Published on 6 April, 2026
Volume 1, No. 1, April 2026

PANE

PANE Journal *of* Physics Education

An International Peer-Reviewed Journal
Advancing Research in Physics Teaching and Learning

A PUBLICATION OF THE PHYSICS ACADEMY OF THE NORTH EAST
Available online - visit https://journals.paneindia.org/index.php/pjpe

e-ISSN:

Authors should visit https://journals.paneindia.org/index.php/pjpe/submission for submission of manuscripts. The submission of a paper in PANE Journal of Physics Education should imply that the paper has not been published previously nor has been sent for publication elsewhere. For manuscript preparation, the authors must adhere to "Submission Guidelines for Authors" given at https://journals.paneindia.org/index.php/pjpe/sgfa.

Manuscripts and related matters may be addressed to the Editor-in-Chief, PANE Journal of Physics Education, Guwahati, Assam, India. Email: editorpjpe@gmail.com

Volume 1, No. 1, April 2026

PANE Journal *of* Physics Education

A PUBLICATION OF THE PHYSICS ACADEMY OF THE NORTH EAST
Available online - visit https://journals.paneindia.org/index.php/pjpe

e-ISSN:

BOARD OF EDITORS

CONTENTS

IS THE VECTOR POTENTIAL REAL?

Dipankar Kaundilya

Department of Physics, IISc Bangalore, Bangalore-560012, Karnataka, India

Corresponding Author: kaundilyadipankar@gmail.com

Abstract: This article explores the fundamental relationship between symmetry principles and topological structures within a consistent theoretical framework. By demonstrating that gauge transformations are canonical, we establish that the vector potential is not merely a mathematical auxiliary, but rather it functions as a primary coordinate of the Hamiltonian and quantum mechanical descriptions. Through a mathematical analysis of solenoid and toroidal geometries, we derive the Aharonov-Bohm (AB) phase shift and energy spectra, illustrating how nontrivial topology induces measurable physical effects in strictly field-free regions. The discussion extends to the Dirac quantization condition via the Wu-Yang construction of magnetic monopoles and concludes by framing the potential-centric view as a necessary precursor to Lattice Gauge Theory and the geometric connections of General Relativity. This synthesis reinforces the shift in physical ontology from local fields to global, gauge-invariant holonomies as the foundational entities of modern physics.

Keywords: Aharonov-Bohm Effect, Gauge Symmetry, Topology, Vector Potential, Dirac Monopole, Holonomy.

1 Introduction

Symmetry principles are foundational to physics, where Abelian gauge symmetry gives rise to electromagnetism (EM). The unification of fundamental interactions, such as the electroweak theory, relies heavily on more complex, non-Abelian gauge fields and mechanisms for broken symmetry. In quantum mechanics (QM), local gauge transformations of wavefunctions necessitate the introduction of vector potentials, highlighting a deep link between seemingly abstract mathematical concepts and physical reality.

Historically, electromagnetism initially relegated potentials to the status of mathematical auxiliaries, favoring the directly measurable $\mathbf{E}$ and $\mathbf{B}$ fields. However, the advent of quantum mechanics revealed that the phase of the wavefunction couples directly to the potential $\mathbf{A}$, suggesting a deeper, non-local reality[1]. This article reexamines this relationship and proves that the vector potential is a real physical entity. This view is established in three steps, first by proving the canonical nature of gauge transformations and second, by demonstrating that local phase symmetry in quantum mechanics necessitates the existence of the gauge field; and third, by analyzing the topological influence of the Aharonov-Bohm effect in solenoid and toroidal geometries. We bridge these foundational concepts with modern developments in Dirac monopoles and Lattice Gauge Theory [2]. Ultimately, we argue for a potential-centric view underpinned by gauge theory and topology, moving toward the geometric descriptions found in General Relativity.

2 Gauge transformation is canonical

It is perhaps surprising that the Hamiltonian equations for a particle in an electromagnetic field involve potentials rather than fields [1]. Fields are defined in terms of forces, while potentials lack an obvious operational definition and are not uniquely defined [1]. In quantum mechanics, potentials govern both orbits and phase relations, affecting interference patterns even in field-free regions [1]. This clarifies the use of potentials in quantum mechanics, though it remains less clear in classical mechanics [1].

The potentials $\mathbf{A}$ and $\mathbf{\Phi}$ determine the fields according to the relations:

$$B = \nabla \times A \quad \text{and} \quad E = -\nabla\Phi - \frac{\partial A}{\partial t} \tag{2.1}$$

New potentials can be introduced via a gauge transformation:

$$A' = A + \nabla\Lambda \quad \text{and} \quad \Phi' = \Phi - \frac{\partial \Lambda}{\partial t} \tag{2.2}$$

where $\Lambda(r,t)$ is an arbitrary differentiable function [1].

This gauge transformation leaves the values of B and E unchanged, demonstrating that A and Φ are not uniquely determined by the forces they exert [1]. This non-uniqueness can simplify computations but may also introduce complications, as even simple fields can become time-dependent and complex under a gauge transformation with an arbitrary Λ [1].

Such transformations encompass a variety of physical symmetries. For instance, a rotation in phase space, defined by $Q = q\cos\alpha - p\sin\alpha$ and $P = q\sin\alpha + p\cos\alpha$, preserves the canonical relations while merely shifting the perspective of the observer. Similarly, hyperbolic transformations, such as the Bogoliubov transformation used in many-body theory, redistribute the identity of particles and excitations while maintaining the underlying commutation relations.

A transformation $(p_i, q_i) \to (P_i, Q_i)$ is canonical if it satisfies the following Poisson brackets:

$$[P_i, P_k]_{q,p} = 0 = [Q_i, Q_k]_{q,p} \quad \text{and} \quad [Q_i, P_k]_{q,p} = \delta_{ik} \tag{2.3}$$

Considering the Hamiltonians $H(p_i, q_i, t)$ and $H'(P_i, Q_i, t)$ derived from Lagrangians L and L', we have:

$$H(p_i, q_i, t) = \sum_j p_j \dot{q}_j - L(q_i, \dot{q}_i, t) \tag{2.4}$$

and

$$H'(P_i, Q_i, t) = \sum_j P_j \dot{Q}_i - L'(q_i, \dot{q}_i, t) \tag{2.5}$$

where P_i is related to p_i by a derivative of a function $f(q_i, t)$:

$$P_i = \frac{\partial L'}{\partial \dot{q}_i} = p_i + \frac{\partial}{\partial \dot{q}_i}\frac{df(q_i,t)}{dt} = p_i + \frac{\partial}{\partial q_i} f(q_i, t) \tag{2.6}$$

and $Q_i = q_i$.

To show that the transformation $(p_i, q_i) \to (P_i, Q_i)$ is canonical, we calculate the Poisson brackets for (P_i, Q_i) :

$$[Q_i, Q_k]_{q,p} = \sum_j \left[\frac{\partial Q_i}{\partial q_j}\frac{\partial Q_k}{\partial p_j} - \frac{\partial Q_i}{\partial p_j}\frac{\partial Q_k}{\partial q_j}\right] = 0 \tag{2.7}$$

as $\frac{\partial Q_i}{\partial p_j} = 0$.

$$[P_i, P_k]_{q,p} = \sum_j \left[\frac{\partial P_i}{\partial q_j}\frac{\partial P_k}{\partial p_j} - \frac{\partial P_i}{\partial p_j}\frac{\partial P_k}{\partial q_j}\right] = \frac{d}{dt}\left[\frac{\partial^2 f}{\partial q_k \partial q_i} - \frac{\partial^2 f}{\partial q_i \partial q_k}\right] = 0 \tag{2.8}$$

and finally

$$[Q_i, P_k]_{q,p} = \sum_j \left[\frac{\partial Q_i}{\partial q_j} \frac{\partial P_k}{\partial p_j} - \frac{\partial Q_i}{\partial p_j} \frac{\partial P_k}{\partial q_j} \right] = \delta_{ik} \tag{2.9}$$

Thus, gauge transformations that connect Lagrangians differing by a total time derivative of coordinate and time are canonical.

By proving that gauge transformations are canonical, we establish that the vector potential is not merely a mathematical convenience but a consistent coordinate of the Hamiltonian framework. This formal equivalence allows the potential to maintain its structural integrity as we transition from the classical equations of motion to the quantum mechanical wavefunction. That is the generalized coordinates (the potentials $\mathbf{A}$, $\mathbf{\Phi}$) we use to build the Schrödinger equation are as fundamental as the positions and momenta in classical description. Consequently, the mathematical freedom inherent in the choice of gauge provides the necessary degree of freedom to accommodate the local phase symmetries of quantum theory.

Once we grasp this idea, then we can appreciate that while the canonical momenta $\mathbf{p}$ is the fundamental coordinate of the Hamiltonian, it is the kinetic or mechanical momentum $\mathbf{\Pi} = \mathbf{p} - \mathbf{qA}$ that couples to the particle's velocity. We can easily verify the commutation relation, $[\Pi_i, \Pi_j] = q\epsilon_{ijk}B_k$, i.e. the kinetic momenta commute only if the magnetic field is zero. This reinforces the concept that the vector potential $\mathbf{A}$ functions like a coordinate in a field free region.

3 Local Phase Symmetry and the Gauge Field

This mathematical freedom within the gauge transformation finds its physical counterpart in the phase of the quantum mechanical wavefunction. To trace a deep link between Electromagnetism and Quantum Mechanics, we observe that the arbitrary nature of the vector potential $\mathbf{A}$ is not just a classical redundancy, but the very mechanism that allows the overall phase of Ψ to be changed locally without affecting physical observables. This suggests that the "background" gauge field is inextricably tied to the local symmetry of the particle's probability amplitude.

Let's analyze the vector potential.

$$B = \nabla \times A \tag{3.1}$$

$$A(x) \to A'(x) = A(x) + \nabla\Lambda(x) \tag{3.2}$$

Now we write the equation in the integral form:

$$\oint_\Gamma A \cdot dx = \iint_{S(\Gamma)} B \cdot ds = F(\Gamma). \tag{3.3}$$

That is, the line integral of the vector potential around a closed loop Γ is equal to the surface integral of the magnetic field over the area $S(\Gamma)$ enclosed by the loop. The magnetic flux $F(\Gamma)$ is then invariant under the Gauge transformation because it is directly related to B. Note also that the gradient term $\nabla\Lambda(x)$ does not contribute to the line integral over a closed loop.

The integral form in Equation (3.3) reveals that the vector potential is more than a computational aid; it functions as a **connection** on a $U(1)$ **fiber bundle**. In this geometric view, the phase of the wavefunction is not a global constant but a local property that "rotates" as the particle moves through space. The physical result of traversing a closed loop Γ is the acquisition of **holonomy**—the net group element $H(\Gamma) = \exp(\frac{iq}{\hbar} \oint_\Gamma \mathbf{A} \cdot d\mathbf{l})$. Even when the local magnetic field B (the curvature of the bundle) vanishes along the path, the holonomy can remain non-trivial $(H(\Gamma) \neq 1)$ due to the **non-trivial topology** of the space. This identifies the Aharonov-Bohm effect as a global manifestation of gauge symmetry, providing a direct conceptual bridge to the **Wilson loops** of lattice gauge theory and the geometric connections of **General Relativity** discussed in the concluding sections of this article.

Now, let's analyze the phase of a wavefunction. We are free to change the phase under the following prescription [3]:

$$\Psi(x) \to \Psi'(x) = e^{i\theta}\Psi(x) \tag{3.4}$$

Note that the relevant physical quantities, e.g., the probability density $|\Psi|^2$, remain unchanged.

Something very interesting happens if we choose the phase angle θ independently in different places. That is, QM admits the local phase transformations:

$$\Psi(x) \rightarrow e^{i\theta(x)}\Psi(x) \tag{3.5}$$

where the angle θ is now allowed to be a function of the position x. Quantities such as probability density $\Psi^*\Psi$ and expectation values $\int \Psi^* V \Psi d^3x$ remain unchanged even under local phase transformations.

What about probability currents $\nabla\Psi$ and kinetic energy operators ∇^2 in the Schrödinger equation?

$$\nabla\Psi(x) \rightarrow \nabla\Psi'(x) = \nabla[e^{i\theta(x)}\Psi(x)] = e^{i\theta(x)}\nabla\Psi(x) + i\nabla\theta(x)e^{i\theta(x)}\Psi(x) \tag{3.6}$$

Note that the second term does not comply with the fact that everything simply changes by a phase factor. However, for a charged particle, the conjugate momentum p is replaced by[4]:

$$p \rightarrow p - qA(x) \tag{3.7}$$

This substitution is known as the **Minimal Coupling**. For a while it may appear as a post-hoc rule to link the particle's motion to the EM field, but it is necessary because it is the only coupling that allows the extra gradient term produced by a local phase shift $(\nabla\theta(x))$, to be exactly canceled by the gauge transformation of the vector potential $(\nabla\Lambda(x))$. Without this specific term, the Schrödinger equation would not be gauge invariant and the choice of gauge would indeed change the probability density.

Using canonical quantization, p is replaced by the operator $-i\nabla$. Hence, we define a covariant derivative D as:

$$-i\nabla \rightarrow -i[\nabla - iqA(x)] \equiv -iD \tag{3.8}$$

Now we perform the momentum operation on Ψ:

$$(\nabla - iqA)\Psi \rightarrow (\nabla - iqA')\Psi' = e^{i\theta(x)}(\nabla - iqA)\Psi + i\nabla(\theta(x) - q\Lambda(x))e^{i\theta(x)}\Psi \tag{3.9}$$

Note that the extraneous term would cancel out if we set:

$$\Lambda(x) = \frac{\theta(x)}{q} \tag{3.10}$$

Then everything simply acquires the same phase, i.e.,

$$D\Psi \rightarrow D'\Psi' = e^{i\theta(x)}[D\Psi] \tag{3.11}$$

where D' indicates that the covariant derivative is to be constructed using the transformed vector potential A'.

Thus, if the original wavefunction Ψ satisfies the Schrödinger equation:

$$\left[-\frac{1}{2m}(\nabla - iqA)^2 + V\right]\Psi(x) = E\Psi(x) \tag{3.12}$$

then the transformed wavefunction Ψ' also satisfies the same equation with A' and the same eigenvalue E, and hence both would describe the same physics.

A transformation that has no effect on the physics—that is to say, one that maps a solution of the equation into another solution—is called a **symmetry**; if the transformation can be implemented at different points, the symmetry is said to be a **local symmetry** or a **Gauge symmetry**.

This invariance demonstrates that the vector potential $\mathbf{A}$ is more than a mathematical auxiliary; it is a fundamental requirement for the local gauge symmetry of the wavefunction. While classical intuition suggests that the physical influence of electromagnetism is restricted to regions where the fields $\mathbf{E}$ and $\mathbf{B}$ are non-zero, this quantum mechanical coupling to the potential implies a "non-local" interaction. To move from this theoretical framework to an empirical proof, we must examine an experimental scenario where the potential acts in a truly field-free region: the Aharonov-Bohm effect.

4 Aharonov-Bohm Effect

The preceding derivation establishes the vector potential $\mathbf{A}$ as a mathematical necessity for the local gauge invariance of the Schrödinger equation. However, to confirm that this "background swirl" is a physical reality rather than a gauge-dependent artifact, we must identify a scenario where $\mathbf{A}$ exerts a measurable influence in the absence of a Lorentz force. This empirical test is provided by the **Aharonov-Bohm (AB) effect**, which occurs when a charged particle travels through a multiply connected space. By analyzing the specific geometries of the **solenoid** and the **toroid**, we can demonstrate how the global topology of the potential induces an observable phase shift, even when the particle is strictly confined to field-free regions.

4.1 Solenoid

We model an infinitely long solenoid as a cylinder of radius R centered on the z-axis, carrying a uniform internal magnetic field $B_0\hat{k}$ while the field outside is zero. Let's calculate the vector potential for the interior region $r \leq R$ and for the exterior region $r \geq R$.

4.1.1 Vector Potential for Solenoid

The vector potential must respect the cylindrical symmetry, because the source (the solenoid) is invariant under rotations about the z-axis and translations along it, the vector potential must reflect this cylindrical symmetry. We assume that it is purely azimuthal, such that $A = A(r)\hat{\phi}$. Then we utilize Stokes's theorem to relate the line integral of A around a circular loop of radius r to the magnetic flux F through the loop:

$$\oint_{\Gamma} A \cdot dl = \oint_{S(\Gamma)} B \cdot ds = \oint_{S(\Gamma)} (\nabla \times A) \cdot ds = F(\Gamma) \tag{4.1}$$

This choice of a purely azimuthal potential $\mathbf{A} = A(r)\hat{\phi}$ is dictated by the cylindrical symmetry of the infinite solenoid. While the magnetic field $\mathbf{B}$ is zero for $r > R$, the continuity of the vector potential across the boundary ensures that the information of the enclosed flux is "carried" into the exterior region as illustrated in Figure 1.

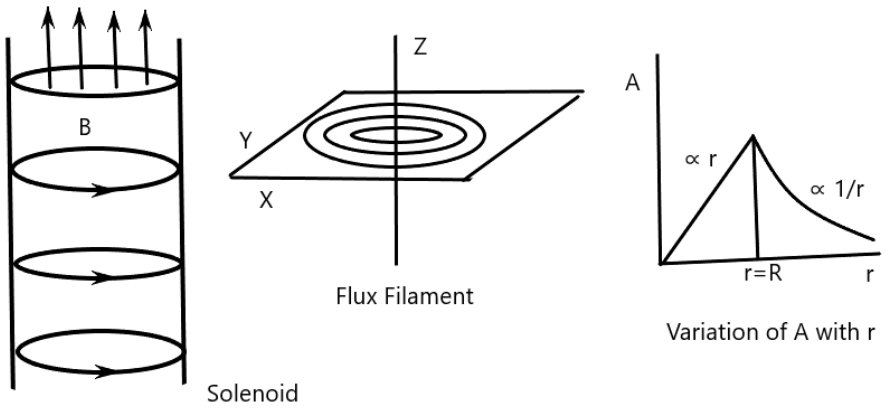


Figure 1: Vector potential of an infinite solenoid. The magnetic field $\mathbf{B}$ is uniform inside ($r < R$) and zero outside ($r > R$). However, the azimuthal vector potential $\mathbf{A}$ is non-zero everywhere, falling off as $1/r$ in the exterior region, which is the physical basis for the Aharonov-Bohm effect.

For the interior region $r \leq R$, the magnetic field inside the solenoid is uniform. The enclosed magnetic flux through a loop of radius r is:

$$\int A \cdot dl = A(2\pi r) = \int_0^r B_0(2\pi r')dr' = B_0\pi r^2 \tag{4.2}$$

$$A = \frac{B_0 r}{2}\hat{\phi} \tag{4.3}$$

For the exterior region $r \geq R$, the magnetic field outside the solenoid is zero. However, the loop of radius r still encloses the total flux contained within the cylinder of radius R:

$$\int A \cdot dl = A(2\pi r) = \int_0^R B_0(2\pi r')dr' = B_0\pi R^2 \tag{4.4}$$

$$A = \frac{B_0 R^2}{2r}\hat{\phi} \tag{4.5}$$

The vector potential for the given solenoid is:

$$A = \begin{cases} \frac{B_0 r}{2}\hat{\phi} & \text{for } r \leq R \\ \frac{B_0 R^2}{2r}\hat{\phi} & \text{for } r \geq R \end{cases} \tag{4.6}$$

Note that at the value at the boundary $r = R, A = \frac{B_0 R}{2}$ matches for the continuity of both inside and outside the solenoid. This proves that the vector potential is a "well-behaved" coordinate that transitions smoothly between the hidden flux and the reachable space.

To check the differentiability of the vector potential at the boundary $r = R$, we calculate the radial derivative dA/dr from both sides using Equation (4.6):

$$\left.\frac{dA}{dr}\right|_{r\to R^-} = \frac{d}{dr}\left(\frac{B_0 r}{2}\right) = \frac{B_0}{2} \quad \text{vs.} \quad \left.\frac{dA}{dr}\right|_{r\to R^+} = \frac{d}{dr}\left(\frac{B_0 R^2}{2r}\right) = -\frac{B_0}{2} \tag{4.7}$$

Since the derivatives $\frac{B_0}{2}$ and $-\frac{B_0}{2}$ are not equal, the vector potential is **not differentiable** at $r = R$, resulting in a kink at the boundary.

4.1.2 Flux Filament

Now, if we shrink the radius of the solenoid to zero, we create a flux filament (an ideal line of flux) along the z axis. When the radius R shrinks to zero while the magnetic field B_0 increases such that the total magnetic flux F_0 remains constant, we create a flux filament. The total flux is defined by:

$$F_0 = \lim_{R\to 0}(B_0\pi R^2) \tag{4.8}$$

In this limit, the magnetic field B becomes a Dirac Delta function along the z axis:

$$B = F_0\delta(x)\delta(y)\hat{k} \tag{4.9}$$

Since the radius $R \to 0$, we only need to consider the region $r > R$. Then $\oint A \cdot dl = F_0$. Assuming azimuthal symmetry $A = A(r)\hat{\phi}$, the path integral over a circle of radius r is $A(r)2\pi r = F_0$, and hence:

$$A(r) = \frac{F_0}{2\pi r} \tag{4.10}$$

This expression is valid for all $r > 0$. Note that the vector potential is singular at $r = 0$, reflecting the infinite density of the flux line at the origin. Consider a vector potential defined by:

$$A_x = -\frac{F_0}{2\pi}\frac{y}{x^2+y^2}, \quad A_y = \frac{F_0}{2\pi}\frac{x}{x^2+y^2}, \quad A_z = 0 \tag{4.11}$$

Let's evaluate its curl in Cartesian coordinates:

$$\nabla \times A = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\hat{i} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\hat{j} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\hat{k} \tag{4.12}$$

Since $A_z = 0$ and A_x, A_y do not depend on z, all derivatives involving z or A_z are zero. Thus, we only have to calculate the z component:

$$\frac{\partial A_y}{\partial x} = \frac{y^2 - x^2}{(x^2 + y^2)^2}, \quad \frac{\partial A_x}{\partial y} = \frac{y^2 - x^2}{(x^2 + y^2)^2} \tag{4.13}$$

Hence, the z component of the curl is also zero. Since all components are zero, $\nabla \times A = 0$ everywhere except at the origin $(0, 0)$ where the vector potential becomes singular. As $B = \nabla \times A$, this represents a region with zero magnetic field but non-zero vector potential.

We can see that A is ill-defined (singular) along the entire z axis. For any point on the z axis, the coordinates are $(0, 0, z)$. The denominator $x^2 + y^2$ becomes zero at $x = 0, y = 0$, meaning the vector potential is singular and the field blows up at the origin for all values of z.

This singularity represents an infinitely thin solenoid located exactly on the z axis. While the curl of A is zero everywhere else, it is a Dirac delta function at the origin:

$$B = \nabla \times A = F_0 \delta(x)\delta(y)\hat{k} \tag{4.14}$$

Because the z axis is ill-defined, we must "puncture" it from our coordinate system. This makes the space multiply connected, a topological requirement for the Aharonov-Bohm phase shift to exist.

Even though the curl is zero in the reachable region, the line integral around any loop Γ enclosing the z axis is non-zero, i.e., $\oint_\Gamma A \cdot dl = F_0$. To use Stokes' theorem, we first evaluate the circulation of A along a circular path Γ of radius R in the x-y plane, centered on the z axis.

We use polar coordinates where $x = R\cos\phi$, $y = R\sin\phi$, and $dl = (-R\sin\phi\hat{i} + R\cos\phi\hat{j})d\phi$. Then the vector potential is:

$$A = -\frac{\sin\phi}{R}\hat{i} + \frac{\cos\phi}{R}\hat{j} \tag{4.15}$$

Thus, the circulation is:

$$\int_0^{2\pi} \left(-\frac{F_0}{2\pi}\frac{\sin\phi}{R}\hat{i} + \frac{F_0}{2\pi}\frac{\cos\phi}{R}\hat{j}\right) \cdot (-R\sin\phi\hat{i} + R\cos\phi\hat{j})d\phi = \frac{F_0}{2\pi}\int_0^{2\pi} (\sin^2\phi + \cos^2\phi)d\phi = F_0 \tag{4.16}$$

According to Stokes' theorem, the line integral around a closed loop is equal to the flux of the curl through the surface enclosed by that loop:

$$\oint_\Gamma A \cdot dl = \iint_{S(\Gamma)} (\nabla \times A) \cdot ds = F_0 \tag{4.17}$$

Since we previously proved that $\nabla \times A = 0$ everywhere except at the origin, the entire value of the integral (F_0) must be concentrated at the singularity on the z axis.

To satisfy the condition that the integral over any area containing the origin is F_0 while the function remains zero elsewhere, we use the two-dimensional Dirac delta function. Thus, the curl of the vector potential on the z axis is:

$$B = \nabla \times A = F_0 \delta(x)\delta(y)\hat{k} \tag{4.18}$$

This represents an infinitely thin solenoid (a magnetic flux filament) carrying a total flux of F_0.

To understand why the z-axis must be 'punctured', consider the difference between a ball and a torus. In a ball (simply connected), any loop of string can be shrunk to a point. In a torus (multiply connected), a loop passing through the center hole cannot be shrunk to a point without leaving the surface. In the AB effect, the magnetic flux inside the solenoid acts as the 'hole'. Even if the particle never touches the flux, the fact that its path encircles the hole makes the interaction 'global' and topologically distinct from a path in a field-free, simply connected space.

4.1.3 Gauge Transformation and Phase Shift

Let's see how this result changes if we apply a gauge transformation. Transform the vector potential A into A' by adding the gradient of a scalar field $\Lambda(x, y, z)$:

$$A' = A + \nabla\Lambda \tag{4.19}$$

Applying the curl operator to both sides:

$$\nabla \times A' = \nabla \times A + \nabla \times (\nabla\Lambda) \tag{4.20}$$

Using the vector identity $\nabla \times (\nabla\Lambda) = 0$, we find:

$$\nabla \times A' = \nabla \times A = F_0\delta(x)\delta(y)\hat{k} \tag{4.21}$$

This demonstrates that the magnetic field is a gauge-invariant quantity. While A changes depending on the gauge, the physical observable—the magnetic flux concentrated on the z-axis—remains exactly the same.

Under a gauge transformation, the wavefunction of a particle undergoes a local phase shift to ensure that the physical predictions of the Schrödinger equation remain unchanged. When we change the potentials:

$$A' = A + \nabla\Lambda, \quad \Phi' = \Phi - \frac{\partial\Lambda}{\partial t} \tag{4.22}$$

The Hamiltonian for a particle with charge q becomes:

$$H' = \frac{1}{2m}(-i\hbar\nabla - qA')^2 + q\Phi' \tag{4.23}$$

To keep the physics invariant, we define the new wavefunction Ψ' as:

$$\Psi' = \Psi e^{i\frac{q}{\hbar}\Lambda} \tag{4.24}$$

This is a local phase shift because Λ depends on space-time. In the Aharonov-Bohm effect, this implies that even in a field-free region ($B = 0$), the non-zero potential A shifts the phase of the wavefunction. This shift is physically observable as an interference pattern displacement because the phase depends on the integral of A along the particle's path. While originally demonstrated with electrons, this effect has been precisely verified in modern contexts using ultracold atoms [5].

Let's calculate the total phase difference $\Delta\phi$ for the two paths around the solenoid. As a particle moves through a region in the presence of a vector potential, its wavefunction picks up a phase factor:

$$\phi = \frac{q}{\hbar}\int \mathbf{A} \cdot d\mathbf{l} \tag{4.25}$$

Consider a beam of charged particles split into two paths, path_1 and path_2, both starting at point S and ending at point P, encircling the z-axis as illustrated in Figure 2.

The phase accumulated along each path is:

$$\phi_1 = \frac{q}{\hbar}\int_{\text{path}_1} \mathbf{A} \cdot d\mathbf{l}, \quad \phi_2 = \frac{q}{\hbar}\int_{\text{path}_2} \mathbf{A} \cdot d\mathbf{l} \tag{4.26}$$

The observable phase difference is:

$$\Delta\phi = \phi_1 - \phi_2 = \frac{q}{\hbar}\left(\int_{\text{path}_1} \mathbf{A} \cdot d\mathbf{l} - \int_{\text{path}_2} \mathbf{A} \cdot d\mathbf{l}\right) \tag{4.27}$$

Reversing the direction of path_2 turns the expression into a closed loop integral around the solenoid:

$$\Delta\phi = \frac{q}{\hbar}\oint_\Gamma \mathbf{A} \cdot d\mathbf{l} \tag{4.28}$$

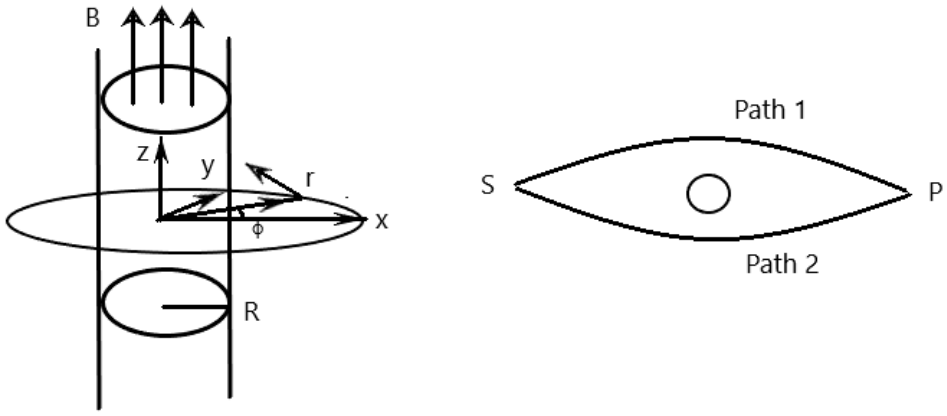


Figure 2: Interference paths encircling a magnetic flux filament. This diagram shows a particle beam splitting at source S into two paths that encircle a central flux filament before recombining at point P. It illustrates how the vector potential $\mathbf{A}$ induces a phase shift even in regions where the magnetic field $\mathbf{B}$ is zero.

Using Stokes' theorem, we know that the loop integral is the total magnetic flux enclosed by the paths. Thus, the total phase shift is:

$$\Delta\phi = \frac{q}{\hbar} F_0 \tag{4.29}$$

This result is extraordinary because it shows that the phase shift depends only on the total flux enclosed, even though the charged particle never passes through the region where the magnetic field is non-zero.

4.1.4 Centrifugal Barrier and Energy Spectrum

Now, suppose there is a repulsive barrier around the z-axis, so we do not have to worry about the singular nature of the vector potential there. The centrifugal repulsive barrier is an effective potential energy term that arises when we transform the kinetic energy from Cartesian into curvilinear cylindrical coordinates. Classically, for a particle with mass m and angular momentum L, the kinetic energy associated with its rotation is:

$$K_{\text{rot}} = \frac{L^2}{2mr^2} \tag{4.30}$$

As the particle approaches the center ($r \to 0$), the energy required to maintain that angular momentum increases quadratically, pushing the particle away from the axis.

In QM, we replace the classical L with the operator $\hat{L}_z = -i\hbar\frac{\partial}{\partial\phi}$. When acting on a wavefunction with angular dependence $e^{in\phi}$, the radial Schrödinger equation sees a term:

$$V_{\text{centrifugal}}(r) = \frac{n^2\hbar^2}{2mr^2} \tag{4.31}$$

This term prevents particles with non-zero angular momentum ($n \neq 0$) from reaching the origin, as the cost in kinetic energy becomes infinite.

However, in the presence of the vector potential $\mathbf{A}$, the kinetic momentum is shifted. The barrier is no longer governed by the canonical angular momentum $n\hbar$, but by the shifted mechanical angular momentum $(n\hbar - \alpha)$; hence, we obtain the effective potential as:

$$V_{\text{eff}} = \frac{(n\hbar - \alpha)^2}{2mr^2} + V(r) \tag{4.32}$$

where $\alpha = \frac{qF_0}{2\pi}$ is the flux coupling constant.

This is a sophisticated way to view the Aharonov-Bohm (AB) effect by observing how the hidden flux modifies the centrifugal barrier. We can write the vector potential in cylindrical coordinates as $\mathbf{A} = \frac{F_0}{2\pi r}\hat{\phi}$. The general vector potential for a solenoid with any flux F_0 in Cartesian components is:

$$\mathbf{A} = \frac{F_0}{2\pi}\left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}, 0\right) \tag{4.33}$$

In cylindrical coordinates (r, ϕ, z), assuming the motion is confined to the x-y plane ($p_z = 0$), the operator $(-i\hbar\nabla - q\mathbf{A})$ in the $\hat{\phi}$ direction is:

$$\hat{p}_\phi = -\frac{i\hbar}{r}\frac{\partial}{\partial\phi} - \frac{qF_0}{2\pi r} \tag{4.34}$$

We apply the Hamiltonian to a wavefunction of the form $\Psi(r, \phi) = R(r)e^{in\phi}$, where n is an integer to ensure the wavefunction is single-valued, $\Psi(\phi) = \Psi(\phi + 2\pi)$. The kinetic energy term $\frac{\hat{p}_\phi^2}{2m}$ acting on the angular part yields:

$$\frac{1}{2mr^2}\left(-i\hbar\frac{\partial}{\partial\phi} - \frac{qF_0}{2\pi}\right)^2 e^{in\phi} = \frac{1}{2mr^2}\left(n\hbar - \frac{qF_0}{2\pi}\right)^2 e^{in\phi} \tag{4.35}$$

Thus, we see that $V_{\text{centrifugal}} \propto \frac{(n\hbar-\alpha)^2}{r^2}$. This treatment reveals rich physics: while the canonical angular momentum is still $n\hbar$, the mechanical angular momentum is $(n\hbar - \alpha)$. Since α can be any real number, the particle behaves as if it has fractional angular momentum.

A profound question arises: how is the axially symmetric potential $\mathbf{A} = \frac{F_0}{2\pi r}\hat{\phi}$ consistent with perturbing the angular momentum? This is resolved by the distinction between canonical and kinetic angular momentum. Because the Hamiltonian is rotationally invariant ($[\hat{H}, \hat{L}_z] = 0$), the canonical angular momentum $\hat{L}_z = -i\hbar\frac{\partial}{\partial\phi}$ remains a conserved quantity with integer eigenvalues $n\hbar$. However, the kinetic angular momentum—the quantity associated with the particle's actual velocity—is shifted:

$$\hat{\Pi}_z = \mathbf{r} \times (\mathbf{p} - q\mathbf{A}) = -i\hbar\frac{\partial}{\partial\phi} - \frac{qF_0}{2\pi} \tag{4.36}$$

The vector potential adds a "background swirl" to the momentum operator. The axial symmetry ensures the shift is uniform for all ϕ, allowing n to remain a good quantum number while shifting the physical angular momentum from its vacuum value.

To find the energy spectrum, we evaluate the Hamiltonian for a particle constrained to a circular path of radius R. Assuming the repulsive barrier $V(r)$ restricts the particle to $r = R$, the Hamiltonian reduces to:

$$\hat{H} = \frac{1}{2mR^2}\left(-i\hbar\frac{\partial}{\partial\phi} - \frac{qF_0}{2\pi}\right)^2 \tag{4.37}$$

Using $\Psi_n(\phi) = \frac{1}{\sqrt{2\pi}}e^{in\phi}$, and defining the magnetic flux quantum $F_q = \frac{h}{q} = \frac{2\pi\hbar}{q}$, we obtain the energy spectrum:

$$E_n = \frac{\hbar^2}{2mR^2}\left(n - \frac{F_0}{F_q}\right)^2 \tag{4.38}$$

This spectrum is periodic in F_0 with period F_q. Increasing the flux by F_q shifts the state n into state $n-1$. At zero flux, $E_n = E_{-n}$, preserving the degeneracy between clockwise and counter-clockwise rotations.

4.2 Toroid

For completeness, let us discuss the same features for a toroid. If we merge the two open ends of a solenoid, we obtain a toroid. We know that the field inside an ideal solenoid is uniform and given by $B = \mu_0 nI$, where n is the number of turns per unit length and I is the current. Similarly, for a toroid with current I, turn density n, and mean radius R, we can calculate the uniform magnetic field inside the core.

4.2.1 Sigmoid Feature of Vector Potential

We consider a circular Amperian loop of radius r inside the toroid, concentric with the center. Due to axial symmetry, the magnetic field $\mathbf{B}$ must be tangential to this loop with a constant magnitude at a fixed radius. The total enclosed current I_{enc} depends on the total number of turns N. Given the mean circumference $L = 2\pi R$, the total number of turns is $N = nL = 2\pi Rn$. Since each turn carries current I, the total enclosed current is $I_{\text{enc}} = NI = 2\pi RnI$. Applying Ampère's law, $\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I_{\text{enc}}$, for $r = R$ we obtain:

$$B(2\pi R) = \mu_0(2\pi RnI) \implies B = \mu_0 nI \tag{4.39}$$

$$B_0 = \mu_0 nI \tag{4.40}$$

For a given turn density n and current I, we define this fixed interior field as B_0. Consequently, the total trapped flux is F_0, allowing for a unified description consistent with the solenoid case.

Thus, the toroid contains the same magnetic field magnitude as the solenoid. However, while the solenoid field is axial, the toroid field is azimuthal (in the $\hat{\phi}$ direction) and entirely trapped within the device. This makes the toroid the "gold standard" for testing the AB effect; unlike the solenoid, which has fringe fields at the ends, the toroid's closed geometry ensures the magnetic field is essentially zero everywhere outside. The non-zero vector potential outside provides the cleanest proof that the phase shift is a result of the potential itself.

For a toroid centered at the origin in the x-y plane, symmetry dictates that the vector potential $\mathbf{A}$ outside must also be azimuthal to ensure its curl is zero while maintaining non-zero circulation. Assuming $\mathbf{A} = A(r)\hat{\phi}$, we use Stokes' theorem for a circular path of radius r:

$$\oint \mathbf{A} \cdot d\mathbf{l} = \int_0^{2\pi} A(r) r d\phi = 2\pi r A(r) = F_0 \tag{4.41}$$

Hence, $\mathbf{A} = \frac{F_0}{2\pi r}\hat{\phi}$. In Cartesian components, using $\hat{\phi} = -\sin\phi\hat{i} + \cos\phi\hat{j}$ and $r^2 = x^2 + y^2$, the vector potential for a toroid with internal flux F_0 is:

$$A_x = -\frac{F_0}{2\pi}\frac{y}{x^2+y^2}, \quad A_y = \frac{F_0}{2\pi}\frac{x}{x^2+y^2}, \quad A_z = 0 \tag{4.42}$$

Let's describe an ideal toroid with mean radius R, cross-sectional radius a, current flowing through the windings I with a number density n. Using Ampere's circuital law, the magnetic field $\mathbf{B}$ inside the toroid at a distance r from the center is directed azimuthally $(\hat{\phi})$ and is given by:

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 I_{\text{enc}} \tag{4.43}$$

$$B(2\pi r) = \mu_0 n(2\pi R)I \implies B = \frac{\mu_0 nIR}{r} \tag{4.44}$$

Magnetic field is zero everywhere outside the core. A toroid divides the space in three distinct regions, the hole, the core and the outside. Magnetic flux $F = \int \mathbf{B} \cdot d\mathbf{s}$ is connected to the vector potential via Stokes' theorem so we can calculate the vector potential $\mathbf{A}$ once we have the magnetic flux, $\oint \mathbf{A} \cdot d\mathbf{l} = F$.

Inside the hole $(r < R - a)$ The flux is zero, because there is no magnetic field present.

$$F = \int \mathbf{B} \cdot d\mathbf{s} = 0 \implies \oint \mathbf{A} \cdot d\mathbf{l} = A(2\pi r) = 0 \implies A = 0 \tag{4.45}$$

Outside the core $(r \geq R + a)$ The flux is the integral of B over the enclosed area. The element of area at a distance r from the center is $ds = 2\sqrt{a^2 - (r-R)^2}dr$, and the field is $B(r) = \frac{\mu_0 nIR}{r}$. Thus the total flux enclosed is:

$$F_0 = \int_{R-a}^{R+a} \frac{\mu_0 nIR}{r} \left(2\sqrt{a^2 - (r-R)^2}\right) dr = 2\mu_0 nIR \int_{R-a}^{R+a} \frac{\sqrt{a^2 - (r-R)^2}}{r} dr \tag{4.46}$$

On evaluating the integral we get $F_0 = 2\pi\mu_0 nIR(R - \sqrt{R^2 - a^2})$. As total flux enclosed by the toroid core is F_0, the vector potential A relates to flux via $\oint \mathbf{A} \cdot d\mathbf{l} = F$. Hence:

$$A = \frac{F_0}{2\pi r} = \frac{\mu_0 nIR}{r}(R - \sqrt{R^2 - a^2})\hat{\phi} \tag{4.47}$$

For the thin shell approximation $(R \gg a)$:

$$A = \frac{\mu_0 nIR^2}{r}\left(1 - \left(1 - \frac{a^2}{R^2}\right)^{1/2}\right) \tag{4.48}$$

Using Binomial approximation:

$$A = \frac{\mu_0 nIa^2}{2r} = \frac{(\mu_0 nI)\pi a^2}{2\pi r} \tag{4.49}$$

Set total flux enclosed $F_0 = (\mu_0 nI)(\pi a^2)$. Thus we obtain $A = \frac{F_0}{2\pi r}\hat{\phi}$.

Inside the core $(R - a \leq r \leq R + a)$ In this region, a circular path of radius r encloses only a partial section of the flux. We integrate from the inner edge $(R - a)$ upto the present position r:

$$F(r) = \int_{R-a}^{r} \frac{\mu_0 nIR}{r'} \left(2\sqrt{a^2 - (r'-R)^2}\right) dr' \tag{4.50}$$

Again by Stokes's law we get $A(2\pi r) = F(r)$:

$$A(r) = \frac{\mu_0 nIR}{\pi r}\left[\int_{R-a}^{r} \frac{\sqrt{a^2 - (r'-R)^2}}{r'} dr'\right]\hat{\phi} \tag{4.51}$$

Let's see how the vector potential within the core is transiting smoothly from 0 value of the hole region to $\frac{F_0}{2\pi r}$ value of the outside region.

This transition occurs because the enclosed flux $F(r)$ is a continuous function of the radial distance r. As we move through the core, our path of radius r captures an increasing slice of the toroid's cross-sectional area as depicted in Figure 3.

To check the continuity of $A(r)$, we have $r = R - a$ at the inner edge. Thus the integration interval becomes $[R-a, R-a]$. The area enclosed is zero, so $F(R-a) = 0$. Thus $A(R-a) = 0$, matching the value with the hole region. At the outer edge we have $r = R+a$. The integration covers the entire cross-section $[R-a, R+a]$. Thus the flux enclosed is the total flux F_0. Thus $A = \frac{F_0}{2\pi(R+a)}$, which matches for the outside region for $r = R + a$.

To check the differentiability of the vector potential we ensure that there are no kinks on the boundaries. The transition is smooth because the integrand contains the term $\sqrt{a^2 - (r-R)^2}$. We know by the definition of flux that $\frac{dF}{dr} = B(r) \times \text{Height}(r)$. At the boundaries $r = R \mp a$, thus the height of the circular cross-section is exactly 0. This means that $\frac{dA}{dr}$ is continuous at the boundaries, ensuring there are no kinks in the vector potential as we enter or leave the core region.

In the thin-shell limit $(R \gg a)$, the growth inside the core is proportional to the area of a circular segment. If we define a normalized coordinate $u = \frac{r-R}{a}$, then limits $[R-a, R+a]$ will change to $[-1, +1]$. And then the vector potential inside the core will scale as:

$$A(u) \propto \frac{1}{2} + \frac{1}{\pi}(u\sqrt{1-u^2} + \arcsin u) \tag{4.52}$$

We can see that at $u = -1$ (inner edge), the term equals to 0, and at $u = +1$ (outer edge), the term equals 1. This $\arcsin u$ behavior is what creates that smooth and elegant sigmoid transition from the hole to outer region for the vector potential.

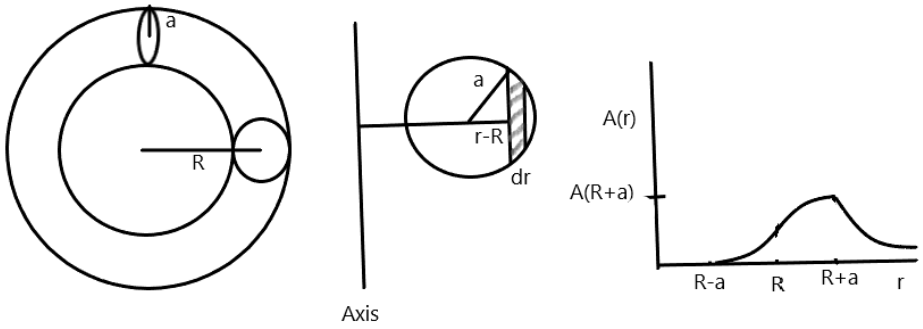


Figure 3: Radial transition of the vector potential $A(r)$. The plot demonstrates the continuous and differentiable sigmoid growth of the potential as it transitions from zero in the hole to its $1/r$ decay in the outside region, following the arcsin behavior in the core.

4.2.2 Phase Shift

In this setup, space is divided into three regions: the hole ($B = 0, A = 0$), the core ($B \neq 0, A \neq 0$), and the outside ($B = 0, A \neq 0$). To verify the AB effect, we utilize two paths—one passing through the hole and the other passing outside the toroid—ensuring the particle never overlaps with the $\mathbf{B}$ field as illustrated in Figure 4.

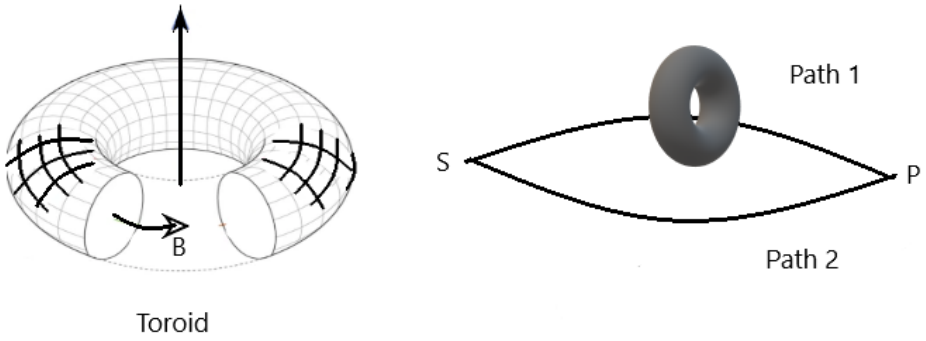


Figure 4: Interference paths around a toroidal magnetic flux. This setup shows paths passing through the center and outside of a toroid. Since the magnetic field is entirely trapped within the core, it proves the Aharonov-Bohm effect by showing the phase shift depends solely on the enclosed flux F_0.

The total wavefunction at the detection point P is the superposition $\Psi = \Psi_1 + \Psi_2$. The relative phase difference $\Delta\phi = \phi_1 - \phi_2$ is the observable that shifts the interference pattern. The phase accumulated along a path is:

$$\phi = \frac{q}{\hbar} \int_{\text{path}} \mathbf{A} \cdot d\mathbf{l} \tag{4.53}$$

The total phase shift is then represented as a closed loop integral Γ enclosing the toroid:

$$\Delta\phi = \frac{q}{\hbar} \oint_{\Gamma} \mathbf{A} \cdot d\mathbf{l} = \frac{q}{\hbar} \iint_{S(\Gamma)} \mathbf{B} \cdot d\mathbf{s} = \frac{qF_0}{\hbar} \tag{4.54}$$

This confirms that the phase shift depends solely on the total enclosed flux F_0, occurring even when the particle never enters the region where $\mathbf{B} \neq 0$. This serves as fundamental evidence that the vector potential $\mathbf{A}$ is a physical entity in quantum mechanics.

4.2.3 Angular Momentum and Energy Spectrum

When we make transition from the classical vector potential of a toroid to the quantum mechanical description of a charged particle interacting with it, we enter into the realm of Aharonov-Bohm effect. Because the magnetic field $\mathbf{B}$ is confined strictly within the core, a charged particle either in the hole or outside regions moves in a field free space ($\mathbf{B} = 0$) but feels the influence of the non-zero vector potential $\mathbf{A}$.

The Hamiltonian for a particle of mass m and charge q in the presence of a vector potential is $H = \frac{1}{2m}(\mathbf{p} - q\mathbf{A})^2$. For a particle constrained to a circular path of radius r in the outside region, where $\mathbf{A} = \frac{F_0}{2\pi r}\hat{\phi}$, the Hamiltonian in cylindrical coordinates becomes:

$$H = -\frac{\hbar^2}{2mr^2}\left(\frac{\partial}{\partial\phi} - i\frac{qF_0}{2\pi\hbar}\right)^2 \tag{4.55}$$

The angular momentum operator in the z direction is $\hat{L}_z = -i\hbar\frac{\partial}{\partial\phi}$. The wavefunction must satisfy the periodicity condition $\Psi(\phi) = \Psi(\phi + 2\pi)$ which leads to the standard integer quantization of the orbital angular momentum $\hat{L}_z\Psi = n\hbar\Psi$, with $n = 0, \pm1, \pm2, \ldots$. The eigenvalues of the mechanical angular momentum (which corresponds to actual velocity) are shifted by an amount $\Delta l_n = \hbar\left(n - \frac{qF_0}{2\pi\hbar}\right)$. Defining the magnetic flux quantum as $F_q = \frac{h}{q}$, the shift is written as $\hbar\left(n - \frac{F_0}{F_q}\right)$.

By substituting the angular momentum eigenvalues into the Hamiltonian, we obtain the energy levels for a particle at a fixed radius r:

$$E_n = \frac{\hbar^2}{2mr^2}\left(n - \frac{F_0}{F_q}\right)^2 \tag{4.56}$$

We infer several features from these results. The energy levels are periodic functions of the total enclosed flux F_0. If the flux is an integer multiple of the flux quantum F_q, the spectrum is identical to the free particle scenario. In a zero-field case n and $-n$ have the same energy. The toroid's vector potential breaks this symmetry unless F_0 is a half integer multiple of F_q. Since the ground state energy depends on the enclosed flux F_0, a particle in the ring will have a non-zero velocity even in the lowest energy state, creating a persistent current.

The AB effect and the use of toroid as an experimental setup have profound philosophical implications concerning the ontology (what is real) and epistemology (how we know it) of fundamental physics. We will touch upon several points hereunder.

4.2.4 Topological Invariance and the Field-Free Interaction

Having analyzed the toroid, we can now resolve the ontological conflict between fields $(\mathbf{E}, \mathbf{B})$ and potentials $(\Phi, \mathbf{A})$. The observable effect (a shift in the interference pattern) due to change in quantum phase even in the field free regions suggests that the vector potential has a signature of reality that transcends its classical role as a mathematical artifact.

Let's define two fundamental ideas in physics. Locality suggests that there is no action at a distance possible and thus physical effects should be caused by local interactions. Gauge Invariance requires that the physical predictions of a theory should not depend on the arbitrary choice of a mathematical gauge. The AB effect presents a conflict between these two ideas. Firstly, if the magnetic field is the only real entity, then its influence must be propagated non-locally to the charged particle passing outside the toroid. And secondly, if the vector potential is the real and local influence, its specific value at a point is not gauge-invariant as it can be changed arbitrarily, which violates the gauge principle for local quantities.

The resolution comes through the interpretation that the AB effect highlights the importance of the gauge potential as a more complete description of EM than the fields alone. The observable outcome of the experiment - the overall phase difference around a closed loop - is a gauge invariant quantity directly related to the total magnetic flux, $\Delta\phi = \frac{qF_0}{\hbar}$. While a specific value of $\mathbf{A}$ at any single point is not physical, the mathematical structure it defines (especially, its holonomy or circulation around a non-contractible loop) is a fundamental observable physical property linked to the topology of the space around the toroid. Attempts to explain the AB effect purely through

non-local forces or interaction energies without potentials have been largely unsuccessful in fully accounting for experimental observations, reinforcing the potential-centric view.

The use of a toroid is crucial because, unlike an infinitely long solenoid, it completely traps the magnetic field inside, creating a truly field free external region for the particles. This setup provides the cleanest empirical evidence that the potentials, not just the fields, are physically significant in QM.

Gauge theory and topology provides the mathematical framework that underpins the AB effect and the physical reality of the vector potential. Gauge theory is built on the concept of symmetry. It suggests that the laws of physics should remain unchanged (invariant) even when certain mathematical descriptions (like potentials or wavefunctions) are transformed locally. In QM, we are free to change the overall phase of a wavefunction arbitrarily without affecting physical quantities like probability density $|\Psi|^2$. This is a global symmetry. But a problem arises if we make the phase angle θ a function of position and time, $\Psi(x) \rightarrow e^{i\theta(x)}\Psi(x)$, the standard Schrödinger equation changes in a way that is not simply a phase factor. The solution to this problem is given by the gauge potential. To restore this local symmetry, we must introduce a new field - the vector potential. This means that a local phase transformation in QM necessitates the idea of a vector potential in EM. The vector potential must simultaneously transform as $\mathbf{A} \rightarrow \mathbf{A}' = \mathbf{A} + \nabla\Lambda$, where Λ is the gauge function related to the phase angle by $\Lambda(x) = \frac{\theta(x)}{q}$. This ensures the Schrödinger equation remains invariant (a symmetry), mapping one valid physical solution to another. This deep connection between local gauge symmetry and the existence of force-carrying potentials is one of the most elegant links between EM and QM.

Topology is the mathematical study of shapes and spaces that remain the same under continuous deformations. In the context of the AB effect, it is essential for understanding why a zero-field region can still produce an effect. The setup for the AB effect (solenoid or toroid) requires the physical space where the particle travels to be punctured or multiply connected. The region of space is not simply connected because it has a hole (the core containing the magnetic field) that a closed loop cannot be shrunk to a single point without crossing a singularity. The vector potential itself is well behaved and curl-free in the reachable region, but it is singular (ill-behaved) precisely on the central axis where the flux filament is located (solenoid). Even if the magnetic field ($\mathbf{B} = \nabla \times \mathbf{A}$) is zero everywhere along the particle's path, the integral of the vector potential around a closed loop enclosing the flux ($\oint \mathbf{A} \cdot d\mathbf{l}$) is non-zero. Stokes' theorem links this circulation to the total flux (F_0) enclosed by the loop. This loop integral is a topological invariant - it doesn't depend on the specific path taken around the core, only on the fact that it encircles the core once. This invariant quantity is what determines the observable phase shift $\Delta\phi = \frac{qF_0}{\hbar}$. Thus, the topology of the experimental setup and the singular nature of the potential at the core are the mathematical requirements for the AB phase shift to exist, confirming that non-local topology plays a critical role in local quantum mechanics.

5 Magnetic Monopole

To model a magnetic monopole of charge g, we consider a semi-infinite flux filament (the string) starting at origin and extending upto $(-\infty)$ along the negative z axis. The end of this string at the origin acts as a source of magnetic flux, appearing as a magnetic monopole to the rest of the space.

5.1 Vector Potentials and Wu-Yang Construction

We have already seen the making of a flux filament from a long solenoid of radius R with internal field $B_0\hat{k}$ and the vector potential in the exterior region ($r \geq R$) is $A = \frac{B_0R^2}{2r}\hat{\phi}$. We define the total magnetic flux as $F_0 = B_0\pi R^2$. We then take the limit $R \rightarrow 0$ while keeping the total flux F_0 constant. Substituting $B_0R^2 = \frac{F_0}{\pi}$ into the exterior vector potential we get $A = \frac{F_0}{2\pi r}\hat{\phi}$. In cartesian coordinates $A_x = -\frac{F_0}{2\pi}\frac{y}{x^2+y^2}$, $A_y = \frac{F_0}{2\pi}\frac{x}{x^2+y^2}$, $A_z = 0$. The magnetic field is $\mathbf{B} = \nabla \times \mathbf{A}$. Outside the z axis, $\nabla \times \mathbf{A} = 0$. However, at the origin, the curl is a Dirac delta function $\mathbf{B} = F_0\delta(x)\delta(y)\hat{k}$. If this filament were semi-infinite (ending at the origin), the leakage of flux F_0 at the endpoint represents a magnetic monopole of strength g related to the flux via Gauss's Law as shown in Figure 5 following the global formulation of gauge fields proposed by Wu and Yang [6].

The total flux F_0 emanating from the end of the string must satisfy Gauss' law for magnetism $\oint \mathbf{B} \cdot d\mathbf{s} = 4\pi g$. Hence we would get $F_0 = 4\pi g$. Note that in the SI units we write $\oint \mathbf{B} \cdot d\mathbf{s} = \mu_0 g$. We will use Gaussian units

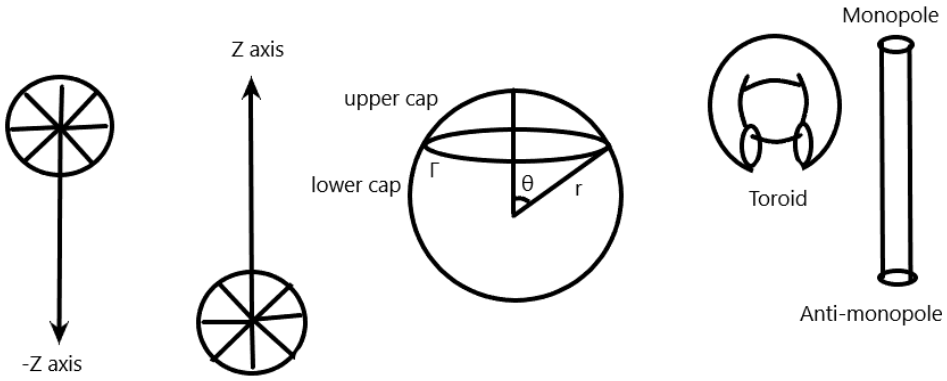


Figure 5: The Dirac string and magnetic monopole. A semi-infinite solenoid (the string) feeds magnetic flux $F_0 = 4\pi g$ into the origin. The endpoint acts as a point source for a radial magnetic field, effectively creating a magnetic monopole g while maintaining a singular vector potential along the negative z-axis.

for our discussion here. This flux is fed into the origin by the semi-infinite filament. For a semi-infinite string on the negative z axis, we require a potential $\mathbf{A}$ whose curl gives the field of a point monopole $\mathbf{B} = \frac{g}{r^2}\hat{r}$, everywhere except the string. Thus a magnetic monopole of charge g at the origin produces a radial magnetic field in the space.

For finding the vector potential we choose a circular path Γ on a sphere of radius r, centered on the z axis at a polar angle θ. Assuming azimuthal symmetry $\mathbf{A} = A_\phi\hat{\phi}$, the integral around the loop of radius $r\sin\theta$ is $\oint_\Gamma \mathbf{A}\cdot d\mathbf{l} = A_\phi(2\pi r\sin\theta)$. By Stokes' theorem, this must be equal to the magnetic flux F passing through the spherical cap S bounded by Γ:

$$F = \int_0^\theta \int_0^{2\pi} (B_r\hat{r})\cdot(r^2\sin\theta' d\theta' d\phi')\hat{r} \tag{5.1}$$

Substituting $B_r = \frac{g}{r^2}$:

$$F = \int_0^\theta \int_0^{2\pi} \left(\frac{g}{r^2}\right) r^2\sin\theta' d\theta' d\phi' = 2\pi g\int_0^\theta \sin\theta' d\theta' = 2\pi g(1-\cos\theta) \tag{5.2}$$

Equating the path integral to the flux $A_\phi(2\pi r\sin\theta) = 2\pi g(1-\cos\theta)$, we get:

$$A_\phi = \frac{g(1-\cos\theta)}{r\sin\theta} \tag{5.3}$$

At $\theta = 0$ (North pole), we have a form of $0/0$. Using L'Hôpital's rule, $\lim_{\theta\to 0}\frac{g\sin\theta}{r\cos\theta} = 0$. Hence $A_\phi \to 0$. Thus the potential is well behaved at the north pole.

At $\theta = \pi$ (South pole), we have denominator $\sin\pi = 0$, while the numerator $1-\cos\pi = 2$. Thus $A_\phi \to \infty$. This singularity represents the semi-infinite flux filament located on the negative z axis, which punctures the space. We can infer that the potential is singular along the path that carries the flux to the monopole. This property makes the space multiply connected allowing the AB phase shift to exist even where the magnetic field is zero.

Let's see how this result changes if we place the string on the positive z axis instead. To move the string to the positive z axis, we change our integration surface to the lower spherical cap (from θ to π). This feeds the flux of monopole from the north instead of south. We apply Stokes' theorem to the same circular path Γ at an angle θ, but this time we integrate the magnetic flux F' through the surface S' that includes the south pole ($\theta' = \pi$). The flux through the lower cap is:

$$F' = \int_\theta^\pi \int_0^{2\pi} \left(\frac{g}{r^2}\right) r^2\sin\theta' d\theta' d\phi' = 2\pi g[-\cos\theta']_\theta^\pi = -2\pi g(1+\cos\theta) \tag{5.4}$$

The negative sign arises because the boundary Γ is traversed in the opposite direction relative to the outward normal of the lower cap. Equating the path integral to this flux:

$$A'_\phi(2\pi r\sin\theta) = -2\pi g(1+\cos\theta) \implies A'_\phi = -\frac{g(1+\cos\theta)}{r\sin\theta} \tag{5.5}$$

At $\theta = \pi$ (south pole), we have a form of $0/0$. Using L'Hôpital's rule, $\lim_{\theta\to\pi} \frac{g\sin\theta}{r\cos\theta} = 0$. Hence $A_\phi \to 0$. Thus the potential is well behaved at the south pole. At $\theta = 0$ (north pole), we have numerator $\cos 0 = 1$, while the denominator is $\sin 0 = 0$. Thus $A_\phi \to \infty$. So we get a singularity at the north pole. This singularity represents the semi-infinite flux filament located on the positive z axis.

Let's compare the two vector potentials:

- $A_N = \frac{g(1-\cos\theta)}{r\sin\theta}$: Well behaved at $\theta = 0$, Singular at $\theta = \pi$.
- $A_S = -\frac{g(1+\cos\theta)}{r\sin\theta}$: Well behaved at $\theta = \pi$, Singular at $\theta = 0$.

In the overlap region, they should differ only by a gauge transformation:

$$A_N - A_S = \frac{g(1-\cos\theta)}{r\sin\theta}\hat{\phi} + \frac{g(1+\cos\theta)}{r\sin\theta}\hat{\phi} = \frac{2g}{r\sin\theta}\hat{\phi} \tag{5.6}$$

5.2 Dirac Quantization

In spherical coordinates, the gradient of the azimuthal angle ϕ is $\nabla\phi = \frac{1}{r\sin\theta}\hat{\phi}$. Therefore, $A_N = A_S + \nabla(2g\phi)$. This shows that shifting the string is equivalent to a gauge transformation with gauge function $\Lambda = 2g\phi$.

To derive the Dirac quantization condition, we require that the shift between the two potentials be a physically invisible gauge transformation. When we transform the vector potential by $\mathbf{A} \to \mathbf{A} + \nabla\Lambda$, the wavefunction must transform by a local phase shift $\Psi_N = \Psi_S e^{i\frac{q}{\hbar}\Lambda}$. Substituting $\Lambda = 2g\phi$:

$$\Psi_N = \Psi_S e^{i\frac{2qg}{\hbar}\phi} \tag{5.7}$$

For the physics to be consistent, the wavefunction Ψ must be single-valued: $\Psi(\phi + 2\pi) = \Psi(\phi)$. Applying this condition to the phase factor:

$$e^{i\frac{2qg}{\hbar}(\phi+2\pi)} = e^{i\frac{2qg}{\hbar}\phi} \tag{5.8}$$

This holds only if $\frac{2qg}{\hbar}(2\pi) = 2\pi n$, where $n = 0, \pm1, \pm2\ldots$. We arrive at the fundamental result:

$$\frac{2qg}{\hbar} = n \implies qg = \frac{n\hbar}{2} \tag{5.9}$$

In SI units, this is written as $qg = \frac{nh}{4\pi\mu_0}$. The existence of a single magnetic monopole g forces all electric charges q in the universe to be quantized in units of $\hbar/2g$.

For a magnetic monopole to truly appear as a point source, the semi-infinite flux filament (the string) that feeds it must be undetectable by any charged particle. If a particle could 'feel' the string via a phase shift as it passed by, the monopole would fail to be spherically symmetric. Requiring the phase shift around the string to be a multiple of (2π) making the string physically invisible—is precisely what forces the quantization of all electric charges in the universe.

The existence of a single magnetic monopole g forces all electric charges q in the universe to be quantized in units of $\frac{\hbar}{2g}$.

This elegant link mirrors the fact that the vector potential and the local phase of the wavefunction reside on an equal footing, where the topology of space (the string) dictates the allowable values of physical constants. This is what is meant when we say that symmetry dictates the laws of physics.

5.3 Monopole Pairs from Toroid

For the sake of completion, let's develop the idea of a monopole from the toroid. Now we make a transition from the open string of a solenoid to a closed topological defect of a toroid. While a solenoid's filament is semi-infinite, a toroid traps flux in a finite, closed volume, providing a model for quantized fluxoids and monopole-antimonopole pairs.

Outside a toroid with internal flux F_0, the vector potential is $\mathbf{A} = \frac{F_0}{2\pi r}\hat{\phi}$. A particle encircling the toroid picks up a phase shift:

$$\Delta\phi = \frac{qF_0}{\hbar} \tag{5.10}$$

If we imagine a toroid whose cross-sectional radius $a \to 0$ while the flux F_0 remains finite, we create a closed flux filament (a loop of flux). This loop creates a "hole" in space. Any path that encircles this loop is topologically non-trivial. From a distance, a small toroidal loop of flux looks like a magnetic dipole. However, unlike a standard dipole made of two poles, this is a current-free topological structure where the magnetic field is strictly zero everywhere except inside the singular loop.

For the vacuum surrounding the toroid to be physically consistent (i.e., for the vector potential to be "removable" by a gauge transformation everywhere except across the loop surface), the phase shift must be an integer multiple of 2π:

$$\frac{qF_0}{\hbar} = 2\pi n \implies F_0 = n\left(\frac{h}{q}\right) \implies F_0 = nF_q \tag{5.11}$$

This $F_q = \frac{h}{q}$ is the magnetic flux quantum. This implies that in a universe with discrete electric charges, trapped magnetic flux in closed loops (like those in superconductors) must be quantized. The connection to the monopole arises if we cut the toroid and pull the ends apart. By stretching the toroid into a long, thin tube, the two ends act as a monopole-antimonopole pair.

The body of the toroid becomes the Dirac string connecting them. If we pull one end to infinity, we recover the semi-infinite solenoid (the single monopole) as we derived earlier.

While the solenoid models the existence of a monopole, the toroid models the quantization of its flux. The toroid proves that the vector potential is not just a mathematical device but a topological swirl. In a toroidal configuration, the monopole is not a point, but the end point of a quantized flux tube, ensuring that magnetic charge and electric charge are fundamentally linked through the topology of the vector potential.

6 Lattice Gauge Theory

Consider electromagnetic fields in space-time, where the coordinates are denoted by $x = (x^0, \vec{x})$, with $x^0 = t$. The vector potential is part of the four-potential $A = (A^0, \vec{A}) = (\Phi, \vec{A})$, where Φ is the scalar potential.

6.1 Link Variables and Parallel Transport

The electromagnetic field tensor is given by:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{6.1}$$

which is the higher-dimensional generalization of a curl. The components of $F_{\mu\nu}$ are related to the electric and magnetic fields, e.g., $F_{10} = E_x$ and $F_{12} = B_z$.

As discussed earlier, the important physical quantity is the line integral of the potential around any closed loop Γ:

$$F(\Gamma) = \oint_\Gamma A_\mu(x)dx^\mu \tag{6.2}$$

which may be regarded as the generalization of the electromagnetic flux. Note that the flux can also be expressed in terms of the fields and their surface integrals in four dimensions.

If the wavefunction is subjected to a phase transformation:

$$\psi(x) \to \psi'(x) = e^{i\theta(x)}\psi(x) \tag{6.3}$$

then it is clear that one should also transform the potential:

$$A_\mu \to A'_\mu = A_\mu + \frac{1}{q}\partial_\mu \theta(x) \tag{6.4}$$

By introducing the covariant derivative $D_\mu = \partial_\mu - iqA_\mu$, we obtain:

$$D_\mu \psi(x) \to D'_\mu \psi'(x) = e^{i\theta(x)}[D_\mu \psi(x)] \tag{6.5}$$

where D'_μ indicates the derivative is constructed from A'_μ. This local transformation keeps the physics intact. This theory concerns the $U(1)$ symmetry group.

Consider a dynamical field ψ with N internal degrees of freedom, expressed as a column vector $(\psi_1, \psi_2, \ldots, \psi_N)^T$. For indistinguishable states, we transform the field via:

$$\psi(x) \to \psi'(x) = T\psi(x) \tag{6.6}$$

where T is an $N \times N$ matrix element of a Lie group G. If T varies with x, and matrices do not commute ($T_1T_2 \neq T_2T_1$), the theory is non-Abelian. This provides the framework for strong interactions (color charges) [7] and electro-weak unification [8].

While the AB effect is an Abelian phenomenon, where the phase shift is simply a number, nature also employs Non-Abelian symmetries. In these theories (like the Strong Force), the 'phase' is replaced by an internal orientation (like color) that can be rotated. Instead of a single number, the transformation requires $N \times N$ matrices that do not commute. This demonstrates that the 'background swirl' of the potential is a universal structural feature shared by all fundamental forces, not just electromagnetism.

A lattice gauge theory discretizes space-time into points $\{x_i\}$ with links (i, j). In this model, $\psi_i = \psi(x_i)$ and transformations are $\psi_i \to \psi'_i = e^{i\theta_i}\psi_i$. The derivative corresponds to:

$$\psi_j - \psi_i = \int_{x_i}^{x_j} \partial_\mu \psi(x) dx^\mu \tag{6.7}$$

We define the lattice link variable:

$$A_{ij} = \int_{x_i}^{x_j} A_\mu(x) dx^\mu \tag{6.8}$$

and the Wilson variables [9]:

$$U_{ij} = U(x_i, x_j) \equiv e^{iqA_{ij}} \tag{6.9}$$

where $U_{ji} = (U_{ij})^{-1}$. The accompanying transformation for link variables is:

$$A_{ij} \to A_{ij} + \frac{1}{q}(\theta_j - \theta_i) \tag{6.10}$$

or, in terms of Wilson variables, $U_{ij} \to e^{i(\theta_j - \theta_i)}U_{ij}$.

6.2 Wilson Loops and Gauge-Invariant Flux

For a closed loop (a K-gon Γ), the discrete analog of the flux is:

$$F(\Gamma) = \sum_\Gamma A_{i_n i_{n+1}} \tag{6.11}$$

This gauge-invariant quantity is expressed through Wilson variables as:

$$\prod_\Gamma U_{i_n i_{n+1}} = \exp\left(iq \sum_\Gamma A_{i_n i_{n+1}}\right) = \exp(iqF(\Gamma)) \tag{6.12}$$

Note that in the specific case where the lattice loop Γ encircles the solenoid or toroid core, the general flux $F(\Gamma)$ reduces to the quantized total flux F_0 derived in Sections 4 and 5.

Nontrivial physics occurs when $F(\Gamma) \neq 0$. In spatial planes, $F(\Gamma)$ relates to magnetic flux; in planes containing the time direction, it relates to electric flux. This Wilson loop is the discrete counterpart to the holonomy $H(\Gamma)$ discussed in Section 3. It confirms that even in a discretized spacetime, the gauge-invariant flux $F(\Gamma)$ remains the fundamental physical observable, reinforcing the potential-centric view established in the solenoid and toroid cases.

7 Geometric Synthesis between EM and GR

While the preceding analysis established the vector potential A_μ as a geometric connection within a $U(1)$ gauge theory, this interpretation gains its most profound justification through its structural equivalence to General Relativity (GR). By moving from the internal phase space of a particle to the manifold of spacetime itself, we can demonstrate that "force" is a universal manifestation of underlying geometric curvature.

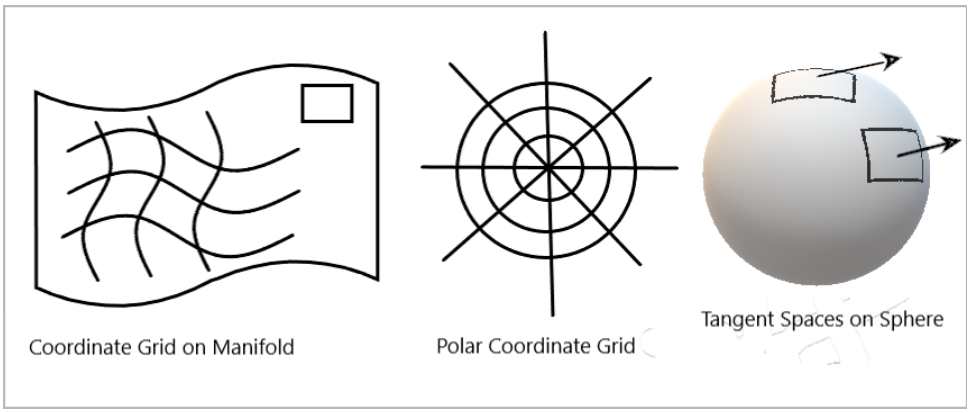


Figure 6: **Geometric Foundations of Curvature.** (Left) A coordinate grid overlaid on a general manifold, illustrating the principle of general covariance. (Center) The equivalence of Cartesian and polar grids on a flat manifold, where the underlying physics remains invariant under coordinate transformation. (Right) Representation of tangent spaces as "private" vector spaces at specific points on a curved manifold, where infinitesimal displacement vectors dx^μ are defined.

7.1 The Manifold and Coordinate Invariance

We define a **manifold** as a topological space that is locally homeomorphic to Euclidean or Minkowski space. On a global scale, the manifold may possess a complex, non-Euclidean topology, yet in a sufficiently small neighborhood, it remains "flat." We can think of it like the Earth: from a satellite, it is a sphere (curved), but to a person standing on a street corner, the ground looks like a flat plane (Euclidean). A Circle locally looks like a straight line. A Sphere locally looks like a flat sheet.

The principle of **General Covariance** dictates that physical laws must remain invariant under coordinate transformations. As seen in Figure [6], a manifold can be described by a variety of grids. For example, a flat 2D plane can be mapped using a Cartesian grid (x, y) or a polar grid (r, θ). While the numerical labels for points change, the physical relationship between them—the geometry—remains invariant.

7.2 The Riemannian Metric and the Concept of Distance

A Riemannian manifold is distinguished by the introduction of a **metric tensor** $g_{\mu\nu}$, which provides a rigorous way to measure distances. For two infinitesimally close points $P(x^\mu)$ and $Q(x^\mu + dx^\mu)$, the square of the interval

ds is defined by a quadratic form:

$$ds^2 = g_{11}dx^1dx^1 + g_{12}dx^1dx^2 + g_{21}dx^2dx^1 + \cdots = g_{\mu\nu}dx^\mu dx^\nu \tag{7.1}$$

The metric $g_{\mu\nu}$ is an $N \times N$ symmetric matrix. In a curved manifold, these coefficients are not constant but vary from point to point. This metric is used to lower indices ($a_\mu = g_{\mu\nu}a^\nu$), while its inverse $g^{\mu\nu}$ is used to raise them ($a^\mu = g^{\mu\nu}a_\nu$).

In Cartesian coordinates (x, y), the metric tensor is the identity matrix:

$$g^{\text{cart}}_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \implies ds^2 = g_{\mu\nu}dx^\mu dx^\nu = dx^2 + dy^2 \tag{7.2}$$

The transformation to Polar coordinates

$$(r, \theta)$$

uses the Jacobian matrix $\mathbf{J}$:

$$\mathbf{J} = \frac{\partial x^i}{\partial x'^j} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \tag{7.3}$$

The metric transforms as a rank-2 covariant tensor:

$$g^{\text{polar}}_{\mu\nu} = \mathbf{J}^T g^{\text{cart}}_{\mu\nu} \mathbf{J} \tag{7.4}$$

Performing the matrix multiplication:

$$g^{\text{polar}}_{\mu\nu} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \tag{7.5}$$

$$g^{\text{polar}}_{\mu\nu} = \begin{pmatrix} \cos^2\theta + \sin^2\theta & -r\sin\theta\cos\theta + r\sin\theta\cos\theta \\ -r\sin\theta\cos\theta + r\sin\theta\cos\theta & r^2\sin^2\theta + r^2\cos^2\theta \end{pmatrix} \tag{7.6}$$

Simplifying via trigonometric identities yields the Polar metric tensor:

$$g^{\text{polar}}_{\mu\nu} = \begin{pmatrix} g_{rr} & g_{r\theta} \\ g_{\theta r} & g_{\theta\theta} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \implies ds^2 = dr^2 + r^2 d\theta^2 \tag{7.7}$$

$g_{rr} = 1$: This tells you that a change in the radial coordinate (dr) translates 1-to-1 into physical distance.

$g_{\theta\theta} = r^2$: This "scales" the angular change. Since a degree of arc is physically longer the further you are from the center ($s = r\theta$), the metric needs that r^2 factor to get the distance right.

$g_{r\theta} = 0$: This tells you the r and θ grid lines are perpendicular (orthogonal). If they weren't, these "off-diagonal" terms would be non-zero.

Even though $dx \neq dr$ and $dy \neq d\theta$, the final physical distance is identical. This is the mathematical proof of Einstein's point: the coordinates are just "scaffolding". The metric tensor automatically adjusts to ensure the physical reality remains constant.

7.3 Tangent Spaces and Basis Vectors

On a curved surface, a finite displacement $x(x^1, x^2, \ldots, x^N)$ is not a vector. However, an infinitesimal displacement $dx(dx^1, dx^2, \ldots, dx^N)$ constitutes a true vector. These vectors do not live on the manifold itself but in a "private" vector space attached to each point, known as the **Tangent Space** (Figure [6], Right).

We define the basis vectors $\vec{e}_\mu$ of the tangent space as the partial derivative operators along the coordinate axes:

$$\vec{e}_\mu = \frac{\partial}{\partial x^\mu} \tag{7.8}$$

To account for the change in both the vector components and the basis vectors, we define the covariant derivative D_ν:

$$D_\nu A^\mu = \partial_\nu A^\mu + \Gamma^\mu_{\rho\nu} A^\rho \tag{7.9}$$

The connection coefficients $\Gamma^\mu_{\rho\nu}$ define how the basis vectors $\mathbf{e}_\rho$ vary across the manifold:

$$\partial_\nu \vec{e}_\rho = \Gamma^\mu_{\rho\nu} \vec{e}_\mu \tag{7.10}$$

The requirement of metric compatibility, $D_\lambda g_{\mu\nu} = 0$ ensures the physics (lengths and angles) remains invariant under transport:

$$\Gamma^\sigma_{\mu\nu} = \frac{1}{2} g^{\sigma\rho} \left(\partial_\nu g_{\rho\mu} + \partial_\mu g_{\rho\nu} - \partial_\rho g_{\mu\nu} \right) \tag{7.11}$$

7.4 Parallel Transport and the Connection

To connect different private vector spaces, we use the **connection 1-form** Γ_μ, which is an $N \times N$ matrix with components $\Gamma^\nu_{\mu\rho}$ known as the **Christoffel symbols**. The index μ describes the direction in which the basis vectors are rotated to align with the next tangent space.

The total parallel transport of a vector over a displacement Δx^μ is then represented by the operator $\exp(\Delta x^\mu D_\mu)$.

To relate the disjoint tangent spaces $T_x M$ and $T_{x+dx} M$, we introduce the connection 1-form Γ_μ, which is an $N \times N$ matrix representing the "rotation" of the basis:

$$(D_\mu \psi)^i = \partial_\mu \psi^i + (\Gamma_\mu)^i{}_k \, \psi^k \tag{7.12}$$

In component form, the connection carries three indices, where μ denotes the direction of displacement:

$$(\Gamma_\mu)^i{}_k = \Gamma^i_{k\mu} \tag{7.13}$$

The transport of a vector ψ along an infinitesimal displacement dx^μ involves the matrix multiplication:

$$\psi^i(x + dx) \approx \psi^i(x) - \Gamma^i_{k\mu} \psi^k dx^\mu \tag{7.14}$$

The matrix Γ_μ describes how the basis vectors $\mathbf{e}_k$ are connected across space:

$$\partial_\mu \vec{e}_k = \Gamma^i_{k\mu} \vec{e}_i \tag{7.15}$$

To connect the vector space $V(x)$ to $V(x + \Delta x)$, we define the parallel transport operator.This operator consists of an ordinary translation and a corrective rotation:

$$\exp(\Delta x^\mu D_\mu) = \exp(\Delta x^\mu \Gamma_\mu) \exp(\Delta x^\mu \partial_\mu) \tag{7.16}$$

Following the "Minimal Coupling Prescription," we define the **Covariant Derivative** D_μ as the sum of the ordinary derivative and the geometric rotation. To first order in Δx, this links the covariant derivative directly to the partial derivative and the connection matrix:

$$D_\mu = \partial_\mu + \Gamma_\mu \tag{7.17}$$

The connection matrix Γ_μ acts on the basis vectors $\vec{e}_j$ such that its components describe the "leakage" into other basis directions:

$$\partial_\mu \vec{e}_j = \Gamma^i_{j\mu} \vec{e}_i \tag{7.18}$$

In this framework, Γ_μ is an $N \times N$ matrix with entries $(\Gamma_\mu)^i{}_j = \Gamma^i_{j\mu}$. The covariant derivative of a vector field ψ is then:

$$D_\mu \psi = (\partial_\mu + \Gamma_\mu)\psi \tag{7.19}$$

7.5 The Non-Commutative Nature of Curvature

Curvature is the result of the non-commutative nature of the covariant derivatives.

$$[D_\mu, D_\nu]\psi = (D_\mu D_\nu - D_\nu D_\mu)\psi \tag{7.20}$$

Expanding $D_\mu D_\nu \psi$:

$$D_\mu(D_\nu \psi) = (\partial_\mu + \Gamma_\mu)(\partial_\nu \psi + \Gamma_\nu \psi) \tag{7.21}$$

$$= \partial_\mu \partial_\nu \psi + (\partial_\mu \Gamma_\nu)\psi + \Gamma_\nu \partial_\mu \psi + \Gamma_\mu \partial_\nu \psi + \Gamma_\mu \Gamma_\nu \psi \tag{7.22}$$

When we subtract the reverse order ($D_\nu D_\mu \psi$), the symmetric terms $\partial_\mu \partial_\nu \psi$ and the mixed terms like $\Gamma_\nu \partial_\mu \psi$ cancel out, leaving:

$$R_{\mu\nu} = \partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu + [\Gamma_\mu, \Gamma_\nu] \tag{7.23}$$

In EM, because the $U(1)$ group is abelian, the commutator $[\Gamma_\mu, \Gamma_\nu]$ vanishes, reducing to the Faraday Tensor, simplifying to the standard field strength. In GR, the non-abelian nature of the connection accounts for the self-interaction of the gravitational field.

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{7.24}$$

7.6 Dynamics and the Unified Force Equation

For a free particle in GR, the linear momentum does not change ($d\vec{p} = 0$). Starting from the conservation of the momentum vector $\vec{p} = p^\mu \vec{e}_\mu$:

$$d(p^\mu \vec{e}_\mu) = (dp^\mu)\vec{e}_\mu + p^\mu (d\vec{e}_\mu) = 0 \tag{7.25}$$

Using the definition of the connection $\partial_\nu \vec{e}_\rho = \Gamma^\mu_{\rho\nu} \vec{e}_\mu$, we expand the second term:

$$(dp^\mu)\vec{e}_\mu + p^\rho \Gamma^\mu_{\rho\nu} dx^\nu \vec{e}_\mu = 0 \tag{7.26}$$

Factoring out the basis $\vec{e}_\mu$ and dividing by the proper time $d\tau$, we use $p^\mu = m\frac{dx^\mu}{d\tau}$:

$$m\left(\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\rho\nu}\frac{dx^\rho}{d\tau}\frac{dx^\nu}{d\tau}\right) = 0 \tag{7.27}$$

If the particle carries a charge q, it couples to the electromagnetic curvature $F^\mu{}_\nu$. The "zero" on the right-hand side is replaced by the Lorentz force, yielding the Unified Equation:

$$m\left(\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\rho\nu}\frac{dx^\rho}{d\tau}\frac{dx^\nu}{d\tau}\right) = qF^\mu{}_\nu \frac{dx^\nu}{d\tau} \tag{7.28}$$

7.7 The Master Comparison: EM vs. GR

The following table summarizes the structural parallels and functional differences between the geometric descriptions of Electromagnetism and General Relativity.

Table 1: Geometric Synthesis: General Relativity (GR) vs. Electromagnetism (EM)

Feature	**General Relativity (GR)**	**Electromagnetism (EM)**
Mathematical Space	Tangent Space (Spacetime Manifold)	$U(1)$ Fiber Bundle (Internal Phase)
Geometric Connection	Christoffel Symbol ($\Gamma^{\nu}_{\mu\rho}$)	Gauge Potential (iqA_μ)
Covariant Derivative	$D_\mu = \partial_\mu + \Gamma_\mu$	$D_\mu = \partial_\mu + iqA_\mu$
Curvature / Field	Riemann Tensor ($R^{\rho}_{\sigma\mu\nu}$)	Faraday Tensor ($F_{\mu\nu}$)
Independent Unknowns	**10 components** (Symmetric $g_{\mu\nu}$)	**4 components** (Four-vector A_μ)
Field Dynamics	**Non-linear, 2nd Order** (Self-sourcing)	**Linear, 2nd Order** (Wave-like)
Commutator	$[D_\mu, D_\nu] = \mathcal{R}_{\mu\nu}$	$[D_\mu, D_\nu] = iqF_{\mu\nu}$
Non-Trivial Transport	Parallel Transport (Vector Rotation)	Holonomy (Phase Shift)
Physical Manifestation	Geodesic Deviation / Gravity	Aharonov-Bohm / Lorentz Force
The "Structural Wall"	**Curvature** $\sim$ **Source** ($G \sim T$)	∂ **of Curvature** $\sim$ **Source** ($\partial F \sim J$)
Quantization Link	Mass-Inertia Equivalence	Dirac String / Charge Quantization

7.8 Field Dynamics and the Structural Wall

Despite these kinematic parallels, the field dynamics reveal a fundamental divergence. Maxwell's equations relate the *derivative* of the curvature to the source ($\partial F \sim J$):

At this level, the theories differ significantly. While EM treats the connection as a field inhabiting spacetime, GR treats the connection as the definition of spacetime itself. This "wall" suggests that while the geometric language is shared, the dynamical origin of gravity is zero-order in curvature, whereas electromagnetism is first-order.

The dynamics of the Electromagnetic field are governed by the derivative of the Faraday tensor:

$$D_\mu F^{\mu\nu} = \partial_\mu F^{\mu\nu} + \cdots = J^\nu \tag{7.29}$$

Note: Since $F \sim \partial A$, this is a second-order equation in terms of the potential A_μ.

In contrast, General Relativity relates the curvature tensor directly to the energy-momentum source:

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu} \tag{7.30}$$

Note: Since $R \sim \partial\Gamma + \Gamma^2$ and $\Gamma \sim \partial g$, this is also a second-order equation, but it links the *geometry itself* to the source.

While the Einstein equation (7.30) appears compact, it represents a **coupled set of 10 non-linear second-order partial differential equations**. This complexity arises because the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}$ is a 4×4 symmetric matrix; since the indices μ and ν each range from 0 to 3, the 16 total components reduce to **10 independent entries** (4 diagonal and 6 off-diagonal). Furthermore, because the Christoffel symbols involve first derivatives of the metric and the Riemann curvature tensor involves derivatives of those symbols, the equations are inherently second-order. The non-linearity is a reflection of the gravitational field's self-interaction—a stark contrast to the Abelian, linear framework of Maxwellian electromagnetism.

In contrast to the 10 non-linear second-order equations of General Relativity, Maxwell's equations (7.29) represent a **linear system of 4 coupled partial differential equations** in terms of the fields, or 2 second-order equations in terms of the four-potential A^μ.

In the Lorenz gauge, the dynamics reduce to the wave equation $\Box A^\mu = \mu_0 J^\mu$, which is **linear** because the electromagnetic field does not act as its own source. This is a fundamental ontological divergence: while the metric $g_{\mu\nu}$ in GR is both the "scaffolding" and the "source" (leading to non-linearity), the vector potential A^μ in

EM is a field that inhabits the scaffolding without deforming it. Consequently, while GR requires 10 independent components to describe the curvature of spacetime, EM requires only **4 components** of the four-potential to describe the "swirl" of the gauge field.

The "Structural Wall" is defined by this mapping:

$$\text{EM:} \quad \text{Derivative of Curvature} \sim \text{Source} \tag{7.31}$$

$$\text{GR:} \quad \text{Curvature} \sim \text{Source} \tag{7.32}$$

The AB effect provides the empirical proof that the potential A_μ is real. The phase shift $\Delta\phi$ acquired by a particle around a closed loop Γ is:

$$\Delta\phi = \frac{q}{\hbar}\oint_\Gamma A_\mu dx^\mu = \frac{q}{\hbar}\iint_{S(\Gamma)} F_{\mu\nu} d\sigma^{\mu\nu} \tag{7.33}$$

This phase shift is an example of **holonomy**—the failure of a state to return to its original value after transport around a loop. In the geometric language of both EM and GR:

$$\text{EM Holonomy:} \quad \exp\left(i\frac{q}{\hbar}\oint A_\mu dx^\mu\right) \tag{7.34}$$

$$\text{GR Holonomy:} \quad \text{Parallel Transport of a Vector } \vec{V} \tag{7.35}$$

The "Structural Wall" of field dynamics tells us *how* the swirl is created, but the Aharonov-Bohm effect tells us that the *swirl itself* is what the particle actually feels.

To address the dynamical mismatch between EM and GR, one may postulate a differential extension of the gravitational field equations. While Maxwell's equations relate the derivative of the curvature to the source ($dF = J$), the Einstein Field Equations relate the curvature tensor directly to the energy-momentum source ($G = T$). By shifting to a differential form ($\partial R = \partial T$):

$$\nabla_\lambda(G_{\mu\nu} - 8\pi G T_{\mu\nu}) = 0 \tag{7.36}$$

we align the two theories under a shared differential mapping. Integration of this form implies $G_{\mu\nu} - 8\pi G T_{\mu\nu} = \Lambda g_{\mu\nu}$, where the integration constant naturally emerges as the Cosmological Constant. While this "differential gravity" points toward a deeper geometric unity, it introduces extraneous solutions and potential instabilities common in higher-derivative theories. Thus, the question of whether the field dynamics should be unified at the level of curvature or its derivative remains an open challenge.

8 Outlook

Any theory is governed by the raw power of imagination, which is termed metaphysics. Metaphysics is the speculative playground where we explore possibilities and inquire about the nature of reality. Theory then comes to terms with the *ontology*—the inventory of entities we commit to in order for a theory to work, such as fields, potentials, flux, and locality.

The structural parallels summarized in Table 1 invite a deeper metaphysical question: if both gravity and electromagnetism are manifestations of geometric curvature, can they be unified within a single manifold? This inquiry leads naturally to the **Kaluza-Klein** framework, which posits that the $U(1)$ fiber bundle of electromagnetism is not merely an "internal" space, but a compactified fifth dimension of spacetime itself[10]. In this ontology, the vector potential A_μ emerges as a component of a higher-dimensional metric tensor, suggesting that the "background swirl" of the AB effect is a localized vibration of an extra-dimensional geometry. While the "structural wall" of field dynamics ($D_\mu F^{\mu\nu} = J^\nu$ vs. $G_{\mu\nu} = 8\pi G T_{\mu\nu}$) remains a significant hurdle, the transition from local fields to global holonomies—reinforced by the recent observation of the gravitational Aharonov-Bohm effect [11]—suggests that the future of physical description lies in a fully integrated, potential-centric geometry where the distinction between "force" and "space" finally dissolves.

Ontology is often constrained by the choices we make of what to select or reject. For example, General Relativity (GR) assumes the equivalence of inertial and gravitational mass, while Quantum Field Theory (QFT) posits the creation and annihilation of particles. The claims of a theory are verified by *epistemology*, which determines the conditions under which an ontology becomes justified knowledge through experimental validation.

A theory is a logically consistent framework that builds from actuality (ontology) and metaphysical principles (symmetry and causality) to create a logical synthesis with predictive power. There exists a feedback mechanism between ontological commitments and epistemological frameworks; experiments force us to rethink our metaphysics and expand our ontology.

For instance, the Aharonov-Bohm effect [1] was a theoretical prediction reached by preferring the ontological commitment of the vector potential $\mathbf{A}$ over magnetic fields. This is mathematically evidenced by the fact that the energy spectrum (Eq. 4.38 and 4.56) is determined solely by the ratio of the total flux to the flux quantum, regardless of the local field strength along the particle's path. The epistemological breakthrough was achieved by Tonomura [12], providing the experimental verification that shifted the vector potential from a mathematical construct to a physical reality.

To extend this toward General Relativity, we approach the concepts of fiber bundles and 1-form connections. The non-uniqueness of the vector potential $\mathbf{A}$ suggests the definition of a private vector space for each point in spacetime. Similar to how a tangent space is defined at every point on a spherical surface, an infinitesimal vector element resides within its own local space. To perform operations like addition or subtraction between vectors at different points, one must parallel transport them to a common space through a projection mechanism defined as a fiber bundle. In this geometric view, the potential acts as a connection that dictates how the phase of a wavefunction rotates during parallel transport. The failure of these transport operations to commute reveals a gauge-theoretic curvature mathematically analogous to **Riemann curvature** in GR. This link is further solidified by the recent experimental observation of a gravitational Aharonov-Bohm effect [11], suggesting that the "background swirl" of the vector potential is a local manifestation of the deep geometric principles governing the curvature of spacetime.

Acknowledgements

The author would like to thank the Physics Academy of the North East (PANE) for providing a platform to discuss these foundational concepts in physics education.

References

[1] Y. Aharonov and D. Bohm, "Significance of electromagnetic potentials in the quantum theory," *Physical review*, vol. 115, no. 3, p. 485, 1959.

[2] Y. Aharonov, I. L. Paiva, Z. Schwartzman-Nowik, A. C. Elitzur, and E. Cohen, "Time-symmetry and topology of the aharonov–bohm effect," *Journal of Physics A: Mathematical and Theoretical*, vol. 56, no. 47, p. 475302, 2023.

[3] P. A. M. Dirac, "Quantised singularities in the electromagnetic field," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 133, no. 821, pp. 60–72, 1931.

[4] J. D. Jackson, *Classical Electrodynamics*. Wiley, New York, 2 ed., 1962.

[5] K. Shinohara, T. Aoki, and A. Morinaga, "Scalar aharonov-bohm effect for ultracold atoms," *Physical Review A*, vol. 66, no. 4, p. 042106, 2002.

[6] T. T. Wu and C. N. Yang, "Concept of nonintegrable phase factors and global formulation of gauge fields," *Physical Review D*, vol. 12, no. 12, p. 3845, 1975.

[7] D. J. Gross and F. Wilczek, "Ultraviolet behavior of non-abelian gauge theories," *Physical Review Letters*, vol. 30, no. 26, p. 1343, 1973.

[8] C.-N. Yang and R. L. Mills, "Conservation of isotopic spin and isotopic gauge invariance," *Physical review*, vol. 96, no. 1, p. 191, 1954.

[9] K. G. Wilson, "Confinement of quarks," *Physical review D*, vol. 10, no. 8, p. 2445, 1974.

[10] T. Kaluza, "Zum unitätsproblem der physik," *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)*, vol. 1921, no. arXiv: 1803.08616, pp. 966–972, 1921.

[11] C. Overstreet, P. Asenbaum, J. Curti, M. Kim, and M. A. Kasevich, "Observation of a gravitational aharonov-bohm effect," *Science*, vol. 375, no. 6577, pp. 226–229, 2022.

[12] A. Tonomura, N. Osakabe, T. Matsuda, T. Kawasaki, J. Endo, S. Yano, and H. Yamada, "Evidence for aharonov-bohm effect with magnetic field completely shielded from electron wave," *Physical review letters*, vol. 56, no. 8, p. 792, 1986.

P. J. Phys. Edu. Vol 1(1), 2026 PANE Journal of Physics Education

A TUTORIAL FOR SPARSE IDENTIFICATION OF NONLINEAR DYNAMICAL SYSTEM (SINDy) FRAMEWORK TO DISCOVER GOVERNING EQUATIONS FROM RAW DATA

Britan Singh[1], Mani K Chettri[1], Vivek Shrivastav[1], Hemam D Singh[2], and Rupak Mukherjee[*1]

[1]*Department of Physics, Sikkim University, Gangtok, Sikkim, India*
[2]*Department of Physics, Netaji Subhas University of Technology, New Delhi, India*

[*]Corresponding Author: rmukherjee@cus.ac.in

Abstract: The discovery of governing equation from time-series data is a challenge in the analysis of nonlinear dynamical systems. In this work, a clear and systematic exposition of the Sparse Identification of Nonlinear Dynamical System (SINDy) framework is presented for data-driven equation discovery. The method is formulated through system representation, construction of a candidate function library, and identification of sparse governing terms. The approach is validated using 1D, 2D, and 3D systems, including the logistic model, nonlinear pendulum, and Lorenz system. The coefficients we found shows that only the actual terms from the true equation are important, unnecessary terms are correctly ignored. These results show that SINDy is an effective method for discovering the governing equations of nonlinear physical systems directly from data.

Keywords: SINDy; Dynamical System; Nonlinear System; Governing Equations

(Received 14 February 2026; accepted 14 March 2026; published 6 April 2026)

1 Introduction

The identification of governing equations has been a fundamental objective in the physical sciences. Traditionally, physical laws are derived from first principles through analytical reasoning, symmetry considerations, and conservation laws [1]. This approach worked well for simple systems, but it runs into problems when dealing with complex, nonlinear behavior [1]. In these cases, the interactions are so complicated that it becomes very difficult to derive the governing equations using traditional methods. However, the advancement of experimental tools, and numerical simulations provide us data rich research environments [2, 3]. Because traditional methods struggle with nonlinear systems, researchers are now turning to data-driven methods. These methods let the data itself reveal the underlying physical law, without needing to start from first principles [4].

Early work in data-driven methods focused on techniques that automatically discover equations from data, such as symbolic regression [5, 6]. These methods showed that it is possible to discover meaningful physical laws, without having prior knowledge of the governing equations. Symbolic regressions frameworks have been successfully applied to recover conservation laws and dynamical models from experimental and simulated datasets [7]. However, these early approaches come with challenges. They often have to search through a large number of possible equations, which demands significant computational power and time. They can also produce very complicated equations that are hard to understand or interpret physically.

In recent years, Sparse Identification of Nonlinear Dynamical System (SINDy) framework has emerged as a powerful for data-driven model discovery [1, 2, 3, 4]. The main idea behind SINDy is that the governing equations are sparse in nature [3, 4]. SINDy works by first estimating derivatives from time series data, then building a large

library of possible candidate functions, and finally using regression technique to pick out only the important terms [3]. This results in a simple, compact model that is easy to interpret. Compared to older methods like symbolic regression, SINDy is much faster and less likely to produce overly complicated equations [3, 4]. Because of this, SINDy is especially useful for uncovering the rules that govern nonlinear and complex systems.

In this work, we present a clear and straightforward explanation of the SINDy method, focusing on how it can be used to discover the governing equations of physical systems. We use classical benchmark systems, including the logistic model, simple pendulum, and Lorenz system, as illustrative examples [5].

2 Dynamical Systems and Time-Series Data

Many physical systems can be described within the framework of dynamical system, where the state of the system evolves with time according to an underlying governing law [1]. Mathematically a dynamical system can be represented as

$$\frac{dX}{dt} = f(X). \tag{1}$$

Here, $X(t) = (x_1, x_2, x_3....)$ denotes the state of the given system, and $f(X)$ in an unknown function that determines how the system evolves. Traditionally the $f(X)$ is derived from first principle based on physical law [5]. In contrast, data-driven approaches assume that the governing function is unknown and must be derived directly from observational data [3, 4]. In real experiments or simulations, we collect time-series data by measuring the system's state at specific time points and these collected data can be arranged into a data-matrix

$$\mathbf{X} = \begin{bmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ . \\ X(t_m) \end{bmatrix}, \tag{2}$$

where m is the number of time samples [4]. In practical situations,the time derivatives of the state variables are usually not measured directly. Therefore, the derivatives must be estimated numerically from the available time-series data [3]. This is an important step because any error in derivative estimation can affect the accuracy of the discovered governing equations [3, 4]. The estimated derivatives can be arranged into a derivative matrix given by

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{X}(t_1) \\ \dot{X}(t_2) \\ \vdots \\ . \\ \dot{X}(t_m) \end{bmatrix}, \tag{3}$$

where $\dot{X}(t_i)$ represents the time derivative of the state vector at time t_i. The main goal is to determine the relationship between $f(X)$ and $\dot{X}$ using the available time series data.

3 Mathematical formulation of "SINDy" method

The "Sparse Identification of Nonlinear Dynamical system" ("SINDy") method is a data-driven framework for method governing equation directly from time-series data [1, 3, 4]. As discussed in earlier section, the system dynamics can be expressed as

$$\dot{X} = f(X), \tag{4}$$

where $X(t) \in \mathbb{R}^n$ denotes the state-vector of the system and $f(X)$ is an unknown nonlinear function governing the temporal equation of the system [4, 8]. In many practical cases, the explicit analytical form of $f(X)$ is not available and must be inferred directly from observational or simulated time-series data. The central idea of the SINDy framework is to approximate the unknown function $f(X)$ using a set of candidate basis functions constructed from the measured data. These candidate functions are organized into library matrix, denoted by

$\Theta(X)$, which contains various possible linear and nonlinear combinations of the state variables. A general form of the library matrix can be written as

$$\Theta(X) = \begin{bmatrix} 1 & x_1(t_1) & x_2(t_1) & x_1^2(t_1) & x_1(t_1)x_2(t_1) & \cdots & \cdot \\ 1 & x_1(t_2) & x_2(t_2) & x_1^2(t_2) & x_1(t_2)x_2(t_2) & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \\ 1 & x_1(t_m) & x_2(t_m) & x_1^2(t_m) & x_1(t_m)x_2(t_m) & \cdots & \cdot \end{bmatrix}. \tag{5}$$

Where each row corresponds to a time snapshot of the system and each column represents a candidate function constructed from the state variables [3, 4]. The first column usually denotes a constant term, while the remaining columns include linear terms, polynomial terms, and other nonlinear functions depending on the expected physics of the system.

The key assumption underlying the SINDy method is that the true governing dynamics are sparse in the space of candidate functions. This means that, although the library matrix may contain a large number of possible terms, only a few small subset of these terms are required to accurately describe our system dynamics [1]. Based on this assumption, the system dynamics can be expanded as a linear combination of the candidate library functions,

$$\dot{X} = \Theta(X)\,\Xi, \tag{6}$$

where Ξ is a matrix of coefficients that tells which terms from the library are actually important. The above equation has a form similar to the standard regression problem, therefore the identification of the governing equations can be treated as a regression problem.

To make this formulation more explicit, the time-series measurements are first arranged into a state data matrix and a corresponding time-derivative matrix. The derivative matrix is given by

$$\dot{X} = \begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \cdot & \dot{x}_n(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \cdot & \dot{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots & \\ \dot{x}_1(t_m) & \dot{x}_2(t_m) & \cdots & \cdot & \dot{x}_n(t_m) \end{bmatrix}, \tag{7}$$

where m denotes the number of time snapshots and n is the number of state variables. Each row corresponds to a time instant, while each column represents the temporal evolution of an individual state variable [3, 4].

Similarly, the library matrix $\Theta(X)$ is constructed from candidate nonlinear functions of the measured state variables and can be written in a general form as

$$\Theta(X) = \begin{bmatrix} 1 & \theta_1(X(t_1)) & \theta_2(X(t_1)) & \cdots & \cdot & \theta_p(X(t_1)) \\ 1 & \theta_1(X(t_2)) & \theta_2(X(t_2)) & \cdots & \cdot & \theta_p(X(t_2)) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 1 & \theta_1(X(t_m)) & \theta_2(X(t_m)) & \cdots & \cdot & \theta_p(X(t_m)) \end{bmatrix}, \tag{8}$$

where p is the number of candidate basis functions, including constant, linear, polynomial, and other nonlinear terms depending on the expected physics of the system [5, 9]

With these definitions, Eq. (6) represents a linear regression problem in a high-dimensional feature space, where the objective is to determine the sparse coefficient matrix

$$\Xi = \begin{bmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1n} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{p1} & \xi_{p2} & \cdots & \xi_{pn} \end{bmatrix}. \tag{9}$$

Each column of Ξ corresponds to the governing equation of a specific state variable, indicating which candidate functions actively contribute to the system dynamics, while the remaining coefficients remain close to zero due to the sparsity assumption.

4 Least Squares Identification

In the present work, the coefficient matrix x_i is obtained using a classical least squares approach. After constructing the library matrix $\Theta(X)$ and the derivative vector $\dot{X}$ from the time series data, the governing equation

$$\dot{X} = \Theta(X)\Xi, \tag{10}$$

is treated as linear regression problem. specifically, the coefficients are computed by solving the normal equations

$$\Theta(X)^T\Theta(X)\,\Xi = \Theta(X)^T\dot{X}, \tag{11}$$

which correspond to the standard ordinary least square solution [4]. In the numerical implementation, the matrices $\Theta^T\Theta$ and $\Theta^T\dot{X}$ are explicitly constructed and the resulting linear system in solved using Gaussian elimination.

Although ordinary least squares regression does not explicitly enforce sparsity, it provides a straightforward and computationally efficient method for identifying the important terms in the governing equations. More broadly, modern machine learning and data-driven modeling approaches have played a significant role in discovering governing equations directly from data [10, 11, 12].

Steps	Description
1	Collect time-series data of the state variables $X(t)$ from experiments or numerical simulations.
2	Estimate the time derivatives $\dot{X}(t)$ using an appropriate numerical differentiation method.
3	Construct the candidate function library matrix $\Theta(X)$ using linear, polynomial, or nonlinear combinations of the state variables.
4	Formulate the regression problem in the form $\dot{X} = \Theta(X)\Xi$.
5	Compute the initial coefficient matrix Ξ using regression methods.
6	Reconstruct the governing equations from the identified coefficients, consistent with sparse representation and compressed sensing principles [13].

Table 1: Step-by-step algorithm of the SINDy method for equation discovery

The SINDy formulation $\dot{X} = \Theta(X)\Xi$ follows the standard framework of sparse regression for dynamical systems described in data-driven modeling literature [4, 10, 14].

5 Illustrative Examples with Code

To demonstrate the effectiveness of the SINDy method [14] we use several classical dynamical systems as examples [1]. These benchmark systems are chosen because their governing equations are well known, which allows us direct comparison between the identified model and the true model.

5.1 Logistic Model

The logistic growth equation,

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right), \tag{12}$$

is a classical nonlinear model that describes population growth in the presence of limited resources [1], where r is the intrinsic growth rate and K is the "carrying capacity" of the system [5]. The term rN represents exponential growth, while the quadratic term $-\frac{r}{K}N^2$ introduces nonlinear saturation as the population approaches the carrying capacity.

Due to the presence of a simple quadratic nonlinearity, the logistic model serves as an ideal benchmark for testing data-driven discovery methods such as genetic programming, symbolic regression and SINDy [8, 15, 16, 17].

To demonstrate the capability of the SINDy method in discovering governing equations from data, we first generate time-series data $N(t)$ using the logistic growth model. The corresponding time derivative $\dot{N}(t)$ is then computed.

In this implementation, a candidate library is chosen as

$$\Theta(X) = [1,\ x,\ x^2,\ x^3,\ x^4,\ x^5,\ \sin(x),\ \cos(x)] \tag{13}$$

which includes polynomial as well as trigonometric terms, The purpose of choosing an extended library is to test whether the SINDy framework can correctly identify the true governing dynamics even in the presence of many irrelevant candidate functions. The SINDy algorithm then determines the coefficient vector associated with each term in the library.

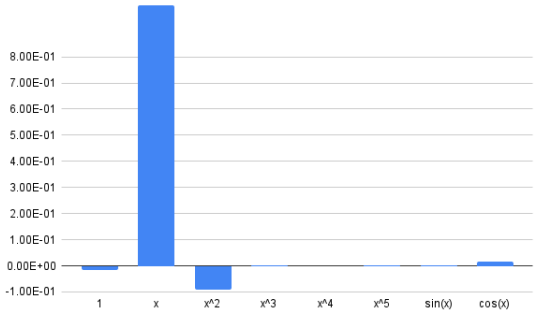


Figure 1: Coefficients identified by SINDy for the logistic growth model.

Fig. 1 illustrates the identified coefficients for the logistic growth model. It is clearly observed that the dominant coefficients correspond to the linear term x and the quadratic term x^2 while the higher order polynomial and trigonometric terms have coefficients close to zero.

5.2 Simple Pendulum

The simple pendulum is a well known nonlinear dynamical system widely studied in classical mechanics and nonlinear dynamics [1], and its equation of motion is given by

$$\ddot{\theta} + \frac{g}{L}\sin(\theta) = 0 \tag{14}$$

where θ denotes the angular-displacement, g, the acceleration-due-to-gravity, and L, the length of the pendulum. Unlike the linearized small angle approximation, the presence of the $\sin(\theta)$ term makes the system inherently nonlinear, providing an ideal model for testing data driven method [5].

In order to apply the SINDy framework, the above second order differential equation (14), must be converted into a system of two first order equations. This is achieved by defining the state variables as

$$x = \theta \tag{15}$$

$$y = \dot{\theta} \tag{16}$$

with this transformation, the simple pendulum dynamics can be expressed as a two dimensional dynamical system, In this case the candidate library is constructed using constant, polynomial and trigonometric terms of the state variables

$$\Theta(X) = [1, x, y, xy, x^2, y^2, \sin(x), \sin(y), ...] \tag{17}$$

This library allows the SINDy algorithm to identify the relevant nonlinear terms governing the pendulum dynamics.

$$\dot{x} = y \tag{18}$$

$$\dot{y} = -\frac{g}{L}\sin(x) \tag{19}$$

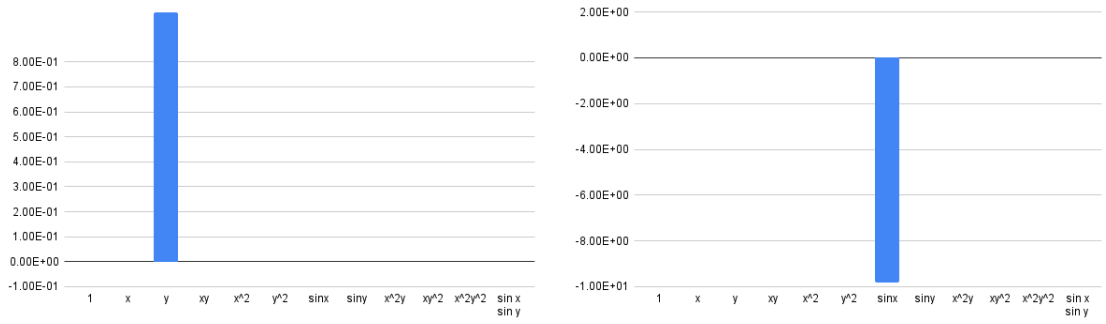


Figure 2: Identified coefficients for the Simple pendulum system using SINDy method.

Fig. 2 Presents the identified coefficient vectors obtained using the SINDy method for the simple pendulum system. The coefficient plot clearly shows that the dominant term in the first equation corresponds to the linear velocity y, which correctly represents the relation $\dot{x} = y$. In the second equation, the largest coefficient is associated with the trigonometric term $\sin(x)$, while the constant, polynomial, and cross terms exhibit coefficients that are close to zero.

5.3 Lorenz System

The "Lorenz system" is a famous example of a nonlinear dynamical system showing chaos [1]. Due to its strong nonlinear coupling and sensitive dependence on initial conditions, it is widely used as a benchmark problem in nonlinear dynamics and data-driven model discovery [5, 14, 18]. The governing equations of the Lorenz system are given by

$$\dot{x} = \sigma\,(y - x) \tag{20}$$
$$\dot{y} = x\,(\rho - z) - y \tag{21}$$
$$\dot{z} = xy - \beta z, \tag{22}$$

where σ, ρ, and β are positive parameters that control the behavior of the system [5].

In this example, synthetic data for the state variables $x(t)$, $y(t)$, and $z(t)$ are generated numerically from the Lorenz equations using standard parameter values. The data are then re-arranged into a state matrix $X(t) = [x(t),\ y(t),\ z(t)]^T$, and the corresponding time derivative are computed to construct $\dot{X}(t)$. A candidate library is formed using constant, linear, quadratic, and intersection terms of the state variables [1, 3, 4]. The SINDy method is then applied to obtain coefficient matrices corresponding to the three governing equations.

In Fig. 3 the results indicate that the dominant terms in the recovered equations correspond to the linear coupling between x and y, the nonlinear interaction term xy, and the linear damping term in the z equation. All the other polynomial and higher order library terms have coefficients close to zero, showing that only a few terms significantly contribute to the system dynamics.

In addition to the coefficient analysis, phase space plot are presented to validate the dynamical behavior of the identified model. For chaotic systems such as the Lorenz system, phase portraits provide an effective way to compare the qualitative dynamics between the true system and the SINDy recovered system. From the phase portrait plot (Fig. 4), it is observed that the trajectories obtained from the SINDy model closely follow the true Lorenz attractor in all state space projection. This close agreement indicates that the SINDy identified model successfully reproduces the essential nonlinear and chaotic dynamics of the system.

6 Conclusion

In this work, we provided a clear and step by step explanation of the SINDy method, which is used to discover the equations governing a dynamical system directly from data. We started by explaining how dynamical systems are typically described mathematically in the framework of nonlinear dynamics [1]. Then, we showed how to build a

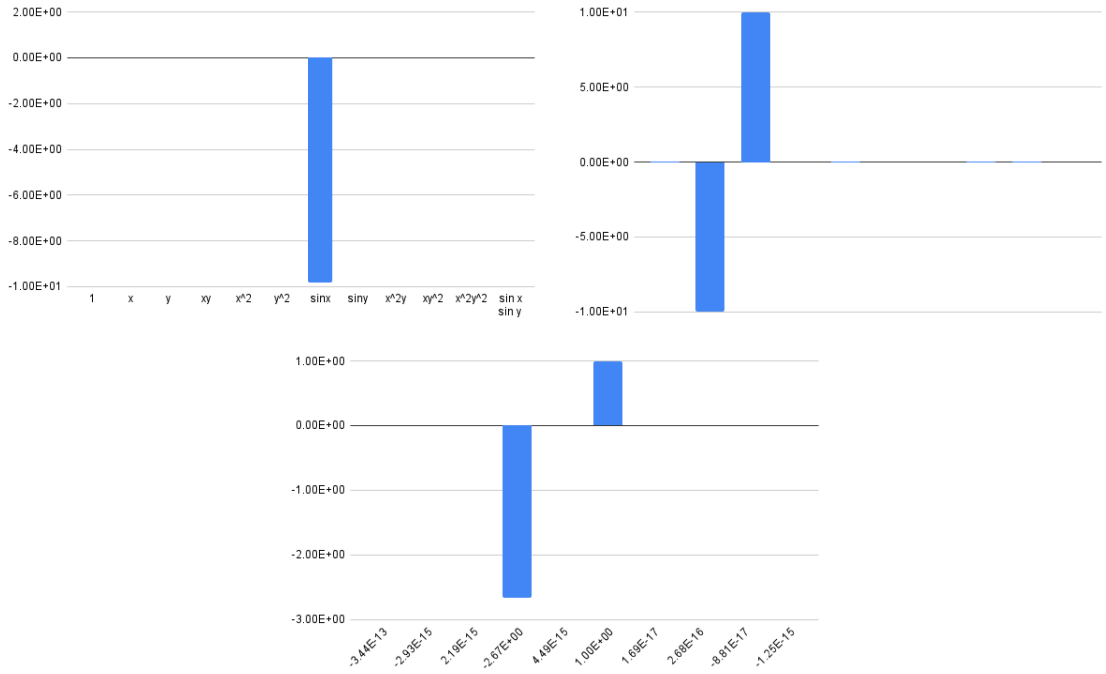


Figure 3: Identified coefficients for the Lorenz system using the SINDy method.

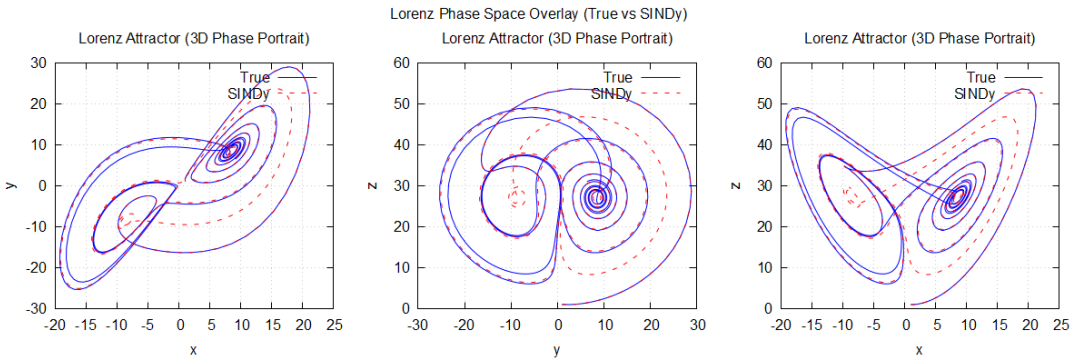


Figure 4: Comparison of true and SINDy reconstructed Lorenz attractor in phase space projections.

library of possible terms and use regression techniques to identify which ones actually appears in the true equations. The effectiveness of the SINDy method was demonstrated using several dynamical systems, including the logistic model, the nonlinear pendulum, and the Lorenz system which serve as representative nonlinear dynamical benchmarks [17]. For each case, the identified coefficient structures showed dominant contributions from the true governing terms, while higher order and redundant library terms were suppressed. This indicates that the method successfully isolates the physically relevant dynamics even when an overcomplete candidate library is used. In, particular, the phase space comparisons for the Lorenz system showed close agreement between the true trajectories and the SINDy reconstructed trajectories which confirms that the identified model successfully captures the structure of the chaotic attractor. Overall, the SINDy framework is a simple, interpretable, and computationally efficient tool for data-driven equation discovery in nonlinear physical systems, consistent with modern machine learning-based scientific modeling approaches [2].

7 Acknowledgment

All the computational work reported in this paper was carried out using the ”Brahmagupta” HPC facility at Sikkim University. One of the authors R.M. acknowledges support provided by IUCCA, Pune, through the visiting asso-

ciate program.

A Appendix

This appendix provides the complete Fortran implementations used for the illustrative examples presented in this work. The codes generate time-series data, construct the candidate library matrix, and compute the coefficient vector. Separate codes are provided for the logistic growth model, nonlinear pendulum, and Lorenz system to ensure reproducibility of the results.

A.1 Logistic Model code

```
program sindy_least_square_logistic_model
  implicit none
  integer, parameter::N = 100
  real(kind=8),dimension(N)::t, x,xdot
  real(kind=8),dimension(N,8)::Theta
  real(kind=8),dimension(8)::xi_x
  real(kind=8)::dt,r,K
  integer::i

 dt=0.01d0
  x(1)=1.0d0
  t(1)=0.0d0
  r=1.0d0
  K=10.0d0
  do i=1,N
     xdot(i)= x(i)*(r-x(i)/K)
     x(i+1)= x(i)+xdot(i)*dt
     t(i+1)= t(i) + dt
  end do

  do i=1,N
     Theta(i,1)= 1.0d0
     Theta(i,2)=x(i)
     Theta(i,3)=x(i)**2
     Theta(i,4)=x(i)**3
     Theta(i,5)=x(i)**4
     Theta(i,6)=x(i)**5
     Theta(i,7)=sin(x(i))
     Theta(i,8)=cos(x(i))
  end do

  call least_squares(N,8,Theta,xdot,xi_x)

  print *, "Sparse vector"
  do i=1,8
     print*,xi_x(i)
  end do

end program

  subroutine least_squares(m,n,A,b,x)
    integer:: m, n
    real(kind=8)::A(m,n),b(m)
    real(kind=8)::x(n)
```

```
    real(kind=8),allocatable::AtA(:,:),Atb(:)
    integer::i,j

    allocate(AtA(n,n),Atb(n))
    AtA=0.0d0
    Atb=0.0d0
do i=1,n
 do j=1,n
        AtA(i,j)=sum(A(:,i)*A(:,j))
       end do
    end do

    do i=1,n
       Atb(i)=sum(A(:,i)*b(:))
    end do

    call solve_linear(n,AtA,Atb,x)
    deallocate(AtA,Atb)
  end subroutine

  subroutine solve_linear(n,A,b,x)
    integer:: n
    real(kind=8)::A(n,n)
    real(kind=8)::b(n)
    real(kind=8)::x(n)
    real(kind=8)::factor
    integer::i,j,k

    do k =1,n-1
       do i=k+1,n
          factor=A(i,k)/A(k,k)
          A(i,k:n)=A(i,k:n)-factor*A(k,k:n)
          b(i) = b(i) - factor*b(k)
       end do
    end do

    do i=n,1,-1
       x(i) = (b(i) - sum(A(i,i+1:n)*x(i+1:n)))/A(i,i)
    end do
  end subroutine
```

A.2 Simple Pendulum code

```
    program sindy_least_square_Pendulum
  implicit none
  integer, parameter:: N=100
  real(kind=8),dimension(N)::t,x,xdot,y,ydot
  real(kind=8),dimension(N,12)::Theta
  real(kind=8),dimension(12)::xi_x,xi_y
  real(kind=8)::dt,g,L
  integer :: i

  dt=0.1d0
  x(1)= 1.0d0
  y(1)= 1.0d0
  t(1)= 0.0d0
```

```
  g=9.81d0
  L=1.0d0

  do i=1,N
     xdot(i)=y(i)
     ydot(i)=-(g/L)*sin(x(i))

     x(i+1)= x(i)+xdot(i)*dt
     y(i+1)= y(i)+ydot(i)*dt
     t(i+1)= t(i)*dt
  end do

  do i = 1, N
     Theta(i,1)= 1.0d0
     Theta(i,2)=x(i)
     Theta(i,3)=y(i)
     Theta(i,4)=x(i)*y(i)
     Theta(i,5)=x(i)**2
     Theta(i,6)=y(i)**2
     Theta(i,7)=sin(x(i))
     Theta(i,8)=sin(y(i))
     Theta(i,9)=x(i)**2*y(i)
     Theta(i,10)=y(i)**2*x(i)
     Theta(i,11)=x(i)**2*y(i)**2
     Theta(i,12)=sin(y(i))*sin(x(i))

  end do

  call least_squares(N,12,Theta,xdot,xi_x)
  call least_squares(N,12,Theta,ydot,xi_y)

  print *, "Sparse vector"
  do i=1,12
     print*,xi_x(i)
  end do

   print *, "Sparse vector"
  do i=1,12
     print*,xi_y(i)
  end do

end program

  subroutine least_squares(m,n,A,b,x)
    integer:: m,n
    real(kind=8):: A(m,n),b(m)
    real(kind=8)::x(n)
    real(kind=8),allocatable::AtA(:,:),Atb(:)
    integer::i,j

    allocate(AtA(n,n),Atb(n))
    AtA=0.0d0
    Atb=0.0d0
do i = 1,n
 do j = 1,n
        AtA(i,j)=sum(A(:,i)*A(:,j))
```

```
      end do
   end do

   do i=1,n
      Atb(i)=sum(A(:,i)*b(:))
   end do

   call solve_linear(n,AtA,Atb,x)
   deallocate(AtA,Atb)
end subroutine

subroutine solve_linear(n,A,b,x)
   integer:: n
   real(kind=8)::A(n,n)
   real(kind=8)::b(n)
   real(kind=8)::x(n)
   real(kind=8)::factor
   integer::i,j,k

   do k =1,n-1
      do i=k+1,n
         factor=A(i,k)/A(k,k)
         A(i,k:n)=A(i,k:n)-factor*A(k,k:n)
         b(i)=b(i) - factor*b(k)
      end do
   end do

   do i=n,1,-1
      x(i) = (b(i)-sum(A(i,i+1:n)*x(i+1:n)))/A(i,i)
   end do
end subroutine
```

A.3 Lorenz system code

```
   program sindy_lorenz_least_squares
implicit none
integer, parameter::N= 1000
real(kind=8),dimension(N)::t,x,y,z,xdot,ydot,zdot
real(kind=8),dimension(N,10)::Theta
real(kind=8),dimension(10)::xi_x, xi_y,xi_z
real(kind=8):: dt,sigma,rho,beta
integer::i

sigma=10.0d0
beta=8.0d0/3.0d0
rho=28.0d0
dt=0.01d0

x(1)= 1.0d0
y(1)= 1.0d0
z(1)= 1.0d0
t(1)= 0.0d0

do i=1,N
   xdot(i) = sigma*(y(i)-x(i))
   ydot(i) = x(i)*(rho-z(i))-y(i)
```

```
      zdot(i) = x(i)*y(i)-beta*z(i)

      x(i+1)=x(i)+dt*xdot(i)
      y(i+1)=y(i)+dt*ydot(i)
      z(i+1)=z(i)+dt*zdot(i)
      t(i+1)=t(i)+dt
   end do

   do i = 1,N
      Theta(i,1)= 1.0d0
      Theta(i,2)=x(i)
      Theta(i,3)=y(i)
      Theta(i,4)=z(i)
      Theta(i,5)=x(i)**2
      Theta(i,6)=x(i)*y(i)
      Theta(i,7)=x(i)*z(i)
      Theta(i,8)=y(i)**2
      Theta(i,9)=y(i)*z(i)
      Theta(i,10)=z(i)**2
   end do

   call least_squares(N,10,Theta,xdot,xi_x)
   call least_squares(N,10,Theta,ydot,xi_y)
   call least_squares(N,10,Theta,zdot,xi_z)

   print*, "spars vector"
   do i = 1,10
      print *,xi_x(i)
   end do

   print*, "sparse vector"
   do i = 1,10
      print *, xi_y(i)
   end do

   print*, "sparse vector"
   do i = 1,10
      print *, xi_z(i)
   end do

end program

subroutine least_squares(m,n,A,b,x)
   implicit none
   integer:: m,n
   real(kind=8):: A(m,n), b(m)
   real(kind=8):: x(n)
   real(kind=8),allocatable :: AtA(:,:), Atb(:)
   integer :: i,j

   allocate(AtA(n,n),Atb(n))
   AtA=0.0d0
   Atb=0.0d0

   do i=1,n
      do j=1,n
```

```
        AtA(i,j)=sum(A(:,i)*A(:,j))
     end do
     Atb(i)=sum(A(:,i)*b(:))
  end do

  call solve_linear(n,AtA,Atb,x)
  deallocate(AtA,Atb)
end subroutine

subroutine solve_linear(n,A,b,x)
  implicit none
  integer::n,i,j,k
  real(kind=8):: A(n,n),b(n),x(n),factor

  do k=1,n-1
     do i=k+1,n
        factor=A(i,k)/A(k,k)
        A(i,k:n)=A(i,k:n)-factor*A(k,k:n)
        b(i)=b(i) - factor*b(k)
     end do
  end do

  do i=n,1,-1
     x(i) = (b(i)- sum(A(i,i+1:n)*x(i+1:n)))/A(i,i)
  end do
end subroutine
```

References

[1] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Westview Press, Boulder, 1994).

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science* **349**(6245), 255–260 (2015).

[3] V. Marx, "Biology: The big challenges of big data," *Nature* **498**(7453), 255–260 (2013).

[4] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, Cambridge, 2019).

[5] S. L. Brunton, J. H. Tu, I. Bright, and J. N. Kutz, "Compressive sensing and low-rank libraries for classification of bifurcation regimes in nonlinear dynamical systems," *SIAM Journal on Applied Dynamical Systems* **13**(4), 1716–1732 (2014).

[6] J. Bongard and H. Lipson, "Automated reverse engineering of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences USA* **104**(24), 9943–9948 (2007).

[7] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, "Spectral analysis of nonlinear flows," *Journal of Fluid Mechanics* **645**, 115–127 (2009).

[8] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science* **324**(5923), 81–85 (2009).

[9] B. C. Daniels and I. Nemenman, "Efficient inference of parsimonious phenomenological models of cellular dynamics using S-systems and alternating regression," *PLoS ONE* **10**(3), e0119821 (2015).

[10] N. Zolman, C. Lagemann, U. Fasel, J. N. Kutz, and S. L. Brunton, "SINDy-RL: Interpretable and Efficient Model-Based Reinforcement Learning," *Nature Communications* **16.1**, 10714 (2025).

[11] J. L. Proctor, S. L. Brunton, B. W. Brunton, and J. N. Kutz, "Exploiting sparsity and equation-free architectures in complex systems (Invited Review)," *European Physical Journal Special Topics* **223**, 2665–2684 (2014).

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B* **58**(1), 267–288 (1996).

[13] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory* **52**(4), 1289–1306 (2006).

[14] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences USA* **113**(15), 3932–3937 (2016).

[15] J. R. Koza, "Genetic Programming: On the Programming of Computers by Means of Natural Selection," MIT Press, Cambridge, MA (1992).

[16] J. R. Koza, "Genetic Programming II: Automatic Discovery of Reusable Programs," MIT Press, Cambridge, MA (1994).

[17] A. Sen, D. P. Ahalpara, A. Thyagaraja, and G. S. Krishnaswami, "A KdV-like advection–dispersion equation with some remarkable properties," *Communications in Nonlinear Science and Numerical Simulation* **17**(11), 4117–4127 (2012).

[18] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of Fluid Mechanics* **656**, 5–28 (2010).

A THEORETICAL REVIEW OF FD STATISTICS IN CLASSICAL LIMIT: EQUIVALENCE OF FD AND MB STATISTICS

Pratikshya Bezbaruah*

Department of Physics, Biswanath College, Chariali-784176, Assam, India

*Corresponding Author: pratikshya@biswanathcollege.in

Abstract: In this brief communication, a systematic analysis on the FD distribution law is presented. It has been elucidated that in classical limit of high temperature or low density, FD statistics reduces to MB statistics. An empirical relation underlying the transition from Quantum to Classical statistics is reported in the present article. The equivalence in the nature of MB and FD distribution functions above certain critical energy value is depicted graphically. The fact has been established in the context of plasma electrons admissible in the near sheath region of laboratory dusty plasma set up. Coulomb coupling parameter for plasma electrons in dusty plasma environment is estimated to be much smaller than unity. The consideration of Maxwellian electrons has correctly accounted for the physical properties of laboratory dusty plasma system. An analytical justification of classical behaviour in fermions is presented in this article.

Keywords: FD statistics; MB statistics; dusty plasma; fermions and Maxwell Boltzmann particles

(Received: 16 March 2026; accepted: 3 April 2026; published: 6 April 2026)

1 Introduction

Statistical Mechanics is a branch of Physics that comprehends the behaviour of many particle systems found in nature [1]. These systems are admissible in terrestrial as well as extraterrestrial world. Statistical Physics bridges the gap between macroscopic properties of a physical system and the dynamical laws governing the evolution of microscopic constituents in a physical system [1]. A system of gas molecules obey Maxwell Boltzmann statistics (MB) which is exclusively applicable for systems that behave classically [2]. The gas molecules stay sufficiently apart thereby manifesting a character of distinguishability [1]. The behaviour of physical systems comprising of bosons (example-photons) is governed by Bose-Einstein (BE) statistics. BE statistics govern the physical behaviour of systems comprising of particles with integral spin. On the other hand, the statistics of systems comprising of half integral spin particles called fermions (example-electrons) is known as the Fermi-Dirac (FD) statistics. BE and FD statistics can explain the nature of systems that behave Quantum mechanically [3]. Thus, the two statistics (BE and FD) are collectively referred to as Quantum statistics. Quantum Statistics is operational only in the limit of low temperature and high density. In this thermodynamic limit, the system is strongly degenerate and Quantum mechanical effects are predominant [2]. The particles retain indistinguishability. However, the system of Bosons or Fermions can behave classically when the temperature of the system is reasonably high and the density is low [3]. Thus, Quantum statistics can get reduced to Classical in high temperature condition or low density regime. However in such continuous limit of transition the particles continue to be indistinguishable [4]. The reduction of Quantum statistics to classical is discussed by Barletti et. al [5] with the derivation of Quantum fluid equations for particles obeying FD and BE statistics. The work analytically shows that a semiclassical expansion of Quantum fluid equations to $O(\hbar^2)$ leads to classical fluid equations [5]. In many particle systems like dusty plasma, the plasma electrons behave classically in certain parameter regime. The electrons that interact with the heavy dust particles levitated in the near sheath region of dusty plasma set up follows MB distribution at high temperature limit. The theoretical work by Bezbaruah et. al [6] has established the presence of inertia less electrons following MB distribution in typical laboratory conditions. It is accounted that the electron susceptibility term in Dielectric Response Function of dusty plasma

is devoid of any dependence on model parameters such as external field strength or inter particle collision frequency. In the context of experimental result, Allen and Chen [7,8] have explained the relevance of Boltzmannian electron in dusty plasma set up. Allen discussed the fact that the electron density obeys MB distribution as one approaches the plasma boundary where the electric term in Lorentz force expression dominates over magnetic term and control the dynamics [7,9]. It is also observed that the MB distribution law obeys an exponential variation with energy and it reveals that the occupation index asymptotically approaches zero as energy value goes to infinity. The distribution law governing Quantum statistics for Bosons and Fermions also have exponential dependence on energy value. However, the distribution law for FD statistics directly refer to probability that a state is occupied by a particle and that for BE and MB statistics give the most probable number per quantum state [10]. The central idea of Statistical Mechanics attempts to evaluate the number of microstates corresponding to a given macrostate describing the system. It can account for the thermodynamic properties of the system under consideration. MB, BE and FD statistics have distinct expressions for computing the number of microstates. The most probable state describes the equilibrium condition in a physical system and the macrostate with highest number of microstates is recognized as the most probable state [3].

The many body systems in the Universe are subjected to the laws of Classical or Quantum statistics. The Distribution law dictates the behaviour of a physical system. Thus, it is of great importance to unravel the physics of Classical and Quantum statistics and to investigate the parameter space in which the two statistics essentially unites.

2 Theoretical Framework

2.1 Description of the Fermi Dirac (FD) system

Let us consider a system of fermions (say electrons) comprising a FD system. 'N' be the total number of fermions in the many body system. The electrons in the system are going to be distributed in some energy levels that constitute the system. One may draw an analogy of the distribution of particles with that of students in desk-bench comprising a classroom. We may suppose that 'ε_1', 'ε_2', 'ε_3'…….'ε_i'…. be the energy levels associated with respective statistical weight factors 'g_1', 'g_2', 'g_3…….'g_i'…..in a system of particles. It is considered that 'n_1' particles will occupy energy level 'ε_1', 'n_2' particles will occupy level 'ε_2' and so on.

Fermions are essentially indistinguishable particles that obey Pauli's Exclusion Principle (PEP). The PEP is connected to anti symmetric nature of wave functions associated with these particles. The matter waves associated with the particles are described by wave functions. The overlapping of these wave functions associated with two fermions tends to cancel each other thereby prohibiting the multiple occupancy of quantum states in FD system.

The number of microstates accessible for the FD system describes the possible number of ways in which the fermions can get arranged in the energy levels and hence it measures the thermodynamic probability. Following the conditions operative [1,2] in the FD system the thermodynamic probability can be stated as-

$$\omega_{FD} = \prod_{i=1}^{k} \frac{g_i!}{n_i!(g_i - n_i)!} \tag{1}$$

where the symbols have their usual meaning.

The above expression for thermodynamic probability can be used to estimate the FD distribution law:

$$f_{FD} = \frac{n_i}{g_i} = \frac{1}{e^{\alpha} e^{\frac{\varepsilon_i}{KT}} + 1} \tag{2}$$

Here, $\alpha = \frac{-\mu}{KT}$ is the Lagrangian multiplier, defined as the ratio of chemical potential to thermal energy. Equation (2) describes the occupation index of FD system. The factor "1" in the denominator ensures that the value of

distribution function f_{FD}cannot exceed unity. It is useful to account for the probability that a quantum state will get occupied by a fermion.

2.2 Reduction of FD statistics to Classical statistics-

Under the limit of high temperature and low density, the behaviour of FD system undergoes a drastic transition. The behaviour of fermions reveals a significant change in the dynamics of the system.

The mean thermal wavelength associated with the particles is defined as [3]-

$$\lambda_T = \frac{h}{\sqrt{2\pi mKT}} \tag{3}$$

At high temperature, the mean thermal wavelength of matter waves associated with the particles (fermions) decreases significantly in comparison to the average inter particle distance r_{av} ($r_{av} \propto n^{-\frac{1}{3}}$, where n is number density). It thereby supresses the wave character associated with the particles. Thus, we approach a limiting condition where $r_{av} \gg \lambda_T$ [2] , the particles are less affected by the wave characteristics and these particles behave independently just like molecules of ideal gas.

In classical systems like the molecules in a gas, we may consider infinitely small size of the cells constituting the phase space. Thus, the statistical weight factor is large enough in comparison to the total number of particles going to be accommodated in a particular energy level. The probability that two particles occupy same quantum state is reasonably small despite the fact that there is no restriction in occupancy imposed for Classical systems.

2.2.1 Equivalence of FD and MB statistics

The number of microstates or the thermodynamic probability of Classical systems obeying Maxwell Boltzmann (MB) statistics is [2]-

$$\omega_{MB} = N! \prod_{i=1}^{k} \frac{{g_i}^{n_i}}{n_{i!}} \tag{4}$$

Incorporating Gibb's correction that accounts for removal of the consideration that classical particles are distinguishable, expression (4) reduces to the form-

$$\omega_{MB} = \prod_{i=1}^{k} \frac{{g_i}^{n_i}}{n_{i!}} \tag{5}$$

Equation (1) and (5) essentially give the same number of microstates under the condition of high temperature and low density. The MB thermodynamic probability can be redefined considering that the individual particles are occupying distinct cells in phase space under a certain thermodynamic condition. The number of ways for arranging 1st particle from a group of 'n_i' in the energy level 'ε_i'is 'g_i', the second particle will preferably choose any cells from a total of '$g_i - 1$', the third will choose any cell from a total of '$g_i - 2$' and so on. Thus, the expression (5) eventually reduces to expression (1).

The MB distribution law can be stated as [1]-

$$f_{MB} = \frac{n_i}{g_i} = \frac{1}{e^{\alpha} e^{\frac{\varepsilon_i}{KT}}} \tag{6}$$

Here, the symbols have their usual meaning. In the FD distribution law i.e equation (2), the unity factor in the denominator can be neglected when-

$$e^{\frac{\varepsilon-\mu}{KT}} \gg 1 \quad (7)$$

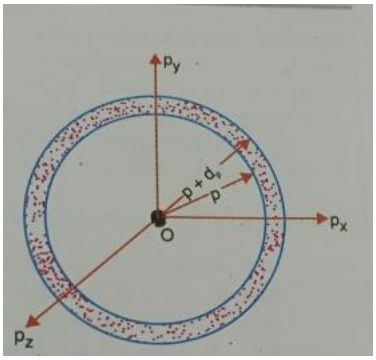
If we approximate that the exponential factor in the LHS of inequality relation (7) is around 10^2 or more than the requirement for reduction of FD distribution law (equation (2)) to MB distribution law (equation (6)) may be fulfilled.

Thus,

$$e^{\frac{\varepsilon-\mu}{KT}} \geq 10^2 \quad (8)$$

$$\text{Or, } \frac{\varepsilon-\mu}{KT} \geq 2\log_e^{10} \quad 9)$$

Inequality relation (9) depicts an empirical relation which can establish equivalence between MB and FD statistics. The equivalence is usually established in a limit where the occupation index lies below unity. The variation of occupation index with energy follows exponential decay.



The statistical weight factor associated with MB and FD system can be computed in phase space as-

$$g(p)dp = \frac{4\pi p^2 V\gamma dp}{h^3} \quad (10)$$

where V is the spatial volume and γ is the spin degeneracy term. The statistical weight factor is a measure of the number of quantum states available in the momentum range 'p' to 'p + dp'. The LHS of Distribution function for MB and FD system (equation (2) and equation (6)) calculate the occupation index. Its value decreases when the statistical weight factor associated with a particular energy level is high. In a FD system, the condition $g_i \gg n_i$ is usually true. High value of statistical weight factor in comparison to number of particles in a FD system substantially reduces the probability that a quantum state is filled with more than one particle. This condition is also readily met at high temperature and low density. In that limit the quantum mechanical effect of degeneracy pressure is insignificant however the particles naturally maintain occupation index below unity and matches with the characteristics of ideal gas molecules in that limit.

2.3 Graphical Analysis

In Figure 1, distribution law for FD system is compared with a MB system by plotting distribution function (occupation index) against normalized energy.

The normalized energy is defined as

$$\varepsilon_n = \frac{\varepsilon-\mu}{KT} \quad (11)$$

The energy of the system is scaled with the thermal energy associated with the system of particles.

The graph is plotted at room temperature, T=300K. The choice of this temperature is relevant in the present context, where the idea is to observe the behaviour of electrons in classical limit (room temperature or above). It

is observed that both the Distribution law (MB and FD) maintains an asymptotic behaviour after a certain value of energy. Thus, the probability to find a particle is zero when the energy of the level approaches infinity. From MB graph it is revealed that the maximum value of occupancy number at room temperature is approximately unity and it decreases exponentially with energy. The graph is observed at $\mu = KT$. When energy is appreciably larger than thermal energy, the distinction between two statistics eventually disappears. It happens at $\varepsilon_n \sim 3$ or above

The behaviour of FD statistics is also studied for relatively high value of chemical potential at T=300K. Figure 2 depicts the nature of Distribution function against normalized energy at $\mu = 2KT$. It is observed that the occupation index value is relatively high at low energy limit when chemical potential is greater than thermal energy. In addition to that, it is observed that the chemical potential can influence the critical value of normalized energy above which the quantum statistics reduces to classical.

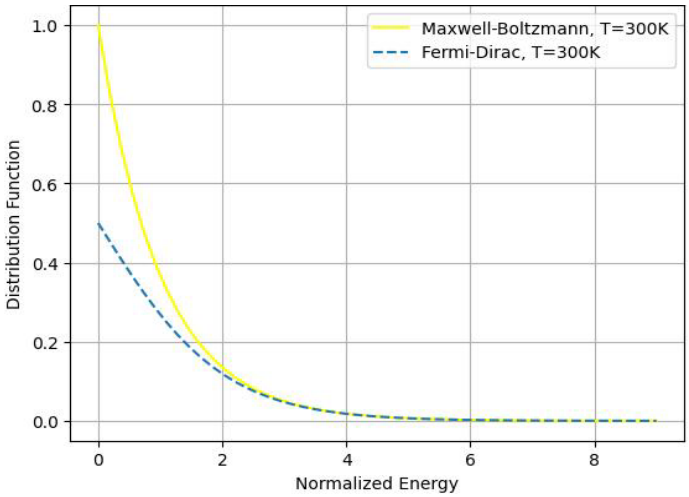


Figure 1: The figure shows the comparison of FD Distribution law with MB distribution law at room temperature (T=300 K) when chemical potential (μ=KT) is comparable to thermal energy.

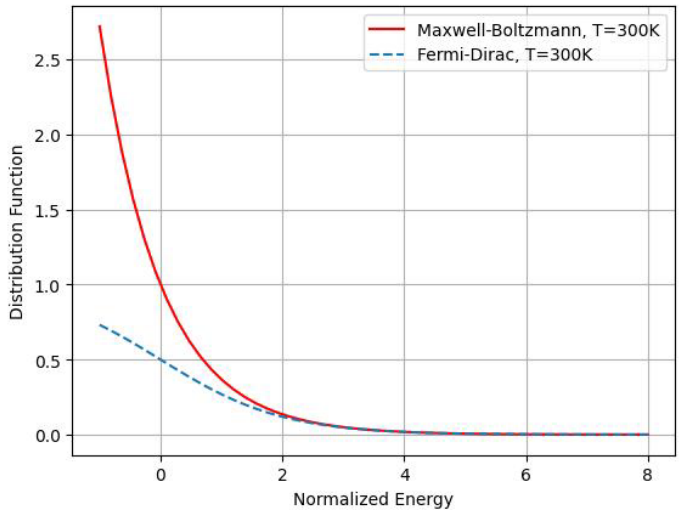


Figure 2: The figure shows the comparison of FD Distribution law with MB distribution law at room temperature (T=300 K) when chemical potential is twice the thermal energy (μ=2KT).

In Figure 1, the deviation of FD statistics from MB reduces less sharply in comparison to that in Figure 2. At room temperature, it is possible to observe the equivalence of FD and MB statistics as ε_n approaches 4. In the classical limit, the fermions have small value of distribution function or occupation index. The thermodynamics

of electron gas in classical limit is similar to ideal gas molecules. The figures also establish the fact that there is negligible probability to find particle at an energy level much higher than the chemical potential of the system.

If the temperature of FD system ($T \ll T_F$) lies much below the Fermi temperature, TF, the Fermi gas behaves quantum mechanically. Under this condition, $\mu = \epsilon_F\left(1 - \frac{\pi^2}{12}\left(\frac{T}{T_F}\right)^2\right)$. Moreover, Fermi energy $\epsilon_F \propto n^{2/3}$. Thus the effect of temperature and density in governing the strength of distribution function (f_{FD}) can be estimated.

2.4 Electrons in laboratory Dusty Plasma (DP) set up

The laboratory dusty plasma set up is an ensemble of electrons, ions and dust that can exhibit collective behaviour. Dusty plasma is used as a model system to diagnose the physical properties of many body system at its atomistic level [6, 9].

In the laboratory dusty plasma systems the typical value of number density for electron species is $n_e \sim 10^{15} m^{-3}$ and spin degeneracy term for electron is 2 [11, 12]. The expression for Fermi energy is obtained considering a Fermi sphere with radius equal to Fermi momentum. Now, the Fermi energy is defined as- $\epsilon_F = \frac{p_F^2}{2m}$ where, $p_F = \left(\frac{3N}{8\pi V}\right)^{\frac{1}{3}} h$, where N is the total number of fermions and V is the spatial volume.

The Fermi energy of electrons in DP system can be calculated using the relation [2]-

$$\epsilon_F = \left(\frac{3n_e}{8\pi}\right)^{\frac{2}{3}} \frac{h^2}{2m} \tag{12}$$

Here, $n_e \sim 10^{15} m^{-3}$, $h = 6.636 \times 10^{-34} JS, m = 9.1 \times 10^{-31} kg$

Thus, $\epsilon_F = 1.3 \times 10^{-29} J$. Now,

$$\epsilon_F = KT_F \tag{13}$$

The Fermi temperature corresponding to the calculated value of Fermi energy for dusty plasma electrons is given as $T_F = 10^{-6} K$. The Fermi temperature is much smaller than typical temperature of dusty plasma electrons ($T_e \sim 10^4 K$). Hence, the electron gas is weakly degenerate and it obeys Maxwell Boltzmann distribution in typical laboratory conditions. Such theoretical considerations have yield experimentally verifiable structural and thermodynamic properties of many body systems [13, 14].

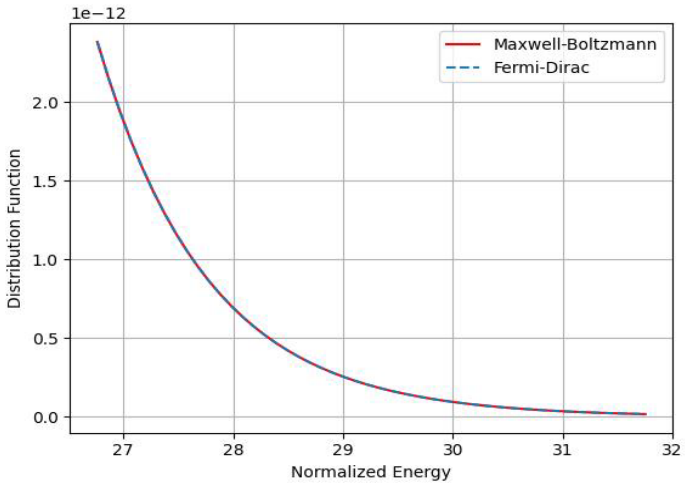


Figure 3: The figure shows the comparison of FD Distribution law with MB distribution law for dusty plasma parameters-$T_e \sim 10^4$K and $n_e = 10^{15}m^{-3}$. Figure depicts the nature of electron gas in typically laboratory dusty plasma parameters.

The plasma electrons in laboratory dusty plasma set up, obeys MB distribution for the entire range of normalized energy when the temperature $T_e \sim 10^4$K and density $n_e = 10^{14} - 10^{15}m^{-3}$. The Coulomb coupling parameter is defined as the ratio of potential to thermal energy of species in dusty plasma environment. It measures the degree of correlation between the particles [15].

$$\Gamma = \frac{e^2}{4\pi\varepsilon_0 r_{av} KT} \tag{14}$$

$$\Gamma = \frac{(1.6\times10^{-19})^2}{4\times3.14\times8.85\times10^{-12}\times10^{-5}\times1.3\times10^{-23}\times10^4} \tag{15}$$

$$\Gamma = 1.7 \times 10^{-4} \ll 1 \tag{16}$$

The above condition justifies the ideal gas approximation and establishes the fact that electrons can behave like ideal gas molecules in classical limit in typical laboratory condition. The chemical potential for electrons in laboratory dusty plasma set up may be defined as-

$$\mu = KT_e \ln\left(n_e \lambda_{T_e}^3\right) \tag{17}$$

Here, $T_e \sim 10^4$K, $n_e \sim 10^{15}m^{-3}$, the mean thermal wavelength λ_{T_e} may be calculated with dusty plasma parameters using equation (3). For the present case,

$$\mu \sim -36.92\times10^{-19} \tag{18}$$

Thus, the chemical potential for Fermi gas in dusty plasma is far less than unity. The result is true for typical laboratory condition. This is another important criteria that satisfies the consideration for MB distributed electrons in Dusty plasma.

3 Conclusions

In the present article, the behaviour and characteristics of FD statistics is studied and it is compared with Classical statistics. The condition for equivalence of two statistics is discussed. FD Distribution law in low temperature limit is also discussed. Thermodynamic parameters have a great role in tuning the occupation index. The reduction of FD statistics to MB statistics is justified in the context of laboratory dusty plasma electrons. The inequality relation (9) is obtained based on approximation and it might depend on thermodynamic parameters. The relation may guide for an approximate value of critical normalized energy beyond which the MB and FD statistics would become equivalent. In the future communication, the author may try to extend the work for a wider parameter space to explore the temperature and density range suitable for maxwellian electrons. Moreover, it will be possible to observe the impact of electron distribution on the thermodynamic properties of many particle systems such as dusty plasma.

The present work is useful to validate the presence of MB distributed electrons in dusty plasma. The analysis of FD statistics in Classical limit justifies the fact. It is well known that the condition $n_e\lambda_{T_e}^3 > 1$ leads to breakdown of the assumption that electrons behave classically. Usually at high enough densities or reasonably low temperature, the above mentioned condition may become true. Under such circumstances, the electrons no longer behave classically. Therefore, the Quantum mechanical effects turn out to be important for governing the dynamics of particles in such thermodynamic conditions.

The non maxwellian electrons are often encountered in space dusty plasmas. In non-thermal plasmas, the propagation of acoustic waves is explained with the consideration of electron distribution that deviates from Classical assumption. In such dusty plasmas, electrons are energetic enough and long tail distribution is accounted [16].

References

[1] R. K. Pathria and P. D. Beale, Statistical Mechanics, 3rd ed. (Elsevier, Amsterdam, 2011), p. 184.
[2] Evelyn Guha, Statistical Mechanics: An Introduction. (Alpha Science International, 2008), p. 246.
[3] Brij Lal, N. Subrahmanyam and P.S. Hemne, Heat Thermodynamics and Statistical Physics. (S Chand and Company Ltd, 2008), p. 665.
[4] S. Fujita, R. Schubert, C. T. Ho and E. K. Lee, Fortschr. Phys 38, 855-873 (1990).
[5] Luigi Berletti and Carlo Cintolesi, Journal of Statistical Physics 148, 353-386 (148).
[6] Pratikshya Bezbaruah and Nilakshi Das, Phys. Plasmas 23, 043701 (2016).
[7] J. E. Allen, Phys. Plasmas 14, 024701 (2007).
[8] Francis F. Chen, Phys. Plasmas 14, 094703 (2007).
[9] L. Nasi and J. L. Raimbault, Phys. Plasmas 17, 113513 (2010).
[10] Arthur Beiser, Concepts of Modern Physics. (McGraw-Hill,2009), p. 556.
[11] R.L. Merlino, A. Barkan, C. Thompson and N.D. Angelo, Phys. Plasmas 5, 1607 (1998).
[12] P. Bezbaruah and N. Das, Phys. Plasmas 25, 053708 (2018).
[13] T. Ott and M. Bonitz, Phys. Rev. Lett. 103, 195001 (2009).
[14] Z. Donko, J. Goree, P. Hartmann, and B. Liu, Phys. Rev. E 79, 026401 (2009).
[15] P. K. Shukla and A. A Mamun, Introduction to Dusty plasma physics. (IOP Publishing Ltd, 2002), p.8-27.
[16] G. Williams, F. Verheest, M.A Hellberg, M.G.M Anowar and I. Kourakis, Phys. Plasmas 21, 092103 (2014).

VIDEO-BASED MOTION TRACKING AND ANALYSIS OF SIMPLE PENDULUM OSCILLATIONS

Lachit Saikia*, and Priyanka Basyach

Department of Physics, Dibru College, Dibrugarh

*Corresponding Author: lachit@dibrucollege.edu.in

Abstract: In this paper, we report a very basic experiment on "Oscillatory motion of a simple Pendulum" using Tracker video analysis and a smartphone camera. Tracker software is fundamentally an ICT-based learning tool and is favoured as it is free, user-friendly, and supports effective learning and teaching. By integrating a smartphone camera into the laboratory setup, the analysis of oscillations is performed at a large angle rather than the traditional small-angle approach. This study findings highlight the feasibility of this approach in studying the motion of a simple pendulum while being simultaneously interactive and inexpensive. Here, the value of "g" (acceleration due to gravity) is determined and the damping coefficient is also calculated. We found that the large-angle approximation did not cause too large an error compared to the original value of "g".

Keywords: tracker; motion-tracking; simple pendulum

(Received 7 March 2026; accepted 1 April 2026; published 7 April 2026)

1 Introduction

In any Physics undergraduate curriculum, the general mechanics laboratory course typically concerns studying various types of oscillations such as a simple pendulum, a compound pendulum, Maxwell's needle, etc. The crucial measurements involved in these experiments are the time taken for a certain number of oscillations which is obtained using a stopwatch. However, manual execution of these experiments evokes lots of human errors which are minimized by recording a large number of observations or measurements[1]. But these pendulum experiments generally deal with the calculation of "g" from the $L^2 - T$ curve only and do not provide information whether the oscillations are simple harmonic in nature, as it is very non-intuitive. Furthermore, no information regarding position can be obtained at various time intervals from this plot. It is therefore essential to involve the students in a more engaging way of doing practicals which would pave the way for adding to their theoretical knowledge and make it more thought-provoking. That can be achieved by the integration of information technologies into the teaching and learning process [1]. In pendulum experiments, the calculations are usually done with the small-angle approximation, but in many cases of compound pendulums, the angle of oscillation is to be considered larger so as to obtain any oscillation at all. This major disadvantage can be overcome using ICT-based approaches such as sensor-based data acquisition termed as micro-computer-based laboratories [2-4] and video capture and analysis, which is termed as video motion-based analysis [5,6]. The first approach while being fascinating and educative, requires a greater learning curve and involves more set-up cost whilst the second approach is easier to learn and available at low cost.

A. Video Motion-based Analysis

A simple smartphone camera with high picture quality can be used to capture videos of the experiments being performed. The motion of the object of interest which is the bob of simple pendulum in our case is recorded as a video followed by further analyses using a motion tracking software called Tracker. Tracker allows the bob's position to be tracked frame by frame thus providing the position and time information required for analysis with good accuracy and precision.[2]. This technique involves less time in experimentation, thereby granting the student more time for analysis and inference of results.

B. Analysis using Tracker software

Tracker is a video analysis package built on the Open Source Physics (OSP) [7] Java framework. It is a user-friendly software where an object's motion is tracked using the video footage shot by web cam/mobile camera. The software is enriched with several features including[2]

1. Tracking the object with position, velocity and acceleration overlays and graphs
2. Multiple reference frames
3. Calibration scales
4. Line profiles for analysis of spectra and interference patterns

C. Theory of Simple Pendulum Oscillations

A simple pendulum consists of a small mass (m), also known as the pendulum bob which is suspended from an inextensible string of negligible mass to a pivot as shown in the figure below,

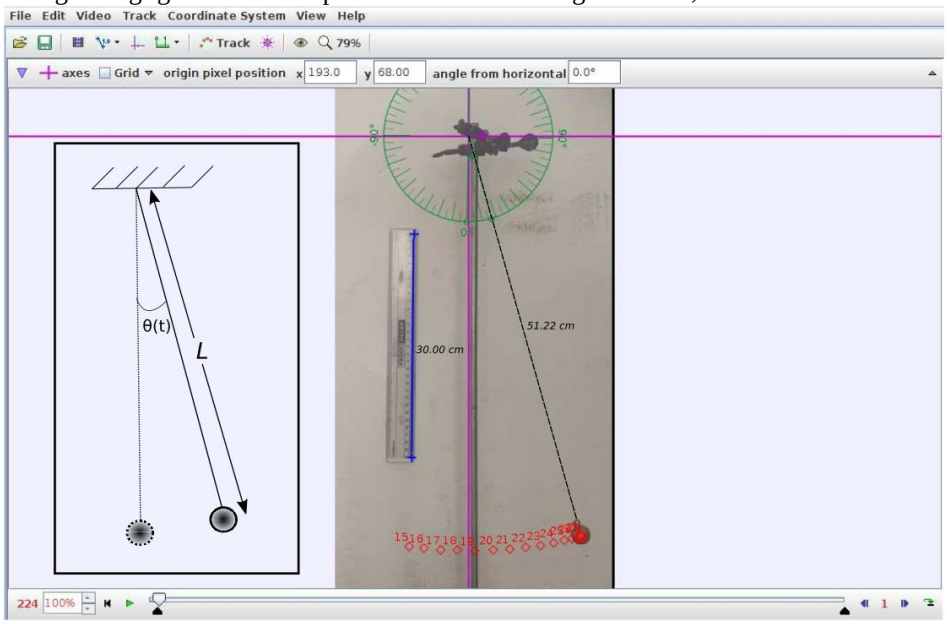


Figure 1: Schematic of a simple pendulum overlaid on a screenshot of a video frame in Tracker

When the bob is released with an initial angle (θ), the oscillation of the pendulum can be described using the following differential equation[1]

$$\frac{d^2\theta}{dt^2} + \frac{g}{L} sin\,\theta = 0 \tag{1}$$

where g is the acceleration due of gravity, L is the length of the pendulum and θ is the amplitude of the angular displacement. While considering small angle approximation, we can consider $sin\theta \simeq \theta$. Using small-angle approximation in the above equation, we get,

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\theta = 0 \tag{2}$$

which is a linear second-order differential equation and its solution is obtained as

$$\theta = A\,sin\,\omega t + B\,cos\,\omega t \tag{3}$$

where ω is the angular frequency of oscillation and given by

$$\omega^2 = \frac{g}{L} \tag{4}$$

$$\omega = \sqrt{\frac{g}{L}} \tag{5}$$

$$T = 2\pi\sqrt{\frac{L}{g}} \tag{6}$$

Generally, for this experiment, the time taken for at least 20 oscillations is recorded using stop watch, and then this step are repeated 3-4 times, following which the average time period is calculated and then the acceleration due to gravity g, is calculated from the slope of $L^2 - T$ the curve. But for experiments involving the large angle oscillations this measurements would result in significant error. For such cases, α - sin θ curve can be plotted and the slope would give the g.

2 Damping Coefficient

In reality, the pendulum cannot keep on oscillating forever due to resistive forces such as air resistance and the amplitude of oscillation will decline due to the frictional force between the string and pivot[1]. Generally, the frictional force is neglected due to the negligible mass of the string and only the air resistance is considered. The equation of motion can be rewritten as

$$\frac{d^2\theta}{dt^2} + \gamma\,\frac{d\theta}{dt} + \frac{g}{L}\,sin\,\theta = 0 \tag{7}$$

where γ is the damping coefficient. Using multiple linear regression fit, we can be obtain the damping coefficient along with acceleration due to gravity.

3 Experiment and Analysis

In our experiment, we set up a small bob suspended from a hinge using a very thin white thread, as shown in figure-1 to observe the oscillations of a simple pendulum. The bob was displaced from its equilibrium position, inducing initial oscillations, and it was confirmed that the motion stays within a plane perpendicular to the direction of observation. A smartphone camera was used to record the oscillations with utmost care taken to minimize errors in capturing the in-plane motion. As shown in the figure, we used a 30 cm scale for calibration to set the distance scale within the video frames. The recorded video was then trimmed to preserve only the relevant portion required for further analysis. The initial angular amplitude was found as approximately 15°, gradually decreasing due to damping.

The recorded video of the pendulum oscillation was imported into Tracker software. Before commencing the tracking process, a coordinate system was assigned to the video frames, where the origin was placed at the point of suspension of the pendulum. As depicted in the figure-1, a measuring tape and a protractor were used to gauge distances and angles within the video frames. After setting the tracking point in a key frame, the software automatically tracked the remaining frames, embodied by red diamond-shaped points with tracking sequence number on top, as shown in the figure-1. These provide x, y position coordinates of the pendulum bob. Each frame was separated by 0.033 seconds, as the video was recorded at 30 frames per second. After obtaining the coordinates of the pendulum bob as a function of time, they were further analysed to determine the angular displacement from the mean position using

$$\theta(t) = tan^{-1}\frac{-y(t)}{x(t)} \tag{8}$$

3 Results

The time variation of angular displacement is plotted in the figure-2. As evident from the plot in the figure-2, the amplitude of oscillation is found to be decreasing with time, analogous to damped pendulum motion. The lower panel in Figure 2 shows the simple harmonic angular displacement of the pendulum for a short time span from t = *100 sec to 110 sec* where amplitude reduction is not so significant apparently; however, reduction in amplitude due to damping present in the system can not be ignored in longer time scale. Thus, this video-based

experiment provides the opportunity for studying the pendulum oscillation in a time range from a few oscillations to hundreds of oscillations, which is not possible in a typical manual laboratory setup.

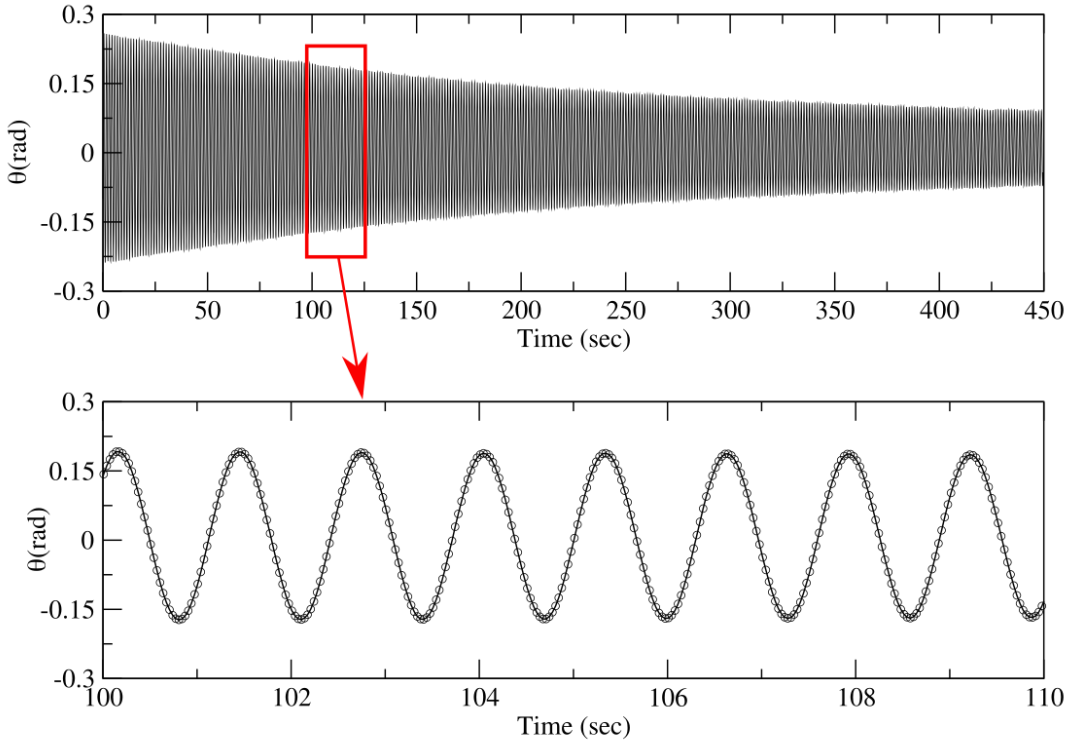


Figure 2: Variation of angular displacement w.r.t time

Using the $\theta(t)$ data, we calculated the angular velocity $\omega = \frac{d\theta}{dt}$ and angular acceleration $\alpha = \frac{d^2\theta}{dt^2}$ and plotted them versus time as shown in Figure 3. The time series ω and α data, obtained in the process, are used to calculate the damping constant γ and $\frac{g}{L}$ term using multiple regression fit. Similar to the lower panel of Figure 2, Figure 4 shows variation ω and α with time from *t = 100 to 110 sec*, where variation of ω is seen to be a smooth sinusoidal function; however, the numerical noise associated with derivative calculation reduces the smoothness for α-time graph.

Overall, as expected, both angular velocity and acceleration reduce with time due to damping. Now using expressions of angular velocity and acceleration, the equation
(7) can be rewritten as

$$\alpha + \gamma\,\omega + \frac{g}{L}\,sin\,\theta = 0$$
$$\Rightarrow\ \alpha = -\gamma\,\omega - \frac{g}{L}\,sin\,\theta$$

We implement multiple linear regression fit on the available time series data of α, ω and $sin\,\theta$ to obtain the coefficients γ and $\frac{g}{L}$. Analysis of our data resulted in a damping coefficient $\gamma = 2 \times 10^{-3}\ kg/s$ and $g = 10.1\ m/s^2$. This analysis was carried out by running a small script using an open-source computation tool named Octave.

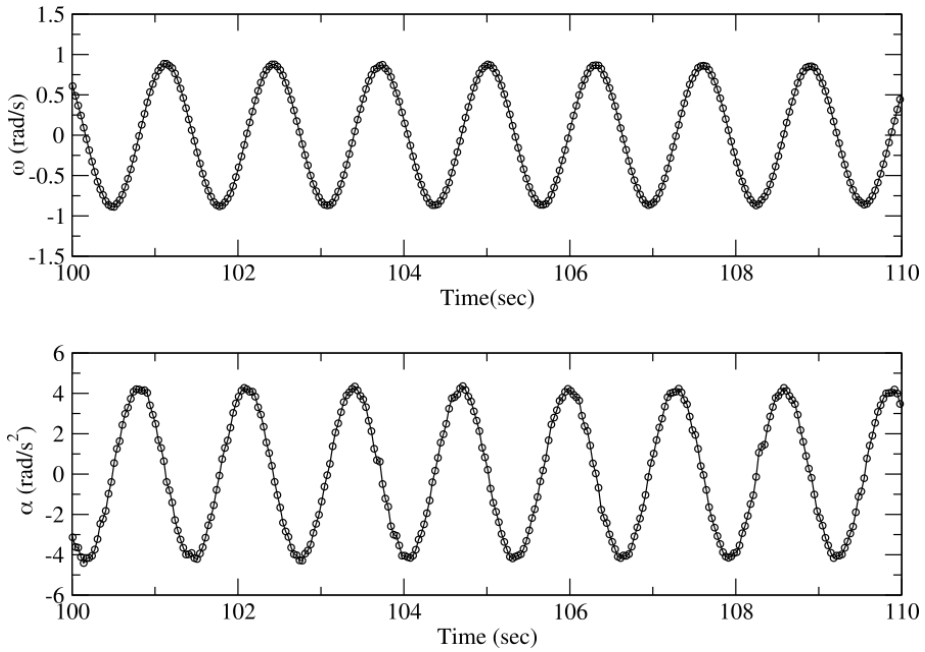


Figure 3: Angular velocity-time & Angular acceleration -time curve

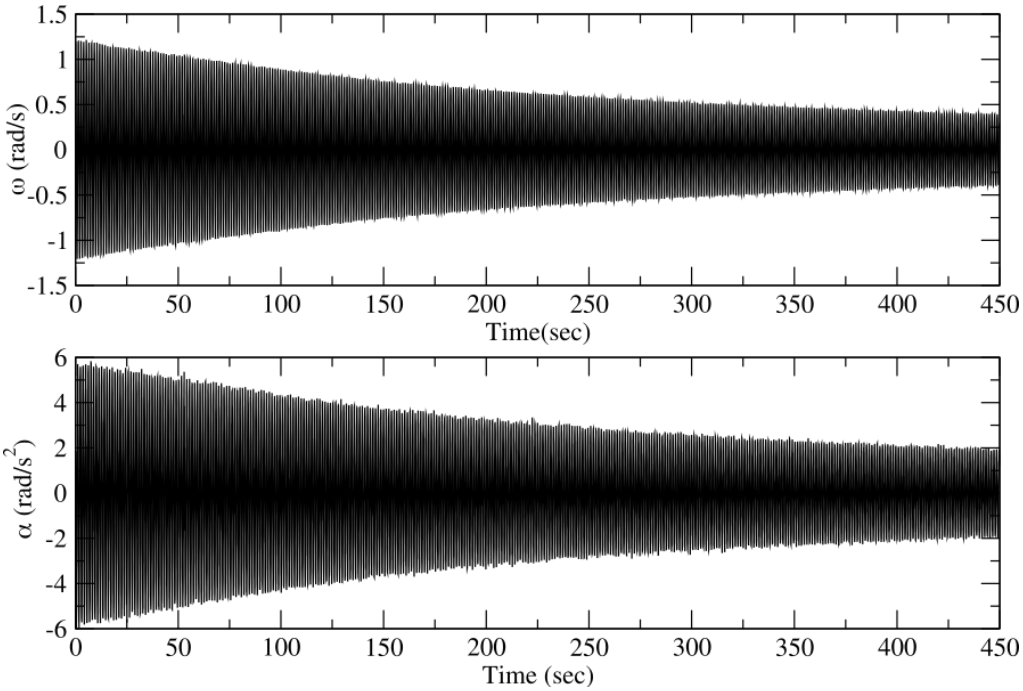


Figure 4: Angular velocity-time & Angular acceleration -time curve from *t = 100 to 110 sec*

4 Conclusion

Thus, in this work, we have employed a non-conventional method to determine the value of "g" and the damping constant γ using a simple pendulum oscillation. The obtained value of acceleration due to gravity is within 3% error, which can be attributed to the precision and accuracy involved in video analysis and the experiment itself. The damping constant lies within the expected range reported in earlier studies. The experiment reported in this article demonstrates a smart way of performing conventional laboratory experiments with a video-based analysis method using Tracker software. This technique can be extended to many other

undergraduate-level laboratory experiments, accommodating a huge amount of data and aiding in the smooth and precise analysis of them.

References

[1] M. H. Ramlia, K. T. Chana and W. F. Yapa, Solid State Science and Technology 24, 297 – 305 (2016).
[2] O. S. K. S. Sastri, International Journal of Advanced Information Science and Technology 27, 19-23 (2014).
[3] K. W. Loo, K. T. Kim, K. L. Tze and T. Ching, Physics Education 50, 436-442 (2015).
[4] C. Sirisathitkul, P. Glawtanong, T. Eadkong and Y. Sirisathitkul, Revista Brasileira de Ensino de Fisica 35, 1504-6 (2013).
[5] D. Brown, American Association of Physics Teachers Summer Meeting, (25-29) July, Ann Arbor, Michigan, (2009).
[6] A. Heck and P. Uylings, The Physics Teacher 48, 176-181 (2010).
[7] https://opensourcephysics.github.io/tracker-website/.

PANE

P. J. Phys. Edu. Vol 1(1), 2026 PANE Journal of Physics Education

HOW BIG IS THE SUN? A MEASUREMENT FROM INDIA'S ADITYA-L1 SPACECRAFT

Rupak Mukherjee[1*], Mani K Chettri[1], Vivek Shrivastav[1], Britan Singh[1], and Hemam D. Singh[2]

[1]*Department of Physics, Sikkim University, Gangtok 737102, Sikkim, India*
[2]*Department of Physics, Netaji Subhas University of Technology, New Delhi 110078, India*

[*]Email: rmukherjee@cus.ac.in

Abstract: We present a step-by-step guide to measuring the Sun's apparent angular radius from real spacecraft images, using Level-1 full-disk data from the Solar Ultraviolet Imaging Telescope (SUIT) onboard India's Aditya-L1 mission. A brightness threshold isolates the solar disk, and an algebraic circle fit to the disk boundary recovers its pixel radius. Three images taken on consecutive days in May 2025 give an angular radius of $936.4 \pm 0.3''$, corresponding to a Sun-observer distance of 1.532×10^8 km (1.024 AU), and a physical solar radius of $(680{,}900 \pm 200)$ km, approximately 2.1% below the accepted value of 695,700 km. The small shortfall is a predictable and well-understood effect of brightness-based limb detection. A comparison with published heliometric, helioseismic, and space-photometric measurements confirms that precision determinations consistently cluster near the IAU 2015 nominal value, and that our 2.1% deviation is fully accounted for by the UV-threshold limb definition used here. A self-contained Python script is provided so that every figure can be reproduced on a standard laptop.

Keywords: Aditya-L1; SUIT; solar radius; image processing; circle fitting; Python; physics laboratory

(Received 5 March 2026; accepted 3 April 2026 ; published 6 April 2026)

1 Introduction

Solar images from space carry much more information than meets the eye. The geometry of any full-disk image encodes the Sun's angular size, the position of its centre on the detector, and the distance between the spacecraft and the Sun at the moment of observation. Extracting these quantities from real data connects textbook geometry directly to live mission results. A key feature exploited in this work is the *solar limb*, the sharp, roughly circular boundary of the solar disk as seen in projection on the sky. Because the Sun's atmosphere becomes optically thin above the photosphere, the line of sight grazes the visible surface at the limb, allowing the disk edge to be identified in an image and its radius measured geometrically.

India's Aditya-L1 is the country's first dedicated solar observatory spacecraft, launched by the Indian Space Research Organisation (ISRO) in September 2023 and positioned at the Sun–Earth Lagrange point L1, approximately 1.5×10^6 km from Earth. From this vantage point it observes the Sun continuously without being interrupted by Earth's shadow. Aditya-L1 carries the Solar Ultraviolet Imaging Telescope (SUIT; [3]), which photographs the full solar disk in twelve ultraviolet bands and makes its data freely available through the ISSDC/PRADAN portal (see footnote 1 in Section 3). The images are stored in the Flexible Image Transport System (FITS) format [5], the standard archival format for astronomical data in which pixel values and observational metadata (telescope pointing, plate scale, spacecraft position) are bundled in a single file readable with widely used Python libraries.

This paper works through the measurement on three consecutive observations (2025 May 18–20). Using three frames rather than one provides a check that the result is stable from day to day, a useful lesson in reproducibility. The paper is organised as follows. Section 2 gives the geometric background. Section 3 describes the data.

Section 4 explains the method. Section 5 presents the results, including a comparison with published solar radius measurements from the literature. Section 6 discusses the accuracy and possible extensions. Section 7 covers data-use requirements. Appendix A summarises the software requirements for reproducibility. Appendix B provides the complete Python script.

2 Physical Background

2.1 Geometric relation between angular radius and distance

When we observe the Sun from a distance D, its physical radius $R_\odot$ subtends an angle $\theta_\odot$ at the observer. At the limb the line of sight is tangent to the photosphere, giving the exact relation

$$\sin\theta_\odot = \frac{R_\odot}{D} \qquad \Longrightarrow \qquad D = \frac{R_\odot}{\sin\theta_\odot}, \tag{1}$$

where $R_\odot = 6.957 \times 10^5$ km is the IAU 2015 nominal solar radius [2]. Because $\theta_\odot \approx 960'' \approx 4.65 \times 10^{-3}$ rad, the small-angle approximation $\sin\theta_\odot \approx \theta_\odot$ is accurate to better than one part in 10^5, giving the simpler working form

$$D \approx \frac{R_\odot}{\theta_\odot\,[\mathrm{rad}]}, \qquad \theta_\odot\,[\mathrm{rad}] = \theta_\odot\,[''] \times \frac{\pi}{648\,000}. \tag{2}$$

In plain words: the Sun's physical radius equals its distance multiplied by its angular size in radians. The smaller the angular size appears, the farther away the Sun must be.

2.2 From pixel radius to angular radius

A camera with plate scale s arcsec per pixel records the solar disk as a circle of R pixels. The plate scale is the angular size of one pixel on the sky, a fixed property of the telescope optics. The angular radius is simply

$$\theta_\odot\,[''] = R\,[\mathrm{px}] \times s\,[\mathrm{arcsec\,px^{-1}}]. \tag{3}$$

In plain words: multiply the radius measured in pixels by the known angular size of one pixel to get the angular radius in arcseconds.

For the SUIT frames used here, the calibrated plate scale is $s = 0.698\,\mathrm{arcsec\,px^{-1}}$, read from the FITS header keyword `CDELT1` (Coordinate DELTa along axis 1), which encodes the angular size of one detector pixel on the sky.

3 Observations and Data

SUIT carries twelve spectral bands in the near-UV range 200–400 nm [3]. The 0971 NB06 narrowband filter used here is centred near 388 nm, where the solar disk is bright and the contrast against the dark background is very clean. Level-1 FITS images have dimensions 4096×4096 pixels. At this processing level the data are geometrically corrected and have valid header metadata, but no further calibration by the authors was required or performed before applying the analysis pipeline described in Section 4. Three header keywords record spacecraft geometry: `DSUN_OBS` (Sun-spacecraft distance in metres), `RSUN_OBS` (angular solar radius in arcseconds computed from the mission ephemeris, the precise positional prediction derived from orbital mechanics), and `RSUN_REF` (the nominal solar radius, 6.957×10^8 m).

Three consecutive frames were downloaded from the PRADAN interface.[1] Table 1 lists dates and key header values. The Sun-spacecraft distance changes by less than 0.05% over the three-day span, so no measurable variation in disk size is expected.

[1] PRADAN Aditya-L1 data portal: `https://pradan.issdc.gov.in/al1`; ISSDC Aditya-L1 data access: `https://www.issdc.gov.in/adityal1.html`; FAQ: `https://pradan1.issdc.gov.in/al1/faq.xhtml`; data-use disclaimer: `https://pradan1.issdc.gov.in/al1/disclaimer.xhtml`; required acknowledgement: `https://pradan1.issdc.gov.in/al1/ack.xhtml`.

Table 1: SUIT 0971 NB06 observations used in this study. `DSUN_OBS` is the Sun-spacecraft distance in astronomical units (AU) as stored in the FITS header; `RSUN_OBS` is the angular solar radius in arcseconds computed from the spacecraft ephemeris. The variation in `DSUN_OBS` across the three days is less than 0.05%, so no measurable change in apparent disk size is expected.

Date (UTC)	File (abbreviated)	DSUN_OBS (AU)	RSUN_OBS (″)
2025 May 18, 13:47	`...000954...0971NB06`	1.00250	956.40
2025 May 19, 11:01	`...000957...0971NB06`	1.00271	955.84
2025 May 20, 13:33	`...000960...0971NB06`	1.00297	956.54

4 Method

The analysis pipeline processes each image through six stages. First, the image array and the relevant header keywords (`CDELT1`, plate scale; `RSUN_OBS`, reference angular radius; `DSUN_OBS`, Sun-spacecraft distance) are read from the FITS file. Second, a binary disk mask is constructed by intensity thresholding (Section 4.1): pixels brighter than a set fraction of the image's peak brightness are labelled "Sun" and the remainder "sky". Third, the mask is cleaned by retaining only the largest connected bright region, discarding stray noise pixels that passed the threshold. Fourth, the solar limb is defined as the boundary of the binary mask, specifically the "Sun" pixels that have at least one "sky" neighbour (Section 4.2). Fifth, a circle is fitted to these limb pixels using the Kåsa algebraic method (Section 4.3), recovering the disk centre and pixel radius. Sixth, to quantify sensitivity to the threshold choice, steps two through five are repeated for five slightly different threshold values and the resulting scatter is taken as the uncertainty (Section 4.4).

4.1 Disk mask by intensity thresholding

The solar disk is far brighter than the background. A threshold $T = f \times \bar{I}_{\text{bright}}$ is applied, where $\bar{I}_{\text{bright}}$ is the median of the top 20th percentile of pixel values and f is a number between 0 and 1. The binary mask is

$$M(x,y) = \begin{cases} 1, & I(x,y) \geq T, \\ 0, & I(x,y) < T. \end{cases} \tag{4}$$

In plain words: every pixel above the threshold is labelled "solar disk" ($M = 1$); every pixel below it is labelled "background" ($M = 0$).

The central threshold fraction $f = 0.26$ is chosen empirically as a value that reliably separates the bright solar disk from the much darker background across all three observations. Values below $f \approx 0.20$ risk incorporating faint background pixels at the image corners into the disk mask, while values above $f \approx 0.35$ begin to exclude genuine near-limb pixels where the intensity has fallen due to limb darkening. The choice $f = 0.26$ sits well within this valid regime, and the sensitivity of the result to the precise value is quantified explicitly in Section 4.4.

Stray bright pixels are removed by keeping only the largest connected region of M, found with `scipy.ndimage.label`.

4.2 Limb extraction

A boundary pixel is one with $M = 1$ that has at least one 4-connected neighbour with $M = 0$:

$$E(x,y) = M(x,y) \;\wedge \big[\, M(x{-}1,y) = 0 \;\vee\; M(x{+}1,y) = 0 \;\vee\; M(x,y{-}1) = 0 \;\vee\; M(x,y{+}1) = 0 \,\big]. \tag{5}$$

In plain words: a limb pixel is a "solar disk" pixel that touches at least one "background" pixel, i.e., it sits right on the visible edge of the Sun.

Four-connectivity (using only the four orthogonal neighbours) is adopted rather than eight-connectivity (which would also include diagonal neighbours) because it produces a single-pixel-wide, unambiguous boundary without including diagonal background pixels that are not truly adjacent to the disk edge. For a nearly circular disk of radius ~ 1342 px, switching to eight-connectivity shifts the limb position by at most one pixel, an effect well below the measured $\sigma_R \approx 4$ px threshold uncertainty.

4.3 Circle fit

A circle with centre (x_0, y_0) and radius R satisfies

$$(x - x_0)^2 + (y - y_0)^2 = R^2, \tag{6}$$

which rearranges to $x^2 + y^2 + Ax + By + C = 0$, so $x_0 = -A/2$, $y_0 = -B/2$, $R = \sqrt{x_0^2 + y_0^2 - C}$. The parameters (A, B, C) are found by solving

$$\underbrace{\begin{pmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{pmatrix}}_{\mathbf{D}} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = - \begin{pmatrix} x_1^2 + y_1^2 \\ \vdots \\ x_n^2 + y_n^2 \end{pmatrix} \tag{7}$$

in the least-squares sense via `numpy.linalg.lstsq`. This Kåsa algebraic fit [1] is simple and fast; for a densely sampled full-disk limb it underestimates R by well under 1 px, negligible here. The implementation first shifts and scales coordinates (centring by the mean and normalising by the RMS radius) before calling `lstsq`, then rescales back, for improved numerical conditioning.

4.4 Uncertainty estimation

Five threshold fractions $f \in \{0.24, 0.25, 0.26, 0.27, 0.28\}$ are tested in turn and the fitted radius recorded for each. The range $[0.24,\ 0.28]$ is chosen to bracket the central value $f = 0.26$ symmetrically by ± 0.04, a span wide enough to sample the sensitivity meaningfully while remaining within the regime where the threshold unambiguously separates disk from background. As visible in Figure 4, the fitted radius decreases monotonically as f increases, because a higher threshold contracts the mask boundary inward by excluding the faintest near-limb pixels. Extending the range significantly beyond these limits, for instance below $f = 0.20$ or above $f = 0.35$, would approach threshold values where the mask either absorbs background noise or erodes genuine limb pixels, causing the curves to diverge more steeply. The uncertainty is taken as the standard deviation of the five fitted radii, $\sigma_R = \mathrm{std}\{R_k\}_{k=1}^{5}$, and the angular uncertainty is $\sigma_\theta = \sigma_R \times s$.

4.5 Plate scale

The header value `CDELT1` $= 0.698\,\mathrm{arcsec\,px^{-1}}$ differs from the nominal $0.700\,\mathrm{arcsec\,px^{-1}}$ by only 0.3%, but over the disk radius of ~ 1342 px this accumulates to $2.7''$. Always reading the calibrated header value is therefore important.

5 Results

5.1 Full-disk images and limb detection

Figure 1 shows the three observations displayed with a 1st–99th percentile intensity stretch. The near-UV limb darkening appears as a smooth radial fade toward the disk edge. Figure 2 shows the detected limb pixels overlaid on each image; the boundary is continuous around the full 360° in every frame.

5.2 Best-fit circle

Figure 3 overlays the best-fit circle on each image. The circle tracks the visible limb closely on all sides, with residuals at the sub-pixel level.

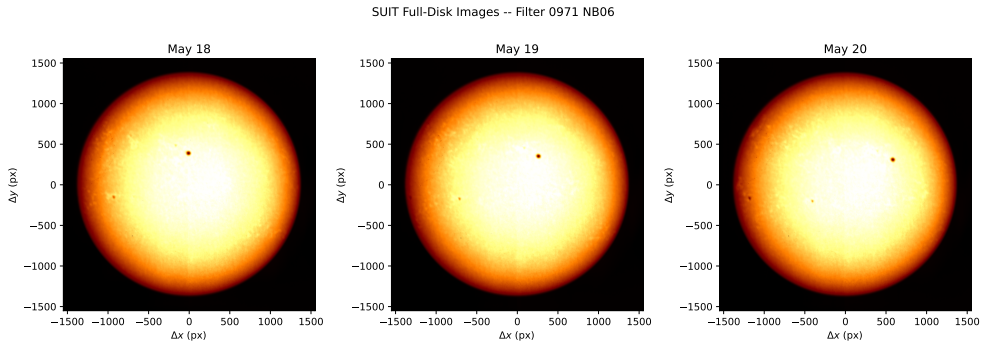


Figure 1: SUIT full-disk images on 2025 May 18, 19, and 20 (left to right), filter 0971 NB06, with a robust 1–99 percentile intensity stretch (`afmhot` colour map). Axes show pixel offsets $(\Delta x, \Delta y)$ from the fitted disk centre. The smooth radial fade toward the edge reflects near-UV limb darkening at 388 nm; the dark dot near the disk centre is a sunspot group.

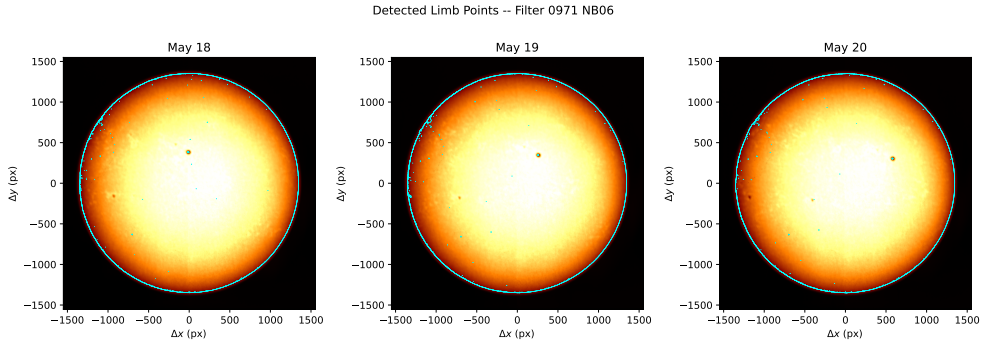


Figure 2: Detected limb pixels (cyan) overlaid on the three SUIT observations at threshold fraction $f = 0.26$. Axes show pixel offsets $(\Delta x, \Delta y)$ from the fitted disk centre. The boundary traces the full 360° circumference in each frame.

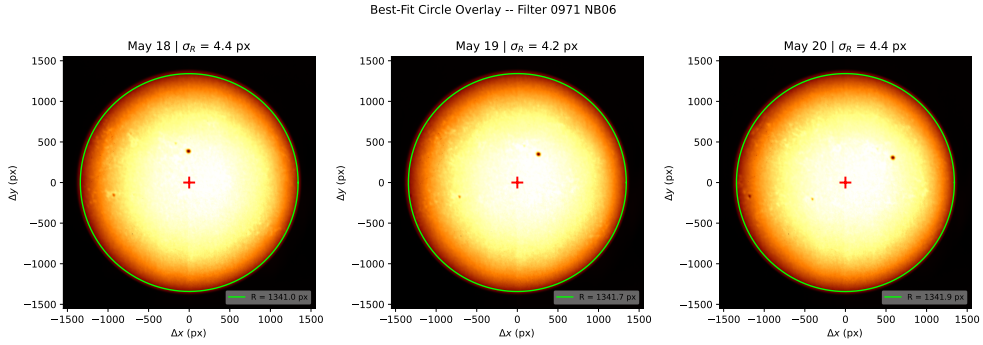


Figure 3: Best-fit circle (green) and disk centre (red cross, at the origin) overlaid on the three SUIT observations. Axes show pixel offsets $(\Delta x, \Delta y)$ from the fitted disk centre, so the cross sits at $(0, 0)$ by construction, consistent with Table 2. The panel title gives σ_R and the legend gives the fitted radius R.

5.3 Threshold sensitivity

Figure 4 shows how the fitted radius varies with threshold fraction. The curves decrease monotonically: a lower threshold fraction admits more near-limb pixels into the disk mask, expanding the boundary and yielding a marginally larger fitted radius, while a higher threshold contracts the boundary by excluding the faintest limb pixels. The standard deviation σ_R across the five tested values is 4.4, 4.2, and 4.4 px for the three observations, representing only $\approx 0.3\%$ of the total fitted radius. This confirms that the measurement is not strongly sensitive to the exact threshold value within the adopted range, and that σ_R provides a realistic estimate of the

threshold-induced uncertainty.

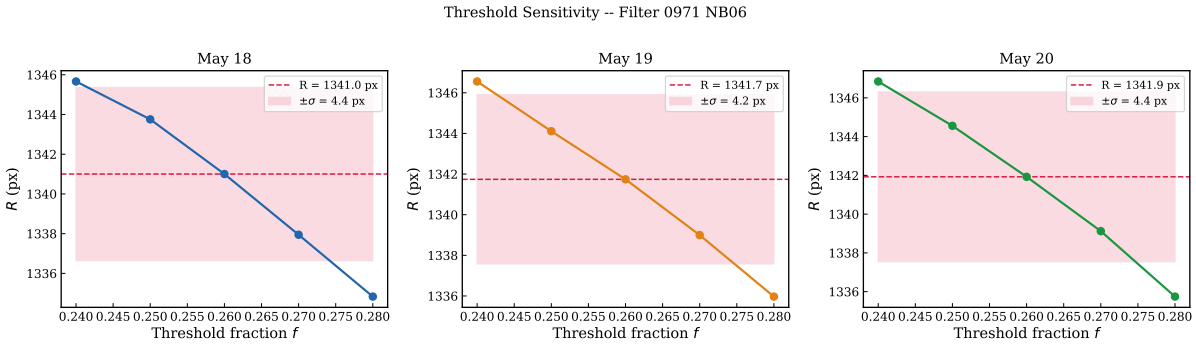


Figure 4: Fitted pixel radius R as a function of threshold fraction f for the three observations (left: May 18; centre: May 19; right: May 20). Each panel plots five data points corresponding to $f \in \{0.24, 0.25, 0.26, 0.27, 0.28\}$. The dashed horizontal line marks the best estimate at $f = 0.26$; the shaded band shows the $\pm\sigma_R$ interval, where σ_R is the standard deviation across the five values. The monotonic decrease of R with increasing f reflects the progressive exclusion of near-limb pixels as the threshold tightens.

5.4 Numerical results and three-day stability

Table 2 lists the full results for each day. Figure 5 shows the day-to-day variation. The fitted radius changes by only 0.9 px over three days ($\sigma = 0.49$ px), and the angular scatter is $0.34''$, both well below one pixel. The disk centre shifts by ≈ 26 px in y_0 on May 19, which is normal pointing variation and does not affect the radius.

Table 2: Fitted solar disk parameters for each observation. x_0 and y_0 are the fitted disk-centre coordinates in the full 4096×4096 detector frame; $R \pm \sigma_R$ is the fitted pixel radius with its threshold-sensitivity uncertainty; $\theta_\odot = R \times s$ is the derived angular radius (plate scale $s = 0.698$ arcsec px^{-1}); `RSUN_OBS` is the ephemeris value from the FITS header; and $\Delta\theta = \theta_\odot - $ `RSUN_OBS` is the residual.

Date	x_0 (px)	y_0 (px)	$R \pm \sigma_R$ (px)	$\theta_\odot$ ($''$)	RSUN_OBS ($''$)	$\Delta\theta$ ($''$)
2025-05-18	2069.3	2023.7	1341.0 ± 4.4	936.0	956.40	-20.4
2025-05-19	2060.7	2049.7	1341.7 ± 4.2	936.5	955.84	-19.3
2025-05-20	2063.5	2024.5	1341.9 ± 4.4	936.7	956.54	-19.9
Mean			1341.56 ± 0.49^a	936.4 ± 0.3^a	956.26 ± 0.37	-19.86 ± 0.54

[a] Standard deviation across three independent observations.

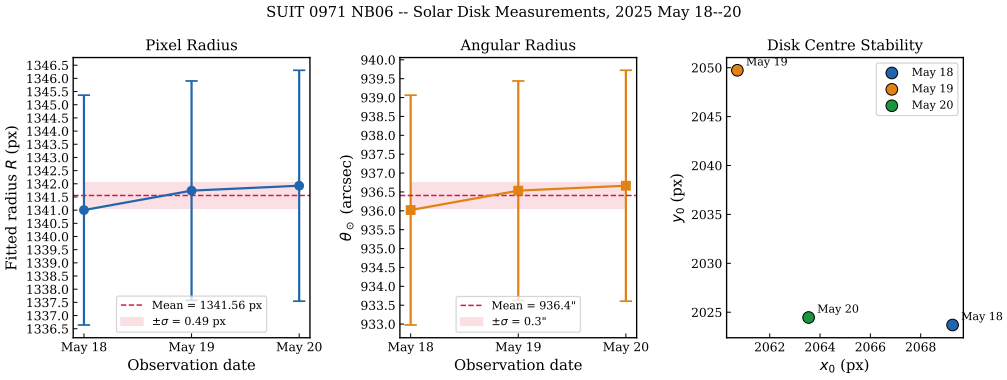


Figure 5: Day-to-day variation of the fitted pixel radius (left), angular radius (centre), and disk centre position (right) over 2025 May 18–20. Error bars in the left and centre panels are σ_R and σ_θ from the threshold sensitivity test; the centre-position panel carries no error bars. The ≈ 26 px scatter in y_0 reflects normal pointing variation of the spacecraft.

5.5 Angular radius comparison with RSUN_OBS

Figure 6 compares the measured $\theta_\odot$ with the pipeline value RSUN_OBS. The measured values fall consistently $\approx 19.86''$ below RSUN_OBS on all three days ($\sigma_{\Delta\theta} = 0.54''$), pointing to a stable systematic bias rather than random scatter.

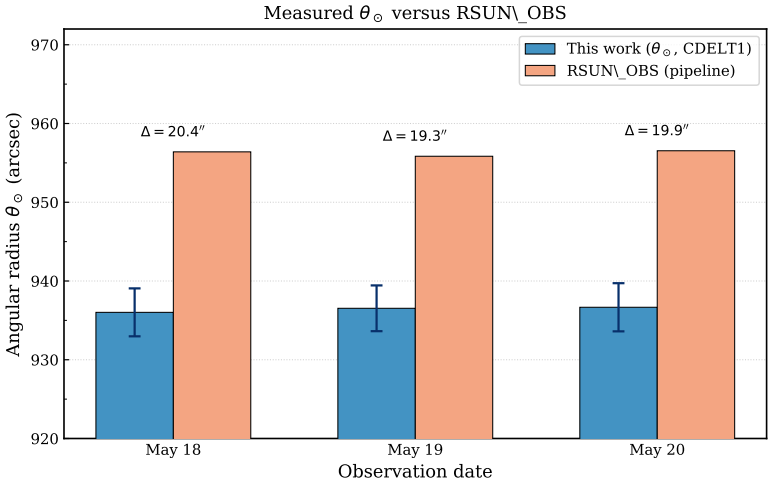


Figure 6: Measured angular radius $\theta_\odot$ (blue bars, with $\pm\sigma_\theta$ error bars) and the ephemeris value RSUN_OBS (orange bars) for the three days. The consistent $\approx 20''$ shortfall of the measured values below the ephemeris is a stable systematic bias attributable to the brightness-threshold limb definition (see Section 5.8).

5.6 Derived Sun-observer distance

Using the mean measured angular radius in Equation (2),

$$D = \frac{6.957 \times 10^5\,\mathrm{km}}{936.4'' \times \pi/648\,000} = 1.532 \times 10^8\,\mathrm{km} = 1.024\,\mathrm{AU}. \tag{8}$$

The spacecraft ephemeris gives DSUN_OBS $= 1.500 \times 10^8$ km (1.003 AU). The derived distance is 3.2×10^6 km (2.16%) larger. Since $D \propto \theta_\odot^{-1}$, the 2.08% underestimate in $\theta_\odot$ propagates directly to a $\approx 2.12\%$ overestimate in D; the small extra 0.04% comes from day-to-day scatter in DSUN_OBS. Figure 7 shows the comparison.

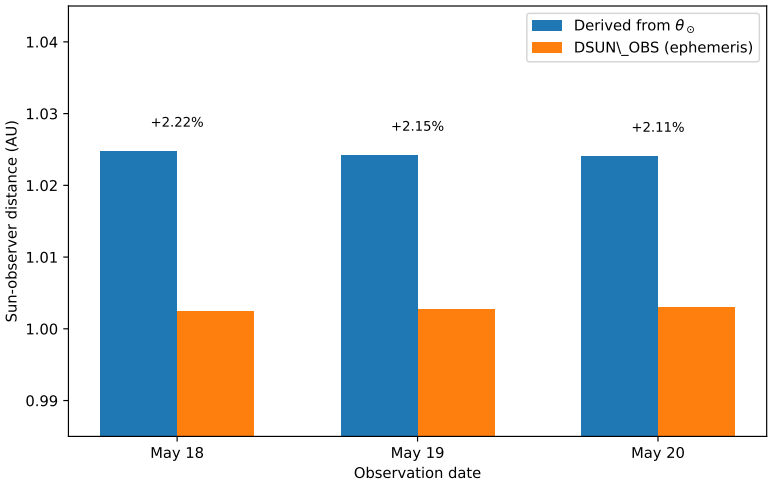


Figure 7: Sun-observer distance derived from $\theta_\odot$ (blue bars) versus the spacecraft ephemeris DSUN_OBS (orange bars). The percentage labels show the fractional overestimate on each day; the $\approx 2.2\%$ excess follows directly from the 2.08% underestimate in $\theta_\odot$, since $D \propto \theta_\odot^{-1}$.

5.7 Summary of measured solar radius in kilometres

The physical solar radius obtained in this work is computed by combining the mean measured angular radius with the mean spacecraft ephemeris distance (DSUN_OBS $= 1.500 \times 10^8$ km):

$$R_\odot^{\text{this work}} = D_{\text{eph}} \times \theta_\odot\,[\text{rad}] = 1.500 \times 10^8\,\text{km} \times \frac{936.4'' \times \pi}{648\,000} = 680{,}900\,\text{km}. \tag{9}$$

The uncertainty propagated from $\sigma_\theta = 0.3''$ is $\sigma_{R_\odot} = D_{\text{eph}} \times \sigma_\theta\,[\text{rad}] \approx 200$ km, giving $R_\odot^{\text{this work}} = (680{,}900 \pm 200)$ km.

Key Result: Solar Radius in Kilometres

Quantity	Value	Notes
Measured angular radius $\theta_\odot$	$936.4 \pm 0.3''$	Mean over 3 days (threshold method)
Our result: $R_\odot^{\text{this work}}$	$(680{,}900 \pm 200)$ km	Using $D_{\text{eph}} = 1.500 \times 10^8$ km
IAU 2015 accepted value $R_\odot$	695,700 km	[2]
Deviation $\Delta R_\odot$	$-14{,}800$ km (-2.1%)	Fully explained by limb-darkening bias

What this means: using a real spacecraft image and a few lines of Python, we measured the physical size of the Sun to within 14,800 km out of a total radius of nearly 700,000 km, a 2.1% agreement with the accepted value. The entire discrepancy is understood and expected from the method used.

5.8 Comparison with published solar radius measurements

Table 3 places the present result alongside a representative selection of published solar radius determinations, spanning classical ground-based heliometry, photoelectric transit timing, helioseismic limb modelling, and modern space photometry. The precision measurements cluster in the range 695,500–696,200 km, consistent with the IAU 2015 nominal value of 695,700 km [2]. Our value of $(680{,}900 \pm 200)$ km lies $\approx 14{,}800$ km (2.1%) below this cluster.

Table 3: Selected solar radius measurements from the literature compared with the present work. Uncertainties are as quoted in the original sources; entries without uncertainties are adopted nominal or classical values.

Method	Instrument / Source	$R_\odot$ (km)	Reference
Photoelectric meridian transit + limb model	Solar Diameter Monitor (HAO)	$695{,}508 \pm 26$	Brown & C-D (1998) [6]
Inflection-point limb + atmosphere model	SOHO/MDI	$695{,}660 \pm 140$	Haberreiter et al. (2008) [7]
Space photometry (Venus transit, 2012)	PICARD/SODISM	$696{,}134 \pm 261$	Meftah et al. (2018) [8]
IAU 2015 nominal	—	695,700	Prša et al. (2016) [2]
UV brightness threshold	*Aditya-L1/SUIT*	$(680{,}900 \pm 200)$	***This work***

The 2.1% shortfall of our result relative to precision determinations is a well-understood consequence of the limb-definition method. Precision techniques, whether photoelectric transit timing combined with limb-darkening models [6], atmosphere-model-based inflection-point location [7], or geometric calibration using a planetary transit [8], all locate the solar limb at or near the steepest intensity gradient in the solar atmosphere, which corresponds closely to the geometric photospheric limb. A brightness threshold, by contrast, identifies the contour where the image intensity crosses a fixed fraction of the peak brightness. In near-UV imagery at 388 nm, strong limb darkening causes intensity to fall off steeply inward of the geometric limb, so the threshold surface lies interior to the true limb; the resulting radius underestimate of $\sim 2\%$ is entirely predictable. The consistency of precision values

in Table 3, each obtained with an independent technique and instrument, further validates the IAU 2015 standard and places our measurement in proper context as a pedagogical result.

6 Discussion

6.1 Reproducibility

The sub-pixel stability ($\sigma_R = 0.49$ px over three days) shows that the pipeline is sensitive enough to detect the $\sim 3\%$ annual variation in the Sun's apparent size as Earth moves from perihelion to aphelion. Applying the same script to monthly SUIT images over a full year and comparing the resulting $\theta_\odot$ curve with Kepler's second law is a natural follow-up project.

6.2 Effect of the Instrumental Point-Spread Function

The instrumental point-spread function (PSF) of SUIT broadens the apparent limb by convolving the true intensity profile with the telescope's diffraction and aberration kernel, in principle shifting the threshold-crossing point and biasing the measured radius. For SUIT at 388 nm, the PSF full-width at half-maximum is of order one pixel, comparable to the sub-pixel Kåsa fitting residuals reported here. Within our threshold-based method the PSF effect is thus subsumed into the overall limb-definition bias discussed in Section 5.8, and does not introduce a significant additional uncertainty beyond the $\sigma_R \approx 4$ px already accounted for by the threshold sensitivity test.

6.3 Future Outlook

The pipeline developed here can be extended in several directions. Applying the same script to monthly SUIT observations over a full year would allow the annual variation in $\theta_\odot$ to be reconstructed and compared with the orbital distance variation predicted by Kepler's second law. Repeating the measurement in a broadband SUIT filter and comparing $\theta_\odot$ across spectral bands would show the wavelength dependence of limb darkening and the height of the emitting layer. On the algorithmic side, replacing the Kåsa circle fit with the Taubin algebraic fit [4] would illustrate the difference between two geometric fitting strategies. Finally, an inflection-point limb detector, which locates the solar limb at the steepest intensity gradient rather than at a fixed brightness threshold, would be expected to yield a radius closer to the precision values in Table 3, providing a direct demonstration of the bias introduced by the threshold method.

7 Data-Use Notes

Aditya-L1 data are provided through ISSDC/PRADAN for non-profit scientific use, with ISRO retaining copyright. Users must follow the PRADAN data-use terms (see footnote 1 for the disclaimer URL). The required acknowledgement text, mandated for publications using these data (see footnote 1 for the acknowledgement URL), is included in the Acknowledgements section below.

Acknowledgements

All computational work presented in this paper was performed on the Sikkim University *Brahmagupta* High Performance Computing (HPC) facility. One of the authors (R.M.) acknowledges support from the Inter-University Centre for Astronomy and Astrophysics (IUCAA), Pune, through the Visiting Associate Programme.

Required Aditya-L1 data acknowledgement (as mandated by PRADAN/ISSDC; see footnote 1): Aditya-L1 is an observatory class mission which is fully funded and operated by the Indian Space Research Organization (ISRO). The mission was conceived and realised with the help from various ISRO centres. The science payloads and science ready data products are realised by the payload PI institutes in close collaboration with ISRO centres. The PI institutes are: Indian Institute of Astrophysics (IIA); Inter University Centre for Astronomy and Astrophysics (IUCAA); Laboratory for Electro-optics Systems (LEOS/URSC); Physical Research Laboratory (PRL); U R Rao Satellite Centre (URSC); and Space Physics Laboratory (SPL/VSSC). We acknowledge the use of data from the Aditya-L1 mission of the Indian Space Research Organisation (ISRO), archived at the Indian Space Science Data Centre (ISSDC).

A Reproducibility Requirements

The following resources are required to reproduce all results and figures in this paper.

What You Need to Reproduce This Work

Data: Three Aditya-L1/SUIT Level-1 FITS images, filter `0971NB06`, dates 2025 May 18–20. Download free of charge from the PRADAN portal at `https://pradan.issdc.gov.in/al1`.

Software: Python 3 with four standard packages: `numpy`, `matplotlib`, `astropy`, `scipy`. All are available via `pip install` or `conda install`.

Hardware: Any standard laptop or desktop computer; no GPU or special hardware required. The full script runs in a few minutes.

Script: The complete, self-contained Python script is given in Appendix B. Running the six blocks in order reproduces every figure in this paper.

B Complete Reproducible Python Script

Run the six blocks below in order with the three SUIT FITS files in the working directory (search the PRADAN portal for Aditya-L1/SUIT Level-1 data, dates 2025 May 18–20, filter `0971NB06`). Every figure in this paper is produced without additional code. Required packages: `numpy`, `matplotlib`, `astropy`, `scipy`.

B.1 Imports and global constants

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
from astropy.io import fits
from scipy.ndimage import label
import os

FITS_FILES = [
    "SUT_T25_0687_000954_Lev1.0_2025-05-18T13.47.35.107_0971NB06.fits",
    "SUT_T25_0687_000957_Lev1.0_2025-05-19T11.01.22.282_0971NB06.fits",
    "SUT_T25_0687_000960_Lev1.0_2025-05-20T13.33.47.638_0971NB06.fits",
]
DATES_SHORT   = ['May 18', 'May 19', 'May 20']
THR_FRAC_LIST = [0.24, 0.25, 0.26, 0.27, 0.28]
AU_KM         = 149597870.7
AU_M          = AU_KM * 1e3
R_SUN_KM      = 6.957e5
OUT_DIR       = "suit_output"
os.makedirs(OUT_DIR, exist_ok=True)
```

B.2 Image-processing functions

Four functions used throughout: `robust_norm` applies a 1–99 percentile stretch for display; `make_disk_mask` thresholds and isolates the largest connected region; `get_limb_points` extracts boundary pixels; `fit_circle_kasa` fits the Kåsa algebraic circle.

```
def robust_norm(img):
    p1, p99 = np.nanpercentile(img, [1, 99])
    if p99 <= p1:
        return img
    return np.clip((img - p1) / (p99 - p1), 0.0, 1.0)

def make_disk_mask(img, thr_frac):
    finite = np.isfinite(img)
    vals   = img[finite]
    ref    = np.nanmedian(vals[vals >= np.nanpercentile(vals, 80)])
    raw    = (img >= thr_frac * ref) & finite
```

```
    lbl, n = label(raw)
    if n == 0:
        raise ValueError("No connected component found.")
    sz    = np.bincount(lbl.ravel())
    sz[0] = 0
    return lbl == sz.argmax(), thr_frac * ref

def get_limb_points(mask):
    m    = mask.astype(np.uint8)
    edge = (m == 1) & (
        (np.roll(m,  1, 0) == 0) | (np.roll(m, -1, 0) == 0) |
        (np.roll(m,  1, 1) == 0) | (np.roll(m, -1, 1) == 0)
    )
    ys, xs = np.where(edge)
    return xs.astype(float), ys.astype(float)

def fit_circle_kasa(x, y):
    if len(x) < 10:
        raise ValueError("Too few limb points.")
    xm, ym = x.mean(), y.mean()
    xn, yn = x - xm, y - ym
    sc     = np.sqrt((xn**2 + yn**2).mean())
    xn    /= sc
    yn    /= sc
    D      = np.column_stack([xn, yn, np.ones(len(xn))])
    z      = xn**2 + yn**2
    (A, B, C), *_ = np.linalg.lstsq(D, -z, rcond=None)
    R2 = (A / 2)**2 + (B / 2)**2 - C
    if R2 <= 0:
        raise ValueError("Circle fit failed.")
    return -A / 2 * sc + xm, -B / 2 * sc + ym, np.sqrt(R2) * sc
```

B.3 Per-file processing

Runs the full pipeline over all five threshold fractions for each file and returns the best-estimate radius, threshold sensitivity σ_R, angular radius, and derived Sun-observer distance.

```
def process_one_file(fpath, thr_list):
    with fits.open(fpath) as h:
        img = np.array(h[0].data, dtype=float)
        hdr = h[0].header
    if img.ndim == 3:
        img = img[0]
    ps       = abs(float(hdr['CDELT1']))
    rsun_obs = float(hdr['RSUN_OBS'])
    dsun_obs = float(hdr['DSUN_OBS'])
    res = []
    for tf in thr_list:
        mask, _ = make_disk_mask(img, tf)
        xs, ys  = get_limb_points(mask)
        if len(xs) > 60000:
            idx    = np.random.default_rng(42).choice(
                         len(xs), 60000, replace=False)
            xs, ys = xs[idx], ys[idx]
        x0, y0, R = fit_circle_kasa(xs, ys)
        res.append(dict(tf=tf, x0=x0, y0=y0, R=R, mask=mask))
    R_arr   = np.array([r['R'] for r in res])
    best    = res[len(res) // 2]
    sigma_R = np.std(R_arr, ddof=1)
    theta   = best['R'] * ps
    sig_th  = sigma_R   * ps
    D_km    = R_SUN_KM / (theta * np.pi / 648000)
    return dict(
        img=img, best=best, res=res,
        R_arr=R_arr, sigma_R=sigma_R,
        theta=theta, sig_th=sig_th,
        rsun_obs=rsun_obs, dsun_obs=dsun_obs,
        D_km=D_km, ps=ps,
```

```
    )

all_data = [process_one_file(f, THR_FRAC_LIST) for f in FITS_FILES]
```

B.4 Figures 1–3: full-disk images, limb points, circle overlay

The crop half-width `r = 1550` px contains the full solar disk ($R \approx 1342$ px) with a ~ 200 px margin. Axes are expressed as pixel offsets $(\Delta x, \Delta y)$ from the fitted disk centre using the `extent` argument of `imshow`, so the disk centre appears at $(0, 0)$ in every panel.

```
def save3(fname, fig):
    fig.savefig(f"{OUT_DIR}/{fname}", dpi=150, bbox_inches='tight')
    plt.close(fig)

t = np.linspace(0, 2 * np.pi, 1200)

fig, axes = plt.subplots(1, 3, figsize=(14, 4.8))
for i, d in enumerate(all_data):
    cx, cy = d['best']['x0'], d['best']['y0']
    ox     = max(0, int(cx) - 1550)
    oy     = max(0, int(cy) - 1550)
    sub    = d['img'][oy:min(4096,oy+3100), ox:min(4096,ox+3100)]
    ext    = [ox-cx, ox+3100-cx, oy-cy, oy+3100-cy]
    axes[i].imshow(robust_norm(sub), origin='lower', cmap='afmhot', extent=ext)
    axes[i].set_xlim(ext[0], ext[1])
    axes[i].set_ylim(ext[2], ext[3])
    axes[i].set_title(DATES_SHORT[i])
    axes[i].set_xlabel(r'$\Delta x$ (px)')
    axes[i].set_ylabel(r'$\Delta y$ (px)')
fig.suptitle('SUIT Full-Disk Images -- Filter 0971 NB06', y=1.01)
fig.tight_layout()
save3('fig1_fulldisk_3panel.pdf', fig)

fig, axes = plt.subplots(1, 3, figsize=(14, 4.8))
for i, d in enumerate(all_data):
    cx, cy      = d['best']['x0'], d['best']['y0']
    ox          = max(0, int(cx) - 1550)
    oy          = max(0, int(cy) - 1550)
    sub         = d['img'][oy:min(4096,oy+3100), ox:min(4096,ox+3100)]
    ext         = [ox-cx, ox+3100-cx, oy-cy, oy+3100-cy]
    xs_b, ys_b  = get_limb_points(d['best']['mask'])
    axes[i].imshow(robust_norm(sub), origin='lower', cmap='afmhot', extent=ext)
    axes[i].scatter(xs_b - cx, ys_b - cy, s=0.4, c='cyan', linewidths=0)
    axes[i].set_xlim(ext[0], ext[1])
    axes[i].set_ylim(ext[2], ext[3])
    axes[i].set_title(DATES_SHORT[i])
    axes[i].set_xlabel(r'$\Delta x$ (px)')
    axes[i].set_ylabel(r'$\Delta y$ (px)')
fig.suptitle('Detected Limb Points -- Filter 0971 NB06', y=1.01)
fig.tight_layout()
save3('fig2_limbpts_3panel.pdf', fig)

fig, axes = plt.subplots(1, 3, figsize=(14, 4.8))
for i, d in enumerate(all_data):
    cx, cy = d['best']['x0'], d['best']['y0']
    ox     = max(0, int(cx) - 1550)
    oy     = max(0, int(cy) - 1550)
    sub    = d['img'][oy:min(4096,oy+3100), ox:min(4096,ox+3100)]
    ext    = [ox-cx, ox+3100-cx, oy-cy, oy+3100-cy]
    xc     = d['best']['R'] * np.cos(t)
    yc     = d['best']['R'] * np.sin(t)
    axes[i].imshow(robust_norm(sub), origin='lower', cmap='afmhot', extent=ext)
    axes[i].plot(xc, yc, color='lime', lw=1.4,
                 label=f"R = {d['best']['R']:.1f} px")
    axes[i].plot(0, 0, 'r+', ms=12, mew=2)
    axes[i].legend(loc='lower right', fontsize=8,
                   framealpha=0.4, edgecolor='none')
    axes[i].set_xlim(ext[0], ext[1])
```

```
    axes[i].set_ylim(ext[2], ext[3])
    axes[i].set_title(
        f"{DATES_SHORT[i]} | $\\sigma_R$ = {d['sigma_R']:.1f} px")
    axes[i].set_xlabel(r'$\Delta x$ (px)')
    axes[i].set_ylabel(r'$\Delta y$ (px)')
fig.suptitle('Best-Fit Circle Overlay -- Filter 0971 NB06', y=1.01)
fig.tight_layout()
save3('fig3_circleoverlay_3panel.pdf', fig)
```

B.5 Figures 4–5: threshold sensitivity and three-day summary

```
fig, axes = plt.subplots(1, 3, figsize=(14, 4.0))
for i, d in enumerate(all_data):
    tf_arr = np.array([r['tf'] for r in d['res']])
    axes[i].plot(tf_arr, d['R_arr'], 'o-')
    axes[i].axhline(d['best']['R'], color='crimson', ls='--',
                    label=f"R = {d['best']['R']:.1f} px")
    axes[i].fill_between(tf_arr,
                         d['best']['R'] - d['sigma_R'],
                         d['best']['R'] + d['sigma_R'],
                         alpha=0.2,
                         label=f"$\\pm\\sigma_R$ = {d['sigma_R']:.1f} px")
    axes[i].set_xlabel('Threshold fraction $f$')
    axes[i].set_ylabel('$R$ (px)')
    axes[i].set_title(DATES_SHORT[i])
    axes[i].legend()
fig.tight_layout()
save3('fig4_sensitivity_3panel.pdf', fig)

R_PX    = np.array([d['best']['R']  for d in all_data])
SIGMA_R = np.array([d['sigma_R']    for d in all_data])
THETA   = np.array([d['theta']      for d in all_data])
SIG_TH  = np.array([d['sig_th']     for d in all_data])
X0      = np.array([d['best']['x0'] for d in all_data])
Y0      = np.array([d['best']['y0'] for d in all_data])

fig = plt.figure(figsize=(13, 4.2))
gs  = gridspec.GridSpec(1, 3, figure=fig, wspace=0.38)

ax0    = fig.add_subplot(gs[0])
R_mean = R_PX.mean()
R_std  = R_PX.std(ddof=1)
ax0.errorbar(DATES_SHORT, R_PX, yerr=SIGMA_R, fmt='o-', capsize=4)
ax0.axhline(R_mean, color='crimson', ls='--',
            label=f'Mean = {R_mean:.2f} px')
ax0.fill_between(DATES_SHORT, R_mean - R_std, R_mean + R_std, alpha=0.15)
ax0.set_ylabel('Fitted radius $R$ (px)')
ax0.set_xlabel('Observation date')
ax0.set_title('Pixel Radius')
ax0.legend()

ax1     = fig.add_subplot(gs[1])
th_mean = THETA.mean()
th_std  = THETA.std(ddof=1)
ax1.errorbar(DATES_SHORT, THETA, yerr=SIG_TH, fmt='s-', capsize=4)
ax1.axhline(th_mean, color='crimson', ls='--',
            label=f'Mean = {th_mean:.1f}"')
ax1.fill_between(DATES_SHORT, th_mean - th_std, th_mean + th_std,
                 alpha=0.15)
ax1.set_ylabel(r'$\theta_\odot$ (arcsec)')
ax1.set_xlabel('Observation date')
ax1.set_title('Angular Radius')
ax1.legend()

ax2 = fig.add_subplot(gs[2])
for i in range(3):
    ax2.scatter(X0[i], Y0[i], s=90, label=DATES_SHORT[i])
    ax2.annotate(DATES_SHORT[i], (X0[i], Y0[i]),
```

```
                 textcoords='offset points', xytext=(7, 4))
ax2.set_xlim(2058, 2072)
ax2.set_xlabel('$x_0$ (px)')
ax2.set_ylabel('$y_0$ (px)')
ax2.set_title('Disk Centre Stability')
ax2.legend()

fig.savefig(f"{OUT_DIR}/fig5_summary_panel.pdf",
            dpi=150, bbox_inches='tight')
plt.close()
```

B.6 Figures 6–7: angular radius and distance comparisons

```
RSUN_OBS = np.array([d['rsun_obs'] for d in all_data])
x = np.arange(3)
w = 0.32

fig, ax = plt.subplots(figsize=(7.5, 4.8))
ax.bar(x - w/2, THETA,    w,
       label=r'This work ($\theta_\odot$, CDELT1)')
ax.bar(x + w/2, RSUN_OBS, w,
       label=r'RSUN\_OBS (pipeline)')
ax.errorbar(x - w/2, THETA, yerr=SIG_TH, fmt='none', capsize=4)
for j in range(3):
    diff = RSUN_OBS[j] - THETA[j]
    ax.annotate(f'$\\Delta={diff:.1f}^{{\\prime\\prime}}$',
                xy=(x[j], max(THETA[j], RSUN_OBS[j]) + 1.5),
                ha='center', va='bottom')
ax.set_xticks(x)
ax.set_xticklabels(DATES_SHORT)
ax.set_ylabel(r'Angular radius $\theta_\odot$ (arcsec)')
ax.set_xlabel('Observation date')
ax.legend()
ax.set_ylim(920, 972)
fig.tight_layout()
fig.savefig(f"{OUT_DIR}/fig6_theta_comparison.pdf",
            dpi=150, bbox_inches='tight')
plt.close()

DSUN_AU  = np.array([d['dsun_obs'] / AU_M  for d in all_data])
D_DER_AU = np.array([d['D_km']     / AU_KM for d in all_data])

fig, ax = plt.subplots(figsize=(7.5, 4.8))
ax.bar(x - w/2, D_DER_AU, w,
       label=r'Derived from $\theta_\odot$')
ax.bar(x + w/2, DSUN_AU,  w,
       label=r'DSUN\_OBS (ephemeris)')
for j in range(3):
    diff_pct = 100 * (D_DER_AU[j] - DSUN_AU[j]) / DSUN_AU[j]
    ax.annotate(f'{diff_pct:+.2f}%',
                xy=(x[j], max(D_DER_AU[j], DSUN_AU[j]) + 0.003),
                ha='center', va='bottom')
ax.set_xticks(x)
ax.set_xticklabels(DATES_SHORT)
ax.set_ylabel('Sun-observer distance (AU)')
ax.set_xlabel('Observation date')
ax.legend()
ax.set_ylim(0.985, 1.045)
fig.tight_layout()
fig.savefig(f"{OUT_DIR}/fig7_distance_comparison.pdf",
            dpi=150, bbox_inches='tight')
plt.close()
```

References

[1] I. Kåsa, "A circle fitting procedure and its error analysis," *IEEE Trans. Instrum. Meas.*, **25**, 8–14, 1976.

[2] A. Prša et al., "Nominal values for selected solar and planetary quantities: IAU 2015 Resolution B3," *Astron. J.*, **152**, 41, 2016.

[3] D. Tripathi et al., "The Solar Ultraviolet Imaging Telescope on board Aditya-L1," *Sol. Phys.*, **300**, 30, 2025. arXiv:2501.02274, https://arxiv.org/abs/2501.02274.

[4] G. Taubin, "Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations, with applications to edge and range image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**(11), 1115–1138, 1991.

[5] W. T. Thompson, "Coordinate systems for solar image data," *Astron. Astrophys.*, **449**, 791–803, 2006.

[6] T. M. Brown and J. Christensen-Dalsgaard, "Accurate determination of the solar photospheric radius," *Astrophys. J.*, **500**, L195–L198, 1998.

[7] M. Haberreiter, W. Schmutz, and A. G. Kosovichev, "Solving the discrepancy between the seismic and photospheric solar radius," *Astrophys. J.*, **675**, L53–L56, 2008.

[8] M. Meftah et al., "Solar radius determined from PICARD/SODISM observations and extremely weak wavelength dependence in the visible and the near-infrared," *Astron. Astrophys.*, **616**, A64, 2018.

PANE Journal *of* Physics Education

PANE Journal of Physics Education is the Education Journal of the Physics Academy of the North East (PANE). It is a peer-reviewed e-journal devoted to publishing original, high-quality research articles on the teaching and learning of physics which features contributions on classroom and laboratory instruction, original insights into derivations in classical and modern physics, apparatus notes, and historical or cultural perspectives, all aimed at sustaining and enhancing the quality of physics education in colleges, universities, and other higher education institutions.

The journal operates as an unpaid open-access journal, providing free access to all published articles for readers and authors.

https://journals.paneindia.org/index.php/pjpe

PANE Journal of Physics Education Volume 1, No. 1, April 2026