= \left(\frac{m}{2\pi i\hbar t}\right)^{1/2}\exp\left(\frac{i\theta_n}{\hbar}\right)\exp\left(\frac{iq}{\hbar c}\int_n A_i dx^i\right) \end{aligned}$$

As before, the total path integral is $I = I_1 + I_2$, and with the solenoid turned on this can be written as

$$\begin{aligned} I &= \left(\frac{m}{2\pi i\hbar t}\right)^{1/2}\exp\left(\frac{i\theta_1}{\hbar}\right)\exp\left(\frac{iq}{\hbar c}\int_1 A_i\, dx^i\right) + \left(\frac{m}{2\pi i\hbar t}\right)^{1/2}\exp\left(\frac{i\theta_2}{\hbar}\right)\exp\left(\frac{iq}{\hbar c}\int_2 A_i\, dx^i\right) \\ &= \left(\frac{m}{2\pi i\hbar t}\right)^{1/2}\exp\left(\frac{i\theta_1}{\hbar}\right)\exp\left(\frac{iq}{\hbar c}\int_1 A_i dx^i\right)\left[1 + \exp(i\Delta)\exp\left[\frac{iq}{\hbar c}\left(\int_2 A_i dx^i - \int_1 A_i dx^i\right)\right]\right] \end{aligned} \tag{4.3}$$

where we have pulled out two phase factors (which will cancel out when we calculate $|I|^2$). The last exponential quantity represents a closed path from the particle source to the detector and back:

$$\int_2 A_i dx^i - \int_1 A_i dx^i = -\oint A_i\, dx^i$$

where the minus sign reminds us that the closed path under consideration is clockwise. Using Stoke's theorem, this integral can be written as the surface integral

$$\oint A_i\, dx^i = \iint \vec{\nabla}\times\vec{A}dS$$

where the surface in question is the cross section of the solenoid. But this is just the magnetic flux Φ_B through the axis of the solenoid, given by

$$\iint \vec{\nabla}\times\vec{A}dS = \iint \vec{B}\cdot\hat{n}\, dS = \Phi_B$$

Equation (4.3) now becomes

$$I = \left(\frac{m}{2\pi i\hbar t}\right)^{1/2}\exp\left(\frac{i\theta_1}{\hbar}\right)\exp\left(\frac{iq}{\hbar c}\int_1 A_i dx^i\right)\left[1 + \exp(i\Delta)\exp\left(\frac{-iq\Phi_B}{\hbar c}\right)\right]$$

The square of this path integral is easily shown to be

$$|I|^2 = \frac{2m}{\pi \hbar t} \cos^2 \left[\frac{1}{2} \left(\Delta - \frac{q\Phi_B}{\hbar c} \right) \right] \tag{4.4}$$

Comparing this result with (4.2), we see that the effect of the solenoid on the charged particles is to shift the pattern in accordance with

$$\Delta \to \Delta - \frac{q\Phi_B}{\hbar c}$$

When Aharonov and Bohm derived this formula there was no way to test this shift experimentally — the required solenoid had to be perfectly shielded and of such a small size (no more than several microns in diameter) that fabrication was impossible. And while there was hardly any doubt that they had made a major discovery, many in the physics community continued to doubt that the vector potential could ever be detected, and no less an authority than Niels Bohr expressed doubt on the Aharonov-Bohm prediction.

5. Experimental Verification of the Aharonov-Bohm Effect

In the years following publication of the Aharonov-Bohm paper, several attempts were made by various researchers to construct a suitable solenoid. For a while, it appeared that microscopic magnetized iron fibers might be workable, and results were obtained that appeared to verify the Aharonov-Bohm prediction. But indisputable results were not achieved until 1986, when Akira Tonomura and his colleagues (following earlier efforts they conducted in 1982) succeeded in producing a 6-micron diameter, micro-fabricated toroidal solenoid utilizing the Meissner effect. The results matched the predicted phase shift perfectly, and the Aharonov-Bohm effect was finally verified experimentally. For his work, Tonomura received the Nishina Memorial Prize, the Asahi Prize, the Japan Academy and Imperial Prize, and the Benjamin Franklin Medal in Physics.

6. Comments

There are few more important quantities that Nature hides so effectively than the four-potential A_μ. It underlies all of electrodynamics, yet its inherent gauge arbitrariness made its physical existence doubtful until long after Maxwell first wrote down his equations.

But arbitrariness in a physical theory is often a sign from Nature that something profound is going on. Indeed, arbitrariness in an action integral actually represents a *mathematical symmetry*, and in a seminal paper written in 1918 the noted German mathematician Emmy Noether proved that mathematical symmetries are responsible for (and equivalent to) *conservation* theorems. A famous example is the local gauge (or phase) symmetry of quantum theory, which the great German mathematical physicist Hermann Weyl subsequently showed is responsible for the conservation of electric charge.

In order to simplify the problem and make it more accessible to the student, the path integral technique used here relied heavily on the free-particle path integral to avoid evaluating the infinite integral. This effectively required "localizing" the paths around the slits and the solenoid, along with the line integral $\int A_i \, dx^i$ itself, but the end result agrees with a more mathematically rigorous analysis. It should be noted by the student, however, that even the free-particle and simple harmonic oscillator problems require some rather cumbersome (if straightforward) algebra. Shankar's text devotes two chapters to the path integral (simple and advanced), and the student is referred to that book for additional information.

References

1. Aharonov, Y., Bohm, D. (1959). "Significance of electromagnetic potentials in quantum theory". Phys. Rev. **115**: 485–491.
2. Sakurai, J. J. (1985). *Modern quantum mechanics*. Addison-Wesley, Menlo Park.
3. Shankar, R. (1994). *Principles of quantum mechanics*, 2nd ed. Plenum Press, New York.
4. Tonomura, A. et al. (1986). "Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave". Phys. Rev. Lett. **56**: 792–795.