

A Simple Derivation of Dirac's Equation

TM Coker MA PhD CEng FIET MInstP*

May 27, 2024

Abstract

This is the first of a series of short papers exploring various aspects of quantum mechanics and quantum field theory. The intent of the full series of articles is to take the reader/student from a basic starting point (somewhere around high school / first year undergraduate maths/ physics/ engineering) to an understanding of relativistic quantum mechanics that would be appropriate for a third/fourth year undergraduate or early stage postgraduate.

This particular article is a derivation of Dirac's Equation.

1 Introduction

1.1 Series Overview

The intent of this series of articles is to allow the reader/student to understand the basic concepts of quantum mechanics and quantum field theory, but with a starting point of comparatively basic maths and physics, such as a first year undergraduate studying maths, physics or engineering might have. The prerequisites are:

- Vectors and basic matrix algebra including eigenvectors and eigenvalues.
- Partial differentiation and vector calculus.
- A purely qualitative knowledge of quantum mechanics and special relativity.

The original motivation for this work was a desire to understand the Higg's Boson and how it somehow "creates" mass. To get to that endpoint it turns out it's necessary to understand special (but thankfully not general) relativity, group theory, Lagrangians and local gauge invariance, Hamilton's Principle, the calculus of variations and electro-weak unification. All these topics, and others, will be introduced along the way.

*Non-affiliated author.

Whilst this article is chronologically the first to be released, it isn't really the first in the series (due to entirely practical reasons), as such there is a very brief section on special relativity, covariance and contravariance that will be covered in more depth in a subsequent article.

“Natural units” where $\hbar = c = 1$ will be used throughout¹, with these units, mass, energy and momentum are dimensionally equal, as are length and time which is useful for the concept of spacetime.

1.2 Covariant and Contravariant Vectors and Special Relativity

Notwithstanding the intent stated above, for a more or less complete understanding of this paper, particularly Sections 2.2 and 2.3, a knowledge of covariant and contravariant vectors is required, particularly as they pertain to special relativity so a basic understanding of the latter is also required.

The issue of covariance and contravariance is not necessarily difficult, but such difficulties as do exist are compounded by different authors using different terminology. For example some texts refer to contravariant vectors (or just vectors) and “one-forms” instead of covariant vectors (Gray (2019) and Barr et al. (2016) for example).

One quick way to access the subtleties is to consider that for spatial co-ordinates the dimension of the co-ordinates is that of *length* (strictly using natural units the dimension is inverse-energy). Vector quantities such as displacement, velocity and so on, where the dimension of the co-ordinate is in the numerator (ie *metres* per second) are typically contravariant. Conversely vectors such as the electrical or magnetic field strength, or the gradient of a scalar field, where the co-ordinate dimension is in the *denominator* (eg volts per metre) are covariant.

Taking the discussion a bit further, whilst all authors invariably refer to vectors as being contravariant or covariant this is not strictly accurate. A vector is simply a vector, what we are really referring to is the co-ordinate system we use to describe the vector. Within the co-ordinate system there are “basis vectors” – for example \hat{i} , \hat{j} and \hat{k} in cartesian co-ordinates – and to describe a vector we resolve it against each of the basis vectors to generate the “components”. However, without going into the detail, there are two ways this can be done, covariantly and contravariantly.

Having created one set of components for one co-ordinate system if we then change the co-ordinate system we expect these components to change. This is where the difference between the two types of vector arises as the different types of component transform differently when we change the co-ordinate system.

A further point we can make is that if we are using Euclidean space where the co-ordinate system is rectilinear, then the difference between covariant and contravariant components largely disappears. But, as we will see below, Special Relativity does not operate in Euclidean space, hence there is a difference and we need to use this difference to understand what is going on.

¹Again a topic to be covered in a later paper, but see almost any quantum mechanical text for a further description

Whilst the above is sufficient for this paper [Fleisch \(2012\)](#) is a good primer for this particular topic.

1.2.1 Special Relativity

This sub-section is a very quick overview of special relativity, such that later sections can be read without needing to consult additional texts. Relativity is the study of physical quantities in different frames of reference. Where these frames of reference are moving *relative* to each other at a constant velocity they are referred to as *inertial* frames of reference and we are studying Special Relativity. The constant relative velocity aspect is crucial, if the velocity is not constant the frames are not inertial and we are studying General Relativity instead, which is much harder.

Quantities measured in one frame of reference can be different when measured in a different frame, they *transform* between the two frames. Here the most obvious example is velocity – a body at rest in one frame is clearly moving in another frame (the example of people on trains is pretty ubiquitous here for some reason). However some quantities do not change between inertial frames, they are invariant or “Lorentz Invariant”. The speed of light may be thought of as invariant, but this isn’t really a good example as the invariance of the speed of light to Lorentz transformations is actually axiomatic for relativity and is the starting point for deriving the rest of the maths.

We now introduce what is called the “metric tensor” ($g_{\mu\nu} = g^{\mu\nu}$), which is a tensor that defines how spacetime is put together. It is the case that for any contravariant vector x^μ we can obtain the covariant vector by contracting with the metric tensor:

$$x_\nu = g_{\mu\nu}x^\mu$$

We should pause here for a moment and consider the message from Section 1.2 above. There we said that it is not the vectors that are covariant or contravariant but their components. Hopefully this makes a bit more sense now – if the vector was intrinsically contravariant, how can we simply multiply it with the metric to make it covariant? The answer is we can’t but what we can do is calculate what the covariant components will be if we know the contravariant ones and the metric.

Unfortunately at this point we have been forced to use tensor notation, which is not entry-level. However we can convert to matrix notation although we may at times need to keep the sub and superscript indices in order to distinguish a contravariant vector from its covariant equivalent. In this case if x is a contravariant vector then we write the covariant form as simply gx , sometimes with brackets.

By definition the Lorentz transformation must leave the scalar product of a contravariant vector with a covariant one unchanged. Using matrix notation the scalar product of two vectors is $x^T y$ so the Lorentz scalar product becomes $x^T(gy)$. If we use Λ as the transformation matrix then requiring the scalar product to be invariant means:

$$\begin{aligned} x^T(gy) &= (\Lambda x)^T g(\Lambda y) \\ &= x^T \Lambda^T g(\Lambda y) \end{aligned}$$

As this must be true for any x and y we get:

$$g = \Lambda^T g \Lambda \quad (1)$$

At this point we would like to know what g is, in order to help let us first note that we could swap the order of the vectors in the scalar product, ie start with $(gx)^T y$ which leads us to:

$$\begin{aligned} (gx)^T y &= (g\Lambda x)^T \Lambda y \\ x^T g^T y &= x^T \Lambda^T g^T \Lambda y \end{aligned}$$

Comparing the last line with Equation (1) leads us to say that g must be symmetric ie $g^T = g$, which is a start. However to proceed further we need to recognise that the vectors we are considering do not live in “ordinary” Euclidean space. If they did then the the magnitude squared of a vector would be $|x|^2 = (t^2 + x_1^2 + x_2^2 + x_3^2)$. Instead we are in *Minkowski* space and in fact $|x|^2 = (t^2 - x_1^2 - x_2^2 - x_3^2)$. This quantity is often called the “spacetime interval” and it is axiomatic to Special Relativity that the spacetime interval is Lorentz Invariant . Proceeding as before:

$$\begin{aligned} x^T(gx) &= \begin{bmatrix} t & x_1 & x_2 & x_3 \end{bmatrix} g \begin{bmatrix} t \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= t^2 - x_1^2 - x_2^2 - x_3^2 \end{aligned}$$

from where it is fairly obvious that

$$g \begin{bmatrix} t \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t \\ -x_1 \\ -x_2 \\ -x_3 \end{bmatrix}$$

and therefore

$$g = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

which, happily, is symmetric as required.

This metric is often referred to by the shorthand $(+, -, -, -)$ and we note that there is an element of convention here, where some authors use $(-, +, +, +)$ instead.

If we now quickly flip back to Euclidean space, then if we follow the argument above we would come up with a metric of $(+, +, +, +)$ which would lead us to see that the components of x^μ are equal to those of x_μ , which is why the distinction in Euclidean space is mostly academic.

All the above was a slightly long-winded way to get us to the point where we can start talking about the four dimensional partial differential operator (or four-gradient) ∂_μ , which for the moment we simply define:

$$\partial_\mu = \left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

It can be shown² that if an object ϕ is a four-vector then $\partial_\mu \phi$ is also a four-vector. In this way we can say that ∂_μ is also a four-vector (or at least behaves like one) and we can therefore go further:

$$\partial^\mu = \left(\frac{\partial}{\partial t}, -\frac{\partial}{\partial x}, -\frac{\partial}{\partial y}, -\frac{\partial}{\partial z} \right)$$

and hence

$$\begin{aligned} \partial^\mu \partial_\mu &= \left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2} \right) \\ &= \left(\frac{\partial^2}{\partial t^2} - \nabla^2 \right) = \square \end{aligned} \quad (2)$$

where \square is known as the d'Alembertian operator. We see that \square is the product of a contravariant four-vector with a covariant one and therefore is itself Lorentz Invariant .

1.3 Canonical Quantisation

Canonical Quantisation (also called first quantisation) is based on the postulate³ that the quantum mechanical wavefunction contains all the information we would ever need to know about a particle, and that we can “access” this information with an operator. More mathematically, the wavefunction is an eigenfunction (or eigenstate) of the operator and the associated eigenvalue is the quantity we seek.

$$\hat{O}\psi = a\psi$$

Here \hat{O} is (obviously) the operator in question and a the relevant eigenvalue and by the way we know from matrix algebra that if a is to be real (ie an actual observable quantity) then the associated matrix (or operator in this context) has to be Hermitian.

For the purposes of this article we will need to know the energy and momentum of a particle, so the operators that we need are:

$$\hat{E} = i \frac{\partial}{\partial t} \quad (3)$$

$$\hat{\mathbf{p}} = -i\nabla \quad (4)$$

2 Dirac's Equation

After these prerequisites we finally turn to Dirac's Equation which is one of the most significant theoretical achievements in quantum mechanics and is the mathematical reason why the wave functions which describe spin-half particles (which are not, as you might expect, called Diracions) are known to be spinors.

Prior to Dirac publishing his famous equation, other researchers had developed various “precursors” and in order to understand Dirac's motivation it

²Proof reserved for a later paper.

³A postulate is an assumption we make, that we can't prove, that we hope is true and no experiment has shown us to be wrong, yet.

will be necessary to discuss these precursors. The detailed derivation here (Section 2.3) largely follows that of Penrose (2004).

2.1 Schrödinger's Equation

Schrödinger's Equation is a semi-classical statement of energy. Applying Equation (3) to a wavefunction ψ we get:

$$i\frac{\partial}{\partial t}\psi = H\psi \quad (5)$$

where H is the total energy of the particle (referred to as the Hamiltonian). The total energy is the sum of kinetic and potential energy and non-relativistically this is:

$$H = \frac{\mathbf{p}^2}{2m} + V \quad (6)$$

where \mathbf{p} is the momentum, the first term is the kinetic energy and V represents some sort of energy potential. After canonical quantisation we get the quantum Hamiltonian *operator*:

$$\hat{H} = \frac{-1}{2m}\nabla^2 + V \quad (7)$$

where \hat{H} is more usually written simply as \mathcal{H} . Substituting Equation (7) into Equation (5) we get:

$$i\frac{\partial}{\partial t}\psi = \frac{-1}{2m}\nabla^2\psi + V\psi \quad (8)$$

which is Schrödinger's Equation in three spatial dimensions.

From Dirac's point of view there are two things wrong with this equation:

- The definition of kinetic energy is non-relativistic.
- The equation as a whole is not Lorentz Invariant as it is first order in the time component and second order in the spatial dimensions.

2.2 The Klein–Gordon Equation

The first attempt at a relativistic wave equation was ironically derived by Schrödinger, but for the reasons explained below he dropped it in favour of his later much more famous equation.

We start with Einstein's energy-momentum equation (which is of course relativistic):

$$E^2 = \mathbf{p}^2 + m^2$$

If we now substitute the operators from Equations 3 and 4 into this equation we get:

$$\frac{\partial^2\psi}{\partial t^2} = \nabla^2\psi - m^2\psi$$

Rearranging this equation and substituting the definition of \square from Equation (2) we get

$$(\square + m^2)\psi = 0 \quad (9)$$

which is known as the Klein-Gordon Equation. For this equation the m is the mass in the rest frame, which nowadays is called the “invariant” mass. From the section on special relativity we know that \square is also invariant meaning the the Klein-Gordon Equation is Lorentz Invariant ⁴.

As it turns out, this equation has pros and cons and nowadays is primarily used to describe spin-0 particles. Based as it is on an equation that is quadratic in energy, it leads to solutions with positive and negative energy values – this is problematic, but not insuperable. More troubling is that it also leads to un-physical negative probabilities. For these reasons both Dirac and Schrödinger kept looking.

2.3 The Dirac Equation

Dirac’s approach was to look for an equation similar to the Klein-Gordon, but one which was first order in all the derivatives. However the problem with this approach is that any solution to such an equation would also need to obey the energy-momentum relationship and therefore would need to be a solution of the Klein-Gordon equation as well. The solution to this conundrum is to seek a quantity that is in a sense the square root of the d’Alembertian or in other words an operator that, when multiplied by itself, equals the d’Alembertian. On the face of it the square root of an operator tends to defeat the imagination, nevertheless Dirac proceeded as follows:

Consider an operator ($\not{\partial}$) that when multiplied by itself equals the d’Alembertian, meaning:

$$\not{\partial}^2 = \square = \left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2} \right) \quad (10)$$

How can this be achieved? The answer is in how you define $\not{\partial}$.

First let’s define a series of mathematical objects known as γ matrices⁵. There has to be four of them: $\gamma^\mu \quad \mu = 0, 1, 2, 3$. Then define $\not{\partial}$:

$$\not{\partial} = (\gamma^0 \frac{\partial}{\partial t} + \gamma^1 \frac{\partial}{\partial x} + \gamma^2 \frac{\partial}{\partial y} + \gamma^3 \frac{\partial}{\partial z}) \quad (11)$$

If you multiply out $\not{\partial}^2$ then Equation (10) is satisfied provided that:

$$(\gamma^0)^2 = \mathbb{I} \quad (12)$$

$$(\gamma^i)^2 = -\mathbb{I} \quad i = 1, 2, 3 \quad (13)$$

$$[\gamma^i \gamma^j + \gamma^j \gamma^i] = 0 \quad i \neq j \quad (14)$$

(where \mathbb{I} is the identity matrix) and if so Equation (9) can be written as:

$$(\not{\partial}^2 + m^2)\psi = 0$$

in which case it can be factorised as

$$(\not{\partial} + im)(\not{\partial} - im)\psi = 0 \quad (15)$$

⁴This is where many authors start to say “manifestly Lorentz invariant” .

⁵Strictly at this point we don’t know they are matrices.

If we ignore the first term in brackets, which seems to imply a negative mass, then write $\not{\partial}$ as $\gamma^\mu \partial_\mu$ then we get the Dirac Equation:⁶

$$(i\gamma^\mu \partial_\mu - m)\psi = 0 \quad (16)$$

and by virtue of Equation (15) solutions, ψ , to the Dirac Equation are also solutions to the Klein–Gordon Equation.

3 Consequences of the Dirac Equation

The principal consequence of the Dirac Equation is that it brings the concept of spinors into quantum physics, which provides the mathematical underpinnings for understanding fermions (ie spin-half particles). Equations 12, 13 and 14 form what mathematicians refer to as a Clifford⁷ algebra, although it was unlikely that Dirac recognised this as Clifford’s work was not widely known even amongst mathematicians of the time.

Equation (14) is what is known as an anti-commutator, often written $\{A, B\} = 0$ meaning that $AB = -BA$. This proves to be fundamental to Quantum Physics (for example, driving us towards the exclusion principle for fermions) but in particular means that the γ quantities cannot be simple numbers and it turns out they need to be matrices of at least dimension 4×4 .

At this point we have not given explicit forms for the γ matrices, and in many cases we don’t need to provided that we respect the Clifford algebra. On a more conceptual level, it’s useful to understand what these γ quantities are, and in that respect we can return to the observation that the matrices are defined by a Clifford algebra, and as such the γ quantities can be considered as *operators* that act on *spinors* therefore the wavefunction ψ is a spinor.

So what is a spinor? It is an object that when rotated through 2π transforms into its *inverse* i.e. it takes two complete rotations to return a spinor to itself. This may seem counter-intuitive, but there are some well known and easy to reproduce examples - the plate trick is one example, but it’s also known as Dirac’s belt trick and by other names too. In the sense that a spinor has to be rotated twice to return to itself, this is what we mean by spin-half, thus the Dirac Equation is the wave equation for spin-half objects, or *fermions*.

3.1 Gamma Matrices

Although Equations 12, 13 and 14 are sufficient to define the properties of the γ matrices, they still leave plenty of room for making choices i.e. there is not a single, unique set of γ matrices. In fact I suspect a mathematician would say they they are not required to be matrices at all but could be any mathematical object that obeys the Clifford Algebra. However, given that matrices are one example, I also suspect that the same mathematician would say that any other choice would be “isomorphic” to the matrices.

⁶The slash notation is actually due to Feynmann and came along after Dirac. If you can think of the γ quantities as a contravariant vector, this explains the superscript μ .

⁷William Kingdon Clifford FRS (1845 – 1879).

Some mathematical logic tells us that if they are matrices they need to be square, but traceless and of *even* rank. They need to be distinct, ie linearly independent, which means they can't be 2×2 (we have to exclude the identity matrix as it has a non-zero trace, and there aren't enough linearly independent 2×2 matrices left). Therefore the simplest objects that are suitable are 4×4 matrices, but higher dimensional matrices are possible.

Because these matrices act on the wavefunction, this requires the wavefunction to have at least four components, and it turns out that this requirement means the wavefunctions have extra degrees of freedom which in turn leads us to deduce that the wavefunctions are describing intrinsic angular momentum, ie spin.

Given the application, it's reasonable to be guided by the physics, in which case we might want the matrices to be Hermitian (or possibly anti-Hermitian) such that they have real eigenvalues. In practice one of the more common approaches is to choose them to be unitary and with this additional requirement we can see immediately from Equation (12) that γ^0 is Hermitian and from (13) that γ^k ($k = 1, 2, 3$) is anti-Hermitian.

The most commonly used forms for γ^μ are the Dirac–Pauli matrices. In block matrix form these are:

$$\gamma^0 = \begin{bmatrix} \mathbb{I} & \mathbb{0} \\ \mathbb{0} & -\mathbb{I} \end{bmatrix} \quad \gamma^1 = \begin{bmatrix} \mathbb{0} & \sigma_1 \\ -\sigma_1 & \mathbb{0} \end{bmatrix} \quad \gamma^2 = \begin{bmatrix} \mathbb{0} & \sigma_2 \\ -\sigma_2 & \mathbb{0} \end{bmatrix} \quad \gamma^3 = \begin{bmatrix} \mathbb{0} & \sigma_3 \\ -\sigma_3 & \mathbb{0} \end{bmatrix}$$

More compactly

$$\gamma^0 = \begin{bmatrix} \mathbb{I} & \mathbb{0} \\ \mathbb{0} & -\mathbb{I} \end{bmatrix} \quad \gamma^i = \begin{bmatrix} \mathbb{0} & \sigma_i \\ -\sigma_i & \mathbb{0} \end{bmatrix} \quad i = 1, 2, 3$$

where the σ quantities are the Pauli spin matrices:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

3.2 Magnetic Moment of the Electron

The potential energy of a moving, electrically charged, particle in a magnetic field is given by the scalar product of the magnetic field strength (\mathbf{B}) and the magnetic moment ($\boldsymbol{\mu}$).

Classically the magnetic moment of a current loop is the current multiplied by the area of the loop so if the particle has charge q moving with speed v in a circle of radius r then the magnetic moment is:

$$|\boldsymbol{\mu}| = \pi r^2 \frac{qv}{2\pi r} \quad (17)$$

Meanwhile the angular momentum of the particle is $|\mathbf{L}| = mvr$, substituting this into Equation (17)

$$|\boldsymbol{\mu}| = g \frac{q}{2m} |\mathbf{L}| \quad (18)$$

where we have included g which is the gyromagnetic ratio ($g = 1$ for the classical case). If we apply this analysis to the electron which has intrinsic spin (ie angular momentum) \mathbf{S} we expect

$$\boldsymbol{\mu} = g \frac{q}{2m} \mathbf{S} \quad (19)$$

(where we have dropped the modulus symbols as both $\boldsymbol{\mu}$ and \boldsymbol{S} are vectors).

If we take the non-relativistic limit of the Dirac Equation it can be reduced to Schrödinger's and from that we can extract the potential energy due to the interaction between the magnetic field and the magnetic moment of the electron. The maths for this is a little involved and beyond the scope of this article (see [Thomson, 2013](#), Appendix B1 for one derivation), suffice to say that Dirac's Equation predicts:

$$\boldsymbol{\mu} = \frac{q}{m} \boldsymbol{S}$$

hence by comparison with Equation (19) $g = 2$. Correctly predicting the value of g in this way was one of the early successes of the Dirac Equation.

The measured value is in fact 2.00232 rounded to 5 decimal places ([Maggiore, 2005](#), Chapter 1) and the small discrepancy between the experimental value and the Dirac prediction can be accounted for in advanced quantum field theory. In fact theory and experiment now agree to 10 decimal places or better, making this one of the most accurately tested of all physical theories.

3.3 Negative Energy Solutions

Although Dirac aimed to eliminate the prediction of negative energy solutions to any wave equation, in this respect he failed with his equation. To address this he initially proposed the idea of the “Dirac Sea” wherein all the negative energy states are actually filled, which is why we don't observe them.

However, this proposal doesn't work for all types of particles and the negative energy solutions to both the Dirac and Klein–Gordon equations are now understood to represent anti-particles (this is known as the Feynman–Stückelberg interpretation).

4 Summary

This article has laid out the maths and physics that lead Dirac to derive his famous equation. The power of the equation is absolutely fundamental to quantum physics making, as it does, a number of key theoretical predictions:

- Electrons are particles with spinor wavefunctions.
- The maths of spinors is the correct treatment for spin-half particles.
- Wavefunctions that are solutions of the Dirac Equation do not lead to un-physical negative probabilities, but remain as solutions to the Klein–Gordon Equation and are therefore relativistically correct.
- Specifically the Dirac Equation correctly predicts (up to small corrections that arise from quantum field theory) the value of the gyromagnetic ratio of the electron.

References and Further Reading

Barr, G., R. Devenish, Walczak, R., and Weidberg, T. 2016. *Particle Physics in the LHC Era*. Oxford University Press.

Fleisch, Daniel. 2012. *A Student's Guide to Vectors and Tensors*. Cambridge University Press.

Gray, Norman. 2019. *A Student's Guide to General Relativity*. Cambridge University Press.

Maggiore, Michele. 2005. *A Modern Introduction to Quantum Field Theory*. Oxford University Press.

Paganini, Pascal. 2023. *Fundamentals of Particle Physics*. Cambridge University Press.

Penrose, Roger. 2004. *The Road to Reality*. BCA.

Thomson, Mark. 2013. *Modern Particle Physics*. Cambridge University Press.

Almost all books on Particle Physics cover this material, often in greater depth. Derivations of the gyromagnetic ratio from the Dirac Equation can be found in many texts, for example [Maggiore \(2005\)](#) and [Paganini \(2023\)](#). This latter Reference is similar to [Thomson \(2013\)](#) but more recent.